Unbiased Group-Level Statistical Assessment of Independent Component Maps by Means of Automated Retrospective Matching

Dave R.M. Langers^{1,2*}

¹Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, Groningen, The Netherlands

²Eaton-Peabody Laboratory, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts

Abstract: This report presents and validates a method for the group-level statistical assessment of independent component analysis (ICA) outcomes. The method is based on a matching of individual component maps to corresponding aggregate maps that are obtained from concatenated data. Group-level statistics are derived that include an explicit correction for selection bias. Outcomes were validated by means of calculations with artificial null data. Although statistical inferences were found to be incorrect if bias was neglected, the use of the proposed bias correction sufficed to obtain valid results. This was further confirmed by extensive calculations with artificial data that contained known effects of interest. While uncorrected statistical assessments systematically violated the imposed confidence level thresholds, the corrected method was never observed to exceed the allowed false positive rate. Yet, bias correction was found to result in a reduced sensitivity and a moderate decrease in discriminatory power. The method was also applied to analyze actual fMRI data. Various effects of interest that were detectable in the aggregate data were similarly revealed by the retrospective matching method. In particular, stimulus-related responses were extensive. Nevertheless, differences were observed regarding their spatial distribution. The presented findings indicate that the proposed method is suitable for neuroimaging analyses. Finally, a number of generalizations are discussed. It is concluded that the proposed method provides a framework that may supplement many of the currently available group ICA methods with validated unbiased group inferences. *Hum Brain Mapp* 31:727–742, 2010. © 2009 Wiley-Liss, Inc.

Key words: functional magnetic resonance imaging (fMRI); independent component analysis (ICA); group-level statistics; bias

Contract grant sponsor: Netherlands Organization for Scientific Research (NWO) and Netherlands Organization for Health Research and Development (ZonMw); Contract grant number: 016.096.011 (VENI grant).

*Correspondence to: Dave R.M. Langers, Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, Groningen, The Netherlands.
E-mail: d.r.m.langers@med.umcg.nl

Received for publication 26 May 2008; Revised 29 July 2009; Accepted 3 August 2009

DOI: 10.1002/hbm.20901

Published online 12 October 2009 in Wiley InterScience (www.interscience.wiley.com).

INTRODUCTION

In recent years, independent component analysis (ICA) has become increasingly popular as a blind source separation tool to analyze multivariate data that are acquired in functional magnetic resonance imaging (fMRI) experiments [Calhoun and Adali, 2006; Hyvärinen and Oja, 2000]. Like the perhaps more familiar principal component analysis (PCA), ICA is a method that decomposes fMRI data sets into components that are the (outer) product of a spatial magnitude map and a corresponding time course. While PCA will result in components that are completely uncorrelated, ICA employs higher-order statistical criteria to

maximize independence, either in the time domain or—more often—in the spatial domain [Calhoun et al., 2001b; Correa et al., 2007]. Conceptually, the maximization of spatial independence is closely related to the minimization of the overlap between the component maps [Langers, 2009]. The application of ICA to fMRI data has been demonstrated to result in components that faithfully represent various brain systems and that therefore are neurophysiologically interpretable [McKeown et al., 1998].

ICA has a key advantage over “conventional” fMRI techniques based on mass-univariate multilinear regression analysis: it does not require any prior specification of expected hemodynamic response functions to model the neural activation related to the stimuli and/or tasks that are employed in the experiment. This means that ICA is flexible enough to analyze experiments in which the characteristics of the responses are hard to predict, like for instance experiments involving complex tasks with multimodal stimuli from the natural environment [Bartels and Zeki, 2005; Kohler et al., 2008; Malinen et al., 2007; van de Ven et al., 2008], or resting state experiments in which no explicit stimuli or tasks are employed at all [Esposito et al., 2008; Jafri et al., 2008; Ma et al., 2007; van de Ven et al., 2004]. Although this property makes ICA an excellent tool for exploratory data analysis, it is much more difficult to employ ICA in a confirmatory manner to test prior hypotheses. This is related to the lack of a natural meaning or order of the various ICA components. These will typically include not only neural activation patterns, but for instance also various artifacts that arise in relation to MR-imaging and physiology [McKeown et al., 2003]. Because neural activation should in practice be largely independent from sources of artifacts, these signal contributions will mostly be separated into different components. Nevertheless, it requires retrospective interpretation to assign meaning to the results that are found. This is in contrast with conventional regression analyses where the interpretation is clear because regressors are designed a priori to reflect the hypothesis under investigation. While this problem can already play a role at the level of individual subjects, it becomes increasingly pressing when analyses at the population or group level are pursued. After all, it is uncertain whether in different subjects or subgroups components can be found with an identical or similar interpretation, and if so, it remains unclear how to properly identify them and incorporate them into a group analysis.

Various procedures have been proposed to overcome this difficulty [Calhoun et al., 2009]. Here, they will be presented according to whether the data of multiple subjects are combined before, during, or after the ICA step.

- **Averaging**

Data from multiple subjects can already be combined before the ICA decomposition takes place. A straightforward approach would be to employ direct averaging [Schmithorst and Holland, 2004]. This assumes that effects of interest are fully repeatable across sub-

jects with regard to their magnitude, their spatial distribution in the brain, as well as with regard to the corresponding dynamics in time. The outcomes of the analysis apply to the group as a whole; no information is available about differences between individual subjects.

- **Concatenation**

All subjects can simultaneously be entered into a single ICA by combining the data of all subjects into an aggregate data set. Subject data can be concatenated along the temporal dimension by linking the time courses [Calhoun et al., 2001c], along the spatial dimension by merging corresponding acquired maps [Svensén et al., 2002] or along a separate third dimension by stacking individual data matrices into a data tensor [Beckmann and Smith, 2005; Lee et al., 2008]. For all three classes of methods, subjects are allowed to differ in the overall magnitude of the components. However, they, respectively, constrain the effects of interest to be identical across subjects with regard to their spatial distribution, their temporal dynamics, or both.

- **Matching**

Finally, various group ICA methods combine the results of subjects only after ICA is carried out on individual subjects. These methods impose no constraints on the independent components, which are allowed to differ with respect to their magnitude, their spatial maps, as well as their temporal dynamics. However, additional steps need to be undertaken to retrospectively match up the components across subjects. Components may be selected by subjective identification [Calhoun et al., 2001a; Malinen et al., 2007; McKeown et al., 1998] or by (semi)automatic classification [Calhoun et al., 2004; De Martino et al., 2007; Esposito et al., 2005; Formisano et al., 2002; Moritz et al., 2003]. Automated matching has also been employed to assess the reproducibility of the outcomes of some nondeterministic ICA algorithms [Himberg et al., 2004; Yang et al., 2008].

Which of these methods is most suitable in a particular context will depend on the validity of the underlying assumptions. For instance, identical spatial distributions can well be assumed for homogeneous subject groups, but may be unacceptable in studies that include patients with brain abnormalities. Temporal dynamics may be the same for stimulus- or task-evoked responses that are related to a fixed paradigm, but these will vary for the spontaneous fluctuations that occur in resting state experiments. In some experiments, responses may be assumed identical within but different across subgroups, and mixed methods may be required [Guo and Pagnoni, 2008; Sui et al., in press].

The temporal concatenation method currently appears to be the most widely applied method. It offers maximum flexibility with regard to response dynamics in time and exploits the more or less fixed modular organization of the

brain in space [Gazzaniga, 1989]. Conceptually, this method parallels the use of fixed effects (FFX) general linear models in regression approaches. These types of methods are suitable to assess the magnitude of effects of interest in the particular dataset under consideration. However, they are unable to generalize these statements directly to the population or group level [Woods, 1996]. Moreover, since identical spatial distributions are assumed in the analysis already, differences between such distributions in subgroups cannot straightforwardly be statistically tested or compared.

These limitations can be circumvented by additional processing steps. Most notably, the extracted individual time courses can be entered into a regression model that is subsequently evaluated on the basis of the full subject data [Calhoun et al., 2001c]. The outcomes of this back-projection step can then be submitted to a group analysis, like in conventional random effects (RFX) regression analyses [Friston et al., 1995]. However, because the regressors are determined a posteriori, assessments are potentially biased. In particular, the outcomes of the back-projection can be predisposed to resemble the original aggregate independent components. This may decrease the observed variance across subjects and overestimate the significance of statistical outcomes. For group ICA methods that employ averaging, spatial concatenation, or tensor-ICA, back-projection is also necessary if inferences at the voxel level are required. Similar problems may then arise during statistical assessment.

Bias has recently received much attention in relation to neuroimaging analyses based on correlation, regression, and classification [Kriegeskorte et al., 2009; Vul et al., 2009]. So far, this discussion has not expanded into the field of ICA. In fact, it appears that performance under the null hypothesis has never been explicitly assessed for any group-level ICA method. This study aims to fill this gap by providing a validated framework for the purpose of making group-level statistical inferences.

A retrospective matching method is proposed to bring the extracted component maps of individual subjects into correspondence with aggregate component maps. Initially, no constraints are placed on the components at the subject-level ICA stage. It is shown that the magnitude of the selection bias that is introduced at the group level as a result of the retrospective matching can be predicted and corrected for. Calculations are performed to test the statistical validity of the method under the null hypothesis, to characterize its discriminatory power in the presence of effects of interest, and to evaluate its practical performance in an actual fMRI experiment.

THEORY

Retrospective Matching

After preprocessing (involving, e.g., slice timing correction, realignment, coregistration, spatial normalization,

smoothing, masking, and detrending), all fMRI data are collected into a single $L \times V$ matrix $\mathbf{Z}_k$ for each subject $k = 1, \dots, K$. The matrix $\mathbf{Z}_k$ contains the time courses of V voxels in its columns and L flattened volume acquisitions in its rows. By means of PCA, these data may be reduced to a desired number of C components, resulting in a decomposition

$$\mathbf{Z}_k \cong \mathbf{F}_k \mathbf{Y}_k. \quad (1)$$

$\mathbf{F}_k$ is an $L \times C$ inflation matrix, and $\mathbf{Y}_k$ is a $C \times V$ deflated data matrix. The decomposition is inexact since not all principal components are typically retained, but optimal in the sense that as much total signal power is preserved as possible. By employing spatial ICA to extract maximally independent components, $\mathbf{Y}_k$ may further be written as

$$\mathbf{Y}_k = \mathbf{A}_k \mathbf{S}_k. \quad (2)$$

$\mathbf{A}_k$ is an estimated $C \times C$ mixing matrix, and $\mathbf{S}_k$ is the corresponding $C \times V$ source matrix with spatial maps.

The matrices $\mathbf{Y}_k$ can be concatenated across subjects into a single $CK \times V$ aggregate data matrix $\mathbf{Y}$ and reduced once more using PCA.

$$\mathbf{Y} \equiv \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_K \end{bmatrix} \cong \mathbf{G} \mathbf{X}. \quad (3)$$

$\mathbf{G}$ is a $CK \times C$ inflation matrix, and $\mathbf{X}$ is a $C \times V$ deflated aggregate data matrix. Aggregate independent components can now be estimated by means of another ICA decomposition.

$$\mathbf{X} = \mathbf{A} \mathbf{S}. \quad (4)$$

At this point, a set of C aggregate independent component maps are available for the group in $\mathbf{S}$, while at the same time C individual independent component maps are available for each subject in $\mathbf{S}_k$. However, these components are not yet known to correspond in any way.

The proposed method proceeds by determining symmetric $C \times C$ product matrices $\mathbf{R}_k = \mathbf{S}_k \mathbf{S}^T$ for all subjects. The entries of $\mathbf{R}_k$ contain the inner products between all pairs of aggregate and individual component maps. Their absolute values form a measure of the similarity between the various maps. The goal is to determine a bijective matching that maximizes the similarities between the paired maps. This may be achieved by maximizing the sum of the absolute values of those entries in $\mathbf{R}_k$ that correspond with matched pairs. This is a classical combinatorial optimization problem that is known as the “(linear) assignment problem.” It can in principle be solved by

Figure 1.

A diagrammatic overview of the retrospective matching method. Individual subject data are reduced by means of principal component analysis (PCA), after which independent component analysis (ICA) is employed to extract meaningful component maps [Eqs. (1) and (2)]. In parallel, subject data are concatenated into an aggregate group data set, which is further reduced with PCA and decomposed with ICA [Eqs. (3) and (4)]. The aggregate component maps are used as templates to guide the matching procedure (although, in principle, templates may also be derived by other means). For each subject, a matrix R_k is constructed

that contains the inner products of the templates with the individual component maps. On the basis of this matrix, individual subject maps are matched to the aggregate maps. All subject maps that are matched to the same aggregate map are subsequently entered into a group-level analysis, resulting in t -value statistical parametric maps $SPM(t)$. The selection bias that results from the retrospective matching is explicitly taken into account on the basis of the number of components C , the aggregate maps S , and the covariance structure Σ of the component maps [Eqs. (10) and (11)].

means of exhaustive search, but computationally more efficient methods are available. For instance, the Hungarian method, also known as the Kuhn–Munkres algorithm, finds a solution in polynomial time [Kuhn, 1955].

Finally, voxelwise statistics regarding the effects of interest can be derived by entering all individual maps that were retrospectively matched to the same aggregate component into a group level analysis. t -value statistical parametric maps $SPM(t)$ can be computed and thresholded to detect significant effects of interest. The entire retrospective matching method is summarized in diagrammatic form in Figure 1.

Selection Bias

Because matching component maps are all selected to show maximal resemblance with the same aggregate map, the expectation value of their mean across all subjects will

be nonzero. The magnitude of the resulting selection bias will be assessed next. It will henceforth be assumed that, under the null hypothesis, the individual independent component maps in all S_k constitute independent and identically distributed (IID) samples from a single multivariate normal distribution $N(0, \Sigma)$. Also, the possible restriction that a single individual component cannot be matched to multiple aggregate components will be ignored: it will be assumed that an aggregate component is matched to the best matching individual component, independently of all other aggregate components.

The method selects the best resembling aggregate template map and individual component map by searching for the pair with the largest absolute inner product (see Fig. 2). For a particular template map s , the expectation value of the best matching individual map m is defined in terms of its (unknown) probability distribution function. This can also be formulated in terms of conditional expectation values with respect to the inner product ms^T .

Figure 2.

A graphical illustration of the selection bias. Under the null hypothesis, the flattened component maps are assumed to comprise samples from a zero-mean multivariate normal distribution. In this 2-D representation, the corresponding points form an ellipsoidal data cloud. Those samples that have a negative inner product with some provided template map $\mathbf{s}$ are flipped with respect to their sign ($\circ$). The sample $\mathbf{m}$ with the largest inner product (circled) will be selected as the best match to the template. A closed form expression for the expectation value of $\mathbf{m}$ can be derived in terms of the template $\mathbf{s}$ and the distribution's covariance structure Σ [see Eq. (10)]. Even though the original multivariate normal distribution centers around zero, the bias of the proposed retrospective matching method can thus be estimated to be nonzero. Similar concerns may hold for other statistical group ICA methods.

$$\begin{aligned} E(\mathbf{m}) &= \int_{\mathbf{m}} [\mathbf{m} \cdot P(\mathbf{m})] d\mathbf{m} \\ &= \int_{\mathbf{m}} [E(\mathbf{m} | \mathbf{m}\mathbf{s}^T = m) P(\mathbf{m}\mathbf{s}^T = m)] d\mathbf{m}. \end{aligned} \quad (5)$$

(Note: $P(\mathbf{m})$ refers to the distribution of the sample with the maximum absolute inner product, not to the known multivariate normal distribution of all original samples!)

Because the maximum sample $\mathbf{m}$ is selected purely on the basis of its inner product with $\mathbf{s}$, its expectation value under a particular fixed conditional constraint $\mathbf{m}\mathbf{s}^T = m$ is precisely equal to the expectation value of the original multivariate normal distribution under that constraint. The conditional expectation value for a normally distributed variable $\mathbf{x}$ satisfies

$$E(\mathbf{x} | \mathbf{x}\mathbf{s}^T = x) = x \frac{\mathbf{s}\Sigma}{\mathbf{s}\Sigma\mathbf{s}^T} \quad (6)$$

[Johnson and Wichern, 2001]. This results in

$$E(\mathbf{m}) = \int_{\mathbf{m}} \left[m \frac{\mathbf{s}\Sigma}{\mathbf{s}\Sigma\mathbf{s}^T} P(\mathbf{m}\mathbf{s}^T = m) \right] d\mathbf{m} = \frac{\mathbf{s}\Sigma}{\mathbf{s}\Sigma\mathbf{s}^T} E(\mathbf{m}\mathbf{s}^T). \quad (7)$$

The scalar quantity $\mathbf{m}\mathbf{s}^T$ is the absolute maximum out of C inner products. Since $\mathbf{s}$ is fixed and all individual maps are assumed to be drawn from a multivariate normal distribution, the inner products themselves form samples from a univariate normal distribution. This distribution's mean is zero and its variance $\sigma^2 = \mathbf{s}\Sigma\mathbf{s}^T$.

By partial integration, it can be shown that the maximum of the absolute values of C independent samples from a univariate zero-mean unit-variance normal distribution has an expectation value M_C equal to

$$M_C = \int_0^\infty \left[1 - \text{erfc}\left(\frac{z}{\sqrt{2}}\right) \right] dz. \quad (8)$$

Although the integral in Eq. (8) cannot be simplified further, M_C can easily be obtained to arbitrary accuracy by means of numerical integration. It only weakly depends on C ; for example, $M_1 = \sqrt{2/\pi} = 0.80$, $M_2 = 2/\sqrt{\pi} = 1.13$, $M_{10} = 1.88$, $M_{20} = 2.17$, $M_{100} = 2.75$. In the context of fMRI, M_C should usually lie in a limited range near 2.0–2.5.

It now follows that

$$E(\mathbf{m}\mathbf{s}^T) = M_C \sqrt{\mathbf{s}\Sigma\mathbf{s}^T}. \quad (9)$$

From Eq. (7), we arrive at the following expression, which provides an exact prediction of the bias that occurs under the null hypothesis when selecting the best matching component map.

$$E(\mathbf{m}) = M_C \frac{\mathbf{s}\Sigma}{\sqrt{\mathbf{s}\Sigma\mathbf{s}^T}}. \quad (10)$$

Equation (10) relies on complete and exact knowledge about the covariance structure of the component maps. Typically, the number of subjects and extracted components is too small and the number of voxels too large to extract a sufficiently accurate estimate for the entire covariance matrix Σ . However, the noise can often be modeled by a D -dimensional Gaussian random field. Here, the observed correlated noise will be assumed equivalent to the convolution of an underlying uncorrelated noise signal with some fixed Gaussian kernel, $\mathbf{k}$, with known full width at half maximum (FWHM). By explicitly writing out the kernel, it can be shown that $\mathbf{s}\Sigma \cong R \cdot \mathbf{s}'\Sigma'$ [see Eq. (A9) the Appendix]. The scalar R provides a measure for the number of voxels that are effectively correlated, analogous to the volume of a resel [Worsley et al., 1992]. If long-range correlations are present, $R \cong (1.505 \cdot \text{FWHM})^D$ [see Eq. (A8) for the exact form]; if spatial correlations are

completely absent, $R = 1$. The vector $\mathbf{s}'$ contains a doubly smoothed template map; symbolically, $\mathbf{s}' = \mathbf{k} \otimes \mathbf{k} \otimes \mathbf{s}$. The matrix Σ' is derived from Σ ; it is strictly diagonal and obeys $\Sigma'_{ii} = \Sigma_{ii}$. In other words, it describes the variance in the voxels, but disregards any covariance due to spatial correlations.

This approximation may now be used to obtain an expression for the bias that does not require estimation of any covariance terms. Equation (10) straightforwardly leads to

$$E(\mathbf{m}) = M_c \sqrt{R} \frac{\mathbf{s}' \Sigma'}{\sqrt{\mathbf{s}' \Sigma' \mathbf{s}'^T}}. \quad (11)$$

Once the bias is calculated from Eq. (11), it may be subtracted from the observed group mean and thus be accounted for in the calculation of the SPM(t).

MATERIALS AND METHODS

Null Data

To experimentally verify the magnitude of the bias under the null hypothesis, calculations with artificial null data were performed in the MatLab programming environment (The MathWorks). In each of 1,000 iterations that were carried out, a group was created that comprised $K = 50$ subjects. Per subject, $C = 20$ components were generated. The component maps each consisted of a 40×40 2-D data matrix with normally distributed zero-mean unit-variance pseudo-random noise. Apart from this default noise distribution, two modified noise distributions were derived (see Fig. 3a). In the first, the noise was modulated by multiplication with a windowing function that varied from 1.0 in the center of the image to 0.0 at the edges. In the second, the noise was smoothed by convolution with a Gaussian kernel (FWHM = 4.0 pixels) while preserving its variance.

To allow the results to be compared across the various iterations, the generated maps were not matched to variable aggregate outcomes, but to an externally imposed fixed template. Two different templates were employed in two sets of calculations (see Fig. 3b). A first template image contained a 4×4 raster of squares, alternatingly with uniform signals levels $+1.0$ or -1.0 , on a background with uniform signal level 0.0 . A second template image contained the "Lena" standard test picture [Munson, 1996] that was converted to black and white, adjusted in brightness and contrast, and resized to 40×40 pixels. Retrospective matching was carried out on the basis of the inner products of the null component maps with the fixed templates, as described in the Theory section. Per subject, the component was selected that was most similar to the template, flipping signs if necessary. Uncorrected SPM(t) were calculated based on the mean and standard error of the matched maps across all subjects. Also, bias was assessed according

to Eq. (11), and corrected SPM(t) were calculated to test the deviation of the components from this prediction.

In Silico Data

To characterize the discriminatory power of the method, additional calculations were carried out on artificial data with known signal components, mimicking fMRI data. A total number of 100 iterations was performed with the following default settings. Groups of $K = 8$ subjects were created. Each subject comprised $L = 100$ acquisition volumes that consisted of $32 \times 32 \times 16$ 3-D data matrices with normally distributed zero-mean unit-variance smoothed pseudo-random noise (FWHM = 2.0 voxels). Four 8×8 square regions of interest (ROIs) were defined that spanned all 16 slices, designated O, I, II, and III, as illustrated in Figure 4a. In ROI O, no signal was added to the noise. In the ROIs I–III, different offsets were added to each acquisition volume separately; the offsets were sampled from zero-mean normal distributions with standard deviations 0.1, 0.2, and 0.3, respectively. As a result, three effects of interest with increasing magnitude were present in these data: their spatial maps coincided with three of the defined ROIs, and their time courses consisted of white noise.

For each subject, PCA was used to reduce the dimensionality of the data to $C = 8$ components. Independent components were extracted by means of the *InfoMax* algorithm [Bell and Sejnowski, 1995]. In addition, the reduced subject data were concatenated, again reduced to eight components, and unmixed into independent components. The individual component maps were matched retrospectively to the aggregate component maps and grouped into sets of corresponding components. Uncorrected SPM(t) were calculated on the basis of the mean and standard error of the sets of matched maps. Also, bias was assessed [Eq. (11)] and corresponding corrected SPM(t) were computed. Voxels were considered significant if in any of the components their t -value exceeded a two-tailed threshold at a confidence level $P = 0.05$. Bonferroni correction for multiple comparisons was performed to account for the number of tested components C ; the multitude of voxels V was not corrected for.

The specificity (SPE) in the uncorrected and corrected SPM(t) was determined by counting the fraction of voxels that did not reach significance in ROI O. The sensitivity (SEN) in relation to the three effects of interest was determined by counting the fraction of voxels that was considered significant in the ROIs I–III. Receiver operating characteristic (ROC) curves were derived by varying the threshold.

Apart from the aforementioned default configuration, additional calculations were performed with modified parameters. The number of subjects and the number of extracted components were varied between 2 and 32. The size of the ROIs was varied from 4×4 to 12×12 . The matrix dimensions were decreased to $16 \times 16 \times 8$ voxels,

Figure 3.

Calculations for artificial null data. (a) Sets of images were generated that contained pseudo-random noise (left), optionally including windowing (middle), or smoothing (right). (b) In separate analyses, the images were matched to two different templates, one containing a raster of squares with alternating intensity and the other containing a real-life picture. The noise image with the highest absolute inner product was selected. (c) Results were averaged over 1,000 such calculations, and the average outcome is shown. The mean image reflects many of the characteristics of the template, but is also affected by the magnitude and smoothness of the noise. (d) The expectation value of the mean image was predicted according to Eq. (11).

No deviations from the image that was actually obtained are visible. (e) To highlight any remaining differences between the actual and the predicted mean image, one was plotted versus the other. The intensities in all voxels strongly cluster around the diagonal, indicating that the prediction is accurate both qualitatively and quantitatively. (f) Finally, t-value statistics were derived across multiple subjects, both with and without a correction for the selection bias. When bias was accounted for, the resulting cumulative density function (cdf) conformed to the theoretical curve (dashed). However, when bias was neglected, deviations were observed for all noise configurations; for smoothed noise, results diverged most widely.

Figure 4.

Calculations for artificial data with true effects. (a) 3-D artificial data were generated that contained pseudo-random smoothed noise. Four rectangular regions were defined (O–III). Pseudo-random offsets were added to model true effects of interest. The root-mean-square power of these signals increased linearly from region I to III (the magnitudes of these signals are exaggerated tenfold in this illustration); no signal was added to region O. (b) The sensitivity (SEN) to the signals in regions I–III is plotted as a function of the specificity (SPE) in region O, both for corrected and uncorrected thresholding. The resulting receiver operating characteristic (ROC) curves show that the discriminatory power increased with increasing signal to noise level across the regions I–III, but also reveal that bias correction leads to a moderate decrease in discriminatory power.

as well as increased to $64 \times 64 \times 32$ voxels (scaling the size of the ROIs accordingly). The number of acquisitions per subject ranged from 25 to 400. The smoothness of the noise varied from zero to $\text{FWHM} = 4.0$ voxels. Finally, results were determined for thresholds between $P = 0.001$ and $P = 0.1$. For each setting, 100 simulations were performed.

In Vivo Data

To assess the performance of the method in a practical setting, the results from an fMRI study on 32 subjects were analyzed. Written informed consent was obtained in approved accordance with the requirements of the institutional committees on the use of human subjects at the Massachusetts Eye and Ear Infirmary, the Massachusetts General Hospital, and the Massachusetts Institute of Technology.

Subjects were placed in supine position in the bore of a 3.0-T MR system (Siemens Allegra), which was equipped with a standard 12-channel transmit/receive head coil. An oblique coronal functional imaging volume was positioned on the basis of high-resolution anatomical images. The functional imaging runs consisted of a dynamic series of T_2^* -sensitive echo planar imaging (EPI) sequences ($\text{TR} \sim 8$ s; $\text{TE} 30$ ms; flip angle 90° ; matrix $64 \times 64 \times 10$; voxel dimensions $3.1 \times 3.1 \times 8.0 \text{ mm}^3$, including a 2.0-mm slice gap). Each run comprised 36 EPI acquisitions, excluding initial start-up scans that were discarded before the analysis. Per subject, between 5 and 11 runs were carried out. A sparse clustered volume scanning paradigm was employed to reduce the influence of acoustic scanner noise [Edmister et al., 1999; Hall et al., 1999; Langers et al., 2005]. The scanner coolant pump and fan were turned off during functional imaging to reduce ambient noise levels. Cardiac triggering on the basis of a pulse-oximeter signal was used to reduce the influence of cardiac motion [Guimaraes et al., 1998]. In one subject, the cardiac signal was too weak to detect reliably, and a fixed TR of 7.5 s was set.

During the functional imaging sessions, a basic audiovisual task was performed by the subjects to control their attention levels. Subjects viewed a projection screen through a mirror. On a uniform gray background, a black cross was displayed that briefly turned red at pseudo-random intervals. Subjects were instructed to press a button device each time they observed such a color change. Responses were monitored to ensure that subjects understood the task and stayed alert, and the achieved hit rate approximately equaled 80–90%. Behavioral results were not analyzed further for the purpose of this study. At the same time, MR-compatible piezoelectric speakers that were mounted in insert ear plugs were used to deliver auditory stimuli. These stimuli consisted of four 32-s broadband noise bursts per run, separated by periods of silence. The stimulus intensities of all bursts were equal within a run,

but differed across runs in a range of 50–80 dB SPL. Subjects were told not to respond to the auditory stimuli.

Preprocessing was performed using various routines from the SPM8 software package (<http://www.fil.ion.ucl.ac.uk/spm/>) [Friston et al., 1995]. The functional image volumes were corrected for motion effects using 3-D rigid body transformations. The anatomical images were coregistered to the functional volumes, and all images were normalized into MNI stereotaxic space using affine transformations. Images were resampled at a 3-mm isotropic resolution and cropped in the anterior-to-posterior direction to a subvolume that comprised the acquired functional data ($-45 \leq y \leq +45$). To improve signal-to-noise ratio characteristics, spatial smoothing was performed on the functional volumes using an isotropic 5-mm Gaussian kernel. Various signal confounds were estimated by means of a linear regression approach. The model included (i) baseline and drift effects (four vectors modeling a third-order polynomial for each separate functional run), (ii) residual motion effects (six vectors containing the estimated translation and rotation parameters in the x -, y -, and z -direction), and (iii) T_1 -relaxation effects (two vectors containing the TRs as well as the squares of their deviations from 8 s). The model was evaluated voxelwise, and the fitted confounds were subtracted from the original data.

Spatial ICA was carried out using the Group ICA of fMRI Toolbox (GIFT) v.1.3e (<http://icatb.sourceforge.net/groupica.htm>) [Calhoun et al., 2001c]. Because baseline effects were already subtracted during preprocessing, subtraction of the global mean prior to PCA and ICA was disabled in GIFT. Individual analyses were set up for all subjects, as well as an aggregate analysis for the group of all 32 subjects as a whole. The processing in GIFT comprised initialization, PCA reduction to 30 components, and ICA extraction by means of the *InfoMax* algorithm [Bell and Sejnowski, 1995; Correa et al., 2007; Esposito et al., 2002]. The extracted independent component maps of all individual subjects were retrospectively matched to those from the group analysis, as described in the theory section. Uncorrected $SPM(t)$ were calculated on the basis of the mean and standard error of the matched maps across all subjects. In addition, the bias as estimated from Eq. (11) was taken into account to derive corrected $SPM(t)$.

RESULTS

Null Data

Null components comprising nothing but random noise were generated, and these were matched to two different templates. The panels in Figure 3 show the outcomes for spatially uncorrelated homogeneous noise (left), uncorrelated but inhomogeneous noise (middle), and spatially correlated homogeneous noise (right). Figure 3a displays representative samples from the generated noise.

In Figure 3b, the two templates are shown (which were identical across all three noise configurations). For each subject, out of 20 generated noise components, the compo-

nent that best resembled the template was selected, flipping signs if necessary. The mean across all subjects was calculated. Figure 3c displays the average results that were obtained over 1,000 iterations. Figure 3d displays the corresponding predictions according to Eq. (11). The images in Figure 3c,d are highly similar, both qualitatively and quantitatively. In particular, the prediction was able to capture the details due to the appearance of the template, as well as those due to variations in the magnitude and smoothness of the noise. Differences appeared to be limited to some remaining noise, mostly invisible in these printed figures. The obtained mean matched components were subsequently plotted as a function of the predicted mean components in a scatter plot, as shown in Figure 3e. (In other words, Fig. 3e plots the intensities in Fig. 3c as a function of those in Fig. 3d; every pixel contributes one point.) Data strongly cluster along the diagonal, indicating that the prediction is accurate. The remaining spread can be attributed to the stochastic character of the generated artificial data.

In each iteration, an $SPM(t)$ was calculated. Both uncorrected and corrected group-level statistics were computed (i.e., deviations from zero or from the predicted bias, respectively). The obtained statistics were collected across all voxels and all iterations, and the cumulative probability distribution (cdf) was constructed. Results are plotted in Figure 3f, together with the theoretical cdf of Student's t . For all three noise configurations, the corrected cdf coincided with the standard curve, but the uncorrected cdf deviated visibly. Although deviations were limited for the two uncorrelated noise configurations, differences were large for the spatially correlated data. When thresholding at $t = \pm 2.01$ (corresponding with a two-tailed significance test at confidence level $P = 0.05$), the null hypothesis was falsely rejected in a fraction of 5.0% of all tests on the basis of the corrected statistics. However, for the uncorrected statistics, effects were considered significant in 6.7% of all tests for both uncorrelated noise configurations, while for the spatially correlated noise configuration no less than 34.2% of all voxels were false positives.

In Silico Data

Figure 4a displays a typical artificial data volume with four ROIs, three of which contain nonzero signals of interest, on a background of smoothed pseudo-random noise. Calculations were carried out for various configurations. Table I contains the SPE and SEN as derived from the positive rates in the four ROIs. With regard to the parameters that were varied, a number of observations could be made. Overall, the SEN correlated positively with increasing numbers of subject, decreasing numbers of extracted components, increasing matrix size, increasing region size, increasing numbers of acquisitions, decreasing smoothness, and more lenient thresholds. With the uncorrected thresholding method, deviant behavior was found for the weakest component in ROI I: the SEN was highest for an

TABLE I. Simulation results for various configuration parameters

Parameters	Uncorrected				Corrected			
	SPE ₀	SEN _I	SEN _{II}	SEN _{III}	SPE ₀	SEN _I	SEN _{II}	SEN _{III}
Subjects (K)								
2	95.1	5.0	5.5	7.1	95.2	4.8	5.1	6.6
4	94.9	5.5	10.1	38.4	95.6	4.5	7.2	30.1
8	94.2	7.2	38.1	99.2	96.0	4.6	21.2	97.2
16	92.2	13.1	89.1	100.0	95.9	5.2	66.0	100.0
32	89.0	30.4	99.9	100.0	96.1	6.8	98.0	100.0
Components (C)								
2	95.0	5.1	52.4	100.0	95.5	4.5	37.0	99.8
4	94.3	6.0	50.8	99.8	95.6	4.2	33.0	99.2
8	94.2	7.2	38.1	99.2	96.0	4.6	21.2	97.2
16	94.0	6.5	21.3	95.5	96.0	4.1	10.1	88.2
32	94.5	5.7	10.5	80.4	96.1	3.7	5.2	64.7
Matrix size (V)								
16 × 16 × 8	83.3	20.1	34.4	81.2	95.4	5.6	8.7	44.1
24 × 24 × 12	92.9	8.3	28.9	92.8	96.0	4.1	10.8	78.1
32 × 32 × 16	94.2	7.2	38.1	99.2	96.0	4.6	21.2	97.2
48 × 48 × 24	94.6	6.4	69.2	99.8	95.5	4.9	58.6	99.5
64 × 64 × 32	94.8	6.8	84.2	99.9	95.4	5.6	79.2	99.8
Region size								
4 × 4	93.5	6.2	8.5	30.9	95.8	4.1	4.7	13.9
6 × 6	93.8	6.6	19.4	93.4	95.7	4.3	8.9	82.2
8 × 8	94.2	7.2	38.1	99.2	96.0	4.6	21.2	97.2
10 × 10	94.2	7.7	51.9	99.5	95.9	4.7	34.6	98.4
12 × 12	94.1	8.1	55.0	99.3	95.7	5.0	39.4	97.9
Acquisitions (L)								
25	94.2	5.9	8.7	22.2	95.9	4.0	4.9	10.8
50	94.2	6.2	13.9	61.5	95.9	4.2	7.0	41.5
100	94.2	7.2	38.1	99.2	96.0	4.6	21.2	97.2
200	94.0	9.2	92.4	100.0	95.8	5.2	83.3	100.0
400	93.5	20.8	100.0	100.0	95.6	11.1	100.0	100.0
Smoothness (FWHM)								
0.0	94.8	6.2	80.5	99.8	95.4	5.4	77.8	99.8
1.0	94.9	6.5	81.8	99.9	95.6	5.3	76.2	99.8
2.0	94.2	7.2	38.1	99.2	96.0	4.6	21.2	97.2
3.0	91.9	10.2	29.7	87.7	96.4	4.2	10.0	65.1
4.0	84.5	18.8	32.1	76.9	95.6	4.9	8.0	36.8
Threshold (p)								
0.001	99.9	0.2	2.6	52.9	99.9	0.1	0.9	37.3
0.010	98.7	1.6	12.9	90.9	99.2	0.9	5.4	80.8
0.050	94.2	7.2	38.1	99.2	96.0	4.6	21.2	97.2
0.100	88.3	13.1	53.8	99.8	91.6	8.8	34.0	99.1

The average specificity (SPE) in ROI 0 and sensitivities (SEN) in ROIs I–III are listed as percentages (%). Hundred numerical simulations were performed with various settings. Results were assessed by means of statistical testing that was either uncorrected or corrected for bias. Parameter settings that were varied included the following: the number of subjects, the number of extracted components, the matrix size, the size of the regions with true effects, the number of acquisitions per subject, the smoothness of the noise as expressed by the full width at half maximum (FWHM) of the Gaussian convolution kernel, and the globally imposed threshold. Default parameter values are shown in bold.

intermediate number of extracted components, for small matrices, and for large smoothness. Using corrected statistics, the latter trends largely disappeared. The uncorrected thresholding method always resulted in higher SEN than the corrected method. However, importantly, the SPE was found to systematically lie below the targeted level for uncorrected thresholding, while it always exceeded that

level for corrected thresholding (the target SPE equals 95% for $P = 0.05$, and $\text{SPE} = 1 - P$ in general).

ROC curves that correspond with the various ROIs are compared between uncorrected and corrected thresholding in Figure 4b. These results were derived from the calculations with default parameter settings. Of course, signal detection improved with increasing signal-to-noise ratio.

Figure 5.

Results for data from an actual fMRI experiment with an audiovisual stimulation paradigm. (a) Component maps from an aggregate ICA with temporal concatenation were scaled to unit root-mean-square signal power. Ten out of thirty components are displayed as aggregate SPM(z) by means of a color code on an anatomical background. Six of these components (1–6) were interpreted to result from neural activation, while four other components (7–10) were deemed MRI-artifacts. (b) Retrospective matching was carried out, and group statistics were determined that did not include any correction for bias. The

corresponding component maps are shown as uncorrected SPM(t). Although some differences are obvious, the interpretation of the components relates well to those from the aggregate analysis. (c) Similar statistics were derived that included an explicit correction for the presence of bias, resulting in corrected SPM(t). Activation was less significant than without bias correction. Still, many of the components of interest could be identified. In particular, auditory activation in response to the employed sound stimuli was consistently found.

For the three signals of interest in the ROIs I, II, and III, the areas under the curves for the uncorrected method equaled 0.5308, 0.8210, and 0.9953, respectively, while they were 0.5148, 0.7561, and 0.9935 for the corrected method. This corresponds with binormal d-prime equal to 0.11, 1.30, and 3.68 (uncorrected thresholding), and 0.05, 0.98, and 3.52 (corrected thresholding). These findings show that correction for bias comes at the cost of a moderate reduction in discriminatory power.

In Vivo Data

The fMRI data were analyzed by means of ICA, both with individual and aggregate data. The individual maps

were retrospectively matched to the aggregate maps. Uncorrected and corrected SPM(t) were derived. Following visual inspection and interpretation of the component maps, 16 components were deemed to be of neural origin. The remaining 14 components were construed to arise from MRI-artifacts. Figure 5 illustrates the spatial maps belonging to a subset of ten components.

Figure 5a shows coronal slices from the aggregate group maps that were converted to SPM(z) by division by the map's root-mean-square signal level. Results were thresholded at $z = \pm 3.3$, corresponding to a two-tailed confidence level $P = 0.001$ (uncorrected for multiple comparisons). Activation was well delineated and confined to specific brain structures. Six neural components

contained strong signals in (1) the superior temporal lobe (auditory processing), (2) the medial temporal lobe (limbic processing), (3) the pericentral gyri (motor/somatosensory processing), (4) the basal ganglia (learning), (5) the left frontal lobe (cognitive processing), and (6) the insula and anterior cingulate gyrus (paralimbic processing); four artifactual components comprised (7) the lower brainstem (pulsatile motion), (8) the ventricles (T_1 -relaxation), (9) the anterior midbrain (homogeneous field distortion), and (10) the extreme superior regions of the frontal lobe and inferior temporal pole (foldover).

In Figure 5b, the corresponding uncorrected $SPM(t)$ are thresholded at $t = \pm 3.6$ (equivalent with $P = 0.001$). The $SPM(t)$ resembled the $SPM(z)$, in the sense that regions with strong signal in the $SPM(z)$ typically also showed up in the $SPM(t)$. This was particularly true for the components that represented neural activation. Nevertheless, some differences were observed. In the $SPM(t)$, the significance level that was reached in local maxima was typically lower, and the maps had a noisier appearance. Also, the relative strength of the activated regions within single maps varied. For instance, in the auditory system (component number 1), activation in the auditory cortex was stronger in the $SPM(z)$, but subcortical activity of the auditory nuclei in the midbrain and thalamus was better detectable in the $SPM(t)$. Similarly, in the limbic system (component number 2), the medio-temporal limbic lobe was strongest in the $SPM(z)$, but the $SPM(t)$ proved more sensitive to associated regions in the thalamus and cingulate cortex. In the artifactual components, deviations between the results of the $SPM(z)$ and $SPM(t)$ were even more noticeable. For the lower brainstem and ventricles (components 7 and 8), effects were much more extensive in the $SPM(t)$. In contrast, for the anterior midbrain and the extreme superior and inferior brain regions (components 9 and 10), effects were less significant in the $SPM(t)$. In component 9, many voxels were revealed with significant opposite effects.

Figure 5c shows the corresponding corrected $SPM(t)$. Not surprisingly, the corrected maps typically revealed fewer significant voxels than the uncorrected ones. For some components, differences were limited (e.g., component 1). Yet, for other components, considerably fewer voxels were detectable after bias correction (e.g., component 4), suggesting that these components were inconsistently found across subjects. Nevertheless, the distribution of the revealed activation patterns was again comparable in the sense that the observed activation maxima were mostly found in the same locations.

DISCUSSION

A retrospective matching method was proposed that employs aggregate ICA outcomes to guide the selection of corresponding components at the subject level. Parametric group-level statistics were derived, taking the presence of

bias into account. The validity of the method was empirically assessed by applying it to artificial null data. Bias correction proved to be necessary to obtain valid results, but it was also shown that the proposed bias correction suffices. The performance of the method was characterized using artificial data with known effects of interest. Again, the corrected thresholding method led to valid outcomes for all configurations that were tried. However, the sensitivity and discriminatory power of the method was found to be lower than without bias correction. Finally, the method was applied to a real fMRI dataset. The outcomes proved that the method is sensitive to neuroimaging effects. In particular, in one of the independent components, the auditory system was found to be strongly activated by sound stimulation. Some other components were more difficult to detect confidently.

Template Extraction

Retrospective matching methods have been used before. Some of these methods match individual subject maps to activation patterns that are expected a priori [Bartels and Zeki, 2005; De Martino et al., 2007; Jafri et al., 2008; van de Ven et al., 2004]. This knowledge may either be explicit (in the form of template maps) or implicit (in the form of general selection criteria to guide subjective component selection). However, such information may not always be available or may be insufficiently precise. Some other methods group similar individual components together blindly. This has for instance been achieved by means of clustering [Esposito et al., 2005; Himberg et al., 2004] or by selecting the most representative individual component maps as templates [Yang et al., 2008]. However, individual component maps can have poor signal-to-noise characteristics and are susceptible to random fluctuations. This may detrimentally affect such blind matching processes. The present method attempts to avoid all of these shortcomings by deriving a set of templates from all available data in an aggregate ICA.

As described in the Theory section, temporal concatenation is used by default to extract aggregate component maps that may serve as templates. This assumes that the spatial distribution of the activation is similar across subjects, which suits most neuroimaging experiments well.

Temporal concatenation in ICA is conceptually similar to the use of FFX models in linear regression approaches. Both methods have in common that they do not allow direct statistical inferences to be made about the significance of observed effects of interest at the group level, like in an RFX analysis. To overcome this limitation, back-projection techniques have been suggested [Calhoun et al., 2001c]. These reconstruct individual component maps from the aggregate outcomes. However, as argued in the Introduction, this approach may be susceptible to bias.

Although in the current method concatenated data are used as well, some important differences between retrospective matching and back-projection exist. Firstly, in

retrospective matching one and the same set of template maps is used in every subject. In contrast, in the temporal concatenation method with back-projection, individual time courses that are already tuned to correspond with the identified aggregate components are fed back into the back-projection step. Secondly, the aggregate maps are only used to select components in the proposed method, but they do not directly affect the calculation of the subject components themselves. Although this may still lead to a selection bias, this effect is predictable and can be accounted for.

The potential presence of bias not only applies to temporal concatenation with back-projection, it also likely plays an important role for many other group ICA methods (the author is not aware of any studies that have investigated this). Conveniently, the retrospective matching framework that is described in this article is not restricted to temporal concatenation per se. It can be generalized to other ways of extracting templates. For instance, when using a fixed stimulus or task paradigm, the use of spatial concatenation, tensor ICA, or averaging to derive the template maps may be preferred. Because no assumptions were made about the templates, the retrospective matching, bias correction, and statistical assessment can subsequently be carried out in identical fashion. Other group ICA methods that do not rely on aggregate data may also be used to provide the necessary set of templates. For instance, prior knowledge can be incorporated by using a fixed set of independently obtained template maps (e.g., from previous studies); the results of clustering methods can be used by entering the cluster centroids into the analysis as templates; or, a set of most representative individual subject maps can be selected. This provides a lot of flexibility and makes the described framework broadly applicable.

Bias assessment and Validity

The proposed matching procedure was carried out on the basis of a matrix $\mathbf{R}_k$ that contained the inner products between all pairs of individual and aggregate maps. Inner products are closely related to correlation coefficients, which have been applied in the context of independent component matching before [Esposito et al., 2005; Yang et al., 2008]. ICA maps are often normalized to zero mean and unit variance, and this would actually render both measures completely equivalent.

Inner products were employed because they are linear operations. Under the null hypothesis, they give rise to parametric statistics. This simplified the description sufficiently such that the magnitude of the bias could be analytically expressed in terms of the number of extracted components, the composition of the template map, and the (co)variance structure of the component maps [Eq. (10)].

Some approximations were necessary to make it possible to estimate these required quantities from the data, especially in the presence of spatial correlations. Nevertheless,

the resulting expression [Eq. (11)] proved accurate. The calculations with artificial null data were set up to maximize the effects of bias by including many subjects and by averaging over many iterations. Even in these extreme conditions, the method was able to predict the resulting bias with high accuracy. While uncorrected statistics were obviously invalid (especially for the configuration with smoothed noise), the outcomes that were corrected for bias conformed to the expected t -statistics without any detectable deviation.

Matching Algorithms

In the described method, subject components were matched to a fixed set of aggregate templates—and thus to each other—by maximizing a joint measure of the similarity of all pairs of subject and group component maps simultaneously. In this case, the sum of the absolute values of the entries in $\mathbf{R}_k$ was maximized. One could consider maximizing the sum of some otherwise transformed similarity measure. For instance, the entries in $\mathbf{R}_k$ may be transformed according to a power law or logarithmic function. Such transformations can be used to reduce the influence of potential outliers. Alternatively, the actual inner product values may be replaced by their rank order to obtain a nonparametric measure. However, these alternatives were not pursued further in this initial study.

The described method employs the Hungarian method to determine a global maximum for the similarity [Kuhn, 1955]. Simpler matching algorithms can be considered. For instance, maps can be paired up one by one, by locating the single entry with the maximum absolute value in $\mathbf{R}_k$. The matched components can subsequently be removed from further consideration by setting all their entries in $\mathbf{R}_k$ equal to zero. By repeating this assignment procedure until all maps are matched up, a bijective matching can be obtained that cannot be improved upon without deteriorating the match for at least some of the components. This algorithm can be modified to match up only a designated subset of templates, as long as the number of aggregate templates does not exceed the number of subject components. (The same holds for the Hungarian method.)

If the matching is not required to be bijective, one could simply match up each group component with the subject component that shows the best resemblance. A subject component may then be matched to more than one group component, while other subject components may not be paired to any group component at all. Still, for each group component, precisely one corresponding individual component would be selected in each subject. This approach can match any number of group components to any number of subject components (though all subjects still need to have identical numbers of components C for the bias to be correctly predicted).

Although results have not been reported for all of these suggested alternative methods, in some provisional analyses they performed satisfactorily as well.

Performance and Power

In none of the calculations with artificial data any excess false positive rates were observed for the corrected method. The SPE was always equal to the required level of 95%, or slightly more. In contrast, uncorrected statistics reached this level for almost none of the calculations that were carried out. This proves that bias effects are generally significant, but can be avoided by the proper corrections.

In the Theory section, bias was predicted for the best match out of C components. In other words, it was assumed that the matching of a particular component is not limited by the presence of other components. If the matching is required to be bijective, this assumption is violated, because two group components are forbidden to be matched to the same subject component. This effectively restricts the number of available components below C and may lead to an overestimation of the bias in Eq. (11). The magnitude of this effect is expected to be limited: the number of components only enters the predicted bias through M_C , and M_C was observed to depend only weakly on C . Moreover, following ICA in the spatial domain, the various aggregate component maps are mutually minimally correlated. As a result, if a subject component well resembles a particular group component, it seems less likely that it resembles another group component at the same time. Finally, as far as this effect does play a role, it should only increase type-II errors (false negatives). These are generally considered much less severe than type-I errors (false positives). Still, this can explain the marginally conservative behavior of the method that was observed in Table I (although this may also partly be attributed to the employed Bonferroni correction).

The SEN was always found to be lower for the corrected method as compared to the uncorrected method. To some degree, this is attributable to the more stringent thresholding that is necessary to reach the required SPE. Yet, bias correction was also found to lead to a moderate reduction in discriminatory power on the basis of the ROC curves. This is likely related to the fact that the bias is strongest in regions with strong effects. A correction for bias will therefore affect the detected regions with strong activation more heavily than the undetected regions with weak or no effects. As a result, it seems inevitable that discriminatory power will decrease when correcting for such bias.

The SEN was found to depend on a number of parameters. Obviously, effects became easier to detect with increasing amounts of evidence from available data (as determined by the total numbers of subjects, the number of acquisitions per subject, the regional volume per acquisition, the number of voxels per region, and the number of voxels per resel). Furthermore, due to the correction for multiple comparisons, signals were less detectable as the number of extracted components increased beyond the true number of effects of interest. At the same time, reducing the number of extracted components below the actual number of effects made the weakest component more difficult to detect.

When applied to real data, the retrospective matching method was able to extract more or less the same components as the aggregate analysis. After correction for bias, various neural components of interest survived. Still, other neural components were weak and resulted in few significant voxels. Similarly, some artifactual components were consistently observed in all subjects (e.g., components 7 and 8), while other components revealed few significant effects (e.g., components 9 and 10).

Such differences are to be expected, since the aggregate SPM(z) have a completely different interpretation than the group-level SPM(t). The SPM(z) represent a voxel statistic and identify effects that are strong in comparison with those in the entire set of voxels. They do not allow any conclusions to be drawn at the group level. The SPM(t), however, determine the significance of effects across all subjects and do allow group-level assessments to be made. If certain effects are very strong in one subject, but absent in other subjects, they may affect the aggregate analysis and show up in the SPM(z), but they should normally not lead to significance at the group level in the SPM(t). Contrariwise, effects that are relatively weak but highly consistent across subjects may show up in the SPM(t) but not in the SPM(z).

It seems likely that some components are difficult to detect at the group level, because they occur insufficiently reliably across subjects. Other components should be easily detectable if they show consistent activity. Indeed, extensive activation was detectable in the entire auditory system, which was of primary interest in relation to the stimulus paradigm.

In this study, statistical maps were determined that reveal whether the null hypothesis can be rejected that no effects of interest are present in a voxel. Another relevant question might be whether effects are different across multiple subgroups, for instance in patients versus matched controls. Although this has not been investigated in this study, it is interesting to point out that such analyses are easy to carry out by means of retrospectively matched component maps. If it can be assumed that the covariance structure Σ of the component maps is identical for the two groups, then the magnitude of the matching bias according to Eqs. (10) or (11) will also be identical in both groups. When testing whether the difference between group means is significantly different from zero, the effects of bias would exactly cancel. This makes any corrections for bias superfluous, simplifying the current method tremendously.

ACKNOWLEDGMENTS

I would like to express my gratitude for the fruitful cooperation with the Melcher Research Group at the Eaton-Peabody Laboratory in Boston (USA) and particularly acknowledge the generous contributions of Jennifer Melcher and Gianwen Gu who provided the fMRI data.

REFERENCES

- Bartels A, Zeki S (2005): Brain dynamics during natural viewing conditions—A new guide for mapping connectivity in vivo. *Neuroimage* 24:339–349.
- Beckmann C, Smith S (2005): Tensorial extensions of independent component analysis for multisubject fMRI analysis. *NeuroImage* 25:294–311.
- Bell AJ, Sejnowski TJ (1995): An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* 7:1129–1159.
- Calhoun VD, Adali T, McGinty VB, Pekar JJ, Watson TD, Pearlson GD (2001a): fMRI activation in a visual-perception task: Network of areas detected using the general linear model and independent components analysis. *Neuroimage* 14:1080–1088.
- Calhoun VD, Adali T, Pearlson GD, Pekar JJ (2001b): Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Hum Brain Mapp* 13:43–53.
- Calhoun VD, Adali T, Pearlson GD, Pekar JJ (2001c): A method for making group inferences from functional MRI data using independent component analysis. *Hum Brain Mapp* 14:140–151.
- Calhoun VD, Adali T, Pekar JJ (2004): A method for comparing group fMRI data using independent component analysis: Application to visual, motor and visuomotor tasks. *Magn Reson Imaging* 22:1181–1191.
- Calhoun VD, Adali T (2006): Unmixing fMRI with independent component analysis. *IEEE Eng Med Biol Mag* 25:79–90.
- Calhoun VD, 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:S163–S172.
- Correa N, Adali T, Calhoun VD (2007): Performance of blind source separation algorithms for fMRI analysis using a group ICA method. *Magn Reson Imaging* 25:684–694.
- De Martino F, Gentile F, Esposito F, Balsi M, Di Salle F, Goebel R, Formisano E (2007): Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. *Neuroimage* 34:177–194.
- Edmister WB, Talavage TM, Ledden PJ, Weisskoff RM (1999): Improved auditory cortex imaging using clustered volume acquisitions. *Hum Brain Mapp* 7:89–97.
- Esposito F, Aragri A, Pesaresi I, Cirillo S, Tedeschi G, Marciano E, Goebel R, Di Salle F (2008): Independent component model of the default-mode brain function: Combining individual-level and population-level analyses in resting-state fMRI. *Magn Reson Imaging* 26:905–913.
- Esposito F, Formisano E, Seifritz E, Goebel R, Morrone R, Tedeschi G, Di Salle F (2002): Spatial independent component analysis of functional MRI time-series: To what extent do results depend on the algorithm used? *Hum Brain Mapp* 16:146–157.
- Esposito F, Scarabino T, Hyvarinen A, Himberg J, Formisano E, Comani S, Tedeschi G, Goebel R, Seifritz E, Di Salle F (2005): Independent component analysis of fMRI group studies by self-organizing clustering. *Neuroimage* 25:193–205.
- Formisano E, Esposito F, Kriegeskorte N, Tedeschi G, Di Salle F, Goebel R (2002): Spatial independent component analysis of functional magnetic resonance imaging time-series: Characterization of the cortical components. *Neurocomputing* 49:241–254.
- Friston KJ, Holmes AP, Poline JB, Grasby PJ, Williams SC, Frackowiak RS, Turner R (1995): Analysis of fMRI time-series revisited. *Neuroimage* 2:45–53.
- Gazzaniga MS (1989): Organization of the human brain. *Science* 245:947–952.
- Guimaraes AR, Melcher JR, Talavage TM, Baker JR, Ledden P, Rosen BR, Kiang NY, Fullerton BC, Weisskoff RM (1998): Imaging subcortical auditory activity in humans. *Hum Brain Mapp* 6:33–41.
- Guo Y, Pagnoni G (2008): A unified framework for group independent component analysis for multi-subject fMRI data. *Neuroimage* 42:1078–1093.
- Hall DA, Haggard MP, Akeroyd MA, Palmer AR, Summerfield AQ, Elliott MR, Gurney EM, Bowtell RW (1999): “Sparse” temporal sampling in auditory fMRI. *Hum Brain Mapp* 7:213–223.
- Himberg J, Hyvärinen A, Esposito F (2004): Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage* 22:1214–1222.
- Hyvärinen A, Oja E (2000): Independent component analysis: Algorithms and applications. *Neural Netw* 13:411–430.
- Jafri MJ, Pearlson GD, Stevens M, Calhoun VD (2008): A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *Neuroimage* 39:1666–1681.
- Johnson RA, Wichern DW (2001): *Applied Multivariate Statistical Analysis*, 5th ed. Englewood Cliffs, NJ: Prentice Hall.
- Kohler C, Keck I, Gruber P, Lie CH, Specht K, Tome AM, Lang EW (2008): Spatiotemporal group ICA applied to fMRI datasets. *Conf Proc IEEE Eng Med Biol Soc* 2008:4652–4655.
- Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009): Circular analysis in systems neuroscience: The dangers of double dipping. *Nat. Neurosci* 12:535–540.
- Kuhn HW (1955): The Hungarian method for the assignment problem. *Naval Res Logistics Q* 2:83–97.
- Langers DRM (2009): Blind source separation of fMRI data by means of factor analytic transformations. *Neuroimage* 47:77–87.
- Langers DRM, Van Dijk P, Backes WH (2005): Interactions between hemodynamic responses to scanner acoustic noise and auditory stimuli in functional magnetic resonance imaging. *Magn Reson Med* 53:49–60.
- Lee J, Lee T, Jolesz FA, Yoo S (2008): Independent vector analysis (IVA): Multivariate approach for fMRI group study. *Neuroimage* 40:86–109.
- Ma L, Wang B, Chen X, Xiong J (2007): Detecting functional connectivity in the resting brain: A comparison between ICA and CCA. *Magn Reson Imaging* 25:47–56.
- Malinen S, Hlushchuk Y, Hari R (2007): Towards natural stimulation in fMRI—Issues of data analysis. *Neuroimage* 35:131–139.
- McKeown MJ, Makeig S, Brown GG, Jung TP, Kindermann SS, Bell AJ, Sejnowski TJ (1998): Analysis of fMRI data by blind separation into independent spatial components. *Hum Brain Mapp* 6:160–188.
- McKeown MJ, Hansen LK, Sejnowski TJ (2003): Independent component analysis of functional MRI: What is signal and what is noise? *Curr Opin Neurobiol* 13:620–629.
- Moritz CH, Rogers BP, Meyerand ME (2003): Power spectrum ranked independent component analysis of a periodic fMRI complex motor paradigm. *Hum Brain Mapp* 18:111–122.
- Munson DC (1996): A note on Lena. *IEEE Trans Image Process* 5:3.
- Schmithorst VJ, Holland SK (2004): Comparison of three methods for generating group statistical inferences from independent

- component analysis of functional magnetic resonance imaging data. *J Magn Reson Imaging* 19:365–368.
- Sui J, Adali T, Pearlson GD, Clark VP, Calhoun VD: A method for accurate group difference detection by constraining the mixing coefficients in an ICA framework. *Hum Brain Mapp* (in press).
- Svensén M, Kruggel F, Benali H (2002): ICA of fMRI group study data. *Neuroimage* 16:551–563.
- van de Ven V, Bledowski C, Prvulovic D, Goebel R, Formisano E, Di Salle F, Linden DEJ, Esposito F (2008): Visual target modulation of functional connectivity networks revealed by self-organizing group ICA. *Hum Brain Mapp* 29: 1450–1461.
- van de Ven VG, Formisano E, Prvulovic D, Roeder CH, Linden DEJ (2004): Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Hum Brain Mapp* 22:165–178.
- Vul E, Harris C, Winkielman P, Pashler H (2009): Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect Psychol Sci* 4:274–290.
- Woods RP (1996): Modeling for intergroup comparisons of imaging data. *Neuroimage* 4:S84–S94.
- Worsley KJ, Evans AC, Marrett S, Neelin P (1992): A three-dimensional statistical analysis for CBF activation studies in human brain. *J Cereb Blood Flow Metab* 12:900–918.
- Yang Z, LaConte S, Weng X, Hu X (2008): Ranking and averaging independent component analysis by reproducibility (RAICAR). *Hum Brain Mapp* 29:711–725.

APPENDIX

In this Appendix, an approximation will be derived for the case when all covariance due to spatial correlations can effectively be attributed to smoothing with a fixed kernel.

It will be assumed that the covariance structure Σ of the component maps $\mathbf{x}$, arranged as row vectors, equals that which would be obtained if an underlying spatially uncorrelated stochastic noise signal $\mathbf{n}$ were spatially convolved with some fixed kernel $\mathbf{k}$ according to $\mathbf{x} = \mathbf{k} \otimes \mathbf{n}$. This is equivalent to

$$x_{\vec{q}} = \sum_{\vec{r}} (k_{\vec{q}-\vec{r}} n_{\vec{r}}). \quad (\text{A1})$$

(The indices with vector arrows symbolically refer to all relevant spatial coordinates simultaneously, and summation is carried out over the entire voxel space.) Making use of the presumed absence of spatial correlations in the underlying noise, it is found that

$$\Sigma_{\vec{p}\vec{q}} = \sum_{\vec{r}} (k_{\vec{p}-\vec{r}} k_{\vec{q}-\vec{r}} \sigma_{\vec{r}}^2), \quad (\text{A2})$$

where σ describes the local root-mean-square power of the underlying uncorrelated noise. It will be assumed that this varies only slowly as a function of location. More pre-

cisely, it will be assumed that it effectively does not vary within the scope of the kernel. As a result, the above equation may be simplified to

$$\Sigma_{\vec{p}\vec{q}} = \sigma_{\vec{q}}^2 \sum_{\vec{r}} (k_{\vec{p}-\vec{r}} k_{\vec{q}-\vec{r}}). \quad (\text{A3})$$

Now, for a symmetrical convolution kernel, the product $\mathbf{s}\Sigma$ involving some arbitrary template map $\mathbf{s}$ will satisfy

$$\begin{aligned} (\mathbf{s}\Sigma)_{\vec{q}} &= \sum_{\vec{p}} \left(s_{\vec{p}} \Sigma_{\vec{p}\vec{q}} \right) = \sigma_{\vec{q}}^2 \sum_{\vec{p}, \vec{r}} (s_{\vec{p}} k_{\vec{p}-\vec{r}} k_{\vec{q}-\vec{r}}) \\ &= \sigma_{\vec{q}}^2 \sum_{\vec{p}, \vec{r}} (k_{\vec{q}-\vec{r}} k_{\vec{r}-\vec{p}} s_{\vec{p}}) = \sigma_{\vec{q}}^2 (\mathbf{k} \otimes \mathbf{k} \otimes \mathbf{s})_{\vec{q}}. \end{aligned} \quad (\text{A4})$$

Although the unsmoothed σ are unknown, the variance of the smoothed signal is contained in the diagonal of Σ , and this can be determined from the data. Equation (A3) then allows σ to be determined from

$$\sigma_{\vec{q}}^2 = \frac{\Sigma_{\vec{q}\vec{q}}}{\sum_{\vec{r}} k_{\vec{r}}^2}. \quad (\text{A5})$$

For an isotropic Gaussian kernel with known full width at half maximum (FWHM)

$$k_{\vec{r}} \propto e^{-\frac{4\ln(2)|\vec{r}|^2}{\text{FWHM}^2}}, \quad (\text{A6})$$

Without loss of generality, the convention will be followed that the kernel is scaled such that it preserves low-frequency fluctuations and offsets.

$$\sum_{\vec{r}} k_{\vec{r}} = 1. \quad (\text{A7})$$

If the kernel is sufficiently smooth (i.e., if the FWHM is sufficiently large), then the summations in the numerator of Eq. (A5) as well as in Eq. (A7) can be replaced by integrals. In D dimensions, it is found that

$$R \equiv \frac{1}{\sum_{\vec{r}} k_{\vec{r}}^2} = \left(\text{FWHM} \sqrt{\frac{\pi}{\ln(4)}} \right)^D. \quad (\text{A8})$$

The interpretation of R is similar to the volume of a resolution element (resel) in Gaussian field theory.

Combining the previous equations, the following approximation is obtained:

$$\mathbf{s}\Sigma = R(\mathbf{k} \otimes \mathbf{k} \otimes \mathbf{s})\Sigma'. \quad (\text{A9})$$

Here, Σ' is a diagonal matrix that is derived from Σ by setting all off-diagonal elements equal to zero.