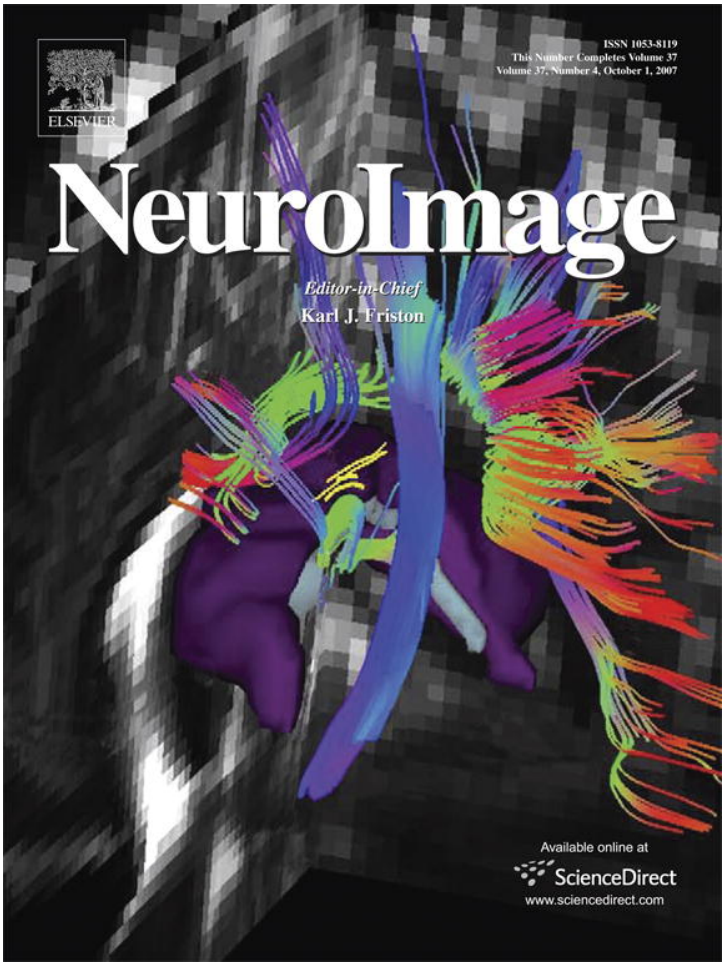

ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 37 (2007) 1178–1185

Conjunction group analysis: An alternative to mixed/random effect analysis

Ruth Heller,[a] Yulia Golland,[b] Rafael Malach,[c] and Yoav Benjamini[a,*]

[a]*Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel*
[b]*Department of Psychology, Hebrew University of Jerusalem, Jerusalem, Israel*
[c]*Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel*

Received 18 April 2007; revised 9 May 2007; accepted 11 May 2007
Available online 9 June 2007

We address the problem of testing in every brain voxel *v* whether at least *u* out of *n* conditions (or subjects) considered shows a real effect. The only statistic suggested so far, the maximum *p*-value method, fails under dependency (unless *u*=*n*) and in particular under positive dependency that arises if all stimuli are compared to the same control stimulus. Moreover, it tends to have low power under independence. For testing that at least *u* out of *n* conditions shows a real effect, we suggest powerful test statistics that are valid under dependence between the individual condition *p*-values as well as under independence and other test statistics that are valid under independence. We use the above approach, replacing conditions by subjects, to produce informative group maps and thereby offer an alternative to mixed/random effect analysis.

Keywords: Combining tests; Dependent contrasts; False discovery rate; Global null hypothesis; Meta-analysis; Multiple comparisons; Pooled significance values

Introduction

Pooling together significance values evaluated under different conditions enables the researcher to either (1) make a stronger scientific statement, as is the case for example if the brain region responds to all conditions, or (2) gain statistical power. As an example for the latter, suppose there is only weak evidence that this brain region responds to each condition, so the individual tests are not convincing evidence by themselves, but since this evidence is consistent across all conditions tested, the evidence is very convincing. In fMRI, the need to pool together significance values arises when looking for brain regions that respond to all of a set of different conditions or to at least some of these conditions. Interestingly, we suggest that the same need arises in multi-subject analysis in order to create an informative group map.

In a typical analysis, a *p*-value map is produced and then a threshold is computed so that only voxels with *p*-values below the threshold are highlighted as findings. The need to correct for multiple comparisons (rather than choose an arbitrary threshold) has been recognized and the false discovery rate (FDR), which is the expected proportion of false rejections, has been adopted in the fMRI community as an appropriate error measure (see Genovese et al., 2002). FDR controlling procedures have been implemented in software packages such as SPM and Brain Voyager. However, if we have two activation maps, and the cut-off of each is calculated separately to control the FDR of the map at q, then the FDR level of the conjunction of activation maps may be larger than $2q$. To see this, consider the following extreme example: assume that the two activation maps overlap only in a proportion q of activations of each, and this is exactly the region of false discoveries for each of the maps. In this example, the conjunction of activation maps is comprised entirely of false discoveries. Choosing more conservative thresholds for each map so that the FDR level of the combined map is maintained is either very conservative, reverting to family-wise error (FWE) control in each of the maps, or very complicated. Thus when studying several *p*-value maps jointly, we suggest to combine *p*-values first and then threshold this pooled map to control the FDR, rather than first threshold each map and then combine the activation maps.

Methods for combining *p*-value maps have already been introduced into the fMRI literature. The maximum *p*-value has been suggested by Nichols et al. (2005) for addressing the problem of looking for brain regions that respond to all of a set of different conditions (i.e., the conjunction hypothesis). Friston et al. (2005) argue that it is enough to look at brain regions that respond to at least a certain number of different conditions and suggest adjusting the distribution of the maximum *p*-value accordingly.

We can view multi-subject analysis as a problem in combining individual *p*-value maps. Traditionally, the fixed effects model and mixed (random) effects model have been in use for combining brains. The fixed effect analysis answers the question whether at least for one subject there is activation in the voxel. The mixed (random) effects analysis answers the question whether the average

* Corresponding author. Fax: +972 3 640 9357.
E-mail address: ybenja@post.tau.ac.il (Y. Benjamini).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.05.051

activation level is higher than 0, as compared to the variability in activation levels between subjects.

The mixed effect model is usually advocated (see e.g., Mumford and Nichols, 2006), since investigators want to infer on the population from which the sample was drawn. For the usual mixed analysis for inference on the population to be valid, it is required that the subject specific effects come from the same normal distribution. However, Mumford and Nichols (2006) mention that it is known that human anatomy is highly variable, and two brains cannot necessarily be matched gyri-to-gyri even when the registration is done manually. The implication from such an observation is that we face at best a mixture of normally distributed effects and zero effects—a highly non-normal distribution. To overcome these limitations of inter-subject registration, spatial smoothing is applied to blur out residual anatomical differences. So the inference can only be on the smoothed signal rather than on the original signal. Thirion et al. (2007) give evidence that even after smoothing the normal distribution assumption is not satisfied in many voxels. Typically, only a proportion of subjects will activate the voxel (see Friston et al. (1999) for a discussion on estimating this proportion). Therefore the assumption of normality required for mixed model inference is violated. Moreover, and possibly as a result of this violation, the mixed model inference has low power to discover activations, especially for small sample sizes.

The fixed effect model usually produces large activation areas. Alternatives to the fixed effect model are suggested in Lazar et al. (2002), where a survey of additional ways to combine brains is given. However, all one can conclude from the fixed effect model or the suggested alternatives is that at least one subject in the sample activated a detected voxel. In the extreme case, the group map may be driven entirely by one subject. Hence, we cannot say that the group's response is significant but only that at least one subject showed activation. Although the statistical power is large, this is a very weak scientific finding from multiple subjects. However, a generalization of the fixed effect analysis that answers whether at least a proportion of subjects activate the voxel is more scientifically meaningful. Moreover, a further generalization to a statement on the population proportion may be even more meaningful (we come back to this point in Final remarks).

In this study, we address the general question of partial conjunction testing, whether at least u out of n conditions or subjects activated the voxel. We suggest combining the p-values at the voxel level into a test statistic that is valid for positive (and many other) dependencies. It coincides with the maximum p-value when testing the (full) conjunction hypothesis. Applying an FDR controlling procedure on this pooled p-value map will control the FDR at the desired level. For the multi-subject problem, or more generally if we have independent p-value maps, we will suggest a generalization of the approach in Lazar et al. (2002), that will enable the investigator to conclude on the least number of subjects that show activation in every voxel. Although from this generalization we only have limited inference, the scientific finding is stronger than that of the fixed model.

This question has been addressed in Friston et al. (2005) assuming independence between the p-value maps, where they suggest using the observed minimum t statistic (i.e., maximum p-value), called $T_{\min}$, and compare it to a threshold based on the distribution of the minimum of $n-u+1$ independent t random variable that controls for the probability of making a type I error even in a single voxel (Family-wise error rate—FWE). The theory of random fields that is traditionally used to control for FWE when testing a single contrast remains valid when n independent random fields are combined into the $T_{\min}$ statistical map (see Worsley and Friston, 2000), making this choice a very natural starting point for developing methods for conjunction analysis. However, this method has very low power even if the brain region responds to all but one condition, as mentioned in McNamee and Lazar (2004). Similarly, Loughin (2004) showed that this method is not able to reject the null hypothesis whenever at least one p-value becomes too large, regardless of how small the other p-values may be. Moreover, unless the conjunction hypothesis is tested, this method is only valid for independent test statistics. Alas, in many neuroimaging studies different conditions are compared to the same control condition, leading to positive dependency between the condition p-values.

The paper is organized as follows. First we present the new proposals for pooling p-values. Then we will give recipes for making multi-contrast and multi-subject inference respectively. We proceed to show the gain in power in using the proposed pooled p-values over using $T_{\min}$ (i.e., the maximum p-value) even when the test statistics are independent via simulations and apply the suggested procedures on the results of an fMRI experiment. Finally, we give our conclusions and additional remarks.

Pooling p-values for testing partial conjunctions

Let k be the (unknown) number of conditions or subjects that show real effect. The problem of testing in every brain voxel v whether at least u out of n conditions or subjects considered show real effects, can be generally stated as follows:

$$H_{0v}^{u/n}: k<u \text{ versus } H_{1v}^{u/n}: k\geq u \tag{1}$$

We shall call $H_{0v}^{u/n}$ the partial conjunction null hypothesis.

In Nichols et al. (2005) and Friston et al. (2005) the n-out-of-n hypothesis, that asks whether all conditions show a real effect in voxel v, is referred to as the test of the Conjunction Null (option "Conjn" in SPM5)

$$H_{0v}^{n/n}: k<n \text{ versus } H_{1v}^{n/n}: k=n \tag{2}$$

and the 1-out-of-n hypothesis test, that tests whether one or more conditions show a real effect in voxel v, is referred to as the test of the Global Null (option "Global" in SPM5)

$$H_{0v}^{1/n}: k=0 \text{ versus } H_{1v}^{1/n}: k\geq 1 \tag{3}$$

Option "Intermed" in SPM5 refers to the case that $1<u<n$.

The researcher should decide what choice of u, $u=1, \ldots, n$ is relevant for the problem at hand, depending on the desired inference. To infer a conjunction of real effects (e.g., brain regions that respond to all of a set of different conditions) (2) should be tested. If finding real effects in less than all conditions is scientifically sufficiently convincing, then taking $u<n$ may be appropriate. For example if n is the number of subjects and each subject is tested for the same effect, then taking $u=n/2$ we may infer from (1) that the effect is real for at least half of the subjects. As another example, for testing n similar mental activities, we may infer from (3) that at least one of them activated the voxel.

Let $p_v^1, \ldots, p_v^n$ be the p-values for testing the n conditions (or subjects) at voxel v. The method to combine the p-values for testing $H_{0v}^{u/n}$ into a valid p-value at the voxel level should depend on the statistical relation between the n p-values in each voxel: if they are independent or dependent, we suggest below ways of combining the p-values. If they are independent, the possibility opens up to utilize many additional tests, and some are highlighted below. Note that the choice of appropriate method for combining p-values relies on the statistical relation between the p-value maps, not the statistical relation of p-values within each map.

Pooling dependent p-values

For the most general dependency structure, we can always rely on the Bonferroni approach. For testing $H_{0v}^{u/n}$, the combined p-value is

$$p_v^{u/n} = (n-u+1)p_v^{(u)} \tag{4}$$

As always, the penalty for its general applicability will be lower power. We therefore suggest the following combined p-value:

$$p_v^{u/n} = \min\left\{(n-u+1)p_v^{(u)}, \frac{(n-u+1)}{2}p_v^{(u+1)}, \ldots, \frac{(n-u+1)}{n-u}p_v^{(n-1)}, p_v^{(n)}\right\} \tag{5}$$

where $p_v^{(u)}$ is the u-th largest p-value in $\{p_v^i: i=1, \ldots, n\}$. This is a generalization of the Simes p-value which originally addresses the global null, i.e., the case $u=1$.

This method can be used when the dependency is more structured. For example, when several conditions are compared to the same control condition. For these and more general structures of positive dependence, Theorem 1 in Benjamini and Heller (2007) shows the combined p-value in Eq. (5) to be a valid one, in the sense that the p-value is uniformly distributed under the partial conjunction null or has a stochastically larger distribution than the uniform. Similarly, Theorem 2 in Benjamini and Heller (2007) shows that for any dependency structure, the combined p-value in Eq. (4) is valid.

For example, suppose that 3 conditions end up with p-values 0.5, 0.022, and 0.01. For testing that all three conditions show a real effect we use $p_v^{3/3}=p_v^{(3)}=0.5$, for testing that at least one condition shows a real effect we use $p_v^{1/3}=\min\{3p_v^{(1)}, 1.5p_v^{(2)}, p_v^{(3)}\}=0.03$ and for testing that at least two conditions show a real effect we use $p_v^{2/3}=\min\{2p_v^{(2)}, p_v^{(3)}\}=0.044$.

Pooling independent p-values

Lazar et al. (2002) discussed several ways available for combining p-values in order to test the global null (i.e., $u=1$) in the context of fMRI analysis. Lemma 1 in Benjamini and Heller (2007) tells us that if we apply these combining p-value methods on the $n-u+1$ largest p-values, the resulting p-value is valid for testing $H_{0v}^{u/n}$. Based on the comparison of the performance of these various combining methods in Lazar et al. (2002) and McNamee and Lazar (2004), we choose to give particular attention to the Stouffer and Fisher methods for combining p-values.

Let $z_{v(1)} \le \ldots \le z_{v(n)}$ be the sorted z-scores corresponding the n p-values $z_{vi}=\Phi^{-1}(1-p_v^i)$. For testing (1), the p-value motivated by the Stouffer method for combining p-values is given by

$$p_v^{u/n} = 1 - \Phi\left(\frac{\Sigma_{i=1}^{n-u+1} z_{v(i)}}{\sqrt{n-u+1}}\right) \tag{6}$$

and the p-value motivated by the Fisher method for combining p-values is given by

$$p_v^{u/n} = P\left(\chi^2_{2(n-u+1)} \ge -2\sum_{i=u}^{n} \log p_v^{(i)}\right) \tag{7}$$

As an example, suppose again that the three independent subject p-values are 0.5, 0.022, and 0.01. For testing that at least one subject shows a real effect, the p-value based on the Stouffer method is $1-\Phi((\tilde{\Phi}^{-1}(0.5)+\tilde{\Phi}^{-1}(0.022)+\tilde{\Phi}^{-1}(0.01))/\sqrt{3})=0.0061$ and the one based on the Fisher method is $P(\chi_6^2 \ge -2(\log(0.5)+\log(0.022)+\log(0.01)))=0.0057$, so the pooled significance value is smaller than all individual p-values. For testing that at least two subjects show a real effect, the p-value based on the Stouffer method $1-\Phi((\tilde{\Phi}^{-1}(0.5)+\tilde{\Phi}^{-1}(0.022))/\sqrt{2}) = 0.077$ and the one based on the Fisher method is $P(\chi_4^2 \ge -2(\log(0.5)+\log(0.022)))=0.061$.

Thresholding a partial conjunction *p*-value map

Once the partial conjunction p-values have been computed at each voxel, a p-value map has been created. At this stage we have to decide how to threshold the map. While choosing the threshold to control the probability of making even one error (i.e., the family-wise error, FWE) is feasible, it is more complicated and more conservative than in the usual setting of thresholding a single map. We choose to control the FDR, as explained in the introduction. On single p-value maps from nueroimaging data, Genovese et al. (2002) argue that the FDR procedure controls the FDR at level q. Their reasoning is that while strict independence between voxel p-values is hard to verify and will often fail with neuroimaging data, the correlations are local and tend to be positive. Benjamini and Yekutieli (2001) prove that for such positive dependency, the BH procedure (Procedure 1 below) controls the FDR. Since $p_v^{u/n}$ is an increasing function of $p_v^1 \le \ldots \le p_v^n$, the positive dependency carries over to the pooled p-value map if Eqs. (6) or (7) are used for combining p-values (see Theorem 3 in Benjamini and Heller (2007) for a proof), so the procedure below controls the FDR at level q. While it is quite likely for positive-dependent maps that this procedure controls the FDR also when the p-values are pooled using Eq. (5), this result has not been formally proved. However, Storey et al. (2004) proves that for local dependencies, this procedure controls the FDR asymptotically. This reasoning carries over to the pooled p-value map for any combining method, since the correlations of the pooled p-values will remain local, so the procedure below controls the FDR (asymptotically) at level q.

Procedure 1. The FDR procedure for testing $H_{0v}^{u/n}$:

1. For every voxel v, $v=1, \ldots, V$, let $p_v^{u/n}$ be the voxel pooled p-value using one of (4)–(7) as appropriate.

2. Sort the p-values $p^{u/n}_{(1)} \le \ldots \le p^{u/n}_{(V)}$.
3. Let $k = \max\{j : p^{u/n}_{(j)} \le (j/V)q\}$. Reject all voxels with $p^{u/n}_v \le (k/V)q$.

Let us now take this general scheme and apply it to multi-contrast analysis. Let n be the number of conditions considered. Calculate the full general linear model (GLM). Next, calculate the p-value map for every contrast of interest separately. For example, if we have three conditions A, B and C, then the p-value map for contrast A will be calculated from the GLM that includes A, B and C. Next, specify the scientifically appropriate number of conditions u ($1 \le u \le n$) for testing $H^{u/n}_{0v}$. Based on the relationship between the contrasts and the experimental design, assess whether the contrast p-values are independent or if not identify whether the dependency structure allows the use of Eq. (5) for combining the contrast p-values (if not, use Eq. (4)). Finally, apply Procedure 1.

Multi-subject analysis

As discussed before, the above methodology of partial conjunction analysis gives the opportunity to create informative group maps by combining the maps of individual subjects while controlling for the FDR across the map.

In preparation for the multi-subject analysis, all subjects need first to be transformed into a common space, e.g., by warping the brain images of the subjects onto a common atlas, using Talairach coordinates. Let n be the number of subjects. First, specify the scientifically appropriate hypothesis per voxel for every subject. The scientifically appropriate hypothesis tests per voxel for every subject can be of the multi-contrast type. Next, calculate a valid p-value per voxel for every subject. Specify the number of people u ($1 \le u \le n$) for testing $H^{u/n}_{0v}$. Since the subjects produce independent observations, use Eqs. (6) or (7) to combine the voxel p-values. Finally, apply Procedure 1.

Procedure 1 can be repeated for all u values of interest (e.g., $u = 1, \ldots, n$). Since the activation map of at least u subjects is a subset of the activation map of at least u' subjects for any $u' < u$, the n activation maps can be superimposed on the same display. For an example of such a display, see Fig. 4 which will be discussed in detail in Applications.

A simulated example

We considered different settings in order to compare the currently used statistic $T_{\min}$ with our proposed methods. A $10 \times 10 \times 10$, $V = 1000$, voxel image was filled with unit variance Gaussian noise for n independent subjects, $n = 3, 5, 10, 15$. For a total activation area of 100 voxels a signal of size μ was added to k subjects in the same voxel, $k = n/2, \ldots, n$.

We pooled the p-values using (5) or (7), then we computed the resulting map threshold using the suggested FDR controlling procedure. For comparison, we computed the p-value map based on $T_{\min}$ and computed the resulting map threshold using the same FDR controlling procedure. Recall that under dependency, the FDR of $T_{\min}$ is not controlled unless the conjunction null is tested, and can be as high as 1. Therefore we only show results under independence.

The simulations show that the power when pooling the p-values using Eqs. (5) or (7) is much larger than when using $T_{\min}$, unless all n subjects activate the voxel (in which case the power is similar). For example, Fig. 1 shows the power versus the signal size μ (standard errors at most 0.0004 for $T_{\min}$ and 0.004 for the other two methods) when testing $H^{5/10}_{0v}$ in a simulation setting in which 7 subjects activate the same voxel (left figure) and $H^{3/10}_{0v}$ in a simulation setting in which 3 subjects activate the same voxel (right figure). $T_{\min}$ has very low power in both settings because even when $H^{5/10}_{0v}$ is false 3 subjects are inactive and when $H^{3/10}_{0v}$ is false 7 subjects are inactive. The identifiable factors that affect the choice between the two combining methods in terms of power are outside the scope of this manuscript.

Applications

High-order object areas

We have applied the new statistical approach to data obtained from 19 subjects, which participated in a well established, visual

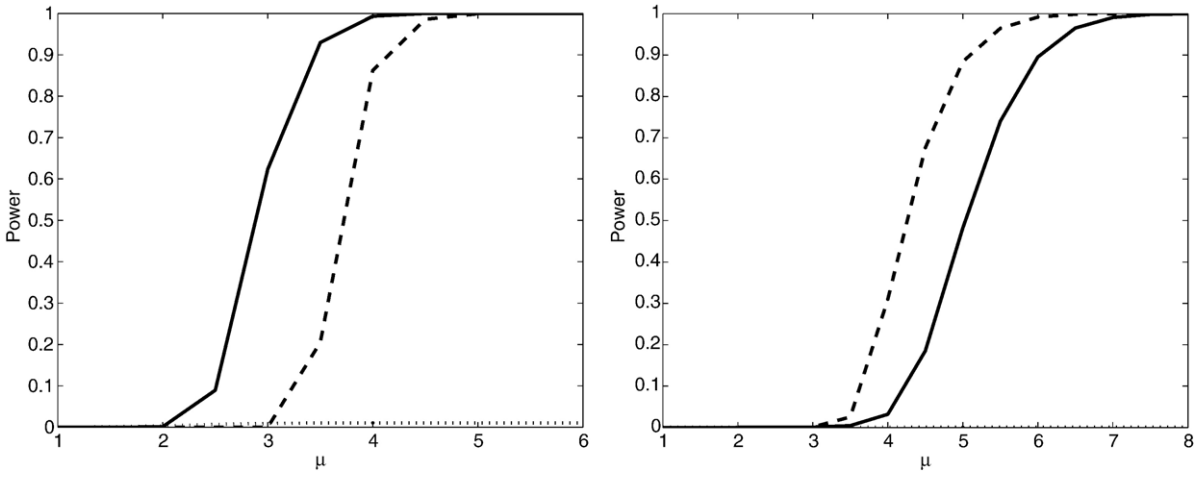

Fig. 1. Power as a function of signal size μ when the FDR level is 0.05 and the combining method is based on (a) Eq. (7) (solid line) (b) Eq. (5) (dashed line) and (c) $T_{\min}$ (dotted line). Left, testing $H^{5/10}_{0v}$ in a simulation setting in which 7 subjects activate the same voxel; the most powerful analysis method is clearly (a), but both (a) and (b) are much more powerful than (c). Right, testing $H^{3/10}_{0v}$ in a simulation setting in which 3 subjects activate the same voxel; the most powerful analysis method is clearly (b), but both (a) and (b) are much more powerful than (c).

localizer paradigm (e.g., Hasson et al., 2003; Levy et al., 2001). An interleaved short block design was used in the experiment. The visual stimuli included line drawings of faces, buildings, common man-made objects, and geometric patterns. Nine images of the same type were presented in each epoch, which lasted 9 s, followed by a 6 s blank screen. A central red fixation point was present throughout the experiment. During the experiment, one or two consecutive repetitions of the same image occurred in each epoch. The subject's task was to covertly report whether the presented stimulus was identical to the previous stimulus or not.

A well studied cortical region which can be consistently revealed using this paradigm, is the object-related lateral occipital complex (LOC) whose most robust functional signature is a preferential activation to images of objects compared to texture patterns. In the original studies (Malach et al., 1995) of the LOC, it was pointed out that the region also showed a preferential activation to other images, such as faces, so it could be of interest to perform a conjunction analysis on different object categories and see if the "core" LOC showed a conjunction response to different object categories. For that aim we performed a multi-contrast analysis on the data of one representative subject. The contrasts included in the analysis were: faces versus patterns, objects versus patterns, and houses versus patterns. For comparison we also produced a GLM analysis on these contrasts that tests whether the average effect of faces, objects and houses is greater than that of patterns. The results are presented in Fig. 2. As can be seen, the new multi-contrast analysis (Fig. 2, bottom) is much more informative compared to the conventional GLM (Fig. 2, top) since it reveals the much wider distribution associated with a single contrast—which includes areas whose selectivity is unique to a single object category, such as the FFA (e.g., Kanwisher et al., 1997), the PPA (e.g., Epstein and Kanwisher, 1998) and other object-related regions (for review see e.g., Hasson et al., 2003; Malach and Levy, 2002). However, when a conjunction of at least two categories is considered (the union of yellow and blue regions) or of all three categories (the blue region), then the delineated regions shrink and become confined to the typical LOC boundaries (Malach et al., 1995; Malach and Levy, 2002).

Another critical issue is to what extent the neuro-anatomical locations of the LOC reproduce in brains of different individuals. Note that precise co-localization is unlikely due to the inter-subject variability in cortical organization. However, one would expect such reproducibility to occur across some subjects, and particularly if the mapped regions are large. As can be seen in the bottom of Fig. 3, the new conjunction analysis nicely reveals significant activations in the expected location of LOC including its larger LO subdivision as well as its ventral pFS focus, evident in the right hemisphere (Malach et al., 1995). Note in white outline the much

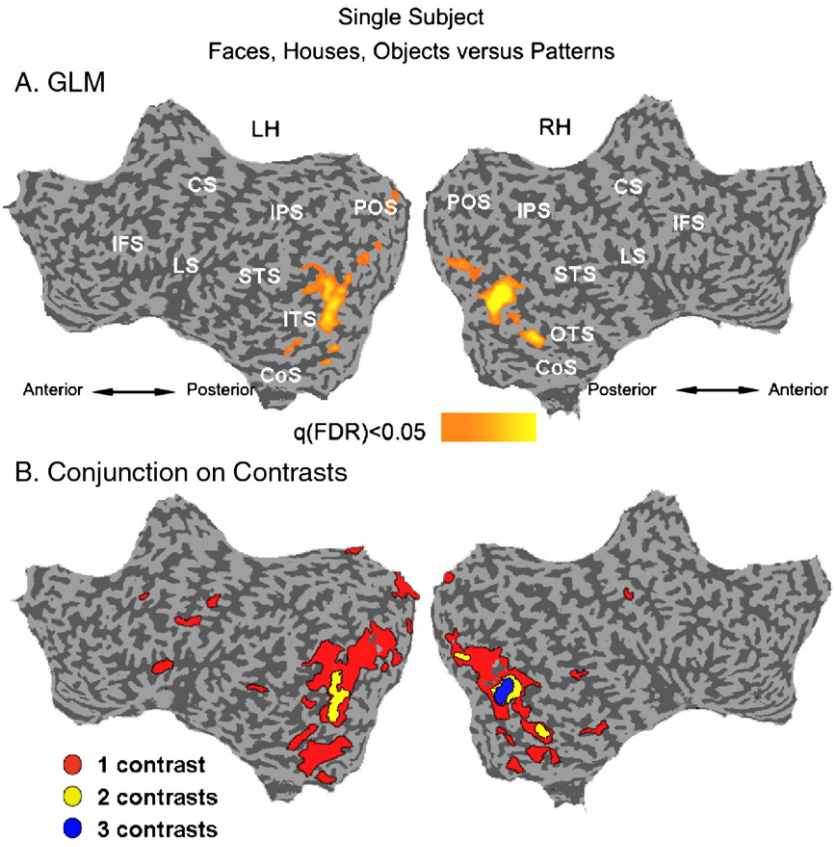

Fig. 2. Activation maps for a single subject presented on unfolded cortical hemispheres. Top: GLM testing whether the average of the faces, objects and houses coefficients is larger than that of patterns with FDR <0.05. Bottom: blue regions activated in all three contrasts with FDR <0.05; yellow or blue regions activated in at least two contrasts with FDR <0.05; red, yellow, or blue regions activated in at least one contrast with FDR <0.05. POS: parietal occipital sulcus; IPS: inferior parietal sulcus; STS: superior temproal sulcus; ITS: inferior temporal sulcus; CoS: collateral sulcus; OTS: occipito temproal sulcus; LS: lateral sulcus; IFS: inferior frontal sulcus; CS: central sulcus.

larger spread of activation when the constraint is relaxed to at least a single subject showing significant activation. The top of Fig. 3 presents the mixed effect analysis. In reporting activation maps from a mixed effect analysis, the standard interpretation is that the regions detected show where the average population activity lies. Comparison with the bottom Fig. 3 reveals that some of the regions discovered by mixed effect analysis were only discovered with the $u=1$ partial conjunction hypothesis, suggesting that these regions are subject specific with strong enough signal to affect the average group activity. As cautioned in the introduction and by the above comparison with our method, the standard interpretation of regions revealed by a mixed effect analysis depends critically on the assumed model.

Face-selective regions

Another well known cortical area, that has been extensively studied, is the Fusiform Face Area (FFA, Kanwisher et al., 1997) which can be typically revealed using the contrast *Face>House*. As can be seen in the bottom of Fig. 4, the new method reveals characteristic, right-lateralized face-selective regions, in lateral occipital region and fusiform gyrus. As expected from the inter-subject brain variability, spatial smoothing enhances the number of subjects which show conjoined activation (Fig. 4 bottom left vs. right). For comparison, a mixed effect analysis was also performed, with FDR <0.05 with and without smoothing.

Computational details

The data presented in Figs. 2–4, top and bottom left, were spatially smoothed with a Gaussian filter of full width half maximum value (FWHM) of 8 mm.

To obtain the multi-subject maps, time series of images of brain volumes for each subject were converted into Talairach space and *z*-normalized. Further details of data analysis and acquisition can be found in Levy et al. (2001).

The general linear model (Friston et al., 1995) consisted of a multiple regression with a regressor for each condition in the experiment, using a box-car shape and assuming a hemodynamic lag of 3 s. The analysis was performed independently for the time course of each individual voxel. After computing the coefficients for all regressors, we performed a test between coefficients of

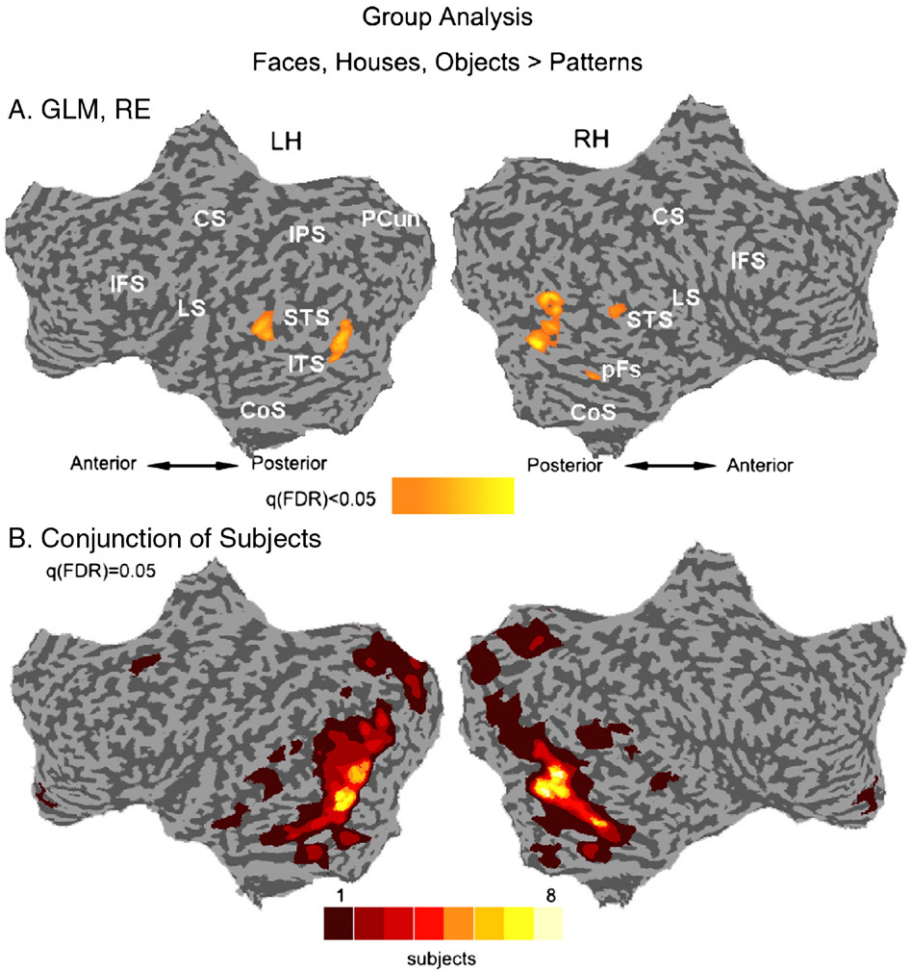

Fig. 3. Activation maps for the group of subjects. Top: mixed effect analysis testing whether the average of the faces, objects, and houses coefficients is larger than that of patterns with FDR <0.05. Bottom: multi-subject analysis testing whether at least one contrast activated the region in at least u subjects with FDR <0.05 ($u=1,\ldots 8$). The intensity represents the minimum number of subjects that activated the region, ranging from at least 1 to 4 in shades of red, and 5 to 8 in shades of orange to white. Regions of high u indicate consistency across subjects. Note that some of the regions detected by the mixed model correspond to $u=1$ only, while others correspond to $u>4$.

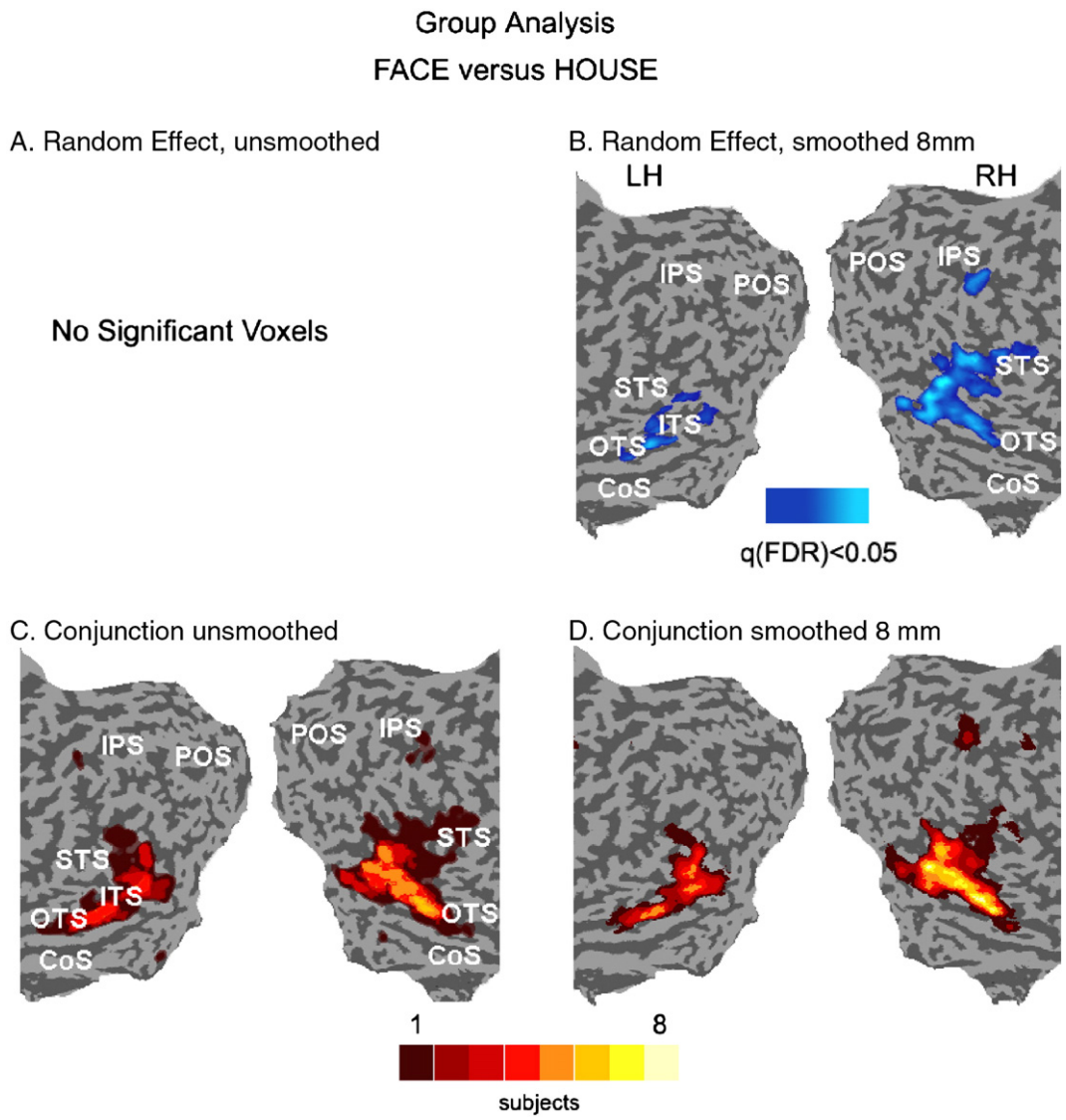

Fig. 4. Activation maps for the group of subjects, testing whether the effect of faces is greater than that of houses. Top: mixed effect analysis with FDR <0.05, un-smoothed (left) and smoothed with 8 mm FWHM (right). Bottom: multi-subject analysis showing activated regions in at least u subjects with FDR <0.05. The intensity corresponds to $u=1,...,8$. Left, un-smoothed image; Right, 8 mm FWHM smoothed image. Regions of high u indicate consistency across subjects. Note that the maximum u is 8 for the smoothed image, but only 6 for the un-smoothed image.

different conditions (e.g., patterns versus buildings, common objects, and faces—the coefficient of the pattern predictor was compared to the average of the building, common object and faces predictors).

The Simes method was used for combining contrast p-values which may be dependent. The p-values motivated by the Fisher and Stouffer methods for combining p-values were used for combining subject p-values utilizing the independence across subjects. While both these methods discovered more activations than the traditional fixed effect analysis, only the Fisher method is shown in Figs. 3 and 4. The decision of when to choose which combining method is a direction for further research.

Final remarks

We found that we can combine brains to discover more active regions than in fixed effect analysis, as well as give a stronger scientific statement. We developed an inference method for testing that at least a proportion u/n of a group of subjects shows a real effect, while controlling for false discoveries using the FDR. Can we further infer on the proportion of the population that has a real effect? A partial answer can be obtained using the following reasoning. For testing whether at least a quarter of the population activate the region, taking into account the background binomial variation, the null hypothesis is rejected with an α level of significance if the estimated sample proportion is larger than $\frac{1}{4}+\sqrt{\frac{\frac{1}{4}\frac{3}{4}}{n}}z_{1-\alpha}$. So if $H_{0v}^{u/n}$ is rejected for $\frac{u}{n}>\frac{1}{4}+\sqrt{\frac{\frac{1}{4}\frac{3}{4}}{n}}z_{1-\alpha}$ we can infer that at least a quarter of the population showed a real effect. However, further attention is needed to the multiplicity of tests at this stage as well. The simplest, yet conservative, adjustment is to use the Bonferroni correction. Less conservative adjustments, thereby reducing the lower bound on the proportion of subjects u/n, can be found in Benjamini and Yekutieli (2005). Another approach is to use ideas

from Friston et al. (1999) to get a lower bound on the proportion of the population that shows an effect in a voxel.

The proportion that shows a real effect will be underestimated because of the inter-subject brain variabilities. Since the analysis is voxel based, it is very sensitive to this variability. Smoothing the data prior to the analysis can increase the overlap of signal between subjects (as shown in the bottom of Fig. 4), but it also smears localized individual signals so the overall evidence may be weaker. Our method of combining brains gives a way to assess the consistency across individuals: the least consistent regions are the regions where we can only say that at least one subject shows an effect, and the consistency increases with the minimum number of subjects u that show a real effect. If we could capture the inter-subject variability better a priori, the estimated proportions would be higher. One step toward this end may be to combine clusters of voxels rather than individual voxels across subjects (see Heller et al. (2006) for controlling the FDR on clusters of voxels). This is a point for further research.

Friston et al. (2005) suggest another approach that depends on the order in which the contrasts are tested: to use small volume adjusted p-values centered on the maximum of the first contrast (e.g., searching over a sphere of 8 mm radius). Note, however, the difficulty in computing the p-values for the second contrast, since these p-values are conditional on the outcome of testing the first contrast. So for example, if after correcting for multiple comparisons all voxels above a threshold t_A are declared as active for contrast A, the p-value for such a voxel when testing for contrast B is $p(T \geq T_B | T_A \geq t_A)$, where T_A and T_B are the test statistics of this voxel. Unless the distribution of these two test statistics is independent, e.g., when contrasts on same data are orthogonal, the p-values cannot be computed since the distribution of T_A under the alternative (i.e., contrast A activates the voxel) is unknown. For finding that at least one (or more) of the dependent contrasts activates the voxel, our multi-contrast approach should be used instead. Note, however, that if the joint distribution of the $n-u+1$ null p-values under $H_{0v}^{u/n}$ was known, a more powerful method could be constructed. Tamhane and Dunnett (1999) give special situations where this is the case. A similar approach may be appropriate in some cases in fMRI.

Acknowledgments

We wish to thank Karl Friston and Tom Nichols for valuable comments on an earlier draft. This study was supported by a grant from the Adams Super Center for Brain Studies, Tel-Aviv University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2007.05.051.

References

Benjamini, Y., Heller, R., 2007. Screening for partial con-junction hypotheses. Technical Report RP-SOR-06-06, URL http://www.math.tau.ac.il/departments/st/.

Benjamini, Y., Yekutieli, Y., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29 (4), 1165–1188.

Benjamini, Y., Yekutieli, Y., 2005. False discovery rate-adjusted multiple confidence intervals for selected parameters. J. Am. Stat. Assoc. 100 (469), 71–93.

Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392, 598–601.

Friston, K., Holmes, A., Poline, J., Grasby, P., Williams, S., Frackowiak, R., Turner, R., 1995. Analysis of fmri time-series revisited. NeuroImage 2 (1), 45–53.

Friston, K., Holmes, A., Worsley, K., 1999. Comments and controversies: how many subjects constitute a study? NeuroImage 10, 1–5.

Friston, K., Penny, W., Glaser, D., 2005. Conjunction revisited. NeuroImage 25, 661–667.

Genovese, C., Lazar, N., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878.

Hasson, U., Harel, M., Levy, I., Malach, R., 2003. Large-scale mirror-symmetry organization of human occipito-temporal object areas. Neuron 37, 1027–1041.

Heller, R., Stanley, D., Yekutieli, D., Rubin, N., Benjamini, Y., 2006. Cluster based analysis of fmri data. NeuroImage 33 (2), 599–608.

Kanwisher, N., McDermott, J., Chun, M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for the perception of faces. Neuroscience 17, 4302–4311.

Lazar, N., Luna, B., Sweeney, J., Eddy, W., 2002. Combining brains: a survey of methods for statistical pooling of information. NeuroImage 16, 538–550.

Levy, I., Hasson, U., Avidan, G., Hendler, T., Malach, R., 2001. Center-periphery organization of human object areas. Nat. Neurosci. 5, 533–539.

Loughin, T., 2004. A systematic comparison of methods for combining p-values from independent tests. Comput. Stati. Data Anal. 47, 467–485.

Malach, R., Levy, I., 2002. The topography of high-order human object areas. Trends Cogn. Sci. 6 (4), 176–184.

Malach, R., Reppas, J., Benson, R., Kwong, K., Jiang, H., Kennedy, W., Ledden, P., Brady, T., Rosen, B., Tootell, R., 1995. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc. Natl. Acad. Sci. U. S. A. 92, 8135–8139.

McNamee, R., Lazar, N., 2004. Assessing the sensitivity of fmri group maps. NeuroImage 22, 920–931.

Mumford, A., Nichols, T., 2006. Modeling and inference of multisubject fmri data. IEEE Eng. Med. Biol. Mag. 25 (2), 42–51.

Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J., 2005. Valid conjunction inference with the minimum statistic. NeuroImage 25, 653–660.

Storey, J., Taylor, J., Siegmund, D., 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. J. R. Stat. Soc., Ser. B Stat. Methodol. 66, 187–205.

Tamhane, A., Dunnett, C., 1999. Stepwise multiple test procedures with biometric applications. Stat. Plan. Inference 82, 55–68.

Thirion, B., Pinel, P., Meriaux, S., Roche, A., Dehaene, S., Poline, J., 2007. Analysis of a large fmri cohort: statistical and methodological issues for group analyses. NeuroImage 35, 105–120.

Worsley, K., Friston, K., 2000. A test for conjunction. Stat. Probab. Lett. 47, 135–140.