Independent component analysis for noisy data —MEG data analysis—

Shiro Ikeda
PRESTO, JST

Keisuke Toyama
Shimadzu Inc.

Abstract

ICA (independent component analysis) is a new, simple and powerful idea for analyzing multi-variant data. One of the successful applications is neurobiological data analysis such as EEG (electroencephalography), MRI (magnetic resonance imaging), and MEG (magnetoencephalography). But there remain a lot of problems. In most cases, neurobiological data contain a lot of sensory noise, and the number of independent components is unknown. In this article, we discuss an approach to separate noise-contaminated data without knowing the number of independent components. A well-known two stage approach to ICA is to pre-process the data by PCA (principal component analysis), and then the necessary rotation matrix is estimated. Since PCA does not work well for noisy data, we implement a factor analysis model for pre-processing. In the new pre-processing, the number of the sources and the amount of the sensory noise are estimated. After the pre-processing, the rotation matrix is estimated using an ICA method. Through the experiments with MEG data, we show this approach is effective.

Keywords ICA; PCA; factor analysis; MDL; MEG; Spatial filters.

1 Introduction

The basic problem of ICA is defined for the noiseless case, where the sources and observations have the following linear relation,

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

$$\boldsymbol{x} \in R^n, \qquad \boldsymbol{s} \in R^m, \qquad A \in R^{n\times m}.$$

The assumptions of an ICA problem are that each component of $\boldsymbol{s}$ is 0 as its mean value, mutually independent and drawn from a probability distribution which is not a normal distribution except for at most one component. In this article, we also restrict m to be smaller or equal to n. This assumption is necessary for the existence of linear solution.

The goal of ICA is to estimate a matrix W which satisfies the following equation,

$$WA = PD \tag{2}$$

$$W \in R^{m\times n}, \qquad P, D \in R^{m\times m}..$$

Here, P is a permutation matrix which has single entry of one in each row and column, and D is a diagonal matrix. This simple problem is solved in the framework of semi parametric approach[1], and giving a lot of interesting theoretical and practical results.

However, when we apply ICA to real world problem, the situation is different from the above ideal case. In many cases, we cannot ignore noise, and the number of the sources m is not known. For example, in the case of neurobiological data such as EEG[12, 17] or MEG[5, 16], the number n of the sensors is large and sometimes around 200, but we believe that the number of the sources is not so large in a macroscopic viewpoint within a short period, and noise is very large. Therefore, eq.(1) is not enough to describe the problem. It is pointed out that, especially when the number of the sources is small, one cannot have a good solution generally[10].

In this article, we discuss the case where there is additive noise in observations as,

$$\boldsymbol{x} = A\boldsymbol{s} + \boldsymbol{\epsilon}. \tag{3}$$

Here, $\boldsymbol{\epsilon}$ is an n-dimensional real valued noise term and we assume that components $\{\epsilon_i\}$ $(i = 1, \cdots, n)$ of $\boldsymbol{\epsilon}$ are mutually independent. We call $\boldsymbol{\epsilon}$ the sensory noise in this article. This is the case of MEG data.

H. Attias has proposed a parametric approach, IFA (independent factor analysis) which solves this problem in the framework of maximum likelihood estimation[3]. It is one of the extensions of basic ICA problem, but is difficult to be applied to biological data when the numbers of sensors and sources are large. We will discuss the relation between our method and IFA (subsection 2.4).

We propose a semi parametric approach to solve this problem. The idea is to use factor analysis for the pre-processing of data. In many ICA algorithms, PCA is used for the pre-processing to make the signals uncorrelated but we replace PCA with factor analysis. By the new pre-processing, the source signals are made to be uncorrelated, and the power of sensory noise and the number of the sources are estimated. After the pre-processing, we use one of ICA algorithms for estimating separation matrix. In the following sections, we describe the factor analysis approach, its relation to ICA, and experimental results.

2 Factor analysis and ICA

2.1 Factor analysis

In factor analysis (see for example [14]), we discuss the case where a real valued n-dimensional observation $\boldsymbol{x}$ is modeled as,

$$\boldsymbol{x} = \boldsymbol{\mu} + A\boldsymbol{f} + \boldsymbol{\epsilon} \tag{4}$$
$$\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\epsilon} \in R^n, \qquad \boldsymbol{f} \in R^m, \qquad A \in R^{n\times m}.$$

The assumptions in factor analysis are that: a) $\boldsymbol{f}$ is normally distributed as $\boldsymbol{f} \sim N(\mathbf{o}, I_m)$, where I_m is the $R^{m\times m}$ identity matrix, b) $\boldsymbol{\epsilon}$ is normally distributed as $\boldsymbol{\epsilon} \sim N(\mathbf{o}, \Sigma)$ where Σ is a diagonal matrix, and c) $\boldsymbol{f}$ and $\boldsymbol{\epsilon}$ are mutually independent. The mean of $\boldsymbol{x}$ is given by $\boldsymbol{\mu}$ which is assumed to be 0 in this article. Without $\boldsymbol{\mu}$, eq.(4) becomes similar to eq.(3).

The goal of factor analysis is to estimate m, A (factor loading matrix), and Σ (unique variance matrix) using the second order statistics of the observation $\boldsymbol{x}$. When m is given, there are various estimation methods for A and Σ. The major ones are ULS (unweighted least squares method) and MLE (maximum likelihood estimation). Both of them are summarized in terms of related loss functions. Suppose we have a data set as $\{\boldsymbol{x}_t\}$

$(t = 1, \ldots, N)$ and let C be the covariance matrix of $\boldsymbol{x}$, $(C = \sum_t \boldsymbol{x}_t \boldsymbol{x}_t^T / N)$. Here, ^{T} denotes transpose. The estimates of ULS $(\hat{A}, \hat{\Sigma})_{\mathrm{ULS}}$ and MLE $(\hat{A}, \hat{\Sigma})_{\mathrm{MLE}}$ are defined as,

$$(\hat{A}, \hat{\Sigma})_{\mathrm{ULS}} = \underset{A,\Sigma}{\operatorname{argmin}} \operatorname{tr} (C - (AA^T + \Sigma))^2, \tag{5}$$

$$(\hat{A}, \hat{\Sigma})_{\mathrm{MLE}} = \underset{A,\Sigma}{\operatorname{argmax}} L(A, \Sigma), \tag{6}$$

$$L(A, \Sigma) = -\frac{1}{2} \left\{ \operatorname{tr} \left(C(\Sigma + AA^T)^{-1} \right) + \log(\det(\Sigma + AA^T)) + n \log 2\pi \right\}.$$

In both loss functions, the matrix A appears only in the form of AA^T. Therefore, cost functions are not affected by any orthogonal matrix (rotation matrix) $U \in R^{m\times m}$, multiplied to A as AU, because $AA^T = AUU^T A^T$. This means we cannot estimate the rotation of A by minimizing a loss function in eq.(5) or (6). One typical method to ignore this ambiguity through the estimation process is to force $A^T A$ to be diagonal and the diagonal components to be in descending order.

The estimation of the rotation is one of the main problems in factor analysis. There have been proposed a lot of methods, but we are not going to use any factor analysis approach. We use an ICA method. This problem is discussed in subsection 2.3.

The matrix A has $n\times m$ parameters and since rotation cannot be determined, it has $n\times m - m(m-1)/2$ meaningful free parameters, and Σ has n parameters. On the other hand, C has $n(n+1)/2$ freedom. Therefore, if $n(m+1) - m(m-1)/2 \leq$n(n+1)/2 holds, we have sufficient conditions to solve eq.(5) or eq.(6). By taking one more condition $m \leq n$ into account, we have the following bound,

$$m \leq \frac{1}{2} \left\{ 2n + 1 - \sqrt{8n+1} \right\}. \tag{7}$$

This is a necessary condition for A and Σ to be estimable (Ledermann[11]). Anderson and Rubin discussed sufficient conditions[2]. We don't go into the detail of the estimation method in this article. But we can use the gradient descent algorithm or Gauss-Newton method[14]. In the case of MLE, we can also use the EM (expectation maximization) algorithm.

The remaining problem is to estimate the number of factors, m. There are many estimation approaches. Some of them are based on the eigenvalues of the covariance matrix. There are other well known approaches which are based on statistical model selection with information criterion such as AIC (Akaike information criteria) and MDL (minimum description length). In this article, we use MDL for the estimation of m in the framework of model selection. Since MDL is based on MLE, we use MLE for the estimation of A and Σ as shown in eq.(6).

MDL is defined as follows,

$$\mathrm{MDL} = -L(\hat{A}, \hat{\Sigma}) + \frac{\log N}{N} \times \text{the number of free parameters.}$$

The number of free parameters in A and Σ is $n(m+1) - m(m-1)/2$, and MDL is defined as,

$$\mathrm{MDL} = -L(\hat{A}, \hat{\Sigma}) + \frac{\log N}{N} \left(n(m+1) - \frac{m(m-1)}{2} \right).$$

We estimate A and Σ for different m, $(1 \leq m \leq \{2n + 1 - \sqrt{8n+1}\}/2)$ and select the set of m, A, and Σ which minimizes MDL.

2.2 Factor analysis as the pre-processing

It is not necessary but, from a practical reason, ICA algorithms are separated into two parts in many cases[6, 8, 9]. One is to pre-process the data such that they become uncorrelated. This part is called sphering or whitening. After the pre-processing, the remaining part of the estimation is the rotation matrix. A lot of algorithms are proposed[4, 6, 8, 9] for the estimation of the rotation matrix.

For sphering, the basic approach is to use PCA[8]. This is based on the second order statistics. Let $P = C^{1/2}$, where $C = PP^T$. By letting,

$$\boldsymbol{x}' = P^{-1}\boldsymbol{x},$$

it is clear that $\sum_t \boldsymbol{x}'_t \boldsymbol{x}'^T_t / N = I_n$. Therefore, components of $\boldsymbol{x}'$ are uncorrelated, and the power of each component of $\boldsymbol{x}'$ is normalized as 1. We can then remove the ambiguity of amplitude.

When the data are noisy, we still assume that the sources are uncorrelated to each other and the power of each is 1, that is, $E[\boldsymbol{s}\boldsymbol{s}^T] = I_m$, where $E[\cdot]$ denotes the expectation. But the matrix $P = C^{1/2}$ does not give pre-processing such that the remaining part is a rotation matrix. We can instead use factor analysis.

As the result of factor analysis, we estimate $\hat{A}$ (with the ambiguity of rotation) and $\hat{\Sigma}$. Let $Q \in R^{m \times n}$ be the pseudo-inverse of $\hat{A}$ where $\hat{A}Q\hat{A} = \hat{A}$ holds. After transforming the data as,

$$\boldsymbol{z} = Q\boldsymbol{x},$$

$\boldsymbol{z}$ becomes sphered data. However, its covariance matrix is not I_m. If we could estimate $\hat{A}$ correctly with the ambiguity of rotation, covariance matrix of $\boldsymbol{z}$ is,

$$E[\boldsymbol{z}\boldsymbol{z}^T] = I_m + Q\Sigma Q^T.$$

This result shows that we can make the part of $\boldsymbol{x}$ due to the sources uncorrelated, but not the sensory noise at the same time. And sphering is completed.

The pseudo-inverse matrices are not unique. We chose the one to minimize the expected norm $E[(\boldsymbol{x} - \hat{A}\boldsymbol{z})^T \hat{\Sigma}^{-1}(\boldsymbol{x} - \hat{A}\boldsymbol{z})]$, which is the difference between $\boldsymbol{x}$ and the reconstructed observation $\hat{A}\boldsymbol{z}$ measured with Σ^{-1}. The solution is,

$$Q = (\hat{A}^T \hat{\Sigma}^{-1} \hat{A})^{-1} \hat{A}^T \hat{\Sigma}^{-1} \tag{8}$$

(for the proof, see [13]). This helps to reduce the sensory noise when we reconstruct the data from independents components. It will be discussed in the next subsection.

2.3 ICA as estimating the rotation matrix

After we pre-processed $\boldsymbol{x}$ by Q, $\boldsymbol{z}$ is uncorrelated except for the part due to the sensory noise. What is left is to estimate the rotation matrix. This is also one big problem in factor analysis. Instead of adopting any factor analysis approach, we use an ICA approach.

In subsection 2.1, we assumed that $\boldsymbol{s}$ ($\boldsymbol{f}$ in the subsection) and $\boldsymbol{\epsilon}$ are normally distributed. We break one of them. We still assume $\boldsymbol{\epsilon}$ is normally distributed, but $\boldsymbol{s}$ is not normally distributed and each component s_i is independent. We can use some ICA algorithm now.

The ICA algorithm we use here should not be affected by the second order statistics since even if data are pre-processed by factor analysis, components of sphered data $\boldsymbol{z}$ still have second order correlations. Therefore, an algorithm based on higher order statistics is preferable. We used JADE (joint approximate diagonalization of eigen-matrices) by

J.-F. Cardoso & A. Souloumiac, which is based on the 4th order cumulant[7]. Suppose a separation matrix $B \in R^{m\times m}$ is estimated by an ICA algorithm. The separated signal $\boldsymbol{y}$ is obtained as,

$$\boldsymbol{y} = B\boldsymbol{z} = BQ\boldsymbol{x} = BQ(A\boldsymbol{s}+\boldsymbol{\epsilon}) = W(A\boldsymbol{s}+\boldsymbol{\epsilon}), \quad W \stackrel{\text{def}}{=} BQ, \quad W{\in}R^{m\times n}. \tag{9}$$

The goal of ICA is to estimate WA to be PD, $(P, D \in R^{m\times m})$ as eq.(2). In the noisy case, we can make signals independent based on the parts due to the sources, but it is impossible to cancel the sensory noise by any linear operation. Our method does not apply nonlinear transformation to reduce the additive sensory noise, but projects the high dimensional observation into smaller dimension of source space with a linear matrix W. And in the process of the linear projection, we compress the sensory noise $\boldsymbol{\epsilon}$ into smaller dimension. The total power of noise will be reduced because each component of sensory noise is mutually independent.

We can also estimate the mixing system with B. Let us denote A estimated by factor analysis as $\hat{A}_{\mathrm{FA}}$, and the new mixing system as $\hat{A}_{\mathrm{ICA}}$ where,

$$\hat{A}_{\mathrm{ICA}} = \hat{A}_{\mathrm{FA}} B^T. \tag{10}$$

This $\hat{A}_{\mathrm{ICA}}$ does not have the ambiguity of rotation.

The mixing matrix $\hat{A}_{\mathrm{ICA}}$ is important when we reconstruct the data from independent sources. Each column of $\hat{A}_{\mathrm{ICA}}$ corresponds to the coefficient of how each independent component contributes to sensory inputs. Let $\hat{A}_{\mathrm{ICA}} = (\hat{\boldsymbol{a}}_{\mathrm{ICA,1}}, \ldots, \hat{\boldsymbol{a}}_{\mathrm{ICA,m}})$, where $\hat{\boldsymbol{a}}_{\mathrm{ICA,i}}$ is an n-dimensional column vector. The source y_i on the sensors are estimated as $\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{a}}_{\mathrm{ICA},i} y_i$.

When we reconstruct $\boldsymbol{x}$ using all the independent sources as,

$$\hat{\boldsymbol{x}} = \hat{A}_{\mathrm{ICA}}\boldsymbol{y} = \hat{A}_{\mathrm{ICA}} W\boldsymbol{x} = \hat{A}_{\mathrm{ICA}} W(A^*\boldsymbol{s}+\boldsymbol{\epsilon}) = \hat{A}_{\mathrm{ICA}} WA^*\boldsymbol{s} + \hat{A}_{\mathrm{ICA}} W\boldsymbol{\epsilon},$$

where A^* is the true mixing matrix. In the original data, the power of the sensory noise is estimated as

$$E[|\boldsymbol{\epsilon}|^2] = \mathrm{tr}\,\hat{\Sigma}. \tag{11}$$

The power of the sensory noise in $\hat{\boldsymbol{x}}$ is estimated as,

$$E[|\hat{A}_{\mathrm{ICA}} W\boldsymbol{\epsilon}|^2] = \mathrm{tr}\,(\hat{A}_{\mathrm{FA}} Q\hat{\Sigma}Q^T \hat{A}_{\mathrm{FA}}^T), \tag{12}$$

from eq.(9) and (10). This quantity is not affected by B, and minimized when Q is chosen as eq.(8) [13].

2.4 Relation to IFA

In this subsection, we discuss the relation between our method and IFA[3]. IFA is a parametric approach to solve the problem in eq.(3). In IFA, the sensory noise distribution is assumed to be normally distributed with 0 mean, but the covariance matrix Σ is not necessarily diagonal, and $m \leq n$ is not assumed.

The characteristics of IFA is that not only the noise distribution, but also the pdf (probability density function) of source signal $\boldsymbol{s}$ is estimated. The original work[3] defined pdf of each source component s_i with a MOG (mixture of Gaussians) which is defined as,

$$p(s_i) = \sum_{r_i=1}^{n_i} w_{i,r_i} \mathcal{G}(s_i - \mu_{i,r_i}, v_{i,r_i}^2) \tag{13}$$

$$\sum_{r_i=1}^{n_i} w_{i,r_i} = 1, \quad w_{i,r_i} \geq 0, \quad \mathcal{G}(s_i - \mu_{i,r_i}, v_{i,r_i}^2) = \frac{1}{\sqrt{2\pi v_{i,r_i}^2}} \exp\left(-\frac{(s_i - \mu_{i,r_i})^2}{2v_{i,r_i}^2}\right).$$

Here, $n_i \geq 2$ is the number of normal distributions in the pdf of source s_i ($i = 1, \cdots, m$), and w_{i,r_i}, μ_{i,r_i}, and v_{i,r_i}^2 ($r_i = 1, \cdots, n_i$) are the mixing weight, mean, and variance of each normal distribution. With these distributions, $p(\boldsymbol{x}, \boldsymbol{s})$ is defined as,

$$p(\boldsymbol{x}, \boldsymbol{s}) = \mathcal{G}(\boldsymbol{x} - A\boldsymbol{s}, \Sigma) \prod_{i}^{m} p(s_i) = \mathcal{G}(\boldsymbol{x} - A\boldsymbol{s}, \Sigma) \prod_{i}^{m} \sum_{r_i=1}^{n_i} w_{r_i} \mathcal{G}(s_i - \mu_{i,r_i}, v_{i,r_i}^2). \tag{14}$$

It is easy to check by expanding eq.(14) that $p(\boldsymbol{x}, \boldsymbol{s})$ is a mixture of $\prod_i^m n_i$ normal distributions. We can define log-likelihood function $\log p(\boldsymbol{x})$ from eq.(14), and the EM algorithm is derived to obtain MLE[3]. Consequently, it is possible to estimate mixing process, the sensory noise distribution and the source distribution. And we can estimate the conditional source distribution $p(\boldsymbol{s}|\boldsymbol{x})$ and reconstruct each source signal from the distribution.

Though IFA seems to be a natural approach, there are practical problems. One problem is the choice of n_i. This is important for the accurate estimation of $p(s_i)$ but it is difficult to choose n_i. Another problem is the calculation cost of the estimation. Large n_i results in slow convergence of estimation iterations. In the process of the EM algorithm or any other gradient descent algorithms, it is necessary to calculate conditional distribution over $n_{\text{all}} \stackrel{\text{def}}{=} \prod_i^m n_i$ components for each data point iteratively. Since $n_i \geq 2$, we have $n_{\mathrm{all}} \geq 2^m$ components. Therefore, the estimation process is not tractable unless m is small. We will show the case where $m = 2$ in section 3, but in the case of MEG brain data analysis, estimated m is more than 15 (section 4.3), and it is impossible to calculate conditional distribution of 2^{16} or more components for each iteration. Finally since estimation process is not easy, we cannot apply statistical model selection method for estimating m.

IFA is theoretically interesting and a general parametric model for noisy observation of linearly mixed independent sources. But for the case of MEG data analysis, it does not give a practical method for estimation. Our method gives a practical method for estimation and it is a semi parametric approach which does not assume particular source distribution.

3 Experiment with synthesized data

First, we used speech data, which are recorded separately and mixed on the computer.

The source data are shown in Fig.1. We have two sources and the power of each source is normalized to be 1 ($E[s_i^2] = 1$). In this experiment, there are 7 sensors and sensory noise is added to each sensory inputs independently as,

$$\boldsymbol{x} = A^* \boldsymbol{s} + \boldsymbol{\epsilon} \tag{15}$$
$$\boldsymbol{x}, \boldsymbol{\epsilon} \in R^7, \qquad \boldsymbol{s} \in R^2, \qquad A^* \in R^{7\times 2}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{o}, \Sigma^*).$$

The observations x_i are shown in Fig.2. The total sum of the power of the source signals is equal to that of $\boldsymbol{\epsilon}$, which means,

$$E[|A^* \boldsymbol{s}|^2] = \mathrm{tr}\ \left(A^* E[\boldsymbol{s}\boldsymbol{s}^T] A^{*T}\right) = \mathrm{tr}\ \left(A^* A^{*T}\right) = \mathrm{tr}\ \Sigma^* = E[|\boldsymbol{\epsilon}|^2]. \tag{16}$$

The data are noisy and it is impossible to use PCA for sphering. And the number of the sources is assumed to be unknown.

First, factor analysis is applied to the data, and the number of the sources m, the mixing matrix $\hat{A}_{\mathrm{FA}}$ and the noise covariance Σ are estimated. In this case, $m \leq 3.7251\ldots$ should be satisfied from eq.(7), and the candidates for the number of the sources are $1, 2,$ and 3.

We used MDL for the estimation of m. For comparison, we show AIC for the candidates in Table 1. Both MDL and AIC selected two as the source number. But AIC gives a very small difference between $m = 2$ and 3. This is the reason why we use MDL for the information criterion in our method. We selected the source number as 2 and obtain the pseudo-inverse of $\hat{A}$ as $Q \in R^{2\times 7}$ by eq.(8). The data $\boldsymbol{x}$ is transformed to $\boldsymbol{z}$ by Q as,

$$\boldsymbol{z} = Q\boldsymbol{x}.$$

Figure 3 shows the sphered data. We can see that the signals are orthogonal, that is, uncorrelated to each other and, pre-processing is almost completed. The remaining problem is to estimate the rotation matrix.

The demixing rotation matrix $W \in R^{2\times 2}$ is estimated with JADE. This is equivalent to making the fourth order cumulant uncorrelated. Finally, we linearly transform the signal as,

$$\boldsymbol{y} = BQ\boldsymbol{x} = W\boldsymbol{x} = WA^*\boldsymbol{s} + W\boldsymbol{\epsilon}. \tag{17}$$

Here, $\boldsymbol{y} \in R^2$ and it still includes compressed noise. However, the original sources are recovered very well (Fig.4).

In this experiment, we know the true mixing matrix A^*, therefore we can calculate the cross-talks as the ratio of diagonal and off diagonal components of the matrix WA^*. The cross talk is 0.92% in $y_1(t)$ and 0.57% in $y_2(t)$.

We tried PCA on the data and selected two dimensional subspace based on the eigenvalues. After reducing the input dimension, we applied JADE to the data and estimated a rotation matrix. For PCA+JADE, the cross talk is 14.8% in $y_1(t)$ and 19.9% in $y_2(t)$.

The estimated covariance matrices and true covariance matrices are shown as 2 dimensional ellipsoids in Fig.5. It is shown that our method gives good estimation, but in the case of PCA+JADE, the signals don't become orthogonal to each other by the pre-processing, and the separation matrix is not estimated correctly.

We also applied IFA for comparison. Two normal distributions are used for modeling each source signal. It took a lot of time for the estimation and the result is shown in Fig.6. The cross talk is 1.92% in $y_1(t)$ and 6.40% in $y_2(t)$. For the estimation of the source signals, we are not using linear transformation, but we estimated the sources by using $p(\boldsymbol{s}|\boldsymbol{x})$ as,

$$\hat{\boldsymbol{s}} = E_{p(\boldsymbol{s}|\boldsymbol{x})}[\boldsymbol{s}] = \int \boldsymbol{s} p(\boldsymbol{s}|\boldsymbol{x}) d\boldsymbol{s}$$

from observation $\boldsymbol{x}$. This is a sort of non-linear transformation from $\boldsymbol{x}$ to $\boldsymbol{s}$, and we cannot have covariance matrix expressions as in Fig.5.

4 MEG data analysis

4.1 The characteristics of MEG data

We applied our method to MEG data. Before going to the detail of the experiment, we discuss the characteristics of the MEG data.

MEG measures the change of magnetic field caused by the brain activity with a lot of (50$\sim$200) coils placed around the brain. Because the change in the magnetic field is directly connected to the nerve activities, MEG can measure the brain activities without any delay. The sampling rate is around 1kHz which is much higher than other techniques such as MRI. We can estimate the location of the sources by solving an inverse problem, and the resolution can be a few mm. Therefore, MEG is a technique which measures the brain activity with high time and spatial resolutions without causing any damage to the brain.

Since the change of the magnetic field caused by the brain activity is extremely small ($\sim 10^{-14}$ T), we need special device called SQUID (super-conducting quantum interference device). The device can detect the brain signal, but the signal contains a lot of environmental noise. We can categorize the noise into two major categories. One is called the artifacts and the other is the quantum mechanical sensory noise. The artifacts include the noise from electric power supply, the earth magnetism, heart beat, breathing and the brain activity which we are not interested in. These artifacts effect on all the sensors simultaneously. On the other hand, the quantum mechanical sensory noise originates from the SQUID itself. SQUID is measuring the magnetic field in liquid Helium. Under this low temperature, sensors cannot avoid to have quantum mechanical noise which is white and independent to each other. The main technique used so far to lower the artifact and the sensory noise is the averaging. The experimentalists usually repeat the same experiment from 100 to 200 times, and then, average the recorded response. But the averaged data still contain a large amount of noise, which are still harmful for the data analysis. After averaging the data, we modeled the data as,

$$\boldsymbol{x} = A\boldsymbol{s} + \boldsymbol{\epsilon} \tag{18}$$

$$\boldsymbol{s}\text{: sources and artifacts,}$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{o}, \Sigma)\text{: quantum mechanical sensory noise.}$$

We then applied our algorithm to MEG data. In the following subsections, we show the results of phantom data and real brain data.

4.2 Experiment with phantom data

We show the result of our algorithm applied to phantom data in this section. The phantom was designed to be roughly the same size as the brain and there is a small platinum electrode inside. We designed the current signal supplied to the electrode to be a 20Hz triangle wave, and averaged the data for 100 trials. Figure 7 shows 5 signals out of 126 active sensors.

The number of the sources and the separation matrix are estimated by our method. We pre-processed the data with factor analysis and estimated the number of the sources by MDL. In this experiment, the number of the sources is estimated as 3. After the pseudo-inverse matrix is multiplied to the observed data, rotation matrix is estimated by JADE. The estimated independent signals and their estimated powers in the total observations are shown in Fig.8. We can see that independent component Y_1 has a large power 86.2% and both of Y_2 and Y_3 has less power than 1%. From the figure, we can see that sensory noise has more power than 10% even after averaging over 100 times.

In this experiment, we know the input to the electrode is Y_1 in Fig.8. After selecting the source, we want to reconstruct the signal. The recovered signals are shown in Fig.9. The artifact was removed compared to Fig.7.

The reason we can have better result is not only from subtracting Y_2 and Y_3 from the observation, but as we wrote, we compressed the sensory noise and it helps to reduce the sensory noise. In the original data, the estimated total power of sensory noise is $\mathrm{tr}\,\hat{\Sigma}$ from eq.(11), and the noise in the recovered signal is estimated as $\mathrm{tr}\,(A_{\mathrm{AIC}} W \hat{\Sigma} W^T A_{\mathrm{AIC}}^T)$ from eq.(12). The noise in the recovered signal is usually much smaller than the noise in th original signlas. Even if we don't subtract Y_2 and Y_3, the amount of the sensory noise in the recovered signal is 1.5% of the original sensory noise in this experiment.

Figure 10 shows estimated strength of each source and noise on the sensors. The strength of source is estimated as the component of matrix $\hat{A}_{\mathrm{ICA}}$. The strength of noise is estimated as square root of the diagonal component of the matrix $\hat{\Sigma}$. From the graph, we can see that some sensors contain much more noise and artifacts than the signal even after averaging over 100 trials. This result shows that we cannot trust the intensities of the signals on sensors sometimes. In order to estimate the location of the signal in the brain, it is necessary to know the intensity ratio of the signals on all the sensors, but the result shows that artifacts and the sensory noise make it difficult to know the true ratio.

4.3 Experiment with brain data

Finally, we applied our method to the data of the brain activity evoked by visual stimulation. The expected results of ICA applied to MEG data analysis can be summarized as follows.

1. Separating artifacts from brain signals.
2. The independent brain activities coming from different parts of the brain to be separated.

For the first part, we believe this is possible because the artifacts and the brain signals would be independent. But for the second part, we don't know if brain activities which are coming from different parts are independent to each other. Probably it is more natural to think that they might be dependent. This is a difficult problem of this study forcing us to go further generalization of the ICA framework.

We show the separated independent components obtained b our method in Fig.12. Some kind of visual stimulations are given to a subject. The data are recorded by 114 sensors in this case (because only 114 of 129 sensors were working correctly at that moment). The duration of recording is from 100msec before the stimulation to 412msec after the stimulation with 1kHz sampling rate. The same procedure is applied to one subject for 100 times and we averaged the data. Some inputs from the sensors are shown in Fig.11. It is observed that the averaging reduces the noise but still a lot of noise remain in the data.

We applied our method to the data and 19 independent components were selected by MDL in this experiment. The independent components are shown in Fig.12. The power of each independent component and sensory noise in the observation is shown in Fig.14(left). It shows that even after averaging over 100 times, there are more than 10% of sensory noise. Figure 14(right) shows the sensory noise on each sensor, and we can see some sensors have a lot of noise and some does not have much. The method is applied to the data from different subjects (4 more), and in all the cases, the selected numbers of sources are roughly the same (from 16 to 19).

Figure 13 shows the result of PCA+JADE. We used PCA to compress the data into 19 dimension, and then used JADE. Some independent components are similar to those in Fig.12, but some are not. For example Y_9 in Fig.12 is mainly a harmonic (180Hz) of the electric power supply (60Hz) but Fig.13 does not have it.

Based on the results in Fig.12, we have to separate the sources from the artifacts in which we are not interested. For example, $y_8(t)$ has a large value at the very end of the record. This seems to be a kind of mechanical or software noise. A harmonic of the electric power supply, $y_9(t)$ is an artifact. But we cannot know if they are due to brain responses or not for the rest. Fortunately, this experiment is designed for studying evoked responses by visual stimulation, and we are not interested in the components which have some power before the stimulation was given to the subject. Therefore, we defined a threshold of a power such that, if a signal has some power before the stimulation, we regarded the signal as an artifact. In this experiment, we set the threshold as follows, if the averaged power of first 100msec is equal to or larger than 0.9 times of the averaged power of the whole part of that component, we regarded it as an artifact. We added one more criterion that if the estimated averaged power of an independent component on all the sensors is smaller than a threshold (We set the threshold as 3% and it is shown in Fig.14), we assume the signal is an artifact. In this case, the selected sources are $y_1(t)$, $y_2(t)$, $y_3(t)$, $y_4(t)$, $y_5(t)$, $y_6(t)$, $y_7(t)$, and $y_{11}(t)$. They are shown in Fig.12 with solid boxes.

After picking those sources up, we put them back to the orignal sensor signal space, and the result is shown in Fig.15. The noises are removed and the data becomes clear.

One of the difficulties in evaluating the results of brain data analysis is that we don't know the true signal. This is a big problem in order to know how well our algorithm is working. We can see the cleaned outputs of the sensors as Fig.15, but we further want to know the relationship between the independent components and brain activities more directly.

There are a lot of methods to see the relationships visually. One popular method is the dipole estimation. In the dipole estimation, we describe the brain activities by dipoles in the brain. In the estimation, we have to specify the number of the dipoles, and we solve an inverse problem numerically. Since the choice of the number in this experiment is difficult, we did not use dipole estimation, but we implemented SF (spatial filter) technique[15]. SFs are a set of virtual sensors which are located on a hemisphere defined on the brain. We can estimate the current flows on those virtual sensors which describe the MEG observations well. This is an inverse problem and it can be solved in a form of a linear mapping from the MEG sensors to SFs.

One of the characteristics of SFs is that we can put a set of virtual sensors at any place we want, so that, we can estimate the activities of any place on the brain virtually by a linear projection. But from the nature of this technique, the sensors at the boundary of the hemisphere are not reliable[15]. Therefore, it is important to place the center of the hemisphere at the position of our interests. In our experiment, the part of the brain we are interested in is the visual cortex, and we put 21×21 SFs on a hemisphere, whose center is located at V1.

Figure 16 shows the output of the SFs before and after the method. The original data include a lot of noise and we cannot know if there is useful information (upper row). But after applying factor analysis and ICA, those signals which we are not interested in are removed well (lower row). The response of the brain in this experiment is known to be high around 100msec after the stimulation[15]. And the characteristics is preserved very well.

5 Discussion

In this article, we proposed a new combination of methods having different backgrounds to analyze biological data. The data includes strong sensory noise which is true in most cases of biological data. We applied the algorithm to MEG data, and have shown the approach is effective. We can estimate the number of the sources, and the power of the sensory noise which is independent to each other. This is one of the serious problem which has not been well treated in conventional ICA approaches, and this article gives one effective method. H. Attias has proposed IFA to solve this problem. It gives the source distribution by MOGs and the problem is solved as a parametric estimation. But we have proposed a different method based on the semi parametric approach which is an attractive point of ICA.

This approach is a natural extension of the standard ICA approaches which use PCA as the pre-processing and then higher order statistics for the second step. We assumed the sensory noise is normally distributed, and applied factor analysis for the pre-processing instead of PCA. After the pre-processing, we applied an ICA method based on higher order statistics for estimating the rotation matrix. It is already said that the estimation based on higher order statistics would not be affected by the sensory noise if it is normally distributed. However, we can skip the pre-processing only if the number of the sources is known beforehand. Our approach estimates not only the mixing system but also the number of the sources.

There still remain some open problems. In the factor analysis, there are a lot of methods to estimate the parameters and the number of the sources, and each method has a characteristic. We applied MLE for estimation and MDL for estimating the number of the sources. But there are different combinations, and there might be a method which suits better for some particular problems. We have the same problem for ICA algorithms. We used JADE but there might be a better algorithm. Another problem is the noise distribution. We assumed normal distributions, but if we can have a better model for the sensory noise, the algorithm will be improved further. Another big problem for MEG data is the choice of brain sources. We made some criterion for the choice of sources, but this is only obtained through trial and error. The choice of the thresholds and the choice of criterion itself is an open problem.

We can also check the algorithm from the viewpoint of factor analysis. How to determine the rotation is one of the traditional problems in factor analysis. There are a lot of algorithms proposed for this purpose, but it is not common to use higher order statistics. Therefore our approach gives a new pathway to factor analysis, too.

Acknowledgment

We are grateful to Shun-ichi Amari and Noboru Murata for their comments and suggestions for our research. We also thank Shigeki Kajihara for his effort on the MEG measurements, and Shimadzu Inc. for giving us the MEG data.

References

[1] Amari, S., & Cardoso, J.-F. (1997). Blind source separation – semiparametric statistical approach. *IEEE Trans. Signal Processing*, *45(11)*, 2692–2700.

[2] Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, (Volume 5, pp. 111–150). Berkeley:University of California Press.

[3] Attias, H. (1999). Independent factor analysis. *Neural Computation*, *11(4)*, 803–851.

[4] Bell, A. J., & Sejnowski, T. J. (1995). An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

[5] Cao, J., Murata, N., Amari, S., Cichocki, A., & Takeda, T. (1998). ICA approach with pre & post-processing techniques. *Proceedings of 1998 International Symposium on Nonlinear Theory and its Applications (NOLTA'98)*, (Volume 1, pp. 287–290).

[6] Cardoso, J.-F. (1999). Higher-order contrasts for independent component analysis. *Neural Computation*, *11(1)*, 157–192.

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

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

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

[10] Hyvärinen, A., Särelä, J., & Vigário, R. (1999). Spikes and bumps: Artifacts generated by independent component analysis with insufficient sample size. *Proceedings of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, (pp. 425–429).

[11] Ledermann, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika*, *2*, 85–93.

[12] Makeig, S., Jung, T.-P., & Sejnowski, T. J. (1996). Using feedforward neural networks to monitor alertness from changes in EEG correlation and coherence. D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, (Volume 8, pp. 931–937), The MIT Press, Cambridge, MA.

[13] C. R. Rao. *Linear Statistical Inference and Its Applications Second Edition*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1973.

[14] Reyment, R., & Jöreskog, K. G. (1993). *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press.

[15] Toyama, K., Yoshikawa, K., Yoshida, Y., Kondo, Y., Tomita, S., Takanashi, Y., Ejima, Y., & Yoshizawa, S. (1999). A new method for magnetoencephalography: A three dimensional magnetometer-spatial filter system. *Neuroscience*, *91(2)*, 405–415.

[16] Vigário, R., Jousmäki, V., Hämäläinen, M., Hari, R., & Oja, E. (1998). Independent component analysis for identification of artifacts in Magnetoencephalographic recordings. M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, (Volume 10, pp. 229–235), The MIT Press, Cambridge MA.

[17] Vigário, R. N. (1997). Extraction of ocular artifacts from EEG using independent component analysis. *Electroenceph. clin. Neurophysiol.*, *103*, 395–404.

Table 1: MDL and AIC for the candidates: In the experiment, the candidates source numbers are 1, 2 and 3. The MDL and AIC for these candidates are shown in the table.

# of the sources	1	2	3
MDL	4.0870	3.8911	3.8928
AIC	4.0826	3.8849	3.8850

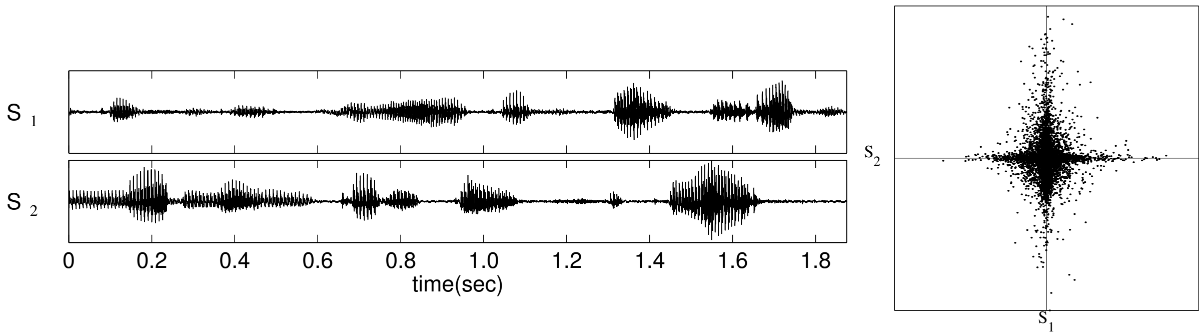


Figure 1: Source sound signals: Both of them are speech signals recorded separately with 16kHz sampling rate. Data are 1.875sec length(30,000data points). Right graph plots the data in the two dimensional space.

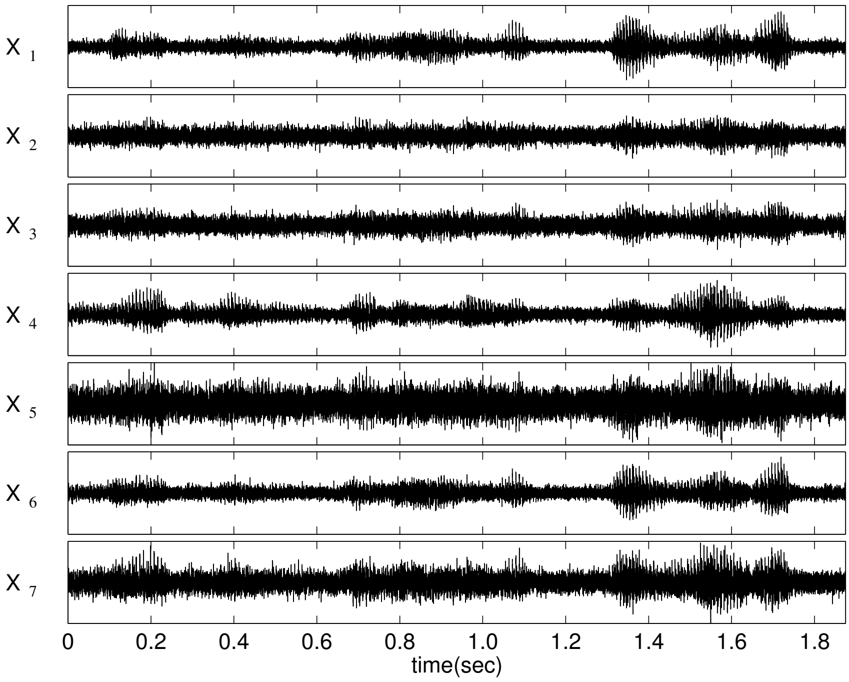


Figure 2: Input signals: Two sources are distributed to seven inputs and sensory noises of normal distributions are added to the seven sensors independently.

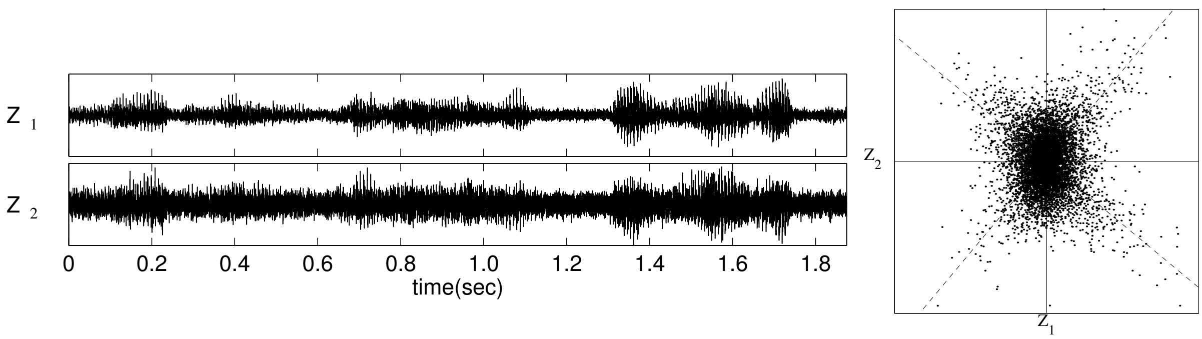


Figure 3: Sphered data: After applying factor analysis to the input signals, two pre-processed time series are obtained.

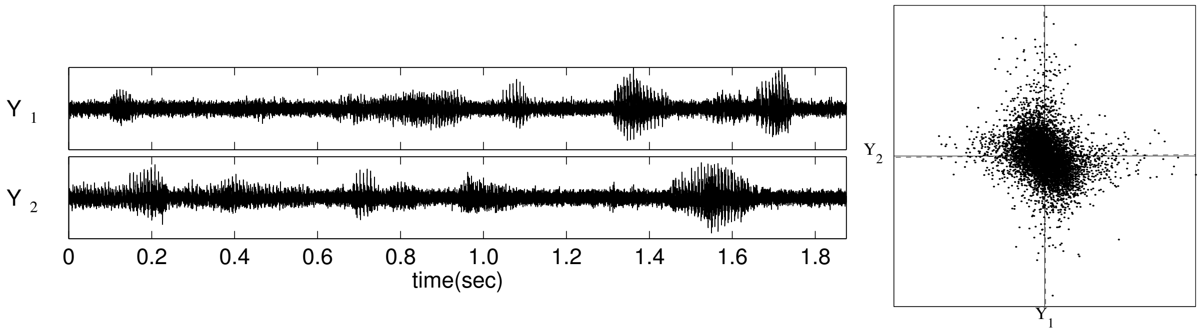


Figure 4: Output data, after using the ICA algorithm: The signals are separated by our method and the outputs are shown.

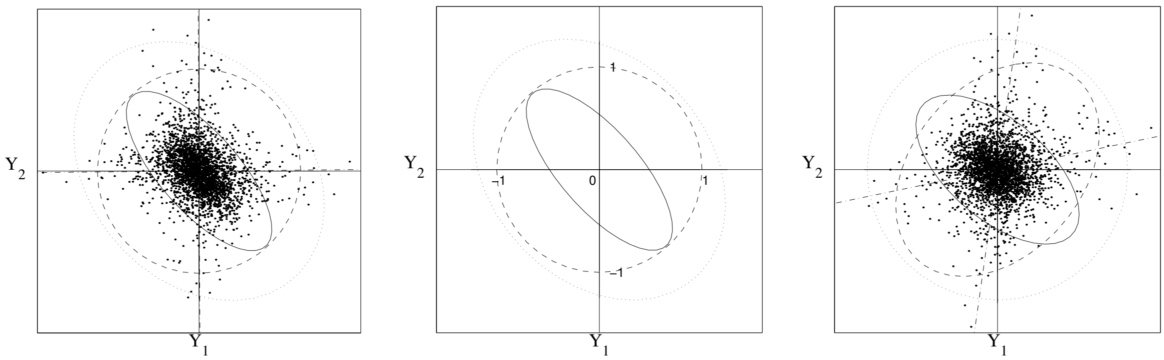


Figure 5: Features of separated signals in the mean of covariance matrices: Separation matrix $W \in R^{2\times 7}$ is estimated by Factor Analysis+ JADE as W_{FA} (left) and by PCA+JADE as W_{PCA}(right). For the comparison, pseudo-inverse of true mixing matrix $W_{\mathrm{true}} = A^{*-}$ is used (center). Solid ellipse shows the covariance matrix of the projected sensory noise term $W_* \boldsymbol{\epsilon}$, dashed ellipse shows the covariance matrix of the projected source signal term $W_* A^* \boldsymbol{s}$. Dot ellipse shows the sum of above two covariance matrices, $\boldsymbol{y} = W_* \boldsymbol{x}$. Dashed dot lines show how the axes of the sources, which are x and y axes in the center figure, was projected by the separation matrix. The data points shown in the left and right graphs were rescaled to match the size.

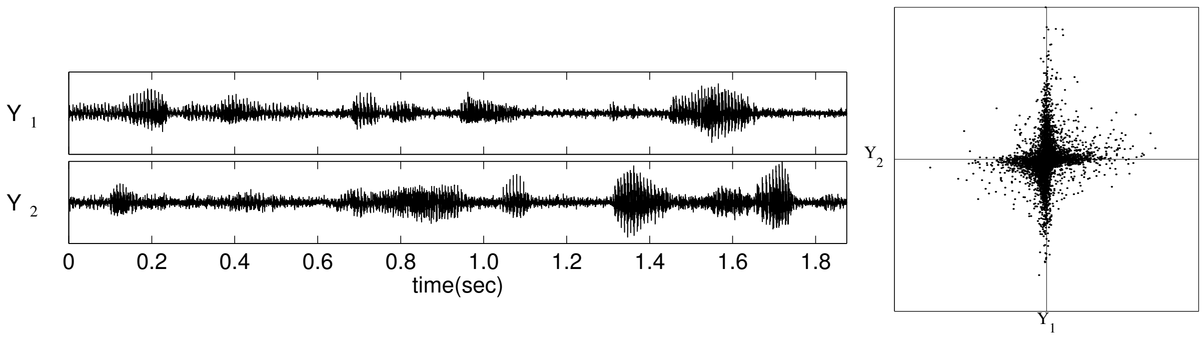


Figure 6: Result of IFA: Source signals are recovered by estimating $E_{p(\boldsymbol{s}|\boldsymbol{x})}[\boldsymbol{s}]$ from each observed data.

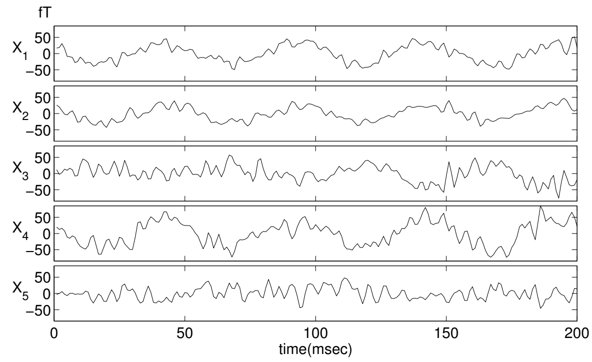


Figure 7: Averaged sensory inputs of phantom data: The phantom has single electrode inside and 20Hz triangle wave is the input. The input amplitude is adjusted to match the signal from a brain.

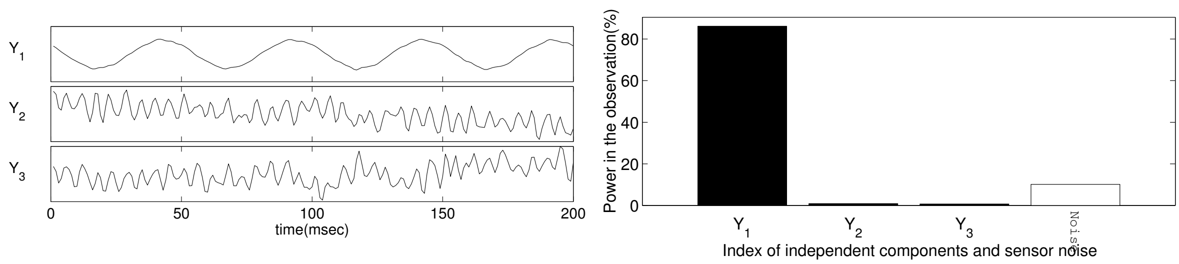


Figure 8: Estimated independent components and their powers: Left figure shows the result of our method. 3 independent components are obtained. First component has 20Hz as its major frequency component, and the rest 2 signals main frequency components are 180Hz. Right figure shows the power of each component in the observation and the estimated sensory noise power in the observation.

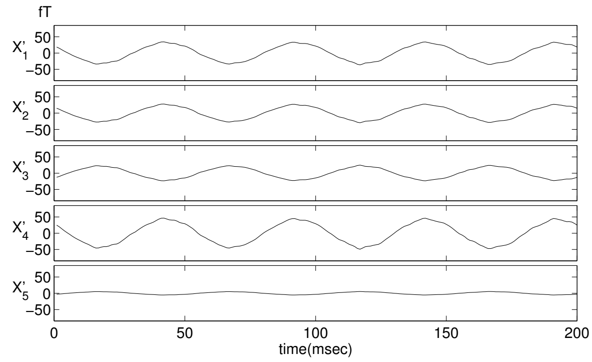


Figure 9: Estimated independent components on the sensors: The independent component y_1 is put back to the original sensors by a linear mapping

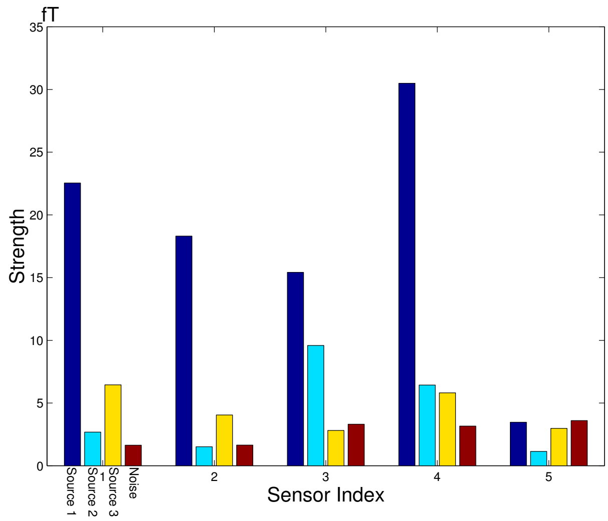


Figure 10: Estimated strength of each source and noise on the sensors: The estimated strength of the 3 independent components in Fig.8 and the sensory noise which is estimated as Σ are shown.

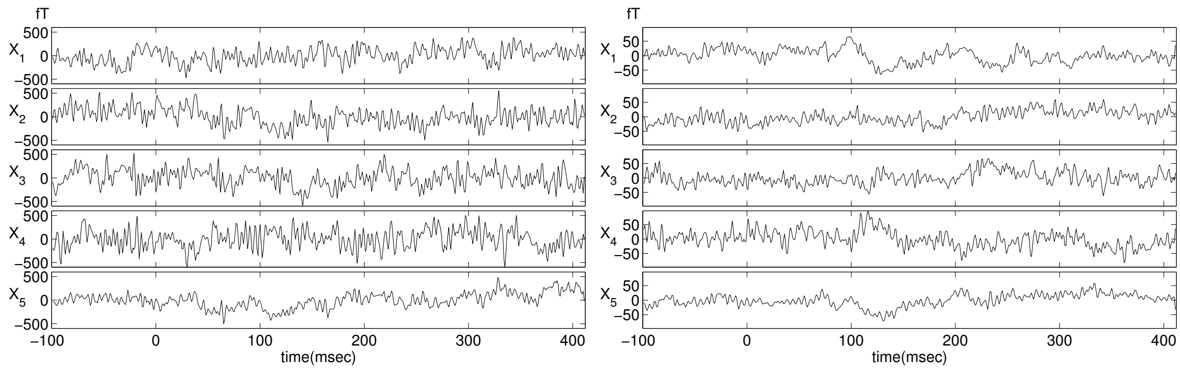


Figure 11: The MEG data: Left side is showing single record of 5 sensors, and the right side is showing the data on those sensors after averaging over 100 trials. 0msec in time axis shows the trigger of the visual stimulation.

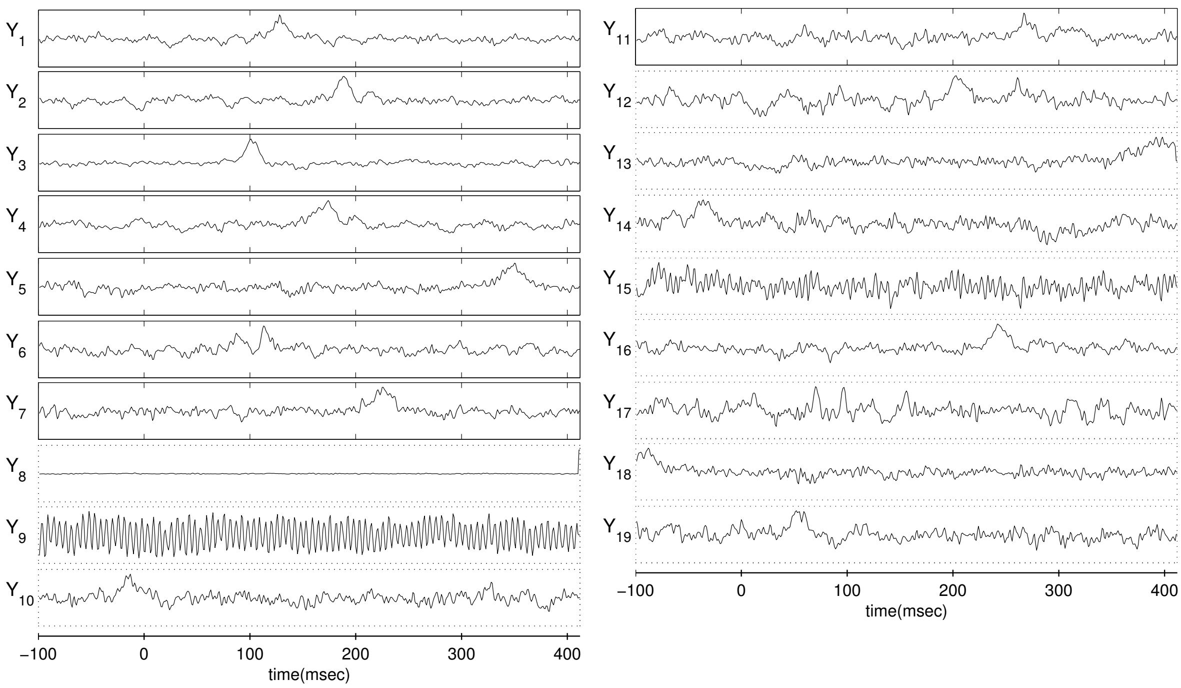


Figure 12: Separated visual evoked response signals: The independent components are aligned along descending order of their averaged powers on the MEG sensors. The signals surrounded by solid boxes are selected as source signals.

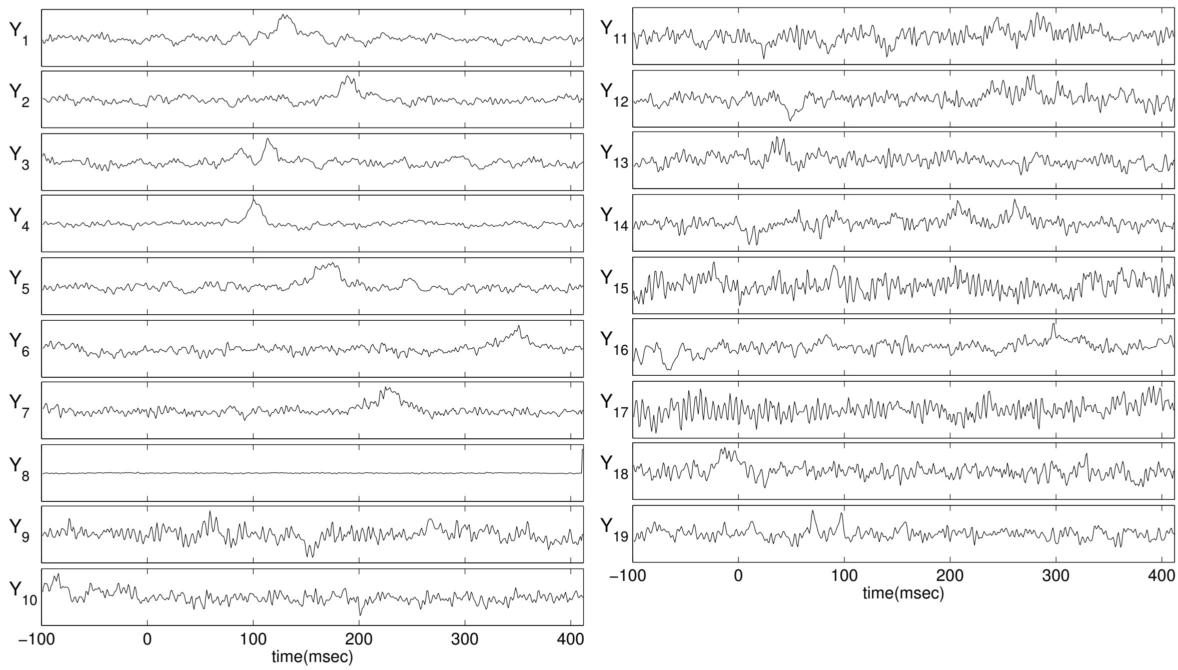


Figure 13: Separation result of using PCA+JADE: Observed data are compressed to 19 dimension using PCA and JADE was applied after PCA.

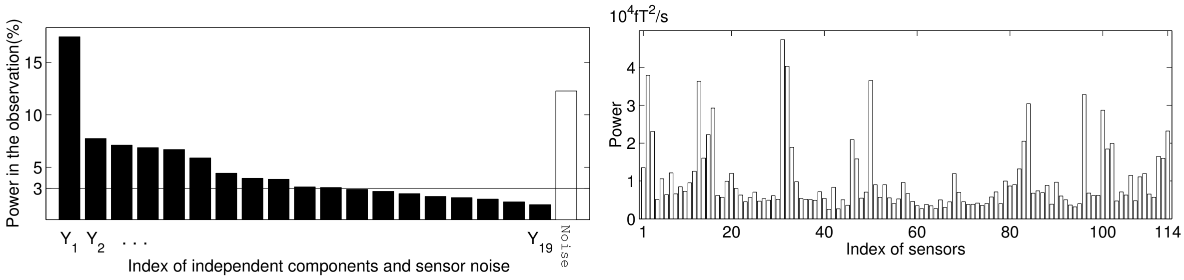


Figure 14: Power of signals and noise: Left figure shows the power of each independent component and sum of the sensory noise powers which is shown as a different color, and the line in the figure shows 3% which was used as the threshold in the selection of sources. Right figure shows the estimated sensory noise on each sensor.

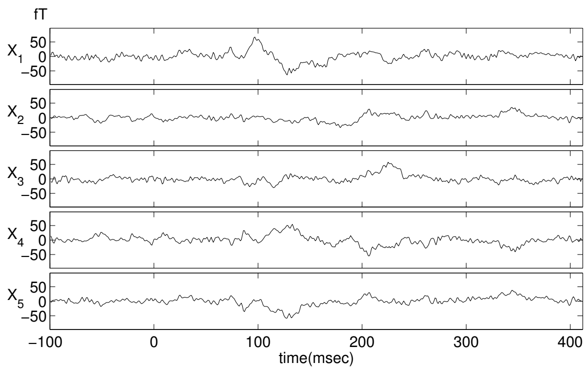
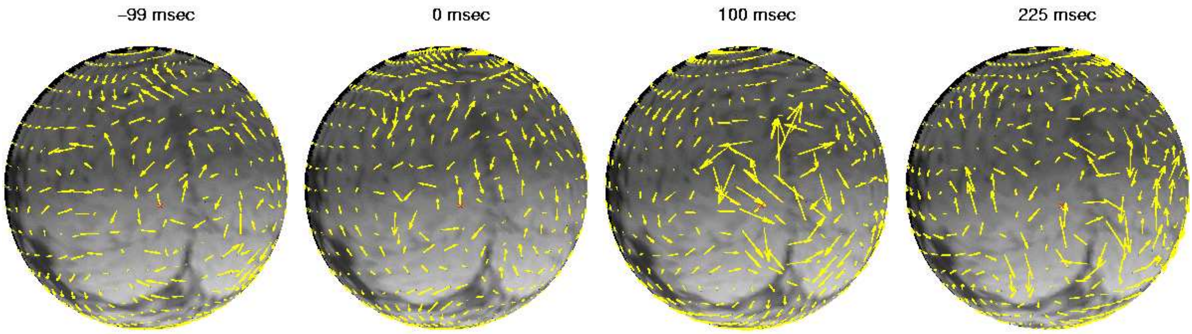


Figure 15: The recovered MEG data by removing the artifacts: The independent components Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, and Y_{11} in Fig.12 are picked up and put back to the sensors by linear mapping.

The original data



After applying our method

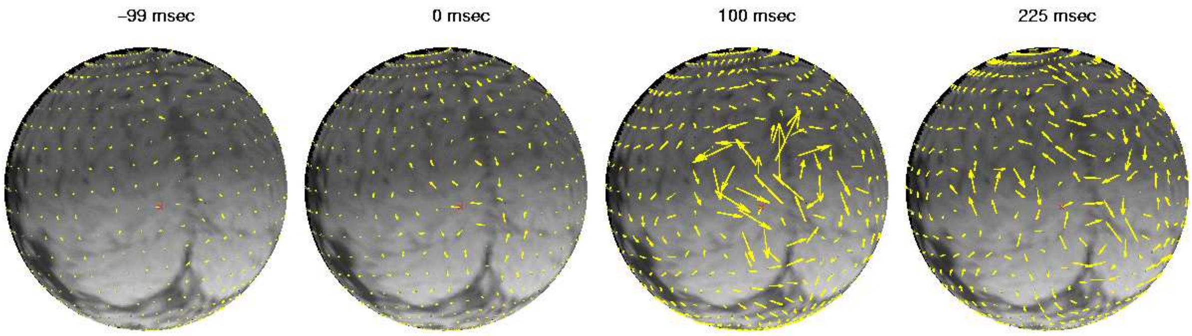


Figure 16: Result of the approach applied to MEG data: Upper row shows the original data and the lower row shows the results of our method. The arrows in the figures are the estimated currents of the virtual sensors. The arrows are superimposed on the figure of the brain obtained by MRI. Signals are recorded from 100msec before to 412msec after the visual stimulation. 0msec in the time axis is the trigger of the stimulation. Red dots in the figures show the position of visual cortex V1.