Comparison between Principal Component Analysis and Independent Component Analysis in Electroencephalograms Modelling

C. Bugli[*,1] and **P. Lambert**[1,2]

[1] Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays, 20, Louvain-la-Neuve, B-1348, Belgium

[2] Unité d'épidémiologie, biostatistique et méthodes opérationnelles, Faculté de Médecine, Université catholique de Louvain, B-1200 Bruxelles, Belgium

Received 4 August 2005, revised 24 March 2006, accepted 26 June 2006

Summary

Principal Component Analysis (PCA) is a classical technique in statistical data analysis, feature extraction and data reduction, aiming at explaining observed signals as a linear combination of orthogonal principal components. Independent Component Analysis (ICA) is a technique of array processing and data analysis, aiming at recovering unobserved signals or 'sources' from observed mixtures, exploiting only the assumption of mutual independence between the signals. The separation of the sources by ICA has great potential in applications such as the separation of sound signals (like voices mixed in simultaneous multiple records, for example), in telecommunication or in the treatment of medical signals. However, ICA is not yet often used by statisticians. In this paper, we shall present ICA in a statistical framework and compare this method with PCA for electroencephalograms (EEG) analysis.We shall see that ICA provides a more useful data representation than PCA, for instance, for the representation of a particular characteristic of the EEG named event-related potential (ERP).

Key words: EEG; ERP; ICA; PCA; Statistical independence.

1 Introduction

Principal Component Analysis (PCA) is a classical technique in statistical data analysis, feature extraction and data reduction. Given a set of multivariate measurements, the purpose is to find a smaller set of variables with less redundancy, that would give as good representation as possible. The redundancy is measured by correlations between data elements. Therefore, the analysis can be based on second order statistics only. As PCA considers second order moments only it lacks information on higher order statistics. Independent Component Analysis (ICA) is a technique data analysis accounting for higher order statistics. ICA is a generalisation of PCA. Moreover, PCA can be used as preprocessing step in some ICA algorithm. ICA has great potential in applications such as the separation of sound signals, in telecommunication or in the treatment of medical signals. However, ICA is not yet often used by statisticians.

We shall present this technique from a statistical point of view in the first section and compare it to PCA in Section 2.3. Next, we shall compare the performance of PCA and ICA for two applications on electroencephalograms (EEG) modelling (see Makeig et al., 1996 for an application of ICA to EEG analysis). The first application is the reduction of dimensionality (Section 3). The second application

[*] Corresponding author: e-mail: bugli@stat.ucl.ac.be, Phone: +32 10 47 41 60, Fax: +32 10 47 30 32

(Section 4) is the modelling of an interesting characteristic of EEG signals named event related potential (ERP). We shall see that ICA provides a more powerful data representation than PCA in the EEG analysis context.

1.1 Event-related potential

EEG activity is present in a spontaneous way. It is affected by external stimuli (e.g. tone or light flash). The alteration of the ongoing EEG due to stimuli is named an event related potential (ERP).

ERP can be visualised during a short period following the stimulation time, with a response pattern which is more or less predictable under similar conditions. We call an ERP episode, the EEG signals observed several seconds after a stimulation. The amplitude of ERP is low comparing with the ongoing EEG. So, it is common practice to average in time several ERP episodes aiming to increase the signal to noise ratio in order to visualize the evoked activity.

The averaged ERP episodes present some well-known peaks (see Figure 1a). The ones usually pointed are the P100 or P1 (peak latency approximately 100 ms after stimulation), the P200 or P2 ($\sim$200 ms), and the P300 or P3 ($\sim$300 ms) peaks (see Figure 1a). The P100 peak can be observed after non-target and target stimulation. Since this peak does not depend on the task, it has relatively short latency ($\sim$100ms). P200 and P300 are typically only found after the target stimulation. Since these responses are task dependent and have long latency, they are traditionally related to cognitive processes as signal matching, decision making, attention, memory updating, etc. The P300 peak has the highest amplitude and, hence, is easier to detect.

The P300 peak is a good indicator of brain performance. It is often studied by neurophysiologists as an amplitude change or a delay in the occurrence of the peak (This delay is named latency (see Figure 1a).) are signs of memory problem like with Alzheimer's disease or indication that a drug is affecting the brain. P300 can be considered as an expression of the central nervous system (CNS) activity involved in the processing of new information when attention is solicited to update memory representations.

Under the oddball paradigm (Oddball paradigm: Two different stimuli are presented in a pseudo-random order: a frequent non-target stimulus and a deviant stimulus named the target. Subjects are instructed to pay attention to the appearance of the target stimulus.), there are probably two components in the P300 wave. These two components are merged so that it is usually difficult to detect the two corresponding peaks. We only observe the peak of the mixture named P300 until now. The first one of these components might present a peak around 250 ms after stimulus. This component named P3a has generally the maximum amplitude in the frontal area. The second component might present a

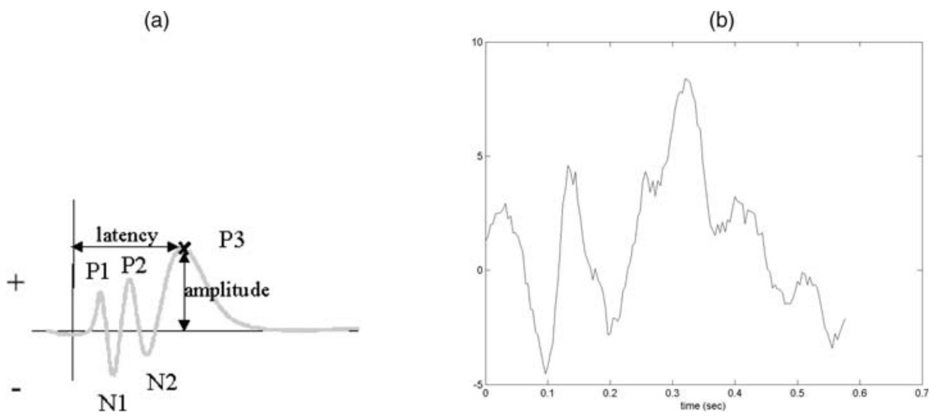


Figure 1 (a) Typical peaks in ERP. (b) Example of time averaged of P300 curve for electrode Cz.

peak around 350 ms after stimulus. This component named P3b has often its maximum amplitude in the parietal area. P3a is supposed to correspond to the stimulus detection. It can be linked up to attention allocation brought by an unexpected stimulus. P3b is supposed to be associated with discrimination, classification, selection and decision (see Polich and Kok, 1995).

Usually, the P300 peak is visualised by averaging the EEG signals over some ERP episodes for the post infrequent stimuli periods on all the electrodes. A global interpretation is made from the results on all the electrodes. Some studies use only some electrodes representative of P3a and P3b phenomena. Most of the time, the P3a pattern is more visible on the Fz electrode (in the frontal area) and the P3b waveform is more visible on the Pz electrode (in the parietal area) as presented in Figure 2. However, the P3a and P3b generators are dispersed in the brain (see Halgren et al. 1995 and 1990). This is why there is a possible influence of P3a and P3b on all the electrodes.

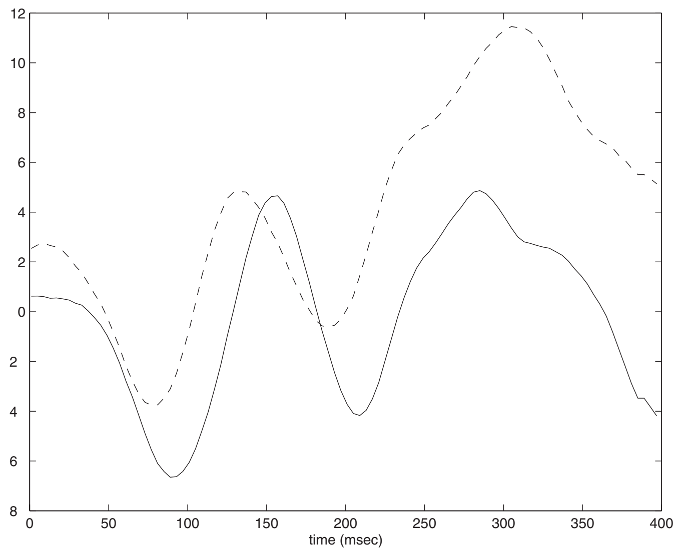


Figure 2 P300 curves obtained by time averaging of electrodes Fz (solid line – frontal position) and Pz (dashed line – parietal position).

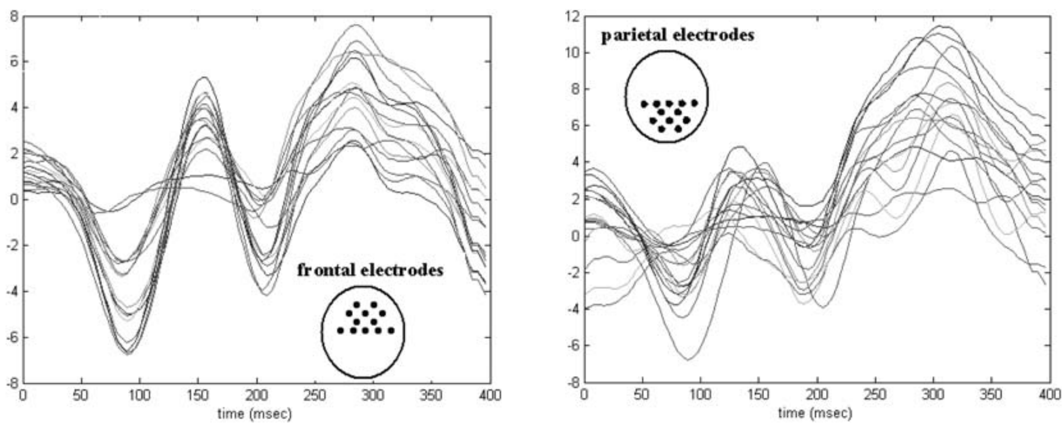


Figure 3 (a) P300 curves for frontal electrodes: large differences between signals. (b) P300 curves for parietal electrodes: large differences between signals.

You can see in Figures 3a and 3b the P300 curves for the different electrodes in respectively the frontal and parietal areas. There are sometimes large differences between signal from electrodes in the same area. Therefore, it seems not sufficient to analyse the P3a and P3b phenomena on only one electrode (Fz or Pz, for example).

2 Independent Component Analysis

Blind signal separation and Independent Component Analysis are techniques of array processing and data analysis, aiming at recovering unobserved signals or 'sources' from observed linear mixtures, exploiting only the assumption of mutual independence between the signals. Typically, the observations are obtained at the output of a set of sensors, each sensor receiving a different mixture of the 'source signals'. The adjective 'blind' stresses the fact that the source signals are not observed and that no information is available about the mixture. This is a useful approach when modelling the transfer from the sources to the sensors is too difficult. It is unavoidable when no reliable a priori information is available about the transfer. The lack of a priori knowledge about the mixture is compensated by a statistically strong but often physically plausible assumption of independence between source signals and of linearity.

The separation of the sources by ICA has great potential in applications such as the separation of sound signals (like voices mixed in simultaneous multiple records, for example), in telecommunication or in the treatment of medical signals (You can see examples in all these fields on the web site: http://web.media.mit.edu/ paris/ica.html (accessed in January 2005).). In recent years, there has been an explosion of interest in the application and theory of ICA, including in statistical applications (see Hastie and Tibshirani, 1991).

In this Section, we shall present ICA and propose mutual information as a way to measure the dependence between 2 random variables. Details about one algorithm named JADE were presented in Cardoso and Souloumiac (1993).

2.1 The ICA game

Independent Component Analysis is a method of extraction of statistically independent signals (called sources) generating simultaneously observed sequences of data. Hastie and Tibshirani (Hastie and Tibshirani, 1991) give a good introduction to ICA. A famous application of ICA is in the 'cocktail party problem', in which a listener is faced with the problem of separating the independent voices chattering a cocktail party.

Each observed signal is assumed to be a linear combination of statistically independent signals. More specifically, we assume the existence of n independent unobserved signals $s_1(t), \dots, s_n(t)$ and the observation of p linear mixtures $x_1(t), \dots, x_p(t)$, at times $t \in [t_0, t_1, \dots, t_{\max}] = \mathcal{T}$ such that $x_i(t) = a_{i1}s_1(t) + a_{i2}s_2(t) + \dots + a_{in}s_n(t)$, for each $i = 1, \dots, p$. This is compactly represented by the matrix equation:

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t)\ \forall t \in \mathcal{T} \tag{1}$$

where $\boldsymbol{s}(t) = [s_1(t), \dots, s_n(t)]^T$ is a $n \times 1$ column vector collecting the n source signals and $\boldsymbol{x}(t) = [x_1(t), \dots, x_p(t)]^T$ is a $p \times 1$ column vector collecting the p observed signals with $n \le p$.

The "mixing matrix" A, a full rank $p \times n$ matrix, contains the linear mixture coefficients. The blind signal separation problem consists in recovering the sequence of source vectors $\{\boldsymbol{s}(t) : t \in \mathcal{T}\}$ using only the observed data $\{\boldsymbol{x}(t) : t \in \mathcal{T}\}$, the assumption of independence between the unobserved components of the input vector $\boldsymbol{s}(t)$ and possibly some a priori information about the probability distribution of the inputs. It can be formulated as the computation of an $n \times p$ "separating matrix" W such that

$$\boldsymbol{u}(t) = W\boldsymbol{x}(t) \tag{2}$$

is an estimate, up to a multiplicative constant and a permutation, of the vector $\boldsymbol{s}(t)$ of the source signals.

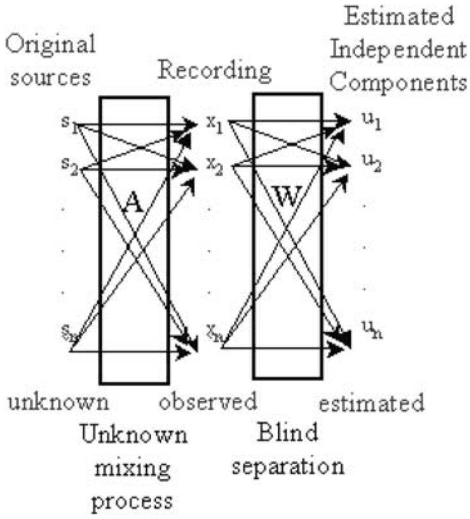


Figure 4 ICA Game.

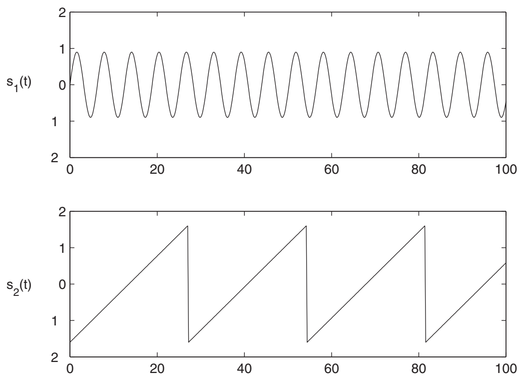


Figure 5 Original sources: $\boldsymbol{s}_1(t)$ and $\boldsymbol{s}_2(t)$.

In practice, an ICA algorithm provides the matrix W such that the signals estimates $\boldsymbol{u}(t)$ are "as statistically independent as possible" (see ICA game in Figure 4).

For example, suppose that we have 2 independent signals $\boldsymbol{s}_1(t)$ and $\boldsymbol{s}_2(t)$ like in Figure 5. After multiplication of these 2 sources by a mixing matrix A, we obtain the observed mixtures of the source signals in Figure 6. ICA find a matrix W such that the components of $\boldsymbol{u}(t) = W\boldsymbol{x}(t)$ are independent. The estimates of the original source signals are estimated by only using the observed signals. As one can see in Figure 7, the original signals were estimated, up to a permutation and a multiplicative constant.

Note that we explain the method with n sources and p observed signals with $n \leq p$. So, we can ask the method to compute less components than observed signals. Then, the dimensions of the matrices of mixing and unmixing are different. We shall suggest a method to determine the required number of independent components in Section 3.2.

2.2 Problem statement and assumptions

In this Section, we first remind the concept of statistical independence. Next, we shall propose mutual information as a way to measure the dependence between 2 random variables. For more details about the contents of this Section, see Herault and Jutten (1991), Comon (1994) and Lee et al. (2000).

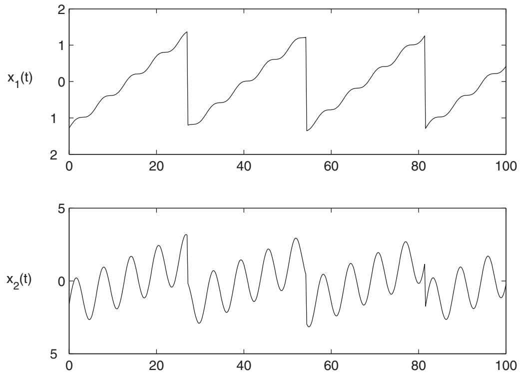


Figure 6 Observed Mixtures: $\boldsymbol{x}_1(t)$ and $\boldsymbol{x}_2(t)$.

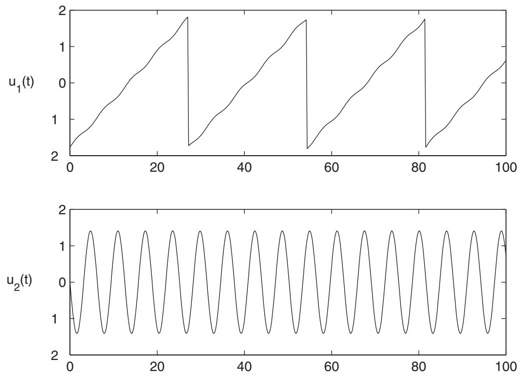


Figure 7 Sources approximated by ICA: $\boldsymbol{u}_1(t)$ and $\boldsymbol{u}_2(t)$. The original signals were very accurately estimated, up to a permutation and a multiplicative constant.

Two random variables X and Y are statistically independent if and only if

$$E(X^p Y^q) - E(X^p)\, E(Y^q) = 0 \; \forall \, p, \quad q > 0\,. \tag{3}$$

It is an extension of the concept of non-correlated random variables for which $E(XY) - E(X)\,E(Y) = 0$. In the ICA problem, the n sources share a n-dimensional probability density function (p.d.f.) $p(s)$ such that $p(s) = \prod_{i=1}^{n} p_i(s_i(t))$, where $p_i(s_i)$ is the density of the i-th source.

We shall now show that mutual information measures the difference in information obtained by observing the 2 random variables separately or jointly. Information theory teaches us that signals x and y are statistically independent if and only if their mutual information $I(x, y)$ equals 0. Indeed, the mutual information I of a p-dimensional random vector X is defined by:

$$I(X) = \int p(x) \log \frac{p(x)}{\prod_{i=1}^{n} p_i(x)} \, \mathrm{d}x\,. \tag{4}$$

It is always non-negative. If the sources are independent, i.e. $p(s) = \prod_{i=1}^{n} p_i(s_i(t))$, then the mutual information of the sources is clearly equal to zero. In practice, the various ICA algorithms will propose different strategies to calculate the "separating matrix" W such that the information of the obtained signals $I(u) = I(Wx)$ is minimal. The ICA algorithm rely on different approximations of the mutual information. Many of the popular approaches to ICA are based on entropy (see Hastie and Tibshirani, 1991). We shall first introduce the concept of entropy and next show its link with mutual information:

$$I(X, Y) = H(X) - H(X \mid Y). \tag{5}$$

The entropy of a p-dimensional random vector X, with density $p(x)$, is defined by:

$$H(X) = H(p(x)) = -\int p(x) \log p(x) \, \mathrm{d}x. \tag{6}$$

It is the basic concept of information theory. The more "random", i.e. unpredictable and unstructured, the variable is, the larger its entropy. A fundamental result of information theory is that a Gaussian variable has the largest entropy among all random variables of equal variance. This shows that the Gaussian distribution is "the most random" or the least structured of all distributions. Entropy is small for distributions that are clearly concentrated on certain values, i.e., when the as a distribution is very "spiky". The joint entropy $H(X, Y)$ of two random variables X and Y is defined as $H(X, Y) = H(p(x, y)) = -\int\int p(x, y) \log p(x, y) \, \mathrm{d}x \, \mathrm{d}y$. It may be interpreted as the information that X and Y jointly encode because it specifies the average uncertainty of a joint measurement of two variables. So, minimizing the mutual information is equivalent to maximizing the information that the signals jointly encode i.e. the joint entropy. Finally, the conditional entropy $H(X \mid Y)$ of the random variables X and Y is defined as $H(X \mid Y) = -\int\int p(x, y) \log p(x \mid y) \, \mathrm{d}x \, \mathrm{d}y$. It measures the remaining uncertainty in X given a knowledge of Y. The mutual information is related to the joint or conditional entropy. One can see that the mutual information of two random variables X and Y equals:

$$\begin{aligned} I(X, Y) &= \int\int p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \, \mathrm{d}x \, \mathrm{d}y \\ &= -H(X, Y) + H(X) + H(Y). \end{aligned} \tag{7}$$

Thus, the mutual information is the difference in the information obtained by observing X and Y separately or jointly. Moreover,

$$\begin{aligned} H(X, Y) - H(Y) &= -\int\int p(x, y) \log p(x, y) \, \mathrm{d}x \, \mathrm{d}y + \int p(y) \log p(y) \, \mathrm{d}y \\ &= -\int\int p(x, y) \log p(x \mid y) \, \mathrm{d}x \, \mathrm{d}y = H(X \mid Y). \end{aligned} \tag{8}$$

Therefore, $I(X, Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$. The information that Y tells us about X is the reduction in uncertainty about X due to the knowledge of Y. So, the information $H(X)$ encoded by X that cannot be obtained by observing Y is $I(X, Y)$.

Gaussian mixture are not separable using ICA. Indeed, linear combinations of statistically independent data components have non-gaussian distribution and linear combination of Gaussian signals has Gaussian distribution. Majority of ICA algorithm maximize non-gaussianity. As discussed in Hyvärinen et al. (2001), there is no way how we could infer the mixing matrix from the mixture. The phenomenon that the orthogonal mixing matrix cannot be estimated for Gaussian variables is related to the property that uncorrelated jointly Gaussian variables are necessarily independent. Thus, in the case of Gaussian independent component we can only estimate ICA model up to orthogonal transformation. In other words, the matrix A is not identifiable for Gaussian independent component. With Gaussian variables, all we can do is whiten the data. PCA is the classic choice.

In this Section, we showed that minimizing mutual information allow to realize statistical independence. Moreover, we showed that mutual information is related to conditional entropy: minimizing mutual information is equivalent to maximizing conditional entropy. Therefore, a lot of ICA algorithms use this relation to setup an objective function which can be optimised to obtain statistically independent sources. Cardosos and Souloumiac (1993) presented an example of another type of ICA algorithm named JADE.

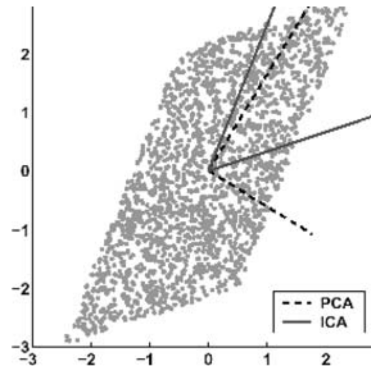


Figure 8 Comparison between ICA and PCA.

2.3 Comparison with principal component analysis

In the previous Sections, we defined the ICA problem and explained the way to find statistically independent projections the signals. In this section, we shall compare ICA with PCA.

PCA is a method used to find an orthogonal frame of reference directions with axes determined by the second order statistics of the original data vectors. There are two major differences with ICA. On the one hand, the frame of reference onto which one projects (represents) the multivariate data with ICA is no more inevitably orthogonal. On the other hand, the direction of the axes in ICA are not only computed from the second order statistics like in PCA but also from higher orders statistics. The advantage of ICA is quite visible from the example in Figure 8 where the data structure is more naturally explained using non orthogonal axes. The data consist of 2D uniformly distributed samples with independent components. Both methods find a new set of basis vectors for the data. PCA maximizes the variance of the projected data along orthogonal directions. ICA correctly finds the two vectors onto which the projections are independent.

Another difference is the ordering of the components. In PCA, the first principal component accounts for as much of the variability in the data as possible, and each successive orthogonal component accounts for as much of the residual variability as possible. It can be used to decrease the dimension of the problem by considering only the first components which explain most of the variance in the data. With ICA, we must first choose the number of sources to compute. By changing the dimensions of the unmixing matrix to estimate, we have the possibility to choose the number of independent sources computed by the algorithm. The obtained independent sources depend on the postulated number of sources. This is not the case in PCA.

3 Application to Dimensionality Reduction of Electroencephalograms

In this Section, we shall use PCA and ICA to reduce dimensionality. In Section 3.1, we shall present the method to reduce dimension of a problem using PCA or ICA. Next, we shall discuss about the choice of the dimensionality of the reduced space (Section 3.2).

We shall use the JADE algorithm for these 2 applications. Note that the performance of an ICA algorithm depends essentially on the distribution of the original sources. In the case of the EEG, we don’t have strong prior information about the original sources. So, the choice of JADE is arbitrary and principally motivated by its frequent use in the EEG analysis.

In the following, we shall use a Matlab package[1] to perform the JADE algorithm. This package was designed to facilitate EEG analysis and to visualize the estimated sources graphically. Implementations of other algorithms are also proposed.

[1] http://www.tsi.enst.fr/icacentral/algos.html

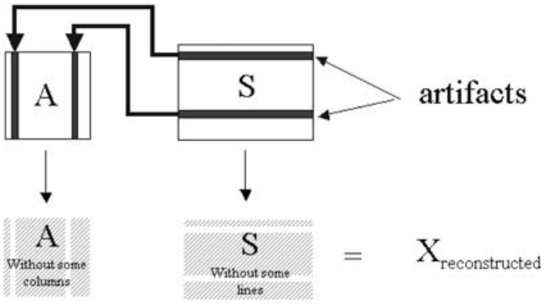


Figure 9 Reconstruction without some independent components.

3.1 Dimensionality reduction

Suppose that s is the matrix of principal components (from PCA) or the matrix of independent sources (from ICA).

The principal components or independent sources are assumed to be linear combinations of the original signals. In the same way, the original signals are linear combinations of the principal components or the independent sources obtained by respectively PCA or ICA. We can rebuild the signals using $x(t) = W^{-1}s(t) = \hat{A}s(t)$ for each $t \in \mathcal{T}$, which in matrix notation corresponds to: $X = W^{-1}S = \hat{A}S$.

We can decide to reconstruct the signals after discarding some principal components or independent sources. In this case, one only has to multiply the selected lines of the matrix S by the corresponding selection of column in matrix $\hat{A}$.

For example, suppose that we have 3 signals $\boldsymbol{x}_1^T$, $\boldsymbol{x}_2^T$ and $\boldsymbol{x}_3^T \in \mathbb{R}_N^1$. We consider the matrix X of observations : $X = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \boldsymbol{x}_3^T)^T$.

By PCA or ICA, we can compute 3 principal components or independent components $\boldsymbol{s}_1^T$, $\boldsymbol{s}_2^T$ and $\boldsymbol{s}_3^T$ yielding the matrix S: $S = (\boldsymbol{s}_1^T, \boldsymbol{s}_2^T, \boldsymbol{s}_3^T)^T$.

If we want to rebuild the original signals X, one only has to use : $X = W^{-1}S = \hat{A}S$.

If we want to reconstruct the signals without the second principal component or independent source $s_2{}^T$, we multiply the matrix s without the second line by the inverse matrix $\hat{A} = W^{-1}$ without the second column. We obtain 3 signals:

$$\hat{X} = \begin{pmatrix} \hat{\boldsymbol{x}}_1^T \\ \hat{\boldsymbol{x}}_2^T \\ \hat{\boldsymbol{x}}_3^T \end{pmatrix} = \begin{pmatrix} \hat{a}_{11} & \hat{a}_{13} \\ \hat{a}_{21} & \hat{a}_{23} \\ \hat{a}_{31} & \hat{a}_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{s}_1^T \\ \boldsymbol{s}_3^T \end{pmatrix}. \tag{9}$$

Figure 9 summarizes the procedure of dimensionality reduction.

Of note, the observed mixture will depend on the number of unobserved sources. If we separate the same observed mixture into 2 or 3 independent sources, the unmixing matrices $A_{(1,3)}$ or $A_{(1,2,3)}$ will be different: $A_{(1,3)}$ is not simply the matrix $A_{(1,2,3)}$ without the second row and second column.

3.2 Choice of the dimensionality of the reduced space

We showed how to reduce dimensionality. The question is now to choose the dimension of the reduced space. It's rather easy to determine the number of principal component needed in PCA. A measure of how well the first n principal components ($n \le p$) explain variation is given by the relative proportion:

$$\phi_n = 1 - \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{10}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of the covariance matrix of the centered data: X^TX.

For ICA, the problem is more difficult because the variance of the independent sources is not related to the eigenvalues of the covariance matrix of the centered data. Generally, we compute as many sources as possible. However, ICA allows to choose the number of sources calculated. When ICA is used to compute as many sources as electrodes (We use data kindly made available by the pharmaceutical company Eli Lilly (Lilly Clinical Operations S.A., Louvain-la-Neuve, Belgium). We have 720 available EEG. For each recording, 28 EEG leads are recorded using an ear linked reference. So, we can compute up to 28 independent sources.), we reconstruct the observations exactly. When the ICA algorithm is used to compute less components than electrodes, an error is made in the reconstruction of the signal.

We wondered how this error changes with the requested number of sources calculated by the ICA algorithm. For each of the 720 available EEGs, we have computed n independent components, for $n = 1, \ldots, 28$ (where 28 is the number of electrodes). The squared differences between the original and reconstructed data were computed. We noticed in each case that this error decreases rapidly between 1 and 6 or 7 components (an example is given in Figure 10a).

We define the percentage of reconstruction performed using the sum of the squared differences: for n computed sources ($n = 1, \ldots, 28$), we define d_n as

$$d_n = \sum_{\text{electrodes}} \sum_{\text{time}} (x - x_{\text{reconstructed with } n \text{ sources}})^2 , \tag{11}$$

and consider

$$r_n = \left(\frac{1 - d_n}{\max\limits_{1 \le n \le 28} (d_n)} \right) = 100 \left(\frac{1 - d_n}{d_0} \right) , \tag{12}$$

where d_0 is equal to $s^2 = \sum\limits_{\text{electrodes}} \sum\limits_{\text{time}} (x)^2$ the estimator of the variance. The sum over time mean the sum for each of the 128 time points: $t \in [t_0, t_1, \ldots, t_{127}] = \mathcal{T}$. So, r_n equals 100% when we reconstruct with as many components as electrodes and r_n is equal to 0% when we reconstruct with only one component (i.e. when the reconstruction error is maximum). When we plot the percentage of reconstruction r_n versus the number of sources used in that reconstruction, for each of the 720 avail-

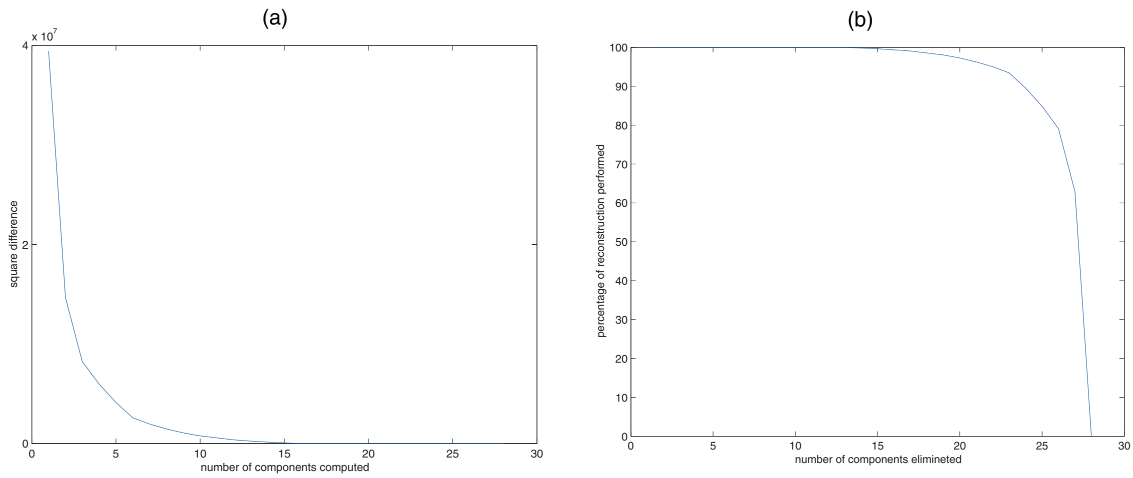


Figure 10 (a) Squared difference between the original and the reconstructed data (for 1 EEG): With more than 7 components computed, the square difference becomes close to zero. (b) Percentage of reconstruction performed versus the number of sources eliminated: With more 20 components eliminated, the percentage of reconstruction decreases dramatically.

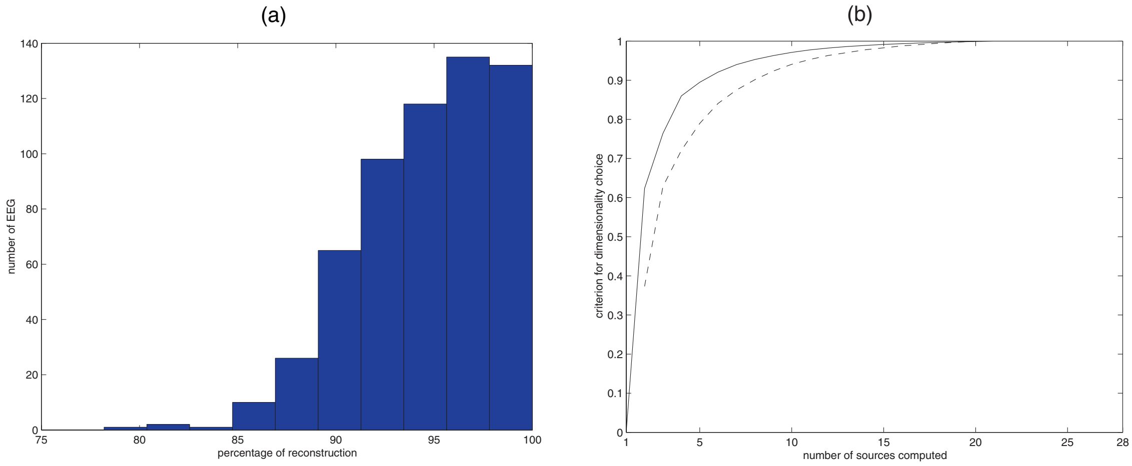


Figure 11 (a) Histogram of the 720 percentages of reconstruction performed with 8 sources: 8 sources reconstruct always more than 80% of the data. (b) The criterion for the choice of the principal components based on the eigenvalues: ϕ_n for $n = 1, \ldots, 28$ (solid line) and the criterion based on the percentage of reconstruction: r_n for $n = 1, \ldots, 28$ (dashed line). According to the value of ϕ_n we would choose only 3 principal components and 6 independent sources using the value of the percentage of reconstruction r_n.

able EEGs, we see that there is always more than 80% of reconstruction performed with 8 independents components. An example can be found in Figure 10b. The histogram of the percentages of reconstruction performed with $n = 8$ sources for the 720 EEGs is displayed in Figure 11a.

Examination of similar histograms for different numbers of sources suggests that 8 independent sources are sufficient to recover most of the EEG signals. In other settings, the required number of components might be different. In the analysis, we shall focus on this number of components to detect and quantify treatment effect on the brain. Notice that we select the same number of principal or independent components if we apply the r_n criterion for the choice of principal components (PCA) and for the choice of the independent sources (JADE using PCA as preprocessing step). Therefore, ICA does not lose information compared to PCA. Moreover, PCA decompositions are computed faster than ICA decompositions. Therefore, it would be interesting to use the r_n criterion applied to PCA decomposition to choose the dimension of the reduced space and to compute the ICA decomposition for the selected number of sources. Finally, as you can see in Figure 11b, the number of principal components needed to obtain more than 80% of reconstruction (r_n criterion) is larger than the number needed to explain more than 80% of variance of the data (ϕ_n criterion). The reason for this is that ϕ_n is related to variance while r_n is related to the Mean Squared Error (MSE) which is equal to the sum of the variance and the squared bias. Therefore, the criterion r_n is more strict because r_n control both variance and bias and, so, need more components. In the example of Figure 11b, we would choose only 3 principal components according to achieve $\phi_n \geq 80\%$ and 6 principal components (or independent sources) using the value of the percentage of reconstruction $r_n \geq 80\%$.

In this Section, we showed an application of ICA and of PCA to dimensionality reduction. We defined a criterion which can be used for both ICA and PCA reduction of dimensionality. In the next Section, we shall present another application of ICA and PCA to EEG analysis.

4 Application to Event-Related Potential Modelisation

In this Section, we shall introduce an interesting characteristic of EEG signals named event related potential (ERP) and in particular the P300 peak. We shall explain that Independent Component Analysis provides an analysis of these peaks allowing to summarize information from several electrodes. Moreover, we shall show that ICA provides a better analysis than PCA because, with ICA, 2 physiological phenomena are clearly extracted and separated from the observed signals.

4.1 Analysis of ERP using PCA vs ICA

In this Section, we shall first analyse ERPs using PCA. Next, we shall compare these results with the analysis of ERPs using ICA.

We shall now look for ERPs features in the principal components. We averaged the principal components over the 400 ms intervals after the infrequent stimuli in the same way as when dealing with the electrodes raw signals to detect P300 curves. That means that we averaged signals for several ERP episodes after stimuli. We saw that the first of the averaged principal components is highly correlated with the P300 curve. The second principal component does not present the same pattern. For example, you can see in figure 12a a P300 curve averaged for electrode Cz and in Figure 12b the average of the first principal component. The two curves are very similar.

We shall now examine ERPs from 8 ICA sources computed using the JADE algorithm with exactly eight sources requested.

We averaged each of the 8 independent sources over the 400 ms intervals after the infrequent stimuli in the same way as when dealing with the electrodes raw signals to detect P300 curves. We saw that 2 of these 8 averaged sources are highly correlated with the P300 curve. For example, you can see in Figure 13a a P300 curve averaged for electrode Cz and in Figure 13b the average of these 2 particular sources. You can see the same characteristics for the 2 curves (3 peaks followed by valleys at almost the same moment). When comparing these 2 sources, we see that their peaks do not occur simultaneously (see figure 13b) with the peak around 300 ms occurring earlier for one of the source (solid line). The other 6 sources do not present the same pattern.

When we averaged 28 independent sources over the 400 ms intervals after the infrequent stimuli in the same way as we dealing with the electrodes raw signals to detect P300 curves, we saw more than 2 of these 28 averaged sources are lightly correlated with the P300 curve. In fact, in that case, the 2 sources founded from the 8 independent sources are split into several signals which are less correlated with the P300 curve. Therefore, the dimensionality reduction is very interesting to summarize information in addition to reduce the number of signal to analyse.

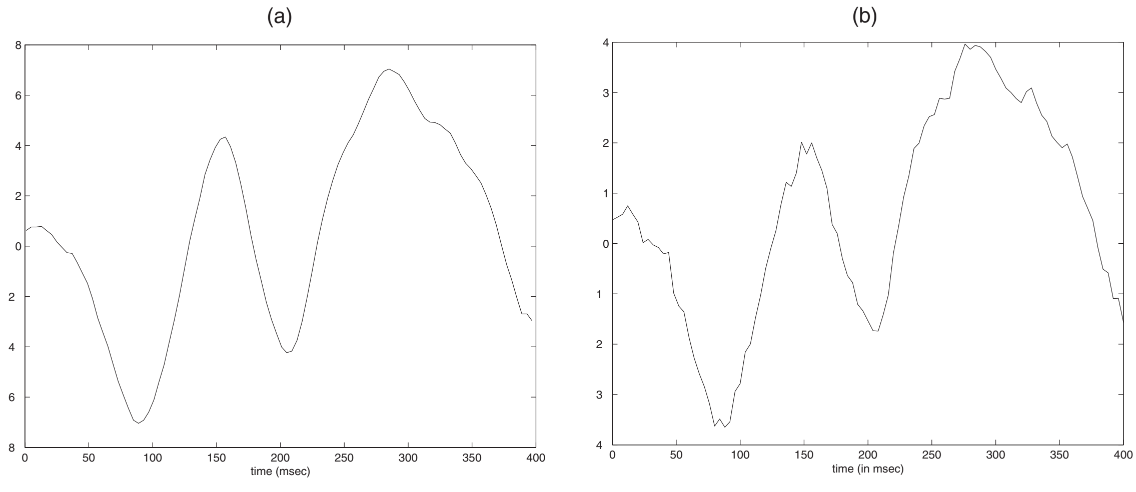


Figure 12 (a) Averaging of electrode Cz (central area). (b) Averaging of the first principal component: this curve is highly correlated with the P300 curve.

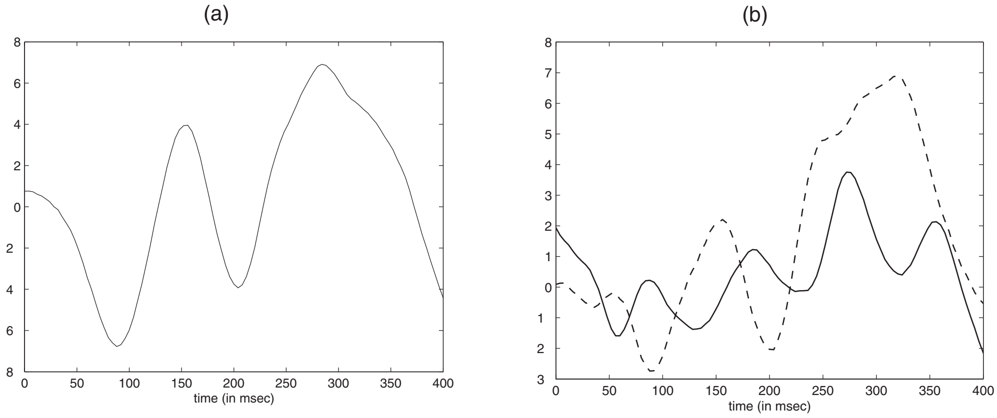


Figure 13 (a) Averaging of electrode Cz (central area). (b) Averaging of the independent components selected because there are highly correlated with the P300 curve.

It is interesting to visualise the contribution of each of the selected sources in the reconstruction of the signals at the 28 electrodes. This can be done by examining the inverse of the unmixing matrix W ($W^{-1} = \hat{A}$). The observations $\boldsymbol{x}(t)$ are a linear combination of the independent sources $\boldsymbol{s}(t)$. Therefore, the coefficient a_{ij} of the mixing matrix A gives the contribution of source j for the reconstruction of the signal at electrode i. One can represent these contributions for each source by a topographic map of each source as a 2-D circular view. A colour is associated to each level of the coefficients of A, with co-interpolation on a fine Cartesian grid used to select the colour between the electrodes. To visualize the topographic maps, we use a GUI-based Matlab toolbox for independent component analysis called EEGLAB and created by Makeig et al.

From the visual inspection of the topographic maps (see Figure 14) for these 2 sources, we see that one is located in the parietal area and the other in the frontal area. Therefore, these 2 sources are computed from the signals respectively from the frontal and parietal electrodes. The first principal component found by PCA (which is high correlated with P300 curve) has a fronto-central activation. The first principal component is probably a mixture of the 2 independent sources found by ICA.

An analysis based on the first principal component or on the 2 selected independent sources is much more attractive than the classical approach analysing the P300 amplitude and latency for all (or

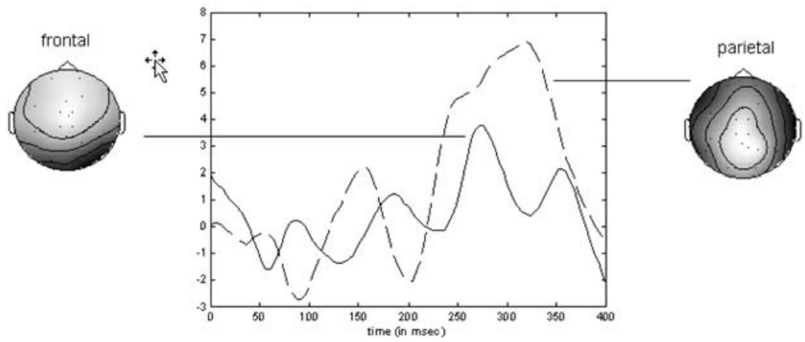


Figure 14 The P300 peak occurs earlier for the frontal source (solid line, left topographic map) than for the parietal source (dashed line, right topographic map).

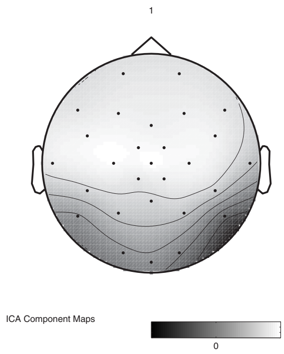
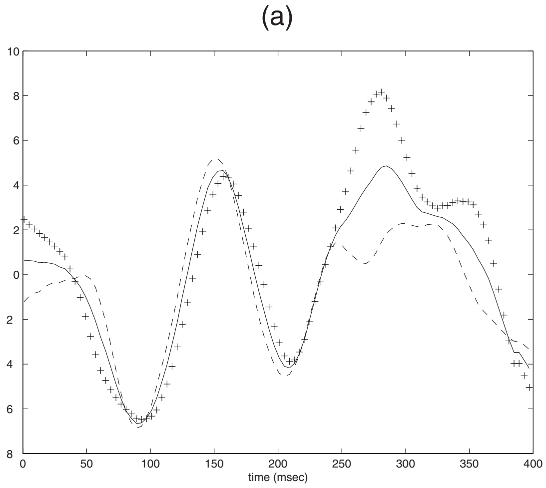


Figure 15 Topographic map of the first principal component.

some) electrodes simultaneously. Indeed, the independent sources turn to summarize the information from a whole area and not only from a particular point (corresponding to only one electrode, Fz or Pz for example).

We shall now show the main influence of these 2 sources. Our 2 sources summarize the information from respectively the frontal and parietal electrodes. In Figure 16a and 16b, we represent the P300 curves on the electrodes Fz and Pz classically used to analyse the P3a and P3b phenomena (continuous lines). On the same graphs, we represent also the same curves plus and minus the frontal (Figure 16a) and parietal (figure 16b) source respectively, multiplied by the coefficient of these sources in the reconstruction of the signals from these electrodes (electrode Fz in Figure 16a and Pz in Figure 16b). So, we can see the main influence of these 2 sources. It is evident in Figure 16a and 16b that the 2 sources have the most influence in the P3a and P3b peak area. A major advantage of our approach is that it summarizes the P3a and P3b phenomena using only two sources that can be used to reconstruct most of the P300 curves on the frontal and parietal areas. Moreover, the 2 phenomena are now clearly extracted and separated from the observed signals.



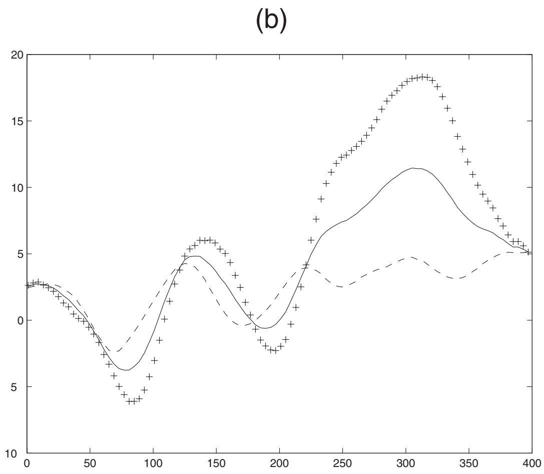


Figure 16 (a) Signal from electrode Fz + and − a coefficient × the "frontal" source: the larger differences in the P3a peak area. (b) Signal from electrode Pz + and − a coefficient × the "parietal" source: the larger differences in the P3b peak area.

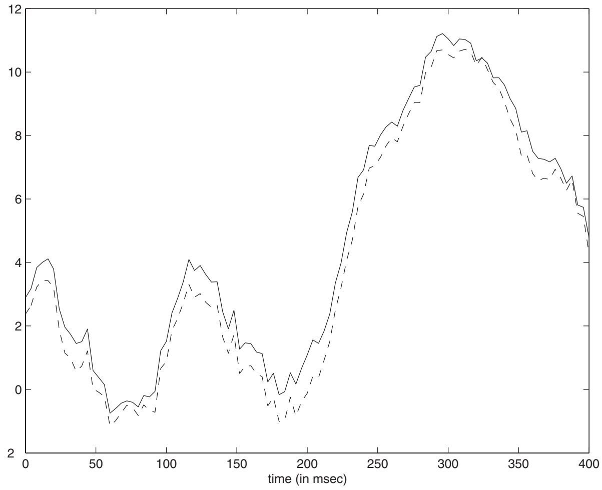


Figure 17 The most important difference between ERP curve computed on data (solid line) and on data reconstructed from 8 independent sources (dashed line). This difference is still very small.

Finally, we can wonder if the dimensionality reduction reduces information on the ERP. We can measure this loss by comparing the ERP curves calculated from the original data to the ERP curves computed on data reconstructed from the 8 independent sources. For the previous example, the most important differences are presented in Figure 17. You can see that differences are very small.

In this Section, we showed that an analysis based on the results of PCA or ICA is much more attractive than the classical approach analysing the P300 amplitude and latency for all (or some) electrodes simultaneously. Moreover, ICA provides a more powerful data representation than PCA in the ERP analysis context. Indeed, using PCA, we saw that the first of the averaged principal components is highly correlated with the P300 curve. Using ICA, we found 2 independent sources to explain the P300 curves. These 2 sources summarize the P3a and P3b phenomena. These 2 phenomena are clearly extracted and separated from the observed signals. From the inspection of the topographic maps, we can conclude that the first principal component can be seen as a mixture of the 2 independent sources found by ICA.

5 Conclusion

We have investigated the usefulness of ICA to reduce problem's dimensionality. We proposed a efficient criterion to choose the dimension of the reduced space. Eight independent sources are probably sufficient to recover most of the EEG signals for the illustrating study. We can thus analyse these 8 sources instead of the correlated signals from 28 electrodes. We found that only 2 of these 8 sources are required to analyse the P300 curve. So, we restrict the analysis to 2 independent signals instead of 28 (one for each electrode). These 2 components summarize the information in the parietal and frontal areas and enable the analysis of 2 phenomena (P3a and P3b) mixed in the original signals. With PCA, we found only one independent component needed to analyse the P300 peak. However, that analysis does not separate the P3a and P3b phenomena. The entire procedure is automatic and easily set up. ICA is very promising to analyse multidimensional biomedical signals and much more efficient than PCA in the EEG analysis context.

Acknowledgements We thank Eli Lilly for making the data set available to us. Moreover, Céline Bugli thanks Eli Lilly for financial support through a Mecenat research grant. The Université catholique de Louvain, Louvain-la-Neuve, Belgium is gratefully acknowledged for financial support through a FSR research grant and the 'Fonds National pour la Recherche Scientifique' (FNRS), Belgium is gratefully acknowledged for financial support through a FRIA research grant. Philippe Lambert thanks the IAP network nr P5/24 of the Belgian state (Federal Office for Scientific, Technical and Cultural Affairs).

References

Cardoso, J.-F. and Souloumiac, A. (1993). Blind Beamforming for non Gaussian Signals. *IEEE-Proceedings-F* **140(6)**, 362–370.

Comon, P. (1994). Independent Component Analysis, a New Concept? *Signal Processing* **36**, 287–314.

Scott Makeig's home page: http://www.sccn.ucsd.edu/ scott/ (accessed in May 2006).

Baudena, P., Halgren, E., Heit, G., and Clarke, J. (1995). Intracerebral potentials to rare target and distractor auditory and visual stimuli. III. Frontal cortex. *Electroencephalography and Clinical Neurophysiology* **94**, 251–264.

Halgren, E., Baudena, P., Clarke, J., Heit, G., Liégeois, C., Chauvel, P. et al. (1995). Intracerebral potentials to rare target and distractor auditoryand visual stimuli. I. Superior temporal plane and parietal lobe. *Electroencephalography and Clinical Neurophysiology* **94**, 191–220.

Halgren, E., Baudena, P., Clarke, J., Heit, G., Marinkovic, K., Devaux, B. et al. (1995). Intracerebral potentials to rare target and distractor auditoryand visual stimuli. II. Medial, lateral and posterior temporal lobe. *Electroencephalography and Clinical Neurophysiology* **94**, 229–250.

Smith, M., Halgren, E., Sokolik, M., Baudena, P., Musolino, A., Liégeois-Chauvel, C. et al. (1990). The intracranial topography of the P3 event-related potential elicited during auditory oddball. *Electroencephalography and Clinical Neurophysiology* **76**, 235–248.

Hastie, T. and Tibshirani, R. (1991). *Generalized Additive Models for Medical Research*, Chapman and Hall, London.

Herault, J. and Jutten, J. (1991). Blind Separation of Sources, Part I: An Adaptative Algorithm Based on Neuromimetic architecture. *Signal Processing* **24**, 1–10.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*, Wiley, New-York.

Lee, T.-W., Girolami, M., Bell, A.J., and Sejnowski, T. J. (2000). A Unifying Information-Theoric Framework for Independent Component Analysis. *Computers and Mathematics with Applications* **39**, 1–21.

Makeig, S., Bell, A. J., Jung, T. P., and Sejnowski, T. J. (1996). Independent Component Analysis of Electroencephalographic Data. *Advances in Neural Information Processing Systems*, 145–151.

Polich, J. and Kok, A. (1995). Cognitive and Biological Determinants of P300: an Integrative Review. *Biological Psychology* **41**, 103–146.

