Published in final edited form as:
Neuroimage. 2009 May 1; 45(4): 1199–1211. doi:10.1016/j.neuroimage.2008.12.038.

Identifying reliable independent components via split-half comparisons

David M. Groppe[a,*], **Scott Makeig**[b], and **Marta Kutas**[c]

[a] Department of Cognitive Science, 0515, University of California San Diego, 9500 Gilman Dr. La Jolla, CA 92093-0515, USA

[b] Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California, San Diego, USA

[c] Department of Cognitive Science and Department of Neurosciences, University of California, San Diego, USA

Abstract

Independent component analysis (ICA) is a family of unsupervised learning algorithms that have proven useful for the analysis of the electroencephalogram (EEG) and magnetoencephalogram (MEG). ICA decomposes an EEG/MEG data set into a basis of maximally temporally independent components (ICs) that are learned from the data. As with any statistic, a concern with using ICA is the degree to which the estimated ICs are reliable. An IC may not be reliable if ICA was trained on insufficient data, if ICA training was stopped prematurely or at a local minimum (for some algorithms), or if multiple global minima were present. Consequently, evidence of ICA reliability is critical for the credibility of ICA results. In this paper, we present a new algorithm for assessing the reliability of ICs based on applying ICA separately to split-halves of a data set. This algorithm improves upon existing methods in that it considers both IC scalp topographies and activations, uses a probabilistically interpretable threshold for accepting ICs as reliable, and requires applying ICA only three times per data set. As evidence of the method's validity, we show that the method can perform comparably to more time intensive bootstrap resampling and depends in a reasonable manner on the amount of training data. Finally, using the method we illustrate the importance of checking the reliability of ICs by demonstrating that IC reliability is dramatically increased by removing the mean EEG at each channel for each epoch of data rather than the mean EEG in a prestimulus baseline.

Introduction

Independent component analysis (ICA) refers to a family of unsupervised learning algorithms that learn to linearly decompose a multivariate data set into maximally independent components (Hyvärinen et al., 2001; Makeig et al., 1996). ICA has often been used to analyze electroencephalographic (EEG) and magnetoencephalographic (MEG) data because, like source localization algorithms (Baillet et al., 2001), ICA can potentially identify and separate individual EEG/MEG sources and thereby vastly improve the informativeness of EEG/MEG data (Ghahremani et al., 1996; Makeig et al., 2000; Tang et al., 2005). However, because the decomposition is learned from EEG/MEG statistics, a complication of using ICA is that the resulting independent components (ICs) might not be reliable. ICs may not be reliable, for example, if the algorithm is not trained on sufficient data or if there are multiple global minima in the algorithm's objective function. The latter

*Corresponding author. Fax: +1 858 534 1128. dgroppe@cogsci.ucsd.edu (D.M. Groppe).

occurs when a subspace of the data can be decomposed in multiple, equally independent ways (Fig. 1). Moreover, if an ICA algorithm (e.g., extended infomax ICA—Lee et al., 1999) learns iteratively, unreliable ICs may be produced by stopping training prematurely at a local minimum in the algorithm's objective function. In practice, these problems with IC reliability can be triggered by inadequate preprocessing of the data (e.g., including spatially non-stereotyped noisy data periods such as those produced by electrode movement and failure to mark major changes in state such as slow wave vs. REM sleep) or using a relatively large number of sensors (e.g., 256). Consequently, evidence of IC reliability is critical for establishing the credibility of ICA results.

Several techniques for assessing IC reliability have been used in practice or have been proposed. In practice, several researchers have reported similar ICs across different participants. In these reports, identification of similar ICs has been accomplished manually (Debener et al., 2005a,b; Srivastava et al., 2005), by automated/semi-automated clustering (Fogelson et al., 2004; Makeig et al., 2002, 2004), or via some objective criteria (Debener et al., 2007; Joyce et al., 2004; Onton et al., 2005). This approach is well motivated in that a typical goal of ICA applications is to identify ICs that frequently occur in the population of interest. However, this approach is problematic because inter-participant differences (e.g., differences in head shape, cortical folding, and alertness) can make homologous ICs from two individuals appear superficially quite different and non-homologous ICs appear similar. In principle, source localization (Baillet et al., 2001) could remedy this problem, but the inherent ambiguities of source localization can confound this approach in practice.

The complications of individual variability can obviously be avoided by assessing the reliability of ICs separately for each participant. Moreover, because individual anatomical differences (Homan et al., 1987) violate ICA's assumption that the scalp topographies of EEG/MEG sources are static throughout the data, ICA is most accurately applied separately to each individual's data as is typically done. Consequently, testing the reliability of an individual's ICs is valuable regardless of any later testing of IC reliability across participants. Specifically, assessing the reliability of an individual's ICs allows one to tune the application of ICA to enhance IC reliability (e.g., one might use this process to determine the necessary amount of data and/or how best to pre-process the data). Also, by focusing on ICs proven to be reliable, one might prevent unreliable ICs from obfuscating subsequent across-participant IC analyses.

A handful of methods for assessing the reliability of an individual's ICs have been proposed. Meinecke et al. (2002) and Himberg et al. (2004) suggested different methods of using bootstrap resampling to estimate IC reliability. Himberg et al. (2004) also suggested using different initial conditions for iterative ICA algorithms, and Harmeling et al. (2004) proposed applying ICA multiple times to a participant's data after adding noise to determine which ICs are reliable.[1] All of these proposals have some shortcomings. In particular, they each require running ICA a large number of times to avoid inaccurate results due to unrepresentative resampling, excessive noise, or inopportune initial conditions. For example, Efron and Tibshirani (1993, pg 188) recommend using 1000 bootstrap samples of a dataset when computing bootstrap confidence intervals.[2] For some ICA algorithms (e.g., extended

[1]More specifically, the method of Harmeling et al., called "noise injection," attempts to identify reliable ICs by adding white Gaussian noise to the data and re-running ICA to determine which ICs are affected. ICs that are most altered by the added noise are deemed the least reliable.

[2]Specifically, Efron and Tibshirani recommend performing an analysis 1000 times when computing bias-corrected and accelerated bootstrap confidence intervals. Meinecke, Himberg, Harmeling and colleagues did not specify how many times one should run ICA when using their reliability algorithms. Himberg et al. (2004) ran ICA 15 times per data set to illustrate their method. Harmeling (personal communication) said that he often starts by running ICA 100 times per data set, though he noted that ideally one would keep running ICA until the reliability algorithm's grouping matrix stops changing.

infomax ICA), running ICA even a moderate number of times (e.g., 100) per experimental participant may require a prohibitive amount of time using computational resources currently available to many researchers. While this is not true of some ICA algorithms (e.g., second-order blind identification, see below, can be run a few hundred times in a reasonable amount of time), some researchers prefer to use slow algorithms and there is some evidence suggesting that extended infomax ICA is more accurate than faster algorithms such as second-order blind identification and FastICA (Delorme et al., 2007; Makeig and Delorme 2004).

A second problem with these three proposals is that when assessing an IC's reliability they only consider the activation time series of the ICs and ignore their other principal feature, their scalp topographies. The activation and scalp topography of an IC are somewhat independent (see subsequent section), thus while the activation of an IC may be quite reliable its scalp topography may be unreliable, thereby leading to spurious inferences about its anatomic origin. Finally, the reliability metrics of Harmeling, Himberg, and Meinecke and colleagues' three algorithms are hard to interpret. While their metrics are sensible and straightforward, it is nonetheless difficult to evaluate how conservative/permissive any specific reliability criterion is for each metric. For instance, Harmeling et al. (2004) proposed a root-mean-squared angle distance (RMSAD) measure of an IC's reliability, which is zero for maximally reliable ICs and increases with decreasing reliability. Thus while an IC with a RMSAD of .01 is more reliable than one with an RMSAD of .05, it is difficult to know how important that difference is and to know what a reasonable threshold value would be for defining reliable and unreliable ICs.

In this paper, we present an alternative method for identifying reliable ICs from a data set by running ICA on user-defined split-halves of the data. This approach avoids the problem of between-participant variation, its computing demands are manageable (one must perform ICA only three times for each data set), it considers both IC activations and scalp topographies, and it uses a reliability criterion (an alpha level for the test of a null hypothesis) that is probabilistically interpretable. In the following sections, we first briefly review ICA and outline the reliability algorithm. Subsequently, we explain and illustrate each step of the algorithm using an example data set and extended infomax ICA. As evidence of the method's validity, we then show that the method can perform comparably to more time intensive bootstrap resampling and is reasonably dependent on the amount of training data. Finally, using the algorithm, we illustrate the utility of checking IC reliability by showing that removing the DC component of each epoch of data can dramatically improve IC reliability relative to that obtained when the mean EEG in a prestimulus period is used to baseline each epoch.

Background: independent component analysis of EEG/MEG data

As typically applied to EEG/MEG data, ICA algorithms learn an $n \times n$ full-rank "unmixing matrix," $\mathbf{W}$, that linearly decomposes an n-sensor data set, $\mathbf{x}(t)$, into an n-dimensional time series, $\mathbf{u}(t)$, where t represents discrete moments in time:

$$\mathbf{u}(t)=\mathbf{W}\mathbf{x}(t). \tag{1}$$

The ith dimension of $\mathbf{u}(t)$ defines the "activation" of the ith IC and the goal of an ICA algorithm is to learn an unmixing matrix that makes these activations maximally temporally independent by some measure. For example, extended infomax ICA (Lee et al., 1999), learns $\mathbf{W}$ by iteratively changing an initial estimate of $\mathbf{W}$ to maximize the joint entropy between the different dimensions of $\mathbf{u}$ under a particular source probability density model

(thereby minimizing the mutual information between any subspaces of independent dimensions). In contrast, another ICA algorithm, second-order blind identification (Belouchrani et al., 1997) analytically derives an unmixing matrix such that the dimensions of **u** are uncorrelated at multiple time lags (here, decorrelation is used as a rough approximation of independence).

In addition to its activation, each IC is also characterized by a scalp topography obtained from the "mixing matrix," **A**, which is the inverse of the unmixing matrix:

$$\mathbf{x}(t)=\mathbf{A}\mathbf{u}(t)=\mathbf{W}^{-1}\mathbf{u}(t). \tag{2}$$

The *i*th column of **A** is the scalp topography of the *i*th IC, which specifies the relative polarity and magnitude of the IC's contribution to each electrode's data and can be used to identify likely anatomical origins of the IC (Debener et al., 2005b; Tang et al., 2002; Tang et al., 2005).

For the purposes of this report, it is important to note two properties of ICA. The first of these is that the magnitude and polarity of an IC's activation and scalp topography are somewhat arbitrary. To illustrate, consider two unmixing matrices **W** and **W′** that are identical except that the first row of **W** is −2 times the first row of **W′**:

$$\mathbf{w}_2=-2\mathbf{w}'_1. \tag{3}$$

The output of the first component of **W** will be twice as large as that of **W′** and of opposite polarity. However, scaling a random variable does not affect its independence relations with other random variables (Hyvärinen et al., 2001), so **W** and **W′** are equally valid by ICA criteria. The corresponding mixing matrices, **A** and **A′**, will have the inverse relationship. The first column of **A** will be −.5 times the first column of **A′**:

$$\mathbf{a}_1=-\frac{\mathbf{a}'_1}{2}. \tag{4}$$

Thus, while the scaling of one of the principal features of an IC (its activation or scalp topography) is arbitrary, the scaling of the other is then fixed (i.e., the polarity and magnitude of an IC is distributed across **w** and **a**). To deal with this ambiguity when comparing the activations or scalp topographies of multiple ICs some type of normalization must be done (e.g., Eqs. (5) and (7)).

The other ICA property worth noting is that the two principal features that define an IC, its activation and scalp topography, are somewhat independent. In other words, while the activation of an IC may be learned quite reliably, the topography of that IC might not be very reliable and vice versa. To illustrate, the reliability of the ICs from a 64 channel EEG data set (from one participant in Experiment 1, see Appendix) was estimated using the aforementioned algorithm developed by Himberg et al. (2004) called ICASSO. FastICA was run on 100 bootstrap samples and the resulting 6400 ICs were grouped according to their activations into 53 clusters as this was the number of clusters that minimized ICASSO's *R*-index (i.e., a guide to the number of latent clusters). Based on ICASSO's cluster quality index for each cluster, 30 clusters were chosen as corresponding to reliable ICs. Subsequently, the topography and activation similarity between each pair of ICs in each cluster was measured using the distance metrics described in Eqs. (5) and (6) (see below). The median similarities of these two features for each cluster are plotted in Fig. 2. Within-

cluster activation and topography similarity are clearly correlated. Thus if an IC's activation is reliable (i.e., reproduces in the bootstrap samples) its topography also tends to be reliable and vice versa. However, there are exceptions. The topography of Cluster A, composed of ICs accounting for blink and ocular muscle activity, appears quite reliable but its activation is not. Conversely, the topography of Cluster B, composed of ICs accounting largely for noise at a single electrode, is not very reliable but the activation is relatively reliable.

Assessing IC reliability via split-half comparisons

Perhaps the simplest, most intuitive way to assess the reliability of any empirical result is to replicate the experiment that generated it. A cheap approximation of true replication is to split a data set into two comparable halves. This logic is the basis of our proposed five-step test of IC reliability:

1. Perform ICA on the full data set.
2. Split the data set into two comparable halves.
3. Perform ICA on each half of the data.
4. Attempt to pair each IC from the full data set with a homologous IC from each of the halves forming triplets of possibly homologous ICs.
5. If all three members of an IC triplet are significantly similar (i.e., approximately homologous), then the member of the triplet from the full data set is deemed reliable.

This algorithm is conceptually similar to resampling based approaches for estimating reliability, such as the bootstrap and the jackknife (Efron and Tibshirani 1993), which essentially re-perform an analysis on a large number (e.g., 1000) of possible subsets of a data set. Rather than using randomly chosen subsets of the data (the bootstrap) or systematically rotating out subsets of the data (the jackknife), using split-halves privileges two particular subsets. This has the disadvantage of being influenced by any bias in the method used to split the data. Furthermore, it has less power than the bootstrap and jackknife, which can analyze more than 50% of the data in each subset (i.e., it is harder to extract the latent ICs of a dataset using 50% of the data than using more than 50%). In the present context, these disadvantages are offset by the minimal computational demands of analyzing only two subsets.

While the proposed algorithm is conceptually simple, the details of Steps 2, 4, and 5 are far from obvious and are explained in the following subsections:

Step 2: splitting the data set into two comparable halves

The sources of data variation of greatest interest to any EEG/MEG analysis are the experimental manipulations (e.g., target vs. standard stimuli in an oddball task). Thus it is important to split the data such that each experimental condition is equally represented in each half. A potentially larger source of data variation is the experimental participant's alertness. This tends to diminish throughout the course of an experiment and can have large consequences for the participant's scalp potentials/magnetic fields and behavior. As the experiment progresses participants are more likely to blink, to relax their muscles, to make mistakes, and to elicit EEG/MEG alpha activity. To attempt to equally represent both of these factors in the two halves of the data, we propose separating the odd and even trials of each experimental condition.

Step 4: attempt to pair each IC from the full data set with a homologous IC from each of the halves forming triplets of possibly homologous ICs

Triplets of possibly homologous ICs are formed by pairing each IC from the full data set with an IC from each half of the data. How is one to pair ICs from two comparable data sets? Identifying a possibly homologous pair of ICs first requires an IC similarity or distance metric. To determine if two ICs are homologous it is important to compare both their activations and scalp topographies; however, to find a pair of possibly homologous ICs, it is sufficient to look at only one of these features. Since the dimensionality of IC topographies is orders of magnitude less than that of IC activations and is easier to visualize, we propose using one minus the absolute value of the cosine similarity metric, $\text{dist}_{\text{topo}}$, to compare IC scalp topographies as a first step in forming pairs of ICs.

$$\text{dist}_{\text{topo}}(i, j) = 1 - \left| \frac{\mathbf{a}_i^t \mathbf{a}_j}{\|\mathbf{a}_i\| \|\mathbf{a}_j\|} \right|. \tag{5}$$

This distance metric deals with the aforementioned IC scaling ambiguity by effectively normalizing the scalp topographies to unit root-mean square (RMS) magnitude and setting their polarities to minimize distance.[3]

This leads to following pairing algorithm:

1. Compute the scalp topography distance, $\text{dist}_{\text{topo}}$, between each possible pair of ICs to quantify their similarity.
2. Pair the two most similar ICs and remove them from further consideration.
3. Repeat Step 3 until all the ICs are paired.

This procedure "greedily" pairs up ICs with the most similar scalp topographies, which we believe is preferable to the more computationally savvy solution of minimizing the total distance of all pairs (Kuhn, 1955). Since it is likely that not all of the ICs replicate across decompositions, we have found that attempting to minimize the distance of all pairs tends to sacrifice highly similar pairs to avoid highly dissimilar pairs (Supplemental Fig. 1).

Step 5: identify homologous ICs

Once possibly homologous triplets of ICs have been identified, what remains to be determined is which triplets are similar enough to be considered equivalent. Constructing such a criterion requires an IC similarity or distance metric, an appropriate null hypothesis, and an appropriate critical region of the distribution of that metric under the null hypothesis. Each of these is discussed in turn below.

Step 5.1: an IC distance metric—As already mentioned, assessing IC reliability requires the comparison of both IC scalp topographies and activations. For comparing IC scalp topographies, we use the $\text{dist}_{\text{topo}}$ metric described above. The complementary distance metric for two IC activations, u_i and u_j, is:

$$\text{dist}_{\text{act}}(i, j) = \max(f(i, j), f(j, i)) \tag{6}$$

[3]Using cosine similarity is preferable to using Pearson's *r* or rank correlation as cosine similarity preserves the orientation of the scalp topography vector while being invariant only to vector scale. In contrast, Pearson's *r* distorts the vector by removing the mean and rank correlation is invariant to monotonic transformations (not just changes in scale).

where:

$$f(i,j)=\frac{\sum_t (u_i(t)\,\|a_i\|\,\mathrm{sign}(a_i^t a_j) - u_j(t)\|a_j\|)^2}{\sum_t (u_i(t)\,\|a_i\|)^2}. \tag{7}$$

This is the maximum normalized sum squared difference (i.e., residual variance) between the pair of IC activations. This measure equals zero for identical activations and grows towards infinity as the activations become increasingly dissimilar. Note, for the IC activations to be comparable, they need to be derived from the same set of scalp data, **x**(t):

$$u_j(t)=\mathbf{w}_i\mathbf{x}(t) \tag{8}$$

$$u_j(t)=\mathbf{w}_j\mathbf{x}(t). \tag{9}$$

Using residual variance (Eq. (7)) as a similarity metric has the advantage that the difference between activations is scaled in proportion to the magnitude of one of the ICs. Thus small differences between large activations will be down-weighted relative to small differences between small activations. However, the residual variance between two activations depends on which IC's activation is in the denominator. We choose to normalize by the smaller activation in Eq. (6) (i.e., normalizing by the smaller activation produces a larger value in Eq. (7)). Doing otherwise would lead ICs with large activations to appear rather similar to any IC with a small activation.

Step 5.2: a null hypothesis—While it is relatively straightforward to quantify the similarity of two ICs with metrics like the ones just described, it is not at all clear how similar a pair of ICs needs to be before they can be considered homologous. Ideally, one would like to test the null hypothesis that the two ICs are not homologous, but a priori, there seems to be no way to know what the distribution of non-homologous IC similarities should be. An obvious alternative would be to use a permutation test (Fisher et al., 1993, pg. 60), which does not require assumptions about the distribution of the null hypothesis. That is, one could randomly permute the features of each IC (e.g., the scalp topography weights) thousands of times and compute the similarity of each permuted pair. If the original ICs are more similar than $(1-\alpha)$* 100% of the permuted pairs, then the original pair would be declared homologous with a *p*-value less than α. In practice, this method is too permissive. For example, using the data of one participant (from Experiment 1 described in the Appendix) we compared the scalp topographies of the 64 ICs from the data to one another. Because the ICs all came from the same decomposition, none of them should be homologous to any of the others. However, of the 2016 possible pairs, we found that 6% of the pairs were more similar than 4000 random permutations of those pairs. Thus using a permutation test with an alpha level of 1/4001 would produce a false positive rate of 6/100 (over 200 times what it should be).

Another alternative to assuming the distribution of a null hypothesis is the family of "bootstrap" methods, where one approximates the distribution of a random variable from the "empirical distribution" of one's data (Efron and Tibshirani, 1993). It is this general approach that we adopt here. Specifically, to approximate the distribution of the similarity of non-homologous IC pairs, we use the similarity of all possible IC pairs from two different

ICA decompositions. Presumably some of the IC pairs are homologous, but at most there will be n homologous pairs out of the n^2 possible pairs. So the bias will be relatively small, and to the extent that it does bias the approximation, it will make the approximation more conservative (i.e., the test will be less likely to declare two ICs homologous). Note, for each participant, $3n^2$ IC pair distances are obtained (n^2 for each of the three comparisons: "Half A" to "Half B", Half A to the whole data, Half B to the whole data). To increase the stability of the empirical distribution, we combine the $3n^2$ IC pair distances from all study participants.

To illustrate, the empirical distribution of scalp topography distances of ICs derived from 64 channel data from sixteen participants (see Appendix: Experiment 1) are plotted in Fig. 3 (Top). The mode of the distribution is approximately one, maximally dissimilar. As distance decreases, the frequency of pairs decreases until there is a small spike in frequency near zero, maximum similarity. This spike presumably reflects homologous pairs. The empirical distribution of IC activation distances is also plotted in Fig. 3 (Bottom). Again, there is a small peak at the maximally similar end of the scale. The mode of the distribution is around 2.23. This is slightly greater than the expected distance between two uncorrelated zero-mean activations of equal variance, two.

Step 5.3: a critical region—Given a distance distribution, we must select a critical region for rejecting the null hypothesis (i.e., that a pair of ICs are no more similar than a pair of ICs chosen at random) at a particular α level. Because it is necessary to test both IC features, the critical region is a segment of the two dimensional joint distance distribution (e.g., Fig. 4). Obviously the critical region should cover the most similar corner of the joint distribution, but several shapes are possible (e.g., rectangular, a quarter circle). We use an "L" shaped region (demarcated by pink boxes in Fig. 4), which favors IC pairs with one highly similar feature over pairs with two somewhat similar features.

We believe that the area of the critical region should be just big enough such that all ICs from one decomposition could be significantly similar to exactly one IC from the other decomposition. This requires the critical region to contain $1/n$ of the empirical distribution. To accomplish this, we construct the critical region from two rectangles. The rightmost edge of one rectangle is set at the 10th percentile of the scalp topography distance distribution (. 28 in Fig. 4) and the highest edge of the other is set to the 10th percentile of the activation distance distribution (2.22 in Fig. 4). The other edge of each rectangle is then grown, one sample at a time, until both rectangles together contain $1/n$ of the empirical distribution.

It is important to note that in practice, the Type I error rate of this test will be higher than $1/n$, even if the empirical distribution does accurately approximate the true distribution. This is because the ICs are first paired according to their scalp topographies and then tested, thereby increasing the chances of rejecting the null hypothesis. While this bias makes it impossible to know what the true Type I error rate of the test is, it does not invalidate the method. This is because activation similarity is also required to reject the null hypothesis, but activations are not used to pair ICs.

An example: visual oddball EEG data

To more clearly explain this conceptually simple but somewhat complicated method for assessing IC reliability, we illustrate the method here using 64 channel EEG data collected from 16 participants performing two visual target detection (oddball) tasks. Each task simply required the participant to silently count the number of occurrences of an infrequent class of target stimuli among frequent "standard" stimuli. The two tasks differed only in the nature and presumed difficulty of the target/standard discrimination. Prior to analysis, the EEG was

parsed into 1 second epochs (100 ms prestimulus to 900 ms poststimulus). Further data acquisition and preprocessing details are given in the Appendix (Experiment 1).

Subsequent to running ICA on each participant's data (which produced one set of 64 ICs per participant), the epochs of each participant's data time locked to targets and standards in the two target detection tasks were separated to form four sets of trials (two stimulus types for each of the two tasks). Each set of trials was then sorted according to their time of occurrence and odd and even numbered trials in each set were split to form two halves of data, "Half A" and "Half B." ICA was applied separately to each half, producing two additional sets of 64 ICs/participant. The distances of all possible pairs of ICs from Half A to Half B, from Half A to the whole data set, and Half B to the whole data set were computed (Step 5.1). These distance measurements from all sixteen participants were combined to approximate the distribution of distances of randomly paired ICs (Step 5.2; Figs. 3–4).

Next,ICs from each half were paired (Step4) with an IC from the full data set, producing triplets of ICs (e.g., Fig. 5, Supplemental Figs. 2–4). If each member of the triplet was significantly similar (Step 5.3) to the other two members of the triplet, then the IC from the whole data set was considered reliable.

Fig. 5 and Supplemental Figs. 2–4 present the scalp topographies of all the ICs from Participant 1's whole data and the scalp topographies of each IC's match from the two halves. ICs are numbered in decreasing order of the mean variance of their scalp-projected activations (Makeig et al., 1997). Thus ICs that greatly contribute to the scalp data (e.g., IC 1, shown in Fig. 5, which accounts primarily for blink activity) have low numbers. Of the 64 ICs from the whole data set, the method finds 33 to be reliable. The reliable ICs are generally the most physiologically plausible (e.g., ICs 4, 10, and 14 in Fig. 5) and typically have large scalp-projected activations (e.g., only one of ICs 49–64 is reliable), but there are some exceptions. For example, IC 7's scalp topography is consistent with a single midline dipolar source and it contributes a relatively large proportion of the scalp data. However, there are no ICs with similar scalp topographies in the decompositions of either half of the data. ICa 6 and ICb 18 are similar to complementary halves of IC 7, but the activations of ICa 6 and ICb 18 are quite different ($dist_{act}$ =3.99; Supplemental Fig. 5). IC 3 is a less physiologically plausible component but its activation appears more strongly at the scalp. It also does not replicate in either half of the data. The ICs from the halves of the data with the most similar scalp topographies are ICa 3 and ICb 2, but the activations of IC 3 and ICb 2 are rather different ($dist_{act}$ =2.28; Supplemental Fig. 6) and their scalp topographies suggest distinct anatomical sources.

Some of the ICs that did not meet the criteria for reliability are less clearly unreliable. For example, IC 30 replicated in Half B and paired with a similar component from Half A, ICa 22 (Supplemental Figs. 2 and 7). However, the scalp topography of ICa 22 was not quite similar enough to ICb 34 ($dist_{topo}$ =.29); thus it was deemed unreliable. A similar example is IC 8, which replicated in Half B but not in Half A (Fig. 5). Both IC 30 and IC 8 might be reliable, but there were not enough data to robustly learn them in both halves of the data. Resampling methods, such as the aforementioned bootstrap and jackknife, that use more than 50% of the data when re-running ICA may be better at extracting these borderline reliable ICs.

Comparison with bootstrap resampling

As already mentioned, using split-halves of the data has the disadvantages of (1) being influenced by any bias in the method used to split the data and (2) using only half of the data in the two subsamples. To evaluate how serious these disadvantages are, the split-half

comparison reliability test was compared to bootstrap resampling, which has neither of these shortcomings.

This comparison was made using the visual target detection data just described. For each participant, 1000 bootstrap samples were created by randomly selecting η epochs from the original data set with replacement (where η is the number of epochs in the original data set). Extended infomax ICA was then applied to each bootstrap sample using the same parameters as in the initial decomposition, producing 1000 sets of "bootstrap ICs" for each participant. Bootstrap ICs from each of the 1000 bootstrap decompositions were uniquely paired with an IC from the original data set using their scalp topographies (Step 4). The scalp topography and activation distances of each pair were computed using the metrics previously described. The median distance of each original IC to its 1000 bootstrap counterparts was used to quantify the reliability of the two features of that IC. The original ICs were used to group the ICs from each bootstrap sample primarily to facilitate comparison of the bootstrap and split-half results. Median similarity was used to mitigate the influence of extreme outlying values. Just as with the bootstrapped data sets, the median distance of each IC from each half of each participant's data and its corresponding IC from the original decomposition was computed as well.

Fig. 6 shows that the two resampling methods produce highly correlated reliability estimates. In other words, the ICs whose scalp topographies and activations reproduce the best across the two split-halves also tend to be the most reproducible according to the 1000 bootstrap samples. Thus using split-halves approximates the general results of the more time intensive bootstrap resampling for these data.

Although reliability estimates from the two resampling techniques are similar, one might expect that bootstrap resamples would find more ICs to be reliable since they can include more than 50% of the data in each bootstrap sample. To test this possibility, all ICs whose median scalp topography and activation distances met the significance criteria used in the previous section (i.e., [$dist_{topo} < .28$ and $dist_{act} < 1.96$] or [$dist_{topo} < .13$ and $dist_{act} < 2.22$]) were deemed reliable. On average, 45 ICs per participant (SD=10) were reliable using bootstrap median feature distances. In comparison, 44 ICs per participant (SD=7) were reliable using the split-half median feature distances.[4] Using a one-tailed paired *t*-test, this difference failed to reach significance ($t(15)=.92$, $p=.19$). Consequently, there appears to be no clear advantage to the more time intensive bootstrap resampling. With smaller or noisier data sets this probably would not be the case.

Reliability vs. amount of data

Because the reliability criteria proposed here are estimated from the data, there is a danger that its performance may greatly degrade with decreasing amounts of data. In particular, since the similarity criteria are loosened to include *n* pairs of ICs, it is possible that the algorithm will still find a large number of ICs to be reliable even with insufficient data.

To test for this possibility, we assessed the reliability of ICA decompositions of four data sets using decreasing amounts of data: the previously described 64 channel visual target detection data set (Experiment 1), another 64 channel data set (Experiment 2), and two 30 channel data sets (Experiments 3 and 4; all data set details are given in the Appendix). For the data sets from Experiments 1 and 2, extended infomax ICA was performed on 100%,

[4]This is a larger number of reliable ICs than found by the split-half comparison reliability test reported in the previous section (which found 33 ICs to be reliable). This is because the previous test required that three pairs of ICs be significantly similar for the original IC to be deemed reliable. Here, only the median similarity need be strong enough to gain a rating of reliability. This was done to facilitate comparison with bootstrap resampling.

66%, 50%, and 25% of the full data sets. For the 30 channel data sets, extended infomax ICA was performed on 100%, 66%, 40%, 20%, and 2% of the full data sets. To split the data into *b*% subsets, epochs were first divided by stimulus type and sorted in order of occurrence. Subsequently, every *c*th epoch (where c=200/b rounded to the nearest integer) and the subsequent epoch were placed into the same subset.[5] This approximately balanced the number of epochs of each stimulus type and the latencies at which the epochs were recorded across subsets. For example, for each participant in Experiment 4, this procedure produced 1, 3, 5, 10, and 100 data subsets each including 100%, 66%, 40%, 20%, and 2% of the data respectively.

Fig. 7 shows that the number of ICs in these data sets judged reliable decreased as the amount of data decreased. Indeed, for all four experiments there was an approximately linear relationship between the log of the number of time points per channel^2, and the percentage of reliable ICs. Thus, in the range of data set sizes explored here, the results of the ICA reliability algorithm appears to depend on the amount of data in a reasonable manner. Given that EEG/MEG data sets collected in most cognitive experiments would commonly fall above the lower end of this range[6], it appears that the algorithm should perform reasonably well in practice.

It also worth noting that these results provide researchers a sense of the amount of data necessary to use extended infomax ICA. For example, Onton et al. (2006) suggested that a minimum of 20 time points per channel^2 is necessary for confidently applying ICA while noting that more data generally help. These results corroborate that 20 time points per channel^2 produce a moderate percentage of reliable ICs and they suggest that to increase the percentage of reliable ICs much beyond this range, considerably more data are needed (since the percentage of reliable ICs appears to be a function of the log of the number of time points per channel^2). However, it is not clear if these results will generalize to data from a larger number of sensors or to different sampling rates (and concomitant pass band).

Improving IC reliability

To apply ICA to a data set, a number of practical decisions need to be made: (1) how to preprocess the data, (2) which ICA algorithm to use, and (3) what values of the algorithm's parameters to use. The reliability test presented here can be used to explore different answers to these questions to improve ICA results. While reliability is not a sufficient condition for a successful ICA decomposition, it is a necessary condition and tuning ICA to improve reliability may improve the quality of ICs in general.

For example, extended infomax ICA was applied to Participant 7's data from Experiment 3 and only nine of the 30 ICs were judged to be reliable (Supplemental Figs. 8 and 9). This was remarkably low given the large amount of data (472 time points per channels^2), the fact that the data from the other seven participants produced 20 reliable ICs on average, and the fact that many of the unreliable ICs appeared physiologically plausible (e.g. ICs 4, 8, and 12). Attempts to improve the reliability of Participant 7's decomposition by removing outlier epochs or by using second-order blind identification instead of extended infomax ICA did not improve results. Eventually it was discovered that IC reliability for these data could be improved dramatically by removing the mean of each epoch instead of the mean of the 100 ms prestimulus baseline (a common EEG/MEG preprocessing step) before applying ICA. Applied to the zero-mean epoch data, extended infomax ICA returned 25 reliable ICs

[5]For example, if c=4, then the first, fifth, ninth, etc. epochs would be placed in a subset along with their immediately following epochs (i.e., second, sixth, tenth, etc.).

[6]One hour of continuous EEG/MEG data, recorded at 250 HZ will provide 879, 220, 55, and 14 time points/channel^2 for 32, 64, 128, and 256 channel data (respectively). For comparison with Figure VI, the $\log_{10}$ of those ratios are 2.9, 2.3, 1.7, and 1.1.

(Supplemental Figs. 10 and 11). Comparison of the original ICs and the more reliable ICs (Supplemental Figs. 12 and 13) reveals that many of the same ICs were extracted in both decompositions. Only three of the original ICs had no convincing counterpart in the second decomposition (Supplemental Fig. 13, bottom row). Thus, removing the mean of each epoch appears to have enhanced the strength of the latent independent components so that they could be extracted from less data. Presumably, preprocessing the data so as to enhance the ability of ICA to find latent components increases the accuracy of IC activations and scalp topographies as well.

While this data set is an extreme case, removing the mean of each epoch produced more reliable ICs than removing the 100 ms prestimulus baseline for each participant in all four experiments listed in the Appendix (48 data sets in total). On average, zero-mean epoch data produced over 26% more reliable ICs (Fig. 8). It is not clear what causes this difference. Removing the mean of each epoch acts as a leaky high-pass filter, zeroing the DC component and dampening low frequency components. Perhaps the low frequency EEG components (at least when viewed in approximately one second epochs) are more variable and/or have a statistical structure that is difficult for ICA to uniquely decompose (e.g., similar to the hypothetical sources in Fig. 1). On the other hand, it could be that removing each epoch's mean does not aid ICA so much as removing the prestimulus baseline impairs it. Raw scalp potentials are highly variable, especially when EEG artifacts are present (as is generally the case when using ICA). Using only a small window to normalize the voltage of each epoch may enhance that variability and mask the latent sources ICA is sensitive to.

Whatever the root cause proves to be, removing epoch means is clearly an improvement over the more conventional 100 ms prestimulus baseline and demonstrates the utility of using IC reliability to explore ICA parameters. In practice, researchers have applied ICA to epoched data after removing a prestimulus baseline (Debener et al., 2005a; Makeig et al., 2004) and to continuous data with various degrees of bandpass filtering (Debener et al., 2005b; Onton et al., 2005; Tang et al., 2005). The reliability test presented here provides a simple, relatively time-efficient way to test whether these and/or various alternative data preprocessing steps, ICA algorithms, or ICA parameter values improve the robustness and possibly the accuracy of ICA results.

Discussion

To summarize, ICA is a useful technique for analyzing EEG/MEG data, but, for several reasons, the resultant ICs might not be reliable and it is thus essential to assess their reproducibility. This was clearly illustrated by the ICA results from Participant 7's data from Experiment 3 (see previous section), which was substantial (relative to the number of channels^2) but produced only a few reliable ICs when a 100 ms prestimulus baseline was used in pre-processing.

While some researchers may be tempted to assess IC reliability by simply re-running ICA a second time using different initial conditions or every other data point and judging reproducibility by eye, such crude checks are most likely problematic. Specifically for many ICs, it probably won't be obvious if they were learned in both decompositions nor will only a second decomposition necessarily give a good sense of how reliable each IC is. This was illustrated in the split-half test of one participant's data from a visual oddball experiment (Fig. 5, Supplemental Figs. 2–7) that found several ICs that were borderline reliable across all three decompositions (e.g., IC 2) or replicated in one half of the data but not the other (e.g., IC 8). Moreover, running ICA multiple times simply using different initial conditions will probably over-estimate IC reliability, since the training data are always the same.

Consequently, researchers are well-advised to use an objective, more-principled method to assess IC reliability. The split-half reliability algorithm presented herein appears to be an efficient and effective such method. Although some steps of the algorithm are complicated, the logic is intuitive: ICs from the original data set that also appear when ICA is run on comparable halves of the data are deemed reliable. This algorithm improves upon some existing methods (Harmeling et al., 2004; Himberg et al., 2004; Meinecke et al., 2002) in that it considers both scalp topographies and activations of ICs, uses a probabilistically interpretable threshold for IC reliability, and only requires applying ICA three times per data set, which is critical for relatively slow ICA algorithms like infomax ICA.

There are three key potentially problematic drawbacks to the proposed algorithm. The first of these is that splitting data sets in half may make the algorithm overly conservative by biasing it to underestimate IC reliability. While the comparison of IC reliability from split-half comparisons with that derived from bootstrap resampling found no clear evidence that using only half of the data reduced the number of reliable ICs, with smaller or noisier data sets this would probably not be the case. In situations where the amount of data is limited (e.g., those obtained from some clinical populations), more time-intensive tests of IC reliability may be worth the additional computational labor.

A second drawback of this algorithm relative to the reliability algorithms proposed by Meinecke et al. (2002) and Harmeling, et al. (2004), is that it tests only for the reliability of *single* ICs, whereas those other two algorithms can detect reliable single ICs and subspaces of the data that span *multiple* ICs. For example, two somewhat dependent EEG/MEG sources with distinct topographies could be decomposed by ICA as two unreliable ICs. However, the subspace spanned by those two ICs could be reliable and could be identified by the algorithms of Meinecke et al. and Harmeling et al. While this is a limitation of the split-half algorithm, the results presented here suggest that a large percentage of ICs are individually reliable (around 45% and 95% for the 64 and 30 channel data sets respectively —Fig. 8) and often researchers are primarily interested in single ICs, not in subspaces of ICs (e.g., Debener et al., 2005b, Onton et al., 2005, Tang et al., 2005). So, in practice the split-half algorithm should still be quite useful.

The final potential drawback of this algorithm and other proposed ICA reliability algorithms (Debener et al., 2005b, Onton et al., 2005, Tang et al., 2005) is that it is not known how accurate their results are. Ideally one would like to know how often the algorithm mistakenly deems an IC as reliable when its features are unreliable enough to produce qualitatively erroneous inferences (e.g., mislocalization of the IC to a qualitatively different area of cortex or drawing an erroneous conclusion about its dynamics) and how often the algorithm mistakenly deems an IC as unreliable when its features are reliable enough to allow qualitatively accurate inference. The fact that the number of ICs deemed reliable by the algorithm presented here increases with the amount of data is qualitatively consistent with how an accurate reliability algorithm should behave. However, to quantitatively assess how accurate this and other ICA reliability algorithms are, when applied to EEG/MEG data, it will be necessary to apply them to simulated EEG/MEG data. Unfortunately such a project is complicated by the fact that it is not clear what the parameter values (e.g., number and size of sources) of realistic simulations of EEG/MEG should be. Indeed, to our knowledge, no existing IC reliability algorithms have been tested on such data. With continued advances in our understanding of the structure of EEG/MEG source activity (Nunez and Srinivasan 2006; Rowe et al., 2004), realistic simulations should be constructed to evaluate IC reliability algorithms and the many variant ICA algorithms themselves.

Finally, we note one way in which the algorithm presented here might be improved. Specifically, using more complicated IC features based on the physical origins of the EEG/

MEG instead of the straightforward $dist_{topo}$ and $dist_{act}$ metrics used here might improve the algorithm's performance. For example, comparing the estimated source locations of ICs might be a better metric for comparing their scalp topographies than $dist_{topo}$. As the anatomical source of an IC, rather than its scalp topography, is often what researchers would really like to know, the distance between estimated sources of potentially homologous ICs would directly reflect the goal of the analysis. Moreover, the distance between estimated sources would be easier to interpret than a unitless distance metric like $dist_{topo}$. However, estimating the location of sources is generally a non-trivial under-determined problem (Baillet et al., 2001) that requires making assumptions (e.g., number of sources) that may not be valid. For this reason, we used the $dist_{topo}$ metric here, but other features and metrics are surely worth exploring.

EEGLAB compatible Matlab code for implementing the algorithm can be downloaded from: http://www.cogsci.ucsd.edu/~dgroppe/eeglab.html.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Laura Kemmer for providing the data in Experiments 3 and 4. We would also like to thank Virginia de Sa, Patrick Gallagher, Rey Ramirez, Jason Palmer, and two anonymous reviewers for helpful comments on the work presented in this paper. This research was supported by US National Institute of Child Health and Human Development grant HD22614 and National Institute of Aging grant AG08313 to Marta Kutas and by a Center for Research in Language Training Fellowship and an Institute for Neural Computation Training Fellowship to David Groppe.

References

Baillet S, Mosher JC, Leahy RM. Electromagnetic brain mapping. IEEE Signal Process Mag. 2001; 18(6):14–30.

Bell AJ, Sejnowski TJ. An information–maximization approach to blind separation and blind deconvolution. Neural Comput. 1995; 7(6):1129–1159. [PubMed: 7584893]

Belouchrani A, Abed Meraim K, Cardoso JF, Moulines E. A blind source separation technique using second order statistics. IEEE Trans Signal Process. 1997; 45:434–444.

Debener S, Makeig S, Delorme A, Engel AK. What is novel in the novelty oddball paradigm? Functional significance of the novelty P3 event-related potential as revealed by independent component analysis. Cogn Brain Res. 2005a; 22(3):309–321.

Debener S, Ullsperger M, Siegel M, Fiehler K, von Cramon DY, Engel AK. Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. J Neurosci. 2005b; 25(50):11730–11737. [PubMed: 16354931]

Debener S, Strobel A, Sorger B, Peters J, Kranczioch C, Engel AK, Goebel R. Improved quality of auditory event-related potentials recorded simultaneously with 3-T fMRI: removal of the ballistocardiogram artefact. NeuroImage. 2007; 34 (2):587–597. [PubMed: 17112746]

Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004; 134 (1):9–21. [PubMed: 15102499]

Delorme A, Sejnowski TJ, Makeig S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. NeuroImage. 2007; 34 (4):1443–1449. [PubMed: 17188898]

Efron, B.; Tibshirani, R. An Introduction to the Bootstrap. Chapman & Hall; New York: 1993.

Fisher, NI.; Lewis, T.; Embleton, BJJ. Statistical Analysis of Spherical Data. Cambridge University Press; New York: 1993.

Fogelson N, Loukas C, Brown J, Brown P. A common N400 EEG component reflecting contextual integration irrespective of symbolic form. Clin Neurophysiol. 2004; 115(6):1349–1358. [PubMed: 15134702]

Ghahremani, D.; Makeig, S.; Jung, T-P.; Bell, AJ.; Sejnowski, TJ. Independent Component Analysis of Simulated EEG Using a Three-Shell Spherical Head Model. Institute for Neural Computation, University of California; San Diego, La Jolla, California, USA: 1996. Report nr 96-01

Groppe, DM. Unpublished PhD Dissertation. University of California; San Diego: 2007. Common independent components of the P3b, N400, and P600 ERP components to deviant linguistic events.

Harmeling S, Meinecke F, Müller KR. Injecting noise for analysing the stability of ICA components. Signal Process. 2004; 84(2):255–266.

Himberg J, Hyvärinen A, Esposito F. Validating the independent components of neuroimaging time series via clustering and visualization. NeuroImage. 2004; 22 (3):1214–1222. [PubMed: 15219593]

Homan RW, Herman J, Purdy P. Cerebral location of international 10–20 system electrode placement. Electroencephalogr Clin Neurophysiol. 1987; 66(4):376–382. [PubMed: 2435517]

Hyvärinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. J. Wiley; New York: 2001.

Joyce CA, Gorodnitsky IF, Kutas M. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. Psychophysiology. 2004; 41 (2):313–325. [PubMed: 15032997]

Kemmer L, Coulson S, De Ochoa E, Kutas M. Syntactic processing with aging: an event-related potential study. Psychophysiology. 2004; 41 (3):372–384. [PubMed: 15102122]

Kuhn H. The Hungarian method for the assignment problem. Nav Res Logist Q. 1955; 2:83–97.

Lee TW, Girolami M, Sejnowski TJ. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Comput. 1999; 11(2):417–441. [PubMed: 9950738]

Makeig, S.; Delorme, A. Physiological Plausibility and Stability of Independent Component Analysis. Cognitive Neuroscience Society Annual Meeting Program; San Francisco. 2004.

Makeig, S.; Bell, AJ.; Jung, T-P.; Sejnowski, TJ. Independent component analysis of electroencephalographic data. In: Touretzky, D.; Mozer, M.; Hasselmo, M., editors. Advances in Neural Information Processing Systems. Vol. 8. MIT Press; Cambridge, MA: 1996. p. 145

Makeig S, Jung T-P, Bell AJ, Ghahremani D, Sejnowski TJ. Blind separation of auditory event-related brain responses into independent components. Proc Natl Acad Sci U S A. 1997; 94(20):10979–10984. [PubMed: 9380745]

Makeig, S.; Jung, T-P.; Ghahremani, D.; Sejnowski, TJ. Independent component analysis of simulated ERP data. In: Nakada, T., editor. Integrated Human Brain Science. Elsevier; New York: 2000.

Makeig S, Westerfield M, Jung TP, Enghoff S, Townsend J, Courchesne E, Sejnowski TJ. Dynamic brain sources of visual evoked responses. Science. 2002; 295 (5555):690–694. [PubMed: 11809976]

Makeig S, Delorme A, Westerfield M, Jung T-P, Townsend J, Courchesne E, Sejnowski TJ. Electroencephalographic brain dynamics following manually responded visual targets. PLoS Biol. 2004; 2(6):e176. [PubMed: 15208723]

Meinecke F, Ziehe A, Kawanabe M, Müller KR. A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. IEEE Trans Biomed Eng. 2002; 49(12):1514–1525. [PubMed: 12549733]

Nunez, PL.; Srinivasan, R. Electric Fields of the Brain: The Neurophysics of EEG. 2. Oxford University Press; Oxford; New York: 2006. p. 611

Onton J, Delorme A, Makeig S. Frontal midline EEG dynamics during working memory. NeuroImage. 2005; 27 (2):341–356. [PubMed: 15927487]

Onton J, Westerfield M, Townsend J, Makeig S. Imaging human EEG dynamics using independent component analysis. Neurosci Biobehav Rev. 2006; 30(6):808–822. [PubMed: 16904745]

Rowe DL, Robinson PA, Rennie CJ. Estimation of neurophysiological parameters from the waking EEG using a biophysical model of brain dynamics. J Theor Biol. 2004; 231(3):413–433. [PubMed: 15501472]

Srivastava G, Crottaz-Herbette S, Lau KM, Glover GH, Menon V. ICA-based procedures for removing ballistocardiogram artifacts from EEG data acquired in the MRI scanner. NeuroImage. 2005; 24 (1):50–60. [PubMed: 15588596]

Tang AC, Pearlmutter BA, Malaszenko NA, Phung DB, Reeb BC. Independent components of magnetoencephalography: localization. Neural Comput. 2002; 14(8):1827–1858. [PubMed: 12180404]

Tang AC, Sutherland MT, McKinney CJ. Validation of SOBI components from high-density EEG. NeuroImage. 2005; 25 (2):539–553. [PubMed: 15784433]

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi: 10.1016/j.neuroimage.2008.12.038.

Appendix B

The EEG data used in this paper came from the following four experiments. All experimental participants participated in exchange for academic credit and/or pay after providing informed consent. The University of California, San Diego Institutional Review Board approved all studies.

Extended infomax ICA (Lee et al., 1999) was individually applied to each participant's data from each experiment using the *binica* function from the EEGLAB Toolbox (Delorme and Makeig, 2004). The extended infomax ICA algorithm was set to estimate the number of subgaussian sources after every training block. Besides this, the default ICA training parameters were used. Specifically, ICA training stopped when the cosine of the angle between the unmixing matrix change of the current and of the previous training step was less than 1e−7 or after 512 training steps (whichever came first). The initial ICA learning rate was .0001. The ICA output unit bias was updated online and the initial state of the ICA unmixing matrix was a sphering matrix (two times the inverse of the principal square root of the data covariance matrix).

Experiment 1

The data in this experiment consisted of 64 channel EEG recorded from 16 participants while they performed alternating blocks of two visual target detection (oddball) tasks. All electrodes were tin and were referenced to the left mastoid. Electrode impedances were kept below 5 KΩ. EEG was processed through Grass amplifiers set at a bandpass of .01–100 Hz, continuously digitized at 250 Hz, and stored on hard disk for later analysis. In each target detection task, the participants silently counted occurrences of a rare class of target stimuli among a series of common "standard" stimuli. Stimulus onset asynchrony (SOA) was 1000–1500 ms. EEG data was re-referenced offline to the algebraic mean of the left and right mastoids and divided into 1 second, non-overlapping epochs extending from 100 ms before to 900 ms after stimulus onset. Each epoch was 50 Hz low-pass filtered and the mean of each epoch was removed. After filtering, individual epochs were rejected via a combination of visual inspection and objective tests designed to detect blocking, drift, and extreme values. These tests are included in the EEGLAB Toolbox (Delorme and Makeig, 2004). After epochs were rejected, the mean number of epochs per participant was 643 (minimum: 548). For full experiment details see Groppe (2007).

Experiment 2

The data from Experiment 2 were obtained from the same participants and in the same recording session as Experiment 1. In this experiment, participants read sentences presented

one word at a time (SOA of 300 ms) on a computer monitor. 25% of the sentences ended in a semantically anomalous word (e.g., “The tweeting in the shed sounded like baby tasks.”), 25% of the sentences ended in a grammatically anomalous word (e.g., “Once a month, Carol goes to the theater with I.”), and the remaining sentences were well-formed. 1300 ms after the last word of the sentence, participants were asked to indicate if sentences contained anomalies. EEG pre-processing was the same as that used in Experiment 1 (epochs were time locked to word onset). After rejecting epochs polluted by recording artifacts, the mean number of epochs per participant was 844 epochs (minimum: 660). Due to the short SOA between words, some epochs of data slightly overlapped. On average, 99.73% of the time points of each participant’s data occurred in only one epoch (minimum: 99.67%). For full experiment details see Groppe (2007).

Experiment 3

The data in this experiment consisted of 30 channel EEG recorded from eight participants while they performed alternating blocks of three visual target detection (oddball) tasks. EEG recording parameters were the same as those in Experiments 1–2. In each target detection task, the participants silently counted occurrences of a rare class of target stimuli among a series of common “standard” stimuli. SOA was between 1000 and 1500 ms. EEG data was re-referenced offline to the algebraic sum of the left and right mastoids and divided into approximately 1 second, non-overlapping epochs extending from 100 ms before to 892 ms after stimulus onset. EEG preprocessing was identical to that used for Experiment 1, save for the fact that the data from two channels of one participant’s EEG were completely removed due to excessive artifacts. After epochs were rejected, the mean number of epochs per participant was 1641 (minimum: 1197).

Experiment 4

The participants in Experiment 4 were the same eight participants who participated in Experiment 3. However, the data from this experiment was recorded in a separate recording session and the task was different. Specifically, participants read sentences presented one word at a time (SOA of 500 ms) on a computer monitor and 50% of the sentences contained ungrammatical words. After the last word of each sentence, participants were asked to indicate if the sentences were grammatical or not and to sometimes answer questions about the meaning of the sentence by pressing buttons. EEG recording and preprocessing parameters were the same as those used in Experiment 3, save that for two participants the data from one channel were completely removed due to excessive artifacts and epochs of data were time locked to word and question onset. After rejecting epochs polluted by recording artifacts, the mean number of epochs per participant was 2185 epochs (minimum: 1765). Due to the short SOA between words, some epochs of data overlapped. On average, 80% of the time points of each participant’s data occurred in only one epoch (minimum: 79%). For full experimental details see Kemmer et al. (2004).

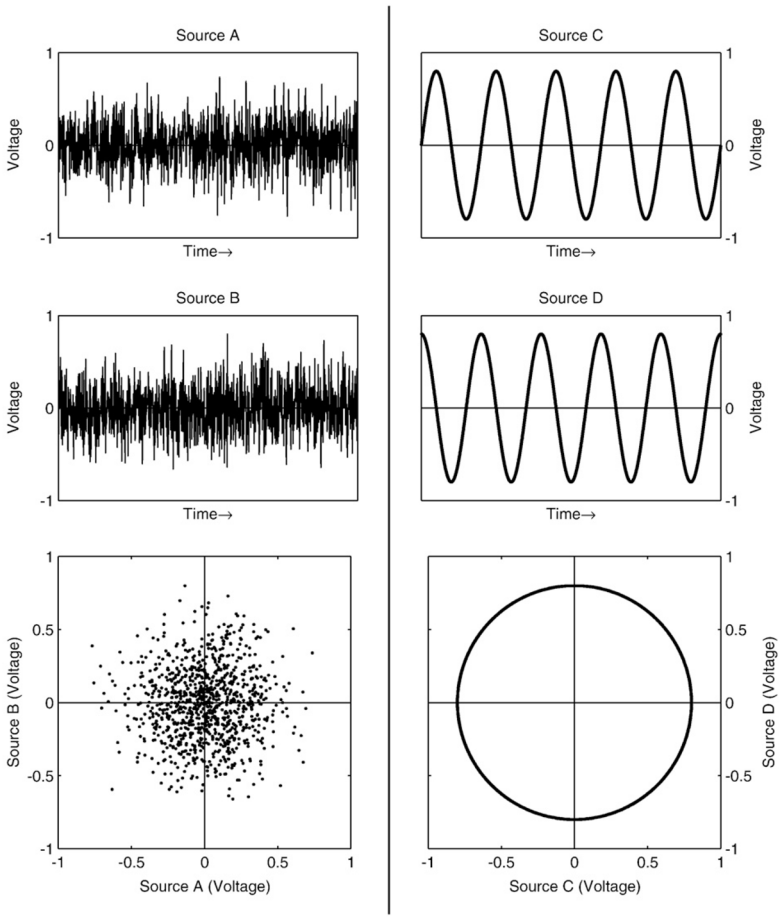

Fig. 1.

Example pairs of hypothetical scalp potential sources that ICA algorithms cannot reliably decompose: white Gaussian sources (left), sin(*t*) and cos(*t*) where *t* indexes time (right). An infinite number of equally independent decompositions are possible for such data (Bell and Sejnowski, 1995; Meinecke et al., 2002).

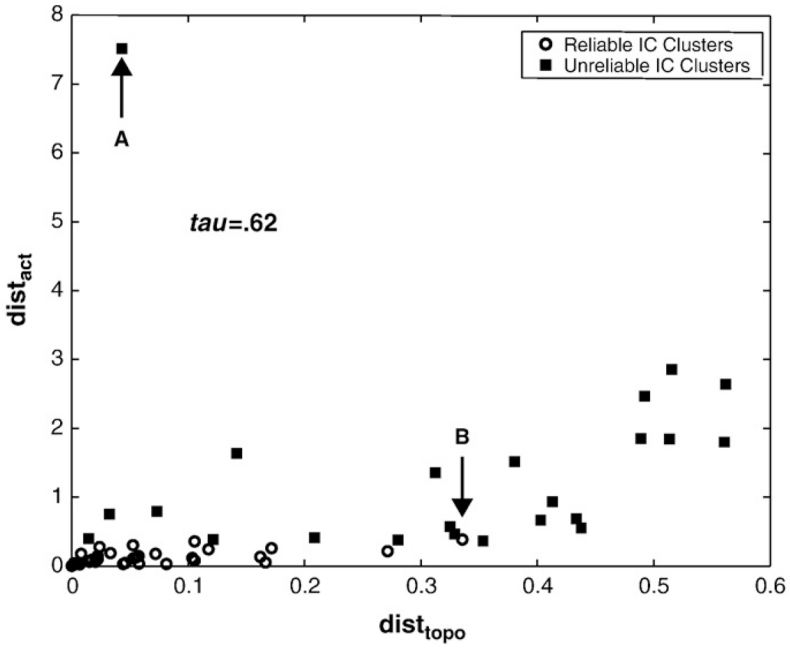

Fig. 2.

Scatterplot of median within-cluster IC scalp topography and activation distances. The 53 clusters were formed from ICs produced by 100 applications of FastICA to 100 bootstrap samples of a data set using the ICASSO algorithm (Himberg et al., 2004) to assess IC reliability. Clusters corresponding to reliable ICs (according to the ICASSO cluster quality index) are represented by circles. “A” and “B” indicate example clusters with one relatively reliable and one unreliable IC feature. The rank correlation between measures is quantified with Kendall’s *tau*.

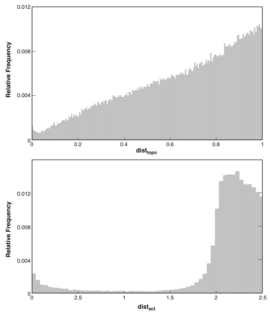

Fig. 3.

Histograms of topography (top) and activation (bottom) distances for all possible IC pairs. Activation distance distribution tapers off to the right towards infinity (not shown). Histograms are derived from 196,608 IC pairs (i.e., 16 participants, 3 ICA decompositions per participant, 64 ICs per decomposition).

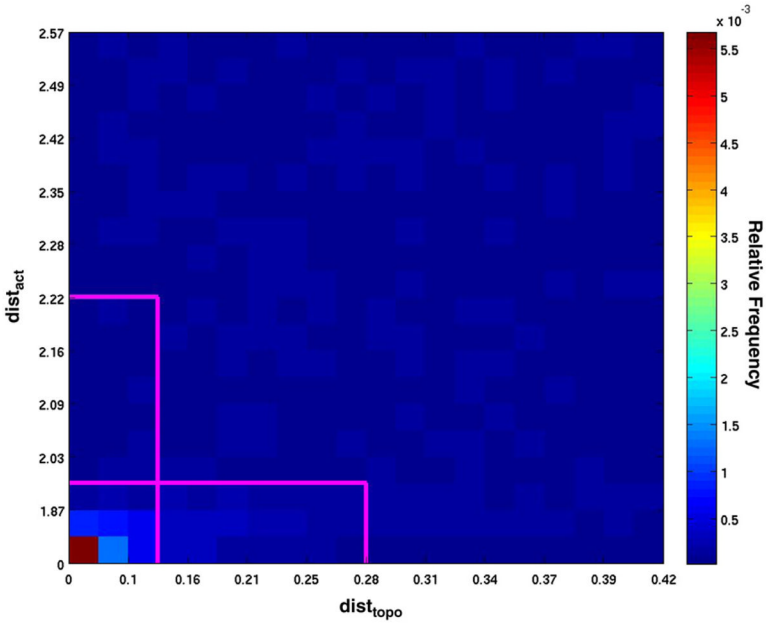

Fig. 4.

Joint distribution of IC pair topography and activation distances in Fig. 3. To make the two distance metrics comparable, they are binned in one percentile increments (e.g., 2% of pairs have topographies that are .1 away from one another or closer). The distribution continues to the right and up (not shown). Pink rectangles indicate “L” shaped critical region that contains 1/64 of all samples. The top right corner of the horizontal rectangle is at .28, 1.96. The top right corner of the vertical rectangle is at .13, 2.22. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

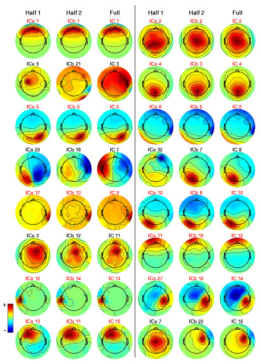

Fig. 5.

Topographies of the top 16 ICs from Participant 1's full data from Experiment 1 (third and sixth column) and their matches from both split-halves of the data. For visualization, unitless topography weights have been normalized so that each topography's maximal absolute weight is one and the 64 electrodes are not shown. Topography weights below the head's equator are plotted progressively beyond the radius of the head. Topography polarities are chosen to maximize similarity with the IC from the full data set. Triplets of reliable ICs are named in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

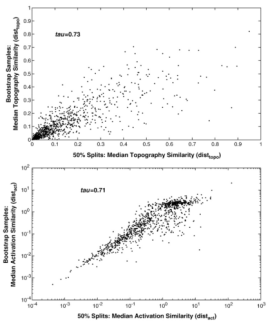

Fig. 6.

IC topography (top) and activation (bottom) reliability estimates using split-half comparisons and bootstrap samples are highly correlated. Rank correlations were quantified using Kendall's tau. Activation distances are plotted on log scales. 1024 samples per plot (e.g., 64 ICs per each of 16 participants).

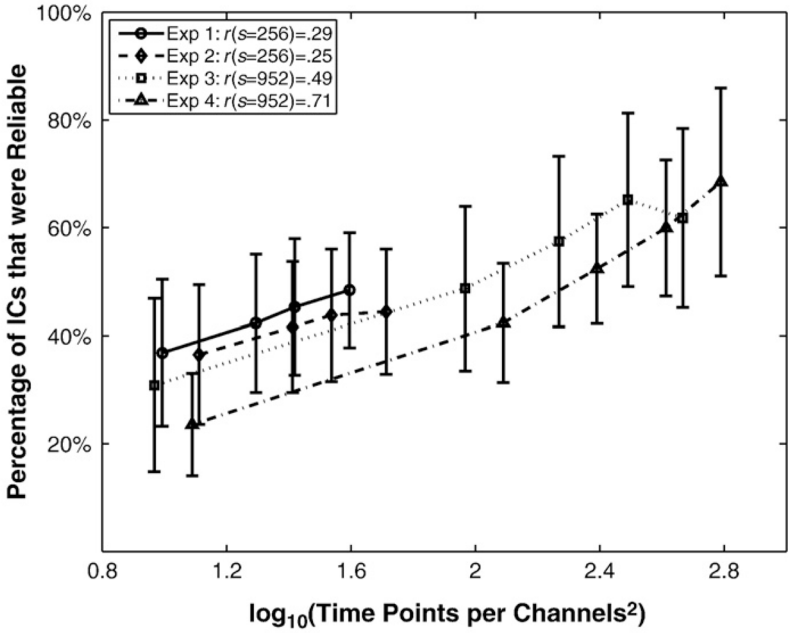

Fig. 7.

Mean percentage of reliable ICs (error bars are standard deviations) as a function of data set size. Modest to strong correlations (see legend) show that the percentage of reliable ICs decreases linearly with a decrease in the log of the number of data points/channel^2. *s* (see legend) indicates sample size.

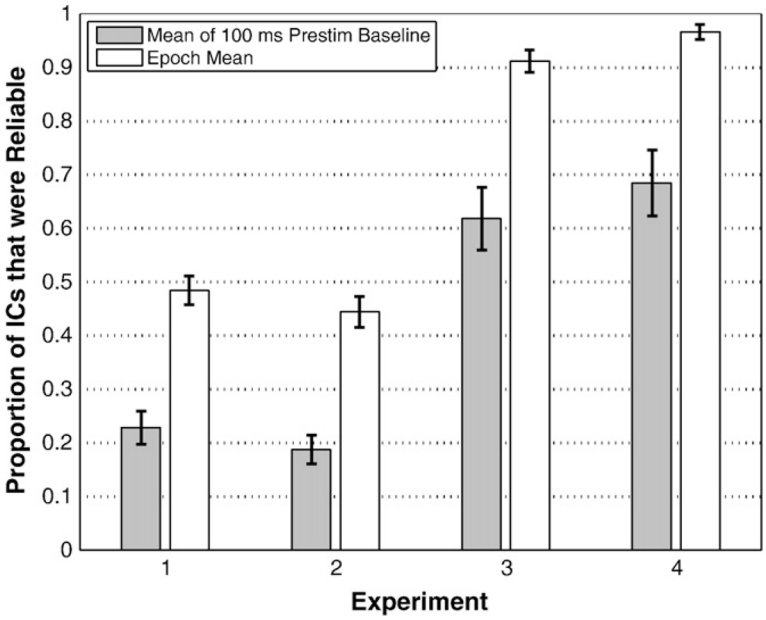

Fig. 8.

Mean proportion of ICs that are reliable per participant (error bars are standard error). Grey bars represent ICs from data that had the 100 ms prestimulus baseline removed from each epoch. White bars represent ICs from data that had the mean of each epoch removed. Experiments 1 and 2 consist of 64 channel data. Experiments 3 and 4 consist of approximately 30 channel data. Zero mean epoch data produce significantly more reliable ICs than 100 ms prestimulus baselined data (Experiments 1and 2: two-tailed sign test (16), k=16, p=3e−5; Experiments 3 and 4: two-tailed sign test(8), k=8, p=.008).