NeuroImage 58 (2011) 122–136

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Testing the ICA mixing matrix based on inter-subject or inter-session consistency

Aapo Hyvärinen *

Dept of Mathematics and Statistics, Dept of Computer Science/HIIT, and Dept of Psychology, University of Helsinki, Finland

ARTICLE INFO

Article history:
Received 9 February 2011
Revised 6 May 2011
Accepted 17 May 2011
Available online 17 June 2011

ABSTRACT

Independent component analysis (ICA) is increasingly used for analyzing brain imaging data. ICA typically gives a large number of components, many of which may be just random, due to insufficient sample size, violations of the model, or algorithmic problems. Few methods are available for computing the statistical significance (reliability) of the components. We propose to approach this problem by performing ICA separately on a number of subjects, and finding components which are sufficiently consistent (similar) over subjects. Similarity is defined here as the similarity of the mixing coefficients, which usually correspond to spatial patterns in EEG and MEG. The threshold of what is "sufficient" is rigorously defined by a null hypothesis under which the independent components are random orthogonal components in the whitened space. Components which are consistent in different subjects are found by clustering under the constraint that a cluster can only contain one source from each subject, and by constraining the number of the false positives based on the null hypothesis. Instead of different subjects, the method can also be applied on different recording sessions from a single subject. The testing method is particularly applicable to EEG and MEG analysis.

Introduction

Independent component analysis (ICA) has been successfully used for analyzing brain imaging data. In particular, analysis of data recorded at rest (Beckmann et al., 2005; de Pasquale et al., 2010; Kiviniemi et al., 2003; van de Ven et al., 2004), or during natural stimulation (Bartels and Zeki, 2004; Hasson et al., 2004) has received a lot of attention recently. Such data cannot be easily analyzed by ordinary supervised methods based on regression with a stimulus function, and unsupervised methods such as ICA may be needed.

Despite its success, there is a fundamental problem which has not been satisfactorily solved in the theory of ICA. ICA provides a number of components many of which may be just random effects due to small sample size, noise or other violations of the model, algorithmic problems such as local minima, etc. Thus, the estimation methods should be complemented by testing methods. Few methods are available for computing the statistical significance (reliability) of the components. It has been proposed that one can randomize the data by bootstrapping and see how the ICA results change (Himberg et al., 2004; Meinecke et al., 2002). However, it is not clear how to determine any meaningful thresholds or p-values in such bootstrapping methods. Moreover, even if a component is robust with respect to randomization of the data, its neuroscientific validity is by no means guaranteed if the analysis is done on a single subject.

Group ICA methods attempt to increase the validity of the components by analyzing data from many subjects (Beckmann and Smith, 2005; Calhoun et al., 2009; Esposito et al., 2005). Typically, the goal is to find components which are sufficiently similar over many subjects. Such components are more likely to be of interest for further analysis, although some of them may still be artifacts. However, most group ICA methods, reviewed by (Calhoun et al., 2009), do not provide any selection of components: the number of independent components they compute is given a priori by the investigator. Thus, these methods do not even attempt to analyze the statistical reliability of the components.

An exception is the group ICA method by (Esposito et al., 2005) which rejects components which are not found in sufficiently many subjects in sufficiently similar form. Nevertheless, it is not clear how to define "sufficient" in that method, i.e. how to set the thresholds, so the method cannot quantify the reliability of the components in a statistically principled way.

A related problem in ICA research is component selection and sorting. After computing ICA, further analysis may require going manually through all of them which is time consuming, especially in a group analysis. Methods for selecting interesting components have been proposed based on their statistical properties (Formisano et al., 2002; Hyvärinen et al., 2010), or intersubject consistency (Malinen and Hari, Submitted for publication). However, a more principled and fundamental way of selecting components would be to use a testing procedure to select components which are statistically significant. Any other method of selection would be more naturally applied to the subset of significant components.

* Dept of Computer Science, P.O. Box 68, FIN-00014 University of Helsinki, Finland. Fax: +1 358 9 191 51120.
E-mail address: Aapo.Hyvarinen@helsinki.fi.

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.05.086

Here, we propose a method for testing the inter-subject consistency of components in a group ICA setting. We perform ICA separately on a number of subjects, and find clusters of components which are sufficiently similar across subjects, not unlike (Esposito et al., 2005). The method does not force components in different subjects to be similar, as happens in many group ICA methods. Our main contribution here is to develop a theory in which the threshold for what components can be considered "sufficiently similar" is obtained by defining a null hypothesis, which allows us to apply basic principles of statistical estimation theory. In particular, we control the false positive rates of the detected clusters of components, and the false discovery rates of joining components to the clusters. Thus, we obtain a method that determines which components are reliable (significant) enough in a statistically principled way. The method is here developed for the case where the similarity is defined as the similarity of the columns of the mixing matrix, which is typically pertinent in EEG and MEG analysis. It is also applicable to data from a single subject if the recordings are divided into sessions (or segments), and the consistency between such sessions is analyzed.

Mathematical theory

General setting

Assume we have recordings from r subjects. The data from each subject is stored in a data matrix $\mathbf{X}_k$ where $k = 1,\ldots,r$ is the index of the subject. If we do temporal ICA, as is typical with EEG and MEG, the rows of $\mathbf{X}_k$ are the channels and the columns are time points. If we do spatial ICA, as is more typical with fMRI, the rows of $\mathbf{X}_k$ are the volumes (time points) and the columns are the voxels. The theory we present here is equally valid in the case where we have r sessions of recordings from a single subject, but for simplicity, we present the method using the terminology of the multi-subject case.

We assume that the data for each subject follows an ICA model with its own mixing matrix $\mathbf{A}_k$ and independent components $\mathbf{S}_k$:

$$\mathbf{X}_k = \mathbf{A}_k\,\mathbf{S}_k. \tag{1}$$

We estimate ICA separately on each subject, thus obtaining a decomposition

$$\mathbf{X}_k = \hat{\mathbf{A}}_k\,\hat{\mathbf{S}}_k \tag{2}$$

where it is important to distinguish the estimates $\hat{\mathbf{A}}_k$ and $\hat{\mathbf{S}}_k$ from the actual values $\mathbf{A}_k$ and $\mathbf{S}_k$.

We develop here a testing procedure which uses the columns of $\hat{\mathbf{A}}_k$ as the vectors characterizing each subject. In temporal ICA as typically applied on EEG and MEG, these are the spatial patterns. In spatial ICA as typically applied in fMRI, they are the time courses. (Thus, the method presented here is not directly applicable to inter-subject analysis of spatial patterns in spatial ICA of fMRI.) We denote the obtained estimates of those columns by $\mathbf{a}_{ik}, i = 1,\ldots,n$ with n denoting the number of independent components. The number of independent components is here fixed to be equal to the dimension after principal component analysis (PCA), which is performed as part of the ICA estimation. (However, the mixing matrices $\mathbf{A}_k$ are in the original space, i.e. the PCA preprocessing has been inverted.)

The goal is then to determine if the different subjects have significantly similar $\mathbf{a}_{ik}$. Note that because of the permutation indeterminacy of ICA, we cannot hope that the indices i in different subjects correspond to each other; we have to search for the best matching intersubject pairs in some way.

In the following, an important aspect is the well-known division of the ICA estimation into two parts: we can estimate ICA by first doing a preliminary whitening of the data (often accompanied by a PCA dimension reduction), and then estimating an orthogonal ICA transform. Thus, the whitening reduces the ICA transform into an orthogonal matrix.

Null hypothesis, or model of inter-subject randomness

The purpose of our null hypothesis, or H_0, is to model the situation where the estimates of the mixing matrix $\hat{\mathbf{A}}_k$ have no inter-subject consistency, so those estimates in different subjects have random relations.

Since our null hypothesis is formulated on the estimates of the parameters, it includes two different elements of randomness. First, it could be that the actual mixing matrices $\mathbf{A}_k$ are completely different in different subjects. This models the real underlying inter-subject variability of the brain activity patterns due to anatomical and physiological differences. Second, it could be that the actual activity patterns are similar in different subjects, but the estimates $\hat{\mathbf{A}}_k$ of the mixing matrices $\mathbf{A}_k$ are very bad and thus effectively random, due to problems in the estimation algorithm. The estimation algorithm can fail because the data does not follow the ICA model, the sample size is too small, there is too much noise, or due to algorithmic problems, as discussed in more detail by Himberg et al.(2004). Our test will consider these two sources of randomness with equal emphasis.

It is important to incorporate just the right amount of randomness in the null hypothesis. We do not want to assume, for example, that the estimates of the mixing matrix are just white noise, because this would introduce too much randomness and the null hypothesis would be too easily rejected. In EEG/MEG, the spatial patterns cannot be assumed to be white noise because different channels are correlated due to volume conduction if for no other reason. Thus, we want to introduce the smallest meaningful amount of randomness in the null hypothesis.

To model the randomness due to anatomical and physiological differences, we assume that the actual mixing matrices $\mathbf{A}_k$ are generated randomly. To introduce a controlled amount of randomness in this (hypothetical) generation of the $\mathbf{A}_k$, we reason as follows: since our goal is to specifically consider the intersubject consistency of the independent components as opposed to the covariance structure of the data, we assume that the recordings $\mathbf{X}_k$ have the same covariance structure. Then, the matrices $\mathbf{A}_k$ are necessarily linked by an orthogonal transformation: $\mathbf{A}_k = \mathbf{A}_0\,\mathbf{U}_k$ where $\mathbf{U}_k$ is an orthogonal matrix, and $\mathbf{A}_0$ is some underlying mixing matrix (which could be taken equal to be any of the $\mathbf{A}_k$). To obtain the maximum amount of randomness in this setting, we assume that $\mathbf{U}_k$ is random and follows a uniform distribution in the set of orthogonal matrices.

Next we model the randomness due to the estimation procedure. Again, since we are interested in modeling the randomness in the ICA estimation as opposed to the covariance structure or its estimation, we assume that only the latter part of the ICA estimation procedure (finding an orthogonal transform) produces random results. Thus we assume that under H_0, the estimated spatial patterns are orthogonal transformations of the underlying spatial patterns, i.e. $\hat{\mathbf{A}}_k = \mathbf{A}_k\mathbf{U'}_k$ for some orthogonal matrix $\mathbf{U'}_k$. Again, $\mathbf{U'}_k$ is assumed to be uniformly distributed in the set of orthogonal matrices.

Thus, we see that we can model both kinds of randomness (variability of brains, and variability of ICA estimation) by the same idea of considering the estimated mixing matrix to be a random orthogonal transformation of some underlying mixing matrix. In fact, we have $\hat{\mathbf{A}}_k = \mathbf{A}_0\,\mathbf{U}_k\,\mathbf{U'}_k$. The product of two uniformly distributed orthogonal matrices is again uniformly distributed in the set of orthogonal matrices, as is well-known in the theory of random matrices.

Thus, we can rigorously formulate the distribution of the model parameters under H_0: under H_0, the mixing matrix $\mathbf{A}_k = [\mathbf{a}_{1k}, \mathbf{a}_{2k}, \ldots, \mathbf{a}_{nk}]$ for the k-th subject has the same distribution as $\mathbf{A}_0\,\mathbf{U_k}$ where $\mathbf{U}_k$ is a random matrix uniformly distributed in the set of orthogonal $n \times n$ matrices, and $\mathbf{A}_0$ is a fixed matrix. The $\mathbf{U}_k$ for different subjects are mutually independent. To use this null hypothesis in practice, it is not necessary to specify or estimate the matrix $\mathbf{A}_0$, as will be seen below.

Defining similarity of components

Our test is based on similarities of the vectors $\mathbf{a}_{ik}$ estimated for different subjects. The similarity of two vectors is defined as the Euclidean similarity which uses a weighting given by a stabilized inverse of the "global" covariance matrix. We define the global covariance matrix of the vectors as

$$\mathbf{C} = \frac{1}{nr} \sum_{ik} \mathbf{a}_{ik}\, \mathbf{a}_{ik}^T \tag{3}$$

which is in fact equal to the covariance of the data computed over all subjects, assuming they all have the same number of data points. Then, we define PCA on the set of the vectors $\mathbf{a}_{ik}$ in the usual way: the PCA is given by the reduced matrices $\mathbf{D}_0$ and $\mathbf{E}_0$ which are obtained as the dominant diagonal entries and columns of the matrices in the eigen-value decomposition $\mathbf{C} = \mathbf{EDE}^T$. The dimension of $\mathbf{D}_0$ and $\mathbf{E}_0$ is fixed as the same n as the dimension of the original data after its PCA dimension reduction.

Using the global covariance matrix, we define the similarities of the vectors $\mathbf{a}$ as follows:

$$\gamma_{ij,kl} = \frac{|\mathbf{a}_{ik}^T \mathbf{R} \mathbf{a}_{jl}|}{\sqrt{\mathbf{a}_{ik}^T \mathbf{R} \mathbf{a}_{ik}}\sqrt{\mathbf{a}_{jl}^T \mathbf{R} \mathbf{a}_{jl}}} \tag{4}$$

where

$$\mathbf{R} = \mathbf{E}_0 \mathbf{D}_0^{-1} \mathbf{E}_0^T. \tag{5}$$

The similarity γ is related to the well-known Mahalanobis similarity, but for the sake of numerical stability, we take the inverse of the covariance "inside" the PCA subspace only. We further take the absolute value in Eq. (4) because of the sign indeterminacy of independent components.

While the use of the Mahalanobis distance has some general justifications in machine learning, our main reason for using this special weighting of the distances is the following property (proven in the Appendix):

Theorem 1. *Under* H_0*, each* γ *follows the (marginal) distribution of the absolute value of an element of an orthogonal matrix uniformly distributed in the set* $n \times n$ *of orthogonal matrices.*

The point here is that under H_0, the distribution of the similarities does *not* depend on any model parameters, such as the covariances or the hypothetical matrix $\mathbf{A}_0$. It only depends on the PCA dimension n. This greatly simplifies the computation of p-values, which we consider next.

Finding significant similarities

After computing all the similarities, we want to determine which similarities are statistically significant.

Null distribution of similarities

First, we need to determine in detail the null distribution of the similarities based on Theorem 1. The starting point is the following theorem (proven in the Appendix):

Theorem 2. *Assume* $\mathbf{U}$ *is a random matrix which follows the uniform distribution in the set of orthogonal* $d \times d$ *matrices. Denote by* u *one entry in the matrix. Then the transformed variable*

$$t = \frac{u\sqrt{d-1}}{\sqrt{1-u^2}} \tag{6}$$

follows a Student's t-distribution with $d-1$ *degrees of freedom, and* u^2 *follows a beta distribution with parameters* $\left(\frac{1}{2}, \frac{d-1}{2}\right)$.

Knowing the distribution of such simple transformations of u under the null hypothesis allows us to determine the chance level of the similarities γ, which are distributed as the elements of a random orthogonal matrix according to Theorem 1. In particular, we can transform the similarities to p-values. (See the end of the Appendix for notes on numerical computation of the p-values.) Using the p-values, we could easily define a test with a controlled false-positive rate (FPR). However, we must take into account the fact that we are computing many similarities, and if we just use ordinary testing based on uncorrected fixed false-positive rate according to the distribution given above, we are likely to get many more false positives as is well-known in the theory of multiple testing.

Corrections for multiple testing

We propose to approach the problem of multiple testing by a combination of two different approaches.

False discovery rate for connections. In general, we use the concept of false discovery rate (FDR), proposed by Benjamini and Hochberg (1995), instead of false positive rate because using the false positive rate leads to very conservative (insensitive) results in the case of a large number of tests, as has been previously shown in the context of brain imaging by Genovese et al.(2002). The FDR is defined as the number of false positives divided by the total number of positives.

Denote by n_γ the number of truly significant similarities, which we assume to be much smaller than the total number of similarities. If we use a corrected significance level $\alpha_{\mathrm{FD}}^{\mathrm{corr}}$ in the test, we get approximately $\alpha_{\mathrm{FD}}^{\mathrm{corr}} m$ false positives where m is the total number of tested similarities, and we assume independence of the tests. Thus, to control the proportion of false positives (FDR) to be below a given level α_{FD}, we should have

$$\frac{\alpha_{\mathrm{FD}}^{\mathrm{corr}} m}{n_\gamma} \leq \alpha_{\mathrm{FD}} \tag{7}$$

where we omit adding the number of false positives in the denominator because it is assumed to be small enough. Thus, we should take

$$\alpha_{\mathrm{FD}}^{\mathrm{corr}} = \alpha_{\mathrm{FD}} \frac{n_\gamma}{m}. \tag{8}$$

It turns out that we do not need to have explicit estimate of n_γ to perform this testing with a controlled FDR. We use the well-known Simes' procedure (Benjamini and Hochberg, 1995; Simes, 1986) to find the threshold without computing n_γ. The number of tests m can be obtained simply by counting similarities considered; a simple formula will be given below.

False positive rate for clusters. Nevertheless, we prefer to control the classical false-positive rate for the *existence* of a component which is consistent (a cluster).

We do this because controlling the FPR is usually preferable to controlling the FDR, if it does not make the test too conservative. In particular, inferring the existence of a consistent component which does not actually exist can be considered a rather serious error from the viewpoint of neuroscientific interpretation. In contrast, it may be less serious to infer that a given subject has a component which really exists in the group although actually not for that subject. So, it makes sense to be more conservative in testing the existence of consistent components. In our simulations and experiments, using the FPR for the consistent components (clusters) seemed to be sensitive enough, and not too conservative.

To control the FPR of clusters, we use a simple Bonferroni correction. We compute the approximately corrected $\alpha_{\mathrm{FP}}^{\mathrm{corr}}$ threshold simply as

$$\alpha_{\mathrm{FP}}^{\mathrm{corr}} = \frac{\alpha_{\mathrm{FP}}}{m}. \tag{9}$$

To calculate the number of tests m needed in Formula (9), consider that we are basically taking the maximum over all the elements of the similarity matrix, excluding connections inside the same subject. The matrix is symmetric which reduces the degrees of freedom by one half. So, we obtain the degrees of freedom as

$$m = \frac{n^2 r(r-1)}{2}. \tag{10}$$

Some idea of the difference between the FDR and Bonferroni corrections can be obtained from our MEG experiments below, in which the factor on the right-hand-side of Eq. (8) is of the order of 1/1000…1/100. Using Bonferroni correction as in Eq. (9), the factor would be much smaller, approximately 10^{-5}. In fact, the difference between the two thresholds is exactly the additional factor n_γ, which was typically between 100 and 1000 in our MEG experiments.

Both Bonferroni correction and Simes' procedure make the assumption of independence of the tests. The validity of this assumption is certainly questionable but it can be considered a useful first approximation. The simulations below will shed light on whether it is reasonable.

Intersubject clustering of similar components

We can use the similarities considered strong (significant) enough in different clustering methods. Here, we develop a rather simple one similar to hierarchical clustering using a single-linkage strategy.

A cluster of components which are consistent over subjects is found by starting with the pair of components which is the most similar in terms of having the smallest p-value. Whether the similarity is sufficient is tested based on the corrected FPR threshold given in Eq. (9). Further components are added to this cluster based on the strongest similarity between a candidate component and the components already in the cluster, until no more components with significant similarities according to the FDR criterion, implemented by Simes' procedure, are found. In the clustering, it is obviously always forbidden to put two vectors from the same subject in the same cluster.

Note that we do not require that the cluster should contain a component from all the subjects because this is unrealistic: many interesting brain sources are likely to be found in some subjects but not all of them. In some cases, it may be interesting to search for clusters which include only a couple of vectors. Thus, we allow the cluster size to be completely determined by the data.

Once a cluster has been found, we can find more by a simple "deflation" procedure. We simply re-run the clustering but ignore all the vectors which have already been clustered. This is a rather heuristic procedure, and its effects of FPR and FDR will be investigated next. Also, the algorithm will be described in more detail below.

Corrections needed because of deflation

Above, all p-values were computed under H_0 which says that there are no clusters of consistent components in the data. However, if we consider data with, say, ten clusters, we need to take the effects of deflation, i.e. the interaction between clusters into account. The false-positive rate for the existence of the 11th cluster is, in fact, different from the false-positive rate for finding the first cluster, as will be seen below. That is, the p-values we computed above are strictly correct only for finding one cluster.

To formalize this, we define a parameterized version of the null hypothesis, $\mathrm{H'}_0(\mathrm{k})$. Under $\mathrm{H'}_0(\mathrm{k})$, k components are present and equal in all the subjects. The existence of k ideal clusters in the data simply means that the dimension n of the data is reduced by k. Thus, we can re-apply the method and use the distribution under H_0 again, taking the new dimension into account in the computation of the p-values. Reducing the dimension effectively reduces the randomness in the data which is seen in larger p-values. Thus, it is important to take this change into account to control the FPR and FDR.

While in the ideal case, finding one cluster of components has simply the effect of reducing the dimension of the data, in practice, the total effect of such deflation is more complex. This is so especially because we do not require a cluster to have r components, and the components are not exactly equal in different subjects. Thus, to take this effect into account more precisely, we define the "effective" PCA dimension for each pair of subjects k, k' based on the number of clusters which include components from both subjects:

$$\tilde{n}(k,k') = n - \left\{\#\text{ of clusters } C \text{ such that } \mathbf{a}_{ik}, \mathbf{a}_{jk'} \in \mathbf{C} \text{ for some } i,j\right\}. \tag{11}$$

The effective PCA dimension essentially quantifies the actual randomness in the data. When computing the p-values of the similarities, the corresponding $\tilde{n}$ should be used in the parameters of the beta or Student distributions.

Ideally, the effective PCA dimension should be computed before computing any p-values and doing any clustering. This may be impossible, however, because it depends on the clustering. We proceed here by updating the estimates of $\tilde{n}$ at every deflation step, i.e. after each formation of a new cluster. This seems to be an appropriate approximation because the clustering uses the smallest p-values first. Thus, it should be enough to make the correction which tightens the thresholds only during the formation of the first clusters, when clustering will be attempted with larger p-values. Furthermore, we reiterate the clustering to further fine-tune the internal parameters, as will be described below.

Description of the algorithm

Finally, we describe the resulting algorithm in detail. It proceeds as follows:

1. Parameters fixed by the ICA results are the PCA dimension of the data n and number of subjects r. Parameters fixed by the investigator for the testing procedure are the false positive rate for clusters α_{FP} and the false discovery rate for similarities α_{FD}.
2. Set the initial effective PCA dimension $\tilde{n}(k,l) = n$ for all k, l. Define m as in Eq. (10).
3. Compute the global covariance $\mathbf{C}$ as in Eq. (3) and the similarities $\gamma_{ij,\,kl}$ as in Eq. (4) for all $i, j = 1,\ldots, n$ and $k, l = 1,\ldots, r$, $k \neq l$. Set similarities of vectors in the same subject to zero.
4. Define the set of found clusters S to be empty. Set the variable $u_{ij,\,kl} = 1$ for all i, j and $k \neq l$; this variable tracks which similarities are still valid (not deflated away).
5. Transform the similarities into p-values by

$$p_{ij,kl} = 1 - B_I\left(\gamma_{ij,kl}^2, \frac{1}{2}, \frac{\tilde{n}(k,l)-1}{2}\right) \tag{12}$$

   where B_I is the regularized incomplete beta function (i.e. the cdf of the beta distribution).
6. Find the smallest p-value $p_{ij,\,kl}$ over all i, j, k, l such that $u_{ij,\,kl} = 1$. Denote the minimizing indices as I, J, K, L.

7. This p-value is significant if

$$p_{IJ,KL} < \frac{\alpha_{FP}}{m}. \tag{13}$$

8. If the p-value is significant according to Eq. (13), define the new cluster C initially as the set of those two vectors: $C = [(I, K), (J, L)]$.
- Otherwise, no more clusters can be found: abort the algorithm and output S as the set of significant clusters.
9. Perform Simes' procedure on the p-values. That is, sort the p-values, and consider the h-th smallest p-value $p(h)$ significant if

$$p(h) \leq \frac{\alpha_{FD} h}{m}. \tag{14}$$

10. Search for the smallest p-value $p_{ij,\,kl}$ which was found significant according to Eq. (14) and which is such that either (i, k) or (j, l), but not both, is in C and $u_{ij,\,kl} = 1$ (i.e. the connection is "going out" from the cluster and not deflated away).
11. If such a p-value could be found, denote the minimizing indices as I, J, K, L, and add the vector which is connected to the cluster by $\gamma_{IJ,\,KL}$ to C, and go back to step 10.
- Otherwise, store C in the set of found clusters S. Set $u_{ij,\,kl}$ to zero for all similarities to and from vectors in C (deflation). Update the effective dimensions $\tilde{n}$ as in Eq. (11). Go back to step 5.

To further fine-tune the clustering, we propose to run the clustering algorithm a second time, using the internal parameters $\tilde{n}$ obtained at the first run of the algorithm. This has the benefit that the computation of all the clusters is using the same estimates of the internal parameters.

Public-domain Matlab code implementing the algorithm is available at www.cs.helsinki.fi/u/ahyvarin/code/isctest/.

Computational complexity

To analyze the computational complexity of the resulting algorithm, we begin by noting that the computations done in the clustering method are relatively simple searches for the largest elements. After the initial computation of the similarities, no sophisticated matrix operations are done. The sorting of the p-values is the only operation which does not have linear complexity in the number of similarities. On the other hand, the number of the similarities is quite large, proportional to n^2r^2. The similarities are manipulated and searched through for every cluster, and thus we need to multiply this by the number of clusters. The number of clusters could be assumed to be proportional to n.

As a first approximation, we might thus assume that time needed for computation is proportional to n^3r^2. This may not be quite the case in theory because typical sorting algorithms would require $O(n^2r^2(\log n + \log r)$ operations, but the difference may be insignificant in practice.

In fact, we have found that the main bottleneck in the method, using a simple PC, is in the memory needed to store the similarities and quantities which are derived from the similarities, such as p-values, indices of which p-values are deflated away, and related temporary quantities. This memory complexity is clearly of the order n^2r^2.

We will consider these issues in more detail in the simulations below.

Experimental methods

Simulation 1: artificial data

As a first validation of the testing procedure we conducted simulations with purely artificial data. The main goal was to compute the false positive and false discovery rates under the null hypotheses and see if it is well controlled in spite of the many approximations made in the development of the testing procedure. We operated in the space of orthogonal rotations, thus neglecting the ICA estimation part.

The data PCA dimension had the values 20 and 50, while the number of subjects was either 6 or 20. We generated random orthogonal matrices in the (hypothetical) PCA space, where each column of the orthogonal matrix corresponds to one component, and computed the similarities. Then, we ran the testing algorithm.

We used the following five scenarios which all could potentially give rise to different kinds of errors:

1. There was no inter-subject consistency at all: all components in all subjects were generated independently of each other.
2. Half of the components were equal in all subjects, and half of the components were completely random. In other words, half of the components had perfect inter-subject consistency, while the other half had zero consistency.
3. Half of the subjects had all equal components, while the other half had components which were completely random, independent of each other and of the first half of subjects. In other words, half of the subjects had perfect inter-subject consistency for all components, while the other half of the subjects had no consistent components.
4. Half of the components were equal in all subjects. Moreover, for half of the subjects, all the components were consistent.
5. For half of the subjects, half of the components were consistent (equal over subjects).

In scenarios 1 and 2, the typical errors would be that the algorithm finds a false positive cluster, usually with just two components. In scenario 3, the typical error would be adding one falsely "discovered" component to one of the clusters. In scenarios 4 and 5, both kinds of errors are equally possible.

The false positive and discovery rates in the testing method were set at $\alpha_{FP} = \alpha_{FD} = 0.05$ and 500 different sets of orthogonal matrices ("data sets") were generated in each of the 2×2×5 = 20 different conditions.

We computed what we call the "actual" FPRs and FDRs as the proportion of data sets in which one of the following errors occurred when compared to the true generating mechanism: either there was a false positive cluster, or a component was added to a cluster although it did not belong there. Note that these error rates are the rates which are relevant in practice; they do not exactly correspond to the error rates $\alpha_{FP} = \alpha_{FD}$ in the theoretical development. For example, even if a similarity falsely exceeds the FDR threshold, it may not lead to a false clustering: it is possible that neither of the corresponding components could be added to an existing cluster because the more stringent FPR threshold was not exceeded for sufficiently similar components.

Simulation 2: semi-realistic data with varying inter-subject consistency

As a second validation of the testing procedure, we used artificial data where the ground truth is known, but went through all the steps of practical data analysis, including ICA estimation. The number of subjects was fixed to 11, the data dimension to 204, the dimension after PCA dimension to 40, and $\alpha_{FP} = \alpha_{FD} = 0.05$.

We first chose a "common" mixing matrix $\mathbf{A}_0$ as a basis for inter-subject consistency. While we could have generated $\mathbf{A}_0$ completely randomly, we chose to introduce some more realism to the simulation by taking as $\mathbf{A}_0$ a mixing matrix estimated from MEG data (see below). Then, the mixing matrices for different subjects were created by adding inter-subject variability to this common matrix.

Intersubject variability was created by adding gaussian "noise" to the common mixing matrix $\mathbf{A}_0$, using different noise samples for each subject. Note that this noise has little to do with measurement noise in a brain imaging device, since it is added on the parameters and not on the signals. The level of noise added to the mixing matrix, which we call intersubject noise, was varied. Furthermore, we completely destroyed inter-subject consistency for one half of the components by replacing half of the columns (same columns for each subject) by random gaussian noise. The variance of the noise was chosen so that the norms of the columns of the mixing matrix were not changed on average.

The independent components were generated as Laplacian i.i.d. signals with 10,000 time points. The standard deviation of each independent component (or equivalently, the norm of the corresponding column of the mixing matrix) was set to a random number uniformly distributed between 0.5 and 1.5. Finally, the independent components were mixed for each subject.

For each level of inter-subject variability, 100 randomized trials were conducted, in which the common mixing matrix $\mathbf{A}_0$ was randomly picked from a set of 11 different estimated mixing matrices (corresponding to different subjects in the MEG experiments below).

Note that an intersubject noise level of 1 essentially means a signal-to-noise ratio of 1 in creation of the individual mixing matrices from the common one. Thus, when the intersubject noise level is larger than one, the "intrasubject" part of the common part of the mixing matrix is larger than the "intersubject" one.

We then estimated ICA for each subject separately using the FastICA algorithm, and tested inter-subject consistency as explained above. To compare the results with the ground truth, we assigned each estimated vector (column of estimated mixing matrix) to one of the columns of the original common mixing matrix $\mathbf{A}_0$, by finding the maximum correlation coefficient (in absolute value) between the estimated vector and the columns of $\mathbf{A}_0$.

We computed two quantities as a function of intersubject noise:

- the number of times the null hypothesis was rejected, and
- the number of clusters found by the method.

Further, we assessed the quality of the clusters found by dividing them into different categories:

- "Perfect" cluster: the cluster has one vector from each subject, and each vector is assigned to the same column of $\mathbf{A}_0$.
- "Correct" cluster: each of the vectors in the cluster was assigned to the same column of $\mathbf{A}_0$, but it does not contain a vector from all the subjects.
- "Incorrect" cluster: it contains vectors which were assigned to different columns of $\mathbf{A}_0$.

Ideally, the number of clusters found would be equal to 20, which is the number of consistent clusters (one half of the PCA dimension, 40). Further, the clusters would all be perfect, and the null hypothesis would be rejected in 100% of the cases.

Simulation 3: semi-realistic data with two subject groups

We further conducted a variant of the preceding simulation to investigate the behavior of the algorithm when there are two different groups of subjects. Instead of adding general inter-subject "noise" to the mixing matrix as in Simulation 2, we added a random perturbation to the mixing coefficients of the consistent components for subjects with indices 6,…,11 so that the perturbation was the same for all subjects (but different for different components). The random perturbation models the difference between the groups consisting of subjects 1,…,5 and 6,…,11. The numbering of the subjects is arbitrary, so this models the general case where the subjects can be divided into two groups and we do not know the grouping.

The norm of the random perturbation was increased in the same way and with the same values as in Simulation 2. We analyzed the clustering in the same way as in Simulation 2. In this case, the meaning of "incorrect" and "perfect" clusters is not quite well-defined, but they serve as useful quantitative measures of the behavior of the algorithm. We further analyzed the data by simply plotting individual clustering results for comparing them with the group structure.

Simulation 4: computational complexity

Next we evaluated the computational complexity of the method to determine which numbers of subjects r and PCA dimensions n are feasible.

To evaluate the computational complexity of the method, we can of course use quite artificial data. However, we cannot use random noise because then the clustering method would find no clusters and there would be not much to compute. So, we decided to generate data from scenario 5 of Simulation 1, which is arguably the most realistic.

We used the same values for the number subjects and PCA dimensions, $n=r$. The values used in the different trials were 8, 16, 32, 64, and 128. The computations were done in Matlab on a rather ordinary Linux PC with two cores of 2.66 GHz each, and 2.4 GB of memory available.

We computed the CPU time needed as well as the memory needed. The memory usage considered only the memory needed for storing the explicit variables, i.e. the final values of any Matlab operations neglecting any intermediate results, and thus clearly provides a lower bound only.

Experiments on MEG data

Inter-subject consistency

We next applied the method on magnetoencephalographic (MEG) data consisting of 204 gradiometer channels measured by the Vectorview helmet-shaped neuromagnetometer at the Brain Research Unit of the Low Temperature Laboratory of Aalto University, Finland.[1] The recordings were of spontaneous activity in 11 healthy volunteers, who received alternating auditory, visual, or tactile stimulations interspersed with rest blocks, taken from Ramkumar et al.(in press). The MEG recordings had a prior approval by the Ethics Committee of the Helsinki and Uusimaa Hospital District.

Noise and artifacts were reduced by the signal space separation method (Taulu et al., 2004) which also reduced the effective dimension of the data to 64. At the same time, the data was downsampled from the initial sampling frequency of 600 Hz to 150 Hz. All 64 dimensions were used in ICA, and PCA dimension reduction was not performed. Thus, 64 independent components were estimated for each subject using Fourier-ICA (Hyvärinen et al., 2010). Each topographic distribution of an independent components on the gradiometer channels is a 204-dimensional vector $\mathbf{a}_{ik}$, and these are used in the intersubject consistency testing. We set $\alpha_{\text{FP}}=\alpha_{\text{FD}}=0.01$.

To further analyze the results, we found in each cluster the most representative component (the one with minimum sum of Euclidean distances to other components) and computed its cortical distribution using the minimum norm estimate, as well as the Fourier spectrum.

Finally, we analyzed the modulation by each stimulation modality (some divided into subcategories) by computing the difference of the logarithm of average energy in each stimulation block and the preceding rest block, separately for each component in the cluster. The differences were converted to (uncorrected) z-scores.

Inter-session consistency

We further applied the method to three different recordings of a single subject obtained in the same set of experiments as the one above. In addition to the dataset consisting of naturalistic stimulations interspersed with rest, we also had two more recording sessions. In

[1] I'm very grateful to Pavan Ramkumar and Riitta Hari for giving access to these measurements.

one of them the subject was resting with eyes open and fixated, while in the other the subject received the same kind of naturalistic stimulation as above but without any rest in between. The analysis of the results was identical to the analysis in the inter-subject consistency experiments, subjects being simply replaced by sessions. However, we did not compute the activity modulation by stimulation because there were not enough different blocks to achieve statistical significance in that respect.

Results

Simulation 1: artificial data

The “actual” false-positive rates and false-discovery rates (as defined in Methods) are shown in Fig. 1. We can see that they are all less than the required 5%, and well controlled in spite of the various approximations done in developing our method.

Thus, the approximations in the computation of the p-values seem to lead to conservative testing, so the FPR and FDR do not need to chosen particularly small.

Perhaps one could argue that the most realistic case is scenario 5 in which the inter-subject consistency is always limited in the sense that component is found in all the subjects, and in addition there are subjects with no consistency with the others. In this scenario, the FPR and FDR were of the order of 1%, which shows some tendency to conservative testing since we set $\alpha_{FD} = \alpha_{FP} = 0.05$.

Simulation 2: semi-realistic data with varying inter-subject consistency

The probability of rejection of the null hypothesis is shown in Fig. 2 a). We see that for a reasonable intersubject noise, i.e., some inter-subject consistency, the method always correctly rejected the null hypothesis. However, with really small inter-subject consistency (noise level of 2), the null hypothesis was no longer always rejected, and intersubject consistency was no longer detected.

The numbers of clusters found, divided into the different categories, are shown in Figs. 2 b)–c). We see that for reasonably small intersubject noise (0.25 or 0.5), i.e. reasonably large consistency, almost all clusters are perfect and there are 20 (or close) of them as expected.

For a larger noise level (=1), clusters are fewer and they are not perfect; one the average, 15% of them were correct only. With a very high noise level (=2), there were only two clusters on the average. No incorrect clusters were observed in these simulations, but this is presumably subject to a lot of random fluctuation and not a conclusive result.

Thus, incorrect clustering seems to happen very rarely: when intersubject consistency is negligible, the method does not usually group the components into incorrect clusters. Instead, it simply finds very few clusters. This is of course what the method was supposed to do, and indicates that clusters considered significant can be trusted to some extent.

Simulation 3: semi-realistic data with two subject groups

The results for Simulation 3 are shown in Fig. 3. We see that first, when the difference between the groups is small (0.25 or 0.5), the method simply ignores the difference between the groups: it clusters corresponding components from all subjects into one cluster, which is thus “perfect” in our classification. In contrast, with the highest group difference (2), the method creates almost 40 “correct” clusters and few perfect ones, as well as a few incorrect ones.

A closer examination of the clusters reveals the expected result that each subject group has its own clusters in the case of high group difference, with components from 5 or 6 subjects in each. This is shown in Fig. 3 d), in which the clustering structure is given for one randomly selected trial in the case of the highest group difference. The method found 38 clusters, 36 of which contain exactly the subjects of one group (up to two random errors), and 2 contain the subjects from both groups. Since the group difference was randomized for each component, for some components the group difference seems to have been too small to be detected.

Thus, any group differences of the subjects can lead to splitting of the clusters by groups, provided the group differences are strong enough.

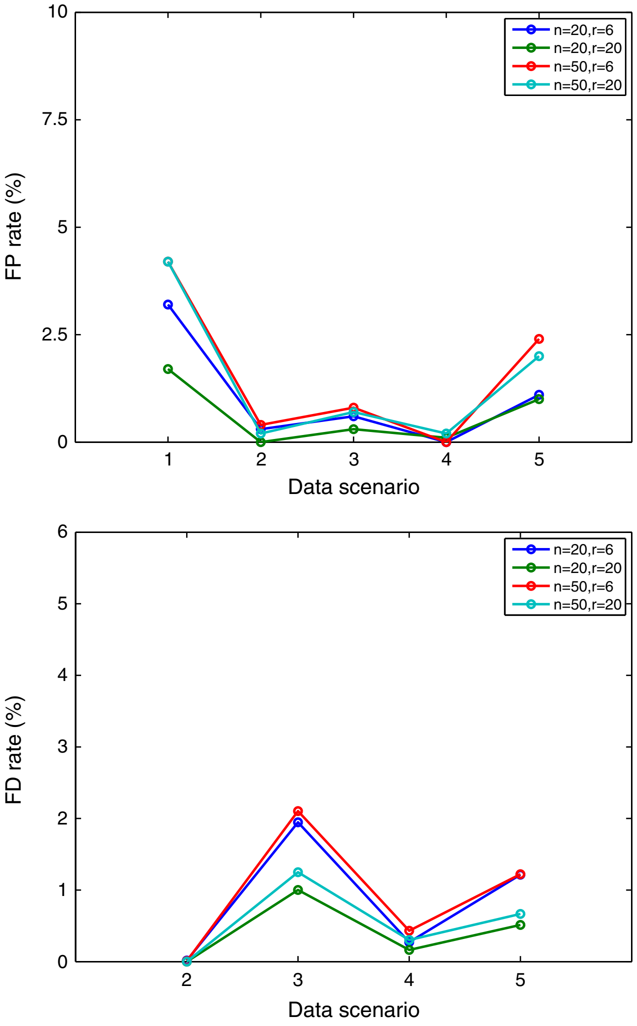

Fig. 1. Simulation 1: false positive rates and false discovery rates for simulated data. Different settings of data dimension and number of subjects are given in different colors. The data scenarios are explained in detail in the text, briefly: 1: no consistent components, 2: half of components consistent for all subjects, 3: all components consistent for half of the subjects, 4: for half the subjects, all components consistent and half of the components consistent for the rest of the subjects, 5: for half of the subjects, half of the components were consistent. The desired false positive and discovery rates were set to$\alpha_{FP} = \alpha_{FD} = 0.05$. For scenario 1, FDR cannot be meaningfully computed since the number of true positives is zero.

Simulation 4: computational complexity

The computational requirements are shown in Fig. 4. Note that both plots are in log–log scale with base 2. We can see that the memory requirement increases rather exactly proportionally to n^2r^2, which is seen as the increase of memory by a constant of $16 = 2^4$ when increasing n and r by a factor of two. Regarding computation time, the progression is also close to linear. The slope of the line is close to 5 (in logarithms) for the largest dimensions computed, thus

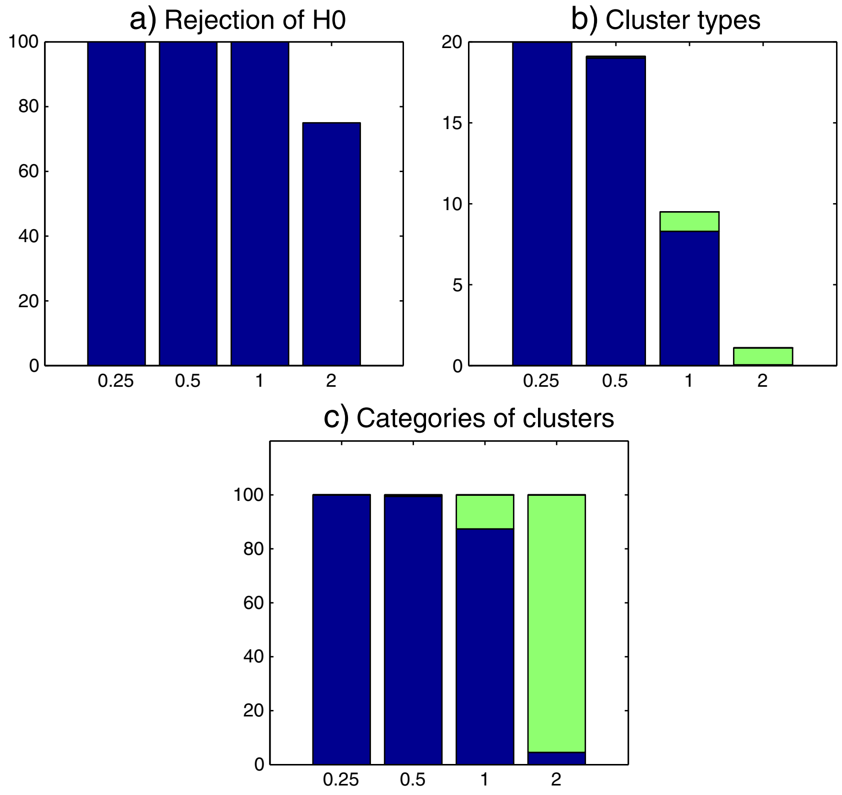

Fig. 2. Simulation 2: semi-realistic data with half of the components consistent to varying degrees. Horizontal axis is noise level. a) The probability (percentages) of rejection of null hypothesis as a function of the intersubject noise level. b) The average number of clusters found per trial, including their division into categories. Blue: "perfect" cluster, Green: "correct" cluster, Red: "incorrect cluster". c) Proportion of different cluster types among all clusters found, in percentages.

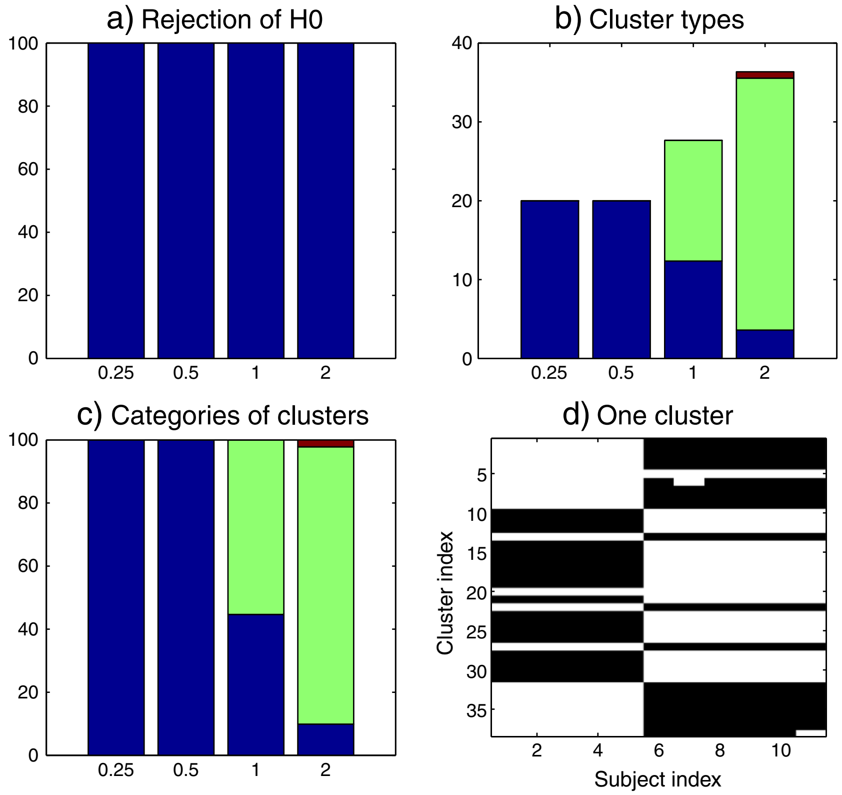

Fig. 3. Simulation 3: Semi-realistic data with half of the components consistent to varying degrees, and subjects divided into two groups. Horizontal axis is magnitude of group difference in a)–c). a) The probability (percentages) of rejection of null hypothesis as a function of the group difference. b) The average number of clusters found per trial, including their division into categories. Blue: "perfect" cluster, Green: "correct" cluster, Red: "incorrect cluster". c) Proportion of different cluster types among all clusters found, in percentages. d) One randomly selected clustering for highest group difference. White means the subject has a component belonging to the cluster, black means there is none.

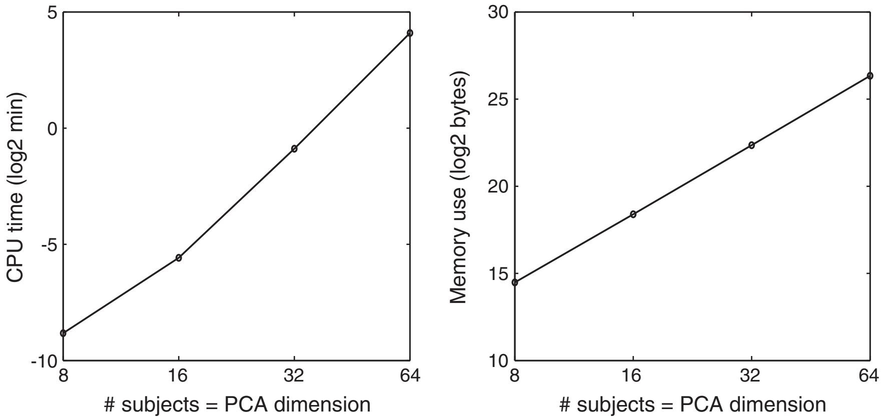

Fig. 4. Simulation 4: computational complexity. Left: CPU time in minutes (log_2 scale), Right: memory required in bytes (log_2 scale). The case $n = r = 128$ was infeasible because it would have required more memory than was available.

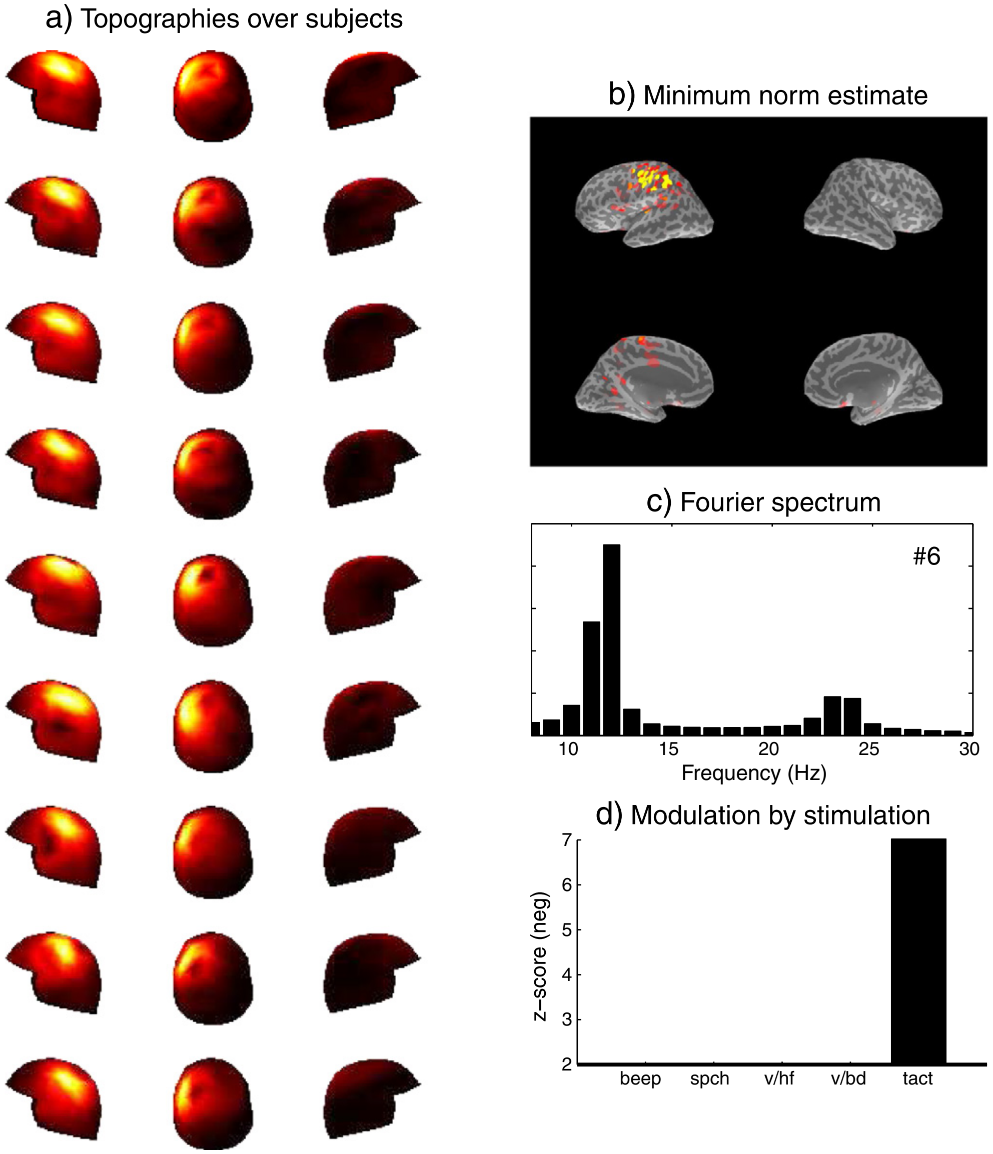

Fig. 5. One cluster of sources found in real group MEG data. The cluster is modulated by tactile stimulation. a) Topographic plots for all components in the cluster, in the same order in which they were added to the cluster. b) Cortical projection using minimum norm estimate, c) Fourier spectrum, d) modulation by stimulation modalities. Cortical projection and Fourier spectrum shown for one representative component, whose row number in a is given inside the plot in c. Modulation was computed for the following stimulation groups: beep: auditory beeps, tsfspch: speech, v/hf: visual stimulation showing hands and faces, v/bd: visual stimulation showing buildings, tact: tactile stimulation.

approximately conforming with the theoretical prediction of n^3r^2 complexity.

The case $n = r = 128$ could not be computed due to lack of memory in our PC. In fact, an extrapolation of the line plotted shows that it would have required approximately 1.4 GB of memory to store the variables. Since this value does not take into account the memory needed for temporary storage of internal variables, the required computations were impossible with the 2.4 GB of free memory we had. However, it would not have been difficult for us to find a computer with the required memory capacity, so the case $n = r = 128$ is not impossible, and presumably already possible in more advanced hardware configurations.

Extrapolating the CPU time, we see that the expected computation time in the case of $n = r = 128$ would have been less than 10 h, which would still have been feasible. Thus, the computational bottleneck is really in the memory requirements.

MEG data

Inter-subject consistency

When applied on the naturalistic stimulation data from 11 subjects and a PCA dimension of 64, the method found 43 reliable clusters. The distribution of cluster sizes (not shown) was rather uniform from 2 to 11, with a slight overrepresentation of clusters of two components. The clusters included a total of 239 components, which is 34% of the total number of estimated components.

We show three manually selected clusters in Figs. 5–7. Fig. 5 shows a typical Rolandic cluster strongly modulated by tactile stimulation. Fig. 6 shows a temporal component mainly modulated by speech stimulation. Fig. 7 shows an occipital visual component. Many further components modulated by sensory input were found as well, typically in the occipital and parietal cortices. Some of the clusters seemed to be ocular or muscular artifacts.

Inter-session consistency

When applied on single-subject data with three different sessions, the method found 32 clusters. 25 of them were of size three, and the rest of size two. Two of the clusters are shown in Figs. 8 and 9. Clearly, the spatial patterns are very similar across sessions.

Discussion

We proposed a method for testing the statistical significance or reliability of independent components based on the consistency of the

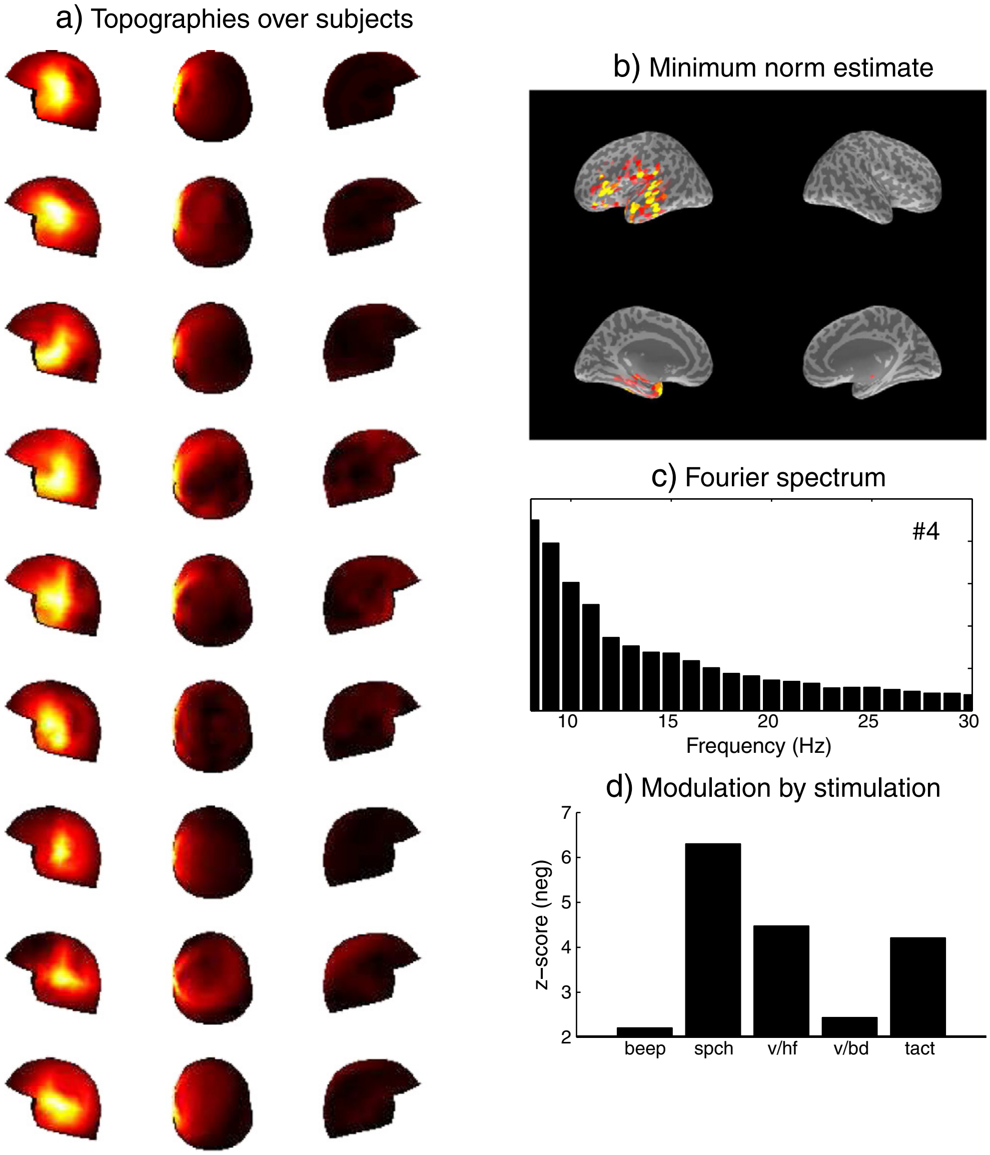

Fig. 6. Another cluster found in real group MEG data, modulated by auditory stimulation. See Fig. 5 for legend.

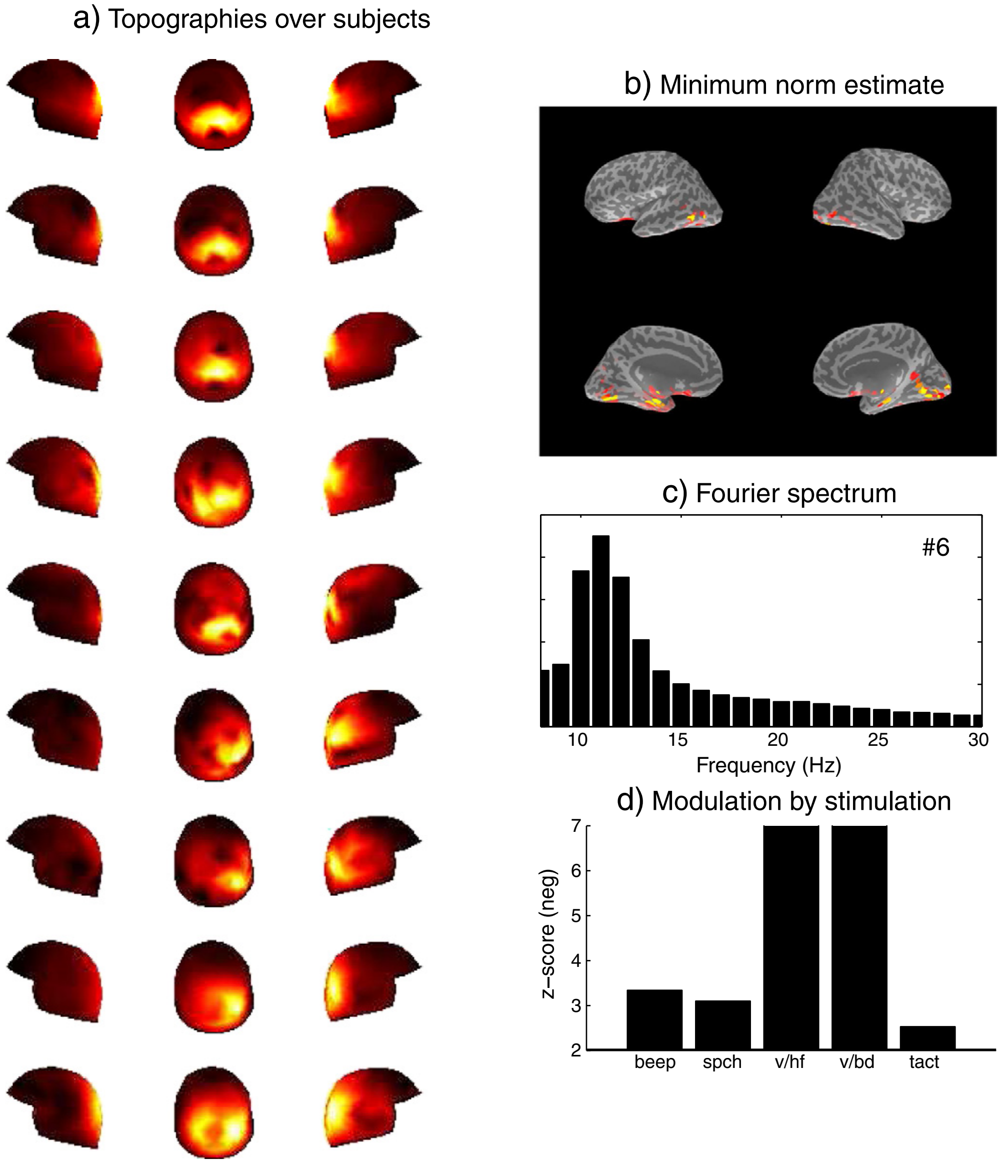

Fig. 7. Another cluster found in real group MEG data, modulated by visual input. See Fig. 5 for legend.

columns of the mixing matrix over subjects or sessions. While clustering of components to solve the group ICA estimation problem has been proposed before (Esposito et al., 2005), our contribution here was to derive statistically principled thresholds to determine if a cluster is reliable or not. We were able to derive such thresholds in closed form, controlling the false-positive rates for clusters and false discovery rates for including components in the clusters. Due to the complexity of the ICA model, the algorithm had to resort to a number of approximations which means that the control of the error rates is not exact. However, according to the simulations, the control of error rates was good, and experiments on real MEG data gave plausible results as well.

Intersubject consistency in ICA theory

The multi-subject scenario has received little attention in the general literature of ICA theory, and it is often considered more of a nuisance in the theoretical literature, although its importance is clear in the context of neuroimaging. Most of the methods for group ICA have been developed, in fact, in the neuroimaging literature.

One implication of the work presented here is that having many subjects is actually very useful for ICA even on a theoretical level, since it leads to a method of testing components which is both intuitively appealing and mathematically principled. Our statistical test discards many of the components and shows which ones are worthy of further attention because they are more consistently found than would be by chance. This is in contrast to most group ICA methods, reviewed by Calhoun et al.(2009), which do not provide any selection of the components. Our method is closely related to the one by Esposito et al. (2005); we improve on that work by replacing an arbitrarily set threshold by a statistically principled one which controls the error rates. A related method based on split-half analysis of the group was recently proposed by Varoquaux et al.(2010) but they did not provide a principled threshold either.

We want to emphasize that widely-used group ICA methods based on concatenation of the individual data matrices are based on the implicit assumption that the same components are present in all subjects. However, this assumption is usually not validated in any way, so there is no guarantee that a given component is really present and meaningful in all the subjects, or even many of them. It is possible

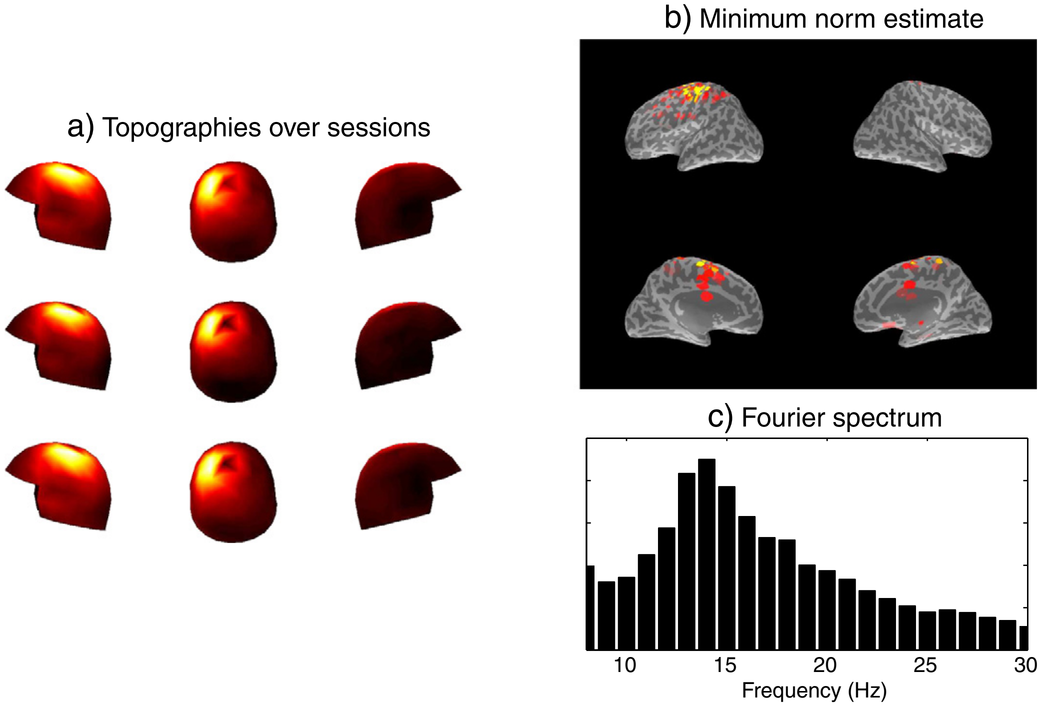

Fig. 8. One cluster of sources in single-subject data with three different sessions. See Fig. 5 for legend. (Modulation by stimulation was not computed here.).

that the ICA algorithm simply ignores some, or even most, of the subjects when estimating a given component. One way to make sure that a component is present in several subjects (and to find out in which subjects) is to compute ICA *separately* for each subject and then analyze the intersubject consistency of the results.

Technical notes and future work

An important extension of the current method would be to consider the case where we are interested in similarities of spatial patterns estimated by spatial ICA. This is, in fact, the most frequent

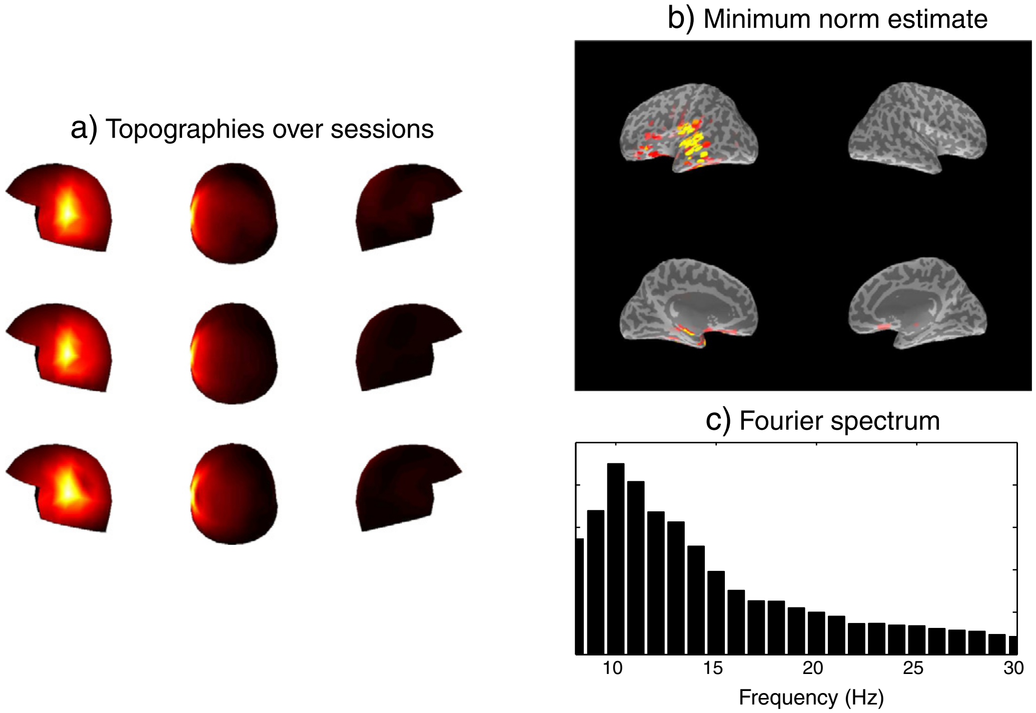

Fig. 9. Another cluster found in single-subject data with three different sessions. See Fig. 5 for legend. (Modulation by stimulation not computed here.).

type of application of group ICA methods, and dominant in the fMRI literature. We are hopeful that the framework can be extended in that direction, and will pursue that goal in the near future. The method as presented here is, in fact, applicable to spatial ICA of fMRI data if one is interested in the inter-subject consistency of time courses, i.e. inter-subject synchronization (Hasson et al., 2004; Malinen and Hari, Submitted for publication).

Our null hypothesis implies some specific equalities of the parameters. In particular, it implies that the subjects have the same covariance matrices, and thus the same PCA subspaces and whitening matrices. While this may be contradictory with their empirical estimates, it is justified by the logic given above for introducing a minimum amount of randomness. If the data for different subjects is actually generated by a random process which has less restrictions, the similarities are less likely to attain any thresholds we compute here. Thus, any thresholds for comparing inter-subject consistency using our H_0 are conservative.

This leads to the question of whether the assumption of equal covariances may actually lead to overly conservative testing. One has to note that the subspace of components which are consistent has, by definition, approximately the same covariance matrix for all the subjects (up to possible differences in scaling of the components). So, the subspace in which the covariance is clearly different is likely to correspond to inconsistent components. Thus, the test is likely to be more conservative for accepting false positives only from subspaces in which there is no consistency. This should not be a problem if it is not likely to be much more conservative in rejecting H_0 for consistent components.

We further assumed in the case of multiple testing that we can approximate the tests to be independent. Alternatively, we might use an FDR procedure which does not make any such assumption (Benjamini and Hochberg, 1995), but we have found (results not shown) that such variants make the test far too conservative. In simulations reported above, we found that our test is not too permissive in spite of this approximation. In fact, it would rather seem to be slightly too conservative, and another topic for future research is to find methods that make the FPR rate closer to the desired one.

Our framework can also be used for analyzing different recordings of the same subject in different conditions, for example, in rest or under different kinds of stimulation. Ultimately, one could even divide a single, long recording into segments and analyze which components are found in many segments. Thus, we can actually determine principled p-values in the general context of the ICASSO method (Himberg et al., 2004) which is applicable to any ICA analysis. It should be noted that our theory cannot be used with bootstrapping samples because such samples have considerable overlap so the complete inter-session randomness of H_0 would be quite unrealistic. Instead, we have to use disjoint subsets of the data points. Because of time correlations in the data, the subsets should also be temporally contiguous, as opposed to random subsets of time points, to make the complete randomness in the null hypothesis a realistic baseline. (Related work based on splitting the data into two halves can be found in Groppe et al., 2009.)

Our clustering method is a simple modification of a classic one: hierarchical clustering by single linkage. The modification consists of allowing at most one cluster member from each subject. There is no reason why any other clustering method could not be used. An obvious option would be to use other variants of hierarchical clustering, in particular complete linkage. Variants of k-means clustering should be easily applicable as well. Whether any benefit can be derived from such variants is another interesting question for future research.

The definition of FDR was here based on the number of similarities considered true although they are false. It should be noted that this is not the number of components falsely clustered. In principle, it is possible that the number of falsely clustered components is larger than given by FDR because if one component is falsely clustered, it may bring other, similar components to the same cluster. However, in our simulations, the opposite seemed to happen, and the number of falsely clustered components was actually smaller than the FDR. In fact, it is not quite clear how to define the number of "falsely clustered" components in the first place, since if the method merges two small clusters, it is not clear whether one should consider falsely clustered all the components in the two clusters, all the components in one of the clusters, or only one component. Our FDR definition considers, in fact, that only one error has been committed.

In general, any estimation method should be accompanied by a testing method in practical data analysis. In the case of ICA, the testing has been long neglected. We hope that the present work and any related future work contribute to correcting that oversight.

Acknowledgments

I'm grateful to Pavan Ramkumar for comments on the manuscript and further experiments. This work was funded by the Centre-of-Excellence on Algorithmic Data Analysis and the Computational Sciences program of the Academy of Finland (Project # 134846).

Appendix A. Proof of Theorem 1

We have

$$\mathbf{C} = \frac{1}{r} \sum_k \mathbf{A}_k \mathbf{A}_k^T = \frac{1}{r} \sum_k \mathbf{A}_0 \mathbf{U}_k \mathbf{U}_k^T \mathbf{A}_0^T = \mathbf{A}_0 \mathbf{A}_0^T \tag{A.1}$$

and thus, for $k \neq l$

$$\mathbf{A}_k^T \mathbf{R} \mathbf{A}_l = \mathbf{U}_k^T \mathbf{A}_0^T \left(\mathbf{A}_0 \mathbf{A}_0^T \right)^+ \mathbf{A}_0 \mathbf{U}_l = \mathbf{U}_k^T \mathbf{U}_l = \mathbf{U} \tag{A.2}$$

where $+$ is the Moore–Penrose pseudoinverse, and $\mathbf{U}$ is uniformly distributed in the set of orthogonal matrices for $k \neq l$. The denominators $\mathbf{a}_{jl}^T \mathbf{R} \mathbf{a}_{jl}$ are all one because they correspond to $\mathbf{U}_k^T \mathbf{U}_l$ with $k = l$, which is identity.

Appendix B. Proof of Theorem 2

The theorem is considered well-known by some authors, and is closely related to results by Fisher(1915) and Anderson(1984). The variable u is, in fact, closely related to the correlation coefficient between two samples drawn from two distributions with zero correlation, since our similarity is a normalized dot-product just like a correlation coefficient. Its distribution has been treated extensively in that context, but since we are unable to find an accessible reference considering this particular case, we provide a complete proof here.

Here, we call a d-sphere a set of the form $\left\{ \mathbf{x} \in \mathbf{R}^d \,|\, \|\mathbf{x}\| = r \right\}$. The volume of the d-sphere is $C(d)r^{d-1}$ where $C(d)$ is a constant that depends on d. We don't need to calculate C because its effect would be essentially to normalize the pdf and we can do that afterwards.

By symmetry considerations, we can see that the distribution of an element of a uniformly distributed $d \times d$ orthogonal matrix is the same as the distribution of an element of an d-dimensional vector $\mathbf{u}$ uniformly distributed on the unit d-sphere (i.e. d-sphere with $r = 1$). From now on, u thus refers to one element of such a random vector.

Consider a fixed value $u_0 > 0$ for $|u|$. Parameterize it as $u_0 = cos\ \alpha$. The probability $P(u \geq u_0)$ is proportional to the volume of the "cap" of the d-sphere which is obtained for angles $|\alpha| \leq arccos\ u_0$. This is illustrated in Fig. 10. The volume of the set in question can be obtained by integrating over the volumes of the segments corresponding to the part of the sphere between the angles $[\alpha, \alpha + \delta]$ (red and magenta

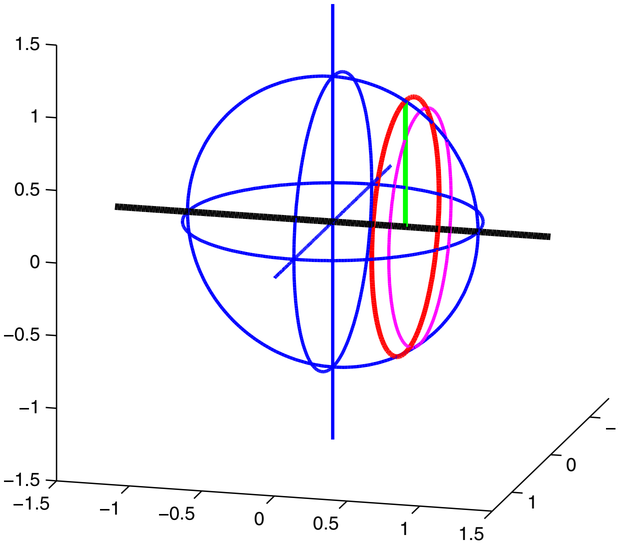

Fig. 10. Illustration of the pdf calculation in a three-dimensional space. The entry u takes values on the black axis. Consider the value at the point $u_0 = cos\alpha$ where the axis meets the green line. The probability that $|u|$ is larger than this is proportional to the volume of the part of the d-sphere which is to the right of the red circle. The volume can be computed by integrating over the volumes of the $d-1$-spheres like the red and magenta circles. The radius of the red circle is $sin\ \alpha$.

circles in Fig. 10). The volume of such a segment, which is essentially an $d-1$-sphere, equals $C(d-1)(\sin\alpha)^{d-2}\delta$. Thus, we have

$$P(|u| \geq u_0) = C'(d)\int_0^{\arccos\, u_0}(\sin\alpha)^{d-2}d\alpha \tag{B.1}$$

for some constant C' which depends on d only.

We make the transformation of variable

$$t = \cos^2\alpha. \tag{B.2}$$

Since

$$\frac{d}{dx}\arccos x = \frac{-1}{\sqrt{1-x^2}} \tag{B.3}$$

we obtain

$$d\alpha / dt = -\frac{1}{2}(1-t)^{-1/2}t^{-1/2} \tag{B.4}$$

and thus

$$\begin{aligned} P(|u| \geq u_0) &= P\left(u^2 \geq u_0^2\right) = C''(d)\int_0^{u_0^2}\left(\sqrt{1-t}\right)^{d-2}(1-t)^{-1/2}t^{-1/2}dt \\ &= C''(d)\int_0^{u_0^2}t^{\frac{1}{2}-1}(1-t)^{\frac{d-1}{2}-1}dt. \end{aligned} \tag{B.5}$$

Here, we recognize in the integrand the unnormalized pdf of the beta distribution with parameters $\left(\frac{1}{2},\frac{d-1}{2}\right)$. The constant C'' thus has to be the proper normalizing constant. Thus, we have proven that u^2 follows the beta distribution. Next, we make the transform to Student's distribution. From the cdf in (B.5) we obtain the pdf of u as

$$p(u) \propto u^{2\times\left(\frac{1}{2}-1\right)}\left(1-u^2\right)^{\frac{d-1}{2}-1}u = \left(1-u^2\right)^{\frac{d-1}{2}-1} \tag{B.6}$$

where the multiplying u at the end of the pdf comes from the change of measure when going from u^2 to u. We use the notation $\propto$ to indicate that the expression of the pdf is missing the normalization with respect to td. The inverse and the volume element of the transformation in Eq. (6) are given by

$$u = \frac{t}{\sqrt{d-1}\sqrt{1+\frac{t^2}{d-1}}} = \frac{t}{\sqrt{d-1+t^2}} \tag{B.7}$$

$$\frac{du}{dt} = \frac{1}{(d-1+t^2)^{3/2}} \tag{B.8}$$

and thus we have

$$\begin{aligned} p(t) &\propto \left(1-\frac{t^2}{d-1+t^2}\right)^{\frac{d-1}{2}-1}\frac{1}{(d-1+t^2)^{3/2}} \\ &= \left(\frac{d-1}{d-1+t^2}\right)^{\frac{d}{2}-\frac{3}{2}}\frac{1}{(d-1+t^2)^{3/2}}. \end{aligned} \tag{B.9}$$

From which we finally obtain

$$p(t) \propto \left(1+\frac{t^2}{d-1}\right)^{-\frac{(d-1)+1}{2}} \tag{B.10}$$

which shows that t follows Student's t distribution with $d-1$ degrees of freedom.

Let us note that if u is complex-valued, as proposed, for example, in Hyvärinen et al.(2010), numerical simulations indicate that $|u|^2$ follows a beta distribution with parameters $\left(1,\frac{2d-1}{2}\right)$, but we do not have an analytical proof for this property.

Appendix C. Numerical computation of p-values

The numerical computation of either of the cumulative distribution functions (cdfs) given in the theorem is fundamentally based on the incomplete beta function, which is essentially the cdf of the beta distribution. We have

$$P\left(u^2 \leq x\right) = B_I\left(x,\frac{1}{2},\frac{d-1}{2}\right) \tag{C.1}$$

where B_I is the regularized incomplete beta function

$$B_I(y,a,b) = \frac{\int_0^y t^{a-1}(1-t)^{b-1}dt}{B(a,b)} \tag{C.2}$$

and B is the (ordinary) beta function $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$. Computation of the incomplete beta function, as well as its inverse, is efficiently implemented in many software platforms for scientific computation. In fact, the cumulative distribution function of Student's distribution is typically computed using the incomplete beta function, which is why the incomplete beta function may be preferable, although any difference in accuracy or speed may be very small.

Whichever distribution is used, due to the corrections for multiple testing, we need to compute the value of the cdfs for values very close to 1, which easily leads to numerical problems. This is because we are in fact interested in the values of one minus the cdf, and this difference will not be properly presented in the value of the cdf if the difference is very small. Such a problem can be solved using the upper option of Matlab's betainc function, or by computing the value of the cdf of the t-distribution for $-t$ instead of t.

References

Anderson, T.W., 1984. An Introduction to Multivariate Statistical Analysis2nd ed. Wiley, New York.

Bartels, A., Zeki, S., 2004. The chronoarchitecture of the human brain — natural viewing conditions reveal a time-based anatomy of the brain. NeuroImage 22, 419–433.

Beckmann, C.F., Smith, S.M., 2005. Tensorial extensions of independent component analysis for group fMRI data analysis. NeuroImage 25 (1), 294–311.

Beckmann, C.F., DeLuca, M., Devlin, J.T., Smith, S.M., 2005. Investigations into resting-state connectivity using independent component analysis. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 360 (1457), 1001–1013.

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

Calhoun, V.D., Liu, J., Adali, T., 2009. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. NeuroImage 45 (1), S163–S172.

de Pasquale, F., Penna, S.D., Snyder, A.Z., Lewis, C., Mantini, D., Marzetti, L., Belardinelli, P., Ciancetta, L., Pizzella, V., Romani, G.L., Corbetta, M., 2010. Temporal dynamics of spontaneous MEG activity in brain networks. Proceedings of the National Academy of Science (USA) 107, 6040–6045.

Esposito, F., Scarabino, T., Hyvärinen, A., Himberg, J., Formisano, E., Comani, S., Tedeschi, G., Goebel, R., Seifritz, E., Salle, F.D., 2005. Independent component analysis of fMRI group studies by self-organizing clustering. NeuroImage 25 (1), 193–205.

Fisher, R.A., 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10 (4), 507–521.

Formisano, E., Esposito, F., Kriegeskorte, N., Tedeschi, G., Salle, F.D., Goebel, R., 2002. Spatial independent component analysis of functional magnetic resonance imaging time-series: characterization of the cortical components. Neurocomputing 49 (1–4), 241–254.

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

Groppe, D.M., Makeig, S., Kutas, M., 2009. Identifying reliable independent components via split-half comparisons. NeuroImage 45 (4), 1199–1211.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640.

Himberg, J., Hyvärinen, A., Esposito, F., 2004. Validating the independent components of neuroimaging time-series via clustering and visualization. NeuroImage 22 (3), 1214–1222.

Hyvärinen, A., Ramkumar, P., Parkkonen, L., Hari, R., 2010. Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis. NeuroImage 49 (1), 257–271.

Kiviniemi, V., Kantola, J.-H., Jauhiainen, J., Hyvärinen, A., Tervonen, O., 2003. Independent component analysis of nondeterministic fMRI signal sources. NeuroImage 19 (2), 253–260.

Malinen, S., Hari, R., Submitted for publication. Comprehension of audiovisual speech: Data-based sorting of independent components of fMRI activity.

Meinecke, F., Ziehe, A., Kawanabe, M., Müller, K.-R., 2002. A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. IEEE Transactions on Biomedical Engineering 49 (12), 1514–1525.

Ramkumar, P., Parkkonen, L., Hari, R., Hyvärinen, A., in press. Characterization of neuromagnetic brain rhythms over time scales of minutes using spatial independent component analysis. Human Brain Mapping.

Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.

Taulu, S., Kajola, M., Simola, J., 2004. Suppression of interference and artifacts by the signal space separation method. Brain Topography 16, 269–275.

van de Ven, V.G., Formisano, E., Prvulovic, D., Roeder, C.H., Linden, D.E., 2004. Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. Human Brain Mapping 22 (3), 165–178.

Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J.B., Thirion, B., 2010. A group model for stable multi-subject ICA on fMRI datasets. NeuroImage 51, 288–299.