Independent component analysis of short-time Fourier transforms for spontaneous EEG/MEG analysis

Aapo Hyvärinen[*,a], Pavan Ramkumar[b], Lauri Parkkonen[b], Riitta Hari[b,c]

[a]*Dept of Mathematics and Statistics, Dept of Computer Science, HIIT, and Dept of Psychology,University of Helsinki, Finland*
[b]*Brain Research Unit, LTL, Helsinki University of Technology, Finland*
[c]*Advanced Magnetic Imaging Centre, Helsinki University of Technology, Finland*

Abstract

Analysis of spontaneous EEG/MEG needs unsupervised learning methods. While independent component analysis (ICA) has been successfully applied on spontaneous fMRI, it seems to be too sensitive to technical artifacts in EEG/MEG. We propose to apply ICA on short-time Fourier transforms of EEG/MEG signals, in order to find more "interesting" sources than with time-domain ICA, and to more meaningfully sort the obtained components. The method is especially useful for finding sources of rhythmic activity. Furthermore, we propose to use a complex mixing matrix to model sources which are spatially extended and have different phases in different EEG/MEG channels. Simulations with artificial data and experiments on resting-state MEG demonstrate the utility of the method.

Key words: Magnetoencephalography (MEG), independent component

[*] Contact address: Dept of Computer Science, P.O.Box 68, FIN-00014 University of Helsinki, Finland. Fax: +358-9-191-51120, Tel: +358-9-191-51415, email: Aapo.Hyvarinen@helsinki.fi

analysis, brain rhythms, resting state

1. Introduction

Recently, analysis of brain activity in resting state (Raichle et al., 2001; Kiviniemi et al., 2003; Beckmann et al., 2005; van de Ven et al., 2004), or during natural stimulation (Bartels and Zeki, 2004; Hasson et al., 2004) has attracted a lot of attention. Such experimental paradigms are a step towards more everyday-life-like recordings which are not constrained to responses to overly simplistic stimuli or tasks. However, in these experiments the computational or statistical analysis of the data is very challenging because there is no simple stimulus sequence to which the measured activity can be compared (correlated). Therefore, unsupervised or exploratory analysis methods have to be used.

Independent component analysis (ICA) and other blind source separation methods have been successfully applied for separating spatially independent sources in functional magnetic resonance imaging (fMRI) data measured in resting state (Beckmann et al., 2005; van de Ven et al., 2004) or during natural stimulation (Bartels and Zeki, 2004). However, the application of ICA on spontaneous electroencephalography (EEG) or magnetoencephalography (MEG) does not seem to be straightforward, and only few studies have successfully separated sources of spontaneous brain activity with ICA. Typically, ICA is very successful in finding artifacts (Jung et al., 2000; Vigário et al., 2000), but less successful in finding components related to brain activity.

We think that the main reason for this problem is that ICA essentially finds the components whose amplitudes have the most non-Gaussian distributions, while the interesting sources in EEG/MEG are often not very far from Gaussian. Consider amplitude-modulated oscillatory activity as illustrated in Figure 1. The amplitude distribution of the modulated sinusoid is determined by two conflicting properties. On the one hand, the amplitude distribution of the underlying oscillation (a sinusoid) has a strongly negative kurtosis (-1.5) since its histogram is bimodal. When such a distibution is modulated, it moves towards sparseness (Beale and Mallows, 1959), i.e. its kurtosis increases. However, when a negative kurtosis increases, it actually gets closer to zero and thus the distribution becomes more Gaussian, as is shown in the histogram of a modulated signal in Figure 1 d). The modulation has to be quite strong in order for the modulated signal to display a strong degree of non-Gaussianity: even the rather strong modulation shown in Figure 1 e) does not produce a large kurtosis (only 1.47 in this case).

Since ICA algorithms can be interpreted as maximizing the non-Gaussianity of the components, they are thus biased towards finding artifacts. This seems to explain why it is difficult to find oscillatory components with basic ICA, although such components are usually the main target of investigation in studies of spontaneous EEG/MEG.

Even if the degree of non-Gaussianity of oscillatory sources were strong enough for their separation by basic ICA, artifacts tend to be even more non-Gaussian. This is a problem because the number of independent components that needs to be estimated to recover oscillatory components can be quite large. Thus, we still need a method for selecting, among the many components estimated by ICA, the interesting ones, i.e. those related to brain activity. Non-Gaussianity would not be informative regarding the “interest-

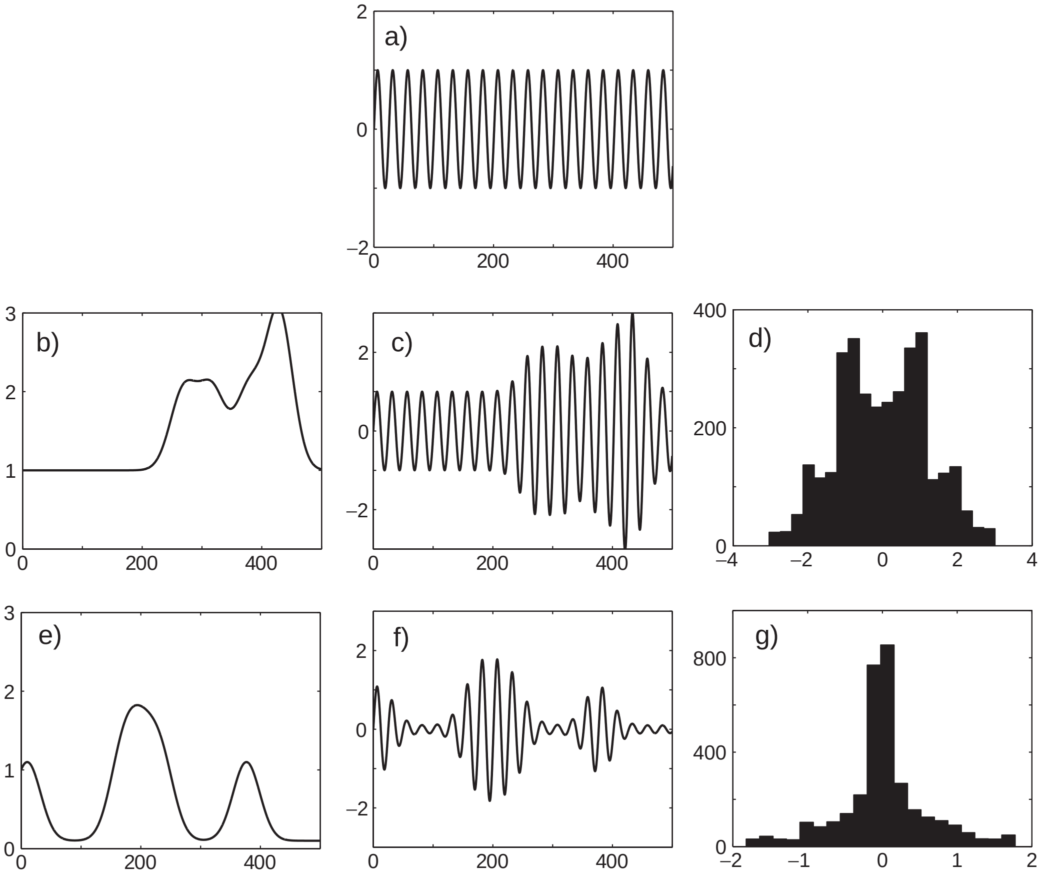

Figure 1: Illustration of amplitude-modulation of a sinusoid and its effect on the non-Gaussianity of the sources. a) A sinusoid, whose kurtosis is ≈ -1.5. b) An envelope which modulates the oscillation in a. c) The resulting modulated oscillation, i.e. a multiplied by b. d) The amplitude histogram of the modulated signal in c. It is not very non-Gaussian, which is reflected in its (normalized) kurtosis which equals -0.51. e) An envelope which has stronger fluctuations than the one in b. f) The resulting modulated oscillation, i.e. a multiplied by e. d) The histogram of the more strongly modulated signal in e. It is somewhat more non-Gaussian, which is reflected in its (normalized) kurtosis which is 1.47.

ingness" of the components. In fact, many blind source separation algorithms can be interpreted as finding sources which maximize some measure of "interestingness", and therefore separation of components and ranking them are essentially two viewpoints on the same problem: how to define a useful measure of "interestingness" which is also a valid separation criterion. Ideally, having a properly defined separation criterion, we can estimate just a subset of components which maximize the criterion of separation or interestingness, and sort the components by the value of the separation criterion, thus solving the separation and sorting problems at the same time.

Another problem with straightforward application of ICA to spontaneous EEG/MEG is that it assumes that each oscillatory source is observed in all the channels at the same phase (or with π radian phase shift, corresponding to just flipping the sign of the signal). While this is true for a single current dipole source, it would be interesting to find sources which are spatially more distributed. Such sources could consist of several dipoles whose time courses are closely correlated but with small time differences. They would presumably be seen in the different channels with constant phase differences; i.e. the oscillations in the different channels would be phase-locked but would not have the same phase. Basic ICA would split such sources into several components, while it would be useful to have them grouped in a single component.

Here, we propose a new method of blind source separation based on a measure of interestingness which is maximized by amplitude-modulated oscillatory activity, and allows for constant phase-differences in the oscillation. The method is based on the following three ingredients: 1) we use short-time

Fourier transforms of the data to probe oscillatory activity, 2) we use a complex mixing matrix to model different phases in spatially extended sources and 3) the analysis is performed using the complex-valued version of FastICA with a robust measure of non-Gaussianity (Bingham and Hyvärinen, 2000).

2. Model

2.1. ICA of short-time Fourier transformed data

We start by assuming a linear instantaneous mixing model like in ordinary ICA. Denote by $x_{c,\tau}$ the measured data where c is the channel index and τ is the time index. Each channel is assumed to be a linear superposition of underlying source signals (independent components) s_p

$$x_{c,\tau} = \sum_{p=1}^{P} a_{c,p} s_{p,\tau} \tag{1}$$

The sources are assumed to be statistically independent stochastic processes.

We propose to transform the EEG/MEG data in a very simple way: take short time-windows of the data, and replace each window by its Fourier transform. For regularly spaced time indices t_τ, we take windows $(x_{c,t_\tau}, x_{c,t_\tau+1}, \ldots, x_{c,t_\tau+w-1})$ where w is the size of the temporal window. We compute the Fourier transform for each such window, and basically replace the original data window with the Fourier transform; the precise definition of such replacement will be given in what follows.

This approach leads to data with the following three indices:

- c is the channel index, as above
- t is the index of the window, i.e. the temporal location within the experiments (with a lower temporal resolution than τ)

- f is the index of the Fourier transform coefficient *inside* the short time window.

The maximum values for the indices are denoted by C, T, F, respectively.

An important point is that in spite of the three indices we consider the data as an ordinary two-dimensional matrix in which each row (corresponding to a given channel c) contains the data for all the possible values of the t and f indices. Thus, our Fourier-transformed data consists of a matrix $\hat{x}_{c,tf}, c = 1 \ldots C, t = 1, \ldots, T, f = 1 \ldots F$. The fact that we collapse the t and f indices together is indicated by the lack of comma between the indices. In other words, we do *not* treat the transformed data as a three-dimensional object in the spirit of three-way data analysis methods (Miwakeichi et al., 2004); instead, we concatenate the short-time Fourier transforms one after the other, so we still have a two-dimensional data matrix.

A fundamental property of the linear mixing model in Equation (1) is that our transformation of the data does not in any way change the mixing model. This is a general property of linear time-filtering and similar transformations of the data (Hyvärinen et al., 2001). Thus, we have the following model for the Fourier-transformed data:

$$\hat{x}_{c,tf} = \sum_{p=1}^{P} a_{c,p} \hat{s}_{p,tf} \tag{2}$$

where the $a_{c,p}$ are the same mixing parameters as in Equation (1), and the $\hat{s}_{p,tf}$ are random coefficients, like in most source separation methods. Again, the $\hat{s}_{p,tf}$ are considered as a two-dimensional matrix so that the combined index tf replaces the time index τ in the original data. This means simply that their short-time Fourier transforms are concatenated one after the other.

Note that the data $\hat{x}$ are complex-valued by the conventional definition of the Fourier transform. The mixing coefficient $a_{c,p}$ are real-valued here, but in Section 2.3 we will further propose that the mixing coefficient are also allowed to take complex values to extend to linear mixing model.

The model in Equation (2) is quite similar to a complex-valued ICA model, whose estimation has been considered in the source separation literature.[1] For example, a variant of FastICA was proposed by Bingham and Hyvärinen (2000). In the following, we call this ICA of short-time Fourier transforms "Fourier-ICA" for short.

2.2. Sparseness and oscillatory activity

The important point here is that after the short-time Fourier transformation, the principles of ICA estimation can be interpreted in a new light. One

[1] Our model is, however, not equivalent to the complex-valued ICA model because the sources cannot be considered strictly independent in general. This is because of the influence of the frequency which is a sampling index in this representation. We can construct the following counterexample in which the original sources are independent but the transformed ones are not. Take two statistically independent narrow-band oscillatory sources s_1 and s_2 with the same peak frequency (say, 10 Hz). The short-time Fourier transforms have peaks in the same places, i.e. for the same combinations of the indices t and f. Thus, if we compute the covariance of the absolutes values of the $\hat{s}_1$ and $\hat{s}_2$, i.e. $\frac{1}{TF}\sum_{tf}|\hat{s}_1||\hat{s}_2| - \frac{1}{TF}\sum_{tf}|\hat{s}_1|\frac{1}{TF}\sum_{tf}|\hat{s}_2|$, this covariance is positive. This is in contradiction with independence, because for independent random variables, any nonlinear transformations are uncorrelated. Fortunately, this dependence seems to be weak, and does not seem to be significant in practice. Furthermore, this dependence can be considered as an artifact of our Fourier representation in which f is considered as a sampling index.

interpretation of ICA estimation is that it maximizes the non-Gaussianity of the sources; for most sources (those which are super-Gaussian or sparse) this means maximization of the sparseness.

What does sparseness mean for our transformed data? In fact, it means two different things, both of which are in line with our goal of finding sources of oscillatory activity:

1. Few frequency bands with non-zero energy. If the Fourier coefficients are zero for most frequency bands, then the distribution of the Fourier coefficients (taken over all frequencies and windows) is sparse.
2. Amplitude-modulation of signals. It is well-known (Beale and Mallows, 1959), and illustrated in Figure 1, that amplitude modulation of signals increases sparseness of the original signal. It also increases the sparseness for each single frequency band in the short-time Fourier transform, because the amplitude-modulation of the signals modulates the Fourier transforms as well (at least if it is slow enough so that the short-time Fourier spectrum is not significantly smeared by the modulation).

Thus, sparseness of the Fourier coefficients is a meaningful objective if we want to separate oscillatory signals, or if we want to find the most interesting sources among the separated ones.

Further, this means that the objective function which measures the "interestingness" of the source is given by the theory of ICA estimation, here the complex-valued case. We use the following objective function proposed by Bingham and Hyvärinen (2000):

$$J(\hat{s}_p) = \frac{1}{TF} \sum_{t,f} -\log(1 + |\hat{s}_{p,tf}|^2) \tag{3}$$

which can be interpreted as a measure of sparseness which is not sensitive to outliers. As we will see in the experiments below, it is important to use a measure of non-Gaussianity which is robust against outliers instead of, for example, the more classic kurtosis measure. The measure in Eq. (3) is robust, i.e. insensitive to outliers, because the logarithmic function does not grow fast when going far from zero, as opposed to the fourth power used in kurtosis. Here, it is assumed that the $\hat{s}_p$ is standardized to zero mean and unit variance. Maximization of this measure, and estimation of the complex-value ICA model are simultaneously performed by the complex-valued FastICA algorithm (Bingham and Hyvärinen, 2000).

We emphasize here that defining a good measure of "interestingness" is important to solve one practical problem in application of ICA: the method can give a large number of components, and it may be time-consuming for the researcher to go through all of them to find the useful components. If our objective function really corresponds to the interestingness of the components in our application, we can do two things: First, we can estimate a smaller number of components than the dimension of the data (even smaller than the dimension after PCA); second, we can rank the estimated components according to the objective function, and thus suggest which components should be considered for further analysis. We will see the utility of both of these ideas in the simulations below.

In contrast to basic ICA, our method thus uses information on the temporal correlations of the sources. A number of blind source separation methods have been developed based on such correlations (Belouchrani et al., 1997; Ziehe and Müller, 1998; Hyvärinen et al., 2001). Those methods can also separate Gaussian sources. However, they have the serious drawback that they cannot separate sources if at least two of the sources have identical temporal correlation structure, i.e. the same Fourier spectra.[2] This is why their direct application to spontaneous EEG/MEG seems weakly justified. In the analysis of spontaneous EEG/MEG, blind source separation methods (see Experiments below) typically find several sources with the same spectral characteristics, which implies that methods based on temporal correlations alone cannot be expected to yield any original sources in such a case, only arbitrary mixtures of such spectrally identical sources.

In contrast, our method combines temporal correlations (narrow-band characteristic) with amplitude modulations to provide a separation methods which can further separate spectrally identical sources. Thus, it attempts to combine the best properties of source separation methods based on non-Gaussianity and temporal correlations. In the Appendix, we provide a mathematical analysis of Fourier-ICA and show, in particular, that it can separate even Gaussian sources based on their temporal correlations alone.

2.3. Modelling phase differences by complex-valued mixing

We further propose that the coefficients $a_{c,p}$ are allowed to take complex values. The meaning of such a complex-valued model is that a source with index p can be observed in the channels with different phases, which are given by the phases of $a_{c,p}$ for different c, as will be explained next.

[2] Note that by the well-known Wiener-Khinchin theorem, if two signals have identical Fourier power spectra, then they also have identical temporal correlations structure (i.e. identical autocorrelation function), and vice versa. Strictly speaking, the theorem applies to stationary processes only.

Assume that in our original mixing model in Equation (1), the source oscillations are recorded in different channels with delays $\delta(c,p)$ which are small with respect to the period of the oscillation:

$$x_{c,\tau} = \sum_{p=1}^{P} a_{c,p} s_{p,\tau-\delta(c,p)} \tag{4}$$

We emphasize that we do not model delays due to the propagation of the signals, since such delays are extremely small. Instead, our goal is to model spatially extended sources of rhythmic activity, or distributed sources consisting of multiple focal sources with constant phase relations. Such sources (which could also be called networks) may result in oscillations with slightly different phases in different channels, if different parts of the source have phase-locked activity but with time differences. We do not propose a biophysical mechanism of how such phase differences arise, but our experiments reported below give some support to such a signal model.

It is well-known in the theory of Fourier analysis that the Fourier transform of a lagged signal $s_{\tau-n}$ is obtained by multiplication by a function of modulus one:

$$\hat{s}_{\tau-n}(f) = \hat{s}_{\tau}(f) \exp(2\pi i n f) \tag{5}$$

Furthermore, we assume that the sources are quite localized in the Fourier space, i.e., narrow-band. Then, the dependence of $\exp(2\pi i n f)$ on f has little effect because we have

$$\hat{s}_{\tau}(f) \exp(2\pi i n f) \approx \hat{s}_{\tau}(f) \exp(2\pi i n f_0) \tag{6}$$

where f_0 is the dominant frequency of an oscillatory source $s(\tau)$. When we take the short-time Fourier transform of the lagged mixing model in (4), we thus have

$$\hat{x}_{c,tf} = \sum_{p=1}^{P} a_{c,p} \exp(2\pi i f_0(p)\delta(c,p))\hat{s}_{p,tf} \tag{7}$$

Here, we can combine the original real-valued mixing coefficients $a_{c,p}$ and the new terms $\exp(2\pi i f_0(p)\delta(c,p))$ as new complex-valued mixing coefficients, denote them by $a'_{c,p}$. Thus, the mixing model in the Fourier domain in Equation (2) can be approximated by a model which is formally exactly the same as Equation (2), but the mixing coefficients $a_{c,p}$ are *complex-valued.*

It should be emphasized that this approximation is only valid for narrow-band sources. In the general case, we would need a separate phase parameter for each frequency band (Anemüller et al., 2003), which would greatly increase the number of parameters in the model (10-fold even if we only had 10 frequency bands), and thus make its estimation more difficult. Since most sources of interest in spontaneous EEG/MEG are relatively narrow-band, this approximation seems reasonable, and it has the benefit of keeping the number of parameters quite low.

As we can see in Equation (7), the magnitudes of the complex-valued mixing coefficients $a'_{c,p}$ are the same as those of the original coefficients $a_{c,p}$ since $\exp(2\pi i f_0(p)\delta(c,p))$ has modulus one and only changes the phase (argument) of the complex coefficient.

As is well-known in the theory of ICA, the global phase or magnitude of a component cannot be estimated in the model. This is because we can always multiply any source $\hat{s}_p$ by a scalar constant q_p, and divide the coefficients $a_{c,p}$ by the same q_p, and the observed data are not changed in any way. Thus, all the estimated magnitudes and phases of each source are only defined relative to each other, i.e. we can only estimate the phase *differences* between $a_{c,p}$

and $a_{c',p}$, or their *relative* magnitudes.

2.4. Reliability analysis

It is also useful, although not necessary, to incorporate in the method an algorithmic reliability analysis as in (Himberg et al., 2004). The estimation is repeated for many times for each data set, using either different random samples of time windows or different random initial points, or both. This allows an analysis of the reliability of the estimates: a minimum requirement for a component to be reliable is that it is obtained in (almost) all these randomized runs.[3] If a component is obtained only a few times, it may be an algorithmic artifact, possibly a small local maximum of the objective function, or a purely random effect due to a small sample size. It is important to investigate this possibility when one uses data-analysis algorithms based on maximization of complicated objective functions, since function maximization can be very difficult and unreliable.

3. Simulations and Experiments

We applied the method on three different kinds of simulated data, as well as on real MEG data.

[3] This rather primitive quantification of reliability is the best that we have in the theory of ICA so far. Hopefully, future research will provide methods which quantify reliability in a more principled manner, for example, in terms of statistical significance (p-values). See Groppe et al. (2009) for related work.

3.1. General methods

To make the simulations as realistic as possible, we processed the simulated data almost exactly the same way as the real MEG data. The hypothetical "sampling frequency" in the simulations was 150 Hz, which was the sampling frequency of the real MEG data after some preprocessing independent of the present method.

All datasets were processed by the following steps:

1. The data were downsampled to 75 Hz using Matlab's `decimate` function which includes an interpolation.
2. Windows of a duration of 1 second were taken, half overlapping (i.e. with an interval of 0.5 s between the starting points of the windows).
3. A Fast Fourier transform (FFT) was performed on each window. Only coefficients in the range from 5 Hz to 30 Hz were retained, i.e. the data were effectively band-pass filtered.
4. Outliers were rejected by computing the logarithms of the norms of each FFT-transformed window, and rejecting any window whose log-norm was larger than the mean plus three standard deviations. (This was only done with real MEG data.)
5. The means of the Fourier-transformed channels were subtracted to make all signals zero-mean.
6. The dimension of the data was reduced by principal component analysis (PCA) to 25, except in Simulations 1 and 2, where no dimension reduction was performed because the initial data dimension was smaller than 25. The choice of this dimension was rather ad hoc, obtained by trying out a few different values in pilot experiments.

7. Fourier-ICA was performed using the complex-valued FastICA algorithm (Bingham and Hyvärinen, 2000) with the objective function in Equation (3). The number of components estimated was typically smaller than the PCA dimension.

As part of the FastICA algorithms, the data are whitened; however, we do not consider that a part of the preprocessing but rather a part of the estimation method itself.

For comparison, we also applied the ordinary (time-domain) FastICA algorithm on the datasets (Hyvärinen, 1999). In this case, steps 2 and 3 above were omitted. As a replacement to the band-pass filtering in step 3, we performed a corresponding time-domain filtering of the data. Outliers were detected in step 4 in each 1-s window.

3.2. Simulation 1: Validation of objective function

In the first simulation, we compared Fourier-ICA with basic (time-domain) ICA using artificial data, and evaluated the effect of using robust vs. non-robust non-Gaussianity measures.

3.2.1. Methods

We synthetized three source signals which resemble rhythmic activity, and another three source signals which resemble artifacts encountered in MEG data. The source signals had a length of 10,000 time points, and they are shown in Figure 2. They were preprocessed as described in Section 3.1.

We computed the values of the robust objective function of Fourier-ICA, in Equation (3), of these source signals. Before the computation, each source signal was preprocessed as described above (steps 1–6). For comparison, we computed the values of objective functions of basic (time-domain) ICA, using the objective functions of the form $\sum_t G(s_t)$ where the sum is taken in the time-domain only. Two different forms of G were used: First, the function G was set to the (negative) log cosh function, which is used in FastICA (Hyvärinen, 1999) with the nonlinearity "tanh" as well as in the infomax algorithm (Amari et al., 1996; Bell and Sejnowski, 1995). Second, we set G to be the fourth power, which is equivalent to using kurtosis in of FastICA (Hyvärinen and Oja, 1997). This choice corresponds to the nonlinearity "pow3" or third power, and is also used in a number of other source separation methods including JADE (joint approximate diagonalization of eigenmatrices) by Cardoso and Souloumiac (1993). Furthermore, we computed the kurtosis of the Fourier-transformed data, which is also a valid objective function in complex ICA. In all these computations, the preprocessed source signals were normalized to unit variance.

Since the maximization of ICA objective functions is computationally quite difficult, the computed values of the objective function do not strictly determine the actual behaviour of source separation algorithms. This is why we made source separation simulations with the corresponding four algorithms (Fourier-ICA using robust objective or kurtosis, and time-domain FastICA with tanh or kurtosis/pow3) using the different objective functions. For this purpose, we mixed these signals with random (real-valued, normally distributed) coefficients $a_{c,p}$, and then applied the three algorithms on the mixed data. To have a more reliable assessement of the algorithms, we also re-generated the random parts (noise, and locations of artifact spikes) of the source signals, as well as the mixing coefficients, and re-ran the algorithms

100 times. The algorithms were told to find three source signals (with the largest values of the objective function). The results were analyzed by considering the product of the estimated separating matrix and the true mixing matrix, normalizing each row of this matrix to unit norm, and counting the number elements in the matrix which were larger than 0.95 in absolute value; this we considered the number of source signals successfully separated.

3.2.2. Results and Discussion

The values of the objective functions are shown in Figure 3. We can see that for Fourier-ICA using the robust measure in Equation (3), the oscillatory ("brain") source signals have larger values of the objective function as compared to the artifactual signals. In contrast, for the basic (time-domain) ICA objective functions, the artifactual signals take larger values, which is also the case for Fourier-ICA using kurtosis. Thus, our proposed objective function is able to better indicate which source signals are likely to be oscillatory: Oscillatory signals have relatively higher values of the robust objective function (higher sparseness) than artifactual signals. In contrast, kurtosis is quite sensitive to artifacts even in the Fourier domain.

The results of the corresponding source separation simulations are summarized in Figure 4. We see that Fourier-ICA using (3) quite often found oscillatory brain signals (243 out of the maximum possible 300, or 77%), whereas the time-domain methods preferred artifactual signals, finding approximately 17% and 11% of brain signals. The performance of Fourier-ICA based on kurtosis was between these two cases. Thus, also on a practical level, our new method seems to be efficient in finding the oscillatory brain signals and ignoring artifacts.

The performance is not 100%, but in practice one can run the algorithm many times to circumvent this problem, as explained in Sections 2.4 and 3.5.1. Alternatively, one can estimate the full set of independent components and then use the objective function to sort them; results in Figure 3 suggest that in this case the performance would be close to 100%.

3.3. Simulation 2: Modelling phase differences by complex-valued mixing

In simulation 2, we investigated the utility of using a complex-valued mixing matrix to recover source signals when the sources are distributed, leading to phase differences (delays).

3.3.1. Methods

We used the three oscillatory source signals in Simulation 1, depicted as the three first source signals in Figure 2.

First, a real-valued random 3×3 mixing matrix was generated from normally distributed variables. Then, a 3×3 matrix of delays were randomly generated by uniformly sampling integer delays between 0 and 7. The maximum delay of 7 samples corresponds to 93 ms which is close to the period of the 10-Hz oscillations, i.e. 100 ms. Again, the length of the signals was 10,000 time points.

The data were preprocessed as in Section 3.1. Source separation was performed by Fourier-ICA using the robust measure in (3) in two conditions: estimating a real-valued mixing matrix, or a complex-valued one.

The evaluation of separation is a bit more involved because the algorithm gives phases as complex numbers whereas the data were generated using a real-valued mixing matrix with delays. Thus, we cannot directly compare

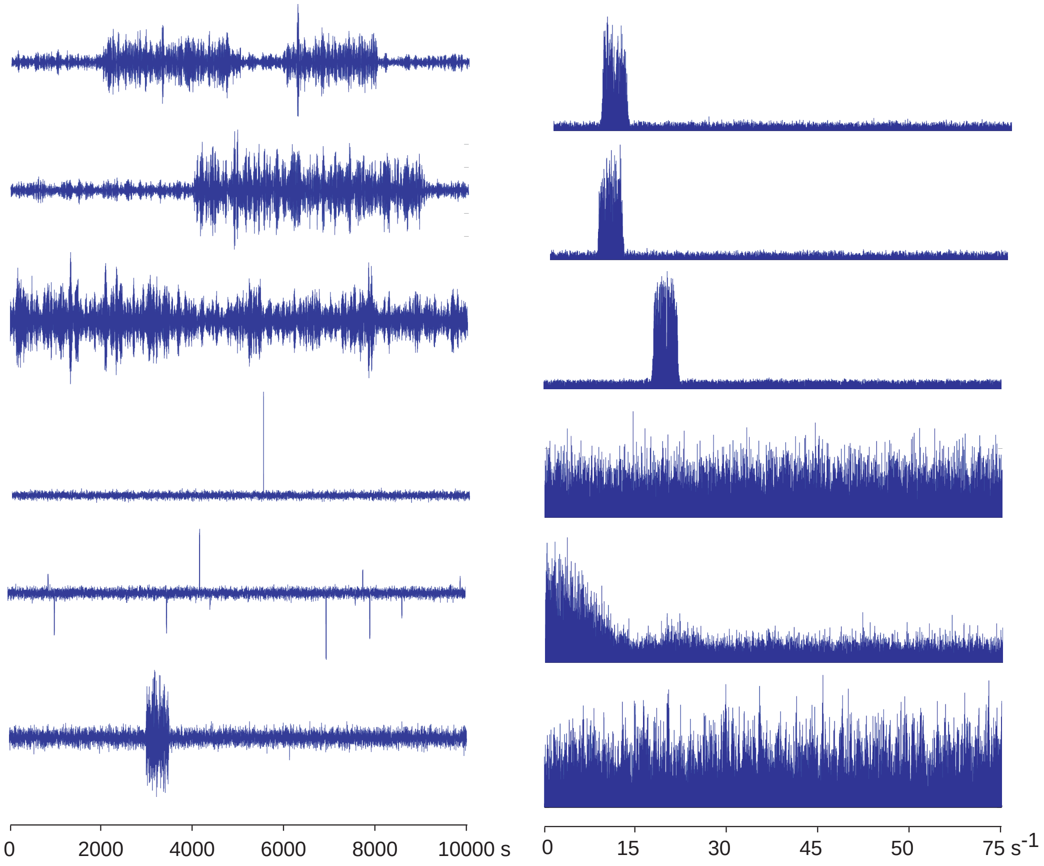

Figure 2: The six simulated source signals used to compare different objective functions. Left: the original signals in time domain, Right: their Fourier amplitudes. Horizontal scales are sample index and Hz, respectively; units on vertical scale are arbitrary. The first three source signals are amplitude-modulated band-pass signals resembling brain oscillations, with carrier frequencies of 10, 10, and 20 Hz. The latter three source signals resemble artifacts: a large spike artifact, several spikes, and a muscular artifact consisting of a short period of strong gaussian noise, respectively.

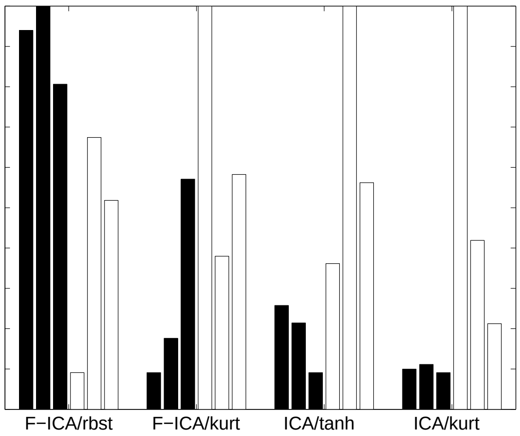

Figure 3: Simulation 1: The values of the objective functions for Fourier-ICA vs. FastICA and robust vs. non-robust sparseness measures. F-ICA/rbst: Fourier-ICA with robust measure in Equation (3), F-ICA/kurt: Fourier-ICA with kurtosis, ICA/rbst: ICA with nonlinearity tanh, ICA/kurt: ICA with kurtosis. For each method, the three black bars give the objective function values for the brain source signals and the three white bars give the values for the artifactual source signals. Source signals are in the same order as in Figure 2. The actual scale is different for each method, but since the actual values do not matter, the values for each method have been rescaled to a common scale from 0.1 (smallest value for the method) to 1 (largest value for the method).

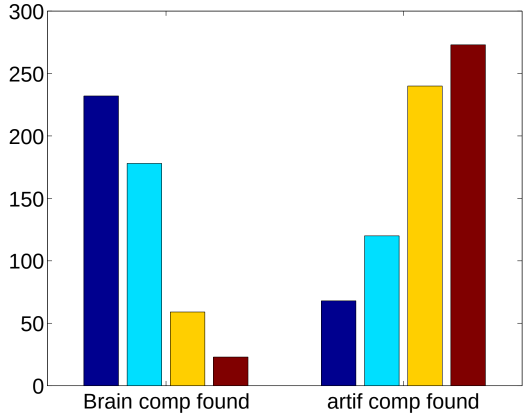

Figure 4: Simulation 1: Separation results for Fourier-ICA using robust objective in (3) or kurtosis, and FastICA with the nonlinearities tanh and pow3 (i.e. kurtosis). The bars show the numbers of brain components and artifact components found over 100 repetitions. Note that the algorithm was told to find only three components in each trial. Blue: Fourier-ICA using the robust objective function in Eq. (3). Cyan: Fourier-ICA with kurtosis. Yellow: Basic (time-domain) ICA with tanh. Red: Basic ICA with kurtosis (pow3)

the estimated mixing matrix with the original one. A simple way of comparing the mixing matrices is to compare the *absolute values* of the estimated and original mixing matrices, ignoring the comparison of estimated phases vs. original delays, which is not straightforward. Thus, we computed the product of the inverse of the absolute values of the estimated mixing matrix with the absolute values of the original mixing matrix. This should be close to a permutation matrix if separation is successful, and the closeness to a permutation matrix can be quantified as in ordinary blind source separation trials: We use the same method as in Simulation 1 to compute the number of successfully separated components.

Furthermore, we computed the correlation coefficients of the estimated source signals and original source signals in the Fourier domain. These correlation coefficients tell us how well the source signals themselves were estimated. Again, we thresholded their absolute values as in Simulation 1 to compute the number of separated components.

3.3.2. Results and Discussion

Based on the absolute values in the mixing matrix, the percentage of correctly estimated components was 85% in the case of a complex-valued mixing model, and 42% in the case of a real-valued one. Using the evaluation based on correlation coefficients, the percentages were 93% and 34%, respectively.

Thus, using a complex-valued mixing matrix greatly enhanced the separation capability in the case of distributed source signals. The separation capability was not 100% presumably because modelling delays with a complex-valued mixing matrix is only an approximation. Some of the errors are, of course, due to the finite data length as well.

3.4. Simulation 3: Oscillatory current dipoles and real MEG noise

In Simulation 3, we mimicked measurements from an MEG device. Brain sources were simulated by a set of current dipoles, and real noise was added.

3.4.1. Methods

We simulated data from a 306-channel MEG device (Elekta Neuromag Oy, Helsinki, Finland) comprising 102 magnetometers and 204 planar gradiometers in a helmet-shaped array. A standard forward model with a spherical volume conductor was used (Sarvas, 1987). Three cortical current dipoles were defined as follows: The time courses were oscillations at constant frequencies, amplitude modulated by different smoothed boxcar functions. Before modulation, Gaussian noise was added to the boxcar function with a signal-to-noise ratio of 1. The locations (approximate) and frequencies of the dipoles were: the right sensorimotor cortex at 10 Hz, the left sensorimotor cortex at 19 Hz and the right visual cortex at 10 Hz. The length of the simulated data was 120 s, and the sampling rate was 150 Hz.

To obtain realistic noise, 120 s of data measured from the empty magnetically shielded room with the 306-channel MEG device was added on the signals generated by the simulated current dipoles with a signal-to-noise ratio of 1.

The reliability analysis (see Section 2.4) was done according to the ICASSO framework (Himberg et al., 2004) (Simulations 1 and 2 included no reliability analysis because in completely artificial data all components are usually reliable). The default settings were used except that we used the complete-linkage strategy in the hierarchical clustering. The reliable components were selected using the stability index I_q defined by Himberg et al. (2004). It is an index between 0 and 1 which measures the reliability of a component. In our experiment with real MEG data, only components for which this index was larger than 0.75 were included in the analysis.

The components considered reliable by the randomization were finally sorted (ranked) according to the values of the corresponding objective function, large values first.

To visualize the time courses of the sources, we computed the norm of the source in each 1-s window. We also computed the average Fourier spectra of each source by averaging across time windows. To visualize the spatial distribution of the sources, we computed the sum of squares of the real and imaginary parts of the $a_{c,p}$ corresponding to the two orthogonal planar gradiometer channels in the same spatial location. This gives a measure of the spatial "weight" of a source in a given sensor location, and we plot it on a topographic helmet. All these quantities are in arbitrary units since each component can only be estimated up to a global scaling factor (Hyvärinen et al., 2001).

Since Fourier-ICA with kurtosis had poor performance in Simulation 1, we did not consider it anymore. We only used Fourier-ICA with the robust measure of non-Gaussianity. As above, we used for comparison, basic ICA with the two nonlinearities (tanh and kurtosis/pow3). Since the data dimension is large even after PCA, and one of our goals is to reduce the number of estimated components by concentrating on the most interesting ones, the number of components to be estimated was set to 5.

To further compare with other ICA algorithms than FastICA, we applied the JADE algorithm (Cardoso and Souloumiac, 1993) to the same Fourier-transformed data, using a complex-valued mixing matrix. To compare with source separation methods based on temporal correlations, we applied the SOBI (second-order blind identification) blind source separation method (Belouchrani et al., 1997) on the data (without Fourier transform), with real-valued mixing matrix, and 500 time lags. These two methods do not allow separation of a smaller number of components, so we had to estimate a full set of 25 components and develop a criterion for choosing the most "interesting" ones. For JADE, we used kurtosis, since the method is based on higher-order cumulants. For SOBI, we used the sum of squares of autocorrelations of a source, which is a relatively heuristic measure of the total amount of temporal structure in a source. No reliability analysis was performed because these methods are deterministic (at least in the authors' implementation, which we used here[4]).

3.4.2. Results and Discussion

The results for the different algorithms (Fourier-ICA, basic ICA with tanh, basic ICA with kurtosis, Fourier-ICA with JADE, SOBI) are shown in Figures 5–9. The reliability analysis found four reliable sources. Fourier-ICA found the three sources corresponding to the simulated current dipoles and ranked them as the first three, showing that the method found exactly what it was supposed to. ICA with tanh found only one of the current dipoles and failed to rank even that one as the first one. ICA with kurtosis found none of the current dipoles.[5]

[4] `http://www.tsi.enst.fr/~cardoso/stuff.html`

[5] In fact, FastICA with kurtosis often failed to converge, which sometimes happens when the data are very far from the specifications of the ICA model.

Fourier-ICA using JADE instead of FastICA was also successful in finding all the three oscillatory components. SOBI found one of the oscillatory components but was not able to separate the 10-Hz oscillations from each other, as predicted by the theory of second-order separation discussed above. (Here we show the four signals ranked most interesting; we give the same number of signals as the reliability analysis indicated with the preceding methods.)

All the methods also found a physical artifact of unknown origin, e.g. signal #4 in Figure 5.

Thus, the index of interestingness in Fourier-ICA was successful in both finding and ranking the sources, while basic (time-domain) ICA was not able to separate most of them properly, and ranking was not successful either. The performance of SOBI clearly showed the limitations of methods based on temporal correlations only. Fourier-ICA with JADE performed equally well as FastICA with the robust measure in this simulation.

3.5. Experiments on real MEG data

Finally, we applied the new method on real resting-state MEG data.

3.5.1. Methods

The raw data consisted of 5 minutes of the 306 MEG signals obtained from a healthy volunteer resting eyes closed (Ramkumar et al., 2007). The subject participated after informed consent, and the MEG recordings had a prior approval by the Ethics Committee of the Helsinki and Uusimaa Hospital District. The initial sampling frequency was 600 Hz. The signal space separation method (Taulu et al., 2004) was used to reduce noise, and the data were downsampled to 150 Hz. Magnetometer channels were exluded

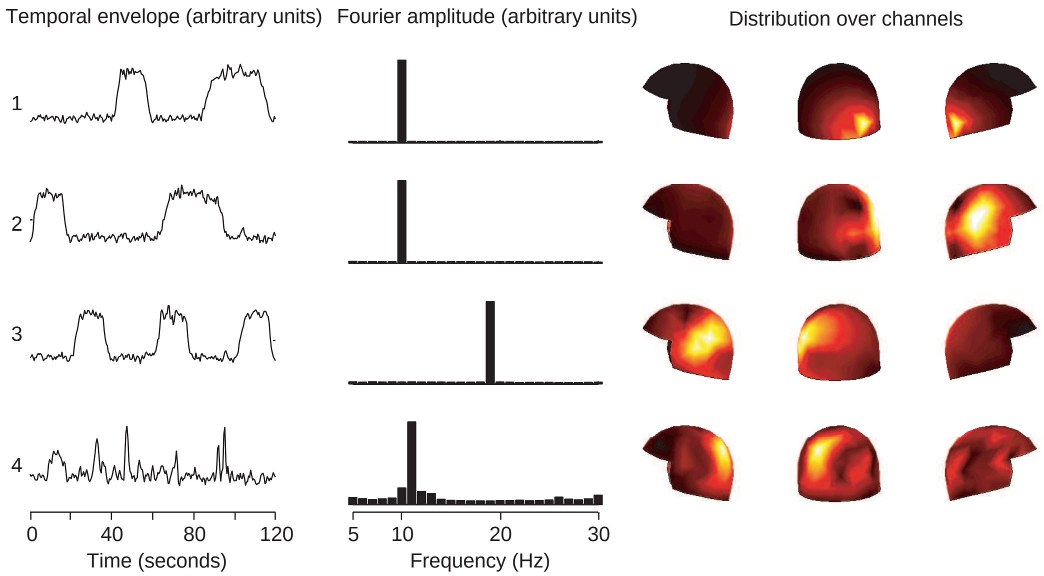

Figure 5: Simulation 3: The reliable sources found by Fourier-ICA sorted according to the criterion in Eq. (3). Left: time courses (envelopes), middle: power spectra, right: spatial distributions.

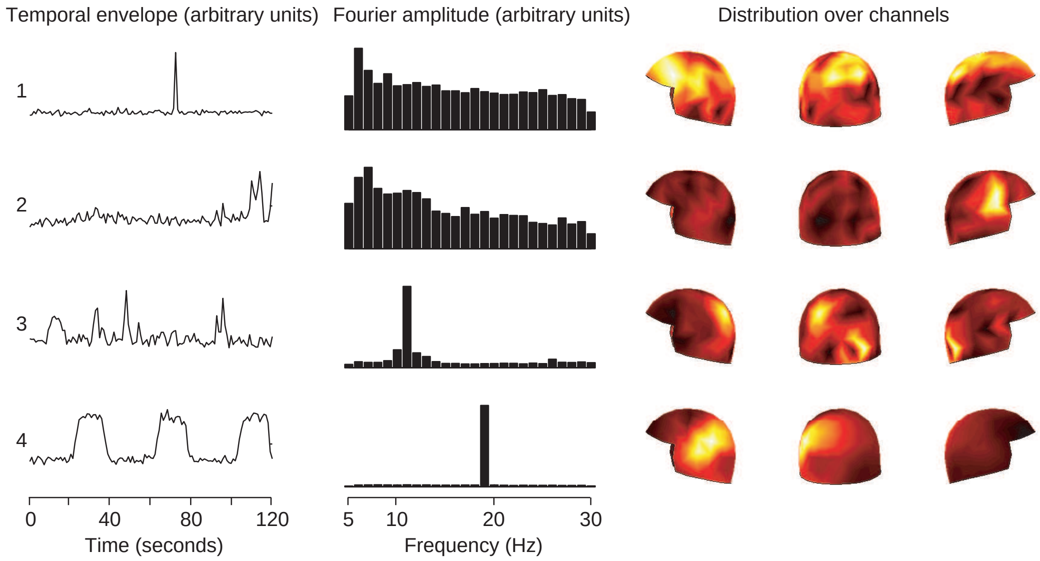

Figure 6: Simulation 3: The reliable sources found by basic ICA and nonlinearity tanh, sorted according to the corresponding objective function. Left: time courses (envelopes), middle: power spectra, right: spatial distributions.

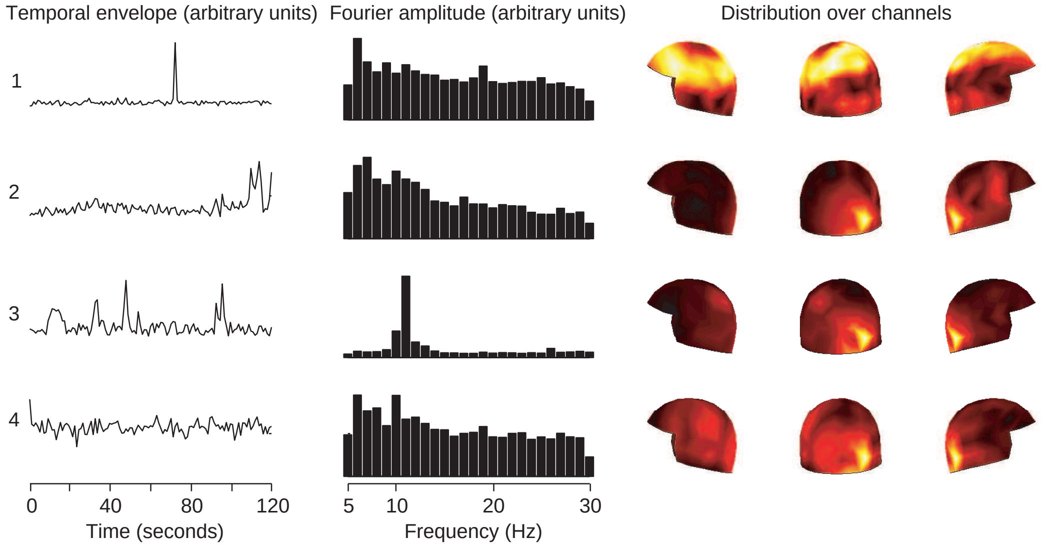

Figure 7: Simulation 3: The reliable sources found by basic ICA and kurtosis, sorted according to the values of kurtosis. Left: time courses (envelopes), middle: power spectra, right: spatial distributions.

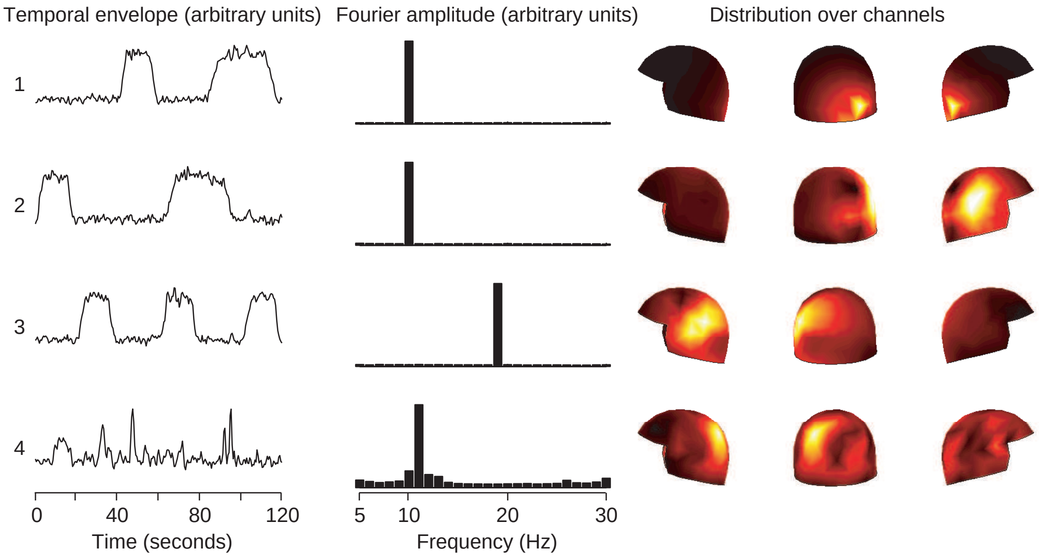

Figure 8: Simulation 3: The most "interesting" sources found by Fourier-ICA using JADE instead of FastICA, sorted according to the values of kurtosis. Left: time courses (envelopes), middle: power spectra, right: spatial distributions.

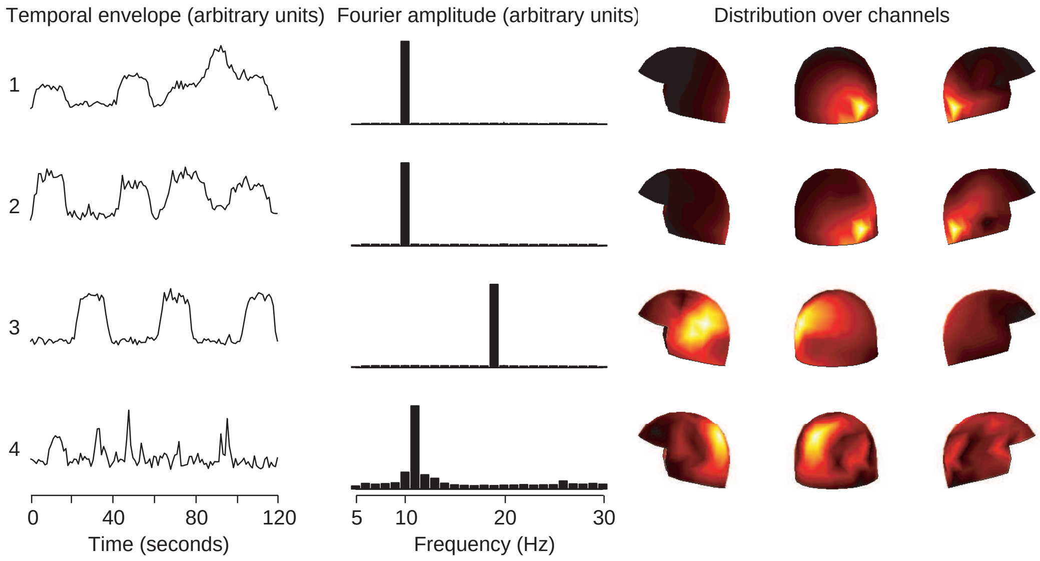

Figure 9: Simulation 3: The most "interesting" sources found by SOBI, sorted according to the amount of temporal structure. Left: time courses (envelopes), middle: power spectra, right: spatial distributions.

from the analysis due to their wide-spread lead fields, leaving the 204 planar gradiometer channels.

The data were analyzed as described above in Section 3.1. The number of estimated components was set to 10; since our goal is to develop a method which directly finds the most interesting sources, this number was chosen to be significantly smaller than the number of dimensions after PCA.

Fourier-ICA with the robust measure was performed on this data. Again, for comparison, basic (time-domain) ICA was also applied to the same data, with the two nonlinearities tanh and pow3 (kurtosis), and for further comparison, SOBI and Fourier-ICA with JADE were performed. Reliability analysis, sorting of the components, and visualization were performed as described in Section 3.4.

The phase differences were visualized as follows. As noted above, an important indeterminacy in ICA is that for each source signal, the global phase cannot be estimated, i.e. each column $a_{\cdot,p}$ is estimated up to a phase rotation. This is not a problem since we are only interested in phase *differences* anyway, but for visualization, we have to fix that phase rotation for each source signal. Here, we rotate the phases as close to zero as possible, and thus we can interpret them directly as the difference from a global average phase. A second problem is that this approach still gives two phase difference values for each sensor location (which have two gradiometers each). We choose simply to plot the one which has the larger absolute value.

3.5.2. Results and Discussion

Components found by Fourier-ICA are shown in Figures 10–11, and components found by basic (time-domain) ICA are shown in Figures 12–13.

Based on visual inspection, all components found by Fourier-ICA seem to be physiologically meaningful and not biological or technical artifacts. We see two rhythmic mu components (#2 and #3 in Figure 10, and #1 and #4 in Figure 11) and a number of approx. 10-Hz components.

Using a complex-valued mixing matrix (Figure 10) does lead to recovery of sources which have time lags, as was our hypothesis. The complex-valued mixing matrix seems to produce "cleaner" results than using a real-valued mixing matrix (Figure 11) in the sense that the source signals correspond to spatially more contiguous regions. This can be seen by comparing the topographic plots in Figs. 10 and 11.

Many of the components found by basic ICA seem to be artifacts, based on visual inspection of their topographic distributions, time courses, and frequency contents. Some of them are ranked high in interestingness by their objective function. In fact, among the four components ranked most interesting (top rows), two seem to be artifacts (#1 and #2 in Figure 12 and #1 and #4 in Figure 13). Only one mu component seems to be found (#5 in Figure 12, #7 in Figure 13).

The JADE ICA algorithm (Fig. 14) failed to clearly separate even both mu rhythms, instead finding many artifacts. Thus, it seems necessary to use a robust nonlinearity in the ICA part of Fourier-ICA, whereas JADE is based on non-robust higher-order cumulants. SOBI (Fig. 15) gave components which seem to be, based on visual inspection, rather similar to those of Fourier-ICA. However, the theoretical results and Simulation 3 indicate that components with similar spectral contents are not likely to be successfully separated, and many of the components here have a spectrum which is almost identical to the spectrum of some other component (in fact, among the remaining 15 sources not shown here, many more similar spectra are likely to be found). Thus, the results of SOBI cannot be trusted.

To conclude, Fourier-ICA, when used with a robust measure of non-Gaussianity, seems to find mainly sources of rhythmic activity, whereas other methods concentrate to a large extent on artifacts. A different problem prevents the application of SOBI: The basic statistical assumption allowing separation by SOBI is violated with this data.

4. General discussion

We proposed to apply ICA on the (complex-valued) short-time Fourier transforms of EEG/MEG. When combined with a robust measure of non-Gaussianity, the method finds sources which are maximally narrow-band and/or amplitude-modulated. Simulations with artificial data and experiments with real MEG data show that such a method is able to separate sources of rhythmic brain activity better than basic ICA, or second-order blind source separation methods. These results are in contrast to most existing literature which has used ICA merely to remove artifacts from EEG and MEG, and separation of spontaneous brain activity into source signals has not been very successful. Moreover, our method provides a principled way of ranking the components, so that the analyst may not need to go manually through dozens or even hundreds of components to find the interesting ones.

It is well-known in the theory of ICA that one can apply a wavelet transform or a related time-frequency decomposition to the data before ICA. Zibulevsky and Pearlmutter (2001) proposed to perform a time-frequency decomposition with the specific goal of making the data and the components sparser. Anemüller et al. (2003) used a short-time Fourier transform of EEG data in a rather similar framework to ours, but in the context of evoked potentials. A related application on fMRI was presented by Anemüller et al. (2006). Further work on convolutional models for EEG signals is by Dyrholm et al. (2007). However, application of such methods to improve source separation in spontaneous EEG/MEG seems to be lacking.

Although ICA, in theory, can separate any independent components which are mixed according to the assumptions of the model, the statistical performance with real data is strongly affected by the model for the signals. Theoretically, such effects could be seen in the asymptotic variance of the estimators. In practice, such analysis may not be very relevant because most of the errors in practical analysis may be due to violations of the model assumptions such as independence and linearity, and the effect of such violations is quite difficult to analyze. Intuitively, however, it seems reasonable that if the signal model is richer, in our case capturing the oscillatory time structure in addition to the amplitude distribution, the method is likely to be better in separating sources. Perhaps a more useful way of analyzing such improvement is to use sophisticated simulations related to the application domain considered, and real data for which the expected results are known to some degree. This was our approach in this article. At the same time, one has to remember that any separation algorithm is biased towards finding certain kinds of sources. We have argued here that basic ICA is biased towards signals with non-Gaussian amplitudes, and thus towards artifacts. Fourier-ICA is also biased towards certain sources, for example those which are narrow-band, and this bias has to be kept in mind when interpreting the results.

Our results on real MEG data suggested that Fourier-ICA is able to decompose rhythmic brain activity into components in a new way. In particular, the method recovered a large number of components for the prominent approx. 10 Hz activity. The significance of such findings needs to be confirmed by experiments with several subjects. On the other hand, based on visual inspection, modelling delays with a complex mixing matrix seems to improve the results, although such improvement is hard to quantify since we do not know the ground truth, i.e. the actual structure of the underlying sources in the real MEG data. This is another important topic for future research.

Investigations of EEG/MEG recordings in resting state, or with natural stimulation, have not advanced nearly as much as the corresponding fMRI analyses. One of the main reasons may have been the lack of good separation methods. We hope that the method proposed here is an important step in that direction.

References

Amari, S.-I., Cichocki, A., Yang, H., 1996. A new learning algorithm for blind source separation. In: Touretzky, D. S., Mozer, M. C., Hasselmo, M. E. (Eds.), Advances in Neural Information Processing Systems 8. MIT Press, Cambridge, MA, pp. 757–763.

Anemüller, J., Duann, J.-R., Sejnowski, T. J., Makeig, S., 2006. Spatio-temporal dynamics in fMRI recordings revealed with complex independent component analysis. Neurocomputing 69, 1502–1512.

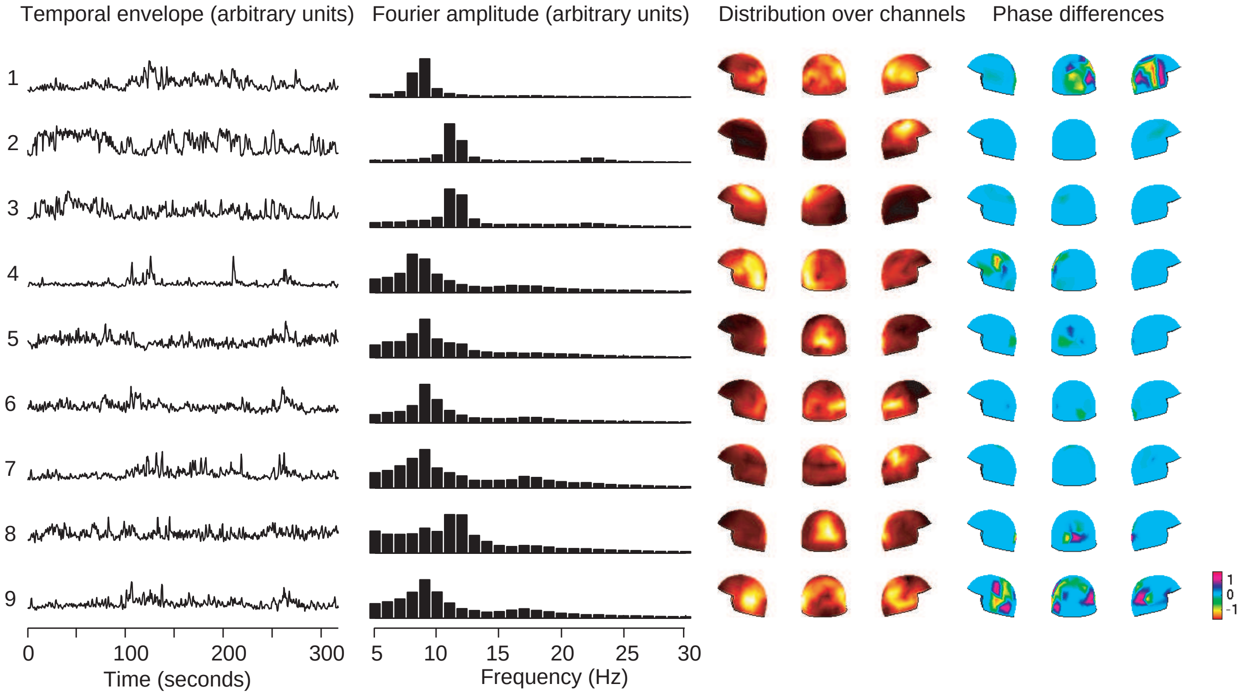

Figure 10: Real MEG recording: Reliable sources obtained by Fourier-ICA, with a complex mixing matrix. Each row is one source; the sources are ordered by the objective function, i.e. the most “interesting” ones in the top. First column: the time course of the source (envelope). Second column: the Fourier amplitude spectrum. Third column: The magnitudes of the sources at the topographic sensor helmet. Fourth column: The phase differences of the sources. Cyan means zero phase difference (see colour bar). Phase differences are only plotted for channel locations in which the magnitude was large.

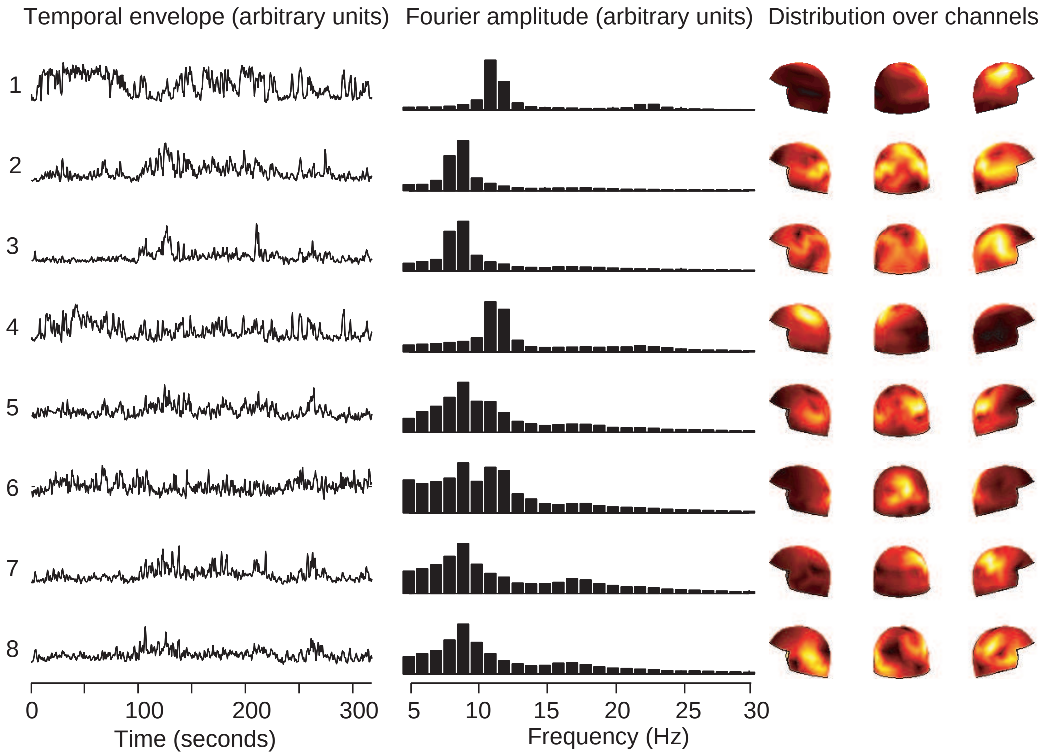

Figure 11: Real MEG recording: The reliable sources obtained by Fourier-ICA, with a *real-valued* mixing matrix. Left: time courses (envelopes), middle: Fourier spectra, right: topographic distributions.

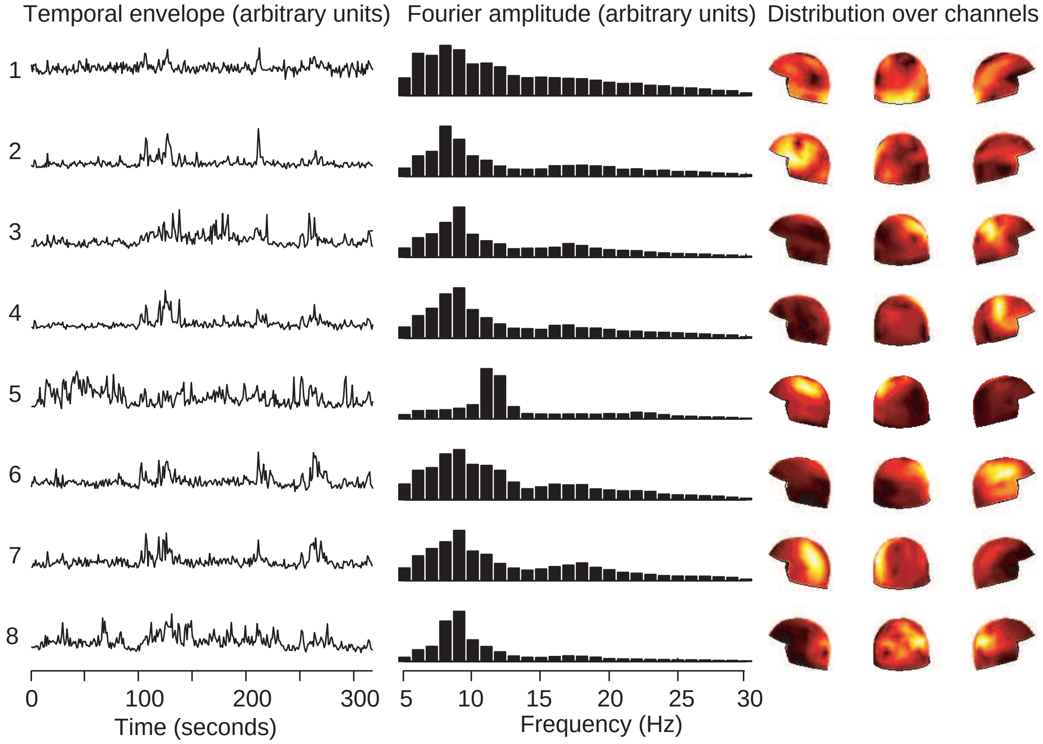

Figure 12: Real MEG recording: The reliable sources obtained by basic ICA, with *tanh* nonlinearity. Left: time courses (envelopes), middle: Fourier spectra, right: topographic distributions.

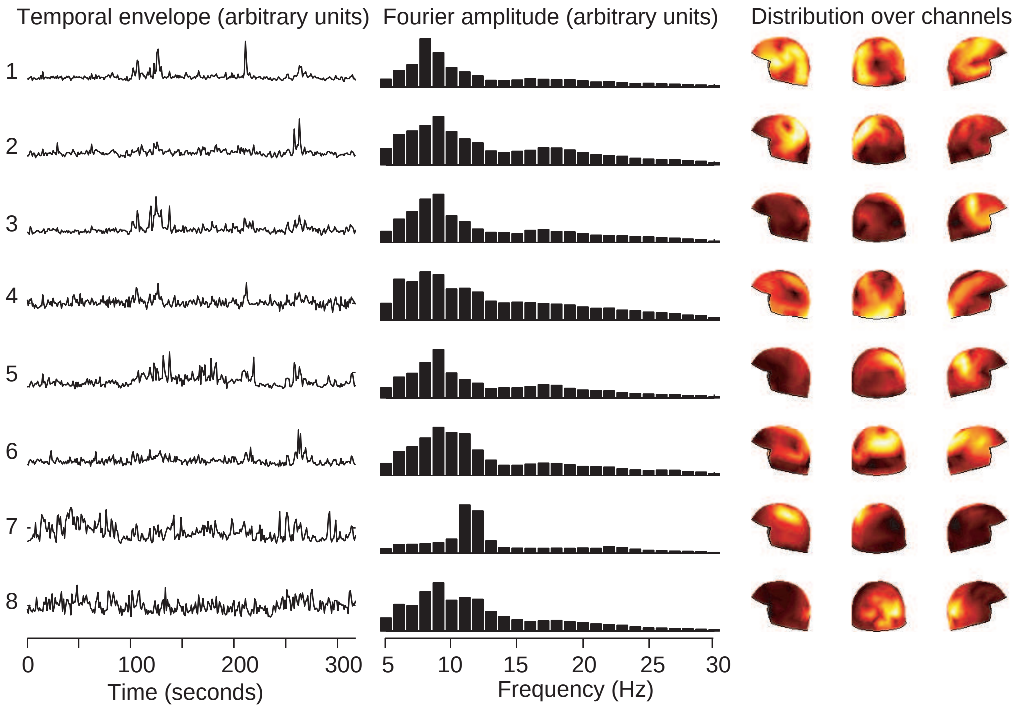

Figure 13: Real MEG recording: The reliable sources obtained by Basic ICA, with kurtosis. Left: time courses (envelopes), middle: Fourier spectra, right: topographic distributions.

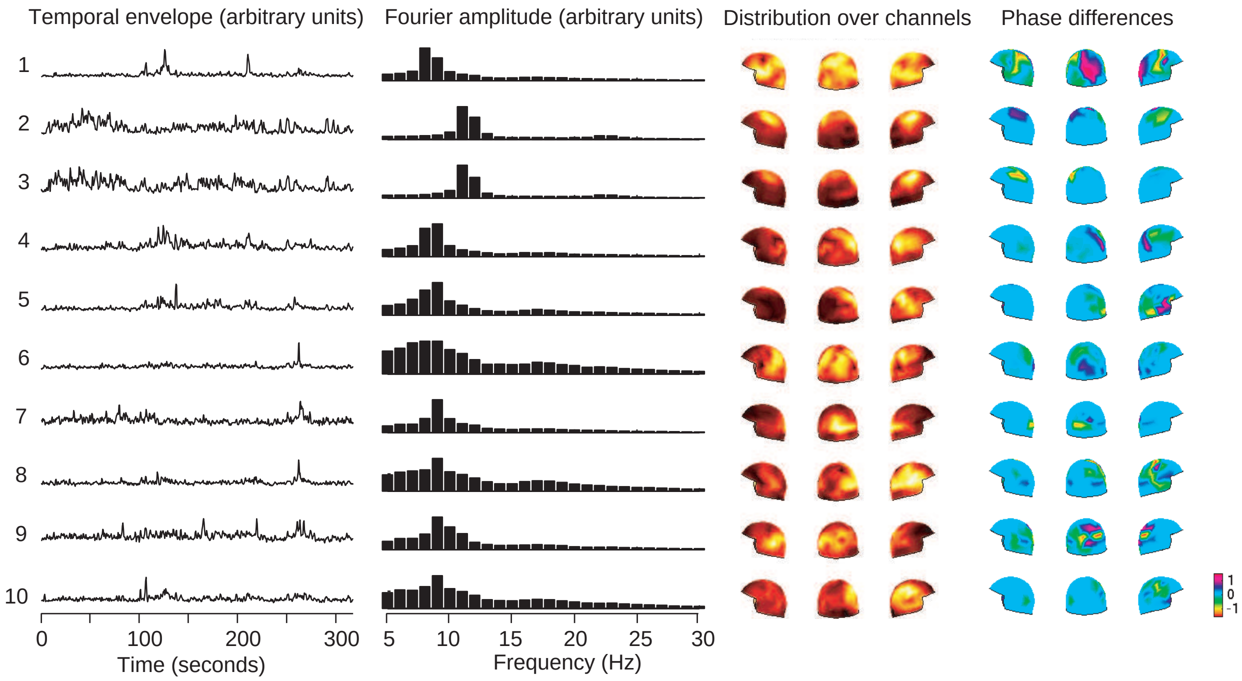

Figure 14: Real MEG recording: The most interesting sources obtained by Fourier-ICA using JADE instead of FastICA, with a complex mixing matrix. For legend see Fig. 10.

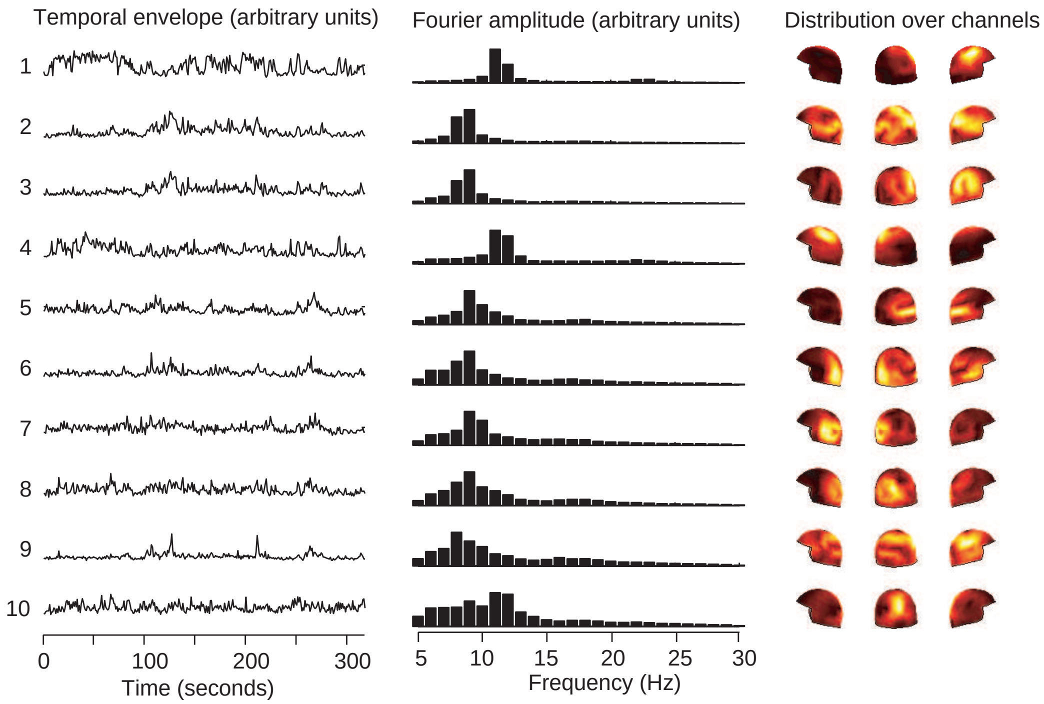

Figure 15: Real MEG recording: The most interesting sources obtained by SOBI. Left: time courses (envelopes), middle: Fourier spectra, right: topographic distributions.

Anemüller, J., Sejnowski, T. J., Makeig, S., 2003. Complex independent component analysis of frequency-domain electroencephalographic data. Neural Networks 16, 1311–1323.

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

Beale, E. M. L., Mallows, C. L., 1959. Scale mixing of symmetric distributions with zero means. The Annals of Mathematical Statistics 30 (4), 1145–1151.

Beckmann, C. F., DeLuca, M., Devlin, J. T., Smith, S. M., 2005. Investigations into resting-state connectivity using independent component analysis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 360 (1457), 1001–13.

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

Belouchrani, A., Meraim, K. A., Cardoso, J.-F., Moulines, E., 1997. A blind source separation technique based on second order statistics. IEEE Trans. on Signal Processing 45 (2), 434–444.

Bingham, E., Hyvärinen, A., 2000. A fast fixed-point algorithm for independent component analysis of complex-valued signals. Int. J. of Neural Systems 10 (1), 1–8.

Cardoso, J.-F., Souloumiac, A., 1993. Blind beamforming for non Gaussian signals. IEE Proceedings-F 140 (6), 362–370.

Comon, P., 1994. Independent component analysis—a new concept? Signal Processing 36, 287–314.

Delfosse, N., Loubaton, P., 1995. Adaptive blind separation of independent sources: a deflation approach. Signal Processing 45, 59–83.

Dyrholm, M., Makeig, S., Hansen, L. K., 2007. Model selection for convolutive ica with an application to spatio-temporal analysis of eeg. Neural Computation.

Groppe, D. M., Makeig, S., Kutas, M., 2009. Identifying reliable independent components via split-half comparisons. NeuroImage, in press.

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

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

Horn, R. A., Johnson, C. R., 1985. Matrix Analysis. Cambridge University Press.

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10 (3), 626–634.

Hyvärinen, A., Hurri, J., 2004. Blind separation of sources that have spatiotemporal variance dependencies. Signal Processing 84 (2), 247–254.

Hyvärinen, A., Karhunen, J., Oja, E., 2001. Independent Component Analysis. Wiley Interscience.

Hyvärinen, A., Oja, E., 1997. A fast fixed-point algorithm for independent component analysis. Neural Computation 9 (7), 1483–1492.

Jung, T. P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., Sejnowski, T., 2000. Removing electroencephalographic artifacts by blind source separation. Psychophysiology 37, 163–178.

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

Miwakeichi, F., Martnez-Montes, E., Valdés-Sosa, P. A., Nishiyama, N., Mizuhara, H., Yamaguchi, Y., 2004. Decomposing EEG data into space-time-frequency components using parallel factor analysis. NeuroImage 22 (3), 1035–45.

Raichle, M. E., MacLeod, A. M., Z., S. A., Powers, W. J., Gusnard, D. A., Shulman, G. L., 2001. A default mode of brain function. Proc. National Academy of Sciences (USA) 98, 676–682.

Ramkumar, P., Parkkonen, L. T., He, B. J., Raichle, M. E., Hämäläinen, M. S., Hari, R., 2007. Identification of stimulus-related and intrinsic networks by spatial independent component analysis of MEG signals. Abstract presented at the Society for Neuroscience Meeting, San Diego, California.

Sarvas, J., 1987. Basic mathematical and electromagnetic concepts of the bio-magnetic inverse problems. Physics in Medicine and Biology 32, 11–22.

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

van de Ven, V. G., Formisano, E., Prvulovic, D., Roeder, C. H., Linden, D. E., 2004. Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. Human Brain Mapping 22 (3), 165–78.

Vigário, R., Särelä, J., Jousmäki, V., Hämäläinen, M., Oja, E., 2000. Independent component approach to the analysis of EEG and MEG recordings. IEEE Trans. Biomedical Engineering 47 (5), 589–593.

Zibulevsky, M., Pearlmutter, B. A., 2001. Blind source separation by sparse decomposition in a signal dictionary. Neural Computation 13 (4), 863–882.

Ziehe, A., Müller, K.-R., 1998. TDSEP—an efficient algorithm for blind separation using time structure. In: Proc. Int. Conf. on Artificial Neural Networks (ICANN'98). Skövde, Sweden, pp. 675–680.

A. Mathematical analysis of Fourier-ICA

Here, we provide rigorous mathematical analysis on two aspects of Fourier-ICA: a decomposition of kurtosis which justifies the decomposition in Section 2.2, and a rather suprising theorem which states that the method can separate even Gaussian sources.

A.1. Decomposition of kurtosis

We assume that the total recording time T is infinite as typical in analysis of blind source separation methods. Denote by $\hat{s}_j$ a random variable which takes randomly and with equal probability the values $\hat{s}_{j,tf}$, i.e. its expectation is average over t and f. Likewise, denote by $\hat{s}_{j,f}$ a random variable which takes randomly and with equal probability the values $\hat{s}_{j,ft}$, i.e. its expectation is average over t with fixed f.

We rescale the components so that $\mathrm{var}(\hat{s}_j) = E\{|\hat{s}_j|^2\} = 1$. Now, consider the kurtosis of $\hat{s}_j$. We use the following definition of kurtosis for complex-valued variables with random phase (Bingham and Hyvärinen, 2000),

$$\mathrm{kurt}(z) = E\{|z|^4\} - 2(E\{|z|^2\})^2, \tag{8}$$

where in contrast to real-valued data, we find a coefficient equal to 2 instead of 3. Due to scaling of the $\hat{s}_j$, we have

$$\mathrm{kurt}(\hat{s}_j) = E_{t,f}\{|\hat{s}_j|^4\} - 2 \tag{9}$$

We would like to express this as a function of the frequency-specific kurtoses $\mathrm{kurt}(\hat{s}_{j,f})$. This can be accomplished as

$$\begin{aligned}\mathrm{kurt}(\hat{s}_j) &= \frac{1}{F}\sum_f E_t\{|\hat{s}_{j,f}|^4\} - 2 \\ &= \frac{1}{F}\sum_f \left[E_t\{|\hat{s}_{j,f}|^4\} - 2\left(E_t\{|\hat{s}_{j,f}|^2\}\right)^2\right] + \left[\frac{2}{F}\sum_f \left(E_t\{|\hat{s}_{j,f}|^2\}\right)^2 - 2\right]\end{aligned} \tag{10}$$

which can be further expressed as

$$\mathrm{kurt}(\hat{s}_j) = \frac{1}{F}\sum_f \mathrm{kurt}(\hat{s}_{j,f}) + 2\left[\mathrm{var}(\hat{s}_{j,f})\right]^2 - 2 \tag{11}$$

This equation shows that the "total" kurtosis over t *and* f can be decomposed as the average of the kurtoses in each frequency band plus (two times) the average of the variances of the frequency bands. This corresponds to the two aspects of non-Gaussianity in Section 2.2

A.2. Gaussian signals: proof of separation

If the sources are Gaussian processes, the first term on the right-hand side in (11) is zero. However, the second term is not, and the kurtosis is not usually zero, as stated in the following Theorem:

Theorem 1. *Assume that the $s_{j,\tau}$ are Gaussian processes. Then, kurt$(\hat{s}_j)$ is non-negative, and zero if and only if $s_{j,\tau}$ is white noise.*

PROOF. Consider a fixed j, and denote $b_f = \left(E_t\{|\hat{s}_{j,f}|^2\}\right)^2 = (\mathrm{var}(\hat{s}_{j,f}))^2$. What we need to prove is that $g(\mathbf{b}) = \frac{1}{F}\sum_f b_f^2 - 1 \geq 0$ and with equality if and only if b_f is constant for all f. We know that the constraint $\frac{1}{F}\sum_f b_f = 1$ must hold by definition of unit variance of the transformed sources. The Lagrangian for this constrained optimization problem gives the condition for extremum as

$$\frac{2}{F}\mathbf{b} - \frac{\lambda}{F}\mathbf{c} = 0 \tag{12}$$

where $\mathbf{c}$ is a vector of all ones. This gives the admissible solution $\mathbf{b} = \mathbf{c}$, at which point the function g attains zero. It is geometrically obvious that this is a minimum, since g is essentially the Euclidean norm, the constraint set is a simplex, and the extremum is attained at the centerpoint of the simplex. Thus, for all other values of b_f, $g(\mathbf{b}) > 0$, and the Theorem is proven.

It may seem surprising that Gaussian signals can be transformed into non-Gaussian ones by a linear transformation, since linear transforms of Gaussian variables are Gaussian. This apparent contradiction is solved when one realizes that when we "concatenate" the Fourier transforms one ofter the other, which is equivalent to summing or averaging over f, the distribution of the transformed sources $\hat{s}_j$ is effectively a mixture of Gaussians. The Gaussian distributions in the mixture cannot all have the same variances unless the data is white noise. Such "scale mixtures of Gaussians" are well-known to be non-Gaussian (Beale and Mallows, 1959).

Thus, the kurtoses of the transformed sources can be assumed to be positive. This would seem to indicate that we can apply any (complex-valued) non-Gaussianity-based ICA method on the transformed data, and we would get the original sources. This is not a valid conclusion, however, because the transformed sources are not independent. This is because of frequency acts as one of the sampling indices. For example, if we know that one source has a large value for a randomly picked column in the matrix $\hat{s}_{j,tf}$, it means that the frequency in that column is one of the preferred frequencies of that source. This gives us information about the value in that column for another source, assuming we know its power spectrum. Thus, the value for one source gives information on the value of another, and statistical independence is violated.

However, in spite of the dependence, ordinary non-Gaussianity-based ICA methods do separate sources under the additional conditions of distinct autocorrelations, as we will prove next. Consider maximization of the following objective function, sum of kurtoses of the estimated sources:

$$K(\mathbf{W}) = \sum_{j=1}^{n} \mathrm{kurt}(\sum_{c} w_{jc}\hat{x}_c) \qquad (13)$$

where $\mathbf{W}$ is the estimate of the inverse of the mixing matrix $\mathbf{A}$. where the kurtosis is computed using expectation over t and f. The maximization of such an objective function is essentially what happens when the Fourier-transformed data is input to kurtosis-based ICA algorithms (Comon, 1994; Delfosse and Loubaton, 1995; Hyvärinen and Oja, 1997), since all the kurtoses of the transformed sources are non-negative according to Theorem 1.

A most interesting mathematical result is the following theorem:

Theorem 2. *Assume the following:*

1. *The original source signals $s_{j,\tau}$ are stationary Gaussian processes.*
2. *The temporal correlations of the sources are distinct in the following sense: The matrix of the Fourier energies*

$$m_{f,j} = E_t\{|\hat{s}_{j,f}|^2\} \tag{14}$$

   has full column rank. (In particular, this requires that the number of frequencies F in the Fourier transform must be equal to or larger than the number of sources.)

3. *The matrix $\mathbf{W}$ is constrained so that the $\sum_c w_{jc}\hat{x}_c$ are uncorrelated and of unit variance (as typical in ICA).*

Then, the objective function K in Equation (13) is maximized when the $\mathbf{W}$ equals the inverse of the mixing matrix $\mathbf{A}$, with rows in arbitrary order and possibly multiplied by -1.

Proof. Consider the linear combination of the Fourier-transformed data: $\sum_c w_c \hat{x}_{c,tf}$. This is also a linear combination of the original sources with some coefficients which depend on the mixing matrix.

$$\sum_c w_{k,c}\hat{x}_{c,tf} = \sum_j (\sum_c w_{k,c}a_{c,j})\hat{s}_{j,tf} = \sum_j q_{k,j}\hat{s}_{j,tf} \tag{15}$$

where we denote the coefficients in parentheses by $q_{k,j}$.

We start by decomposing the kurtoses in K as in (11). Note that the decomposition is valid for any random variable, and therefore also for $\sum_c w_c\hat{x}_{c,tf}$. The kurtoses of the frequency bands (the first terms) are zero because the time points in the windows are jointly Gaussian (Assumption 1 in theorem), and thus any linear transformation of them is also Gaussian. Thus, we only need to consider the second term (the last constant term being immaterial). Transforming the w to the q as in (15), we have

$$K = 2\sum_{j=1}^{n}\sum_{f=1}^{F}\left(E_t\{|\sum_k q_{j,k}\hat{s}_{k,f}|^2\}\right)^2 + \text{const.} \tag{16}$$

By the stationarity part of Assumption 1 of the Theorem, the covariance of $s_{k,t}$ and $s_{k,t'}$ depends only on $t-t'$, and the Fourier basis vectors are the eigenvectors of the covariance matrix of the time windows. Thus, the coefficients $\hat{s}_{k,f}$ and $\hat{s}_{k,f'}$ in the Fourier transform are uncorrelated for $f \neq f'$. This implies

$$K = 2\sum_{f=1}^{F}\sum_{k=1}^{n}\sum_{j=1}^{n}(q_{j,k}^2 m_{f,k})^2 + \text{const.} \tag{17}$$

Denote $b_{i,j} = q_{j,i}^2$, and denote by $\|.\|^2$ the square of the Frobenius norm of the matrix, i.e. the sum of the squares of all the elements. The objective function can then be expressed in matrix form as

$$K = 2\|\mathbf{MB}\|^2 + \text{const.} \tag{18}$$

Now, we can apply the following lemma, which slightly generalizes the lemma in (Hyvärinen and Hurri, 2004), which is also related to Lemma 15 in (Comon, 1994):

Lemma 1. *Consider a $n \times n$ matrix $\mathbf{B}$ that is doubly stochastic, i.e. the sums of rows and the sums of columns are all equal to one. Take any $F \times n$ matrix $\mathbf{M}$ that has full column rank. Then for the Frobenius norm it holds:*

$$\|\mathbf{MB}\|^2 \leq \|\mathbf{M}\|^2 \tag{19}$$

with equality if and only if $\mathbf{B}$ is a permutation matrix.

Proof of Lemma: According to a theorem by Birkhoff (Horn and Johnson, 1985), we can represent a doubly stochastic matrix as a finite convex sum of permutation matrices: $\mathbf{B} = \sum_s \alpha_s \mathbf{P}_s$ with $\alpha_s > 0$ and $\sum_s \alpha_s = 1$. The converse also holds. The set of doubly stochastic matrices is thus a compact convex polygon with extreme points $\mathbf{P}_s$. On the other hand, the square of the Frobenius norm $\|\mathbf{MB}\|^2$ is a strictly convex function of $\mathbf{B}$ because $\|.\|^2$ is trivially strictly convex, and the multiplication (from the left) by $\mathbf{M}$ is an injective linear transformation since $\mathbf{M}$ has full column rank. Thus, the maxima are obtained at the extreme points, i.e. when $\mathbf{B}$ is a permutation matrix, which proves the lemma.

To continue with the proof of the Theorem: In (18), $\mathbf{B}$ is doubly stochastic since it consists of the squares of an orthogonal matrix by Assumption 3, and $\mathbf{M}$ has full column rank by Assumption 2 of the Theorem. Applying the Lemma, we see that K is maximized exactly when $\mathbf{B}$ is a permutation matrix. This means that $\mathbf{Q} = \mathbf{WA}$ is a signed permutation matrix, and the theorem is proven.

The theorem basically says that even Gaussian sources are separated by maximization of K. The indeterminacy of the sign is ubiquituous in source separation. The condition that the source signals must have *distinct* temporal correlations is well-known in the theory of blind source separation using temporal correlation, sometimes called “second-order” source separation (Belouchrani et al., 1997; Ziehe and Müller, 1998; Hyvärinen et al., 2001).