NeuroImage **15,** 407–421 (2002)
doi:10.1006/nimg.2001.0986, available online at http://www.idealibrary.com on IDEAL®

Spatiotemporal Independent Component Analysis of Event-Related fMRI Data Using Skewed Probability Density Functions

J. V. Stone,[1] J. Porrill, N. R. Porter, and I. D. Wilkinson*

*Psychology Department and *Academic Department of Radiology, Sheffield University, Sheffield S10 2TP, England*

Received May 29, 2001

We introduce two independent component analysis (ICA) methods, spatiotemporal ICA (stICA) and skew-ICA, and demonstrate the utility of these methods in analyzing synthetic and event-related fMRI data. First, stICA simultaneously maximizes statistical independence over both time and space. This contrasts with conventional ICA methods, which maximize independence either over time only or over space only; these methods often yield physically improbable solutions. Second, skew-ICA is based on the assumption that images have skewed probability density functions (pdfs), an assumption consistent with spatially localized regions of activity. In contrast, conventional ICA is based on the physiologically unrealistic assumption that images have symmetric pdfs. We combine stICA and skew-ICA, to form skew-stICA, and use it to analyze synthetic data and data from an event-related, left–right visual hemifield fMRI experiment. Results obtained with skew-stICA are superior to those of principal component analysis, spatial ICA (sICA), temporal ICA, stICA, and skew-sICA. We argue that skew-stICA works because it is based on physically realistic assumptions and that the potential of ICA can only be realized if such prior knowledge is incorporated into ICA methods.

INTRODUCTION

Independent component analysis (ICA) (Jutten and Herault, 1988; Bell and Sejnowski, 1995) is becoming increasingly popular as a tool for analyzing biomedical data (McKeown *et al.*, 1998; Makeig *et al.*, 1997; Porrill *et al.*, 2000; Biswal and Ulmer, 1999; McKeown, 2000). However, any method will provide poor results if the assumptions on which it is based are violated. As demonstrated in Calhoun *et al.* (2001), ICA works well on fMRI data if experimental protocols are specifically designed to meet certain statistical independence criteria, but fails if these criteria are not met. However, the extent to which a method fails depends on the robustness of the method with respect to violations of its underlying assumptions. Indeed, it is widely accepted that, strictly speaking, many data sets violate the assumptions of conventional second order statistical methods, such as SPM (SPM99, 1999). Despite this, the robustness of such methods with respect to relatively minor violations ensures that useful statistical results can still be obtained.

Here, we argue that conventional ICA can be made robust by incorporating physically realistic assumptions into the method. As with second-order statistical methods, the assumptions on which the method is based are usually violated to some extent, but such violations are typically minor if the assumptions are essentially physically correct. Conventional ICA embodies the assumption that data can be decomposed into underlying sources that are independent over space (McKeown *et al.*, 1998) or time (Makeig *et al.*, 1997; Calhoun *et al.*, 2001) and that the probability density functions (pdf) of these sources are highly kurtotic and symmetric (McKeown *et al.*, 1998; Calhoun *et al.*, 2001). Neither of these assumptions are warranted for most fMRI data sets and are motivated mainly by historical precedent and mathematical convenience. In contrast, our method is based on the assumptions that the underlying sources of fMRI data are not *strictly* independent over time nor space and that the underlying spatial sources of fMRI data have skewed probability density functions (Suzuki *et al.*, 2001).

We begin with a nontechnical description of ICA methods and review the characteristics of each method. This is followed by mathematical description of the analysis methods used in this paper (e.g., principal component analysis and the general linear model). MatLab code for the ICA methods described here can be downloaded from http://www.shef.ac.uk/~pc1jvs/.

THEORETICAL METHODS: SPATIAL AND TEMPORAL ICA

ICA can be used in two complementary ways to decompose an image sequence into a set of images and a corresponding set of time-varying image amplitudes.

[1] http://www.shef.ac.uk/~pc1jvs.

1053-8119/02 $35.00

TABLE 1

Magnitude of Correlation between the Left/Right Time Course Extracted by Each Method and the Predicted Activation Sequence (PAS) for the Left and Right Hemisphere

Method	Left hemisphere	Right hemisphere
GLM	0.58	0.68
PCA	0.42	0.40
sICA	0.43	0.24
tICA	0.48	0.40
stICA	0.51	0.25
Skew-sICA	0.49	0.44
Skew-stICA	0.52	0.46

Note. See text for details.

Spatial ICA (sICA) seeks a set of mutually independent component (IC) source images and a corresponding (dual) set of unconstrained time courses (e.g., McKeown *et al.*, 1998). In contrast, *temporal* ICA (tICA) seeks a set of IC source time courses and a corresponding (dual) set of unconstrained images (e.g., Makeig *et al.*, 1997).

Critically, the unconstrained nature of these dual signals permits them to adopt physically improbable forms in order that their corresponding ICs are made statistically independent. That is, ICA can appear to find statistically independent source signals in a data set by making use of the extra degrees of freedom implicit in the unconstrained dual signals, *even if the data do not consist of statistically independent source signals.* ICA can therefore find maximally independent signals which are not the underlying sources. Thus, the independence of extracted (spatial or temporal) signals can be achieved at the cost of physically improbable forms for their unconstrained (temporal or spatial, respectively) dual signals. A case in point may be exemplified in McKeown (2000), where it is emphasized that, while the images extracted by sICA are approximately independent, their corresponding dual time courses can be highly correlated. The correlations between these time courses may be much larger than that between the underlying temporal sources precisely because the images found by sICA are forced to be as independent as possible.

In contrast, spatiotemporal ICA is based on the physically realistic assumption that there exist small dependences between different spatial source signals and between different temporal source signals. For example, if two different brain regions responded independently to a stimulus, but with slightly different latencies, then their activations would be statistically correlated. Any method, such as tICA, which clustered brain regions according to their temporal correlations would incorrectly associate both regions with a single "hybrid" time course which did not reflect the activations of either region accurately.

Spatiotemporal ICA

Spatiotemporal ICA (stICA) places the ICs and their dual signals on an equal footing. The critical difference between ICA (e.g., sICA, tICA) and stICA is as follows. ICA can achieve independence over space (sICA) *or* time (tICA) by "sacrificing" the dual temporal (sICA) or dual spatial (tICA) signals, respectively. In contrast, stICA maximizes the degree of independence over space *and* time, without necessarily producing independence in either space *or* time. That is, stICA permits a tradeoff between the mutual independence of images and the mutual independence of their corresponding time courses.

Skew-ICA

The nature of the signal extracted by ICA is specified in terms of a user-defined model pdf. For example, the standard model pdf used with ICA extracts signals with high kurtosis pdfs, a characteristic common to many natural source signals, such as speech. However, when using spatial ICA with fMRI data, there is no *a priori* reason to suppose that the pdfs of source images have high kurtosis. Indeed, it may be argued that source images are likely to consist of spatially localized features surrounded by an homogeneous background. Such source images have skewed pdfs (Suzuki *et al.*, 2001). Accordingly, we introduce skew-ICA, which utilizes a skewed pdf model.

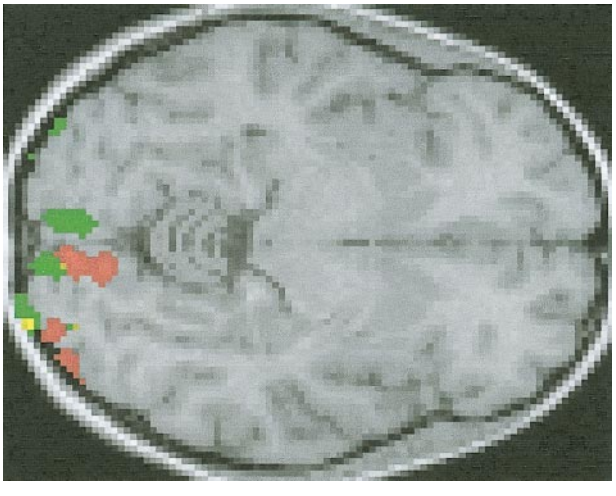

FIG. 1. Areas identified by the GLM. Activation overlaid on a structural scan of the subject. Activation induced by left and right hemifield activation are shown in green and red, respectively, and overlapping activity is shown in yellow. This image was derived from a thresholded spatial image; only pixels in the top 5% of the absolute distribution of the image values are shown.

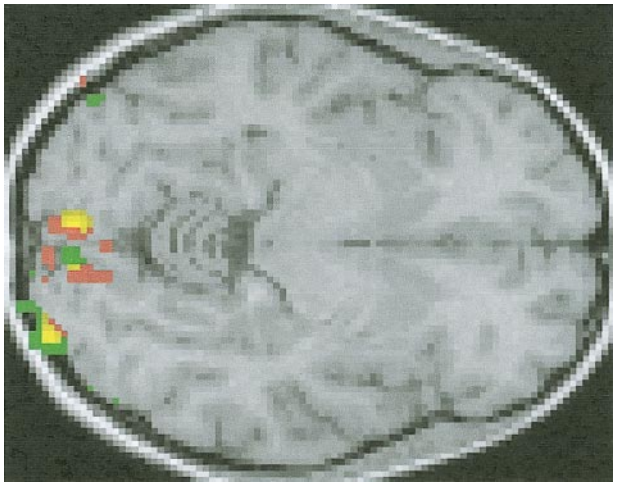

FIG. 3. Areas identified by PCA. See Fig. 1 for key.

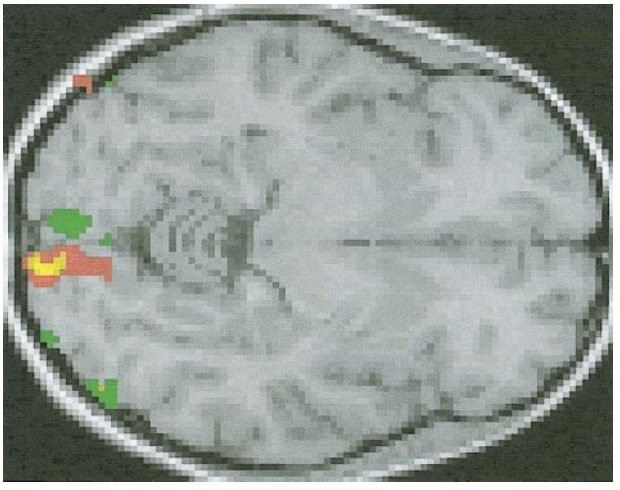

FIG. 5. Areas identified by tICA. See Fig. 1 for key.

LINEAR MODELS: PRINCIPAL COMPONENT ANALYSIS (PCA), ICA, AND GENERAL LINEAR MODEL (GLM)

In order to compare ICA results with conventional techniques, we analyzed our data using a number of methods, each of which provides a linear decomposition. Each method is used to obtain a set of images and a corresponding set of temporal sequences. These temporal sequences are used to evaluate each method with respect to either the temporal sources for the synthetic data set or the model fMRI time courses for the experimental protocol used.

A full description of ICA, based on the method of Bell and Sejnowski (1995), can be found in Appendix 1. As described in Appendix 1, the nature of signals extracted by ICA is determined by the particular form of model pdf σ' employed. Accordingly, each ICA method described in this section is defined in terms of one or more model pdfs.

We represent a sequence of n images as an $m \times n$ matrix $X = (x_1| \ldots |x_n)$, where each image contains m pixels.

Principal Component Analysis

PCA was used for two reasons: in order to provide a comparison for ICA and to provide a reduced rank data set as input to ICA and to the GLM. PCA was executed using the singular value decomposition (SVD) method. SVD provides a factorization,

$$X \approx UDV^t, \tag{1}$$

where U is an $m \times k$ matrix of $k \leq m$ eigenimages, V is an $n \times k$ matrix of k eigensequences, and D is a diagonal matrix of singular values $\lambda^{1/2}$. The approximation is exact if $k = n$. Each singular value corresponds to the

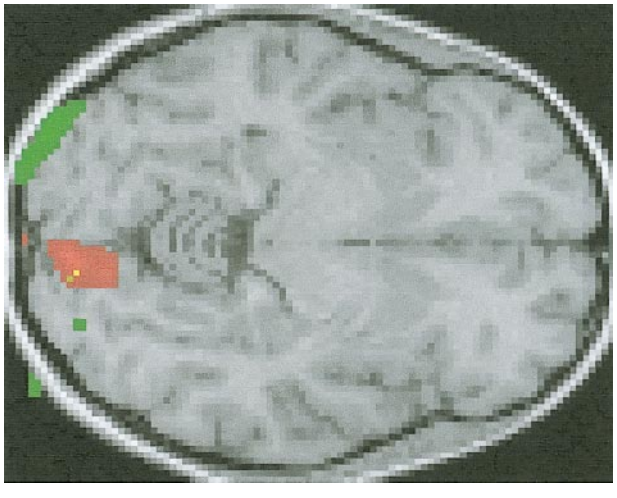

FIG. 4. Areas identified by sICA. See Fig. 1 for key.

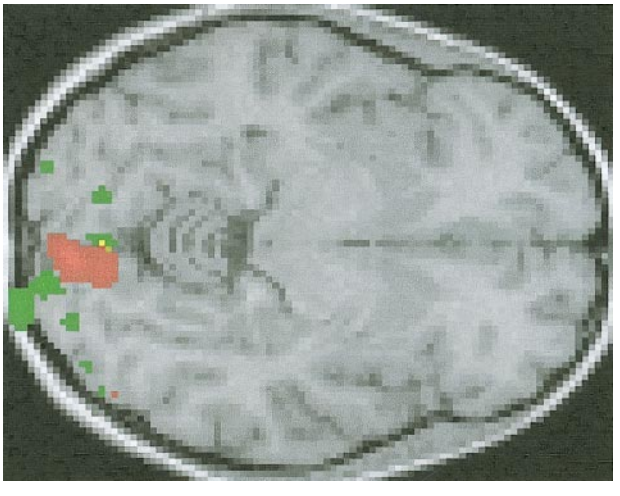

FIG. 6. Areas identified by stICA. See Fig. 1 for key.

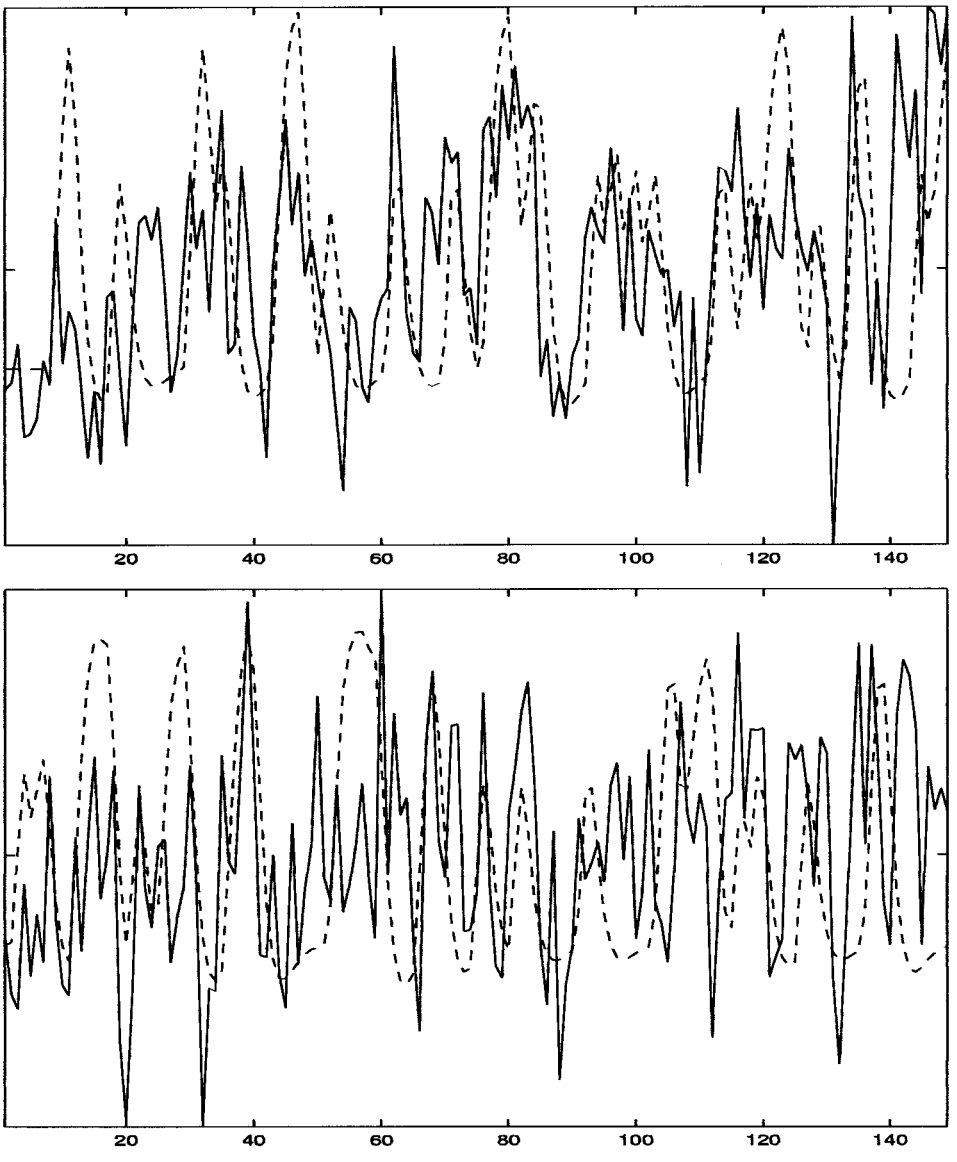

FIG. 2. Time courses associated with images (shown in Fig. 1) extracted by the general linear model (GLM) (see text for details). These time courses are induced by stimuli presented in the left and right visual hemifield (shown in the top and bottom figures, respectively). Solid line, mean time course of pixels associated with stimulation of the left (or right, resp.) visual hemifield, shown in red (or green, resp.) in Fig. 1. Dashed line, predicted activation sequence (PAS) for visual areas in the right (or left, resp.) hemisphere, based on the time course of visual stimulation in the left (or right, resp.) visual hemifield. For display purposes, time courses have been normalized to have zero mean and unit variance.

square root of an eigenvalue λ. For later use, we define $\tilde{X} \approx X$ as

$$X \approx \tilde{X} = UDV^t = (UD^{1/2})(VD^{1/2})^t = \tilde{U}\tilde{V}^t. \qquad (2)$$

Independent Component Analysis

Briefly, spatial ICA finds independent images S and a corresponding set of dual unconstrained time courses T, whereas temporal ICA finds independent time courses T and a corresponding set of dual unconstrained images S.

Spatial ICA. sICA embodies the assumption that each image in $\tilde{U}$ is composed of a linear combination of k spatially IC images $\tilde{U} = S\tilde{A}_S$, where $\tilde{A}_S$ is a $k \times k$ mixing matrix and S is an $m \times k$ set of k statistically independent images $S = (\mathbf{s}_1 | \ldots | \mathbf{s}_k)$. sICA decomposes $\tilde{U}$ into k IC images $\mathbf{y}_S = \tilde{U}W_S$. The unmixing matrix W_S is a permuted version of $\tilde{A}_S^{-1}$, such that each column in $\mathbf{y}_S$ is a scaled version of exactly one column in S.

ICA achieves this by maximizing the entropy $H(\mathbf{Y}_S)$

of $\mathbf{Y} = \sigma_S(\mathbf{y}_S)$, where σ_S approximates the cumulative density function (cdf) of each of the spatial ICs (see Appendix 1). The model pdf used to evaluate $H(\mathbf{Y}_S)$ for sICA is $\sigma'_S = \text{sech}^2 y$, which defines the standard ICA high-kurtosis pdf.

The k dual time courses A_S associated with the k IC images $\mathbf{y}_S$ can be recovered as follows. If $\tilde{X} = \mathbf{y}_S A_S = \tilde{U}\tilde{V}^t$ and $\tilde{U} = y_S W^{-1}$, then

$$\tilde{X} = \mathbf{y}_S W_S^{-1} \tilde{V}^t = \mathbf{y}_S A_S, \tag{3}$$

where each row of $A_S = W_S^{-1}\tilde{V}^t$ contains one time course.

Temporal ICA. tICA embodies the assumption that each eigensequence in $\tilde{V}$ is a linear combination of k temporally independent sequences $\tilde{V} = T\tilde{A}_T$, where $\tilde{A}_T$ is a $k \times k$ mixing matrix and T is an $n \times k$ set of k statistically independent temporal sequences $T = (\mathbf{t}_1 | \ldots | \mathbf{t}_k)$. tICA decomposes V into k IC sequences $\mathbf{y}_T = \tilde{V} W_T$. W_T is a permuted version of A_S^{-1}, such that each column vector in $\mathbf{y}_T$ is a scaled version of exactly one column in T. This is achieved by maximizing the entropy $H(\mathbf{Y}_T)$ of $\mathbf{Y}_T = \sigma_T(\mathbf{y}_T)$, where σ_T approximates the cdf of the temporal source signals. The model pdf used to evaluate $H(\mathbf{Y}_T)$ for tICA is $\sigma'_T = \text{sech}^2 y$, which defines the standard ICA high-kurtosis pdf.

The k dual images A_T associated with the k IC sequences $\mathbf{y}_T$ can be recovered as follows. If $\tilde{X}^t = \tilde{V}\tilde{U}^t$ and $\tilde{V} = y_T W_T^{-1}$, then

$$\tilde{X}^t = \mathbf{y}_T W_T^{-1} \tilde{U}^t = \mathbf{y}_T A_T, \tag{4}$$

where each row of $A_T = W_T^{-1}\tilde{U}^t$ contains one image.

Spatiotemporal ICA. Given the eigendecomposition $\tilde{X} = \tilde{U}\tilde{V}^t$, stICA embodies the assumption that each eigenimage in $\tilde{U}$ is a linear combination of k spatially independent images S, *and* each eigensequence in $\tilde{V}$ is a linear combination of k temporally independent sequences T

$$\tilde{X} = S\Lambda T^t, \tag{5}$$

where S is an $m \times k$ matrix of k mutually independent images, T is an $n \times k$ matrix of mutually independent sequences, and Λ is a diagonal scaling matrix. Λ is required to ensure that S and T have amplitudes appropriate to their respective cdfs σ_S and σ_T.

If $\tilde{X} = \tilde{U}\tilde{V}^t$, then two $k \times k$ unmixing matrices W_S and W_T exist such that $S = \tilde{U}W_S$ and $T = \tilde{V}W_T$

$$\tilde{X} = S\Lambda T^t = \tilde{U}W_S\Lambda(\tilde{V}W_T)^t = \tilde{U}W_S\Lambda W_T^t\tilde{V}^t = \tilde{U}\tilde{V}^t, \tag{6}$$

implying that $W_S\Lambda W_T^t = I$, from which we can derive

$$W_T = (W_S^{-1})^t(\Lambda^{-1})^t. \tag{7}$$

The matrices W_S and W_T can be found by simultaneously maximizing a function h_{ST} of the spatial and temporal entropy of extracted signals. The function h_{ST} to be maximized is defined

$$h_{ST}(W_S, \Lambda) = \alpha H(\mathbf{Y}_S) + (1 - \alpha)H(\mathbf{Y}_T), \tag{8}$$

where α defines the relative weighting afforded to spatial and temporal entropy. We used $\alpha = 0.5$ for results presented here.

The temporal entropy is $H(\mathbf{Y}_T)$, where $\mathbf{Y}_T = \sigma_T(\mathbf{y}_T)$ and $\mathbf{y}_T = \tilde{V}W_T$ is a set of extracted temporal signals. The spatial entropy is $H(\mathbf{Y}_S)$, where $\mathbf{Y}_S = \sigma_S(\mathbf{y}_S)$ and $\mathbf{y}_S = \tilde{U}W_S$ is a set of extracted spatial signals. The functions σ_S and σ_T are assumed to be the cdfs of the individual spatial and temporal signals, respectively, and their derivatives σ'_S and σ'_T are their corresponding pdfs. The model pdf used to evaluate $H(\mathbf{Y}_S)$ and $H(\mathbf{Y}_T)$ are $\sigma'_S = \text{sech}^2 y$ and $\sigma'_T = \text{sech}^2 y$ (respectively), which defines the standard ICA high-kurtosis pdf.

Skew-sICA. The source images of interest usually contain localized features, such as vascular activity, on a homogeneous background. Such images are characterized by the skewness (rather than the kurtosis) of their pdfs. Accordingly, a convenient functional form for a skew-pdf σ'_{SK} is

$$\sigma'_{SK}(y; a, b) \propto \exp\left(\frac{a-b}{2}y - \frac{a+b}{2}\sqrt{y^2+1}\right), \tag{9}$$

which has the asymptotic behaviors $\sigma'_{SK} \approx \exp(ay)$ in the limit $y \to -\infty$, and $\sigma'_{SK} \approx \exp(-by)$ in the limit $y \to +\infty$. The exponential decay rates ensure that it shares the robustness properties of the more conventional pdf $\text{sech}^2 y = d \tanh y/dy$. The constants a and b define the skewness of the distribution; we used $a = 4$ and $b = 1$.

We cannot guarantee that the chosen values for a and b are consistent with the spatial sources. However, it is known that an approximate match between the model pdf used in ICA and the source pdfs is a sufficient condition for ICA when the sources are exactly independent (Cardoso, 2000). For the nonindependent sources considered here, a pdf with high skewness constitutes a generic model of spatial sources, just as a high-kurtosis pdf constitutes a generic model of speech signals (Bell and Sejnowski, 1995). The time courses are obtained in the same manner as described for sICA.

Skew-stICA. Skew-stICA was implemented by maximizing the function h_{ST} defined in Eq. (8), except that the skewed spatial pdf σ'_{S_K} defined in Eq. (9) was used for $H(\mathbf{Y}_S)$.

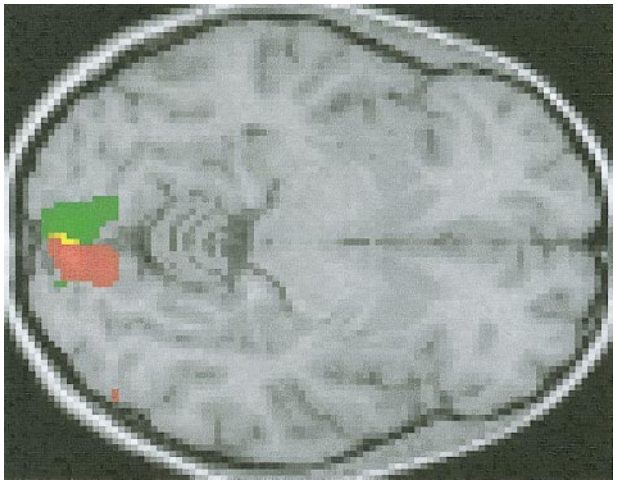

FIG. 7. Areas identified by skew-sICA. See Fig. 1 for key.

General Linear Model

In order to compare the results of ICA with a standard method, we analyzed our data using a GLM. Given an $(n \times m)$ set $\tilde{X}^t$, containing n images in its rows, and an $(m \times p)$ design matrix G, the general linear model is $\tilde{X}^t = G\beta + \epsilon$, where ϵ is gaussian noise. A least squares estimate of the regression coefficients β can be obtained as $\beta = (G^tG)^{-1}G^t\tilde{X}^t$, where β is a $(p \times m)$ matrix, with one image per row.

Results for GLM are displayed as follows. For fMRI data, the extracted images are the rows β, and the time courses are given by the mean time course of the top 5% of the pixels with the largest magnitudes in each row of β. For the synthetic fMRI data, each image is simply one row of β, and each time course is the corresponding column of the estimated time courses $\hat{G} = X^t\beta^t(\beta\beta^t)^{-1}$.

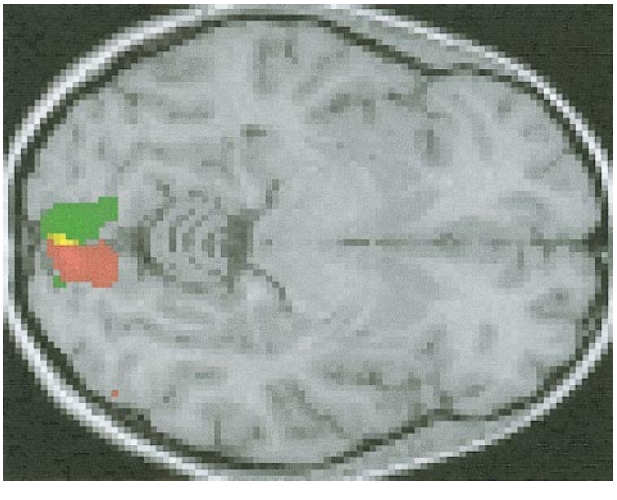

FIG. 8. Areas identified by skew-stICA. See Fig. 1 for key.

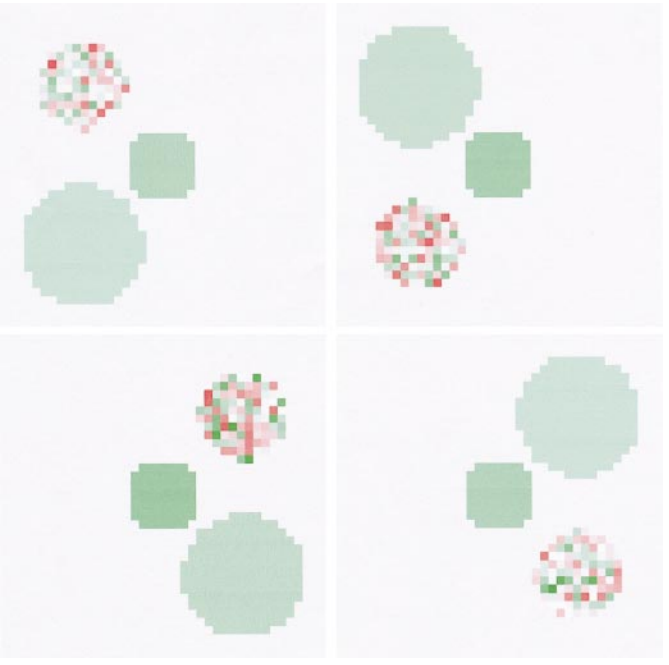

FIG. 10. Source images used for synthetic fMRI data. See text for details. Red denotes negative pixel values, and green denotes positive pixel values in all synthetic fMRI images.

MATERIALS AND METHODS

The ICA methods described above were tested on synthetic data and on data obtained from an event-related left/right visual hemifield stimulation fMRI experiment, using a single 27-year-old female subject.

fMRI Data

Stimuli. The stimuli consisted of three types of events: a checkerboard presented either (1) to the left or (2) to the right of a fixation cross and (3) a null event (i.e., fixation cross only). All three events were presented with equal probability every 3 s, giving a mean stimulus onset asynchrony (SOA) of 9 s. Each checkerboard stimulus was counterphased at 4 Hz and was presented for 1 s. Stimuli were projected onto a translucent screen 2.5 m away from the subject, subtending a visual angle of 9° and viewed via a 45° angled mirror.

Procedure. The subject was instructed to look at the fixation cross, which was present at all times. The anatomy of the visual system ensures that stimuli presented to one side of the fixation point stimulates the hemisphere on the opposite side of the brain. Therefore, the left and right stimuli should produce a corresponding response in the right and left visual hemispheres, respectively.

fMRI data acquisition. The fMRI data were obtained from a 1.5-T Eclipse (Marconi Medical Systems, Highland Heights, OH) using a gradient recalled, single-shot echo-planar imaging (EPI) sequence (TE = 40

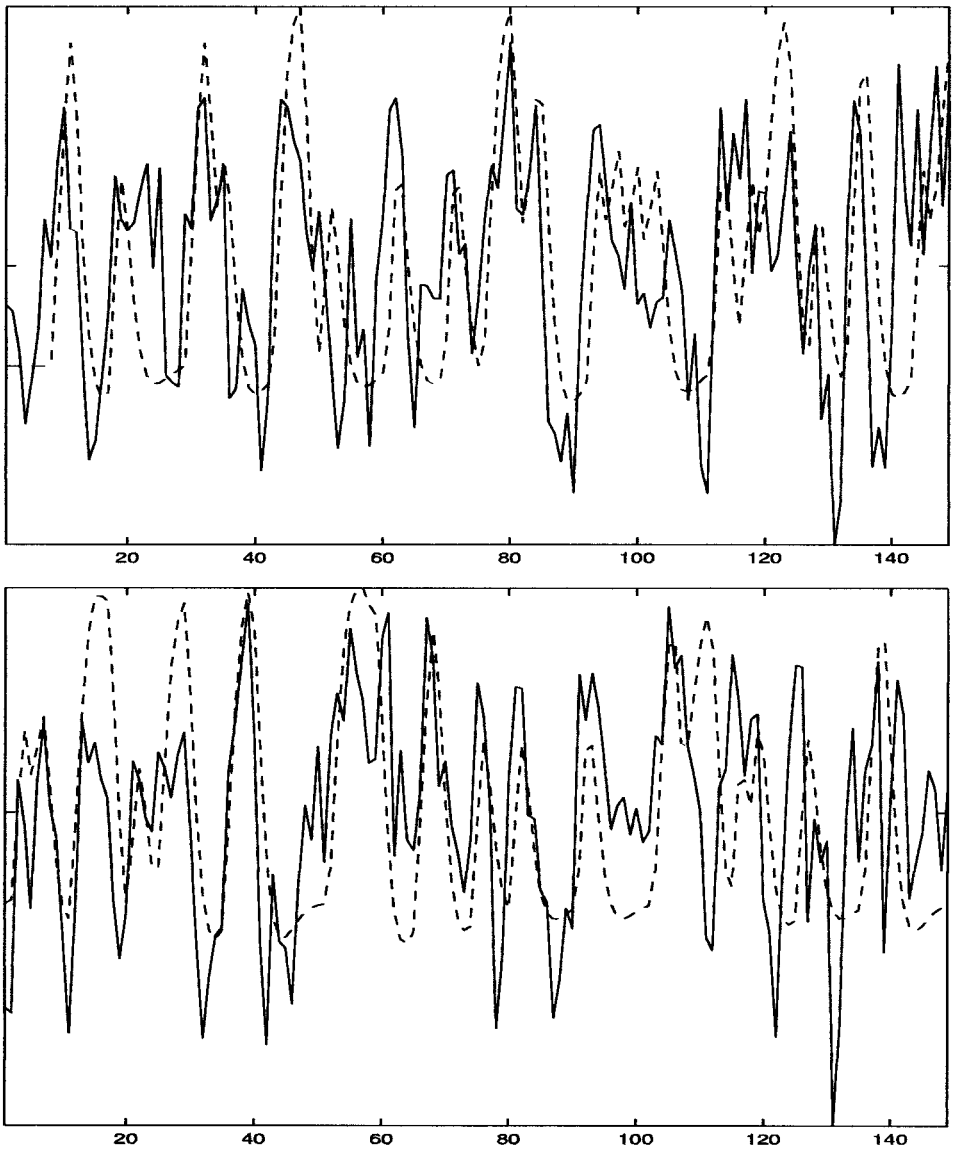

FIG. 9. Skew-stICA time courses. Time course activation induced by stimuli presented in the left and right visual hemifield (shown in the top and bottom figures, respectively). Solid lines, time courses (ICs) of pixels associated with stimulation of the left (or right, resp.) visual hemifield, shown in red (or green, resp.) in Fig. 8. Dashed line, predicted activation sequence (PAS) for visual areas in the right (or left, resp.) hemisphere, based on the time course of visual stimulation in the left (or right, resp.) visual hemifield. See Results for details. For display purposes, time courses have been normalised to have zero mean and unit variance.

ms; TR = 2 s). A set of 153 T2* brain volumes (20 contiguous slices, thickness = 5 mm; matrix = 128 × 128; FOV = 240 mm) were acquired during each 306-s run. However, the first 4 volumes were not used in the analysis to avoid saturation effects, making a total of 149 volumes available for analysis.

Analysis. We used the SPM99 software package (SPM99, 1999) to identify areas associated with visual stimulation. As expected, these areas were confined to the lingual gyri of area V2. In order to reduce the run time associated with analysis, a 79 × 18-pixel section of one slice of data from this region of interest (ROI) was extracted for further analysis (each pixel is 2 × 2 mm). This yielded a data set $X3 = [79 \times 18 \times 149]$. These data were extracted after motion correction and normalization to the SPM EPI template. The location of the slice was −2.6 mm (Talairach coordinates).

Data preprocessing. The 3D data set $X3 = [79 \times 18 \times 149]$ was transformed to a 2D data set by collapsing across one spatial dimension to yield $X = [1422 \times 149]$. X was preprocessed so that each row (i.e., time) and each column (i.e., space) had zero mean. We applied PCA to X using the SVD method. We use a standard criterion to discard eigenvectors with eigenvalues less than unity (with each eigenvalue λ normalized such that $\Sigma_i \lambda_i = 149$). This resulted in $k = 18$ of 149

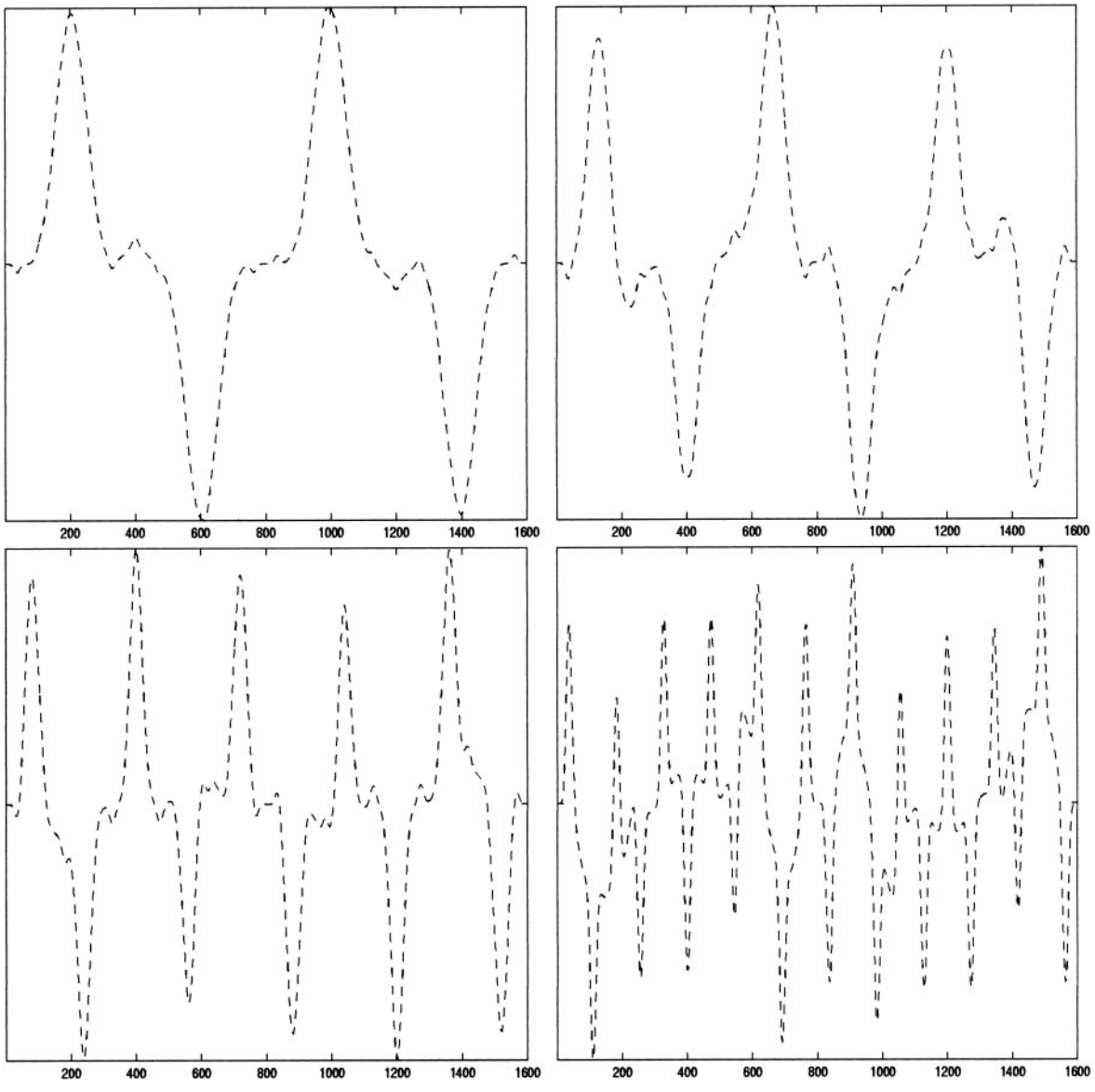

FIG. 11. Source time courses used for synthetic fMRI data. See text for details.

eigenvectors being retained. These eigenvectors define a data set $X \approx \tilde{X}$, as defined in Eq. (2). Columns of $\tilde{U}$ and $\tilde{V}$ contain the first 18 scaled eigenimages and eigensequences of X, respectively.

Synthetic fMRI Data

We synthesized data that were designed to emulate the gross properties of fMRI data. Briefly, the data contained four nonindependent, noisy, spatial source signals (images) S modulated by four corresponding nonindependent temporal source signals T (see Figs. 10 and 11).

Each spatial source contained three circular regions, A, B, and C. Regions A and B had uniform intensity, and region C was gaussian noise. All spatial sources contained a region A in a different position; on their own, these regions would ensure that the spatial sources were independent. However, all spatial sources also contained a region B in the same central position, and the spatial sources were therefore not spatially independent. Additionally, all spatial sources contained a region C of gaussian noise in a different position. Note that mixtures of these sources tend to have overlapping spatial sources in every quadrant of each image mixture (see Fig. 14).

Temporal source signals T were created as follows. Four supergaussian signals $\mathbf{z}$ were created by raising four sinusoids to the power 7. Small correlations between these signals were created using a matrix $M = \beta I + (1 - \beta)R$, where I is the identity matrix, R is a

TABLE 2

Magnitude of Correlation between the Four Time Course Extracted by Each Method and the Temporal Source Signals

Method	Source 1	Source 2	Source 3	Source 4	Mean
GLM	0.87	0.93	0.95	0.84	0.90
PCA	0.76	0.55	0.77	0.81	0.72
tICA	0.87	0.90	0.91	0.72	0.85
sICA	0.48	0.75	0.40	0.71	0.59
stICA	0.87	0.89	0.91	0.72	0.85
Skew-sICA	0.83	0.90	0.88	0.81	0.86
Skew-stICA	0.87	0.92	0.94	0.83	0.89

Note. See text for details.

random gaussian matrix, and $\beta = 0.85$. The temporal source signals are defined as $T = M\mathbf{z}$. These temporal source signals were therefore not temporally independent.

The spatial sources S consisted of a set of four 40×40-pixel images (described above). These spatial sources were modulated with four corresponding time courses T (each of which was 1600 time steps) to form a 1600×1600 data set $X = ST^t$, where each column of X was formed by concatenating rows in a single image.

We applied PCA (using SVD) to the $n = 1600$ images in X (see Fig. 14). X was preprocessed so that each row and each column had zero mean. SVD was applied to X to obtain $\tilde{X} = \tilde{U}\tilde{V}^t$, where columns of $\tilde{U}$ and $\tilde{V}$ contain the first four eigenimages and eigensequences of X, respectively. These four eigenvectors were used as input to ICA methods.

Function Optimization

For each ICA method, a solution was obtained using a conjugate gradient method (Williams, 1991). Each solution was obtained in under 350 conjugate gradient line searches, which required approximately 2 min on a 450 MHz (Intel PIII) PC. All solutions are based on the same (random) initial $k \times k$ unmixing matrix W_S.

RESULTS

fMRI Data

Results were evaluated as follows. The correlations between the time courses extracted by each method and the predicted activation sequence (PAS) for each hemisphere were computed. The PAS is given by a signal which is the convolution of the visual stimulation time course with a canonical hemodynamic response HR function.[2] The visual stimulation time course for one hemisphere consists of a delta function for each stimulus presented to that hemisphere. We compared the sequences extracted by each method (PCA, sICA, tICA, stICA, skew-sICA, and skew-stICA) with the PAS for each hemisphere. Results are summarized in Table 1. Note that, for those methods that yielded more than two time courses (e.g., sICA), the correlations with the highest magnitudes are displayed.

The correlations in Table 1 were obtained as follows. For the GLM, the correlation between the PAS for each hemisphere and the mean time course of a set P of pixels was computed (see Fig. 2). The set P was defined as those pixels with the highest magnitude in each image extracted by the GLM. For all methods except the GLM, exactly k time courses were identified, as described above. For a given method, the two correlation values given in Table 1 were obtained as follows. The correlation between each time course and each of the two (left/right) PASs was computed. Those two time courses with the largest magnitude correlations were defined as the *estimated time courses* $\mathbf{y} = \{y_L, y_R\}$ for a given method. In each of the Figs. 1–8, the red and green activations correspond to $\mathbf{y}_L$ and to $\mathbf{y}_R$, respectively, for a given method.

All results are based on analysis of a reduced rank data $\tilde{X} = \tilde{U}\tilde{V}$ set, having $k = 18$ eigenvectors of the original data set X.

GLM. The results from the GLM are intended to act as a baseline for comparison with other methods. Note that this is the only model-based method used in this paper. It should not be surprising that, of the methods used here, the correlation between the extracted GLM time courses and the PASs is largest (see Fig. 2). Despite this, the images identified by the GLM show regions of activation well outside of the primary visual areas (see Fig. 1).

Note that the regions of activation identified by the GLM are similar to those identified by SPM99, the main difference being that SPM99 did not identify regions in one hemisphere induced by visual stimulation of the ipsilateral visual hemifield. This difference is due to the fact that the raw image magnitude of each pixel identified in Fig. 1 is not normalized with respect to its underlying variance.

PCA. Figure 3 shows eigenimages, corresponding to the eigensequences $\mathbf{y}$ extracted by PCA. The eigenimages are qualitatively similar to results obtained with the GLM, the main difference being the larger degree of left/right activation overlap here.

sICA. Figure 4 shows regions of activation, which correspond to $\mathbf{y}$ for sICA. Whereas left hemifield stimulation produces activation which is well localized to the primary visual areas of the right hemisphere, right hemifield stimulation produces activation exclusively outside of the primary visual areas.

tICA. Figure 5 shows regions of activation, which correspond to $\mathbf{y}$ for tICA. As with GLM and PCA, it

[2] As defined in SPM99 (1999).

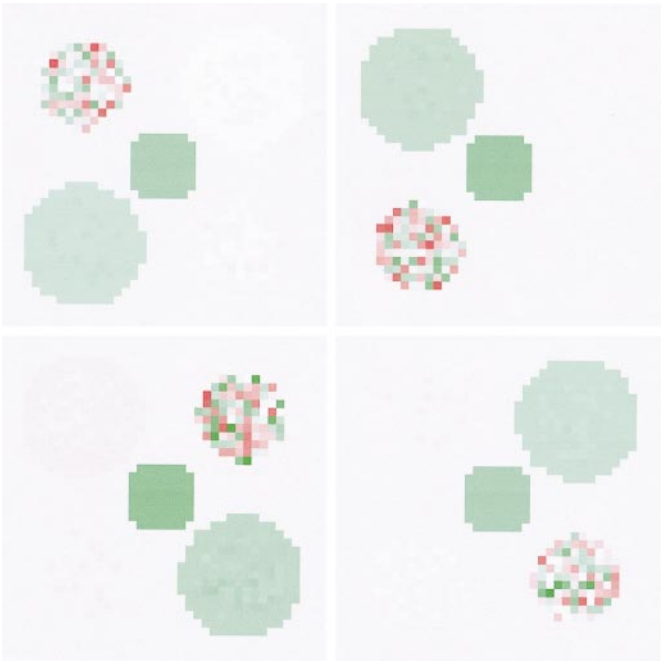

FIG. 12. GLM images extracted from synthetic fMRI data (see "General Linear Model" for details).

appears that stimuli in the right hemifield generate activity in both hemispheres, as well as activity outside of the primary visual areas. In contrast, stimuli in the left hemifield appears to produce activity restricted to the primary visual areas of the right hemisphere.

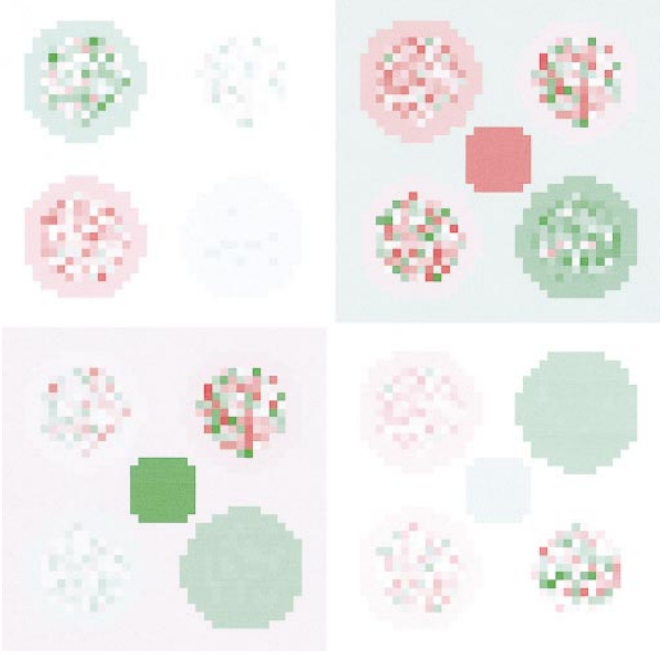

FIG. 14. PCA images extracted from synthetic fMRI data. These four eigenimages and their corresponding eigensequences (not shown) were used as input to all ICA methods.

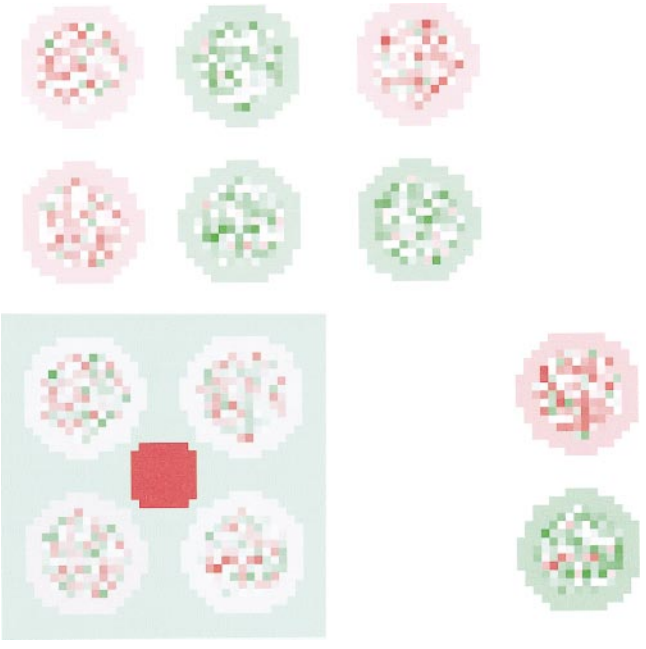

FIG. 15. sICA images extracted from synthetic fMRI data.

stICA. Figure 6 shows regions of activation, which correspond to **y** for stICA. Whereas left hemifield stimulation produces activation which is well localized to the primary visual areas of the right hemisphere, right hemifield stimulation produces widely scattered activation in both hemispheres.

Skew-sICA. Figure 7 shows two regions of activation, which correspond to **y** for skew-sICA. Left and

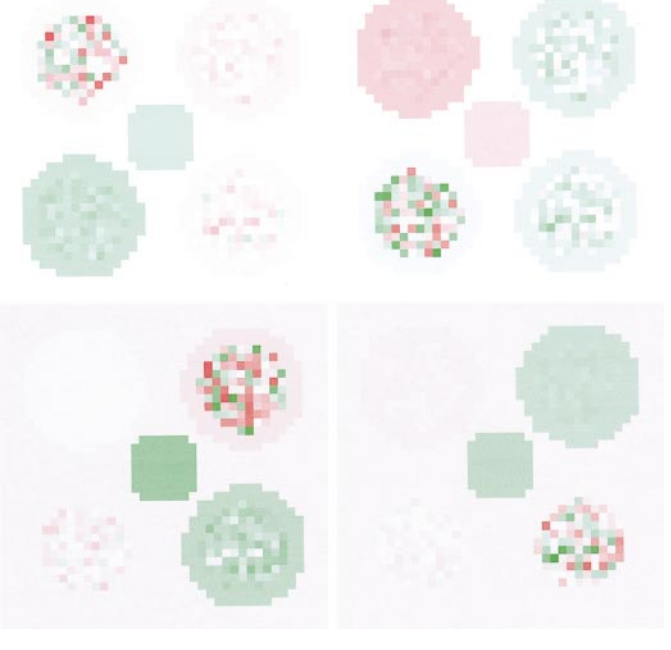

FIG. 16. Images corresponding to temporal ICs extracted by tICA from synthetic fMRI data.

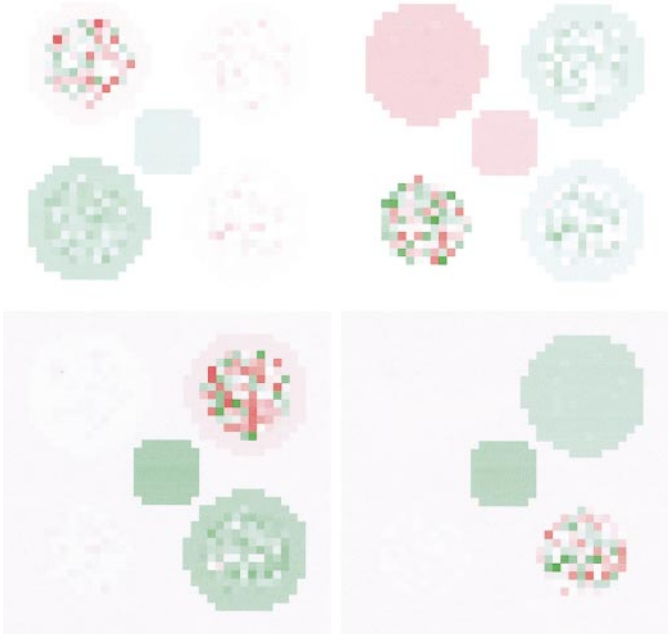

FIG. 17. Images extracted from synthetic fMRI data by stICA.

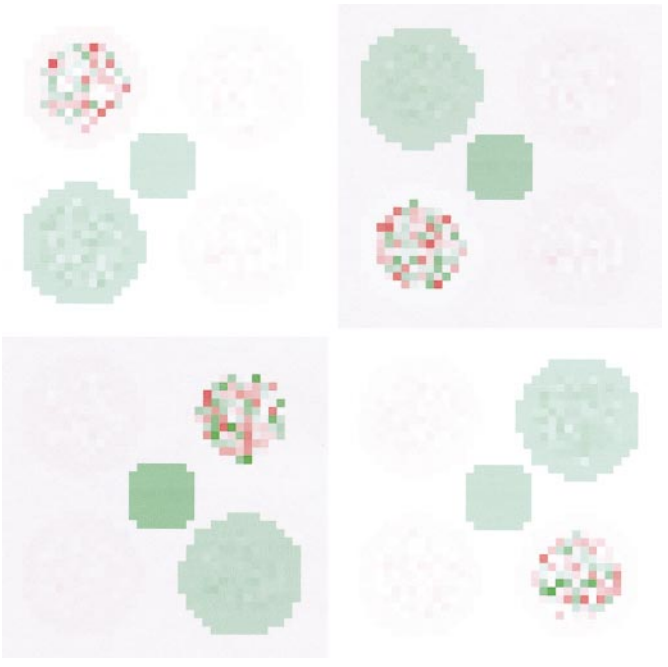

FIG. 19. Images extracted from synthetic fMRI data skew-stICA.

right visual hemifield stimulation produce activation only in the contralateral primary visual areas of each hemisphere, although a small region of overlap can be seen between the hemispheres.

Skew-stICA. Figure 8 shows two regions of activation, which correspond to **y** for skew-stICA. Left and right visual hemifield stimulation produce activation only in the contralateral primary visual areas of each hemisphere, although a small region of overlap can be seen between the hemispheres. It is possible that this overlap found here and for skew-sICA results from the spatial smoothing applied to the data. Although the time course correlations are numerically larger than those obtained with skew-sICA, the results obtained with skew-sICA and skew-stICA are very similar. The skew-stICA time courses (ICs) of the left and right regions shown in Fig. 8 are shown in Fig. 9.

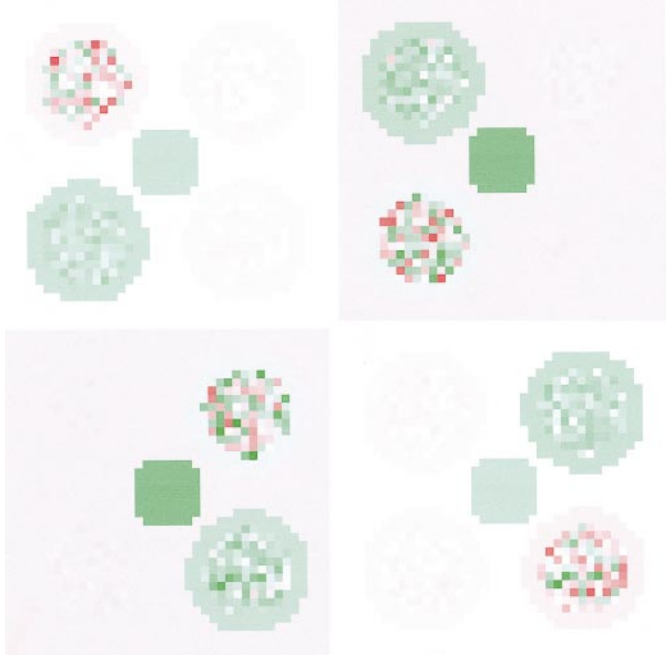

FIG. 18. Images extracted from synthetic fMRI data by skew-sICA.

Synthetic fMRI Data

Results were evaluated as for the fMRI data (Figs. 10 and 11), except the PASs are the synthetic temporal sources. Correlations between extracted time course and the PASs are shown in Table 2.

GLM. The spatial modes are a close match to the spatial sources (see Fig. 12), and the correlations between the extracted GLM time courses and the temporal sources are high (see Fig. 13).

PCA. Figure 14 shows eigenimages, corresponding to the eigensequences **y** extracted by PCA. The spatial PCs do not recover the source images, and the temporal PCs are poorly correlated with the PASs. These eigenmodes were used as input to all ICA methods.

sICA. Figure 15 shows regions of activation, which correspond to **y** for sICA. The sources are not kurtotic (an assumption implicit in the tanh model pdf used with sICA), but sICA imposes kurtosis on the images it extracts. This results in three of four extracted images containing equal amounts of negative and positive ac-

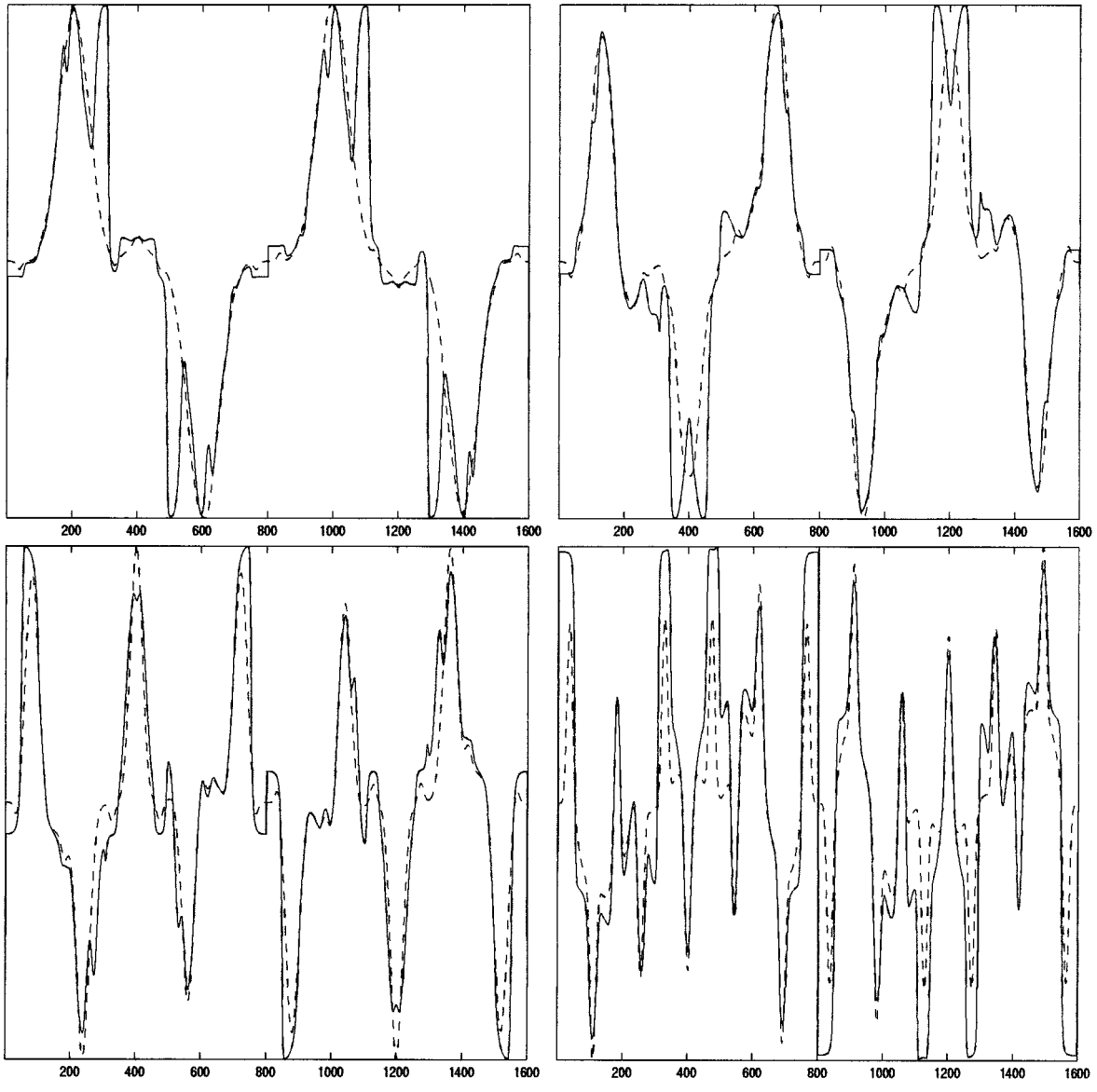

FIG. 13. GLM time courses extracted from synthetic fMRI data, corresponding to Fig. 12. Dashed line, predicted activation sequence (PAS) (i.e., temporal source signal). Solid line, time course estimated by GLM course of GLM (see "General Linear Model"). For display purposes, time courses have been normalised to have zero mean and unit variance.

tivation (red and green in Fig. 15), plus large regions of zero-valued pixels, which approximates a kurtotic pdf. The fourth extracted image contains the central region which is common to all underlying source images (see Fig. 10). A side effect of this is that extracted temporal sequence are also distorted, which have a mean correlation with sources of $r = 0.59$.

tICA. Figure 16 shows regions of activation corresponding to **y** for tICA. Even though the correlations of the extracted sequences with the PASs are high, the corresponding images are a poor approximation to the source images.

stICA. Figure 17 shows regions of activation corresponding to **y** for stICA. This method produces essentially the same result as that of tICA because the high-kurtosis spatial pdf of stICA is inconsistent with the skew-pdf of the spatial source images.

Skew-sICA. Figure 18 shows regions of activation corresponding to **y** for skew-sICA. Compared to tICA and stICA, there is a slight improvement in the temporal correlations between extracted and source sequences. There is also a marked improvement in the quality of the extracted spatial images, with source regions being more clearly defined.

Skew-stICA. Figure 19 shows regions of activation corresponding to **y** for skew-stICA. Skew-stICA is an improvement on skew-sICA, both temporally (Fig. 20) and spatially. Less noise is present in the extracted

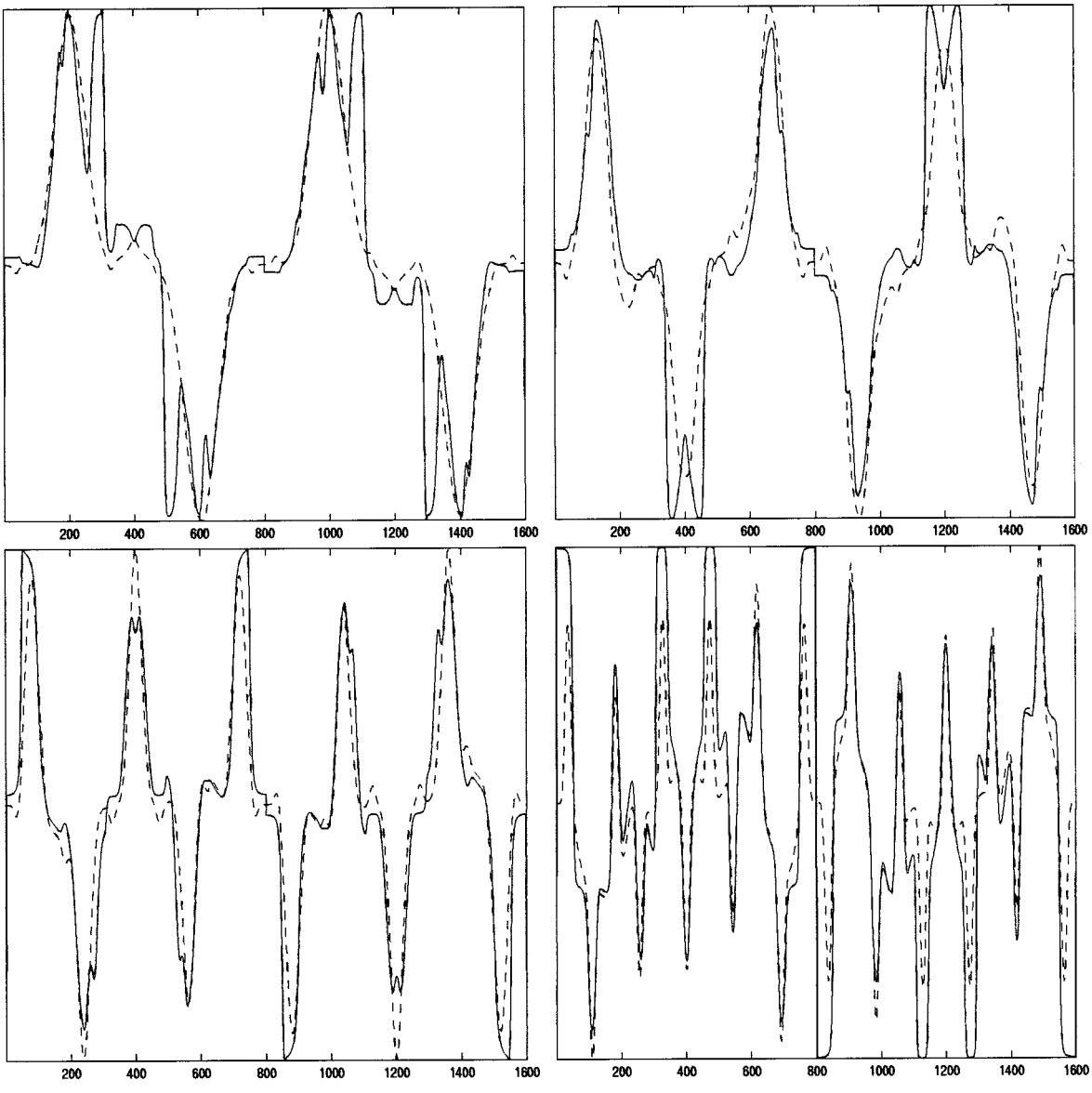

FIG. 20. Time courses extracted from synthetic fMRI data by skew-stICA.

images, and the correlations between temporal ICs and extracted sequences are highest of all the methods, except GLM.

DISCUSSION

ICA is clearly a powerful method, but its successful application to biomedical data depends critically on the validity of the statistical assumptions implicit in the method. Accordingly, our approach is to incorporate physically realistic constraints into the solutions obtained by ICA. Overall, results for the fMRI and synthetic data are qualitatively similar, and the following applies to both data sets.

The correlations between the PAS (i.e., time course dictated by the experimental protocol for fMRI data) and the time courses of all methods used here are large. However, time courses extracted by GLM, PCA, tICA, and stICA are associated with images which are inconsistent with the PAS for both fMRI and synthetic data. In contrast, time courses extracted by skew-sICA and skew-stICA are associated with well-defined images which are consistent with the PASs. It therefore appears that stICA alone does not perform as well as skew-sICA nor skew-stICA. Critically, for the fMRI data, both skew-sICA and skew-stICA provide results which appear physically more plausible than those obtained with the traditional GLM method.

Note that the GLM method requires an explicit time course model of the experimental protocol, whereas skew-sICA and skew-stICA both embody a generic model of the statistical structure of spatial ICs (i.e.,

skewed pdf). Despite this difference in the underlying models of GLM and skew-stICA, GLM accounts for a mean of 40% ($= (0.58^2 + 0.68^2)/2$) of the variance in the fMRI data, whereas skew-stICA accounts for 24% ($= (0.52^2 + 0.46^2)/2$) of the variance.

sICA represents a benchmark method for fMRI data (McKeown *et al.,* 1998), although the current paper is the first application of ICA to *event-related* fMRI data. For our fMRI and synthetic data, results from skew-sICA and skew-stICA are clearly superior to results from sICA. In this respect, the strategy of incorporating physically realistic assumptions into ICA appears to have some merit.

It is possible that different results could have been obtained using second order blind source separation (BSS) methods (Ziehe *et al.,* 2000; Stone, 2001), because such methods make minimal assumptions regarding the pdfs of data. However, as noted in the Introduction, any method will fail if the assumptions on which it is based are violated. In contrast to ICA methods, BSS second-order methods assume zero correlations between extracted time courses at user-specified lags (Ziehe *et al.,* 2000) or that different extracted time courses have zero correlations and their temporal derivatives have zero correlation (Stone, 2001). While such assumptions are quite general (and in this respect, similar to the assumptions of ICA), there is no *a priori* reason to suppose that they are more likely to be valid than those of ICA.

There is no unambiguous divide between model-driven (e.g., GLM, SPM) and data-driven methods (e.g., PCA, ICA). All methods which are apparently data driven imply a particular statistical model, whether or not this model is made explicit. For example, the model implicit in PCA is that different modes are gaussian and uncorrelated, whereas ICA's model is that different modes are nongaussian and independent. The main difference between model- and data-driven methods such as SPM and ICA is one of degree and involves the specificity of the model implied by these methods.[3] Whereas SPM uses an explicit time course as a model of brain activation, an ICA model consists of a generic statistical constraint on the spatial and/or temporal form of brain activation. We argue that, if the model implicit in ICA is essentially valid then the results obtained by ICA may be superior to those obtained with conventional model-driven techniques such as SPM.

The results presented in this paper provide evidence consistent with the argument that so-called data-driven methods can provide results which are qualitatively similar to those of model-driven methods, such as SPM.

[3] In practice, model-driven methods can associate a level of statistical significance with a model of a given data set, whereas data-driven techniques usually cannot.

APPENDIX 1

Independent Component Analysis

The description of ICA given here is based on Bell and Sejnowski (1995). Consider the outputs $\mathbf{x} = (x_1, \ldots, x_N)^T$ of N measurement devices (e.g., microphones), such that each output variable x_i is a linear mixture of N independent source signals $\mathbf{s} = (s_1, \ldots, s_N)^T$ $\mathbf{x} = A\mathbf{s}$, where A is an $(N \times N)$ mixing matrix. We wish to find an $(N \times N)$ unmixing matrix W such that each recovered component y_i in $\mathbf{y} = W\mathbf{x}$ is one of the original signals $\mathbf{s}$.

ICA recovers estimates $\mathbf{y} = (y_1, \ldots, y_N)^T$ of the source signals $\mathbf{s} = (s_1, \ldots, s_N)^T$ by finding an unmixing matrix W which maximizes the entropy $H(\mathbf{Y})$ of the joint distribution $\mathbf{Y} = \{Y_1, \ldots, Y_N\} = \{\sigma_1(y_1), \ldots, \sigma_N(y_N)\}$. Each function σ_i has the same form as the cdf of the corresponding source signal s_i. However, in many cases it is sufficient to approximate these cdfs (Bell and Sejnowski, 1995; Cardoso, 2000). The entropy of a multidimensional signal $\mathbf{Y} = \sigma(W\mathbf{x})$ is equal to the entropy of $\mathbf{x}$ plus the change in entropy ΔH induced by the mapping $\mathbf{x} \rightarrow \mathbf{Y}$. The change in entropy ΔH is given by the expected value of $\ln|J|$, where $|\cdot|$ denotes the absolute value of the determinant of a Jacobian matrix, which defines the ratio of infinitesimal volumes in $\mathbf{x}$ and $\mathbf{Y}$. Specifically, the entropy $H(\mathbf{Y})$ is related to the entropy $H(\mathbf{x})$ by $H(\mathbf{Y}) = H(\mathbf{x}) + E[\ln|J|]$, where J is the Jacobian $J = \partial\mathbf{Y}/\partial\mathbf{x}$. Given that $\mathbf{Y} = \sigma(W\mathbf{x})$, and that σ is fixed, the output entropy $H(\mathbf{Y})$ can be maximized by adjusting the entries in the matrix W.

Using the chain rule, we can evaluate $|J|$ as

$$|J| = \left|\frac{\partial \mathbf{Y}}{\partial \mathbf{x}}\right| = \left|\frac{\partial \mathbf{Y}}{\partial \mathbf{y}}\right| \left|\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right| = \prod_{i=1}^{N} \sigma_i'(y_i)|W|, \qquad (10)$$

where $\partial\mathbf{Y}/\partial\mathbf{y}$ and $\partial\mathbf{y}/\partial\mathbf{x}$ are Jacobians, and $\sigma_i' = \partial\sigma_i/\partial y_i$ is (by definition) the pdf of the ith source signal. Substituting Eq. (10) in Eq. (6) yields

$$H(\mathbf{Y}) = H(\mathbf{x}) + E[\sum_{i=1}^{N} \ln \sigma_i'(y_i)] + \ln|W|. \qquad (11)$$

The term $E[\sum \ln \sigma_i'(y_i)]$ can be estimated given n samples from the distribution defined by $\mathbf{y}$

$$E[\sum_{i=1}^{N} \ln \sigma_i'(y_i)] \approx \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{N} \ln \sigma_i'(y_i^{(j)}). \qquad (12)$$

The entropy $H(\mathbf{x})$ of $\mathbf{x}$ is unaffected by W and can therefore be ignored in the maximization of $H(\mathbf{Y})$. Omitting $H(\mathbf{x})$, and substituting Eq. (12) in (11), yields a simpler function $h(W)$ with the same extrema as $H(\mathbf{Y})$

$$h(W) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{N} \ln \sigma_i'(y_i^{(j)}) + \ln|W|. \qquad (13)$$

Given that $y = W\mathbf{x}$, the derivative of h with respect to the ijth element in W is

$$\nabla_h W_{ij} = [W^{-1}]_{ij} + E\left[\sum_i^N \frac{\sigma_i''(y_i)}{\sigma_i'(y_i)} x_j \right].$$

This can be more succinctly written as $\nabla_h W = W^{-1} + \psi(\mathbf{y})\mathbf{x}^t$, where

$$\psi(\mathbf{y}) = \frac{\sigma'}{\sigma''} = \left[\frac{\sigma'(y_1)}{\sigma''(y_1)}, \ldots, \frac{\sigma'(y_k)}{\sigma''(y_k)} \right]. \qquad (14)$$

If both the function $h(W)$ and its derivative are available (as here) then a second-order technique can be used to obtain solutions relatively quickly. We used a conjugate gradient method for the results reported here.

ACKNOWLEDGMENTS

We thank Ying Zheng for helpful discussions and the Sheffield University Neuroimaging Group and the MRC for supporting this work. We thank three anonymous reviewers for their helpful comments. J.V.S. was supported by a Wellcome Mathematical Biology Research Fellowship during this research.

REFERENCES

Bell, A., and Sejnowski, T. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7:** 1129–1159.

Biswal, B., and Ulmer, J. 1999. Blind source separation of multiple signal sources of fmri data sets using independent component analysis. *J. Comput. Assist. Tomogr.* **23**(2): 265–271.

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. 2001. Spatial and temporal independent component analysis of functional mri data containing a pair of task-related waveforms. *Hum. Brain Mapp.* **13:** 43–53.

Cardoso, J. 2000. On the stability of source separation algorithms. *J. VLSI Signal Process. Syst.* **26**(1/2): 7–14.

Jutten, C., and Herault, J. 1988. Independent component analysis versus pca. In *Proceedings, EUSIPCO*, pp. 643–646.

Makeig, S., Jung, T., Bell, A., Ghahremani, D., and Sejnowski, T. 1997. Blind separation of auditory event-related brain responses into independent components. *Proc. Natl. Acad. Sci. USA* **94:** 10979–10984.

McKeown, M. 2000. Detection of consistently task-related activations in fMRI data with hybrid independent components. *NeuroImage* **11:** 24–35.

McKeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., and Sejnowski, T. 1998. Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task. *Proc. Natl. Acad. Sci. USA* **95:** 803–810.

Porrill, J., Stone, J., Berwick, J., Mayhew, J., and Coffey, P. 2000. Analysis of optical imaging data using weak models and ica. In *Advances in Independent Component Analysis* (M. Girolami, Ed.), pp. 217–233. Springer-Verlag, New York.

SPM99 1999. Statistical parametric mapping (spm99). http://www.fil.ion.ucl.ac.uk/spm/.

Stone, J. 2001. Blind source separation using temporal predictability. *Neural Comput.* **13**(7): 1559–1574.

Suzuki, K., Kiryu, T., and Nakada, T. 2001. Fast and precise independent component analysis for high field time series tailored using prior information on spatiotemporal structure. *Hum. Brain Mapp.*, in press.

Williams, P. 1991. *A Marquardt Algorithm for Choosing the Step-Size in Backpropagation Learning with Conjugate Gradients*, Cognitive Science Research Paper CSRP 229. University of Sussex, UK.

Ziehe, A., Muller, K., Nolte, G., Mackert, B., and Curio, G. 2000. Artifact reduction in magnetoneurography based on time-delayed second order correlations. *IEEE Trans. Biomed. Eng.* **47**(1): 75–87.