TOPOGRAPHIC INDEPENDENT COMPONENT ANALYSIS: VISUALIZING THE DEPENDENCE STRUCTURE

Aapo Hyvärinen, Patrik O. Hoyer and Mika Inki

Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland
http://www.cis.hut.fi/projects/ica/

ABSTRACT

In ordinary independent component analysis, the components are assumed to be completely independent, and they do not necessarily have any meaningful order relationships. In practice, however, the estimated "independent" components are often not at all independent. We propose that this residual dependence structure could be used to define a topographic order for the components. In particular, a distance between two components could be defined using their higher-order correlations, and this distance could be used to create a topographic representation. Thus we obtain a linear decomposition into approximately independent components, where the dependence of two components is approximated by the proximity of the components in the topographic representation.

1. INTRODUCTION

Indendent component analysis (ICA) [9] is a statistical model where the observed data is expressed as a linear transformation of latent variables that are nongaussian and mutually independent. The classic version of the model can be expressed as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ is the vector of observed random variables, $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ is the vector of the independent latent variables (the "independent components"), and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. The problem is then to estimate both the mixing matrix $\mathbf{A}$ and the realizations of the latent variables s_i, using observations of $\mathbf{x}$ alone. Exact conditions for the identifiability of the model were given in [3]; the most fundamental is that the independent components s_i must be nongaussian [3].

In classic ICA, the independent components s_i have no particular order, or other relationships. It is possible, though, to define an order relation between the independent components by such criteria as nongaussianity or contribution to the observed variance; the latter is given by the norms of the corresponding columns of the mixing matrix as the independent components are defined to have unit variance. Such trivial order relations may be useful for some purposes, but they are not very informative in general.

The lack of an inherent order of independent components is related to the assumption of complete statistical independence. In practical applications of ICA, however, one can very often observe clear violations of the independence assumption. It is possible to find, for example, couples of estimated independent components such that they are clearly dependent on each other. This dependence structure is often very informative, and it would be useful to estimate it somehow.

Estimation of the "residual" dependency structure of estimates of independent components could be based, for example, on computing the cross-cumulants. Typically these would be higher-order cumulants since second-order cross-cumulants, i.e. the covariances, are typically very small, and are in fact forced to be zero in many ICA estimation methods, e.g. [3, 7, 4]. A more information-theoretic measure for dependence would be given by mutual information. Whatever measure is used, however, the problem remains as to how such numerical estimates of the dependence structure should be visualized or otherwise utilized. Moreover, there is another serious problem associated with simple estimation of some dependency measures from the estimates of the independent components. This is due to the fact that often the independent components do not form a well-defined set. Especially in image decomposition [1, 12, 6], the set of potential independent components seems to be larger than what can be estimated at one time, in fact the set might be infinite. A classic ICA method gives an arbitrarily chosen subset of such independent components. Thus, it is important in many applications that the dependency information is utilized during the estimation of the independent components, so that the estimated set of independent components is one that can be ordered in a meaningful way.

We propose here that the residual dependency structure of the "independent" components, i.e. dependencies that cannot be cancelled by ICA, could be used to define a *topographic order* between the components. The topographic order is easy to represent by visualization, and has the usual computational advantages associated with topographic maps [10]. We propose a modification of the ICA model that explicitly formalizes a topographic order between the independent components. This gives a topographic map where the distance of the components in the topographic representation is a function of the dependencies of the components. Components that are near to each other in the topographic representation are strongly dependent in the sense of higher-order correlations, or mutual information. This gives a new principle for topographic organization. Furthermore, we derive a learning rule for the estimation of the model. Finally, we show experiments to validate our approach, including feature extraction from images and blind separation of magnetoencephalographic components.

2. TOPOGRAPHIC ICA

2.1. Modelling correlations of energies

The idea in topographic ICA to relax the assumption of the independence of the components s_i in (1) so that components that are *close to each other in the topography* are not assumed to be independent in the model. For example, if the topography is defined by a lattice or grid, the dependency of the components is a function of the distance of the components on that grid. In contrast, components that are not close to each other in the topography *are* independent, at least approximately; thus most pairs of components are independent.

The basic problem is then to choose what kind of dependencies are allowed between near-by components. The most basic dependence relation is linear correlation. However, allowing linear correlation between the components does not seem very useful. In fact, in many ICA estimation methods, the components are constrained to be uncorrelated [3, 7, 4], so the requirement of uncorrelatedness seems natural in any extension of ICA as well.

A more interesting kind of dependency is given by a certain kind of higher-order correlation, namely correlation of energies. This means that

$$\text{cov}\,(s_i^2, s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} \neq 0 \qquad (2)$$

if s_i and s_j are close in the topography. Here, we assume that this covariance is positive. Intuitively, such a correlation means that the components tend to be active, i.e. non-zero, at the same time, but the actual values of s_i and s_j are not easily predictable from each other. For example, if the variables are defined as products of two zero-mean independent components z_i, z_j and a common "variance" variable σ:

$$\begin{aligned} s_i &= z_i \sigma \\ s_j &= z_j \sigma \end{aligned} \qquad (3)$$

then s_i and s_j are uncorrelated, but their energies are not. In fact the covariance of their energies equals $E\{z_i^2\sigma^2 z_j^2\sigma^2\} - E\{z_i^2\sigma^2\}E\{z_j^2\sigma^2\} = E\{\sigma^4\} - E\{\sigma^2\}^2$, which is positive because it equals the variance of σ^2 (we assumed here for simplicity that z_i and z_j are of unit variance). This kind of a dependence has been observed, for example, in linear image features [13, 6]; it is illustrated in Fig. 1.

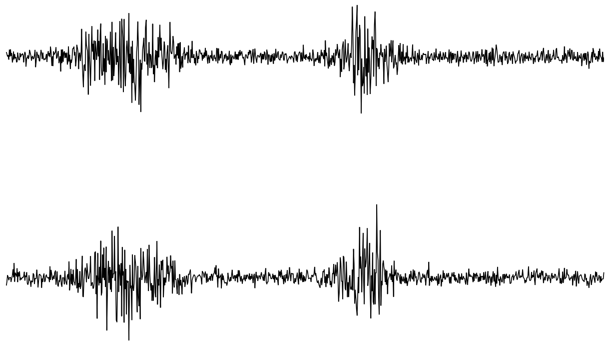

Figure 1: Illustration of higher-order dependencies. The two signals in the figure are uncorrelated but they are not independent. In particular, their energies are correlated. The signals were generated as in (3), but for purposes of illustration, the random variable σ was replaced by a time-correlated signal.

2.2. The generative model

Now we define a generative model that implies correlation of energies for components that are close in the topographic grid. In the model, the observed image patches are generated as a linear transformation of the components s_i, just as in the basic ICA model in (1). The point is to define the joint density of $\mathbf{s}$ so that it expresses the topography.

We define the joint density of $\mathbf{s}$ as follows. The variances σ_i^2 of the s_i are not constant, instead they are assumed to be random variables, generated according to a model to be specified. After generating the variances, the variables s_i are generated independently from each other, using some conditional distributions to be specified. In other words, the s_i are *independent given their variances*. Dependence among the s_i is implied by the dependence of their variances. According to the principle of topography, the variances corresponding to near-by components should be (positively) correlated, and the variances of components that are not close should be independent, at least approximatively.

To specify the model for the variances σ_i^2, we need to first define the topography. This can be accomplished by a neighborhood function $h(i,j)$, which expresses the proximity between the i-th and j-th components. The neighborhood function can be defined in the same ways as with the self-organizing map [10]. The neighborhood function $h(i,j)$ is thus a matrix of hyperparameters. In this paper, we consider it to be known and fixed.

Using the topographic relation $h(i,j)$, many different models for the variances σ_i^2 could be used. We prefer here to define them by an ICA model followed by a nonlinearity:

$$\sigma_i = \phi(\sum_{k=1}^{n} h(i,k)u_k) \tag{4}$$

where u_i are the "higher-order" independent components used to generate the variances, and ϕ is some scalar nonlinearity. The distributions of the u_i and the actual form of ϕ are additional hyperparameters of the model; some suggestions are given in [5]. It seems natural to constrain the u_k to be non-negative. The function ϕ can then be constrained to be a monotonic transformation in the set of non-negative real numbers. This assures that the σ_i's are non-negative.

2.3. Basic properties of topographic ICA

The model as defined above has the following properties. (Here, we consider for simplicity only the case of sparse, i.e. supergaussian, data.) The first basic property is that all the components s_i are uncorrelated, as can be easily proven by symmetry arguments [5]. Moreover, their variances can be defined to be equal to unity, as in classic ICA. Second, components s_i and s_j that are near to each other, i.e. such that $h(i,j)$ is significantly non-zero, tend to be active (non-zero) at the same time. In other words, their energies s_i^2 and s_j^2 are positively correlated. This is exactly the dependence structure that we wanted to model in the first place. Third, latent variables that are far from each other are practically independent. Higher-order correlation decreases as a function of distance. For details, see [5].

3. ESTIMATION OF THE MODEL

3.1. Approximating likelihood

To estimate the model, we can use maximum likelihood estimation. The model is, however, a missing variables model in which the likelihood cannot be obtained in closed form. To simplify estimation, we can derive a tractable approximation of the likelihood.

Fixing all the hyperparameters (nonlinearities etc.) to suitable values [5], we obtain the following approximation $\tilde{L}$ of the log-likelihood. Given T observed data point $\mathbf{x}(t), t = 1, ..., T$, we have

$$\log \tilde{L}(\mathbf{w}_i, i = 1, ..., n) = \sum_{t=1}^{T} \sum_{j=1}^{n} G(\sum_{i=1}^{n} h(i,j)(\mathbf{w}_i^T \mathbf{x})^2) + T \log |\det \mathbf{W}|. \tag{5}$$

where the scalar function G is obtained from the pdf's of p_u, the variance-generating variables, by:

$$G(y) = \log \int \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}uy) p_u(u) \sqrt{h(i,i)u}\, du, \tag{6}$$

and the matrix $\mathbf{W}$ is the inverse of the mixing matrix $\mathbf{A}$.

The function G has a similar role as the log-density of the independent components in classic ICA. If the data is sparse (like natural image data), the function $G(y)$ needs to be chosen to be *convex* for non-negative y [5]. For example, one could use the function:

$$G_1(y) = -\alpha_1 \sqrt{y} + \beta_1, \tag{7}$$

This function is related to the exponential distibution [6]. The scaling constant α_1 and the normalization constant β_1 are determined so as to give a probability density that is compatible with the constraint of unit variance of the s_i, but they are irrelevant in the following.

3.2. Estimation algorithm

The approximation of the likelihood given above enables us to derive a simple gradient algorithm. First, we assume here that the data is preprocessed by whitening to give the white data $\mathbf{z}$, and that the $\mathbf{w}_i$, are constained to form an orthonormal system [3, 7, 4]. This implies that the estimates of the components are uncorrelated.

Thus we can simply derive a gradient algorithm in which the i-th (weight) vector w_i is updated as

$$\Delta \mathbf{w}_i \propto E\{\mathbf{z}(\mathbf{w}_i^T \mathbf{z}) r_i)\} \tag{8}$$

where

$$r_i = \sum_{k=1}^{n} h(i,k) g(\sum_{j=1}^{n} h(k,j)(\mathbf{w}_j^T \mathbf{z})^2). \tag{9}$$

The function g is the derivative of G. After every step of (8), the vectors $\mathbf{w}_i$ are normalized to unit variance and orthogonalized, as in [7, 4], for example.

In a neural interpretation, the learning rule in (8) can be considered as "modulated" Hebbian learning, since the learning term is modulated by the term r_i. This term could be considered as top-down feedback as in [6], since it is a function of the local energies which could be the outputs of higher-order neurons (complex cells).

3.3. Connection to independent subspace analysis

Another closely related modification of the classic ICA model was introduced in [6]. As in topographic ICA, the components s_i were not assumed to be all mutually independent. Instead, it was assumed that the s_i can be divided into couples, triplets or in general n-tuples, such that the s_i inside a given n-tuple could be dependent on each other, but dependencies between different n-tuples were not allowed. A related relaxation of the independence assumption was proposed in [2, 11]. Inspired by Kohonen's principle of feature subspaces [10], the probability densities for the n-tuples of s_i were assumed in [6] to be *spherically symmetric*, i.e. depend only on the norm.

In fact, topographic ICA can be considered a generalization of the model of independent subspace analysis. The likelihood in independent subspace analysis can be expressed as a special case of the approximative likelihood (5), see [5]. This connection shows that topographic ICA is closely connected to the principle of invariant-feature subspaces. In topographic ICA, the invariant-feature subspaces, which are actually no longer independent, are completely overlapping. Every component has its own neighborhood, which could be considered to define a subspace. Each of the terms $\sum_{i=1}^{n} h(i,j)(\mathbf{w}_i^T \mathbf{z})^2$ could be considered as a (weighted) projection on a feature subspace, i.e. as the value of an invariant feature.

4. EXPERIMENTS

4.1. Experiments in feature extraction of image data

A very interesting application of topographic ICA can be found with image data. The data was obtained by taking 16×16 pixel image patches at random locations from monochrome photographs depicting wild-life scenes (animals, meadows, forests, etc.). The mean gray-scale value of each image patch (i.e. the DC component) was subtracted. The data was then low-pass filtered by reducing the dimension of the data vector by principal component analysis, retaining the 160 principal components with the largest variances, after which the data was whitened by normalizing the variances of the principal components. These preprocessing steps are essentially similar to those used in [6, 12].

The neighborhood function was defined so that every neighborhood consisted of a 3×3 square of 9 units on a 2-D torus lattice [10]. (The choice of a two-dimensional grid was here motivated by convenience of visualization only.) The approximation of likelihood in Eq. (5) for 50,000 observations was maximized under the constraint of orthonormality of the filters in the whitened space, using the gradient method (8).

The obtained basis vectors, i.e. columns of the mixing matrix, are shown in Fig. 2. The basis vectors are similar to those obtained by ordinary ICA of image data [12, 1]. In addition, they have a clear topographic organization.

The connection to independent subspace analysis [6], which is basically a complex cell model, can also be found in these results. Two neighboring basis vectors in Fig. 2 tend to be of the same orientation and frequency. Their locations are near to each other as well. In contrast, their phases are very different. This means that a neighborhood of such basis vectors, i.e. simple cells, tends to function as a complex cell.

4.2. Experiments with magnetoencephalographic recordings

Next we did experiments on blind source separation. Two minutes of magnetoencephalographic (MEG) data was collected using a 122-channel whole-scalp neuromagnetometer device. The measurement device and the data are described in detail in [14]. The test subject was asked to blink and make horizontal eye saccades in order to produce typical ocular artifacts and bite the teeth for 20 seconds in order to create myographic artifacts. This 122 dimensional input data was first reduced to 20 dimensions by PCA, in order to eliminate noise and "bumbs", which appear in the data if the dimensionality is not sufficiently reduced [8]. The topographic ICA algorithm was then run on the data using a one dimensional ring-shaped topography. The neighborhood was formed by convolving a vector of three ones with itself four times.

The resulting separated signals are shown in Fig, 3. The signals themselves are very similar to those found by ICA in [14]. As for the topographic organization, we can see that, first, the signals corresponding to bites (#9-#15) are now adjacent. When computing the field patterns corresponding to these signals, one can also see that the signals are ordered according to whether they come from the left or the right side of the head. Second, two signals corresponding to eye artifacts are adjacent as well (#18 and #19). The signal # 18 corresponds to horizontal eye saccades and the signals #19 to eye blinks. Finally, a signal which seems to relate to eye activity has been separated into #17. We can also see signals that do not seem to have any meaningful topographic relations, probably because they are quite independent form the rest of the signals. These include the heart beat (signal #7), and a signal corresponding to a digital watch which was at a distance of 1 m from the magnetometer (signal #6). Thus topographic ICA orders the signals mainly into two clusters, one created by the signals coming from the muscle artifact, and the other by eye muscle activity.

Thus we see that topographic ICA finds largely the same components as those found by ICA in [14]. Using topographic ICA has the advantage, however, that signals are grouped together according to their dependence content.

5. CONCLUSION

We introduced topographic ICA, which is a generative model that combines topographic mapping with ICA. As in all topographic mappings, the distance in the representation space (on the topographic "grid") is related to the distance of the represented components. In topographic ICA, the distance between represented components is defined by the mutual information implied by the higher-order correlations, which gives the natural distance measure in the context of ICA.

The utility of this novel model of topographic organization is clearly seen with natural image data, where topographic ICA gives a linear decomposition into Gabor-like linear features. In contrast to ordinary ICA, the higher-order dependencies that linear ICA could not remove define a topographic order such that nearby cells tend to be active at the same time. This implies also that the neighborhoods have properties similar to those of complex cells. Another application of topographic ICA is found in blind source separation of MEG data, where the method finds clusters of signals related to each other.

6. REFERENCES

[1] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[2] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA, 1998.

[3] P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.

[4] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

[5] A. Hyvärinen and P. O. Hoyer. Topographic independent component analysis. 1999. Submitted, available at http://www.cis.hut.fi/~aapo/.

[6] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 2000. (in press).

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

[8] A. Hyvärinen, J. Särelä, and R. Vigário. Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 425–429, Aussois, France, 1999.

[9] C. Jutten and J. Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

[10] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Berlin, Heidelberg, New York, 1995.

[11] J. K. Lin. Factorizing multivariate function classes. In *Advances in Neural Information Processing Systems*, volume 10, pages 563–569. The MIT Press, 1998.

[12] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[13] E. P. Simoncelli and O. Schwartz. Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In *Advances in Neural Information Processing Systems 11*, pages 153–159. MIT Press, 1999.

[14] R. Vigário, V. Jousmäki, M. Hämäläinen, R. Hari, and E. Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems 10*, pages 229–235. MIT Press, 1998.

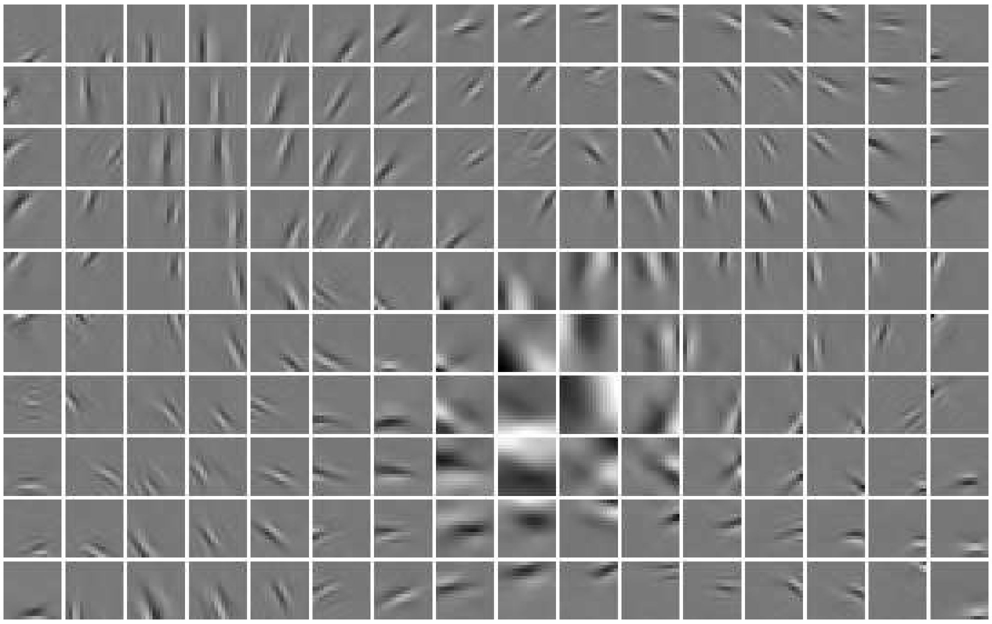

Figure 2: Topographic ICA of natural image data. The model gives Gabor-like basis vectors for image windows. Basis vectors that are similar in location, orientation and/or frequency are close to each other. The phases of nearby basis vectors are very different, giving each neighborhood properties similar to those of complex cells.

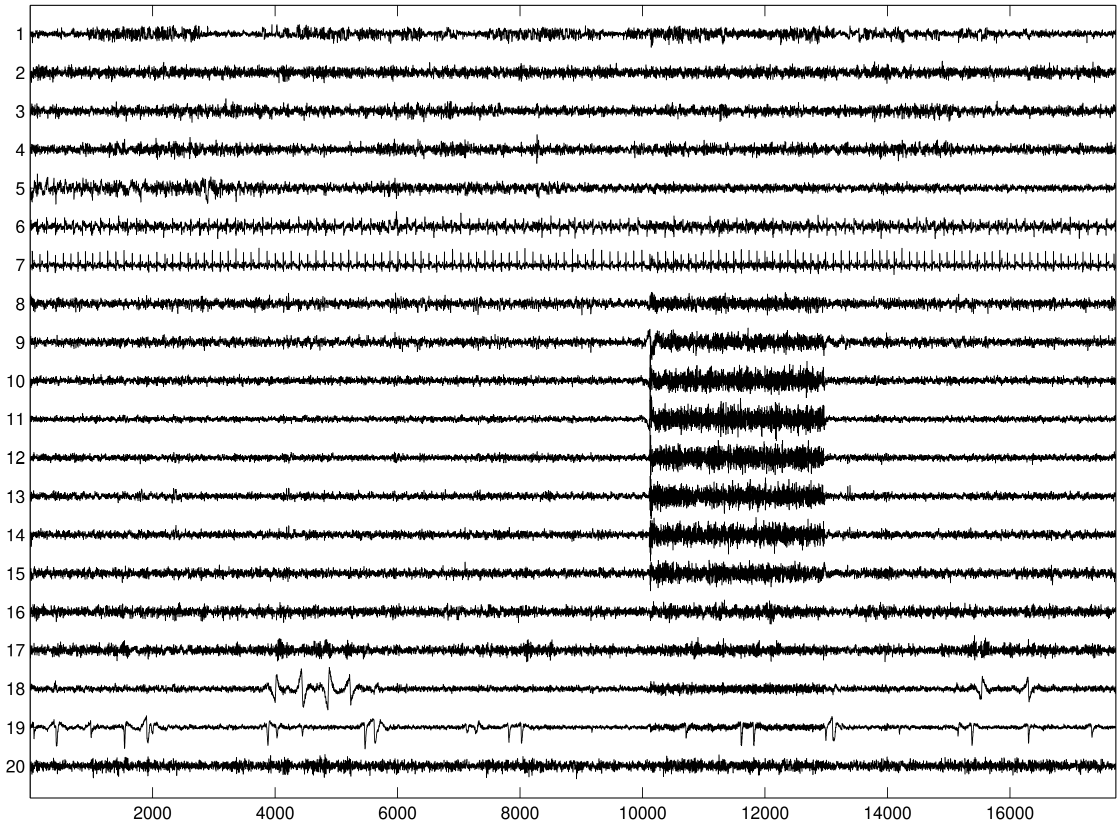

Figure 3: The source signals found by topographic ICA from MEG data.