University of Bristol - Explore Bristol Research

General rights

Human Brain Mapping

Non-Parametric Combination and related permutation tests for neuroimaging

SCHOLARONE™
Manuscripts

Non-Parametric Combination and related permutation tests for neuroimaging

Anderson M. Winkler[a,*], Matthew A. Webster[a], Jonathan C. Brooks[b],
Irene Tracey[a], Stephen M. Smith[a], Thomas E. Nichols[a,c]

[a]*Oxford Centre for Functional* MRI *of the Brain, University of Oxford, Oxford,* UK
[b]*Clinical Research and Imaging Centre, University of Bristol, Bristol,* UK
[c]*Department of Statistics & Warwick Manufacturing Group,*
University of Warwick, Coventry, UK

Abstract

In this work we show how permutation methods can be applied to combination analyses such as those that include multiple imaging modalities, multiple data acquisitions of the same modality, or simply multiple hypotheses on the same data. Using the well-known definition of union-intersection tests and closed testing procedures, we use synchronised permutations to correct for such multiplicity of tests, allowing flexibility to integrate imaging data with different spatial resolutions, surface and/or volume-based representations of the brain, including non-imaging data. For the problem of joint inference, we propose and evaluate a modification of the recently introduced Non-Parametric Combination (NPC) methodology, such that instead of a two-phase algorithm and large data storage requirements, the inference can be performed in a single phase, with reasonable computational demands. We also evaluate, in the context of permutation tests, various combining methods that have been proposed in the past decades, and identify those that provide the best control over error rate and power across a range of situations. We show that one of these, the method of Tippett, provides a link between correction for the multiplicity of tests and their combination. Finally, we discuss how the correction can solve certain problems of multiple comparisons in one-way ANOVA designs, and how the combination is distinguished from conjunctions, even though both can be assessed using permutation tests. We also provide a common algorithm that accommodates combination and correction.

Keywords: permutation tests, non-parametric combination, multiple testing, conjunctions, general linear model

	List of abbreviations
ANOVA	Analysis of variance
CCA	Canonical correlation analysis
CVA	Canonical variates analysis
CMV	**Classical multivariate test** (e.g. MANOVA, CCA)
CTP	**Closed testing procedure**
DTI	Diffusion tensor imaging
DTP	Dual truncated product
EE	Exchangeable errors
EEG	Electroencephalography
FA	Fractional anisotropy
FMRI	Functional magnetic resonance imaging
FDR	False discovery rate
FWER	Familywise error rate
GLM	**General linear model**
ICA	Independent component analysis
ISE	Independent and symmetric errors
IQ	Intelligence quotient
IUT	**Intersection–union test**
JNH	**Joint null hypothesis**
LSD	Least significant difference
MANOVA	Multivariate analysis of variance
MANCOVA	Multivariate analysis of covariance
MD	Mean diffusivity
MRI	Magnetic resonance imaging
MTP-I	**Multiple testing problem I**
MTP-II	**Multiple testing problem II**
NPC	**Non-parametric combination**
PALM	Permutation Analysis of Linear Models
PET	Positron emission tomography
RD	Radial diffusivity
RTP	Rank truncated product
SII	Secondary somatosensory cortex
TFCE	Threshold-free cluster enhancement
TPM	Truncated product method
TS	Tail strength
TTS	Truncated tail strength
UIT	**Union–intersection test**

Those indicated **bold** are used more often throughout the paper.

2

1. Introduction

In this paper we show that permutation tests can provide a common solution to seemingly disparate problems that arise when dealing with multiple imaging measurements. These problems refer to the multiplicity of tests, and to the combination of information across multiple modalities for joint inference. We begin by describing each of these problems separately, then show how they are related, and offer a complete and generic solution that can accommodate a myriad of designs that can mix imaging and non-imaging data. We also present an algorithm that has with amenable computational demands for treating these problems.

1.1. Multiple tests — but not the usual multiplicity

Because in neuroimaging one statistical test is typically performed at each of many thousands of imaging units (e.g., voxels or vertices), the problems related to such multiplicity of tests were recognised almost as early as these techniques were developed (for pioneering examples, see Fox et al., 1988; Friston et al., 1991). There is now a comprehensive body of literature on multiple testing correction methods that include those based on the random field theory, on permutation tests, as well as on other strategies that control the familywise error rate (FWER) or the false discovery rate (FDR) (for reviews, see Nichols and Hayasaka, 2003; Nichols, 2012). However, the multiplicity of tests in neuroimaging can appear in other ways that are less explicit, and most importantly, that have not been fully appreciated or made available in software packages. In the context of the general linear model (GLM, Scheffé, 1959), these *other* multiple tests include:

A. *Multiple hypotheses in the same model:* Testing more than one hypothesis regarding a set of explanatory variables. An example is testing the effects of multiple variables, such as presence of a disease along with its duration, some clinical score, age and/or sex of the subjects, on a given imaging measurement, such as maps from functional magnetic resonance imaging (FMRI) experiments.

B. *Multiple pairwise group comparisons:* Often an initial global (omnibus) test is performed, such as an F-test in the context of analysis of variance (ANOVA), and if this test is significant, subsequent (post hoc) tests are performed to verify which pairwise difference(s) drove the global result, thus introducing a multiple comparisons problem.

3

C. *Multiple models:* Testing more than one set of explanatory variables on one given dataset, that is, assembling and testing more than one design matrix, each with its own set of regressors, which may differ across designs, and each with its own set of contrasts. An example is interrogating the effect of distinct seeds, one at a time, in a resting-state FMRI experiment; another is in an imaging genetics experiment, testing multiple candidate polymorphisms.

D. *Multiple modalities:* Testing separately, in the same study, more than one imaging modality as the response variable, such as FMRI and positron-emission tomography (PET), or different metrics from the same modality, such as various measurements from diffusion tensor imaging (DTI), as fractional anisotropy (FA), mean diffusivity (MD), or radial diffusivity (RD), or the effect of various networks identified using independent component analysis (ICA).

E. *Imaging and non-imaging:* Testing separately, in the same study, imaging and non-imaging measurements as response variables. An example is studying group effects on FMRI and on behavioural or cognitive scores, such as IQ, or disease severity scores, among countless other non-imaging measurements.

F. *Multiple processing pipelines:* Testing the same imaging modality multiple times, each time after a different processing pipeline, such as using filters with different widths for smoothing, or using different strategies for registration to a common space.

G. *Multiple multivariate analyses:* Testing more than one multivariate hypothesis with the GLM in repeated measurements designs, such as in profile analyses, in which the same data allows various different hypotheses about the relationships between explanatory and response variables.

In all these cases, the multiple tests cannot be assumed to be independent, so that the simple FWER correction using the conventional Bonferroni method risks a considerable loss in power. Modelling the degree of dependence between these tests can be a daunting task, and be suboptimal by invariably requiring the introduction of assumptions about the data, which, if at all valid, may not be sufficient. By contrast, robust, generic, multi-step procedures, which do not depend as much on assumptions, or on independence among tests, such as the Benjamini–Hochberg procedure that controls

4

the false discovery rate (FDR) (Benjamini and Hochberg, 1995; Genovese et al., 2002), do not guarantee that the spatial relationship between voxels or vertices within test is preserved when applied across *these* multiple tests, therefore being not as useful as in other settings. More specifically, the difficulty relates to correcting across various distinct imaging tests, while maintaining control across space within any given test, as opposed to controlling just within a single imaging test as commonly done. For the same reason, various multiple testing approaches that are applicable to many particular cases, can hardly be used for the problems we discuss here; extensive details on these tests can be found in Hochberg and Tamhane (1987) and in Hsu (1996).

We call the multiple tests that arise in situations as those listed above "multiple testing problem II" (MTP-II), to allow a distinction from the usual multiple testing problem due to the many voxels/vertices/faces that constitute an image, which we denote "multiple testing problem I" (MTP-I). Methods that can be used in neuroimaging for the MTP-I not always can be considered for the MTP-II, a problem that has remained largely without treatment; for two rare counter examples in which the MTP-II *was* considered, we point to the studies by Licata et al. (2013) and Abou Elseoud et al. (2014).

1.2. Combination of imaging modalities

Acquisition of multiple imaging modalities on the same subjects can allow the examination of more complex hypotheses about physiological processes, and has potential to increase power to detect group differences. Such combination of modalities can refer strictly to data acquired from different instruments (e.g., MRI, PET, EEG), or more broadly, to data acquired from the same instrument using different acquisition parameters (e.g., different MRI sequences, different PET ligands); for an overview, see Uludağ and Roebroeck (2014); Zhu et al. (2014), and for example applications, see Hayasaka et al. (2006); Thomas et al. (2015). Irrespective of which the modalities are, the options in the context of the GLM rest in testing for a single multivariate hypothesis, or in testing for a combination of multiple univariate hypotheses. Single multivariate tests encompass various classical tests, known in particular cases as multivariate analysis of variance (MANOVA), multivariate analysis of covariance (MANCOVA), or canonical correlation/variates analysis (CCA/CVA); these tests will be referred here as *classical multivariate tests*, or CMV.

The combination of multiple univariate hypotheses requires that each is analysed separately, and that these results are grouped together to test, at each voxel (or vertex, or face) a *joint null hypothesis* (JNH); in this context, the separate tests are termed *partial tests.* Different criteria to decide upon rejection of the JNH give rise to three broad categories of combined tests: (I) reject if *any* partial test is significant; (II) reject if *all* partial tests are significant; and (III) reject if some *aggregate measure* from the partial tests is significant. The first of these can be traced back to Tippett (1931), and in current terminology, could be defined as rejecting the joint null hypothesis if any partial test is rejected at the FWER level using the Šidák correction (Šidák, 1967); it also corresponds to a *union–intersection test* (UIT, Roy, 1953). The second is the *intersection–union test* (IUT, Berger, 1982), that in neuroimaging came to be known as *conjunction test* (Nichols et al., 2005). The third offers a trade-off between the two other approaches, and gives rise to a large number of possible tests, each with a different rejection region, and therefore with different sensitivity and specificity profiles; some of these tests are popular in meta-analyses, with the method of Fisher (Fisher, 1932) being one of the most popular, and new approaches are continually being developed. A summary is shown in Table 1, and a brief overview of these and yet other tests, along with bibliographic information, is in Appendix A.

Both cases — a single multivariate test or the combination of multiple univariate tests — can be assessed parametrically when the asymptotic distribution of the test statistic is known, which may sometimes be the case if various assumptions about the data are met. These generally refer to the the independence between observations and between tests, to the distribution of the error terms, and for brain imaging, to yet other assumptions regarding the relationship, across space, between the tests. However, if the observations are *exchangeable*, that is, if their joint distribution remains unchanged after shuffling, then all such assumptions can be eschewed at once, and instead, permutation tests can be performed. The p-values can then be computed for either the classical multivariate tests, or for the combi-

Table 1: *(page 7)* Various functions are available for joint inference on multiple tests. For each method, both its statistic (T) and associated p-value, P are shown. These p-values are only valid if, for each method, certain assumptions are met, particularly with respect to the independence between tests, but sometimes also with respect to underlying distributions. Under exchangeability, the p-values can be computed using permutation tests, and the formulæ in the last column are no longer necessary. The tests are shown in chronological order; see Appendix A for details and bibliographic information.

Method	Test statistic (T)	p-value (P)
Tippett	$\min(p_k)$	$1 - (1-T)^K$
Fisher	$-2\sum_{k=1}^K \ln(p_k)$	$1 - \chi^2(T;\ \nu = 2K)$
Stouffer	$\frac{1}{\sqrt{K}}\sum_{k=1}^K \Phi^{-1}(1-p_k)$	$1 - \Phi(T;\ \mu=0,\ \sigma^2=1)$
Wilkinson	$\sum_{k=1}^K I(p_k \leqslant \alpha)$	$\sum_{k=1}^K \binom{K}{T}\alpha^k(1-\alpha)^{K-k}$
Good	$\prod_{k=1}^K p_k^{w_k}$	$\sum_{k=1}^K w_k^{K-1} T^{1/w_k}\left(\prod_{i=1}^{k-1}(w_k-w_i)^{-1}\right)\left(\prod_{i=k+1}^K (w_k - w_i)^{-1}\right)$
Lancaster	$\sum_{k=1}^K w_k F_k^{-1}(1-p_k)$	$1 - G(T)$
Winer	$\sum_{k=1}^K t_{\mathrm{cdf}}^{-1}(1-p_k;\ \nu_k)\ /\ \sqrt{\sum_{k=1}^K \frac{\nu_k}{\nu_k-2}}$	$1 - \Phi(T;\ \mu=0,\ \sigma^2=1)$
Edgington	$\sum_{k=1}^K p_k$	$\sum_{j=0}^{\lfloor T\rfloor}(-1)^j \binom{K}{j}\frac{(T-j)^K}{K!}$
Mudholkar–George	$\frac{1}{\pi}\sqrt{\frac{3(5K+4)}{K(5K+2)}}\sum_{k=1}^K \ln\left(\frac{1-p_k}{p_k}\right)$	$1 - t_{\mathrm{cdf}}(T;\ \nu=5K+4)$
Darlington–Hayes	$\frac{1}{r}\sum_{k=1}^r \Phi^{-1}(1-p_{(k)})$	Computed through Monte Carlo methods. Tables are available.
Zaykin et al. (TPM)	$\prod_{k=1}^K p_k^{I(p_k\leqslant\alpha)}$	$\sum_{k=1}^K \binom{K}{k}(1-\alpha)^{K-k}\left(I(T>\alpha^k)\,\alpha^k + I(T\leqslant\alpha^k)\,T\sum_{j=0}^{k-1}\frac{(k\ln\alpha-\ln T)^j}{j!}\right)$
Dudbridge–Koeleman (RTP)	$\prod_{k=1}^r p_{(k)}$	$\binom{K}{r+1}(r+1)\int_0^1 (1-t)^{K-r-1} A(T,t,K)\,\mathrm{d}t$
Dudbridge–Koeleman (DTP)	$\max\left(\prod_{r=1} p_{(k)},\ \prod_{k=1}^K p_k^{I(p_{(k)}\leqslant\alpha)}\right)$	$\sum_{k=1}^K \binom{K}{k}(1-\alpha)^{K-k} A(T,\alpha,k) + I(r<K)\binom{K}{r+1}(r+1)\int_0^\alpha (1-t)^{K-r-1} A(T,t,K)\,\mathrm{d}t$
Taylor–Tibshirani (TS)	$\frac{1}{K}\sum_{k=1}^K \left(1 - p_{(k)}\frac{K+1}{k}\right)$	$1 - \Phi(T;\ \mu=0,\ \sigma^2\approx\frac{1}{K})$
Jiang et al. (TTS)	$\frac{1}{K}\sum_{k=1}^K I(p_{(k)}\leqslant\alpha)\left(1 - p_{(k)}\frac{K+1}{k}\right)$	Computed through Monte Carlo methods.

T is the statistic for each method and P its asymptotic p-value. All methods are shown as function of the p-values for the partial tests. For certain methods, however, the test statistic for the partial tests, if available, can be used directly. K is the number of tests being combined, p_k, $k = \{1, 2, \ldots, K\}$ are the partial p-values, w_k are positive weights assigned to the respective p_k, $p_{(r)}$ are the p_k with rank r in ascending order (most significant first), α is the significance level for the partial tests, $I(\cdot)$ is an indicator function that evaluates as 1 if the condition is satisfied, 0 otherwise, $\lfloor \cdot \rfloor$ represents the floor function, χ^2_ν is the cumulative distribution function (cdf) for a χ^2 distribution, with the ν degrees of freedom, t_{cdf} is the cdf of the Student's t distribution with degrees of freedom ν, and t_{cdf}^{-1} its inverse, Φ is the cdf of the normal distribution with mean μ and variance σ^2, and Φ^{-1} its inverse, and F and G are the cdf of arbitrary, yet well chosen distributions. For the two Dudbridge–Koeleman methods, $A(T, a, b) = I(T > a^b)\, a^b + I(T \leqslant a^b)\, T \sum_{j=0}^{b-1}(b\ln a - \ln T)^j/j!$.

7

nation of univariate tests; when used in the last case, the strategy corresponds to Pesarin's method of *non-parametric combination* (NPC, Pesarin, 1990, 2001), discussed below. Exchangeability is assumed only for the observations within each partial test (or for the errors terms of the respective models, see below); exchangeability is not assumed between the partial tests for either CMV or NPC. Moreover, non-independence does not need to be explicitly modelled, either between observations, between partial tests, or across space for imaging data, thus making such tests applicable to a wide variety of situations.

1.3. Overview of the article

We show that a single, elegant permutation solution is available for all the situations described above, addressing the *comparisons* of response variables when these can be put in comparable scale, the *correction* of p-values, via adjustment to allow exact control over FWER in the various multiple testing scenarios described above, and the *combination* of multiple imaging modalities to allow for joint inference. The *conjunction* of multiple tests is a special case in which the null hypothesis differs from that of a combination, even though it can be approached in a similar fashion; because the distinction is quite an important one, it is also discussed.

In the next section we outline the notation used throughout the paper. We then use the definition of union-intersection tests, closed testing procedures, and synchronised permutations to correct for multiple hypotheses, allowing flexibility to mix in the same framework imaging data with different spatial resolutions, surface and/or volume-based representations of the brain, and even non-imaging data. For the problem of joint inference, we propose and evaluate a modification of the NPC, such that instead of two phases and large data storage requirements, the permutation inference can be performed in a single phase, without prohibitive memory needs. We also evaluate, in the context of permutation tests, various combining methods that have been proposed in the past decades, and identify those that provide the best control over error rate and power across a range of situations. We also exemplify the potential gains in power with the reanalysis of the data from a pain study. In the Appendix, we provide a brief historical review of various combining functions, discuss criteria of consistency and admissibility, and provide an algorithm that allows combination and correction in a unified framework.

8

2. Theory

2.1. Notation and general aspects

For a given voxel (or vertex, or face), consider a multivariate GLM:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{Y}$ is the $N \times K$ matrix of observed data, with N observations of K distinct (possibly non-independent) variables, $\mathbf{X}$ is the full-rank $N \times R$ design matrix that includes explanatory variables (i.e., effects of interest and possibly nuisance effects), $\boldsymbol{\beta}$ is the $R \times K$ matrix of R regression coefficients for each of the K variables, and $\boldsymbol{\epsilon}$ is the $N \times K$ array of random errors. Estimates for $\boldsymbol{\beta}$ can be computed by ordinary least squares, i.e., $\hat{\boldsymbol{\beta}} = \mathbf{X}^+\mathbf{Y}$, where the superscript $(^+)$ denotes a pseudo-inverse. One generally wants to test the null hypothesis that a given combination (contrast) of the elements in $\boldsymbol{\beta}$ equals to zero, that is, $\mathcal{H}^0 : \mathbf{C}'\boldsymbol{\beta}\mathbf{D} = \mathbf{0}$, where $\mathbf{C}$ is a $R \times S$ full-rank matrix of S contrasts of coefficients on the regressors encoded in $\mathbf{X}$, $1 \leqslant S \leqslant R$ and $\mathbf{D}$ is a $K \times Q$ full-rank matrix of Q contrasts of coefficients on the dependent, response variables in $\mathbf{Y}$, $1 \leqslant Q \leqslant K$. Often more than one such standard multivariate hypothesis is tested, each regarding different aspects of the same data, and each using a different pair of contrasts $\mathbf{C}$ and $\mathbf{D}$. Not uncommonly, even different sets of explanatory variables are considered, sometimes arranged in entirely different designs. We denote the set of such design matrices as $\mathcal{X} = \{\mathbf{X}\}$, the set of pairs of contrasts for each hypothesis related to that design as $\mathcal{C}_{\mathbf{X}} = \{(\mathbf{C}, \mathbf{D})\}$, and the set of sets of such contrasts as $\{\mathcal{C}_{\mathbf{X}}\}$.

Depending on the values of K, Q, and S, $\mathcal{H}^0$ can be tested using various common statistics. If $K = 1$, or if $K > 1$ and $Q = 1$, the problem reduces to the univariate case, in which a t statistic can be used if $S = 1$, or an F-statistic if $S \geqslant 1$. If $K > 1$ and $Q > 1$, the problem is a multivariate proper and can be approached via CMV when respective multivariate Gaussian assumptions are satisfied; in these cases, if $S = 1$, the Hotelling's T^2 statistic can be used (Hotelling, 1931), whereas if $S > 1$, various other statistics are available, such as the Wilks' λ (Wilks, 1932), the Lawley–Hotelling's trace (Lawley, 1938; Hotelling, 1951), the Roy's largest root(s) (Roy, 1953; Kuhfeld, 1986), and the Pillai's trace (Pillai, 1955); the merits of each in the parametric case are discussed in various textbooks (e.g., Christensen, 2001; Timm, 2002; Anderson, 2003; Johnson and Wichern, 2007), and such tests have been applied to neuroimaging applications (Chen et al., 2014).

The model in Equation 1 can be rewritten as $\tilde{\mathbf{Y}} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\epsilon}}$, where $\tilde{\mathbf{Y}} = \mathbf{YD}$, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\mathbf{D}$ and $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}\mathbf{D}$. If $Q = 1$, this is a univariate model, otherwise it remains multivariate, with $\tilde{\mathbf{Y}}$ having $\tilde{K} = Q$ columns, and the null hypothesis simplified as $\mathcal{H}^0 : \mathbf{C}'\tilde{\boldsymbol{\beta}} = \mathbf{0}$. This null is equivalent to the original, and can be split into multiple partial hypotheses $\mathcal{H}_{\tilde{k}}^0 : \mathbf{C}'\tilde{\boldsymbol{\beta}}_{\tilde{k}} = \mathbf{0}$, where $\tilde{\boldsymbol{\beta}}_{\tilde{k}}$ is the $\tilde{k}$-th column of $\tilde{\boldsymbol{\beta}}$, $\tilde{k} = 1, \ldots, \tilde{K}$. This transformation is useful as it defines a set of separate, even if not independent, partial hypotheses, that can be tested and interpreted separately. We drop heretofore the "$\sim$" symbol, with the modified model always implied.

Non-parametric inference for these tests can be obtained via permutations, by means of shuffling the data, the model, the residuals, or variants of these, in a direct extension from the univariate case (Winkler et al., 2014, Table 2). To allow such rearrangements, some assumptions need to be made: either of *exchangeable errors* (EE) or of *independent and symmetric errors* (ISE). The first allows permutations, the second sign flippings; if both are available for a given model, permutations and sign flippings can be performed together. We use generically the terms *rearrangement* or *shuffling* when the distinction between permutations or sign flippings is not pertinent. These are represented by permutation and/or sign flipping matrices $\mathbf{P}_j$, $j = 1, \ldots, J$, where J is the number of such rearrangements.

Another aspect that concerns permutation tests refers to the use of statistics that are *pivotal*, i.e., that have sampling distributions that do not depend on unknown parameters. Most statistics used with parametric tests (and all the uni- and multivariate examples from the previous paragraph) are pivotal if certain assumptions are met, especially homoscedasticity. Their benefits in non-parametric tests are well known (Hall and Wilson, 1991), and for neuroimaging, pivotal statistics are useful to allow exact correction for the MTP-I.

2.2. Union–intersection and intersection–union tests

Consider the set of p-values $\{p_k\}$ for testing the respective set of partial null hypotheses $\{\mathcal{H}_k^0\}$. A union–intersection test (UIT, Roy, 1953) considers the JNH corresponding to a *global null hypothesis* that *all* $\mathcal{H}_k^0$ are true; if any such partial null is rejected, the global null hypothesis is also rejected. An intersection–union test (IUT, Berger, 1982) considers the JNH corresponding to a *conjunction null hypothesis* (also termed *disjunction of null hypotheses*) that *any* $\mathcal{H}_k^0$ is true; if all partial nulls are rejected, the conjunction null hypothesis is also rejected. In the UIT, the null is the intersection of

Table 2: Joint hypotheses tested with union–intersection and intersection–union of K partial tests. In the UIT, the null is also called *global null hypothesis*, whereas in the IUT, the null is also called *conjunction null hypothesis*.

	UIT	IUT
Null hypothesis ($\mathcal{H}^0$)	$\bigcap\limits_{k=1}^{K} \mathcal{H}_k^0$	$\bigcup\limits_{k=1}^{K} \mathcal{H}_k^0$
Alternative hypothesis ($\mathcal{H}^1$)	$\bigcup\limits_{k=1}^{K} \mathcal{H}_k^1$	$\bigcap\limits_{k=1}^{K} \mathcal{H}_k^1$

the null hypotheses for all partial tests; the alternative is the union of the alternatives. In the IUT, the null is the union of the null hypotheses for all partial tests; the alternative is the intersection of the alternatives. A UIT is significant if the smallest p_k is significant, whereas an IUT is significant if the largest p_k is significant. Figure 1 illustrates the rejection regions for UIT and IUT cases based on two independent t-tests, in which the statistic larger than a certain critical level is considered significant. Table 2 shows the null and alternative hypotheses for each case.

Enlarging the number of tests affects UITs and IUTs differently. For the UIT with a given statistic threshold, more tests increase the chances of false positives, and correction for this multiplicity needs to be applied. In fact, it can be shown that a UIT at a significance level α is equivalent to controlling the FWER at α for the same tests. In other words, a union-intersection procedure is an FWER procedure. For an IUT, in contrast, the procedure does not change with more tests. The conjunction null hypothesis is *composite*, consisting of different parameter settings. For the extreme case that exactly one partial null is true and $K - 1$ effects are real, an IUT is exact for any K; if two or more more partial nulls are true, an IUT becomes increasingly conservative with larger K.

The null hypothesis of the UIT can be rejected if the smallest p_k is significant or, equivalently, its corresponding statistic, that is, the extremum statistic. For tests in which larger statistics provide evidence against the null hypothesis, the relevant extremum is the maximum. Conversely, for tests in which smaller statistics provide evidence against the null, the extremum is

11

the minimum. Clearly, if the most extreme statistic is significant, at least one partial hypothesis is rejected, therefore the global null hypothesis can be rejected without the need to continue testing the other $K-1$ partial hypotheses. The null hypothesis of the IUT can be rejected if the largest p_k is significant or, equivalently, its corresponding least extreme statistic. Clearly, if the least extreme statistic is significant, all partial hypotheses can be rejected, therefore the conjunction hypothesis can be rejected without the need to continue testing all other $K-1$ partial hypotheses.

In brain imaging, the term *conjunction* refers to a test performed when one wants to localise regions where there is signal in all partial tests, that is, a logical AND of all alternative hypotheses (Nichols et al., 2005), and is synonymous with the IUT. In noting the lack of power of such a proper conjunction test, Friston et al. (2005) suggested a *partial conjunction*, in which fewer than all alternatives need to intersect. Using the same notation of Table 1, both approaches have the same statistic, $T = \max(p_k)$, but the p-value of the latter can be computed as T^{K-v+1}, so that the test is a conjunction of at least v alternative hypotheses; if $v = K$, it is an IUT, and if $v = 1$ the null is equivalent to that of a UIT (such a test, however, is inconsistent for a UIT; see Appendix B). Benjamini and Heller (2008) further generalised the procedure by allowing the combination of the largest p-values using any of various possible combining functions, such as those we present in Table 1 and in Appendix A.

2.3. Closed testing

In a *closed testing procedure* (CTP), each $\mathcal{H}_k^0$ is rejected if, and only if, it is significant in its own right at a certain level α, and if all possible sub-JNHs that include the same $\mathcal{H}_k^0$ and comprise some or all of the partial hypotheses (that is, subsets of the global JNH formed by some of the partial tests) are also rejected at α using a suitable test. Various such tests can be considered, including CMVs and NPC (next section).

A CTP guarantees strong control over FWER (Marcus et al., 1976). To produce adjusted p-values, the original method requires that all $2^K - 1$ sub-JNHs are tested[1], a requirement that is computationally onerous, even for a moderate number of tests, a problem aggravated by the large number of tests that are considered in an imaging experiment. There exists, however, a particular test for the sub-JNHs that obviates the need for such a gargantuan

---

[1]From the Pascal triangle: $\sum_{i=1}^{K} \binom{K}{i} = 2^K - 1$.

12

computational venture: the union–intersection test. In a UIT using the extremum statistic, the most extreme of the global JNH that comprises all the K partial tests is also the most extreme of any other sub-JNH that includes that particular partial hypothesis, such that the other joint subtests can be bypassed altogether. As a UIT is also an FWER-controlling procedure, this raises various possibilities for correction of both MTP-I and MTP-II. While such a shortcut can be considered for both parametric (Holm, 1979) and non-parametric cases (Westfall and Young, 1993), for the non-parametric methods using permutation, one additional feature is needed: that the joint sampling distribution of the statistic used to test each of the sub-JNH is the same regardless whether the null is true for all the K partial tests, or just some of them. This property is called *subset pivotality* (Westfall and Young, 1993; Westfall and Troendle, 2008), and it constitutes the multivariate counterpart to the univariate pivotality.

2.4. Non-parametric combination

The NPC consists of testing each of the $\mathcal{H}_k^0$ using shufflings that are performed synchronously for all K partial tests. The resulting statistics for each permutation are recorded, allowing an estimate of the complete empirical null distribution to be constructed for each partial test. In a second stage, the empirical p-values for each statistic are combined, for each permutation, into a *joint statistic*. As such a combined joint statistic is produced from the previous permutations, an estimate of its empirical distribution function is immediately known, and so the p-value of the unpermuted statistic, hence of the joint test, can be assessed. The method was proposed by Pesarin (1990, 1992), and independently, though less generically, by Blair et al. (1994); a thorough description is available in Pesarin (2001) and Pesarin and Salmaso (2010b). An early application to brain imaging can be found in Hayasaka et al. (2006), its use to combine different statistics within the same modality in Hayasaka and Nichols (2004), and a summary description and practical examples are presented in Brombin et al. (2013). The JNH of the combined test is that all partial null hypotheses are true, and the alternative that any is false, which is the same null of a UIT, although the rejection region may differ widely from the example in Figure 1a, depending on the combining function.

The only two requirements for the validity of the NPC are that the partial test statistics have the same direction suggesting the rejection of the null hypothesis, and that they are consistent (see Appendix B). For the com-

13

bining function, it is desirable that (I) it is non-decreasing with respect to all its arguments (which are the p-values p_k, or $1 - p_k$, depending on the combining function), (II) that it approaches its maximum (or minimum, depending on the function) when at least one of the partial tests approaches maximum significance (that is, when at least one p-value approaches zero), and (III) that for a test level $\alpha > 0$, the critical significance threshold is smaller than the function maximum value. These requirements are easily satisfied by almost all functions shown in Table 1, which therefore can be used as combining functions in the framework of NPC (see Appendix B for a discussion on the few exceptions).

One of the most remarkable features of NPC is that the synchronised permutations implicitly account for the dependence structure among the partial tests. This means that even combining methods originally derived under an assumption of independence, such as Tippett or Fisher, can be used even when independence is untenable. In fact, modifications to these procedures to account for non-independence (e.g., Brown, 1975; Kost and McDermott, 2002, for the Fisher method) are made redundant. As the p-values are assessed via permutations, distributional restrictions are likewise not necessary, rendering the NPC free of most assumptions that thwart parametric methods in general. This is why NPC methods are an alternative to CMV tests, as each of the response variables in a MANOVA or MANCOVA analysis can be seen as an univariate partial test in the context of the combination.

2.5. Transformation of the statistics

While NPC offers flexibility in a simple and uncomplicated formulation, its implementation for brain imaging applications poses certain challenges. Because the statistics for all partial tests for all permutations need to be recorded, enormous amounts of data storage space may be necessary, a problem further aggravated when more recent, high resolution imaging methods are considered. Even if storage space were not a problem, however, the discreteness of the p-values for the partial tests becomes problematic when correcting for multiple testing, because with thousands of tests in an image, ties are very likely to occur among the p-values, further causing ties among the combined statistics. If too many tests across an image share the same most extreme statistic, correction for the MTP-I, while still valid, becomes less powerful (Westfall and Young, 1993; Pantazis et al., 2005). The most obvious workaround — run an ever larger number of permutations to break the ties — may not be possible for small sample sizes, or when possible,

14

requires correspondingly larger data storage.

However, another possible approach can be considered after examining the two requirements for the partial tests, and also the desirable properties (I)–(III) of the combining functions, all listed earlier. These requirements and properties are quite mild, and if the sample size is reasonably large and the test statistics homogeneous, i.e., they share the same asymptotic permutation distribution, a *direct combination* based not on the p-values, but on the statistics themselves, such as their sum, can be considered (Pesarin and Salmaso, 2010b, page 131). Sums of statistics are indeed present in combining functions such as of Stouffer, Lancaster, Winer, and Darlington–Hayes, but not others listed in Table 1 and Appendix A. In order to use these other combining functions, most of them based on p-values for the partial tests, and under the same premises, the statistics need to be transformed to quantities that behave as p-values. In the parametric case, these would be the parametric p-values, computed from the parametric cumulative distribution function (cdf) of the test statistic. If the parametric assumptions are all met for the partial tests, their respective parametric p-values are all valid and exact; if the assumptions are not met, these values are no longer appropriate for inference on the partial tests, but may still be valid for NPC, for satisfying all requirements and desirable properties of the combining functions. As they are not guaranteed to be appropriate for inference on the partial tests, to avoid confusion, we call these parametric p-values "u-values".

Another reason for not treating u-values as valid p-values is that they do not necessarily need to be obtained via an assumed, parametric cumulative distribution function for the statistics of the partial tests. If appropriate, other transformations applied to the statistics for the partial tests can be considered; whichever is more accurate to yield values in the interval $[0; 1]$ can be used. The interpretation of a u-value should not be that of a probability, but merely of a monotonic, deterministic transformation of the statistic of a partial test, so that it conforms to the needs of the combining functions.

Transformation of the statistic to produce quantities that can be used in place of the non-parametric p-values effectively simplifies the NPC algorithm, greatly reducing the data storage requirements and computational overhead, and avoiding the losses in power induced by the discreteness of p-values. This simplification is shown in Figure 2, alongside the original NPC algorithm.

Regardless of the above transformation, the distribution of the combined statistic, T, may vary greatly depending on the combining function, and it is always assessed non-parametrically, via permutations. Different

15

distributions for different combining functions can, however, pose practical difficulties when computing spatial statistics such as cluster extent, cluster mass, and even threshold-free cluster enhancement (TFCE, Smith and Nichols, 2009). Consider for instance the threshold used to define clusters: prescribed values such as 2.3 or 3.1 (Woo et al., 2014) relate to the normal distribution and are not necessarily sensible choices for combining functions such as Tippett or Fisher. Moreover, for some combining functions, such as Tippett and Edgington, smaller values for the statistic are evidence towards the rejection of the null, as opposed to larger as with most of the others. To address these practical issues, a monotonic transformation can be applied to the combined statistic, so that its behaviour becomes more similar to, for instance, the z-statistic (Efron, 2004). This can be done again by resorting to the asymptotic behaviour of the tests: the combined statistic is converted to a parametric p-value (the formulas are summarised in Table 1), which, although not valid for inference unless certain assumptions are met, particularly with respect to the independence among the partial tests, are useful to transform, at each permutation, the combined statistic to the z-statistic, which can then be used for inference using cluster extent, mass, or TFCE.

2.6. Directed, non-directed, and concordant hypotheses

When the partial hypotheses are one-sided, i.e., $\mathcal{H}_k^0 : \mathbf{C}'\boldsymbol{\beta}_k > 0$ or $\mathcal{H}_k^0 : \mathbf{C}'\boldsymbol{\beta}_k < 0$, and all have the same direction (either), the methods presented thus far can be used as described. If not all have the same direction, a subset of the tests can be scaled by -1 to ensure a common direction for all.

If the direction is not relevant, but the concordance of signs towards one of them (either) is, a new combining test can be constructed using one-sided p-values, p_k, and another using $1 - p_k$, then taking the best of these two results after correcting for the fact that two tests were performed. For example, for the Fisher method, we would have:

$$T = \max \left(-2 \sum_{k=1}^{K} \ln (p_k), -2 \sum_{k=1}^{K} \ln (1 - p_k) \right) \tag{2}$$

where T is the combined test statistic, with its p-value, P, assessed via permutations.

If direction or concordance of the signs are not relevant, two-sided (non-directed) tests and p-values can be used before combining, that is, ignoring the sign of the test statistic for the partial tests, or using a statistic that is non-directional (e.g., with F-tests for the partial hypotheses). It worth

16

mentioning, however, that it is not appropriate to simultaneously ignore directions of the partial tests *and* use a combination that favours concordant signs. Such a test would lack meaning and would be inadmissible, with examples shown in Appendix C.

Rejection regions for these three cases, for four different combining functions, are shown in Figure 3, as functions of the partial p-values, for $K = 2$ partial tests.

2.7. The method of Tippett

From the various combining functions listed in Table 1, consider the combining function of Tippett (1931), that has statistic $T = \min p_k$ and, when all partial tests are independent, a p-value $P = 1 - (1 - T)^K$. This test has interesting properties that render it particularly attractive for imaging:

– It defines a UIT test: If the minimum p-value remains significant when all tests are considered, clearly the global null hypothesis can be rejected.

– It controls the FWER: Controlling the error rate of a UIT is equivalent to an FWER-controlling procedure over the partial tests.

– If the partial tests are independent, it defines an exact FWER threshold: The function is closely related to Šidák (1967) correction: set $P = \alpha^{\text{FWER}}$, then $T^{\text{FWER}} = 1 - (1 - \alpha^{\text{FWER}})^{\frac{1}{K}}$; one can retain only the partial p-values that satisfy $p_k \leqslant T^{\text{FWER}}$. Adjusted p-values can be obtained similarly through the Šidák procedure, that is $p_k^{\text{FWER}} = 1 - (1 - p_k)^{\frac{1}{K}}$.

– If the partial tests are not independent, it still defines an FWER threshold and adjusted p-values: As a UIT, the Tippett function can be used in a closed testing procedure. Further, it is *the* function that makes CTP with large K feasible in practice; adjusted p-values are obtained with the distribution of the minimum p-value (or of the extremum statistic).

– Because it subsumes correction using the extremum statistic that is already in use in imaging to account for MTP-I, the correction for the MTP-II can be done by pooling the maximum statistics across both space and the set of partial tests. This allows algorithmic advantages that we exploit in the proposed implementation shown in Appendix D.

– It can be used as the combining function with NPC, thus providing a common procedure for correction and for combination of p-values.

17

– It is fast to compute: Taking the extremum statistic or minimum p-value is trivial compared to other functions that require cumulative sums or products, multiple parameters, integrations, or that depend on Monte Carlo simulations.

While the Tippett function is advantageous for all these reasons, note that, even when other combining functions are used for NPC, the extremal statistic (equivalent to the Tippett combining function) is also used for the MTP-I to control FWER over space.

2.8. A unified procedure

Armed with these concepts, and with the modifications to the original NPC algorithm, we are positioned to tackle the various problems identified in the Introduction:

Combination of multiple modalities. With K modalities, all in register and with the same spatial resolution, each is tested separately, using synchronised permutations, and their statistics converted to u-values for each shuffling. These are are combined using a suitable combining function, such as one from those shown in Table 1. The p-values for the combined statistic are produced using the same set of permutations used to assess each test separately. This is the modified NPC algorithm that we propose, shown in Figure 2.

Correction for multiple modalities. With K modalities, which are not necessarily in register, nor with the same resolution, nor of the same type (e.g., some from volumetric, some from surface representations of the brain), or which may not necessarily be all related to imaging (e.g., some imaging and some non-imaging data), each is tested separately using a suitable test statistic. The permutation distribution of the extremum statistic across *all* tests is produced and used to compute FWER-adjusted p-values that simultaneously address the MTP-I and MTP-II.

Correction for multiple designs and contrasts. Each pair of contrasts defined by $(\mathbf{C}, \mathbf{D})$ allows the corresponding design matrix to be partitioned into effects of interest and nuisance effects (Winkler et al., 2014, Appendix A), and also the redefinition of the response variables (Section 2.1). Thus, multiple designs and their respective contrasts can be tested separately. Differently than for the correction for multiple modalities, however, with different contrasts, their respective statistics may possess different asymptotic behaviour

18

(due to, e.g., the contrasts having different ranks, or the designs having different degrees of freedom), thus precluding the use of the distribution of the extremum statistic. When known, the asymptotic behaviour can be used to convert these statistics — univariate or multivariate — to a z-statistic. The distribution of the maximum across the results of the various designs and contrasts can then be computed and used for correction.

Correction for multiple modalities, designs and contrasts. Following the same principles, it is also possible to account for the multiplicity of input modalities, each tested with their respective design and set of contrasts, or each tested versus all designs and contrasts. Each test is applied separately, statistics converted to a z-statistic based on their asymptotic behaviour, and the distribution of the extremum used to obtain adjusted p-values for all in a CTP using a UIT. It is not necessary that all are in register, neither that all use the same kind of image representation of the brain (i.e., volume or surface), nor that they are even all (or any) imaging-related, and can therefore include clinical or behavioural, biomarkers, and other types of data.

Conjunctions. An IUT can be assessed through permutations simply by computing $\max(p_k)$, which is, in its own right, the p-value of the IUT, such that there is no need for transformation into u-values for the assessment of the combined statistic. In the context of imaging, such conjunctions can be used with statistics at every voxel (or vertex or face), thus allowing also certain spatial statistics such as TFCE.

Since combinations and conjunctions are performed at each individual image point, it is necessary that all images have been registered to the same common space and possess similar spatial resolution (Lazar et al., 2002). This can be accomplished through intra-subject and inter-subject registration and resampling. By contrast, correction for the multiplicity of tests uses the maximum statistic across such tests, thus not requiring that the tests match on space, or even that they are all related to imaging. However, they explicitly require pivotal statistics (for pivotality in this context, see Winkler et al., 2014), so that the extreme is taken from statistics that share the same sampling distribution. The statistics used with CMV and NPC are all pivotal and therefore can be used. Spatial statistics, however, lack this property and require similar search volumes and resolutions, even for correction. Moreover, by including information from neighbouring voxels, such as using spatial smoothing or spatial statistics like TFCE (Smith and Nichols,

19

2009), subset pivotality is lost, meaning that strong control of FWER cannot be guaranteed. In practice, though, the power gained by pooling information over space is essential. In the Appendix D we provide an algorithm that generically implements the combination and correction methods presented.

3. Evaluation methods

3.1. Validity of the modified NPC

To assess the validity of the proposed modification to the NPC, we consider one of the simplest scenarios that would have potential to invalidate the method and reduce power: this is the case of having a small number of partial tests, small sample size, and with each partial test possessing substantially different distributions for the error terms. We investigated such a scenario with $K = 2$, varying sample sizes $N = \{8, 12, 20, 30, 40, 50, 60, 70, 80, 120, 200\}$, and different error distributions. Using the notation defined in Section 2.1, response variables were generated for each simulation using the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbf{Y}$ sized $N \times K$. Each modality was simulated as having 500 points, these representing, for instance, voxels or vertices of an image representation of the brain. The errors, $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2]$, were simulated following either a Gaussian distribution with zero mean and unit variance, or a Weibull distribution (skewed), with scale parameter 1 and shape parameter $1/3$, shifted and scaled so as to have expected zero mean and unit variance. Different combinations of error distributions were used: Gaussian for both partial tests, Weibull for both partial tests, or Gaussian for the first, and Weibull for the second partial test.

The response data, $\mathbf{Y}$, were constructed by adding the simulated effects, $\mathbf{X}\boldsymbol{\beta}$, to the simulated errors, where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$, with $\boldsymbol{\beta}_k = [\beta_1, 0]'$, β_1 being either 0 (no signal) or $t_{\text{cdf}}^{-1}(1 - \alpha; N - \text{rank}(\mathbf{X})) / \sqrt{N}$ (with signal), where $\alpha = 0.05$ is the significance level of the permutation test to be performed. This procedure ensures a calibrated signal strength sufficient to yield an approximate power of 50% for each partial test, with Gaussian errors, irrespective of the sample size; for non-Gaussian errors this procedure does not guarantee power at the same level. The actual effect was coded in the first regressor of $\mathbf{X}$, constructed as a vector of random values following a Gaussian distribution with zero mean and unit variance; the second regressor was modelled an intercept. All four possible combinations of presence/absence of effect among the $K = 2$ partial tests were simulated, that is, (1) with no signal in any of the two partial tests, (2) with signal in the first partial test

20

only, (3) with signal in the second partial test only, and (4) with signal in both partial tests.

The simulated data was tested using the Tippett and Fisher methods. The case with complete absence of signal was used to assess error rates, and the others to assess power. The p-values were computed with 500 permutations, and the whole process was repeated 500 times, allowing histograms of p-values to be constructed, as well as to estimate the variability around the heights of the histogram bars. Confidence intervals (95%) were computed for the empirical error rates and power using the Wilson method (Wilson, 1927). The p-values were also compared using Bland–Altman plots (Bland and Altman, 1986), modified so as to include the confidence intervals around the means of the methods.

3.2. Performance of combined tests

We also took the opportunity to compare the combining functions shown in Table 1. While other comparisons have been made in the past (for a list of references, see Appendix A), none included *all* these functions, nor explored their performance under permutation or NPC, and therefore, did not consider the modifications that we introduce to the procedure to render it feasible for imaging applications. In addition, we investigate the performance of two classical multivariate tests, the Hotelling's T^2, and the Wilks' λ, both assessed through permutations.

Four different simulation sets were conducted, named A–D; in all, the number of partial tests being combined could vary in the range $K = 2, \ldots, 16$, and the number of partial tests containing true, synthetic signal could vary in the range $K_s = 0, \ldots, K$. In simulation A, K varied, while K_s was held fixed at 0, that is, no synthetic signal was added. In simulation B, K varied, while K_s was held fixed at 1, that is, just one partial test had signal added. In simulation C, K was held fixed at 16, while K_s varied. Finally, in simulation D, K varied, and K_s was set as equal to K, that is, all partial tests had synthetic signal added. Figure 4 shows graphically how K and K_s varied in each simulation.

The response variables **Y** had size $N \times K$, $N = 20$, that is, simulating measurements for 20 subjects, each with K image modalities (partial tests). Each modality was simulated as having 500 points, these representing, for instance, voxels or vertices. The errors were simulated following either a Gaussian distribution with zero mean and unit variance, or a Weibull distribution, with scale parameter 1 and shape parameter 1/3, shifted and scaled

21

so as to have expected zero mean and unit variance. The response data were constructed by adding to the errors the simulated effects — either no signal, or a signal with strength callibrated to yield an approximate power of 50% with Gaussian errors, irrespective of the sample size, as described above for the simulations that tested the validity of the modified NPC; for the Weibull errors, the signal was further decreased, in all these four simulations, by a factor $5/8$, thus minimising saturation at maximum power in simulation D. The actual effect was coded in the first regressor only, which was constructed as a set of random values following a Gaussian distribution with zero mean and unit variance; the second regressor was modelled as an intercept.

The simulated data was tested using 500 shufflings (permutations, sign-flippings, and permutations with sign-flippings). For all the simulations, the whole process was repeated 100 times, allowing histograms of p-values to be constructed, as well as to estimate the variability around the heights of the histogram bars. Confidence intervals (95%) were computed for the empirical error rates and power using the Wilson method.

3.3. Example: Pain study

While the proposed correction for the MTP-II has a predictable consequence, that is, controlling the familywise error rate at the nominal level, the combination of modalities, designs, and contrasts may not be quite as obvious. In this section we show a re-analysis of the data of the pain study by Brooks et al. (2005). In brief, subjects received, in separate tests, painful, hot stimuli in the right side of the face (just below the lower lip), dorsum of the right hand, and dorsum of the right foot. The objective was to investigate somatotopic organisation of the pain response in the insular cortex using FMRI, and the complete experimental details, stimulation and imaging acquisition protocols, analysis and conclusions can be found in the original publication. Here we sought to identify, at the group level, in standard space, areas within the insula that jointly respond to hot painful stimuli across the three topologically distinct body regions. We used the modified NPC, comparing the combining functions of Tippett, Fisher, Stouffer and Mudholkar–George, as well as the Hotelling's T^2 statistic, and an IUT (conjunction). At the group level, the design is a one-sample t-test, for which only sign flippings can be used to test the null hypothesis. We used twelve of the original subjects, and performed exhaustively all the 4096 sign flippings possible.

22

4. Results

A large number of plots and tables were produced and are shown in the Supplementary Material. The Figures below contain only the most representative results, that are sufficient to highlight the major points.

4.1. Validity of the modified NPC

Both the original and the modified NPC methods controlled the error rates at exactly the level of the test. Such validity was not limited to $\alpha = 0.05$, and the histograms of uncorrected p-values under complete absence of signal were flat throughout the whole $[0, 1]$ interval for both the original and modified NPC methods, using either the Tippett or the Fisher combining functions. A representative subset of the results, for the Fisher method only, and for sample sizes $N = \{8, 12, 20, 40\}$, is shown in Figure 5.

When considering the uncorrected p-values, the modified NPC yielded a mostly negligible increase in power when compared to the original NPC, with the difference always within the 95% confidence interval. Although this slight gain can be hardly observed in the histograms and Bland–Altman plots for the uncorrected p-values, they are clearly visible in the Bland–Altman plots for the p-values corrected across the 500 tests. In these plots, the predominance of smaller (towards more significant) p-values can be seen as a positive difference between the original and modified NPC p-values. A representative subset of the results is shown in Figure 6.

4.2. Performance of combined tests

Representative results demonstrating the performance of the methods of Tippett, Fisher, Stouffer, Mudholkar–George, as well as Hotelling's T^2, is shown in Figure 7. The remaining results are browseable in the Supplementary Material. In the absence of signal (simulation A), all combining functions controlled the error rate at the level of the test or below it, never above, thus confirming their validity. With normally distributed (Gaussian) errors, most functions yielded uniformly distributed p-values, although some functions seemed to converge towards uniformity only as the number of partial tests is increased; this was the case for the methods of Wilkinson, Zaykin, Dudbridge–Koeleman (DTP) and Jiang. With skewed (Weibullian) errors, the error rate was controlled at the test level with the use of permutations; with sign-flippings or permutations with sign-flippings, the combined results tended to be conservative, and more so for the Hotelling's T^2 statistics (and likewise the Wilks' λ).

23

With signal added to just one of the partial tests (simulation B), the method of Tippett was generally the most powerful, followed by the methods of Fisher and Dudbridge–Koeleman (both RTP and DTP variants). As the number of tests was increased, predictably, the power was reduced for all tests. The method of Stouffer did not in general have good performance with skewed errors, presumably because the dependence on z-statistics strengthens the dependence on the assumption of normality of the statistics for the partial tests in the modified NPC. The CMV did not deliver a good performance either, being generally among the least powerful.

With the number of partial tests held fixed, as the number of tests with signal was increased (simulation C), the power of the method of Fisher increased more quickly than of the other methods, although when most of the partial tests had signal, most of the combining functions reached similar power, all close to 100% for both normal or skewed errors. Hotelling's T^2 test was the considerably less powerful than any of the combining functions used with the modified NPC.

As the total number of partial tests and the number of partial tests with signal were both increased (simulation D), almost all combined tests had similar power, and reached saturation (100% power) quickly, particularly for the Weibullian errors, in which the calibration, even after reduction with the $5/8$ factor, yielded power above 50% for each partial test. With Gaussian errors, in which calibration ensured average 50% power, two tests had considerably lower sensitivity: Tippett's and Hotelling's T^2, the last with the remarkable result that power reached a peak, then began to fall as the number of tests kept increasing.

4.3. Example: Pain study

Using a conventional, mass univariate voxelwise tests, assessed through sign flippings, and after correction for multiple testing (MTP-I), only a few, sparse voxels could be identified at the group level for face, hand, and foot stimulation separately, in all cases with multiple distinct foci of activity observed bilaterally in the anterior and posterior insula. However, the joint analysis using the modified NPC with Fisher, Stouffer and Mudholkar–George evidenced robust activity in the anterior insula bilaterally, posterior insula, secondary somatosensory cortex (SII), and a small focus of activity in the midbrain, in the periaqueductal gray area. The combining function of Tippett, however, did not identify these regions, presumably because this method is less sensitive than the others when signal is present in more than

24

a single partial test, as suggested by the findings in the previous section.

The Hotelling's T^2 was not able to identify these regions, with almost negligible, sparse, single-voxel findings in the anterior insula, bilaterally. The conjunction test, that has a different JNH, and searches for areas where all partial tests are significant, identified a single, barely visible, isolated voxel in the right anterior insula.

The above results are shown in Figure 8. Cluster-level maps that can directly be compared to the original findings of Brooks et al. (2005) are shown in the Supplementary Material.

5. Discussion

5.1. Validity of the modified NPC

The modified NPC combines u-values, which are simply parametric p-values here renamed to avoid confusion. The renaming, however, emphasises the fact that the conversion to u-values via a parametric approximation should only be seen as a data transformation, in which the interpretation as a p-value is not preserved due to untenable assumptions. The combination method continues to be non-parametric as the combined statistic is assessed non-parametrically. More importantly, irrespective of the validity of parametric assumptions, any dependence between the tests is accounted for, implicitly, by the combination procedure, without the need of any modelling that could, at best, introduce complex and perhaps untenable assumptions, and at worst, be completely intractable.

The results suggest that, even in the cases in which the modified NPC could have failed, i.e., with small sample sizes and different distributions, the combined statistic controlled the error rate at the level of the test. This control, maintained even in such difficult scenarios, suggests that the modified NPC controls the error rates in general. The results also suggest that the modification increases power, even if such increase is minute in some scenarios. The Bland–Altman plots indicate that gains in sensitivity are more pronounced in the results corrected for the MTP-I, suggesting that the modified method is appropriate not merely due to its expediency for imaging applications, but also for having increased sensitivity compared to the original NPC.

5.2. Performance of combined tests

The results also demonstrate that the NPC method is more powerful than the Hotelling's T^2. The superiority of combined permutation tests when

25

compared to classical multivariate tests has been observed in the literature (Blair et al., 1994), and the fact that power increases as the number of partial tests with signal increases is one of its most remarkable features. While CMV depends on the positive-definiteness of the covariance matrix of the vectors of residuals, such limitation does not apply to NPC (Pesarin and Salmaso, 2010a). As a consequence, although in the comparisons only the Hotelling's T^2 and the Wilks' λ statistics were used (in the simulations, $\mathrm{rank}\,(\mathbf{C}) = 1$), and had their p-values assessed through permutations, similar behaviour can be expected when using other CMVs, such as Pillai's trace (and with $\mathrm{rank}\,(\mathbf{C}) > 1$). With effect, NPC can be used even when the number of variables equals or even greatly exceeds the number of observations, that is, when $K \geqslant N$. In the results shown in Figure 7, this can be noted as a reduction in power that can be seen with the Hotelling's T^2, particularly for simulation D, and this is the case even considering that the test is assessed through permutations.

Regarding the different combining functions, the simulations show that the method of Tippett is the most powerful when signal is present in only a small fraction of the partial tests. For other cases, other combining functions, particularly that of Fisher, tend to be considerably more powerful.

The results also indicate that the use of sign flipping when the errors are not symmetric (a violation of assumptions) tends to produce a conservative test, with error rates below the nominal level, even if the power eventually remained unaltered when compared with permutations. While permutations together with sign flippings did alleviate conservativeness, at least for the Tippett method, the error rate remained below the nominal level. In general, if the errors are known to be skewed, only permutations should be used; if sign flippings are used, the error rate can be expected to be below the nominal level.

5.3. Interpretation of combined tests

The key aspect of the NPC is that these tests seek to identify, *on the aggregate* of the partial tests, a measure of evidence against the JNH, even if only some or none of them can be considered significant when seen in isolation, just as originally pointed out by Fisher (1932):

> *When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than*

26

> *would often have been obtained by chance. It is sometimes desired*
> *(. . .) to obtain a single test of the significance of the aggregate.*

This is the logic and interpretation of all of these combining statistics, with the exception of the conjunction inference. Combination is known to be able to answer questions that could otherwise not be answered be at all, or be answered less accurately if each information source were considered separately (Draper et al., 1992). Here the simulations and the pain study exemplify these aspects, and the improved sensitivity compared to each partial test when seen in separate.

As they depend on fewer assumptions than classical multivariate tests, NPC can be considered whenever the validity of the former cannot be guaranteed. Even when parametric CMV assumptions hold, note that the NPC can have superior power when sample size is small and prevents precise estimation of a covariance.

It should be noted that the aggregation of information follows a different principle than using different measurements separately to interrogate particular aspects of the brain (or of any other experiment or physiological phenomenon). Used judiciously, NPC provides a complete framework that can be used for both the aggregate and for the correction of tests separately, with the valuable feature of being based on minimal assumptions.

5.4. Correction over contrasts and over modalities

Correction over contrasts using synchronised permutations provides a novel solution to the multiple comparisons problem for certain common experimental designs, in particular, for the popular one-way ANOVA layout, that is, when the means of multiple groups are compared. The classical Fisher's protected least significant difference (LSD), that consists of performing an omnibus F-test and only proceeding to the group-wise post hoc tests if this initial test is significant, is known to fail to control the error rate if there are more than just three groups (Hayter, 1986; Hsu, 1996; Meier, 2006), and the failure can be by a wide margin, that grows as the number of groups being compared increases. Even though the same may not happen with other correction methods (e.g., Tukey's range test, Tukey, 1949), the correction done non-parametrically also renders these older, parametric methods, redundant.

The correction over contrasts further obviates methods that are based on what has been termed "logical constraints" among hypotheses (Shaffer, 1986; Hochberg and Tamhane, 1987), as the dependencies among the tests

27

are implicitly taken into account by the correction using the distribution of the extremum across contrasts, with or without concomitant combination or correction across multiple K variables. In fact, the use of an omnibus F-test as a way to guard against multiple testing becomes quite unnecessary.

In the same manner, while combination across multiple modalities is a powerful substitute for classical multivariate tests as shown earlier, the correction across such modalities can replace the post hoc tests that are usually performed after significant results are found with CMVs.

5.5. Pain study

Joint significance is an important consideration when trying to interpret data such as these, that are distinct in some aspects (here, the topography of the stimulation), but similar in others (here, the type of stimulation, hot and painful), strengthening the case for distinct representations in some brain regions, but not in others. In terms of identifying areas with significant joint activity, the results suggest involvement of large portions of the anterior insula and secondary somatosensory cortex. The Fisher, Stouffer and Mudholkar–George combining functions were particularly successful in recovering a small area of activity in the midbrain and periaqueductal gray area that would be expected from previous studies on pain (Reynolds, 1969; Petrovic et al., 2002; Tracey et al., 2002; Roy et al., 2014), but that could not be located from the original, non-combined data.

5.6. Relationship with meta-analysis

Most of the combining functions shown in Table 1 were originally defined based on p-values, and some of them are popular in meta-analyses, such as those of Fisher and Stouffer (Borenstein et al., 2009). Although there are commonalities between these meta-analytical methods and NPC, it is worth emphasising that the two constitute distinct approaches to entirely different problems. In the NPC, the objective is to interrogate joint significance across the multiple observed variables (or multiple designs and contrasts if these are instead combined) when the data for each individual observation is readily available to the researcher. Meta-analyses methods based on p-values, while sometimes using the same combining functions, attempt to identify a joint effect across multiple studies that not have necessarily been performed on the same experimental units, and when the data for the individual observations are not available. Moreover, the p-value of the combined statistic in the NPC is produced through permutations, a procedure that is not available for ordinary meta-analytical methods.

The fact that NPC and meta-analysis form different approaches to separate problems also imply that certain criticisms levelled at the use of certain combined functions in the context of meta-analysis do not extend trivially to NPC. As the simulations show, various of the combining functions more recently developed did not in general outperform older combining methods, such as Fisher and Stouffer, even though these were developed precisely for that purpose, in the context of meta-analyses, or for problems framed as such.

6. Conclusion

We proposed and evaluated a modified version of Non-Parametric Combination that is feasible and useful for imaging applications, and serves as a more powerful alternative to classical multivariate tests. We presented and discussed aspects related multiple testing problems in brain imaging, and proposed a single framework that addresses all these concerns at once. We showed that combination and correction of multiple imaging modalities, designs, and contrasts, are related to each other in the logic of their implementation, and also through the use of the simplest and the oldest of the combining functions, attributed to Tippett.

Appendix A. Brief overview of combining functions

Below are a few details and references for the methods shown in Table 1, plus a few others, presented in chronological order. A number of studies comparing some of these functions in various scenarios have been published (Birnbaum, 1954; van Zwet and Oosterhoff, 1967; Oosterhoff, 1969; Rosenthal, 1978; Berk and Cohen, 1979; Westberg, 1985; Lazar et al., 2002; Loughin, 2004; Whitlock, 2005; Wu, 2006; Won et al., 2009; Bhandary and Zhang, 2011; Chen, 2011; Zaykin, 2011; Chang et al., 2013). Some of these are permutationally equivalent to each other, that is, their rejection region under permutation is the same, and it becomes immaterial which is chosen.

Tippett. This is probably the oldest, the simplest, and the most intuitive of the combination methods, having appeared in the first edition of Tippett's book *The Methods of Statistics* (Tippett, 1931, page 35). The combined test statistic is simply the minimum p-value across all partial tests, and Tippett shows its distribution has a simple closed form.

Fisher. This method appeared in the fourth edition of *Statistical Methods for Research Workers* (Fisher, 1932), and follows the idea of treating the joint probability as the intersection of all partial tests, which is given by their product $\prod_k p_k$. This product, however, is not uniformly distributed, even if the global null hypothesis is true. Using a few properties of the uniform distribution, Fisher showed that twice the negative logarithm of the products follows a χ^2 distribution, with degrees of freedom $2K$.

Stouffer. This method appeared in footnotes in the report of the large sociological study conducted among veterans of the World War ii by Stouffer et al. (1949, page 45, footnote 15, and page 151, footnote 14). The idea is to sum z-scores, normalise the variance of this sum, and from this statistic obtain a p-value for the joint hypothesis.

Wilkinson. The probability of observing r significant p-values at the level α can be computed using a binomial expansion, as proposed by Wilkinson (1951). The statistic is therefore simply r, and the probability does not depend on the actual p-values for the partial tests, but only on r and α.

Good. A generalisation of the Fisher method that assigns arbitrary, unequal positive weights w_k for each of the partial tests, was suggested by Good (1955). The weights are defined according to some criteria, such as the sample size for each of the partial test, the number of degrees of freedom, or some other desirable feature, such as ecological or internal validity (Rosenthal, 1978).

Lipták. Another generalised combined statistic can be produced using the inverse cdf, F^{-1}, of the p_k, summing the values of the statistics, and computing a new p-value for the global null using the cdf G of the sum of the statistics, a method proposed by Lipták (1958). Each summand can be arbitrarily weighted, as in the Good method. In principle, any continuously increasing function with support in the interval $[0,\ 1]$ can be used for F, albeit a more obvious choice is the cdf of the normal distribution, which can be used as both F and G, and which equals the approach to the Stouffer method if all weights are 1.

Lancaster. While the Lipták method generalises combining strategies such as Fisher and Stouffer, the Lancaster method (Lancaster, 1961) further generalises the Lipták approach by allowing different F_k^{-1} for each partial test. Choices for F_k^{-1} include, for instance, the cdf of the gamma distribution

30

with scale parameter $\theta = 2$, possibly with different shape parameters taking the place of the weights for each partial test. If the weights are all positive integers, the p-values can be assessed from the cdf of a χ^2 distribution with degrees of freedom $\nu = 2\sum_k w_k$ (Berk and Cohen, 1979).

Winer. A combination strategy that resembles the Stouffer method, but uses the Student's t statistic, was proposed by Winer (1962, page 44), albeit not found in later editions of the book. The idea is to sum the t statistics for all the partial tests, then normalise the sum so that the resulting statistic follows a standard normal distribution. The normalisation is based on the fact that the variance of the t distribution can be determined from its degrees of freedom ν as $\nu/(\nu-2)$. The method cannot be applied if $\nu_k \leqslant 2$ for any of the partial tests. Moreover, ν_k should not be too small for the normal approximation to be reasonably valid (e.g., $\nu_k \geqslant 10$). The Winer method is a particular case of the Lancaster method.

Edgington. The probability of observing, due to chance, a value equal or smaller than the sum of the partial p-values was proposed by Edgington (1972) as what would be a more powerful alternative to the Fisher method. The method however, lacks consistency (see Appendix B).

Mudholkar–George. It is possible to use a simple logit transformation to compute a statistic that approximates a scaled version of the Student's t distribution, as shown by Mudholkar and George (1979). If the scaling is taken into account, the combined statistic follows a t distribution.

Darlington–Hayes. In a discussion about pooling p-values for meta-analysis, Darlington and Hayes (2000) raised a number of limitations of these methods, and proposed a modification over the method of Stouffer that would address some of these concerns. The modified method, called by the authors as *Stouffer-max*, uses as test statistic the mean of the r highest z-scores, rather than the normalised sum of all the z-scores as in the original method. When $r = 1$, it is equivalent to the Tippett method, whereas when $r = K$, is equivalent to the original Stouffer. The p-values for intermediate values of r can be computed through Monte Carlo simulation, and the authors provided tables with critical values.

Zaykin et al. This method, called *truncated product method* (TPM) was proposed by Zaykin et al. (2002) as a way to combine features of the Fisher and Wilkinson methods. The statistic is the product of only the partial

31

p-values that are significant at the level α, whereas in the Fisher method, all p-values are used. If $\alpha = \min (p_k)$, the approach is equivalent to the Tippett method. If $\max (p_k) \leqslant \alpha \leqslant 1$, the approach is equivalent to the Fisher method. An expression for the p-values that produces exact values was provided by the authors. The expresion, however, is prone to over/underflows for certain combinations of large K and α, and when p-values cannot be obtained analytically, Monte Carlo methods can be used.

Dudbridge–Koeleman. While the Zaykin method combines only the partial tests that are significant at the level α, it is also possible to create a statistic that combines only the most r significant tests, where r is specified in advance. This method was proposed by Dudbridge and Koeleman (2003) and called *rank truncated product* (RTP). The main benefit of this strategy is that it depends only on a predetermined number of partial tests to be rejected, rather than on their p-values, which are random quantities. As with the Zaykin method, for certain combinations of r and large K, the p-values need to be computed through Monte Carlo methods. In the same article, the authors also introduced a combination of the TPM and RTP, and named it *rank-and-threshold truncated product* or *dual truncated product* (DTP). The statistic is the largest of either if these two, and its p-value can be computed analytically or via Monte Carlo methods.

Taylor–Tibshirani. If the p-values are sorted in ascending order, these ranked p-values can be compared to their expectations under the global null hypothesis. Large deviations from the expected values suggest the presence of the effect among the tests. Taylor and Tibshirani (2006) suggested that a measurement of this deviation could be used to infer the overall significance of the tests. The corresponding statistic was termed *tail strength* (TS), and under the assumptions that the global null is true and that the tests are independent, it follows a normal distribution with zero mean and a variance that can be approximated as $1/K$ for large K, from which the p-value can be assessed. When these assumptions are not met, non-parametric methods can be used.

Jiang et al. The statistic of the Taylor–Tibshirani method has a variance that depends asymptotically only on the number of tests. However, the value of the statistic can be small when effect is truly present in only a few partial tests, therefore potentially reducing power. By analogy to the Zaykin method, Jiang et al. (2011) proposed to compute the tail strength

32

using only partial tests with p-values smaller than a certain level α. The method is called *truncated tail strength* (TTS). The analytical form for the distribution is not known, and the authors propose computing the p-value using Monte Carlo or permutation methods.

Li–Tseng. Li and Tseng (2011) proposed a modification of the Fisher method that is used not to test the JNH (hence not shown in Table 1), but to identify which of the partial tests contribute the most to the resulting combined statistic. The authors define a quantity $A_W = -\sum_{k=1}^{K} w_k \ln p_k$, where w_k is a weight that can be either 0 or 1. All possible $2^K - 1$ non-trivial combinations $W = [w_1, \ldots, w_K]$ are evaluated to produce a value for A_W. The respective p-values p_W are computed via permutations, and the W that yields the smallest such p-value over all possible combinations of weights, is the one that identifies the subset among the K tests that contributes the most to the combined p-values.

Appendix B. Consistency of combined tests

A hypothesis test is said to be *consistent* if, for a fixed test level, its power goes to unity as the sample size increases to infinity. The use of a non-consistent combining function to form an NPC test is problematic, as the rejection region may not be reached even if the p-value for one or more of the partial tests approach zero, thus violating the second of the three desirable properties of the combining functions, presented in Section 2.4.

Among the functions shown in Table 1, the notable non-consistent combining functions are the Edgington and Wilkinson (see Appendix A). Also, it should be noted that functions that define conjunctions (IUT), such as those based on $\max(p_k)$, are likewise not consistent in the context of NPC, as the latter serves to test the global null hypothesis. Figure 9 shows rejection regions for some inconsistent combining functions, and variants, similarly as for the (consistent) shown in Figure 3.

Appendix C. Admissibility of combined tests

A combined hypothesis test is said to be *admissible* if there exists no other test that, at the same significance level, without being less powerful to all possible alternative hypotheses, is more powerful to at least one alternative (Lehmann and Romano, 2005). This can be stated in terms of either of two sufficient conditions for admissibility: (I) that rejection of the null for a

33

given p-value implies the rejection of the null for all other p-values smaller or equal than that, or (II) that the rejection region is convex in the space of the test statistic.

Combinations that favour tests with concordant directions (Section 2.6), if used with of non-directional partial tests, create tests that are inadmissible, that is, tests that are not optimal in the sense that there exist other tests that, without being less powerful to some true alternative hypotheses, are more powerful to at least one true alternative. Inadmissibility implies that the test cannot be used, as certain combinations of partial tests lead to nonsensical results, such as rejecting the JNH for some partial p-values, and failing to reject for some p-values that are even smaller. Figure 10 shows rejection regions of inadmissible versions of the combining functions considered in Figures 3 and 9; clearly none of the two conditions above are satisfied. The particular combining function shown in Equation 2 was suggested by Pearson (1933) and used by David (1934), but after a paper by Birnbaum (1954), it was for decades thought to be inadmissible. However, it is in fact admissible (Owen, 2009).

Admissibility is important in that it allows, for more than just two partial tests, combined tests that favour alternative hypotheses with the same direction. Other possibilities favouring alternatives with common direction, such as multiplying together the partial test statistics to produce a combined statistic, work for two partial tests only (Hayasaka et al., 2006).

Appendix D. Implementation

A unified algorithm for combination and correction that is amenable for use with imaging applications is shown below. It has many similarities with the `randomise` algorithm (Winkler et al., 2014), with various modifications to accommodate combination and correction. The p-values adjusted for the multiplicity of tests are computed using the distribution of the extremum statistic, which can be collapsed across modalities and/or designs and contrasts for each case, rendering the algorithm simpler. The notation below is slightly different than that used throughout the paper. The inputs are:

– $\mathbf{Y}$: The input data for each of the K modalities and image points. Each column vector of N observations for the k-th modality is accessed as $\mathbf{Y}[k, \mathbf{v}]$, where $\mathbf{v} = [x, y, z]$ is used to specify the point position in space; this is so without loss of generality for non-imaging data.

34

– $\mathcal{X}$: The set of design matrices $\mathbf{X}$.

– $\{\mathcal{C}_{\mathbf{X}}\}$: The set of sets of contrasts for each design matrix $\mathbf{X}$. Each element of each subset is a pair of multivariate contrasts $(\mathbf{C}, \mathbf{D})$. This definition allows each design to be tested with multiple such pairs of contrasts, and allows various designs to be tested with the same input data.

– $\mathbf{B}$: Definition of multi-level exchangeability blocks, used to define valid shufflings that respect the data structure (Winkler et al., 2015, *in press*).

– $\mathbf{V}$: Definition of the variance groups, useful to compute statistics that are robust to heteroscedasticity.

– EE, ISE: Boolean flags (true/false) indicating whether errors can be treated as exchangeable (EE), allowing permutations, independent and symmetric (ISE), allowing sign-flippings, or both.

– J: Number of permutations to be performed.

– NPCMOD, NPCCON: Boolean indicating whether combination should be performed respectively across modalities, across designs and contrasts, or both.

– FWEMOD, FWECON: Boolean indicating whether familywise error rate correction should be performed respectively across modalities, across designs and contrasts, or both.

The output of interest is the p-value. For simplicity, as shown, the output is always FWER-adjusted across the image points indexed by $\mathbf{v}$, and for the non-combined, further adjusted based on the contrasts and modalities; these are shown in the algorithm topped by a tilde, that is, as "p̃-value", as opposed to simply "p-value". Also for simplicity, p-values for combination of modalities are not shown adjusted for multiple contrasts, nor vice-versa. These can also be obtained following the same logic used for the FWER-adjustment of the non-combined statistics. Uncorrected p-values, useful for correction using false discovery rate (FDR, Benjamini and Hochberg, 1995) can be obtained with trivial modifications.

Algorithm 1: Unified algorithm. See the main text for details.

Require: $\mathbf{Y}, \mathcal{X}, \{\mathcal{C}_{\mathbf{X}}\}, \mathbf{B}, \mathbf{V}$, EE, ISE, J, NPCMOD, NPCCON, FWEMOD, FWECON.

35

1: $\mathcal{P} \leftarrow \mathsf{sync\_perms}(\mathcal{X}, \{\mathcal{C}_{\mathbf{X}}\}, \mathbf{B}, \text{EE}, \text{ISE}, J - 1)$ ▷ Define the permutation set.

2: $\mathcal{P} \leftarrow \{\mathbf{I}, \mathcal{P}\}$ ▷ Ensure first permutation is no permutation.

3: **for** $j = 1, \ldots, J$ **do**   ▷ For each shuffling.

4:    $c \leftarrow 1$ ▷ Counter for the number of designs and contrasts.

5:    **for all** $\mathbf{X} \in \mathcal{X}$ **do**   ▷ For each design matrix.

6:      **for all** $(\mathbf{C}, \mathbf{D}) \in \mathcal{C}_{\mathbf{X}}$ **do**   ▷ For each pair of contrasts.

7:        $\mathbf{Y} \leftarrow \mathbf{YD}$ ▷ Redefine the data, discard $\mathbf{D}$.

8:        $\mathbf{X}^* \leftarrow \mathbf{P}_j \mathbf{X}$ ▷ Shuffle the model.

9:        **for all** $k \in \{1, \ldots, K\}$ **do**   ▷ For each partial test.

10:          **for all** $\mathbf{v}$ **do**   ▷ For each image point.

11:            $\hat{\boldsymbol{\beta}} \leftarrow (\mathbf{X}^*)^+ \mathbf{Y}[k, \mathbf{v}]$ ▷ Estimated regression coefficients.

12:            $\hat{\mathbf{E}} \leftarrow \mathbf{Y}[k, \mathbf{v}] - \mathbf{X}^* \hat{\boldsymbol{\beta}}$ ▷ Estimation residuals.

13:            $\mathbf{G} \leftarrow \mathsf{pivotal}(\mathbf{X}^*, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\epsilon}}, \mathbf{C}, \mathbf{V})$ ▷ Test statistic.

14:            $\mathbf{U}[j, k, c, \mathbf{v}] \leftarrow \mathsf{transform}(\mathbf{G})$ ▷ Transform to u-value.

15:            **if** $j = 1$ **then**   ▷ In the first permutation (no permutation).

16:              $\mathbf{U}_0[k, c, \mathbf{v}] \leftarrow \mathbf{U}[1, k, c, \mathbf{v}]$ ▷ Keep the unpermuted u-value.

17:            **end if**

18:          **end for**

19:          $\mathbf{U}_e[j, k, c] \leftarrow \mathsf{extremum}(\mathbf{U}[j, k, c, \cdot])$ ▷ Extremum across space.

20:        **end for**

21:        $c \leftarrow c + 1$ ▷ Increment counter for the number of designs and contrasts.

22:      **end for**

23:    **end for**

24:    $C \leftarrow c$ ▷ Keep the total number of designs and contrasts for later use.

25:    **if** NPCMOD $\wedge \neg$ NPCCON **then**   ▷ Combine modalities only.

26:      **for all** $c \in \{1, \ldots, C\}$ **do**   ▷ For each design/contrast.

27:        **for all** $\mathbf{v}$ **do**   ▷ For each image point.

28:          $\mathbf{T}[c, \mathbf{v}] \leftarrow \mathsf{combine}(\mathbf{U}[j, \cdot, c, \mathbf{v}])$ ▷ Combined statistic.

29:        **end for**

30:        $\mathbf{T}_e[j, c] \leftarrow \mathsf{extremum}(\mathbf{T}[c, \cdot])$ ▷ Distribution of the extrema across tests.

31:      **end for**

32:    **else if** NPCCON $\wedge \neg$ NPCMOD **then**   ▷ Combine designs/contrasts only.

33:      **for all** $k \in \{1, \ldots, K\}$ **do**   ▷ For each design/contrast.

34:        **for all** $\mathbf{v}$ **do**   ▷ For each image point.

35:          $\mathbf{T}[k, \mathbf{v}] \leftarrow \mathsf{combine}(\mathbf{U}[j, k, \cdot, \mathbf{v}])$ ▷ Combined statistic.

36:        **end for**

37:        $\mathbf{T}_e[j, k] \leftarrow \mathsf{extremum}(\mathbf{T}[k, \cdot])$ ▷ Distribution of the extrema across tests.

38:      **end for**

39:    **else if** NPCMOD $\wedge$ NPCCON **then** ▷ Combine modalities & designs/contrasts.

40:      **for all** $\mathbf{v}$ **do**   ▷ For each image point.

41:        $\mathbf{T}[\mathbf{v}] \leftarrow \mathsf{combine}(\mathbf{U}[j, \cdot, \cdot, \mathbf{v}])$ ▷ Combined statistic.

42:      **end for**

43:      $\mathbf{T}_e[j] \leftarrow \mathsf{extremum}(\mathbf{T}[\cdot])$ ▷ Distribution of the extrema across tests.

44:    **end if**

45:    **if** $j = 1$ **then**   ▷ In the first permutation (no permutation).

46:      $\mathbf{T}_0 \leftarrow \mathbf{T}$ ▷ Keep the unpermuted combined statistic.

47:    **end if**

48: **end for**

49: **if** NPCMOD $\wedge \neg$ NPCCON **then**   ▷ Combine modalities only.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

50: **for all** $c \in \{1, \ldots, C\}$ **do** $\qquad\qquad$ ▷ For each design/contrast.
51: $\quad$ **for all v do** $\qquad\qquad$ ▷ For each image point.
52: $\qquad$ p-value$[c, \mathbf{v}] \leftarrow \mathsf{data\_pval}(\mathbf{T}_0[c, \mathbf{v}], \mathbf{T}_e[\cdot, c])$ $\qquad$ ▷ Combined p-value.
53: $\quad$ **end for**
54: **end for**
55: **else if** NPCCON $\wedge \neg$ NPCMOD **then** $\qquad$ ▷ Combine designs/contrasts only.
56: **for all** $k \in \{1, \ldots, K\}$ **do** $\qquad\qquad$ ▷ For each design/contrast.
57: $\quad$ **for all v do** $\qquad\qquad$ ▷ For each image point.
58: $\qquad$ p-value$[k, \mathbf{v}] \leftarrow \mathsf{data\_pval}(\mathbf{T}_0[k, \mathbf{v}], \mathbf{T}_e[\cdot, k])$ $\qquad$ ▷ Combined p-value.
59: $\quad$ **end for**
60: **end for**
61: **else if** NPCMOD $\wedge$ NPCCON **then** $\quad$ ▷ Combine modalities & designs/contrasts.
62: **for all v do** $\qquad\qquad$ ▷ For each image point.
63: $\quad$ p-value$[\mathbf{v}] \leftarrow \mathsf{data\_pval}(\mathbf{T}_0[\mathbf{v}], \mathbf{T}_e[\cdot])$ $\qquad$ ▷ Combined p-value.
64: **end for**
65: **end if**
66: **if** FWEMOD $\wedge \neg$ FWECON **then** $\qquad$ ▷ Correct over modalities only.
67: **for all** $c \in \{1, \ldots, C\}$ **do** $\qquad\qquad$ ▷ For each design/contrast.
68: $\quad$ **for all** $j \in \{1, \ldots, J\}$ **do** $\qquad\qquad$ ▷ For each shuffling.
69: $\qquad$ $\mathbf{U}'_e[j, c] \leftarrow \mathsf{extremum}(\mathbf{U}_e[j, \cdot, c])$ $\qquad$ ▷ Distribution of the extrema.
70: $\quad$ **end for**
71: $\quad$ **for all** $k \in \{1, \ldots, K\}$ **do** $\qquad\qquad$ ▷ For each modality.
72: $\qquad$ **for all v do** $\qquad\qquad$ ▷ For each image point.
73: $\qquad\quad$ $\tilde{\mathrm{p}}$-value$[k, c, \mathbf{v}] \leftarrow \mathsf{data\_pval}(\mathbf{U}_0[k, c, \mathbf{v}], \mathbf{U}'_e[\cdot, c])$ $\quad$ ▷ Adjusted p-value.
74: $\qquad$ **end for**
75: $\quad$ **end for**
76: **end for**
77: **else if** FWECON $\wedge \neg$ FWEMOD **then** $\qquad$ ▷ Correct over designs/contrasts only.
78: **for all** $k \in \{1, \ldots, K\}$ **do** $\qquad\qquad$ ▷ For each modality.
79: $\quad$ **for all** $j \in \{1, \ldots, J\}$ **do** $\qquad\qquad$ ▷ For each shuffling.
80: $\qquad$ $\mathbf{U}'_e[j, k] \leftarrow \mathsf{extremum}(\mathbf{U}_e[j, k, \cdot])$ $\qquad$ ▷ Distribution of the extrema.
81: $\quad$ **end for**
82: $\quad$ **for all** $c \in \{1, \ldots, C\}$ **do** $\qquad\qquad$ ▷ For each design/contrast.
83: $\qquad$ **for all v do** $\qquad\qquad$ ▷ For each image point.
84: $\qquad\quad$ $\tilde{\mathrm{p}}$-value$[k, c, \mathbf{v}] \leftarrow \mathsf{data\_pval}(\mathbf{U}_0[k, c, \mathbf{v}], \mathbf{U}'_e[\cdot, k])$ $\quad$ ▷ Adjusted p-value.
85: $\qquad$ **end for**
86: $\quad$ **end for**
87: **end for**
88: **else if** FWEMOD $\wedge$ FWECON **then** $\qquad$ ▷ Correct over modalities & des./contr.
89: **for all** $j \in \{1, \ldots, J\}$ **do** $\qquad\qquad$ ▷ For each shuffling.
90: $\quad$ $\mathbf{U}'_e[j] \leftarrow \mathsf{extremum}(\mathbf{U}_e[j, \cdot, \cdot])$ $\qquad$ ▷ Distribution of the extrema.
91: **end for**
92: **for all** $k \in \{1, \ldots, K\}$ **do** $\qquad\qquad$ ▷ For each modality.
93: $\quad$ **for all** $c \in \{1, \ldots, C\}$ **do** $\qquad\qquad$ ▷ For each design/contrast.
94: $\qquad$ **for all v do** $\qquad\qquad$ ▷ For each image point.
95: $\qquad\quad$ $\tilde{\mathrm{p}}$-value$[k, c, \mathbf{v}] \leftarrow \mathsf{data\_pval}(\mathbf{U}_0[k, c, \mathbf{v}], \mathbf{U}'_e[\cdot])$ $\qquad$ ▷ Adjusted p-value.
96: $\qquad$ **end for**
97: $\quad$ **end for**
98: **end for**

37

99: **end if**

Within the algorithm, the functions are:

– *sync_perms*: This function produces a set $\mathcal{P}$ of permutation and/or sign flipping matrices that can be performed sychronously to test a joint null hypotheses about the input data. The synchronisation is always necessary to allow combination/correction over modalities, and it may also be necessary across multiple designs and/or contrasts if these are to be combined/corrected as well. If synchronisation is not necessary for designs and/or contrasts, the algorithm can be modified so that $\mathcal{P}$ can be defined inside the for-loops that iterate over designs and contrasts.

– *transform*: This converts the test statistic into a u-value, thus rendering the NPC method feasible for imaging applications. If no combination is to be performed, the algorithm can be modified to skip this step and work directly with the test statistic.

– *extremum*: For statistics in which larger values are evidence against the null hypothesis, this function takes the maximum. For statistics in which smaller values are indication against the null, this takes the minimum. In either case, it is always the most extreme towards evidence favouring the alternative. This function effectively implements a CTP using an IUT.

– *combine*: This combines the inputs (p- or u-values) into a new, combined statistic. Any of the combining functions from Table 1 can be considered. For the method of Tippett, *combine* and *extremum* are the same.

– *data_pval*: This function produces a p-value based on a set of empirical values for the test statistic after shuffling. This works by computing the fraction of the test statistics after shuffling that is larger or equal than the unpermuted test statistic, while taking care of ties.

The algorithm has four major parts: the first consists of the loop that begins in line 5 of the pseudocode above, and which consists of a simplified version of the randomise algorithm. The second begins with the conditional structure in line 27, which performs the combination and computes distribution of the extremum statistics for each case of NPC, thus also treating the MTP-I. These initial two parts are repeated for each shufflong, in the

38

loop that begins in line 3. The third part begins with the conditional in line 51, that is, once all rearrangements have been performed; in this part, the distributions are used to compute the combined p-values. Finally, the fourth part begins with the conditional in line 68, in which the MTP-II is addressed.

As shown, the algorithm is simplified so as to emphasise the most important aspects of combination and correction. However, various modifications and improvements can be applied for particular circumstances, and for speed, including the partitioning discussed in the Appendix A of Winkler et al. (2014). An open-source working implementation, that can be executed in MATLAB (The MathWorks Inc., 2013) or Octave (Eaton et al., 2014), is available in the tool *Permutation Analysis of Linear Models* (PALM), available for download at www.fmrib.ox.ac.uk/fsl.

Acknowledgements

Abou Elseoud, A., Nissilä, J., Liettu, A., Remes, J., Jokelainen, J., Takala, T., Aunio, A., Starck, T., Nikkinen, J., Koponen, H., Zang, Y. F., Tervonen, O., Timonen, M., Kiviniemi, V., 2014. Altered resting-state activity in seasonal affective disorder. Human Brain Mapping 35 (1), 161–172.

Anderson, T. W., 2003. An Introduction to Multivariate Statistical Analysis. Wiley, Hoboken, NJ, USA.

Benjamini, Y., Heller, R., 2008. Screening for partial conjunction hypotheses. Biometrics 64 (4), 1215–22.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 289–300.

Berger, R. L., 1982. Multiparameter hypothesis testing and acceptance sampling. Technometrics 24 (4), 295–300.

Berk, R. H., Cohen, A., 1979. Asymptotically optimal methods of combining tests. Journal of the American Statistical Association 74 (368), 812–814.

Bhandary, M., Zhang, X., 2011. Comparison of several tests for combining several independent tests. Journal of Modern Applied Statistical Methods 10 (2), 436–446.

Birnbaum, A., 1954. Combining independent tests of significance. Journal of the American Statistical Association 49 (267), 559–574.

Blair, R. C., Higgins, J. J., Karniski, W., Kromrey, J. D., 1994. A study of multivariate permutation tests which may replace Hotelling's t^2 test in prescribed circumstances. Multivariate Behavioral Research 29 (2), 141–163.

Bland, J. M., Altman, D. G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 327 (8476), 307–10.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., Rothstein, H. R., 2009. Introduction to Meta-Analysis. Wiley, West Sussex, UK.

Brombin, C., Midena, E., Salmaso, L., 2013. Robust non-parametric tests for complex-repeated measures problems in ophthalmology. Statistical Methods in Medical Research 22 (6), 643–660.

Brooks, J. C. W., Zambreanu, L., Godinez, A., Craig, A. D. B., Tracey, I., 2005. Somatotopic organisation of the human insula to painful heat studied with high resolution functional imaging. NeuroImage 27 (1), 201–9.

Brown, M. B., 1975. A method for combining non-independent, one-sided tests of significance. Biometrics 31 (4), 987–992.

Chang, L.-C., Lin, H.-M., Sibille, E., Tseng, G. C., 2013. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics 14 (vi), 368.

Chen, G., Adleman, N. E., Saad, Z. S., Leibenluft, E., Cox, R. W., 2014. Applications of multivariate modeling to neuroimaging group analysis: a comprehensive alternative to univariate general linear model. NeuroImage 99, 571–88.

Chen, Z., 2011. Is the weighted z-test the best method for combining probabilities from independent tests? Journal of Evolutionary Biology 24 (4), 926–30.

Christensen, R., 2001. Advanced Linear Modelling, 2nd Edition. Springer, New York, USA.

Darlington, R. B., Hayes, A. F., 2000. Combining independent p values: extensions of the stouffer and binomial methods. Psychological Methods 5 (4), 496–515.

David, F. N., 1934. On the P_{λ_n} test for randomness: Remarks, further illustration, and table of P_{λ_n} for given values of $-\log_{10} \lambda_n$. Biometrika 26 (1-2), 1–11.

Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., Waternaux, C., 1992. Combining information: statistical issues and opportunities for research. National Academy Press Washington, DC, USA.

Dudbridge, F., Koeleman, B. P. C., 2003. Rank truncated product of P-values, with application to genomewide association scans. Genetic Epidemiology 25 (4), 360–6.

Eaton, J. W., Bateman, D., Hauberg, S., 2014. GNU Octave: A high-level interactive language for numerical computations. Network Theory Ltd.
URL http://www.gnu.org/software/octave

Edgington, E. S., 1972. An additive method for combining probability values from independent experiments. The Journal of Psychology 80 (2), 351–363.

Efron, B., mar 2004. Large-scale simultaneous hypothesis testing. Journal of the American Statistical Association 99 (465), 96–104.

Fisher, R. A., 1932. Statistical Methods for Research Workers, 4th Edition. Oliver and Boyd, Edinburgh.

Fox, P. T., Mintun, M. A., Reiman, E. M., Raichle, M. E., 1988. Enhanced detection of focal brain responses using intersubject averaging and change-distribution analysis of subtracted PET images. Journal of Cerebral Blood Flow and Metabolism 8 (5), 642–53.

Friston, K. J., Frith, C. D., Liddle, P. F., Frackowiak, R. S., 1991. Comparing functional (PET) images: the assessment of significant change. Journal of Cerebral Blood Flow and Metabolism 11 (4), 690–9.

Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C., Worsley, K. J., 1999. Multisubject fMRI studies and conjunction analyses. NeuroImage 10 (4), 385–96.

Friston, K. J., Penny, W. D., Glaser, D. E., 2005. Conjunction revisited. NeuroImage 25 (3), 661–7.

Genovese, C. R., Lazar, N. A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15 (4), 870–8.

Good, I. J., 1955. On the weighted combination of significance tests. Journal of the Royal Statistical Society. Series B 17 (2), 264–265.

Hall, P., Wilson, S. R., 1991. Two guidelines for bootstrap hypothesis testing. Biometrics 47 (2), 757–762.

Hayasaka, S., Du, A.-T., Duarte, A., Kornak, J., Jahng, G.-H., Weiner, M. W., Schuff, N., 2006. A non-parametric approach for co-analysis of multi-modal brain imaging data: application to alzheimer's disease. NeuroImage 30 (3), 768–79.

Hayasaka, S., Nichols, T. E., 2004. Combining voxel intensity and cluster extent with

40

permutation test framework. NeuroImage 23 (1), 54–63.

Hayter, A. A. J., 1986. The maximum familywise error rate of Fisher's least significant difference test. Journal of the American Statistical Association 81 (396), 1000–1004.

Hochberg, Y., Tamhane, A. C., 1987. Multiple Comparison Procedures. John Wiley & Sons, Inc., New York, NY, USA.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6 (2), 65–70.

Hotelling, H., 1931. The generalization of Student's ratio. The Annals of Mathematical Statistics 2 (3), 360–378.

Hotelling, H., 1951. A generalized T test and measure of multivariate dispersion. In: Neyman, J. (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. No. 042. University of California Press, Berkeley, pp. 23–41.

Hsu, J. C., 1996. Multiple Comparison: Theory and Methods. Chapman & Hall/CRC, Boca Raton, FL, USA.

Jiang, B., Zhang, X., Zuo, Y., Kang, G., 2011. A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. Journal of Theoretical Biology 277 (1), 67–73.

Johnson, R. A., Wichern, D. W., 2007. Applied Multivariate Statistical Analysis, 6th Edition. Pearson Prentice Hall, Upper Sadle River, NJ USA.

Kost, J. T., McDermott, M. P., 2002. Combining dependent p-values. Statistics & Probability Letters 60 (2), 183–190.

Kuhfeld, W. F., 1986. A note on Roy's largest root. Psychometrika 51 (3), 479–481.

Lancaster, H. O., 1961. The combination of probabilities: an application of orthonormal functions. Australian Journal of Statistics 3 (1), 20–33.

Lawley, D. N., 1938. A generalization of Fisher's z test. Biometrika 30 (1), 180–187.

Lazar, N. A., Luna, B., Sweeney, J. A., Eddy, W. F., Jun. 2002. Combining brains: a survey of methods for statistical pooling of information. NeuroImage 16 (2), 538–50.

Lehmann, E. L., Romano, J. P., 2005. Testing Statistical Hypotheses, 3rd Edition. Springer, New York, NY, USA.

Li, J., Tseng, G. C., 2011. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. Annals of Applied Statistics 5 (2), 994–1019.

Licata, S. C., Nickerson, L. D., Lowen, S. B., Trksak, G. H., MacLean, R. R., Lukas, S. E., 2013. The hypnotic zolpidem increases the synchrony of BOLD signal fluctuations in widespread brain networks during a resting paradigm. NeuroImage 70, 211–222.

Lipták, T., 1958. On the combination of independent tests. A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közléményei 3, 171–197.

Loughin, T., 2004. A systematic comparison of methods for combining p-values from independent tests. Computational Statistics & Data Analysis 47 (3), 467–485.

Marcus, R., Peritz, E., Gabriel, K. R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63 (3), 655.

Meier, U., 2006. A note on the power of Fisher's least significant difference procedure. Pharmaceutical Statistics 5 (4), 253–263.

Mudholkar, G. S., George, E. O., 1979. The logit statistic for combining probabilities. In: Rustagi, J. (Ed.), Symposium on Optimizing Methods in Statistics. Academic Press, New York, pp. 345–366.

Nichols, T., 2012. Multiple testing corrections, nonparametric methods, and random field theory. NeuroImage 62 (2), 811–815.

Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J.-B., 2005. Valid conjunction inference with the minimum statistic. NeuroImage 25 (3), 653–60.

Nichols, T., Hayasaka, S., 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical Methods in Medical Research 12 (5), 419–46.

Oosterhoff, J., 1969. Combination of one-sided statistical tests. Mathematisch Centrum, Amsterdam, The Netherlands.

Owen, A. B., 2009. Karl Pearson's meta-analysis revisited. The Annals of Statistics 37 (6B), 3867–3892.

Pantazis, D., Nichols, T. E., Baillet, S., Leahy, R. M., 2005. A comparison of random field theory and permutation methods for the statistical analysis of MEG data. Neuroimage 25 (2), 383–94.

41

Pearson, K., 1933. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. Biometrika 25 (3/4), 379–410.

Pesarin, F., 1990. On a nonparametric combination method for dependent permutation tests with applications. Psychotherapy and Psychosomatics 54, 172–179.

Pesarin, F., 1992. A resampling procedure for nonparametric combination of several dependent tests. Journal of the Italian Statistical Society 1 (1), 87–101.

Pesarin, F., 2001. Multivariate Permutation Tests, with Applications in Biostatistics. John Wiley and Sons, West Sussex, England, UK.

Pesarin, F., Salmaso, L., 2010a. Finite-sample consistency of combination-based permutation tests with application to repeated measures designs. Journal of Nonparametric Statistics 22 (5), 669–684.

Pesarin, F., Salmaso, L., 2010b. Permutation Tests for Complex Data: Theory, Applications and Software. John Wiley and Sons, West Sussex, England, UK.

Petrovic, P., Kalso, E., Petersson, K. M., Ingvar, M., 2002. Placebo and opioid analgesia–imaging a shared neuronal network. Science 295 (5560), 1737–1740.

Pillai, K. C. S., 1955. Some new test criteria in multivariate analysis. The Annals of Mathematical Statistics 26 (1), 117–121.

Reynolds, D. V., 1969. Surgery in the rat during electrical analgesia induced by focal brain stimulation. Science 164 (878), 444–445.

Rosenthal, R., 1978. Combining results of independent studies. Psychological Bulletin 85 (1), 185–193.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G. E., Wager, T. D., 2014. Representation of aversive prediction errors in the human periaqueductal gray. Nature Neuroscience 17 (11), 1607–1612.

Roy, S. N., Jun. 1953. On a heuristic method of test construction and its use in multivariate analysis. The Annals of Mathematical Statistics 24 (2), 220–238.

Scheffé, H., 1959. The Analysis of Variance. John Wiley and Sons, New York.

Shaffer, J. P., 1986. Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association 81 (395), 826–831.

Smith, S. M., Nichols, T. E., 2009. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44 (1), 83–98.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., Jr., R. M. W., 1949. The American Soldier: Adjustment During Army Life (Volume 1). Princeton University Press, Princeton, New Jersey, USA.

Taylor, J., Tibshirani, R., 2006. A tail strength measure for assessing the overall univariate significance in a dataset. Biostatistics 7 (2), 167–81.

The MathWorks Inc., 2013. MATLAB Version 8.1 (R2013a). Natick, Massachusetts.

Thomas, A. G., Dennis, A., Rawlings, N. B., Stagg, C. J., Matthews, L., Morris, M., Kolind, S. H., Foxley, S., Jenkinson, M., Nichols, T. E., Dawes, H., Bandettini, P. A., Johansen-Berg, H., 2015. Multi-modal characterization of rapid anterior hippocampal volume increase associated with aerobic exercise. NeuroImage (submitted).

Timm, N. H., 2002. Applied multivariate analysis. Springer, New York.

Tippett, L. H. C., 1931. The methods of statistics. Williams and Northgate, London.

Tracey, I., Ploghaus, A., Gati, J. S., Clare, S., Smith, S., Menon, R. S., Matthews, P. M., 2002. Imaging attentional modulation of pain in the periaqueductal gray in humans. The Journal of Neuroscience 22 (7), 2748–2752.

Tukey, J. W., 1949. Comparing individual means in the analysis of variance. Biometrics 5 (2), 99–114.

Uludağ, K., Roebroeck, A., 2014. General overview on the merits of multimodal neuroimaging data fusion. NeuroImage 102, 3–10.

van Zwet, W., Oosterhoff, J., 1967. On the combination of independent test statistics. The Annals of Mathematical Statistics 38 (3), 659–680.

Šidák, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62 (318), 626–633.

Westberg, M., 1985. Combining independent statistical tests. The Statistician 34 (3), 287–296.

Westfall, P. H., Troendle, J. F., 2008. Multiple testing with minimal assumptions. Bio-

42

metrical Journal 50 (5), 745–55.

Westfall, P. H., Young, S. S., 1993. Resampling-Based Multiple Testing: Examples And Methods for p-Value Adjustment. John Wiley and Sons, New York.

Whitlock, M. C., 2005. Combining probability from independent tests: the weighted z-method is superior to Fisher's approach. Journal of Evolutionary biology 18 (5), 1368–73.

Wilkinson, B., 1951. A statistical consideration in psychological research. Psychological Bulletin 48 (3), 156–8.

Wilks, S. S., 1932. Certain generalizations in the analysis of variance. Biometrika 24 (3), 471–494.

Wilson, E. B., 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association 22 (158), 209–212.

Winer, B. J., 1962. Statistical Principles in Experimental Design. McGraw-Hill, New York.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., Nichols, T. E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–97.

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E., Smith, S. M., 2015, *in press.* Multi-level block permutation. NeuroImage.

Won, S., Morris, N., Lu, Q., Elston, R. C., 2009. Choosing an optimal method to combine p-values. Statistics in Medicine 28 (11), 1537–53.

Woo, C.-W., Krishnan, A., Wager, T. D., 2014. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. NeuroImage 91, 412–9.

Wu, S. S., 2006. Combining univariate tests for multivariate location problem. Communications in Statistics: Theory and Methods 35 (8), 1483–1494.

Zaykin, D. V., 2011. Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. Journal of Evolutionary Biology 24 (8), 1836–1841.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S., 2002. Truncated product method for combining p-values. Genetic Epidemiology 22 (2), 170–85.

Zhu, D., Zhang, T., Jiang, X., Hu, X., Chen, H., Yang, N., Lv, J., Han, J., Guo, L., Liu, T., 2014. Fusing DTI and fMRI data: A survey of methods and applications. NeuroImage 102, 184–191.

Figure captions

From next page.

Figure 1: *(a)* Rejection region of a union–intersection test (UIT) based on two independent t-tests. The null is rejected if either of the partial tests has a statistic that is large enough to be qualified as significant. *(b)* Rejection region of an intersection–union test (IUT) based the same tests. The null is rejected if both the partial tests have a statistic is large enough to be qualified as significant.

Figure 2: The original NPC algorithm combines non-parametric p-values and, for imaging applications, requires substantial amount of data storage space. Two modifications simplify the procedures: (I) the statistic t_k for each partial test k is transformed into a related quantity u_k that has a behaviour similar to the p-values, and (II) the combined statistic is transformed to a variable that follows approximately a normal distribution, so that spatial statistics (such as cluster extent, cluster mass, and TFCE) can be computed as usual. The first simplification allows the procedure to run in a single phase, without the need to retrieve data for the empirical distribution of the partial tests.

45

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 3: *Upper row:* Rejection regions for the combination of two partial tests using four different combining functions, and with the p-values assessed parametrically (Table 1). The regions are shown as function of the p-values of the partial tests (p_k). *Middle row:* Rejection regions for the same functions with the modification to favour alternative hypotheses with concordant directions. *Lower row:* Rejection regions for the same functions with the modification to ignore the direction altogether, that is, for two-tailed partial tests.

46

Figure 4: The simulations A–D. Each was constructed with a set of K partial tests, a number of which (K_s) had synthetic signal added.

47

Figure 5: Histograms of frequency of p-values for the simulation without signal in either of the two partial tests (upper panel, blue bars) or with signal in both (lower panel, green bars). The values below each plot indicate the height (in percentage) of the first bar, which corresponds to p-values smaller than or equal to 0.05, along with the confidence interval (95%, italic). Both original and modified NPC methods controlled the error rates at the nominal level, and produced flat histograms in the absence of signal. The histograms suggest similar power for both approaches. See also the Supplemental Material.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
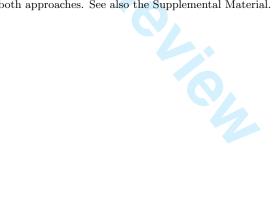51
52
53
54
55
56
57
58
59
60

Figure 6: Bland–Altman plots comparing the original and modified NPC, for both uncorrected and corrected p-values, without signal in either of the two partial tests (upper panel, blue dots) or with signal in both (lower panel, green dots). The values below each plot indicate the percentage of points within the 95% confidence interval ellipsoid. For smaller sample sizes and non-Gaussian error distributions, the methods differ, but the differences become negligible as the sample size increases. In the presence of signal, the modification caused increases in power, particularly for the corrected p-values.

49

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 7: Performance of the modified NPC with four representative combining functions (Tippett, Fisher, Stouffer, and Mudholkar–George) and of one CMV (Hotelling's T^2), using normal or skewed errors, and using permutations (EE), sign flippings (ISE), or both. All resulted in error rates controlled at or below the level of the test. The Tippett and Fisher were generally the most powerful, with Tippett outperforming others with signal present in a small fraction of the tests, and with Fisher having the best power in the other settings.

50

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
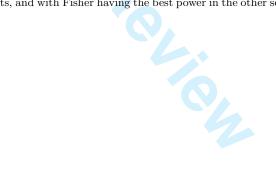51
52
53
54
55
56
57
58
59
60

Figure 8: Without combination, and with correction across voxels (MTP-I), no significant results were observed at the group level for any of the three tests. Combination using the methods of Fisher, Stouffer and Mudholkar–George (M–G), however, evidenced bilateral activity in the insula in response to hot, painful stimulation. A classical multivariate test, Hotelling's T^2, as well as the Tippett method, failed to identify these areas. An intersection-union test (conjunction) could not locate significant results; such a test has a different null hypothesis that distinguishes it from the others. Images are in radio-logical orientation. For cluster-level results, comparable to Brooks et al. (2005), see the Supplementary Material.

51

Figure 9: Examples of inconsistent combining functions for testing the global null hypothesis: *(a)* Addition of p-values for the partial tests (Edgington, 1972); *(b)* Maximum of p-values for the partial tests, with the p-value computed as T^K (Friston et al., 1999, 2005); *(c)* Maximum of p-values for the partial tests, but with the p-value computed as T (Nichols et al., 2005). While the last is not appropriate for testing the global null, it is appropriate for the conjunction null.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 10: *Upper row:* Inadmissible versions of the four consistent combining functions shown in Figure 3 (in the same order). *Lower row:* Inadmissible versions of the three inconsistent combining functions shown in Figure 9 (in the same order). These inadmissible functions arise if one attempts to favour alternatives with the same sign while performing two-tailed partial tests.