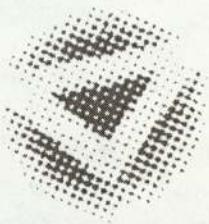
Dynamical Embedding and Feature Extraction of Electroencephalographic Data

Nathalie Christiane Noël

MSc in Pattern Analysis and Neural Networks

Supervisor: Professor David Lowe



Aston University

September 1998

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor David Lowe, for his excellent guidance, stimulating ideas and his ability to always find time for me when there appeared to be none.

I would also like to record my appreciation to Doctor Helen Stone of Sowerby Research Center who aided my understanding of electroencephalograms.

Finally, I am grateful to my parents and all my friends on Aston Campus for their support and assistance.

Chapter 1

Introduction

> ... I shall demonstrate how this tiny sound within, this nothing, contains everything; and how, with the bacillary aid of a single sensation — always the same one, and deformed at that in its very origins — a brain isolated from the world can create a world in itself ...
>
> **Remy de Gourmont** — *Sixtine.*

Measurements of brain activity can be performed by recording the electric potentials on the scalp surface : this is known as *electroencephalography* (**EEG**). EEG analysis has played a key role in the modeling of the brain's cortical dynamics. If several mental states can be reliably distinguished by recognizing patterns in EEG, then is it possible to utilise EEG information to automatically estimate *car driver* or *pilot workload* for example ? By estimating workload, we mean assessing the level of attentiveness or vigilance of a pilot.

There have been many studies of alertness to try to discover whether vigilance may be recognized from single or multi-channel EEG traces ([35], [36], [39]), since the first investigation by Loomis *et al* in 1937 [7]. Most of them have confirmed that, despite sincere intentions, few subjects remain vigilant while engaged in monotonous monitoring tasks.

The analysis of the data is problematic due to the fact that multiple neural generators of the EEG may be simultaneously active and the potentials and electromagnetic

are distorted by the head volume conductor), reference electrode effects and algorithm effects (algorithms that are adopted to reduce volume conduction effects may introduce false coherency estimates) [32].

The fundamental reason why EEG analysis is performed in the frequency domain is because of the belief in the linear nature of the physical sources generating the potential differences measured by the sensors. This linearity suggests that the signal might be decomposed into a sum of sinusoidal components. So we are supposed to obtain a description of the signal in terms of its fundamental frequency characteristic.

Recently, blind source separation by *Independent Component Analysis* (**ICA**) has received attention because of its potential applications in signal processing such as in speech recognition systems, telecommunications and medical processing. The goal of ICA is to recover independent sources given sensor outputs in which the sources have been linearly mixed. In contrast to correlation based solutions such as Principal Component Analysis (**PCA**), ICA not only decorrelates the signals but also reduces higher-order statistical dependencies, attempting to make the signals as independent as possible. The blind source separation problem has been studied by researchers in the field of neural networks [33], [1], [2], [26], [19], [8]. It has also been applied to the particular field of EEG in several studies [35], [36], [39], [25]. All these studies only consider multi-channel EEG recordings to perform the ICA algorithm.

For this study, we have made the choice to consider only single channel EEG data recorded from wake subjects due to the hypothesis that over short segments of EEG data, we can reconstruct the dynamics of the system with a dynamical embedding of one single channel. We will question the use of linear Fourier analysis within the wake state, essentially because of the nature of the noise sources and complexity of the signal. Indeed, we will start from the hypothesis that the complexity in wake EEG is due to the nonlinear interaction of a few degrees of freedom rather than the linear interaction of many degrees of freedom, plus additive noise. This is a dynamical systems perspective which considers the existence of an *underlying data generator* (or *attrac-*

consists of finding the statistically independent sources responsible for a set of data. The most familiar situation is the "cocktail party problem" where there are many speakers, or sources of accoustic signals, and the listener detects mixtures of these signals.

Finally, *topographic mappings* may be viewed as nonlinear, unsupervised feature extraction processes. Here the criterion for selection of features is not to maximise variance or any mutual information, but rather that the *topology* or geometric structure of the data be preserved in the feature space.

In the feature space, delay vectors x are expressed as a linear combination of spanning basis functions v_i and a set of "source" signals $\alpha_i(t)$

$$x(t) = \sum \alpha_i(t) v_i$$

In a Principal Component Analysis embedding, the basis functions v_i are obtained as the eigenvectors of a covariance matrix (see Chapter 3). In an Independent Component Analysis, the expansion basis vectors v_i are instead determined as the independent components of a demixing matrix (see Chapter 4).

In both cases (PCA and ICA), we project the data linearly on the embedding vectors and therefore reduce the dimensionality of our feature space. This also allows us to reduce some of the noise structure (see figure 1.1).

We can now build a model in this reduced (dimension and noise) feature space (see Chapter 6), search for some dynamic structure, identify anomalous behaviour and look for interesting structure which might characterise vigilance.

Figure 1.1 presents a symbolic overview of the structure and function of this thesis.

Chapter 5 details all the results with ICA applied to our raw or filtered data. It discusses these results in terms of frequencies and validates the hypothesis made in **Chapter 3**.

Chapter 6 applies different methods of topographical mappings to the results of ICA. These methods are : Kohonen Self-Organizing Maps, Sammon Mapping, NEUROSCALE The different results are then discussed and the methods are compared.

Chapter 7 concludes this thesis, discusses the major results and gives directions for further research.

sources rather than sources farther from the sensors.

- EEG is most sensitive to correlated dipole layer in ab, de, gh (that is perpendicular fields), less sensitive to correlated dipole layer in hi (that is tangential fields) and insensitive to opposing dipole layer in bcd, efg.

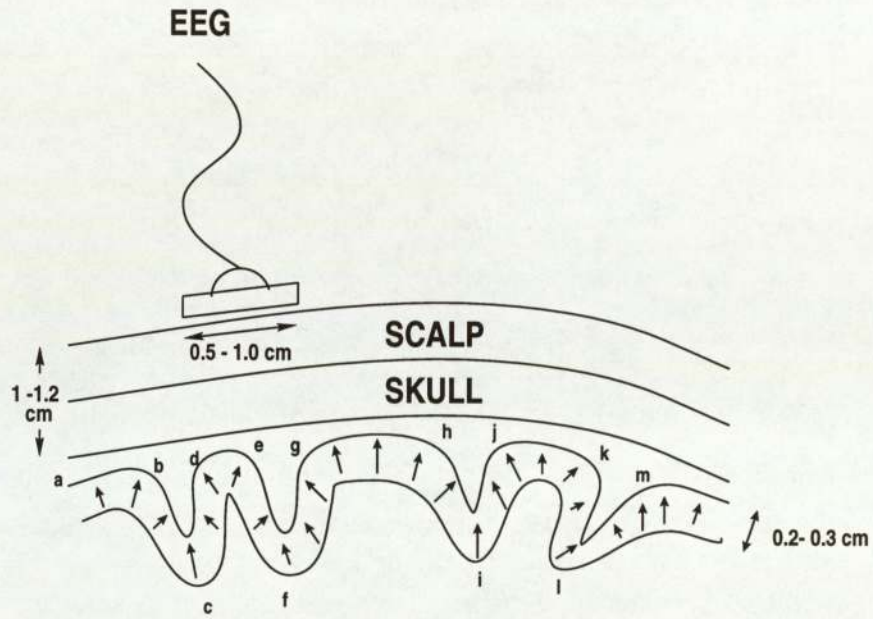


Figure 2.1: Extracranial recording of EEG data

With all the limitations of extracranial recordings, we can wonder if anything valuable can be extracted from EEG signals. However, it has long been appreciated (since the first EEG recording in 1928) that electroencephalography is a genuine measure of conscious experience. As a matter of fact, EEG recording has long been used in medicine as a clinical test for variety of pathologic conditions : epilepsy, Alzheimer's disease, severe head injuries, multiple tumors...

2.4 Brain Waves

Interpreting EEG involves the characterization of wave forms largely defined by their frequency and to a lesser extent by their morphology. The difficulty lies, in part, in recognizing artifacts and also in being able to differentiate normal variants from abnormalities.

2.4.1 Waves defined by frequency

Frequency means the number of waves per second. The frequencies of the EEG waves run from 0.5 per second to hundreds per second. Waves are usually defined by their frequency and are divided, on this basis, into four main groups. Figure 2.2 displays these different types of waves and the following sections describe them more precisely.

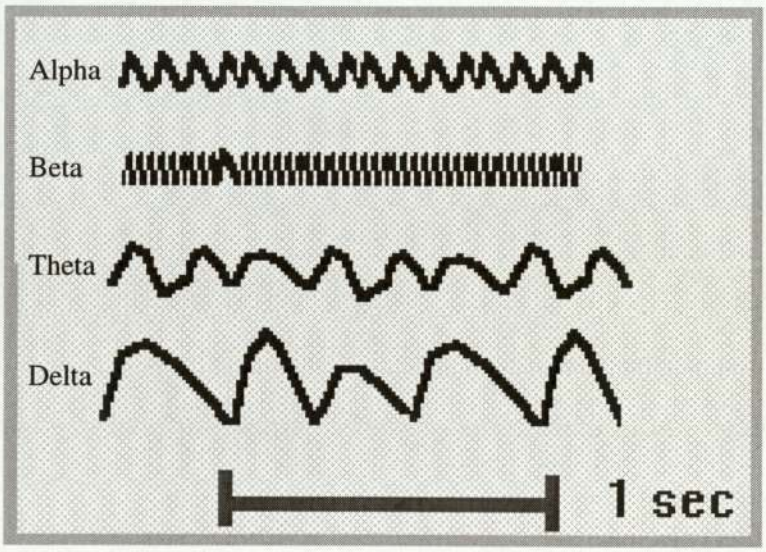


Figure 2.2: Brain waves

in children.

2.4.2 Waves defined by morphology

Certain waves have characteristic forms irrespective of their frequency and are recognizable by their shape; in other instances pair or groups of waves have typical appearances. Single waves that are specially shaped include, for instance spikes or sharp waves - waves that rise rapidly to a point and fall away equally dramatically with a base that is small compared to the wave's amplitude. Some wave forms can be recognized by their morphology and these include two main types:

- Specially shaped waves
- Specially shaped wave complexes

Artifacts

Artifacts are disturbances caused by technical defects - usually transitory. Included are such things as eye movement, electrode movement with loss of contact, muscle activity obscuring the EEG, movements of the head, scratching the scalp, sweating etc...

Normal variants

There are several waves or patterns of waves which are unusual in appearance yet are not significant for abnormality or disease. These waves can be misinterpreted. Amongst the more common ones are mu rhythm, psychomotor variant, lambda waves, POSTS, spindles, vertex waves.

EEG data was collected on an Oxford Instruments Medilog system utilising 8 separate measurement channels at the University of West England on behalf of British Aerospace. The electrodes were sited according to a standard 10-20 system and bipolar potential differences due to electroencephalographic activity were measured.

The data was sampled at 128 Hz and scaled and linearly quantised at one byte per sample. For this study, only signals extracted from scalp locations T5-Oz were used. Cross-channel effects are the subject of future work.

One major problem with using wake EEG is due to the very poor signal-to-noise ratio as a consequence of the overall mental and mechanical activity, giving rise to spurious and background electrical activity. Hence our first approach was to consider approaches to reduce the noise components in the signal and to enhance the information-carrying signal components.

3.2 Nonlinear Phenomena

The separation of linear systems into discrete and continuous systems is also appropriate for nonlinear systems. Our aim is to exhibit a new appreciation for the complexity and richness of behaviour of a system such as brain waves with only a few degrees of freedom.

There are different types of behaviour for a deterministic time series :

- equilibrium
- periodic
- quasiperiodic
- chaotic

These four types of behaviour are called *attractors*. The distinction between these different types of attractors are most obvious when applied to systems with relatively few degrees of freedom and large signal-to-noise ratio. The problem we are going to face is that our electroencephalographic data has a very poor SNR. The generator of our data will thus be quite difficult to characterise.

The general idea is that simple systems (that is systems with a few degrees of freedom) behave simply. This idea is certain for linear systems. However, with nonlinear systems, new phenomena have been described that cannot be predicted by linear theory. So even very simple systems, if nonlinear, can exhibit extremely complex behaviour. Such nonlinear systems are very interesting for appreciating their complexity and richness of behaviour despite their few degrees of freedom. They can exhibit *chaotic and dynamic behaviour* characterized by complexity and sensitivity to the initial state of the system [32].

a particular physiological state of the brain) typically is estimated by expressing the data from our single EEG channel into an M-dimensional space. This is equivalent to constructing a delay vector $X(t)$ from the initial time series

$$X(t) = (X(t-\tau), \ldots, X(t-(M-1)\tau)) \in \mathbf{R}^n$$

where τ is the time increment (or lag) and M is the embedding dimension (or number of lags). According to [15], we choose $\tau = 1$ in the whole following.

A single point in this space is then located by the vector

$$Y_1(t) = (X_1, X_2, \ldots, X_M)$$

and the next location is given by the vector

$$Y_2(t) = (X_2, X_3, \ldots, X_{M+1}), \text{ and so on.}$$

Hence, in a time series consisting of K measured values, $Y(t)$ can assume $L = K-M+1$ discrete values in the embedding space.

For a D-dimensional attractor, the *embedding dimension* M must be at least as large as D. Takens showed in [16] that for a system of D degrees of freedom, we must have $M > 2D + 1$.

We then have to determine the complexity and the window size M of the embedding. In the following study, we will consider $K = 1000$ samples which is worth 10 seconds of EEG data.

3.4 Determining the Complexity

As described before, a delay-space embedding can be used to reconstruct a multi-dimensional representation. But an important question is *"How many dimensions should be used in the representation ?"*.

It was shown in [15] that an analysis of the number of degrees of freedom in X leads to the singular value problem

$$X = S\Sigma C^T \tag{3.2}$$

by the σ_i. Furthermore, as σ_i decreases, the noise-to-signal ratio (inverse of SNR) increases. Thus it is possible to obtain information on the level of noise in the system by studying the eigenspectrum.

So, to make a choice from among all the different singular spectra, we are looking for a change in the curvature that displays the limit between the signal space and the noise space. We expect a general stability of the spectrum as sufficient information content is captured with the window size. *Convergence* of the singular spectrum is the criterion for obtaining a sufficiently large delay window.

Thus, we know that we have found the right kink when the singular spectrum does not change in the signal space when we go on incrementing the window size. This means that the window size is big enough to capture the whole dynamics of the signal generator. However, because of serial correlations in the data, the length of the delay vector is *not equivalent* to the number of degrees of freedom of the data. It is rather determined by the location of the kink on the converged spectrum. The number of degrees of freedom is the dimension of the subspace containing the embedding manifold rather than the dimension of the manifold in itself.

For our given EEG data, the singular spectra seem to converge for a window size of 30. In figure 3.1 (a), we can observe two kinks occuring at the second eigenvalue and at the eigth eigenvalue. We know the first kink is typical of such eigenspectra. They relate to the trend of the time series and gather a very large amount of variance. Therefore, we know that, they are not enough to reconstruct the signal subspace. The kink we are looking for, which shows the singular spectrum of the delay vector for a window size of 30, occurs around 8 eigenvalues.

Figure 3.1 (b) gives us the same singular spectrum by using a delay window of 8. The delay embedding has therefore 8 degrees of freedom. We can see that the curve is smoother in the noise space after the kink at about 5. But there is a residual structure in the noise space after this kink. That implies an intrisic dimensionality of the underlying manifold generating the EEG signals of about 5.

3.5 Principal Component Analysis

Now that we have chosen to build an embedding matrix of size 30 with our input vector, it would be interesting to first perform a Principal Component Analysis on our data as a preliminary linear analysis.

PCA is a commonly used method for analysing data, and it is closely related to some other methods such as least squares methods and factor analysis. The objective of PCA is to find a set of m orthogonal vectors in data space that have the greatest contribution to the data variance. Dimensionality reduction is accomplished by projecting the original data with n-dimensional space onto the m-dimensional subspace spanned by the orthogonal vectors. This projection often retains most of the essential information in the data. PCA is also used to search for clusters. The first principal component is taken along the direction of maximum variance, while the second principal component has to be the subspace perpendicular to the first one and taken along the direction of maximum variance within the subspace. Then the third principal component is a subspace perpendicular to the first two with maximum variance direction, and so on.

The above steps can be generalised that the direction of the kth principal component is along an eigenvector direction of the kth largest eigenvalue of the full *covariance matrix*. Proof can be found in reference [17].

Unfortunately, this method has a number of problems. For example, extreme points in the data set (known as outliers) can generate large errors in the eigenvalues. The structure of the data cannot be recovered, *i.e.*, there is a loss of orientation due to aliasing along the largest variance of two parallel groups of data. Finally, the linearity aspect of PCA will obviously not solve non-linear problems. PCA may yield a relatively large number ($m > 10$) of significant principal components from which we cannot obtain information (ordinarily we can visualise 2-3 dimensional data space). Moreover, a large number of data can significantly increase computational complexity, *e.g.*, computing the inverse of the covariance matrix is typically $\Theta(n^3)$.

Chapter 4

Independent Component Analysis

4.1 Introduction

We have just seen the results of a Principal Component Analysis applied to our data. Recent studies of Independent Component analysis with multi-channel EEG have proven promising [35] [36] [39]. That is why we have chosen to apply it to our single channel EEG data. The fundamental advantage of ICA over PCA is that the signal space is not constrained to be spanned by orthogonal basis vectors and hence the independent sources obtained from Blind Source Separation should have a better *interpretability* in terms of the original EEG problem.

Let us assume that we have some phenomenon which manifests itself through a set of n independent random variables. We shall denote the combination of these variables with a random vector $s = [s_1 s_2 \dots s_n]^T$. Components $s_1, s_2, \dots s_n$ are called *sources* and s is called the *source vector*. This name implies independence : the sources are assumed to be independent sources of information.

Now suppose that the original independent source components are observed via a linear process. Denote the observed random vector by x. Since the process is assumed linear, the relation between s and x can be modelled as

$$x = As \tag{4.1}$$

because any constant multiplying an independent component may be cancelled by dividing the corresponding column of the mixing matrix A by the same constant.

- there is no ordering between the independent components.

4.2 Removing correlations

Assume that our data has zero mean, that is $E\{x\} = 0$. If we can find a linear transformation giving relation (4.2), the independent components of s have zero mean as well. We assume then that the data has this property, that it has been *centered* by removing its mean and that it has been unit-varianced. So the covariance matrix of s is $cov\{s\} = I$, and components of s are uncorrelated. Uncorrelatedness is necessary but not sufficient for independence.

We can accomplish uncorrelatedness by transforming x so that its covariance matrix will be diagonal. If in addition, all components have unit variance (the covariance matrix is unity), the process of accomplishing this is called *whitening* or *sphering.*

Whitening can be done using PCA basis vectors. Let E denote the matrix of principal component basis vectors of random data vector x, i.e., the eigenvectors of $cov\{x\}$, and $D = diag(\xi_1, \ldots, \xi_m)$ a diagonal matrix of corresponding eigenvalues. The new whitened data vector v is given by

$$v = D^{-1/2} E^T x \tag{4.3}$$

Matrix $V = D^{-1/2} E^T$ is a *whitening matrix.* The fact that v is really *white* can be seen from

$$\begin{aligned} cov\{v\} &= E\{D^{-1/2} E^T x x^T E D^{-1/2}\} \\ &= D^{-1/2} E^T cov\{x\} E D^{-1/2} \\ &= D^{-1/2} E^T E D E^T E D^{-1/2} \\ &= I \end{aligned}$$

- Karhunen, Oja, Wang, Vigario and Joutsensalo [19]
- Karhunen and Pajunen [21]
- Girolami and Fyfe [26]
- Pearlmutter and Parra [8]

In Table 4.1, we describe the algorithms we have selected to use before choosing the definite one for the experiments and the study.

	Method of solution	
Mathematical approach	Diagonalization	Fixed point
Fourth order cumulants	`JADE` [23]	Original fixed point [5]
Contrasts based on other nonlinearities	—	Generalized fixed point [4]

Table 4.1: A classification of used ICA algorithms.

One especially important class of algorithms missing from this list is the set of algorithms with foundations in information theory. In section 4.3.2, I will introduce the entropy maximization algorithm of Bell and Sejnowski. Then in section 4.3.3, I will present algorithms based on batch computations. Finally, in section 4.3.4, I will present the *fast-fixed point algorithm* which is a particular method of the generalized fixed-point, I have been using for my experiments on EEG data.

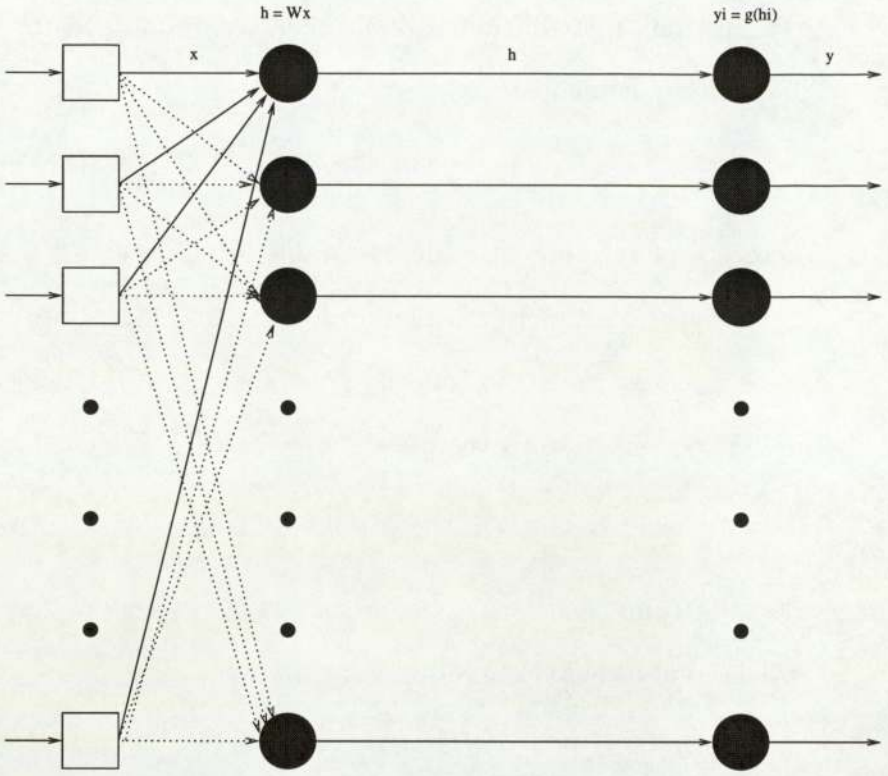


Figure 4.1: Entropy maximization error network

4.3.3 Jade Algorithm

The JADE algorithm of Cardoso and Souloumiac is based on joint approximate diagonilization of eigenmatrices [23]. The ICA problem can be solved by computing the eigenvectors of the cumulant matrix $Q_v(M)$ for any matrix M. JADE diagonalizes a set of eigenmatrices representing the whole cumulant matrix set C_v.

The problem with such an algorithm is that it uses batch tensorial computations. The data we are using is too large for such difficult tensorial computations.

4.3.4 Fast-Fixed Point Algorithm

One way to approach the ICA problem is to try to form an optimization problem that has its solutions as the independent components. We shall call such objective functions *contrast functions.*

We introduce a measure called *kurtosis.* Its value is described to measure the peakedness of the distribution, with peaked distributions giving positive values of kur-

There are two ways of solving this equation. One of them would be by applying standard numerical algorithms. Hyvärinen and Oja have chosen to write the equation in the form :

$$w = scalar \times (E\{x(w^T x)^3\} - 3\|w\|^2 w) \tag{4.12}$$

This is very useful for the scalar takes into account the penalty function which we therefore don't need to compute and hence, we don't need to take into acount the peakness of our signal (sub or super gaussian).

Then, the iteration obtained is very fast.

Using the preceding equation, we derive the following algorithm :

1. $w := rand()$
2. $w := w/\|w\|$
3. $w_{old} := 0$
4. **while** $\|w - w_{old}\| > \varepsilon \bigwedge \|w + w_{old}\| > \varepsilon$
   - $w_{old} := w$
   - $w = E\{v(w^T v)^3\} - 3w$
   - $w := w/\|w\|$

   end

Table 4.2: Hyvärinen and Oja's Fast-Fixed Point Algorithm

The final vector w equals one of the column of the mixing matrix B, which means that one of the non-gaussian independent component has been separated. Thus to estimate n independent components, one needs to run the algorithm n times. We are sure that we estimate each time a different component thanks to the orthogonalizing projection inside the loop.

Hyvärinen and Oja prove that their algorithm has a cubic convergence.

Chapter 5

Experiments and Results

5.1 Introduction

The Independent Component Analysis is ideally suited for performing source separation in domains where :

- the sources are independent
- the propagation delay of the "mixing medium" are negligible
- the sources are analog and have probability density functions not too unlike the gradient of the logistic sigmoid
- the number of independent signal source is the same as the number of sensors

In our case of EEG signal, one scalp electrode picked up correlated signals at different times of the day on a wake human being executing four different tasks. We would like to know what effectively *independent brain sources* generated these mixtures. If we assume that the complexity of EEG dynamics can be modelled as a collection of statistically independent brain processes, the EEG source analysis problem satisfies ICA assumption 1. Since volume conduction in brain tissue is effectively instantaneous, ICA assumption 2 is also verified. Assumption 3 is plausible. But assumption 4 is

5.3 Independent Component Analysis of Raw EEG Data

I have chosen to show here only the most relevant results on task 4 in order to remain as concise as possible.

Figure 5.1 shows the original time serie of 1000 samples we are going to work on.

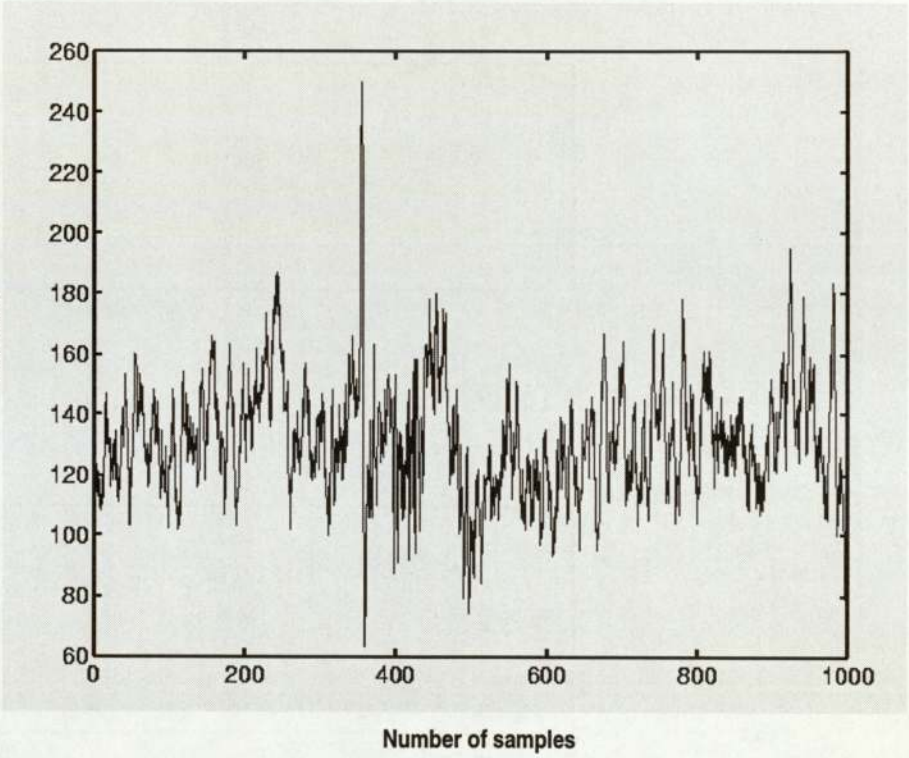


Figure 5.1: 1000 samples time series for task4

Then, we complete an embedding on our signals (let us bear in mind that we use 1000 samples which is approximately worth 8 seconds of EEG recording), we can apply the fast fixed point algorithm.

We chose to compute as many independent components as we have delay vectors. As a matter of fact, we do not know how many sources such a time series is constituted of. We do not want to risk losing any information. Hence, at the end of our computation, which is quite long, we obtain 30 independent components. We project each of our delay vectors on the corresponding independent component in order to obtain the sources we

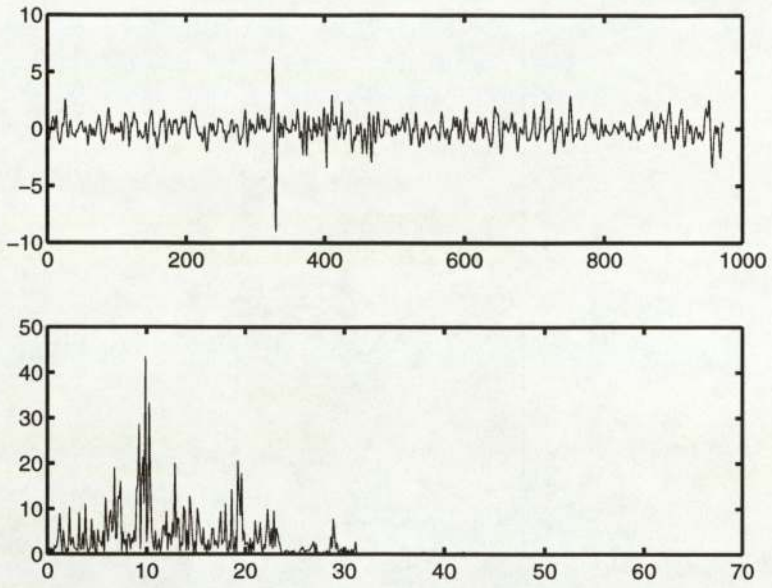


Figure 5.4: 9th source of the raw data and its power spectral density

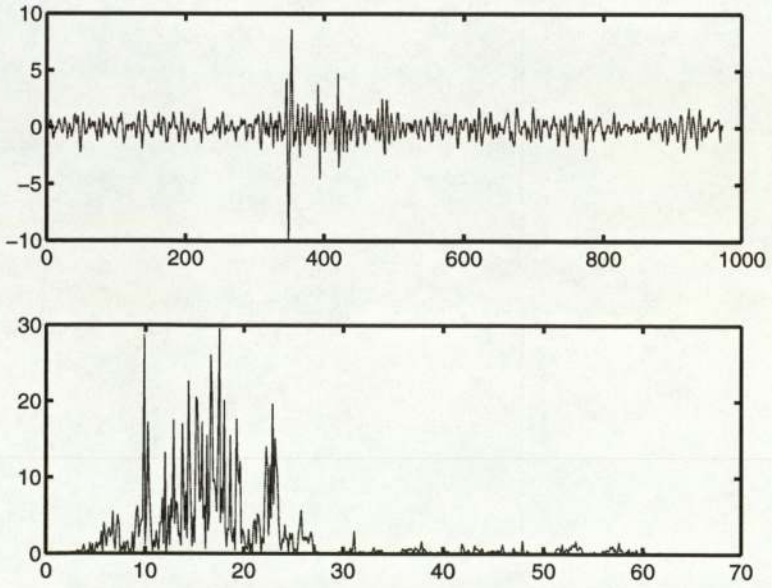


Figure 5.5: 8th source of the raw data and its power spectral density

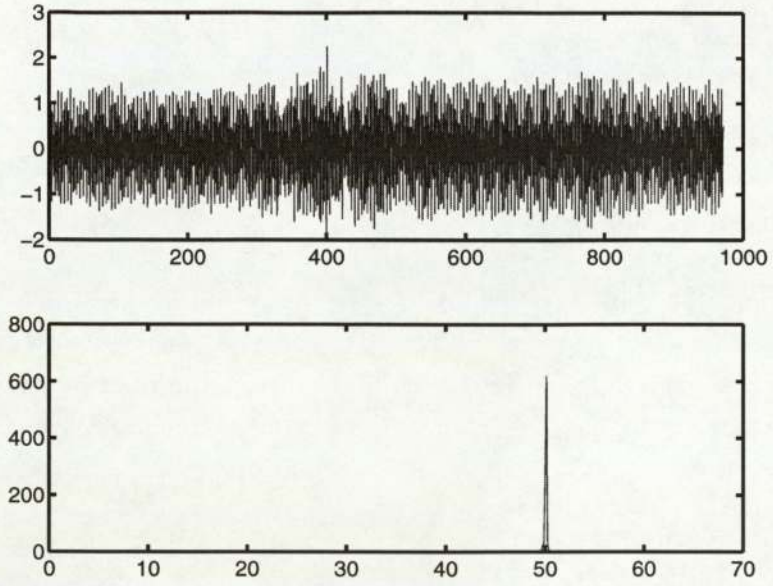


Figure 5.8: 10th source of the raw data and its power spectral density

As none of our resulting sources are ordered, we had to choose a way to classify them. Using the power spectrum of each of them proved to be a good method. As a matter of fact, it seems that the sources showing the brain activity during the different tasks, display different frequency range. Of course, some of them are identical or fall very closely in the same frequency interval. We can then advance that they naturally *cluster* by order of frequency.

Moreover, knowing the approximate frequency of each of the sources and their shape allows us to be able to relate them to an activity of the brain. Let us study each preceding source separately :

- source 16 (figure 5.2) has a very low frequency at about 5 Hz and its morphology can make us think that the ICA has isolated some eye movements.
- source 20 (figure 5.3) shows us the general trend of the signal but if we analyse it more closely we see a frequency range from 5 Hz to 10 Hz which characterizes alpha activity.
- source 9 (figure 5.4) shows high alpha activity and beta activity mixed.
- source 8's morphology (see figure 5.5) displays more spikes and the frequency

chosen to be reasonable (in combination with a reasonable necessary compromise on the shape of the rectangle). The Butterworth filter provides the maximum flatness in the passband (no ripples) which implies the minimum amplitude distortion and is therefore suited for our situation.

They are causal and of various orders, the lowest order being best (shortest) in the time domain, and the higher orders being better in the frequency domain. Well-engineered projects often include Butterworth filters.

Our need is to design a lowpass filter that loses no more than 3 dB in the passband and has at least 50 dB in the stopband because we assume that there is no more human activity after 50 Hz. 45 Hz and 50 Hz are the passband and stopband edge frequencies and the sampling frequency is 128 Hz. The estimated order of the filter is 7. The order was estimated using MATLAB function **butterord**. Figure 5.9 shows the magnitude of the transfer function of our designed filter. The filter was designed using MATLAB function **butter**.

Let $\frac{B(z)}{A(z)}$ denote the transfer function of the Nth-order digital filter. Then, by computing the z-transform of the digital filter, we have :

$$H(e^{jw}) = \frac{B(z)}{A(z)} = \frac{b(1) + b(2)z^{-1} + \cdots + b(n_b + 1)z^{-n_b}}{1 + a(2)z^{-1} + \cdots + a(n_a + 1)z^{-n_a}} \tag{5.1}$$

The vector w is a L-point frequency vector in radians, and H is the L-point complex frequency response vector of the filter $\frac{B}{A}$ given numerator and denominator coefficients in vectors B and A.

Figure 5.10 shows 1000 samples of EEG data after the application of our Butterworth filter. We can notice that the signal is smoother after the filtering of the data compared to figure 5.1.

5.4.2 Results

We apply the Butterworth filter to the raw EEG data, and then construct the embedding matrix of window size 30 with the resulting signal. We then perform the Independent Component Analysis on this matrix like we did previously. We notice that the convergence is faster than for the raw data.

As a result of the Independent Component Analysis, we get a set a 30 independent components on which we project the delay vectors and perform a spectral analysis of the sources. After this analysis, we notice that most of them are very similar to the previous ones we studied in the previous section.

Here is an example of signal we obtain. It is very similar to the source displayed on figure 5.2

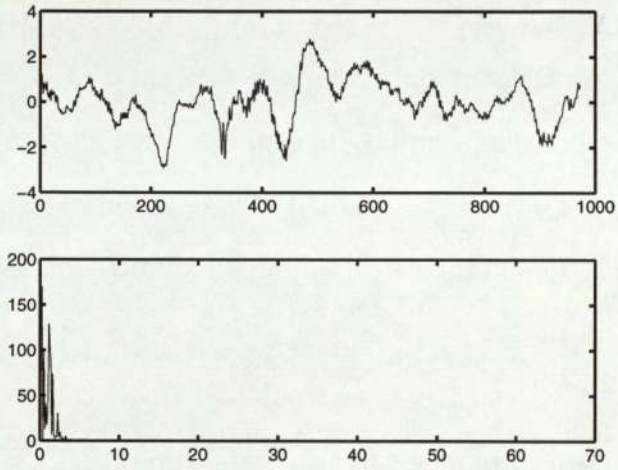


Figure 5.11: 10th source of the filtered data and its power spectral density

The sources are much "cleaner" with filtered data. We got rid of the noise from the fluorescent lamp, and of the extraneous noise above 40 Hz.

the original signals. In appendices A, B, C and D, we display the whole 30 sources and their *power spectral densities* resulting from our experiments on 1000 samples of our data for each task. We notice that after removing the noisy sources (very low or very high frequencies on the power spectral densities) and the redundant sources (same sources with different scaling, inverted sources), we can classify our sources by order of frequencies and that clusters naturally form from there. This allows us to verify our first hypothesis (see Chapter 3 section 3.4) : we can really identify four, five or six clusters in the interesting output sources which carry information in the signal domain.

One question remains : *How to analyse the behavioral significance of such sources ?* We must not forget the first aim of this study : can we simply characterize *vigilance* during the EEG trials just by extracting interesting brain activities from the output sources of the ICA. Unfortunately, not yet. Specialists only could give further explanations of the results, and some more experiments should be conducted on particular samples of the original EEG signals and compare the results with the "apparent" vigilance of the subjects during the trials.

competitive and unsupervised, meaning that no teacher is needed to define the correct output (that is to say the cell into which the input is mapped) for an input. The locations of the responses in the array tend to become ordered in the learning process as if some meaningful nonlinear coordinate system for the different input features were being created over the network.

The SOM was developed by Professor Teuvo Kohonen in the early 1980s [18].

6.2.1 The Self-Organizing Map Algorithm

Assume that the sample data sets have to be mapped onto the array depicted in figure 6.1. The set of input samples is described by a real vector $x(t) \in \mathbf{R}^n$ where t is the index of the sample, or the discrete time coordinate. Each node i in the map contains a model vector $m_i(t) \in \mathbf{R}^n$, which has the same number of elements as the input vector $x(t)$.

The stochastic SOM algorithm performs a regression process. Therefore, the initial values of the components of the model vector, $m_i(t)$, may even be selected at random.

Any input item is thought to be mapped into the location, the $m_i(t)$ of which matches the best with $x(t)$ in some metric. The self-organizing algorithm creates the ordered mapping as a repetition of the following basic task :

1. An input vector $x(t)$ is compared with all the model vectors $m_i(t)$. The best-matching unit (node) on the map, i.e., the node where the model vector is most similar to the input vector in some metric (e.g. euclidean) is identified. This best matching unit is called the winner.

2. The model vectors of the winner and a number of its neighboring nodes in the array are changed towards the input vector according to the learning principle specified below.

The basic idea of the SOM learning process is that, for each sample input vector $x(t)$, the winner and the nodes in the neighbourhood are changed closer to $x(t)$ in the input

data space. During the learning process, individual changings may be contradictory, but the net outcome in the process is that ordered values for the $m_i(t)$ emerge over the array.

Adaptation of the model vectors in the learning process may take place according to the following equations :

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)[x(t) - m_i(t)] & \text{for each } i \in N_c(t) \\ m_i(t) & \text{otherwise} \end{cases}$$

where t is the discrete-time index of the variables, the factor $\alpha(t) \in [0, 1]$ is a scalar that defines the relative size of the learning step, and $N_c(t)$ specifies the *neighbourhood* around the winner in the map array.

At the beginning of the learning process, the radius of the neighbourhood is quite large, but it is made to shrink during the learning. This ensures that the global order is obtained already at the beginning, whereas towards the end, the radius gets smaller, the local corrections of the model vector in the map will be more specific. The factor $\alpha(t)$ also decreases during the training.

One method of evaluating the quality of the resulting map is to calculate the average quantization error over the input samples, defined as $E\{\|x - m_c(x)\|\}$ where c indicates the best matching unit for x. After training, for each input sample vector, the best-matching unit in the map is searched for, and the average of the respective quantization errors is returned.

6.2.2 Experiments and Results

The SOM_PAK program, developed by Kohonen, Hynninen, Kangas and Laaksonen, was used during all the course of these experiments [18].

The Self-Organizing Map was applied to the recognition of topographic patterns on the resulting sources of our former Blind Source Separation (see Chapter 5).

The training set consists of 24 sources of 500 samples each. We do not use any labelling at all to characterize the sources on this set. For each task, we obtain a

square map and by stronger patterns in its top left and bottom right corner. Whereas boring tracking tasks are characterized by lighter patterns in the upper left side and brighter patterns in the bottom left of the upper right corner.

The main problem is that, in that way, we do not determine the kind of brain activity these lighter and brighter patterns correspond to.

6.3 Sammon Mapping

6.3.1 The Algorithm

We said in the introductive section that in a topographic map, N data vectors $\{x_i\}$ in $\mathbf{R}^p$ are transformed into a corresponding set of feature vectors $\{y_i\}$ in $\mathbf{R}^q$ such that $q < p$ and the *geometric structure* of the input data vector remains unchanged. The *Sammon Mapping* is the most intuitive basis for this definition since it is generated by the minimization of an error measure E of the inter-point distances also called STRESS

$$E = \frac{1}{\sum_i \sum_{j<i} d^*_{ij}} \times \frac{\sum_i \sum_{j<i} (d^*_{ij} - d_{ij})^2}{d^*_{ij}} \tag{6.1}$$

where $d^*_{ij} = \|x_i - x_j\|$ is the distance between points i, j in the input data set and $d_{ij} = \|y_i - y_j\|$ is the distance between their images in the map or feature space.

The procedure for performing the transformation is shown in figure 6.3 and summarised in Table 6.1.

Various error minimisation procedures can be used, one of which is the gradient descent procedure. But this procedure can get trapped in local minima.

1. Compute inter-point distances in the original space.
2. Initialise target space by a random number generator.
3. Calculate mapping error between original and target space.
4. Modify coordinate points in the target space by means of a non-linear procedure.
5. Repeat Step 3 until the mapping error is sufficiently small.

Table 6.1: Sammon Mapping's algorithm

Let us now come back to equation (6.1). The $d^*_{ij} - d_{ij}$ term represents a measure of the deviation between the corresponding distances. The Sammon STRESS thus represents an optimal matching of the inter-point distances in the input and map spaces. According to [24], normalising the expression by the first fractional term reduces the sensitivity of the measure to the number of input points and their scaling. Moreover, to render the overall measure dimensionless, the d_{ij} term is included in the denominator of the sum to moderate the domination of errors in large distances over those in smaller distances.

In the standard Sammon Mapping, the STRESS is minimised by adjusting the location of the points y_i directly, according to a gradient-descent scheme. For each point y_i, we define a parameterised non-linear function of the input $f(x_i; w)$, where w is the weight vector. Then the STRESS becomes :

$$E = \sum_{i}^{N} \sum_{j}^{N} (d^*_{ij} - \|f(x_i; w) - f(x_j; w)\|)^2 \tag{6.2}$$

Then, it is straightforward to differentiate E with respect to the mapped coordinates y_i and optimise the map using standard error-minimisation methods. This gives :

$$\begin{aligned} \frac{\partial E}{\partial w_k} &= \sum_{i}^{N} \frac{\partial E}{\partial y_i} \cdot \frac{\partial y_i}{\partial w_k} \\ &= \sum_{i}^{N} \frac{\partial E}{\partial y_i} \cdot \frac{\partial f(x_i; w_k)}{\partial w_k} \end{aligned} \tag{6.3}$$

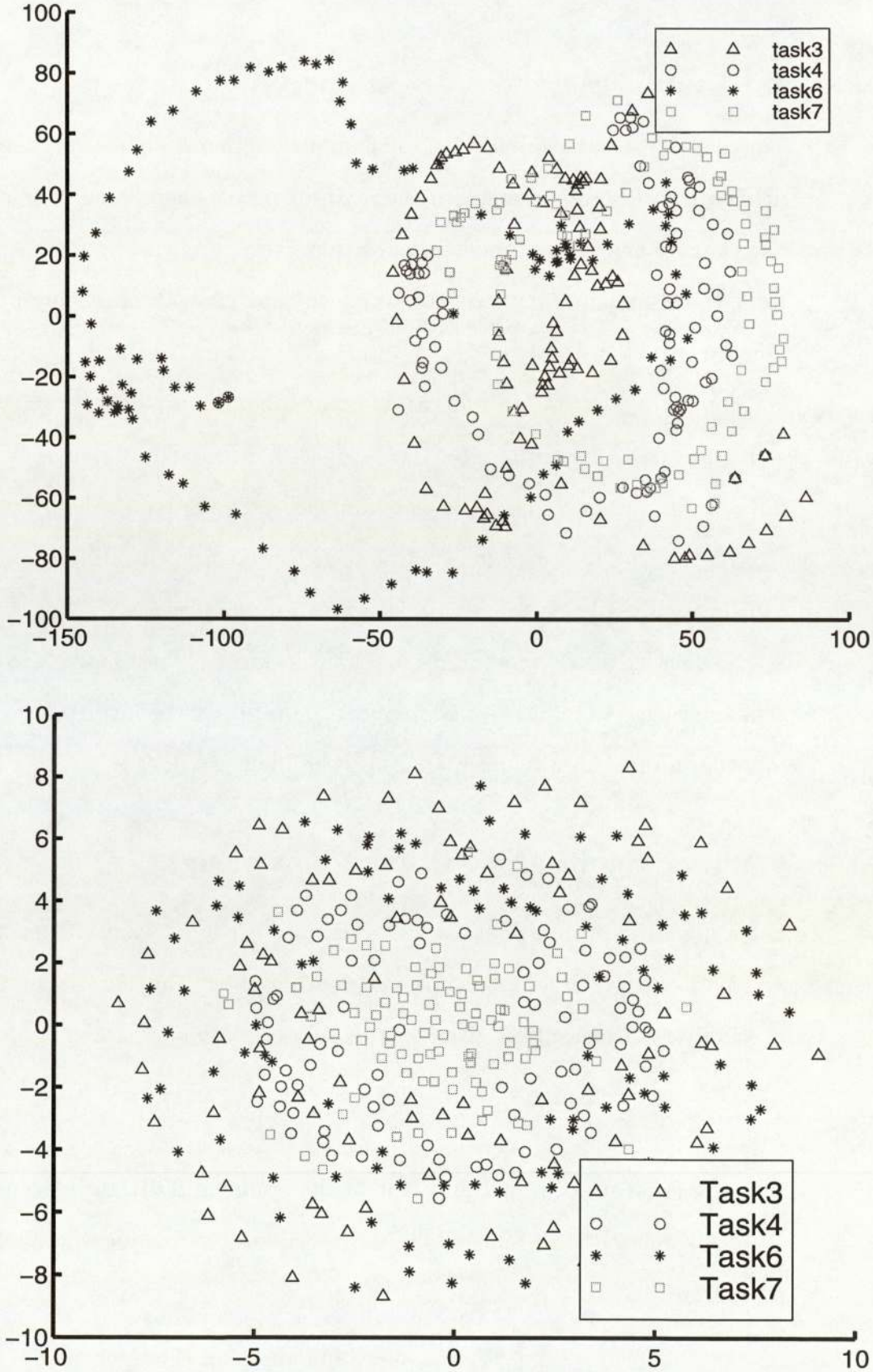


Figure 6.4: Sammon Maps of a projection of our EEG data on the principal components (top) and on the independent components (bottom)

The constant term k is added to the inter-point distances of two pairs of points so that their separation should be exaggerated in the resultant map.

In many problems, there may be further knowledge available regarding class relationships which we call *subjective dissimilarity* and will denote in the following section $S = [s_{ij}]$ corresponding to each d^*_{ij}. The assignment of class dissimilarity means that for every pair of data points in addition to the *objective dissimilarity*, there is a dual *subjective dissimilarity* which stresses alternative knowledge about the data. Thus, one can relate the existence of this set of subjective dissimilarities to a *subjective metric* implicitely defined over the input space.

6.4.3 NeuroScale

NEUROSCALE is a technique which transforms a p-dimensional input space into a q-dimensional feature space ($q < p$) with a feed-forward radial basis function. The network is trained with the same algorithm as with a Sammon Mapping (see Table 6.1) but by minimizing the following stress measure :

$$E = \sum_{i}^{N} \sum_{j<i}^{N} (\delta_{ij} - \|y_i - y_j\|)^2 \tag{6.7}$$

where

$$\delta_{ij} = (1 - \alpha) d^*_{ij} + \alpha s_{ij} \tag{6.8}$$

The parameter α ($0 \leq \alpha \leq 1$) controls the degree to which the subjective metric S influences the output configuration. One can say that it helps finding a nice middle between an unsupervised and a supervised mapping.

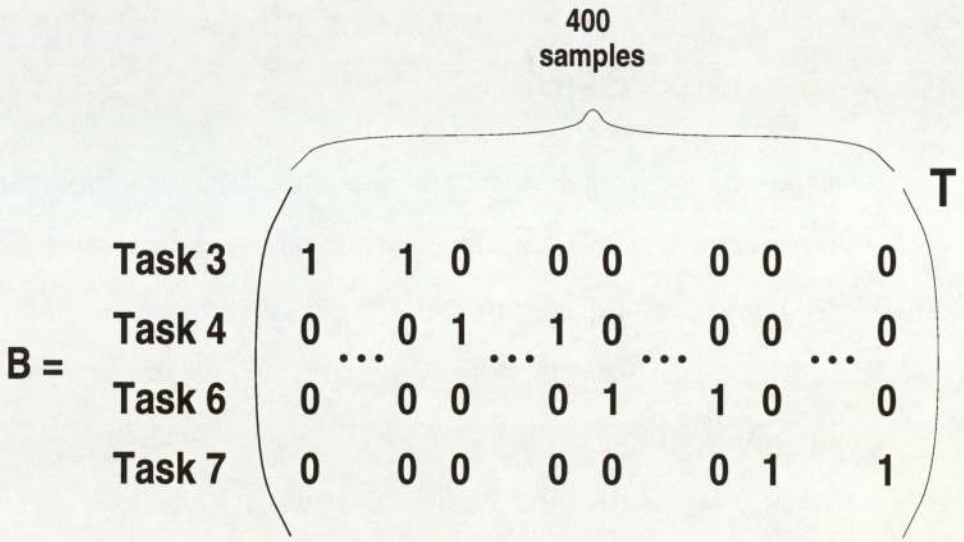
The main difference between the algorithm of NEUROSCALE and the Sammon Mapping algorithm (Table 6.1) is that this latter is fixed, *i.e.*, we know when it has converged. We must alter it in order to : include a calculation of the elements of the input space distance matrix, take into account the particular value of α. If α equals 0, then the algorithm computes a parameterized Sammon Mapping. If α equals 1,

6.4.4 Experiments and Results

To minimize E, a gradient descent algorithm is employed. The network weights are initialised at random.

The data comprises 100 samples for each task taken from the same sets we used for performing the ICA on which we do an embedding in order to look for the independent components. After computing the ICA on the embedded matrix, we obtain a 400×30 matrix of independent sources on which we can perform the NEUROSCALE algorithm.

We also choose a subjective metric which only takes into account the task knowledge relative to the data. So we build a boolean matrix B corresponding to each task in the following way :



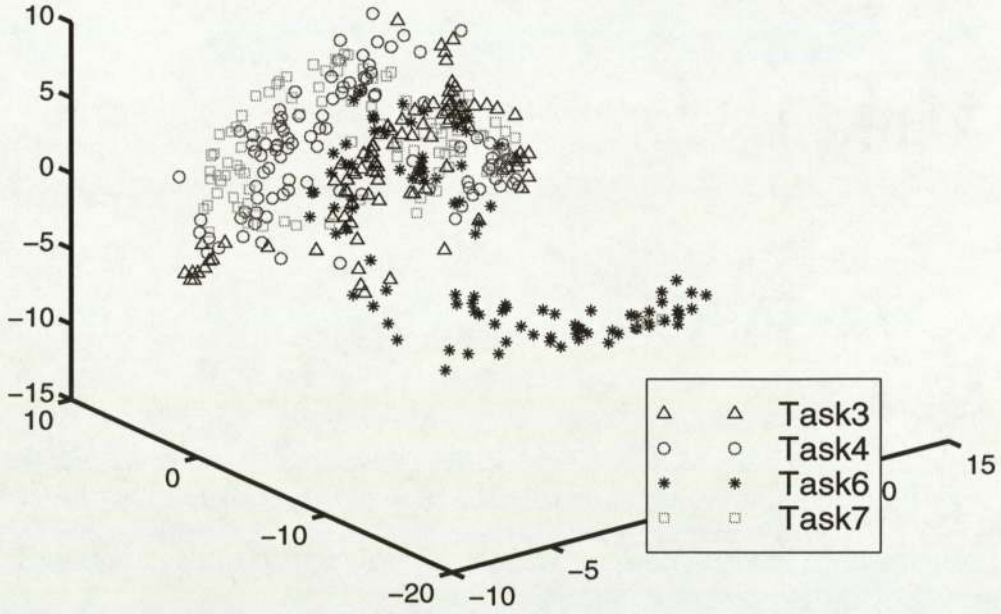
Then, to compute our subjective matrix S, we just have to compute the inter-point euclidean distances of the matrix B.

With $\alpha = 0.5$, we both retain some of the objective spatial topology and impose some task knowledge to the output configuration.

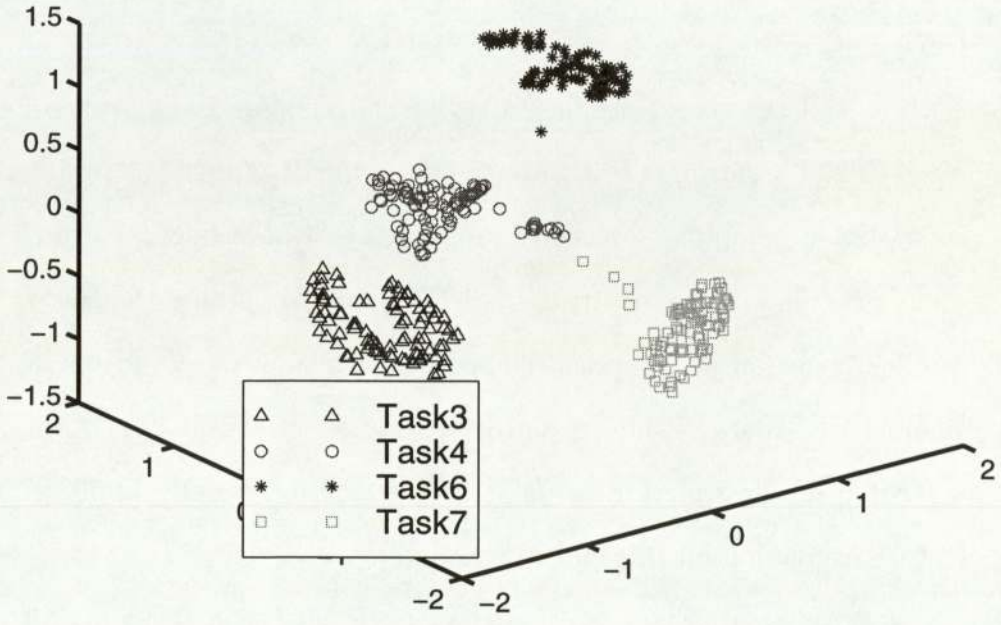
It is interesting to first observe the differences between these two maps and the Sammon Maps (see figure 6.4). We can see that the NEUROSCALE algorithm has split the clusters up even more to reveal some class knowledge. On the projections on the principal components, the algorithm has kept the ordering of the Sammon Mapping,

spiral shape of the principal components). But we are still unable to identify their meaning.

NEUROSCALE displays some advantageous features which the Sammon Mapping does not. By incorporating varying degrees of subjective knowledge, we can influence the extracted feature space. The resulting sphere, with its four distinct clusters and their evolution with the changing of the parameter α gives us a better idea of clustering than the Sammon Mapping. And of course, the NEUROSCALE map is far more representative than the Kohonen SOM. Moreover, there is a cost function associated with a particular mapping which allows us to assess individual maps and to compare them.



(a)



(b)

Figure 6.7: NEUROSCALE of a projection of our EEG data on the principal components (top) and on the independent components (bottom) with $\alpha = 0.9$. The value of the STRESS measure was : 2.98 for case a), 0.13 for case b)

the mixture of sources and remove them in order to extract the information-carrying sources that is the main brain activity corresponding to the behavioural state of the subject when achieving a particular task during a trial.

Moreover, these sources characterising brain activity, also provide justification for our main fundamental hypothesis that there are only a few of them constitutive of a nonlinear system with just a few degrees of freedom. Unfortunately, we are not yet able to extract these interesting sources directly without performing a power spectral analysis on each of them separately to isolate them from the noisy ones as there is yet no existing algorithm to classify these sources.

In Chapter 6, it was reasoned that the Sammon Mapping and NEUROSCALE were most effective strategies for topographic dimension reduction. The main advantage of Kohonen's approach is computational, because realistically, application of the Sammon Mapping is restricted to fewer than 1000 data points.

The feed-forward neural network topographic mapping technique NEUROSCALE , was thus based upon the Sammon Mapping and utilises a radial basis function neural network. Because of this neural network element, it offers the capability of generalisation to new data — a feature absent from Sammon's original algorithm.

An important extension embodied in NEUROSCALE is the capacity to exploit additional information in the mapping process. In standard approaches to topographic mapping, the geometry of the output space is determined solely according to some conventional metric (generally Euclidean) defined over the data space. If alternative information is available — such as class labels — then this may be allowed to influence the mapping (in order to emphasise clustering, for example).

The results shown by both Sammon Mapping and NEUROSCALE are interesting in a way that they show some clustering according to the types of task, but they also prove that there is another important clustering that we are not yet able to interpret and which is not related to the time history of the four tasks. One can suggest that this particular feature of the map comes from the "way" the subject has undertaken

Bibliography

[1] Bell A and Sejnowski T. "An Information-Maximisation Approach to Blind Separation and Blind Deconvolution". *Neural Computation*, 7:1004–1034, 1995.

[2] Cichocki A, Kasprzak W, and Amari S. "Multi-layer Neural Networks with a Local Adaptative Learning Rule for Blind Separation of Sources". *Proceedings of the Research society of Nonlinear Theory and its Aplications (NOLTA)*, 1995.

[3] Cichocki A, Kasprzak W, and Amari S. "Robust Neural Networks with Online Learning for Blind Identification and Blind Separation of Sources". *IEEE Transacions on Circuits and Systems – Fundamental Theory and Applications*, 43, 1996.

[4] Hyvärinen A. "A Family of Fixed-Point Algorithms for Independent Component Analysis". *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3917–3920, 1997.

[5] Hyvärinen A. "Independent Component Analysis by Minimization of Mutual Information". *Technical Report, Helsinki University of Technology, Laboratory of Computer Science*, 1997.

[6] Hyvärinen A and Oja E. "A Fast Fixed-point Algorithm for Independent Component Analysis". *Neural Computation*, 9:1483–1492, 1997.

[7] Loomis AL, Harvey E, and Hobart GA. "Cerebral States during Human Sleep as studied by Human Brain Potentials". *Journal of Experimental Psychology*, 21, 1937.

[19] Karhunen J, Oja E, Wang L, Vigrio R, and Joutsensalo J. "A Class of Neural Networks for Independent Component Analysis". *IEEE Transactions on Neural Networks*, 1997.

[20] Karhunen J and Pajunen P. "Hierarchic Nonlinear PCA Algorithms for Neural Blind Source Separation". *Proceeding of the IEEE Nordic Signal Processing Symposium (NORSIG)*, 1996.

[21] Karhunen J and Pajunen P. "Blind Source Separation and Tracking using Nonlinear PCA Criterion : A Least-squares Approach". *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, 1997.

[22] Mao J and Jain AK. "Artificial Neural Networks for Feature Extraction and Multivariate Observations". *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability*, 1995.

[23] Cardoso JF and Souloumiac A. "Blind Beaforming for Non Gaussian Signals". *IEE-Proceedings*, 140:362–370, 1993.

[24] Sammon JW. "A Nonlinear Mapping for Data Structure Analysis". *IEEE Transactions on Computers*, C–18(5):401–409, 1989.

[25] Knuth KH. "Difficulties Applying Recent Blind Source Separation Techniques to EEG and MEG". *Maximum Entropy and Bayesian Methods*, 1997.

[26] Girolami M and Fyfe C. "Blind Separation of Sources using Exploratory Projection Pursuit". *Proceedings of the International Conference on the Engineering Applications of Neural Networks (EANN)*, 1996.

[27] Tipping ME. *Topographic Mappings and Feed-Forward Neural Networks.* PhD thesis, 1996.

[28] Delfosse N and Loubaton P. "Adaptative Blind Source Separation of Independent Sources : A Deflation Approach". *Signal Processing*, 45, 1995.

[39] Jung T-P, Makeig S, and Sejnowski T. "Using Feedforward Neural Networks to Monitor Alertness from Changes in EEG Correlation and Coherence". In *Advances in Neural Information Processing Systems 8.*, 1996.

[40] Li X, Gasteiger J, and Zupan J. "On the Topology Distorsion in Self-Organizing Feature Maps". *Biological Cybernetics*, 70, 1993.

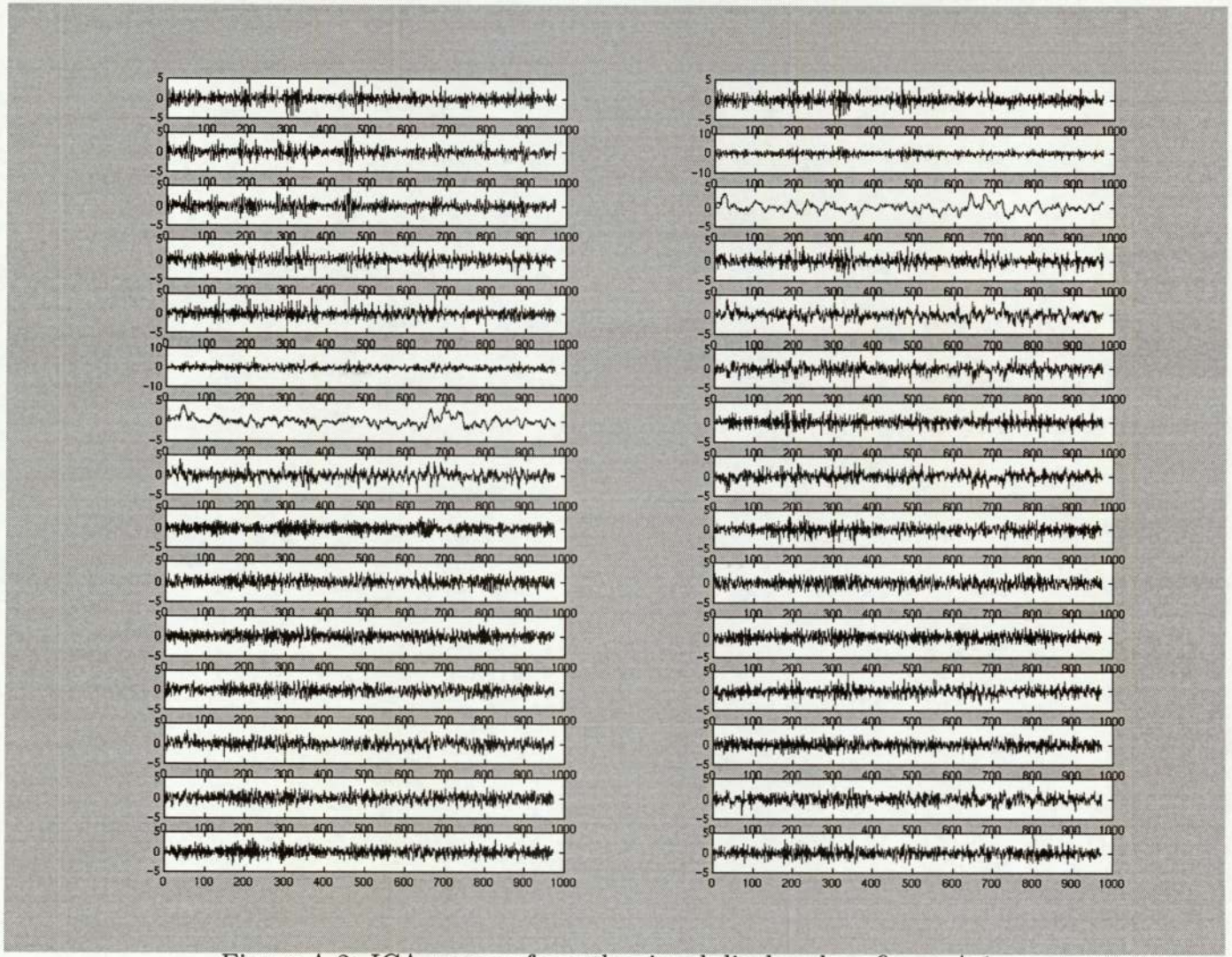


Figure A.2: ICA sources from the signal displayed on figure A.1

Appendix B

Results for task4

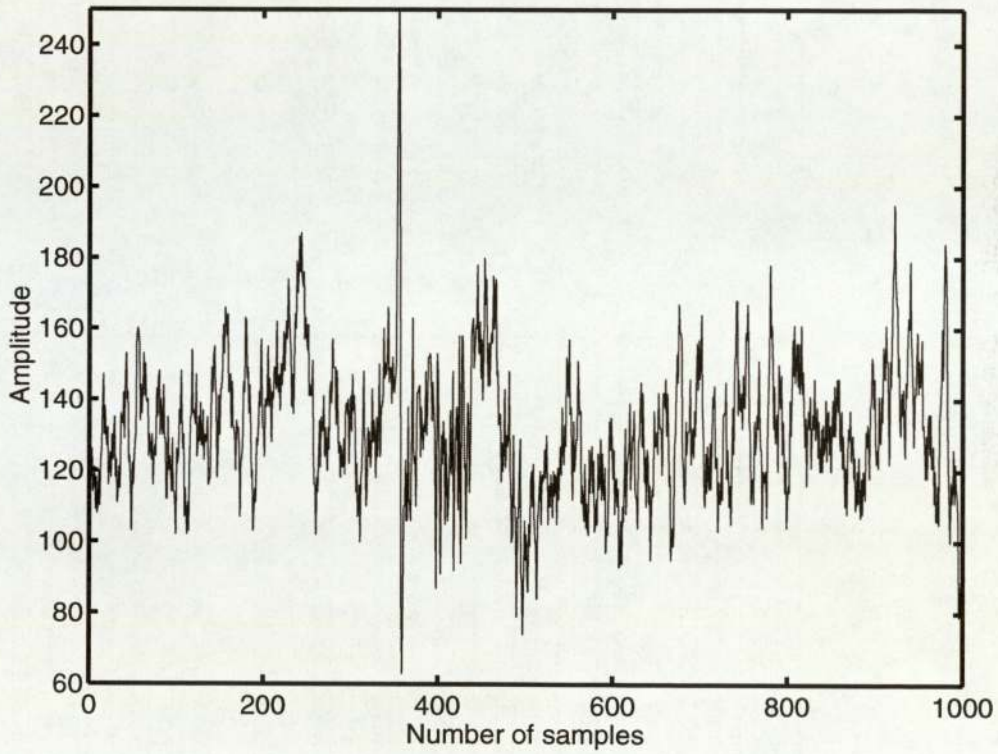


Figure B.1: 1000 samples taken from Task4 on which we apply the ICA algorithm

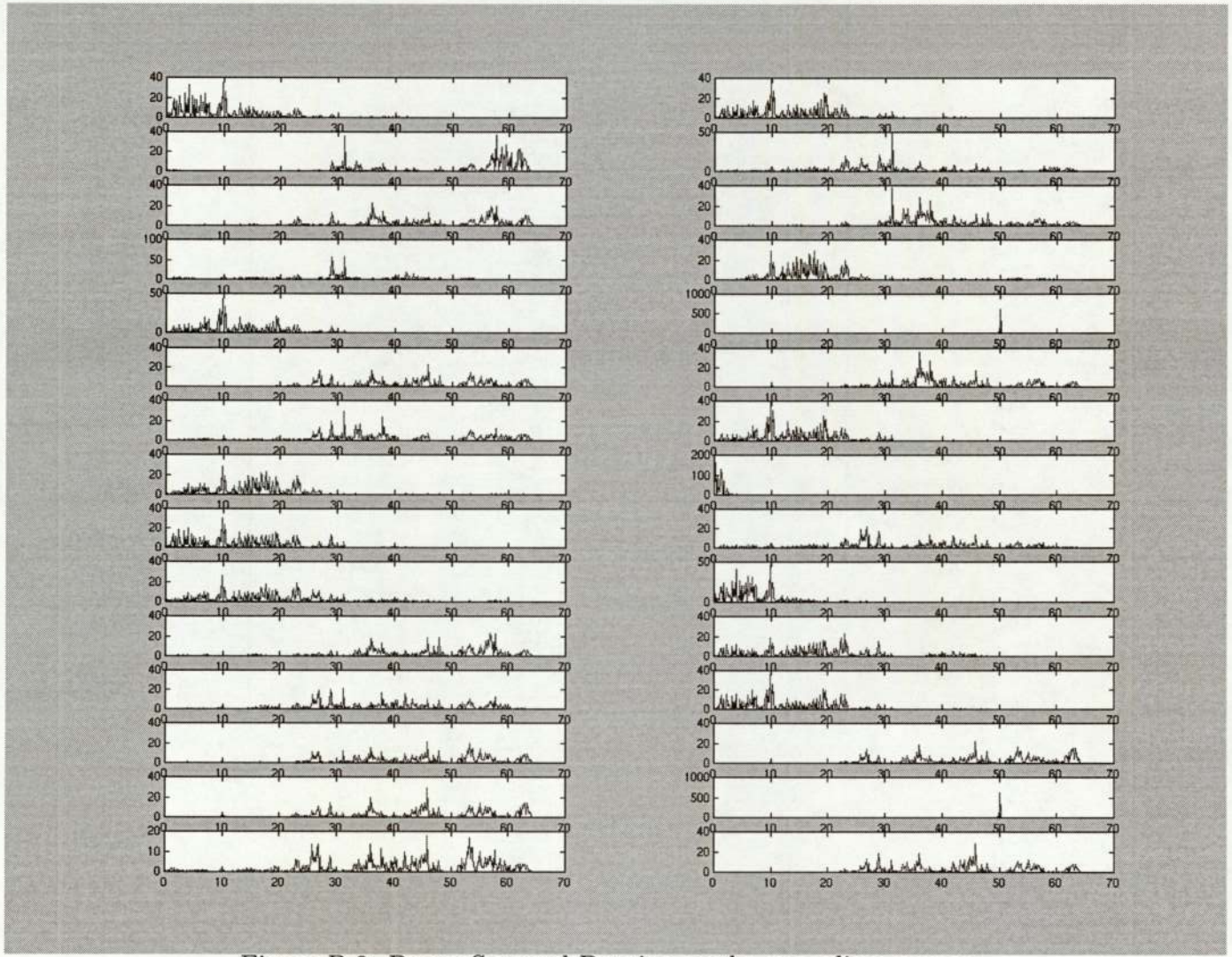


Figure B.3: Power Spectral Density on the preceding sources

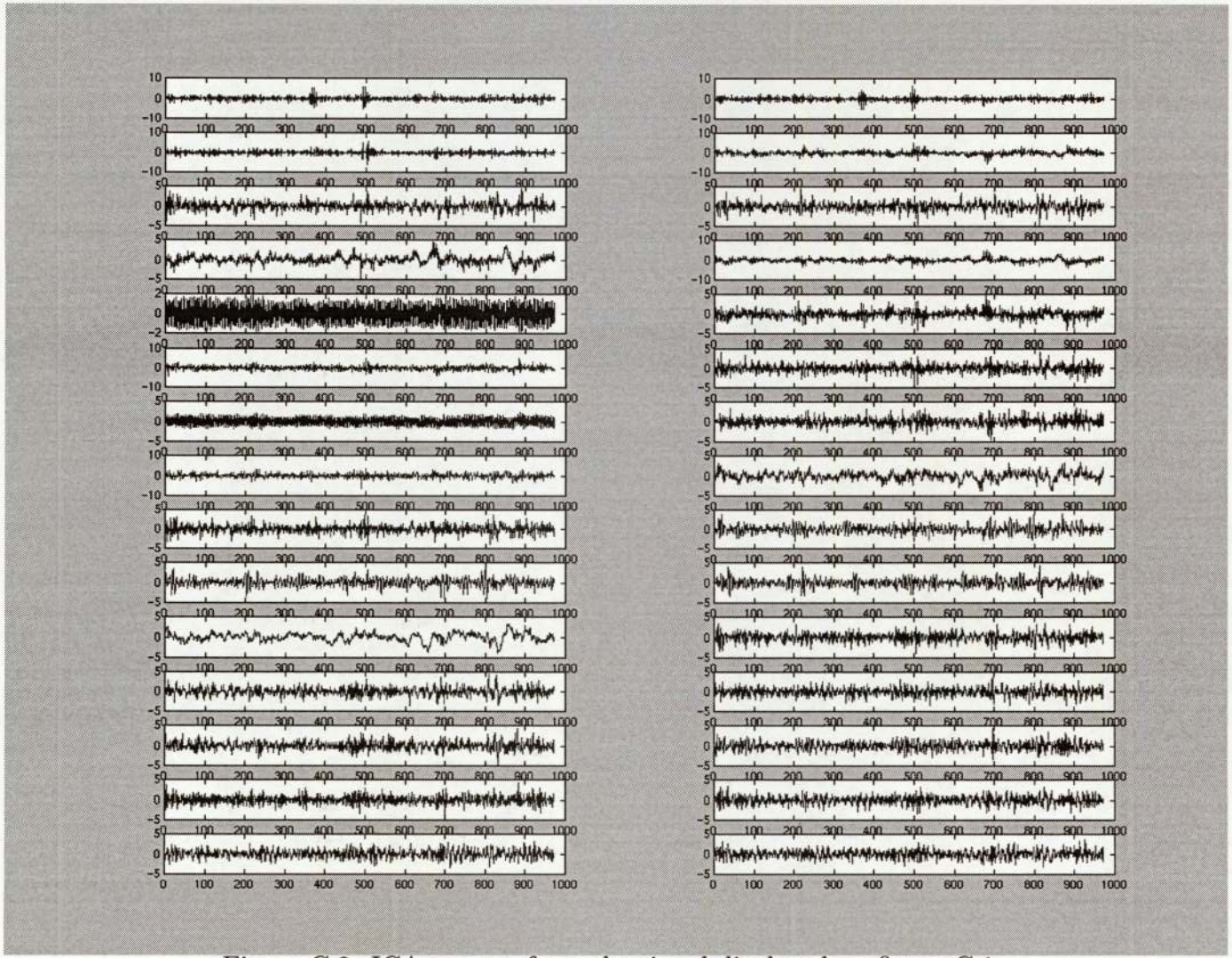


Figure C.2: ICA sources from the signal displayed on figure C.1

Appendix D

Results for task7

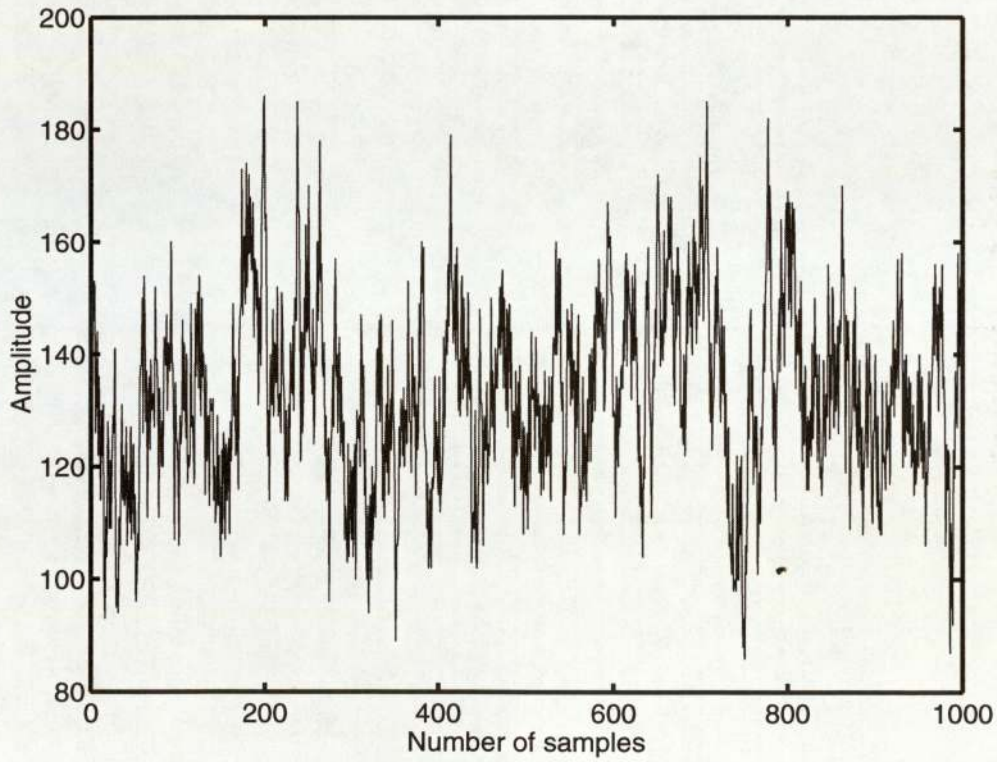


Figure D.1: 1000 samples taken from Task7 on which we apply the ICA algorithm

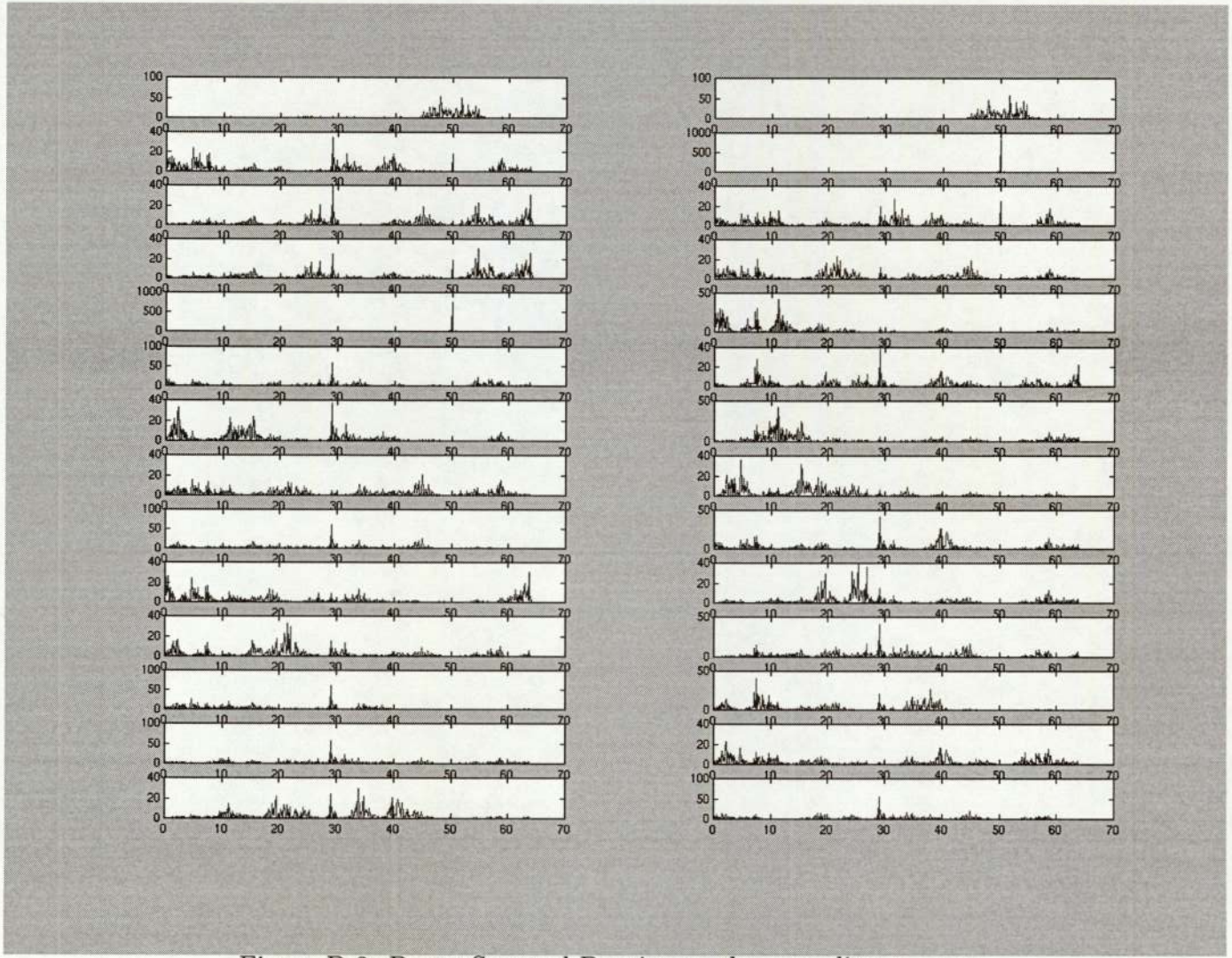


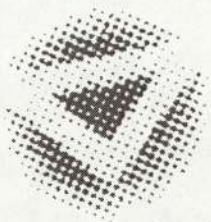
Figure D.3: Power Spectral Density on the preceding sources

Dynamical Embedding and Feature Extraction of Electroencephalographic Data

Nathalie Christiane Noël

MSc in Pattern Analysis and Neural Networks

Supervisor: Professor David Lowe



Aston University

September 1998

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor David Lowe, for his excellent guidance, stimulating ideas and his ability to always find time for me when there appeared to be none.

I would also like to record my appreciation to Doctor Helen Stone of Sowerby Research Center who aided my understanding of electroencephalograms.

Finally, I am grateful to my parents and all my friends on Aston Campus for their support and assistance.

Chapter 1

Introduction

> ... I shall demonstrate how this tiny sound within, this nothing, contains everything; and how, with the bacillary aid of a single sensation — always the same one, and deformed at that in its very origins — a brain isolated from the world can create a world in itself ...
>
> **Remy de Gourmont** — *Sixtine.*

Measurements of brain activity can be performed by recording the electric potentials on the scalp surface : this is known as *electroencephalography* (**EEG**). EEG analysis has played a key role in the modeling of the brain's cortical dynamics. If several mental states can be reliably distinguished by recognizing patterns in EEG, then is it possible to utilise EEG information to automatically estimate *car driver* or *pilot workload* for example ? By estimating workload, we mean assessing the level of attentiveness or vigilance of a pilot.

There have been many studies of alertness to try to discover whether vigilance may be recognized from single or multi-channel EEG traces ([35], [36], [39]), since the first investigation by Loomis *et al* in 1937 [7]. Most of them have confirmed that, despite sincere intentions, few subjects remain vigilant while engaged in monotonous monitoring tasks.

The analysis of the data is problematic due to the fact that multiple neural generators of the EEG may be simultaneously active and the potentials and electromagnetic

are distorted by the head volume conductor), reference electrode effects and algorithm effects (algorithms that are adopted to reduce volume conduction effects may introduce false coherency estimates) [32].

The fundamental reason why EEG analysis is performed in the frequency domain is because of the belief in the linear nature of the physical sources generating the potential differences measured by the sensors. This linearity suggests that the signal might be decomposed into a sum of sinusoidal components. So we are supposed to obtain a description of the signal in terms of its fundamental frequency characteristic.

Recently, blind source separation by *Independent Component Analysis* (**ICA**) has received attention because of its potential applications in signal processing such as in speech recognition systems, telecommunications and medical processing. The goal of ICA is to recover independent sources given sensor outputs in which the sources have been linearly mixed. In contrast to correlation based solutions such as Principal Component Analysis (**PCA**), ICA not only decorrelates the signals but also reduces higher-order statistical dependencies, attempting to make the signals as independent as possible. The blind source separation problem has been studied by researchers in the field of neural networks [33], [1], [2], [26], [19], [8]. It has also been applied to the particular field of EEG in several studies [35], [36], [39], [25]. All these studies only consider multi-channel EEG recordings to perform the ICA algorithm.

For this study, we have made the choice to consider only single channel EEG data recorded from wake subjects due to the hypothesis that over short segments of EEG data, we can reconstruct the dynamics of the system with a dynamical embedding of one single channel. We will question the use of linear Fourier analysis within the wake state, essentially because of the nature of the noise sources and complexity of the signal. Indeed, we will start from the hypothesis that the complexity in wake EEG is due to the nonlinear interaction of a few degrees of freedom rather than the linear interaction of many degrees of freedom, plus additive noise. This is a dynamical systems perspective which considers the existence of an *underlying data generator* (or *attrac-*

consists of finding the statistically independent sources responsible for a set of data. The most familiar situation is the "cocktail party problem" where there are many speakers, or sources of accoustic signals, and the listener detects mixtures of these signals.

Finally, *topographic mappings* may be viewed as nonlinear, unsupervised feature extraction processes. Here the criterion for selection of features is not to maximise variance or any mutual information, but rather that the *topology* or geometric structure of the data be preserved in the feature space.

In the feature space, delay vectors x are expressed as a linear combination of spanning basis functions v_i and a set of "source" signals $\alpha_i(t)$

$$x(t) = \sum \alpha_i(t) v_i$$

In a Principal Component Analysis embedding, the basis functions v_i are obtained as the eigenvectors of a covariance matrix (see Chapter 3). In an Independent Component Analysis, the expansion basis vectors v_i are instead determined as the independent components of a demixing matrix (see Chapter 4).

In both cases (PCA and ICA), we project the data linearly on the embedding vectors and therefore reduce the dimensionality of our feature space. This also allows us to reduce some of the noise structure (see figure 1.1).

We can now build a model in this reduced (dimension and noise) feature space (see Chapter 6), search for some dynamic structure, identify anomalous behaviour and look for interesting structure which might characterise vigilance.

Figure 1.1 presents a symbolic overview of the structure and function of this thesis.

Chapter 5 details all the results with ICA applied to our raw or filtered data. It discusses these results in terms of frequencies and validates the hypothesis made in **Chapter 3**.

Chapter 6 applies different methods of topographical mappings to the results of ICA. These methods are : Kohonen Self-Organizing Maps, Sammon Mapping, NEUROSCALE The different results are then discussed and the methods are compared.

Chapter 7 concludes this thesis, discusses the major results and gives directions for further research.

sources rather than sources farther from the sensors.

- EEG is most sensitive to correlated dipole layer in ab, de, gh (that is perpendicular fields), less sensitive to correlated dipole layer in hi (that is tangential fields) and insensitive to opposing dipole layer in bcd, efg.

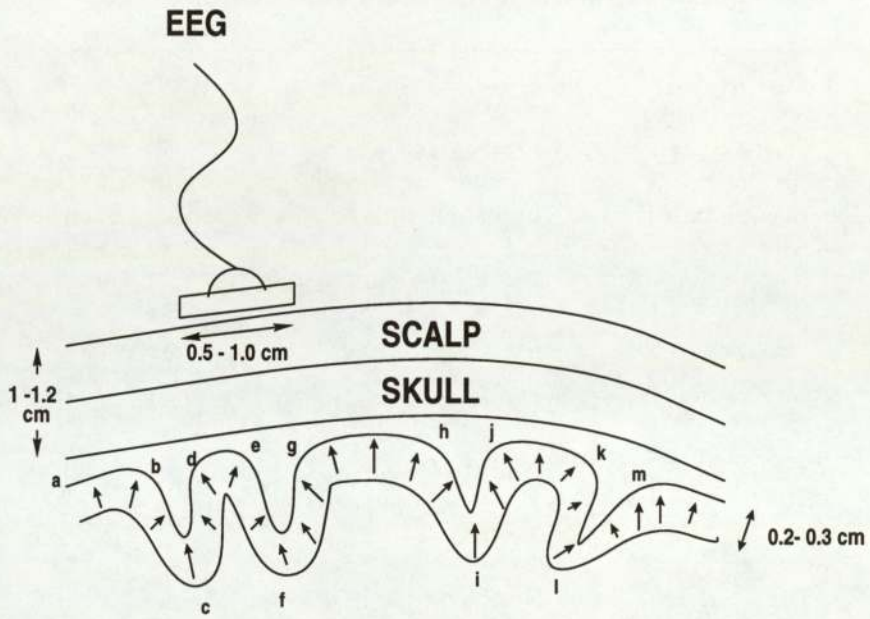


Figure 2.1: Extracranial recording of EEG data

With all the limitations of extracranial recordings, we can wonder if anything valuable can be extracted from EEG signals. However, it has long been appreciated (since the first EEG recording in 1928) that electroencephalography is a genuine measure of conscious experience. As a matter of fact, EEG recording has long been used in medicine as a clinical test for variety of pathologic conditions : epilepsy, Alzheimer's disease, severe head injuries, multiple tumors...

2.4 Brain Waves

Interpreting EEG involves the characterization of wave forms largely defined by their frequency and to a lesser extent by their morphology. The difficulty lies, in part, in recognizing artifacts and also in being able to differentiate normal variants from abnormalities.

2.4.1 Waves defined by frequency

Frequency means the number of waves per second. The frequencies of the EEG waves run from 0.5 per second to hundreds per second. Waves are usually defined by their frequency and are divided, on this basis, into four main groups. Figure 2.2 displays these different types of waves and the following sections describe them more precisely.

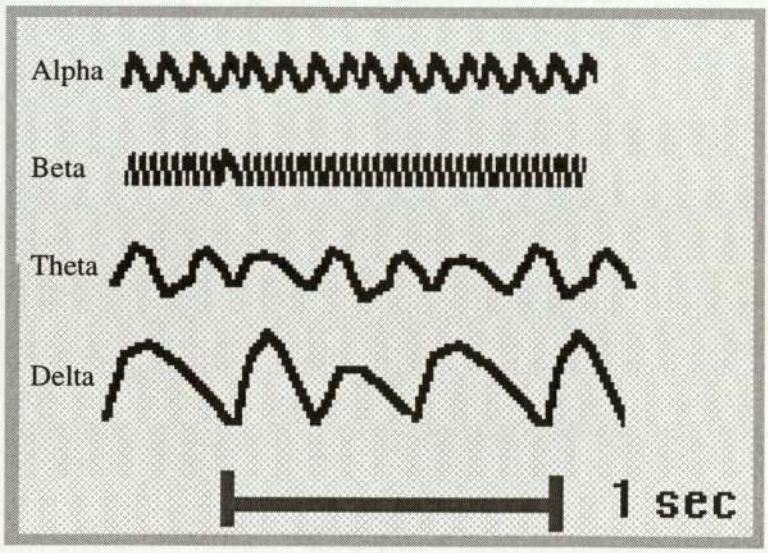


Figure 2.2: Brain waves

in children.

2.4.2 Waves defined by morphology

Certain waves have characteristic forms irrespective of their frequency and are recognizable by their shape; in other instances pair or groups of waves have typical appearances. Single waves that are specially shaped include, for instance spikes or sharp waves - waves that rise rapidly to a point and fall away equally dramatically with a base that is small compared to the wave's amplitude. Some wave forms can be recognized by their morphology and these include two main types:

- Specially shaped waves
- Specially shaped wave complexes

Artifacts

Artifacts are disturbances caused by technical defects - usually transitory. Included are such things as eye movement, electrode movement with loss of contact, muscle activity obscuring the EEG, movements of the head, scratching the scalp, sweating etc...

Normal variants

There are several waves or patterns of waves which are unusual in appearance yet are not significant for abnormality or disease. These waves can be misinterpreted. Amongst the more common ones are mu rhythm, psychomotor variant, lambda waves, POSTS, spindles, vertex waves.

EEG data was collected on an Oxford Instruments Medilog system utilising 8 separate measurement channels at the University of West England on behalf of British Aerospace. The electrodes were sited according to a standard 10-20 system and bipolar potential differences due to electroencephalographic activity were measured.

The data was sampled at 128 Hz and scaled and linearly quantised at one byte per sample. For this study, only signals extracted from scalp locations T5-Oz were used. Cross-channel effects are the subject of future work.

One major problem with using wake EEG is due to the very poor signal-to-noise ratio as a consequence of the overall mental and mechanical activity, giving rise to spurious and background electrical activity. Hence our first approach was to consider approaches to reduce the noise components in the signal and to enhance the information-carrying signal components.

3.2 Nonlinear Phenomena

The separation of linear systems into discrete and continuous systems is also appropriate for nonlinear systems. Our aim is to exhibit a new appreciation for the complexity and richness of behaviour of a system such as brain waves with only a few degrees of freedom.

There are different types of behaviour for a deterministic time series :

- equilibrium
- periodic
- quasiperiodic
- chaotic

These four types of behaviour are called *attractors*. The distinction between these different types of attractors are most obvious when applied to systems with relatively few degrees of freedom and large signal-to-noise ratio. The problem we are going to face is that our electroencephalographic data has a very poor SNR. The generator of our data will thus be quite difficult to characterise.

The general idea is that simple systems (that is systems with a few degrees of freedom) behave simply. This idea is certain for linear systems. However, with nonlinear systems, new phenomena have been described that cannot be predicted by linear theory. So even very simple systems, if nonlinear, can exhibit extremely complex behaviour. Such nonlinear systems are very interesting for appreciating their complexity and richness of behaviour despite their few degrees of freedom. They can exhibit *chaotic and dynamic behaviour* characterized by complexity and sensitivity to the initial state of the system [32].

a particular physiological state of the brain) typically is estimated by expressing the data from our single EEG channel into an M-dimensional space. This is equivalent to constructing a delay vector $X(t)$ from the initial time series

$$X(t) = (X(t-\tau), \ldots, X(t-(M-1)\tau)) \in \mathbf{R}^n$$

where τ is the time increment (or lag) and M is the embedding dimension (or number of lags). According to [15], we choose $\tau = 1$ in the whole following.

A single point in this space is then located by the vector

$$Y_1(t) = (X_1, X_2, \ldots, X_M)$$

and the next location is given by the vector

$$Y_2(t) = (X_2, X_3, \ldots, X_{M+1}), \text{ and so on.}$$

Hence, in a time series consisting of K measured values, $Y(t)$ can assume $L = K-M+1$ discrete values in the embedding space.

For a D-dimensional attractor, the *embedding dimension* M must be at least as large as D. Takens showed in [16] that for a system of D degrees of freedom, we must have $M > 2D + 1$.

We then have to determine the complexity and the window size M of the embedding. In the following study, we will consider $K = 1000$ samples which is worth 10 seconds of EEG data.

3.4 Determining the Complexity

As described before, a delay-space embedding can be used to reconstruct a multi-dimensional representation. But an important question is *"How many dimensions should be used in the representation ?"*.

It was shown in [15] that an analysis of the number of degrees of freedom in X leads to the singular value problem

$$X = S\Sigma C^T \tag{3.2}$$

by the σ_i. Furthermore, as σ_i decreases, the noise-to-signal ratio (inverse of SNR) increases. Thus it is possible to obtain information on the level of noise in the system by studying the eigenspectrum.

So, to make a choice from among all the different singular spectra, we are looking for a change in the curvature that displays the limit between the signal space and the noise space. We expect a general stability of the spectrum as sufficient information content is captured with the window size. *Convergence* of the singular spectrum is the criterion for obtaining a sufficiently large delay window.

Thus, we know that we have found the right kink when the singular spectrum does not change in the signal space when we go on incrementing the window size. This means that the window size is big enough to capture the whole dynamics of the signal generator. However, because of serial correlations in the data, the length of the delay vector is *not equivalent* to the number of degrees of freedom of the data. It is rather determined by the location of the kink on the converged spectrum. The number of degrees of freedom is the dimension of the subspace containing the embedding manifold rather than the dimension of the manifold in itself.

For our given EEG data, the singular spectra seem to converge for a window size of 30. In figure 3.1 (a), we can observe two kinks occuring at the second eigenvalue and at the eigth eigenvalue. We know the first kink is typical of such eigenspectra. They relate to the trend of the time series and gather a very large amount of variance. Therefore, we know that, they are not enough to reconstruct the signal subspace. The kink we are looking for, which shows the singular spectrum of the delay vector for a window size of 30, occurs around 8 eigenvalues.

Figure 3.1 (b) gives us the same singular spectrum by using a delay window of 8. The delay embedding has therefore 8 degrees of freedom. We can see that the curve is smoother in the noise space after the kink at about 5. But there is a residual structure in the noise space after this kink. That implies an intrisic dimensionality of the underlying manifold generating the EEG signals of about 5.

3.5 Principal Component Analysis

Now that we have chosen to build an embedding matrix of size 30 with our input vector, it would be interesting to first perform a Principal Component Analysis on our data as a preliminary linear analysis.

PCA is a commonly used method for analysing data, and it is closely related to some other methods such as least squares methods and factor analysis. The objective of PCA is to find a set of m orthogonal vectors in data space that have the greatest contribution to the data variance. Dimensionality reduction is accomplished by projecting the original data with n-dimensional space onto the m-dimensional subspace spanned by the orthogonal vectors. This projection often retains most of the essential information in the data. PCA is also used to search for clusters. The first principal component is taken along the direction of maximum variance, while the second principal component has to be the subspace perpendicular to the first one and taken along the direction of maximum variance within the subspace. Then the third principal component is a subspace perpendicular to the first two with maximum variance direction, and so on.

The above steps can be generalised that the direction of the kth principal component is along an eigenvector direction of the kth largest eigenvalue of the full *covariance matrix*. Proof can be found in reference [17].

Unfortunately, this method has a number of problems. For example, extreme points in the data set (known as outliers) can generate large errors in the eigenvalues. The structure of the data cannot be recovered, *i.e.*, there is a loss of orientation due to aliasing along the largest variance of two parallel groups of data. Finally, the linearity aspect of PCA will obviously not solve non-linear problems. PCA may yield a relatively large number ($m > 10$) of significant principal components from which we cannot obtain information (ordinarily we can visualise 2-3 dimensional data space). Moreover, a large number of data can significantly increase computational complexity, *e.g.*, computing the inverse of the covariance matrix is typically $\Theta(n^3)$.

Chapter 4

Independent Component Analysis

4.1 Introduction

We have just seen the results of a Principal Component Analysis applied to our data. Recent studies of Independent Component analysis with multi-channel EEG have proven promising [35] [36] [39]. That is why we have chosen to apply it to our single channel EEG data. The fundamental advantage of ICA over PCA is that the signal space is not constrained to be spanned by orthogonal basis vectors and hence the independent sources obtained from Blind Source Separation should have a better *interpretability* in terms of the original EEG problem.

Let us assume that we have some phenomenon which manifests itself through a set of n independent random variables. We shall denote the combination of these variables with a random vector $s = [s_1 s_2 \dots s_n]^T$. Components $s_1, s_2, \dots s_n$ are called *sources* and s is called the *source vector*. This name implies independence : the sources are assumed to be independent sources of information.

Now suppose that the original independent source components are observed via a linear process. Denote the observed random vector by x. Since the process is assumed linear, the relation between s and x can be modelled as

$$x = As \tag{4.1}$$

because any constant multiplying an independent component may be cancelled by dividing the corresponding column of the mixing matrix A by the same constant.

- there is no ordering between the independent components.

4.2 Removing correlations

Assume that our data has zero mean, that is $E\{x\} = 0$. If we can find a linear transformation giving relation (4.2), the independent components of s have zero mean as well. We assume then that the data has this property, that it has been *centered* by removing its mean and that it has been unit-varianced. So the covariance matrix of s is $cov\{s\} = I$, and components of s are uncorrelated. Uncorrelatedness is necessary but not sufficient for independence.

We can accomplish uncorrelatedness by transforming x so that its covariance matrix will be diagonal. If in addition, all components have unit variance (the covariance matrix is unity), the process of accomplishing this is called *whitening* or *sphering.*

Whitening can be done using PCA basis vectors. Let E denote the matrix of principal component basis vectors of random data vector x, i.e., the eigenvectors of $cov\{x\}$, and $D = diag(\xi_1, \ldots, \xi_m)$ a diagonal matrix of corresponding eigenvalues. The new whitened data vector v is given by

$$v = D^{-1/2} E^T x \tag{4.3}$$

Matrix $V = D^{-1/2} E^T$ is a *whitening matrix.* The fact that v is really *white* can be seen from

$$\begin{aligned} cov\{v\} &= E\{D^{-1/2} E^T x x^T E D^{-1/2}\} \\ &= D^{-1/2} E^T cov\{x\} E D^{-1/2} \\ &= D^{-1/2} E^T E D E^T E D^{-1/2} \\ &= I \end{aligned}$$

- Karhunen, Oja, Wang, Vigario and Joutsensalo [19]
- Karhunen and Pajunen [21]
- Girolami and Fyfe [26]
- Pearlmutter and Parra [8]

In Table 4.1, we describe the algorithms we have selected to use before choosing the definite one for the experiments and the study.

	Method of solution	
Mathematical approach	Diagonalization	Fixed point
Fourth order cumulants	`JADE` [23]	Original fixed point [5]
Contrasts based on other nonlinearities	—	Generalized fixed point [4]

Table 4.1: A classification of used ICA algorithms.

One especially important class of algorithms missing from this list is the set of algorithms with foundations in information theory. In section 4.3.2, I will introduce the entropy maximization algorithm of Bell and Sejnowski. Then in section 4.3.3, I will present algorithms based on batch computations. Finally, in section 4.3.4, I will present the *fast-fixed point algorithm* which is a particular method of the generalized fixed-point, I have been using for my experiments on EEG data.

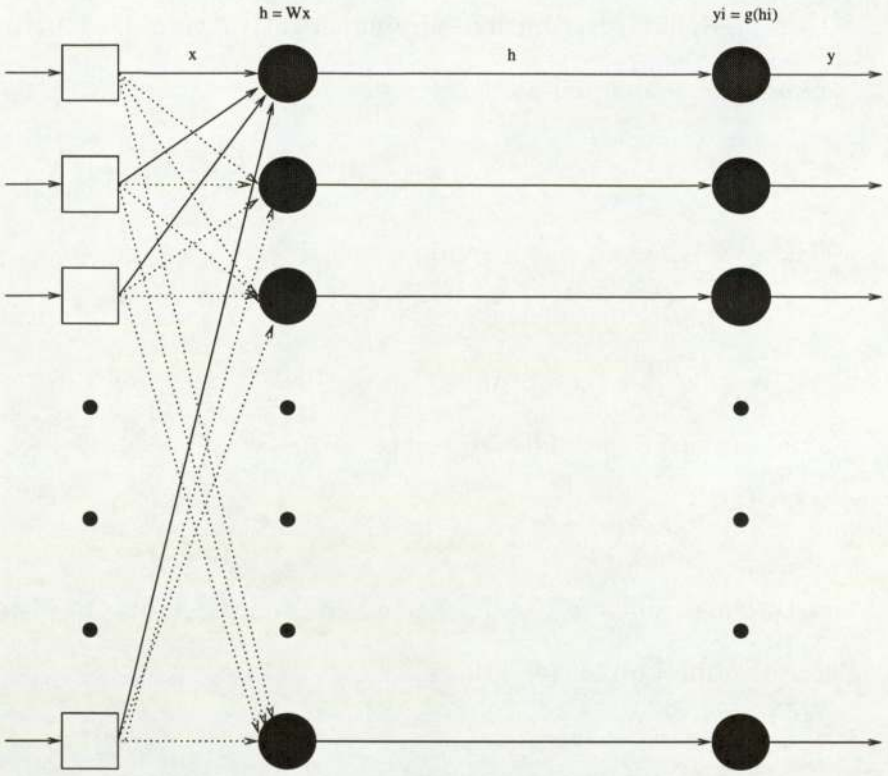


Figure 4.1: Entropy maximization error network

4.3.3 Jade Algorithm

The JADE algorithm of Cardoso and Souloumiac is based on joint approximate diagonilization of eigenmatrices [23]. The ICA problem can be solved by computing the eigenvectors of the cumulant matrix $Q_v(M)$ for any matrix M. JADE diagonalizes a set of eigenmatrices representing the whole cumulant matrix set C_v.

The problem with such an algorithm is that it uses batch tensorial computations. The data we are using is too large for such difficult tensorial computations.

4.3.4 Fast-Fixed Point Algorithm

One way to approach the ICA problem is to try to form an optimization problem that has its solutions as the independent components. We shall call such objective functions *contrast functions.*

We introduce a measure called *kurtosis.* Its value is described to measure the peakedness of the distribution, with peaked distributions giving positive values of kur-

There are two ways of solving this equation. One of them would be by applying standard numerical algorithms. Hyvärinen and Oja have chosen to write the equation in the form :

$$w = scalar \times (E\{x(w^T x)^3\} - 3\|w\|^2 w) \tag{4.12}$$

This is very useful for the scalar takes into account the penalty function which we therefore don't need to compute and hence, we don't need to take into acount the peakness of our signal (sub or super gaussian).

Then, the iteration obtained is very fast.

Using the preceding equation, we derive the following algorithm :

1. $w := rand()$
2. $w := w/\|w\|$
3. $w_{old} := 0$
4. **while** $\|w - w_{old}\| > \varepsilon \bigwedge \|w + w_{old}\| > \varepsilon$
   - $w_{old} := w$
   - $w = E\{v(w^T v)^3\} - 3w$
   - $w := w/\|w\|$

   end

Table 4.2: Hyvärinen and Oja's Fast-Fixed Point Algorithm

The final vector w equals one of the column of the mixing matrix B, which means that one of the non-gaussian independent component has been separated. Thus to estimate n independent components, one needs to run the algorithm n times. We are sure that we estimate each time a different component thanks to the orthogonalizing projection inside the loop.

Hyvärinen and Oja prove that their algorithm has a cubic convergence.

Chapter 5

Experiments and Results

5.1 Introduction

The Independent Component Analysis is ideally suited for performing source separation in domains where :

- the sources are independent
- the propagation delay of the "mixing medium" are negligible
- the sources are analog and have probability density functions not too unlike the gradient of the logistic sigmoid
- the number of independent signal source is the same as the number of sensors

In our case of EEG signal, one scalp electrode picked up correlated signals at different times of the day on a wake human being executing four different tasks. We would like to know what effectively *independent brain sources* generated these mixtures. If we assume that the complexity of EEG dynamics can be modelled as a collection of statistically independent brain processes, the EEG source analysis problem satisfies ICA assumption 1. Since volume conduction in brain tissue is effectively instantaneous, ICA assumption 2 is also verified. Assumption 3 is plausible. But assumption 4 is

5.3 Independent Component Analysis of Raw EEG Data

I have chosen to show here only the most relevant results on task 4 in order to remain as concise as possible.

Figure 5.1 shows the original time serie of 1000 samples we are going to work on.

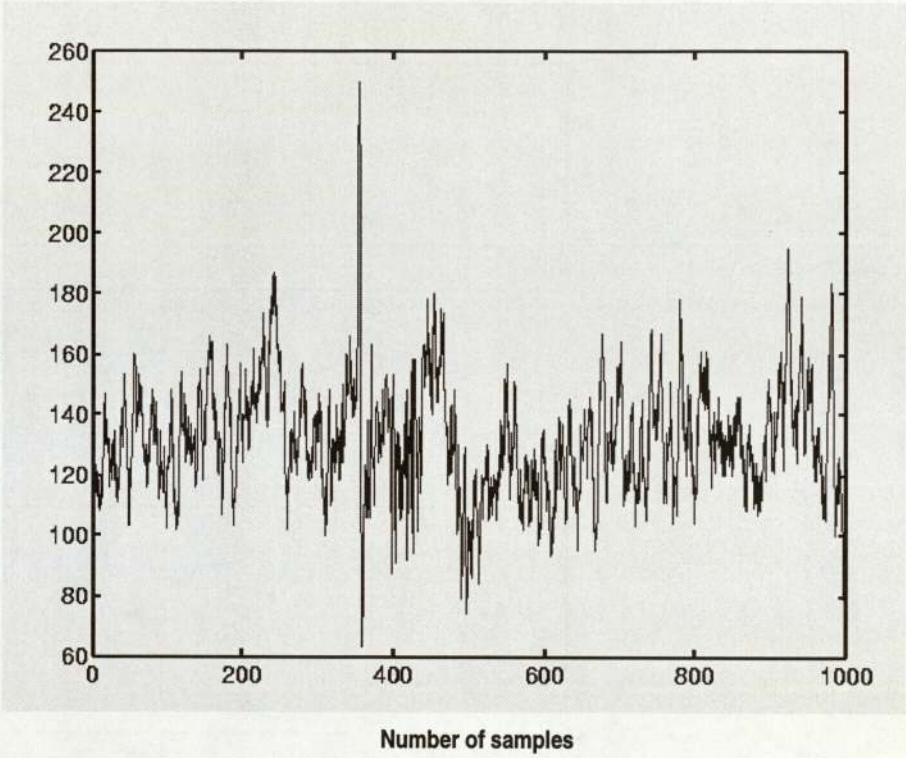


Figure 5.1: 1000 samples time series for task4

Then, we complete an embedding on our signals (let us bear in mind that we use 1000 samples which is approximately worth 8 seconds of EEG recording), we can apply the fast fixed point algorithm.

We chose to compute as many independent components as we have delay vectors. As a matter of fact, we do not know how many sources such a time series is constituted of. We do not want to risk losing any information. Hence, at the end of our computation, which is quite long, we obtain 30 independent components. We project each of our delay vectors on the corresponding independent component in order to obtain the sources we

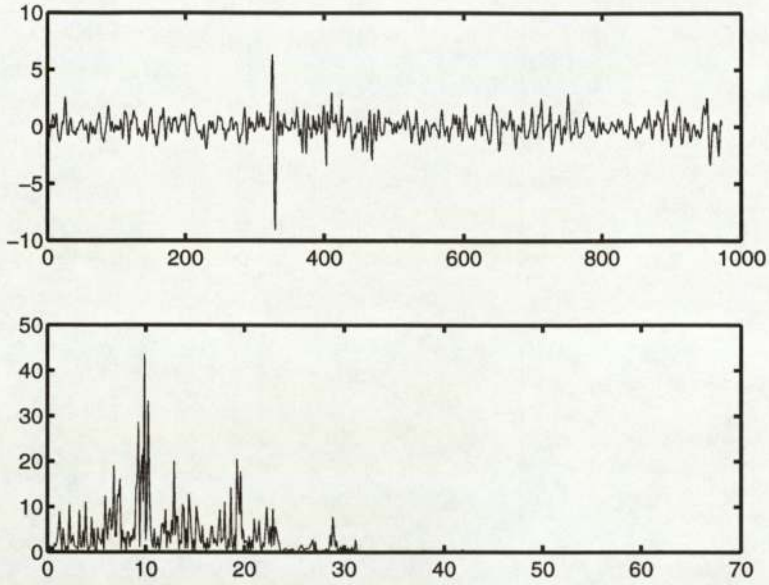


Figure 5.4: 9th source of the raw data and its power spectral density

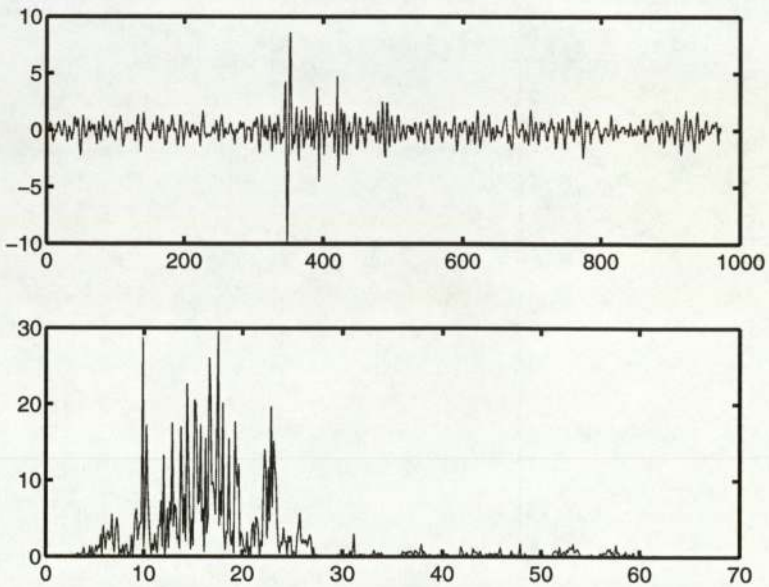


Figure 5.5: 8th source of the raw data and its power spectral density

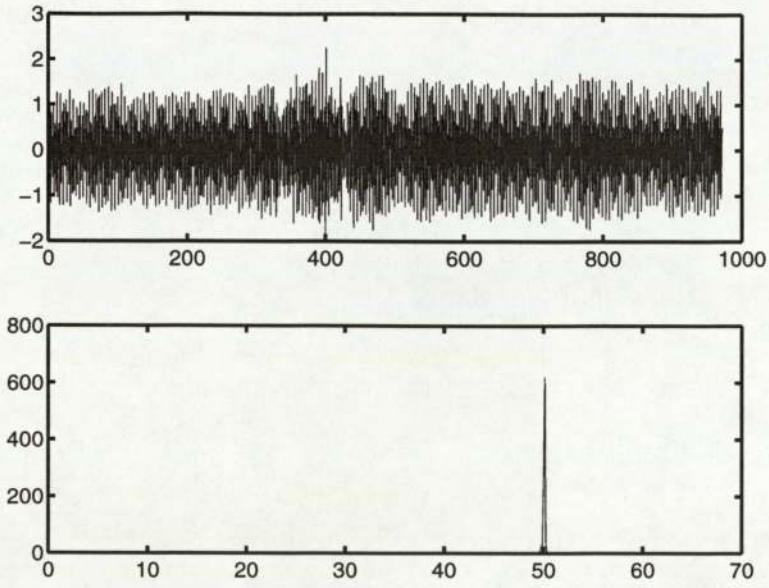


Figure 5.8: 10th source of the raw data and its power spectral density

As none of our resulting sources are ordered, we had to choose a way to classify them. Using the power spectrum of each of them proved to be a good method. As a matter of fact, it seems that the sources showing the brain activity during the different tasks, display different frequency range. Of course, some of them are identical or fall very closely in the same frequency interval. We can then advance that they naturally *cluster* by order of frequency.

Moreover, knowing the approximate frequency of each of the sources and their shape allows us to be able to relate them to an activity of the brain. Let us study each preceding source separately :

- source 16 (figure 5.2) has a very low frequency at about 5 Hz and its morphology can make us think that the ICA has isolated some eye movements.
- source 20 (figure 5.3) shows us the general trend of the signal but if we analyse it more closely we see a frequency range from 5 Hz to 10 Hz which characterizes alpha activity.
- source 9 (figure 5.4) shows high alpha activity and beta activity mixed.
- source 8's morphology (see figure 5.5) displays more spikes and the frequency

chosen to be reasonable (in combination with a reasonable necessary compromise on the shape of the rectangle). The Butterworth filter provides the maximum flatness in the passband (no ripples) which implies the minimum amplitude distortion and is therefore suited for our situation.

They are causal and of various orders, the lowest order being best (shortest) in the time domain, and the higher orders being better in the frequency domain. Well-engineered projects often include Butterworth filters.

Our need is to design a lowpass filter that loses no more than 3 dB in the passband and has at least 50 dB in the stopband because we assume that there is no more human activity after 50 Hz. 45 Hz and 50 Hz are the passband and stopband edge frequencies and the sampling frequency is 128 Hz. The estimated order of the filter is 7. The order was estimated using MATLAB function **butterord**. Figure 5.9 shows the magnitude of the transfer function of our designed filter. The filter was designed using MATLAB function **butter**.

Let $\frac{B(z)}{A(z)}$ denote the transfer function of the Nth-order digital filter. Then, by computing the z-transform of the digital filter, we have :

$$H(e^{jw}) = \frac{B(z)}{A(z)} = \frac{b(1) + b(2)z^{-1} + \cdots + b(n_b + 1)z^{-n_b}}{1 + a(2)z^{-1} + \cdots + a(n_a + 1)z^{-n_a}} \tag{5.1}$$

The vector w is a L-point frequency vector in radians, and H is the L-point complex frequency response vector of the filter $\frac{B}{A}$ given numerator and denominator coefficients in vectors B and A.

Figure 5.10 shows 1000 samples of EEG data after the application of our Butterworth filter. We can notice that the signal is smoother after the filtering of the data compared to figure 5.1.

5.4.2 Results

We apply the Butterworth filter to the raw EEG data, and then construct the embedding matrix of window size 30 with the resulting signal. We then perform the Independent Component Analysis on this matrix like we did previously. We notice that the convergence is faster than for the raw data.

As a result of the Independent Component Analysis, we get a set a 30 independent components on which we project the delay vectors and perform a spectral analysis of the sources. After this analysis, we notice that most of them are very similar to the previous ones we studied in the previous section.

Here is an example of signal we obtain. It is very similar to the source displayed on figure 5.2

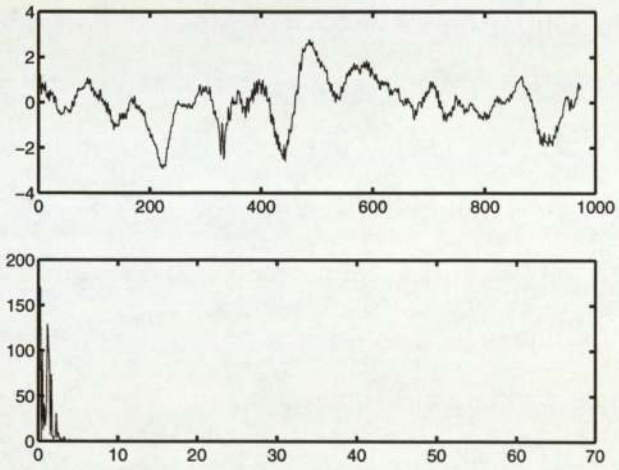


Figure 5.11: 10th source of the filtered data and its power spectral density

The sources are much "cleaner" with filtered data. We got rid of the noise from the fluorescent lamp, and of the extraneous noise above 40 Hz.

the original signals. In appendices A, B, C and D, we display the whole 30 sources and their *power spectral densities* resulting from our experiments on 1000 samples of our data for each task. We notice that after removing the noisy sources (very low or very high frequencies on the power spectral densities) and the redundant sources (same sources with different scaling, inverted sources), we can classify our sources by order of frequencies and that clusters naturally form from there. This allows us to verify our first hypothesis (see Chapter 3 section 3.4) : we can really identify four, five or six clusters in the interesting output sources which carry information in the signal domain.

One question remains : *How to analyse the behavioral significance of such sources ?* We must not forget the first aim of this study : can we simply characterize *vigilance* during the EEG trials just by extracting interesting brain activities from the output sources of the ICA. Unfortunately, not yet. Specialists only could give further explanations of the results, and some more experiments should be conducted on particular samples of the original EEG signals and compare the results with the "apparent" vigilance of the subjects during the trials.

competitive and unsupervised, meaning that no teacher is needed to define the correct output (that is to say the cell into which the input is mapped) for an input. The locations of the responses in the array tend to become ordered in the learning process as if some meaningful nonlinear coordinate system for the different input features were being created over the network.

The SOM was developed by Professor Teuvo Kohonen in the early 1980s [18].

6.2.1 The Self-Organizing Map Algorithm

Assume that the sample data sets have to be mapped onto the array depicted in figure 6.1. The set of input samples is described by a real vector $x(t) \in \mathbf{R}^n$ where t is the index of the sample, or the discrete time coordinate. Each node i in the map contains a model vector $m_i(t) \in \mathbf{R}^n$, which has the same number of elements as the input vector $x(t)$.

The stochastic SOM algorithm performs a regression process. Therefore, the initial values of the components of the model vector, $m_i(t)$, may even be selected at random.

Any input item is thought to be mapped into the location, the $m_i(t)$ of which matches the best with $x(t)$ in some metric. The self-organizing algorithm creates the ordered mapping as a repetition of the following basic task :

1. An input vector $x(t)$ is compared with all the model vectors $m_i(t)$. The best-matching unit (node) on the map, i.e., the node where the model vector is most similar to the input vector in some metric (e.g. euclidean) is identified. This best matching unit is called the winner.

2. The model vectors of the winner and a number of its neighboring nodes in the array are changed towards the input vector according to the learning principle specified below.

The basic idea of the SOM learning process is that, for each sample input vector $x(t)$, the winner and the nodes in the neighbourhood are changed closer to $x(t)$ in the input

data space. During the learning process, individual changings may be contradictory, but the net outcome in the process is that ordered values for the $m_i(t)$ emerge over the array.

Adaptation of the model vectors in the learning process may take place according to the following equations :

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)[x(t) - m_i(t)] & \text{for each } i \in N_c(t) \\ m_i(t) & \text{otherwise} \end{cases}$$

where t is the discrete-time index of the variables, the factor $\alpha(t) \in [0, 1]$ is a scalar that defines the relative size of the learning step, and $N_c(t)$ specifies the *neighbourhood* around the winner in the map array.

At the beginning of the learning process, the radius of the neighbourhood is quite large, but it is made to shrink during the learning. This ensures that the global order is obtained already at the beginning, whereas towards the end, the radius gets smaller, the local corrections of the model vector in the map will be more specific. The factor $\alpha(t)$ also decreases during the training.

One method of evaluating the quality of the resulting map is to calculate the average quantization error over the input samples, defined as $E\{\|x - m_c(x)\|\}$ where c indicates the best matching unit for x. After training, for each input sample vector, the best-matching unit in the map is searched for, and the average of the respective quantization errors is returned.

6.2.2 Experiments and Results

The SOM_PAK program, developed by Kohonen, Hynninen, Kangas and Laaksonen, was used during all the course of these experiments [18].

The Self-Organizing Map was applied to the recognition of topographic patterns on the resulting sources of our former Blind Source Separation (see Chapter 5).

The training set consists of 24 sources of 500 samples each. We do not use any labelling at all to characterize the sources on this set. For each task, we obtain a

square map and by stronger patterns in its top left and bottom right corner. Whereas boring tracking tasks are characterized by lighter patterns in the upper left side and brighter patterns in the bottom left of the upper right corner.

The main problem is that, in that way, we do not determine the kind of brain activity these lighter and brighter patterns correspond to.

6.3 Sammon Mapping

6.3.1 The Algorithm

We said in the introductive section that in a topographic map, N data vectors $\{x_i\}$ in $\mathbf{R}^p$ are transformed into a corresponding set of feature vectors $\{y_i\}$ in $\mathbf{R}^q$ such that $q < p$ and the *geometric structure* of the input data vector remains unchanged. The *Sammon Mapping* is the most intuitive basis for this definition since it is generated by the minimization of an error measure E of the inter-point distances also called STRESS

$$E = \frac{1}{\sum_i \sum_{j<i} d^*_{ij}} \times \frac{\sum_i \sum_{j<i} (d^*_{ij} - d_{ij})^2}{d^*_{ij}} \tag{6.1}$$

where $d^*_{ij} = \|x_i - x_j\|$ is the distance between points i, j in the input data set and $d_{ij} = \|y_i - y_j\|$ is the distance between their images in the map or feature space.

The procedure for performing the transformation is shown in figure 6.3 and summarised in Table 6.1.

Various error minimisation procedures can be used, one of which is the gradient descent procedure. But this procedure can get trapped in local minima.

1. Compute inter-point distances in the original space.
2. Initialise target space by a random number generator.
3. Calculate mapping error between original and target space.
4. Modify coordinate points in the target space by means of a non-linear procedure.
5. Repeat Step 3 until the mapping error is sufficiently small.

Table 6.1: Sammon Mapping's algorithm

Let us now come back to equation (6.1). The $d^*_{ij} - d_{ij}$ term represents a measure of the deviation between the corresponding distances. The Sammon STRESS thus represents an optimal matching of the inter-point distances in the input and map spaces. According to [24], normalising the expression by the first fractional term reduces the sensitivity of the measure to the number of input points and their scaling. Moreover, to render the overall measure dimensionless, the d_{ij} term is included in the denominator of the sum to moderate the domination of errors in large distances over those in smaller distances.

In the standard Sammon Mapping, the STRESS is minimised by adjusting the location of the points y_i directly, according to a gradient-descent scheme. For each point y_i, we define a parameterised non-linear function of the input $f(x_i; w)$, where w is the weight vector. Then the STRESS becomes :

$$E = \sum_{i}^{N} \sum_{j}^{N} (d^*_{ij} - \|f(x_i; w) - f(x_j; w)\|)^2 \tag{6.2}$$

Then, it is straightforward to differentiate E with respect to the mapped coordinates y_i and optimise the map using standard error-minimisation methods. This gives :

$$\begin{aligned} \frac{\partial E}{\partial w_k} &= \sum_{i}^{N} \frac{\partial E}{\partial y_i} \cdot \frac{\partial y_i}{\partial w_k} \\ &= \sum_{i}^{N} \frac{\partial E}{\partial y_i} \cdot \frac{\partial f(x_i; w_k)}{\partial w_k} \end{aligned} \tag{6.3}$$

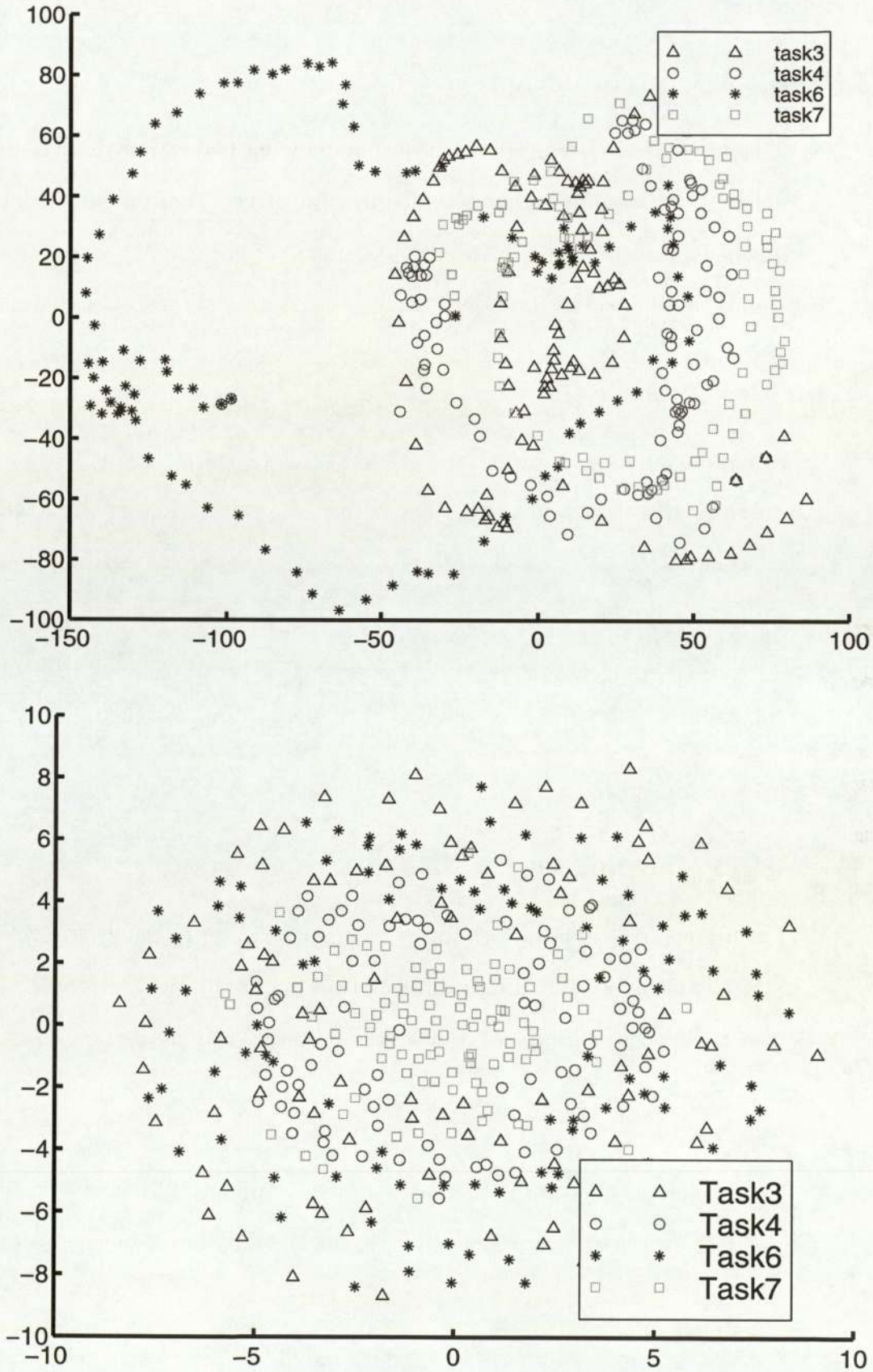


Figure 6.4: Sammon Maps of a projection of our EEG data on the principal components (top) and on the independent components (bottom)

The constant term k is added to the inter-point distances of two pairs of points so that their separation should be exaggerated in the resultant map.

In many problems, there may be further knowledge available regarding class relationships which we call *subjective dissimilarity* and will denote in the following section $S = [s_{ij}]$ corresponding to each d^*_{ij}. The assignment of class dissimilarity means that for every pair of data points in addition to the *objective dissimilarity*, there is a dual *subjective dissimilarity* which stresses alternative knowledge about the data. Thus, one can relate the existence of this set of subjective dissimilarities to a *subjective metric* implicitely defined over the input space.

6.4.3 NeuroScale

NEUROSCALE is a technique which transforms a p-dimensional input space into a q-dimensional feature space ($q < p$) with a feed-forward radial basis function. The network is trained with the same algorithm as with a Sammon Mapping (see Table 6.1) but by minimizing the following stress measure :

$$E = \sum_{i}^{N} \sum_{j<i}^{N} (\delta_{ij} - \|y_i - y_j\|)^2 \tag{6.7}$$

where

$$\delta_{ij} = (1 - \alpha) d^*_{ij} + \alpha s_{ij} \tag{6.8}$$

The parameter α ($0 \leq \alpha \leq 1$) controls the degree to which the subjective metric S influences the output configuration. One can say that it helps finding a nice middle between an unsupervised and a supervised mapping.

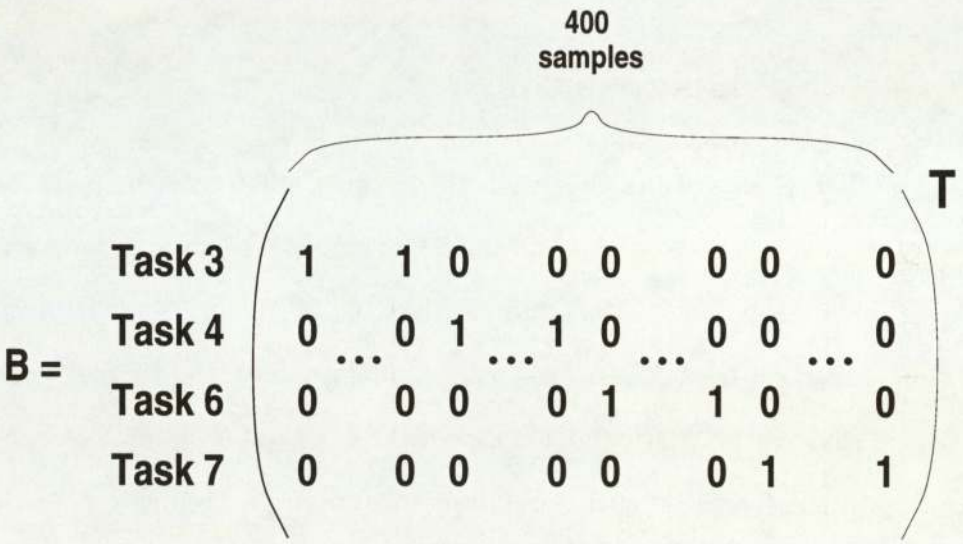
The main difference between the algorithm of NEUROSCALE and the Sammon Mapping algorithm (Table 6.1) is that this latter is fixed, *i.e.*, we know when it has converged. We must alter it in order to : include a calculation of the elements of the input space distance matrix, take into account the particular value of α. If α equals 0, then the algorithm computes a parameterized Sammon Mapping. If α equals 1,

6.4.4 Experiments and Results

To minimize E, a gradient descent algorithm is employed. The network weights are initialised at random.

The data comprises 100 samples for each task taken from the same sets we used for performing the ICA on which we do an embedding in order to look for the independent components. After computing the ICA on the embedded matrix, we obtain a 400×30 matrix of independent sources on which we can perform the NEUROSCALE algorithm.

We also choose a subjective metric which only takes into account the task knowledge relative to the data. So we build a boolean matrix B corresponding to each task in the following way :



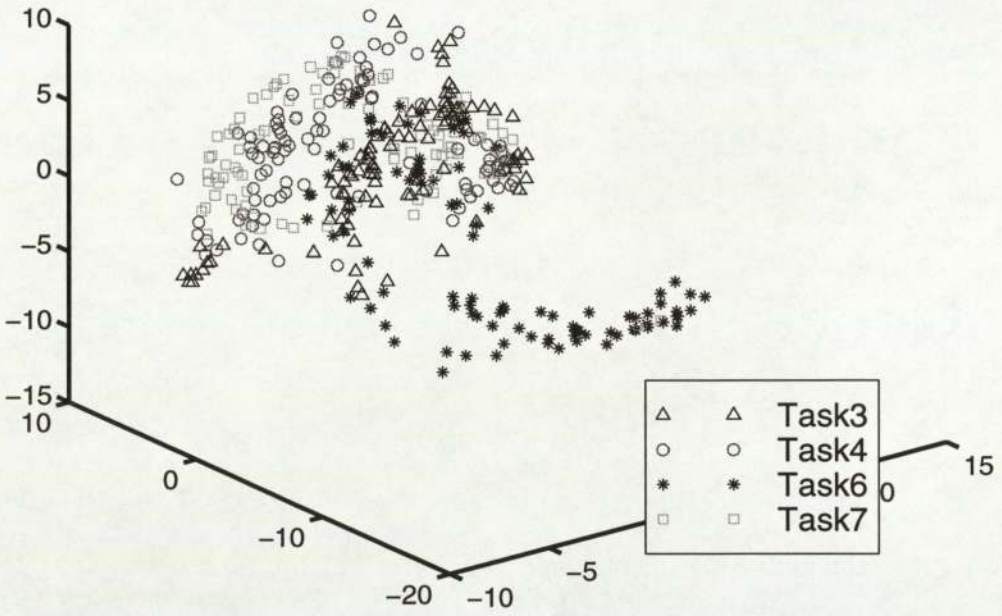
Then, to compute our subjective matrix S, we just have to compute the inter-point euclidean distances of the matrix B.

With $\alpha = 0.5$, we both retain some of the objective spatial topology and impose some task knowledge to the output configuration.

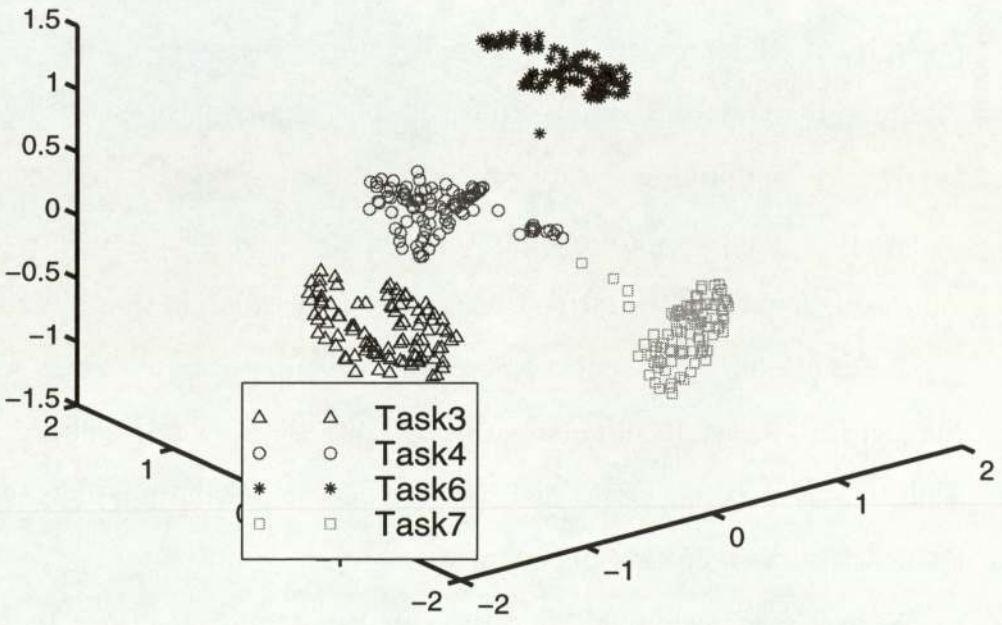
It is interesting to first observe the differences between these two maps and the Sammon Maps (see figure 6.4). We can see that the NEUROSCALE algorithm has split the clusters up even more to reveal some class knowledge. On the projections on the principal components, the algorithm has kept the ordering of the Sammon Mapping,

spiral shape of the principal components). But we are still unable to identify their meaning.

NEUROSCALE displays some advantageous features which the Sammon Mapping does not. By incorporating varying degrees of subjective knowledge, we can influence the extracted feature space. The resulting sphere, with its four distinct clusters and their evolution with the changing of the parameter α gives us a better idea of clustering than the Sammon Mapping. And of course, the NEUROSCALE map is far more representative than the Kohonen SOM. Moreover, there is a cost function associated with a particular mapping which allows us to assess individual maps and to compare them.



(a)



(b)

Figure 6.7: NEUROSCALE of a projection of our EEG data on the principal components (top) and on the independent components (bottom) with $\alpha = 0.9$. The value of the STRESS measure was : 2.98 for case a), 0.13 for case b)

the mixture of sources and remove them in order to extract the information-carrying sources that is the main brain activity corresponding to the behavioural state of the subject when achieving a particular task during a trial.

Moreover, these sources characterising brain activity, also provide justification for our main fundamental hypothesis that there are only a few of them constitutive of a nonlinear system with just a few degrees of freedom. Unfortunately, we are not yet able to extract these interesting sources directly without performing a power spectral analysis on each of them separately to isolate them from the noisy ones as there is yet no existing algorithm to classify these sources.

In Chapter 6, it was reasoned that the Sammon Mapping and NEUROSCALE were most effective strategies for topographic dimension reduction. The main advantage of Kohonen's approach is computational, because realistically, application of the Sammon Mapping is restricted to fewer than 1000 data points.

The feed-forward neural network topographic mapping technique NEUROSCALE , was thus based upon the Sammon Mapping and utilises a radial basis function neural network. Because of this neural network element, it offers the capability of generalisation to new data — a feature absent from Sammon's original algorithm.

An important extension embodied in NEUROSCALE is the capacity to exploit additional information in the mapping process. In standard approaches to topographic mapping, the geometry of the output space is determined solely according to some conventional metric (generally Euclidean) defined over the data space. If alternative information is available — such as class labels — then this may be allowed to influence the mapping (in order to emphasise clustering, for example).

The results shown by both Sammon Mapping and NEUROSCALE are interesting in a way that they show some clustering according to the types of task, but they also prove that there is another important clustering that we are not yet able to interpret and which is not related to the time history of the four tasks. One can suggest that this particular feature of the map comes from the "way" the subject has undertaken

Bibliography

[1] Bell A and Sejnowski T. "An Information-Maximisation Approach to Blind Separation and Blind Deconvolution". *Neural Computation*, 7:1004–1034, 1995.

[2] Cichocki A, Kasprzak W, and Amari S. "Multi-layer Neural Networks with a Local Adaptative Learning Rule for Blind Separation of Sources". *Proceedings of the Research society of Nonlinear Theory and its Aplications (NOLTA)*, 1995.

[3] Cichocki A, Kasprzak W, and Amari S. "Robust Neural Networks with Online Learning for Blind Identification and Blind Separation of Sources". *IEEE Transacions on Circuits and Systems – Fundamental Theory and Applications*, 43, 1996.

[4] Hyvärinen A. "A Family of Fixed-Point Algorithms for Independent Component Analysis". *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97)*, pages 3917–3920, 1997.

[5] Hyvärinen A. "Independent Component Analysis by Minimization of Mutual Information". *Technical Report, Helsinki University of Technology, Laboratory of Computer Science*, 1997.

[6] Hyvärinen A and Oja E. "A Fast Fixed-point Algorithm for Independent Component Analysis". *Neural Computation*, 9:1483–1492, 1997.

[7] Loomis AL, Harvey E, and Hobart GA. "Cerebral States during Human Sleep as studied by Human Brain Potentials". *Journal of Experimental Psychology*, 21, 1937.

[19] Karhunen J, Oja E, Wang L, Vigrio R, and Joutsensalo J. "A Class of Neural Networks for Independent Component Analysis". *IEEE Transactions on Neural Networks*, 1997.

[20] Karhunen J and Pajunen P. "Hierarchic Nonlinear PCA Algorithms for Neural Blind Source Separation". *Proceeding of the IEEE Nordic Signal Processing Symposium (NORSIG)*, 1996.

[21] Karhunen J and Pajunen P. "Blind Source Separation and Tracking using Nonlinear PCA Criterion : A Least-squares Approach". *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, 1997.

[22] Mao J and Jain AK. "Artificial Neural Networks for Feature Extraction and Multivariate Observations". *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability*, 1995.

[23] Cardoso JF and Souloumiac A. "Blind Beaforming for Non Gaussian Signals". *IEE-Proceedings*, 140:362–370, 1993.

[24] Sammon JW. "A Nonlinear Mapping for Data Structure Analysis". *IEEE Transactions on Computers*, C–18(5):401–409, 1989.

[25] Knuth KH. "Difficulties Applying Recent Blind Source Separation Techniques to EEG and MEG". *Maximum Entropy and Bayesian Methods*, 1997.

[26] Girolami M and Fyfe C. "Blind Separation of Sources using Exploratory Projection Pursuit". *Proceedings of the International Conference on the Engineering Applications of Neural Networks (EANN)*, 1996.

[27] Tipping ME. *Topographic Mappings and Feed-Forward Neural Networks.* PhD thesis, 1996.

[28] Delfosse N and Loubaton P. "Adaptative Blind Source Separation of Independent Sources : A Deflation Approach". *Signal Processing*, 45, 1995.

[39] Jung T-P, Makeig S, and Sejnowski T. "Using Feedforward Neural Networks to Monitor Alertness from Changes in EEG Correlation and Coherence". In *Advances in Neural Information Processing Systems 8.*, 1996.

[40] Li X, Gasteiger J, and Zupan J. "On the Topology Distorsion in Self-Organizing Feature Maps". *Biological Cybernetics*, 70, 1993.

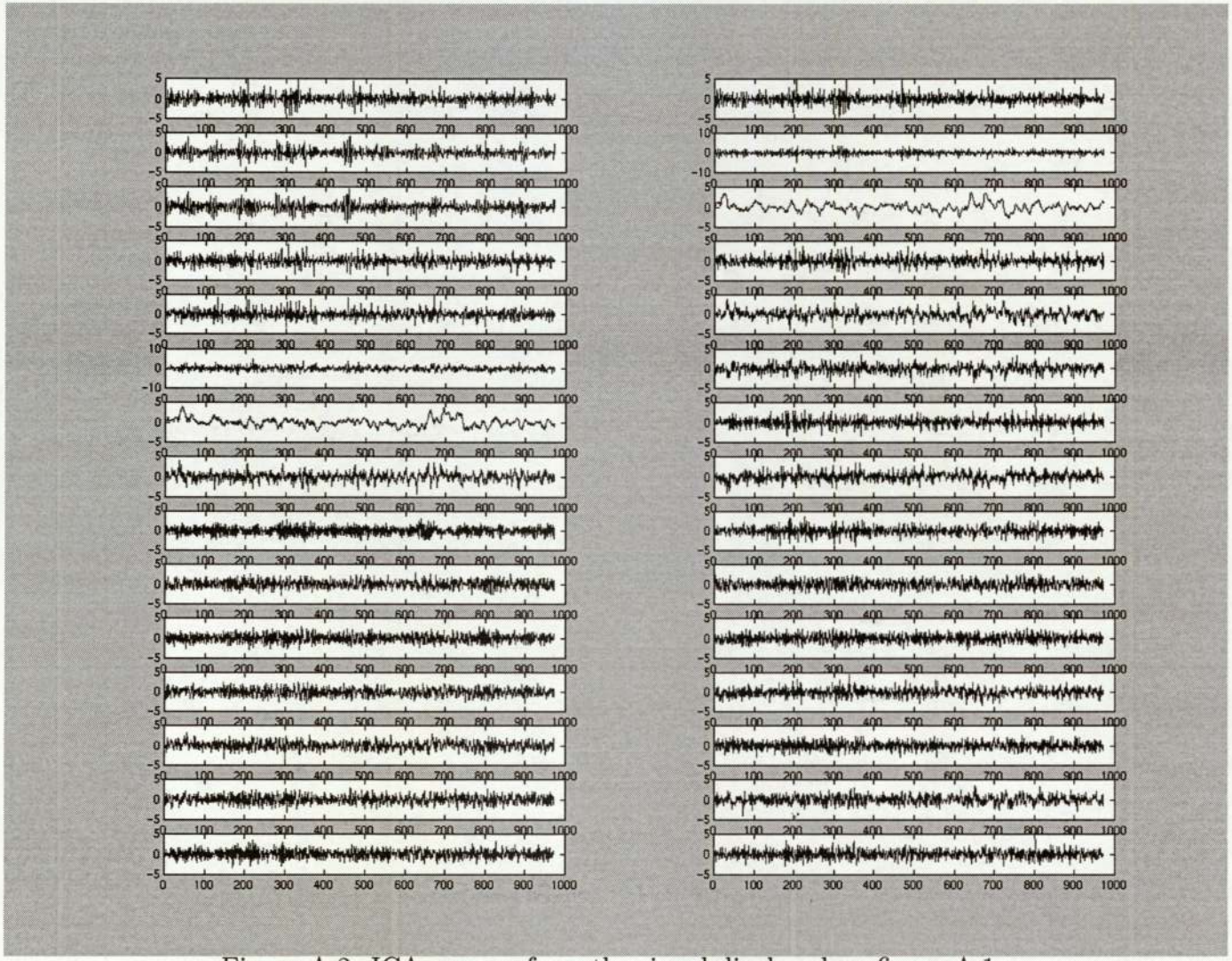


Figure A.2: ICA sources from the signal displayed on figure A.1

Appendix B

Results for task4

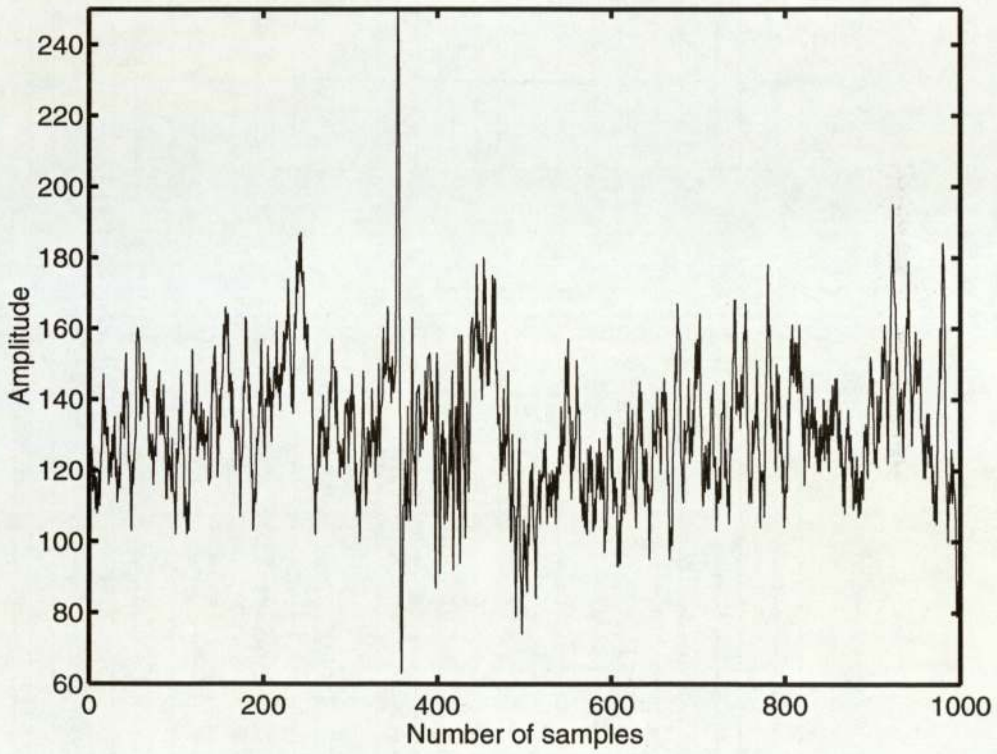


Figure B.1: 1000 samples taken from Task4 on which we apply the ICA algorithm

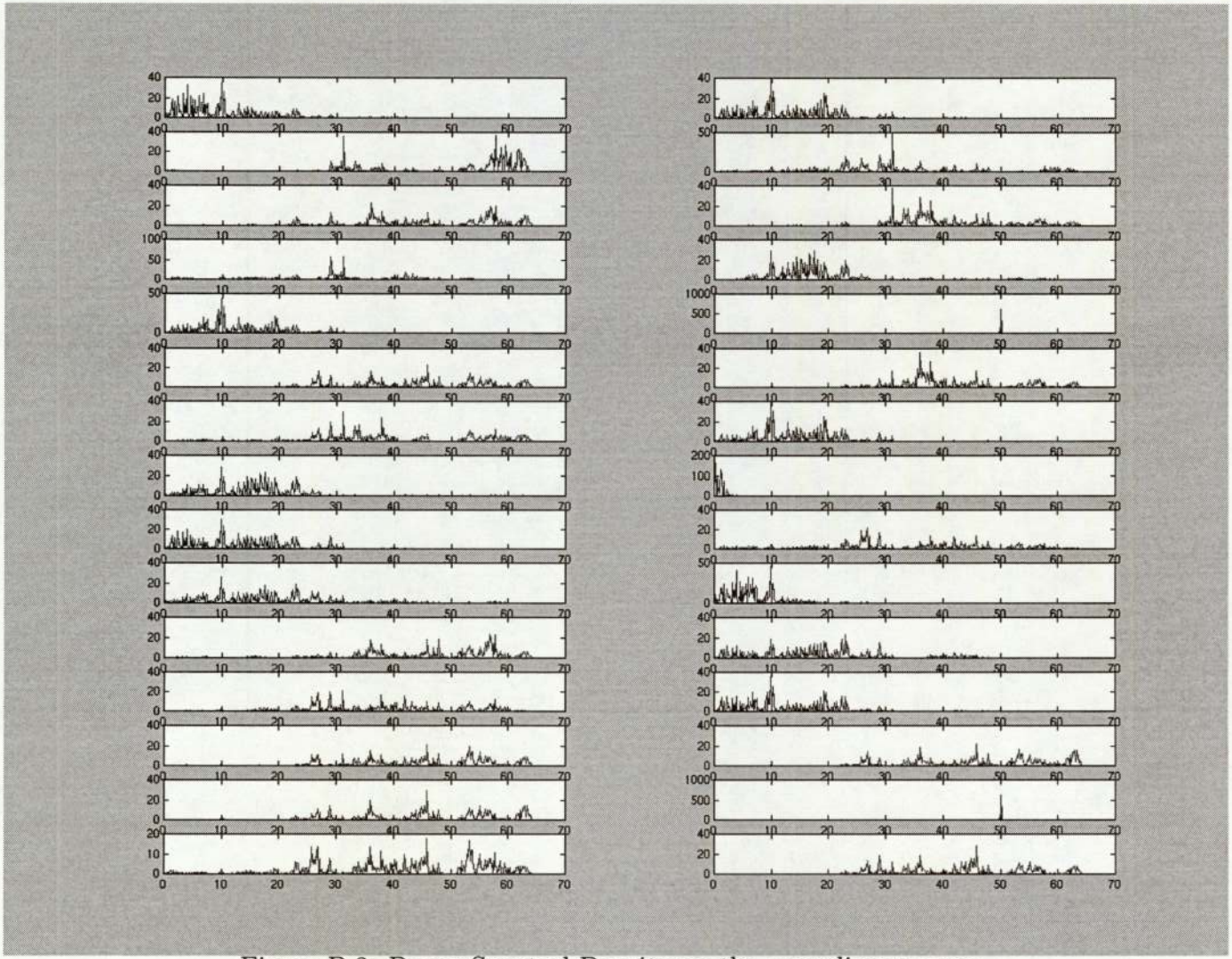


Figure B.3: Power Spectral Density on the preceding sources

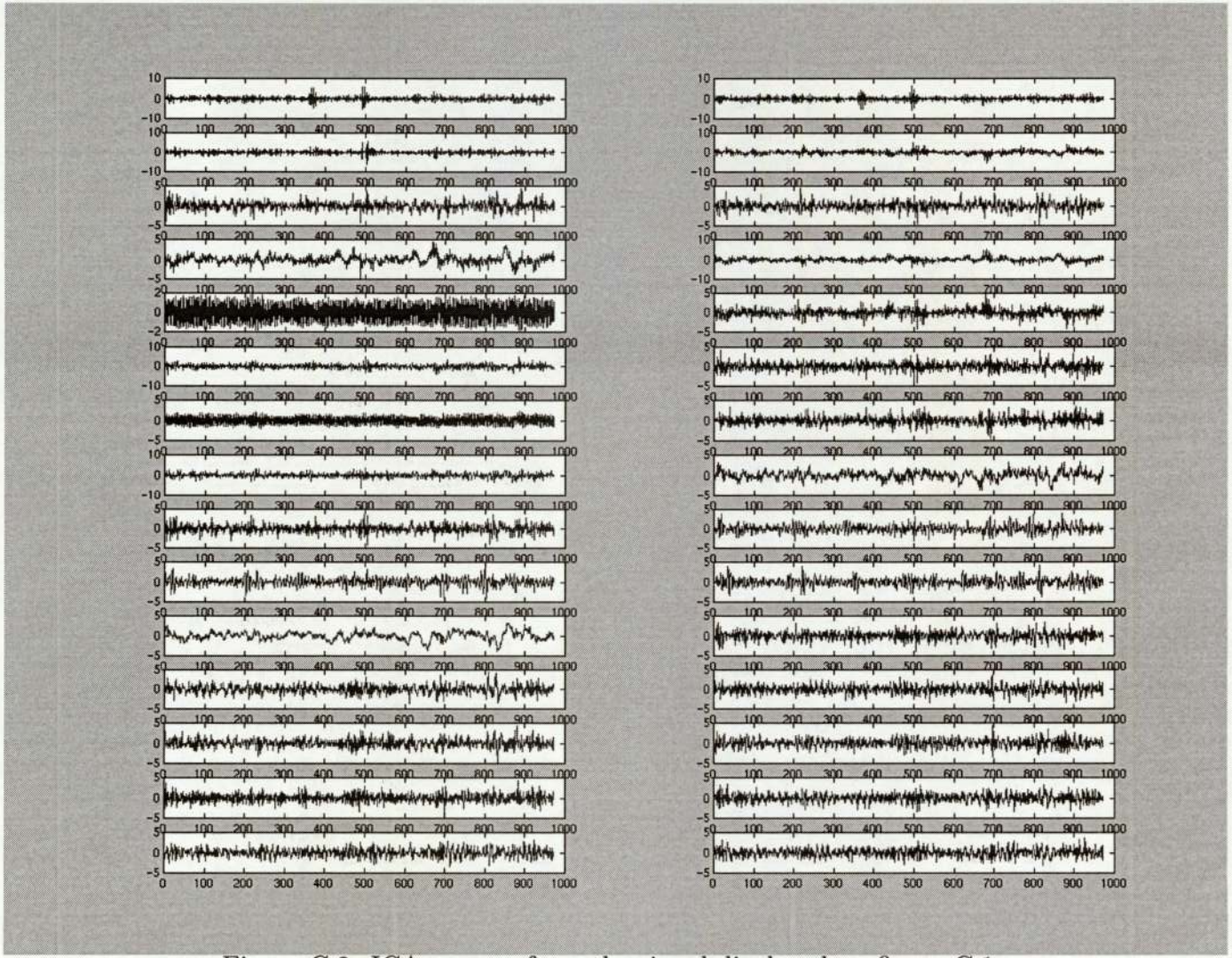


Figure C.2: ICA sources from the signal displayed on figure C.1

Appendix D

Results for task7

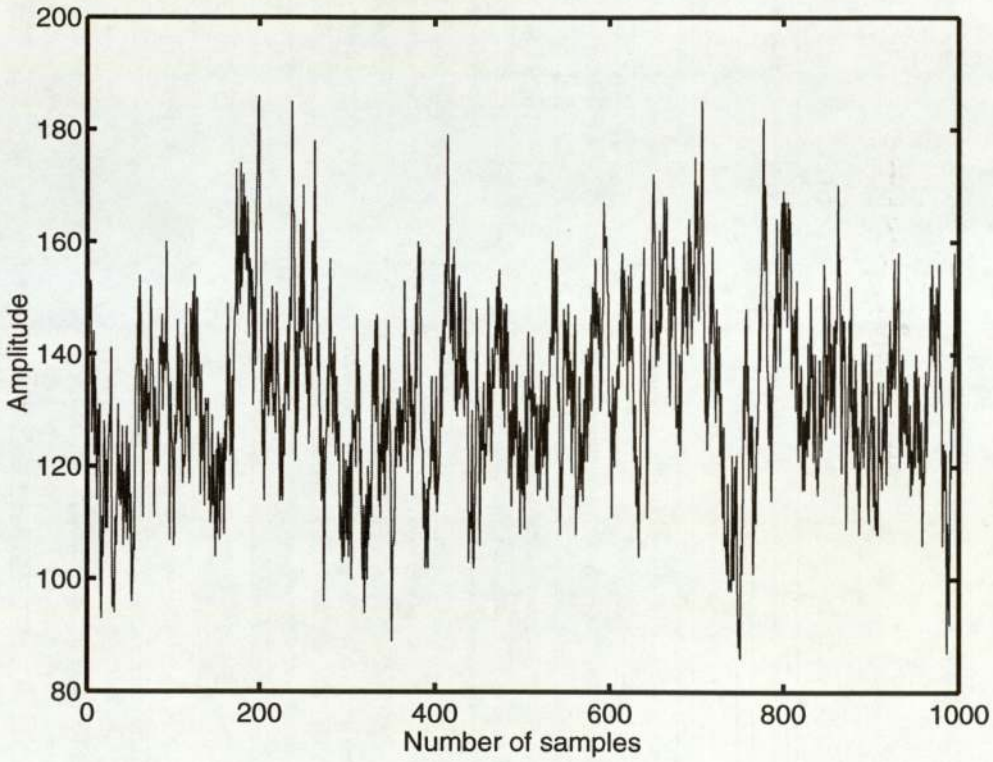


Figure D.1: 1000 samples taken from Task7 on which we apply the ICA algorithm

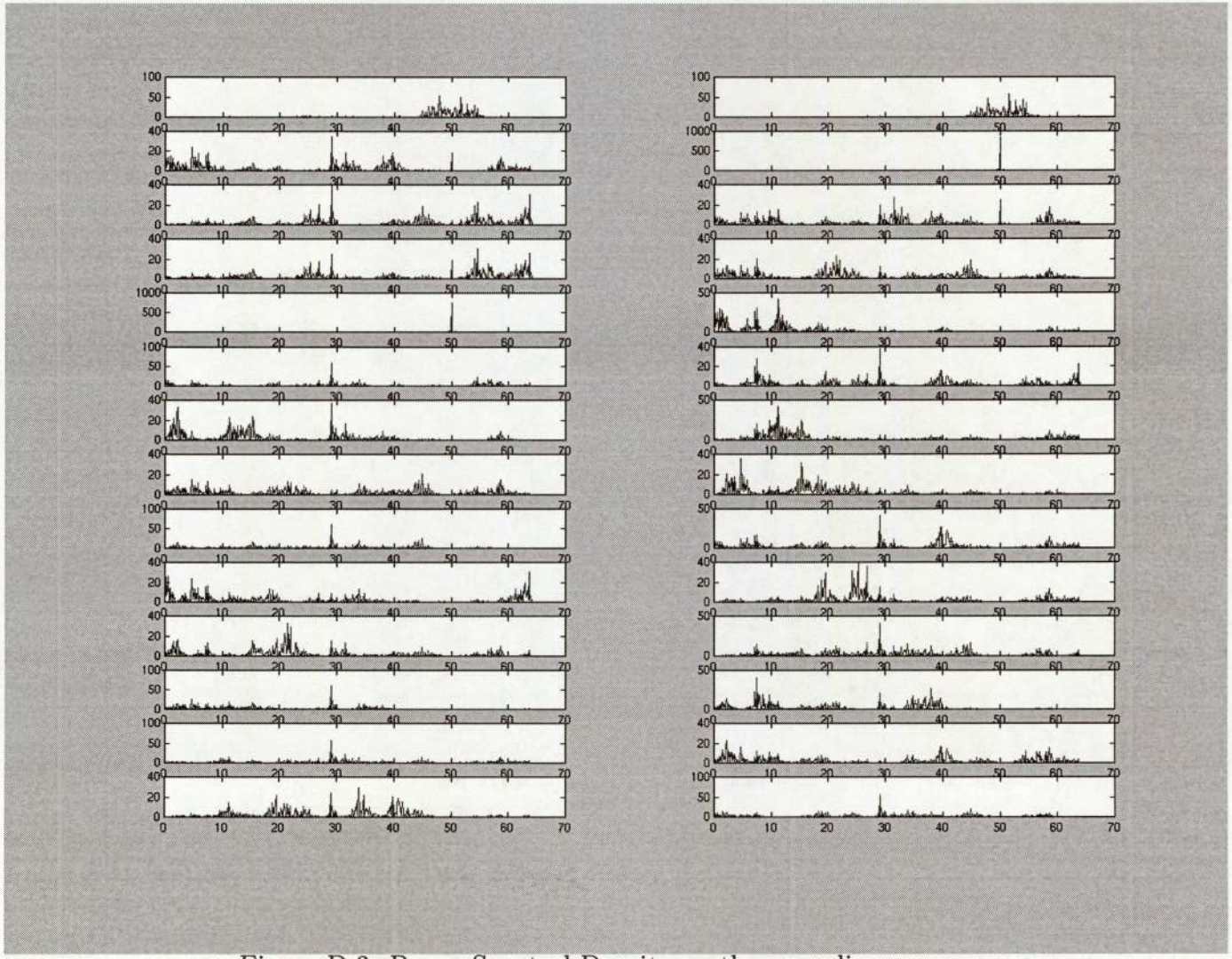


Figure D.3: Power Spectral Density on the preceding sources