DYNAMICAL FACTOR ANALYSIS OF RHYTHMIC MAGNETOENCEPHALOGRAPHIC ACTIVITY

Jaakko Särelä[1], *Harri Valpola*[1], *Ricardo Vigário*[1,2], *and Erkki Oja*[1]

[1]Helsinki University of Technology
Neural Network Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland

[2]GMD - FIRST
Kekuléstr. 7
D - 12489 Berlin, Germany

{Jaakko.Sarela, Harri.Valpola, Ricardo.Vigario, Erkki.Oja}@hut.fi

ABSTRACT

Dynamical factor analysis (DFA) is a generative dynamical algorithm, with linear mapping from factors to the observations and nonlinear mapping of the factor dynamics. The latter is modeled by a multilayer perceptron. Ensemble learning is used to estimate the DFA model in an unsupervised manner. The performance of the DFA have been tested in a set of artificially generated noisy modulated sinusoids. Furthermore, we have applied it to magnetoencephalographic data containing bursts of oscillatory brain activity. This paper shows that DFA can correctly estimate the underlying factors in both data sets.

1. INTRODUCTION

Recent advances in blind source separation (BSS) have provided new and powerful algorithms for the analysis of electro- and magnetoencephalographic signals (EEG and MEG respectively). For a selection of reviews in this field see [1, 2]. In spite of the very useful results obtained using classic BSS approaches, it is often clear that these algorithms fail to fully model the underlying signals. For example, independent component analysis (ICA, [3]) assumes the signals to be random samples from non-Gaussian distributions, which is not a very plausible signal model for time-series with time-structure. On the other hand, algorithms taking implicitly the dynamics into account (see, *e.g.*, [4]), do not provide an explicit generative model for the observed data. Without a generative model, it is difficult to reliably estimate the observation noise. This is a serious shortcoming regarding MEG applications, where the signal-to-noise ratio can be extremely poor.

In this paper we introduce a generative dynamical algorithm for noisy measurements. This algorithm is dynamical factor analysis (DFA), and it exploits a Bayesian treatment called ensemble learning [5, 6]. Ensemble learning provides a general framework to learn generative models from a given data set and it can be used for model selection, *e.g.*, in the determination of the most probable number of underlying factors to the observations. It also provides uncertainties for the learned parameters thus automatically performing regularisation.

Here we will apply DFA to learn factors from both artificially generated signals, consisting of mixtures of modulated sinusoids, and MEG measurements containing bursts of rhythmic activity.

Cortical electromagnetic rhythms have been observed and studied since the early EEG and MEG recordings. It is believed that spontaneous brain rhythms are mainly associated with a cortical resting state, and thought to respond quicker to incoming signals than a silent system would. The spectral content and reactivity of spontaneous brain rhythms are affected, *e.g.*, by vigilance, several brain disorders, development and ageing. Oscillatory brain activity thus gives an overall view of brain function and is therefore routinely monitored in clinical EEG recordings.

A typical way to characterize brain rhythms is through their respective frequency bands. The most common rhythm, present mainly over the parieto-occipital and occipital cortex, has frequencies in the interval 8–13 Hz, and is labeled as α-rhythm. Oscillations within the interval 14–30 Hz are often labeled as β-rhythms. For a more comprehensive discussion regarding EEG and MEG, and their spontaneous rhythms see, *e.g.*, [7, 8].

From a signal processing viewpoint, in particular when applying ICA algorithms in BSS, brain rhythmic activity constitutes a great challenge. Due to their typical burst-like nature, rhythmic activities may present very weak high order moments, rendering them indiscernible from Gaussian processes. These, as we know, would then be very hard to separate, unless additional information is present [9, 10]. Such information should make use of the intrinsic temporal dynamics present in each rhythm.

This work is partially funded by EU BLISS project. RV is funded by EU (Marie Curie Fellowship HPMF-CT-2000-00813)

2. MODEL

The dynamical factor analysis model is a very general model of complex dynamical processes. Observations $\mathbf{x}(t)$ are assumed to be generated by linear mixing $\mathbf{A}$ from hidden states $\mathbf{s}(t)$ including Gaussian white additive noise $\mathbf{n}(t)$. Additionally each state $\mathbf{s}(t)$ for all t is generated from the previous states $\mathbf{s}(t-1)$ by a nonlinear mapping $\mathbf{f}$ with a Gaussian innovation process $\mathbf{m}(t)$. Mappings $\mathbf{A}$ and $\mathbf{f}$ are assumed to be independent of time. Thus we have a two part model:

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \\ \mathbf{s}(t) &= \mathbf{f}(\mathbf{s}(t-1)) + \mathbf{m}(t). \end{aligned} \tag{1}$$

The nonlinear mapping $\mathbf{f}$ is modelled by a two-layer MLP network [11] with sigmoidal tanh's as the hidden layer nonlinearities. This gives the mapping

$$\mathbf{f}(\mathbf{s}) = \mathbf{s} + \mathbf{C}\tanh(\mathbf{B}\mathbf{s} + \mathbf{b}) + \mathbf{c} \tag{2}$$

Note that only the change in the states is modelled by the MLP network. Figure 1 illustrates the model graphically. Dark circles correspond to the observations $\mathbf{x}(t)$, empty circles are the states $\mathbf{s}(t)$ and the circles with sigmoids inside represent the hidden units of the MLP network.

Notice that there is only a single delay in (2). When time-series are predicted directly in the observation space, *i.e.* auto-regressive models are used, (see *e.g.* [11]), it is common to use several time instances of history as the inputs. In state-space models a single delay suffices, because some of the states can represent the dynamics. For example a free projectile movement can be modelled using three previous positions of the moving object, but it can also be modelled using the position, the speed and the acceleration of the object as the states.

2.1. Dynamics in blocks

Factor analysis defines the mapping up to a rotation. This means that the learned states can be mixtures of each others, though they are not correlated [12]. The dynamical mapping defines the rotation, but it is very slow to learn, if the MLP network is fully connected.

In MEG signal analysis it is often crucial to learn the rotation, since we are mainly interested in the characteristics of the states, not in prediction. For this reason the dynamics of the factors is forced to be block-wise (see Fig. 1), which simplifies the network and encourages the model to find independent source processes. If the factors are modulated sinusoids as is the case in rhythmical activity, blocks of two factors suffice.

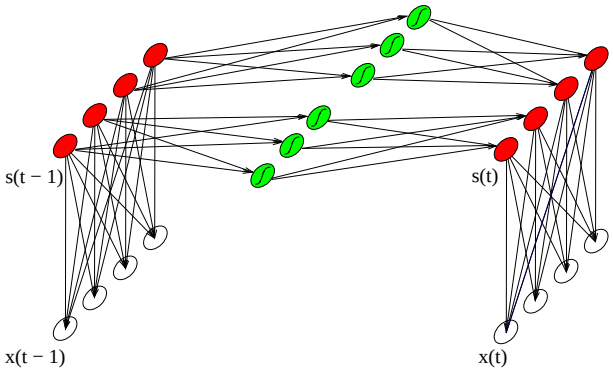

Fig. 1. *Part of the graphical model. Dark units are states, empty units the observations and the ones with a sigmoid inside correspond to the MLP dynamics. Direction of the arrows correspond to the direction of the causality (observations are caused by states and next states are caused by previous states).*

3. ENSEMBLE LEARNING

The aim of learning in Bayesian framework is to calculate the posterior density function $p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})$, where $\mathbf{X} = (\mathbf{x}(1), \ldots, \mathbf{x}(T))$, $\mathbf{S} = (\mathbf{s}(1), \ldots, \mathbf{s}(T))$ and $\boldsymbol{\theta}$ contains all the model parameters. The posterior is obtained from the Bayes' theorem: $p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X}|\mathbf{S}, \boldsymbol{\theta})p(\mathbf{S}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{X})$.

The likelihood of the observations given the model (1,2) can be written as:

$$\begin{aligned} p(\mathbf{X}|\mathbf{S}, \boldsymbol{\theta}) &= \prod_{i,t} p(x_i(t)|\mathbf{s}(t), \boldsymbol{\theta}) \\ &= \prod_{i,t} N\left(x_i(t);\ \mathbf{a}_i\mathbf{s}(t), \exp(2v_i)\right), \end{aligned} \tag{3}$$

where $N(x;\ \mu, \sigma^2)$ denotes a Gaussian distribution over x with mean μ and variance σ^2, $\mathbf{a}_i$ is the ith column vector of the mixing matrix $\mathbf{A}$ and v_i is a hyperparameter specifying the noise variance. The likelihood $p(\mathbf{S}|\boldsymbol{\theta})$ of the states $\mathbf{s}$ is specified similarly using the function $\mathbf{f}$ instead of the linear mapping $\mathbf{a}_i$. All the parameters of the model have hierarchical Gaussian priors. For example the noise parameters v_i of different components of the data share a common prior [13].

Ensemble learning [5, 6] is a recently developed method for fitting a parametric approximation to the exact posterior density function $p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})$. The true posterior is approximated by a density $q(\mathbf{S}, \boldsymbol{\theta})$ with a simple factorial form. The misfit of the approximation is measured by Kullback-Leibler divergence between the approximation and the true posterior:

$$D(q(\mathbf{S}, \boldsymbol{\theta})||p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})) = \mathrm{E}_{q(\mathbf{S}, \boldsymbol{\theta})}\left[\log \frac{q(\mathbf{S}, \boldsymbol{\theta})}{p(\mathbf{S}, \boldsymbol{\theta}|\mathbf{X})}\right]. \tag{4}$$

The posterior distribution can be written as $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{X}) = p(\boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{X})/p(\boldsymbol{X})$. The normalizing term $p(\boldsymbol{X})$ cannot usually be evaluated, and therefore the actual cost function used in ensemble learning is

$$C = \mathrm{E}\left[\log \frac{q(\boldsymbol{S}, \boldsymbol{\theta})}{p(\boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{X})}\right] = D(q(\boldsymbol{S}, \boldsymbol{\theta})||p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{X})) - \log p(\boldsymbol{X}) \geq -\log p(\boldsymbol{X}). \quad (5)$$

Ensemble learning has been previously used for nonlinear state-space models [14, 12], where both observation and dynamical mappings are nonlinear. This more general model can naturally learn linear mappings too, but we prefer linear model of observation since it is highly plausible in MEG [8].

3.1. Form of the posterior approximation

The cost function can be minimized efficiently if a suitably simple factorial form for the approximation is chosen. We use $q(\boldsymbol{\theta}, \boldsymbol{S}) = q(\boldsymbol{\theta})q(\boldsymbol{S})$, where $q(\boldsymbol{\theta}) = \prod_i q(\theta_i)$ is a product of univariate Gaussian distributions, *i.e.* the distribution for each parameter θ_i is parameterized with mean $\bar{\theta}_i$ and variance $\hat{\theta}_i$. These are the variational parameters of the distribution to be optimized.

The approximation $q(\mathbf{S})$ takes into account the posterior dependences between $s_i(t)$ at consecutive time instances. It has a form $q(\boldsymbol{S}) = \prod_i [q(s_i(1)) \prod_t q(s_i(t)|s_i(t-1))]$. The value $s_i(t)$ depends only on $s_i(t-1)$ at previous time instant, not on the other $s_j(t-1)$ with $j \neq i$. The distribution $q(s_i(t)|s_i(t-1))$ is a Gaussian with mean that depends linearly on the previous value: $\mu_i(t) = \bar{s}_i(t) + \breve{s}_i(t-1, t)(s_i(t-1) - \bar{s}_i(t-1))$, and variance $\mathring{s}_i\ (t)$. The variational parameters of the distribution are $\bar{s}_i(t)$, $\breve{s}_i(t-1, t)$ and $\mathring{s}_i\ (t)$.

3.2. Learning scheme

Minimisation of the cost function (5) is based on iterative gradient-based search. In general the learning proceeds in batches. After each sweep through the data the distributions $q(\boldsymbol{S})$ and $q(\boldsymbol{\theta})$ are updated. There are slight changes to the basic learning scheme in the beginning of training. The hyperparameters governing the distributions of other parameters are not updated to avoid pruning away parts of the model that do not seem useful at the moment.

If some initial guesses can be given to the factors, the learning process is much faster. In this case we are interested in rhythmical activity, *i.e.* signals with varying time structure and simple frequency content.

We suggest that DFA would be initialized with band-pass filtered principal components of the data. Cut-off frequencies for the band-pass filters can be set manually looking at the power spectrum of the principal component or the process can be automated by finding important bumps in it. Also prior information can be used to set the filters.

For periodical signals, robust prediction of the dynamics of the signal is achieved, when there are, in addition to the real signal, another factor $\pi/2$ shifted to the original one. This can be achieved if the frequency responses of the band-pass filters are set to be zero on band $[\pi, 2\pi]$. Then inverse Fourier transform does not become real, but has real and imaginary parts. Real part corresponds to the normal band-pass filtered signal, and the imaginary part corresponds to $\pi/2$ shift of it. It is suspected that the factor initialized using the imaginary part will not be connected to the observations at all, but is used only for the dynamics.

4. EXPERIMENTS

4.1. Artificial modulated sinusoids

To test the algorithm, we generated three modulated oscillation signals, following the model:

$$s(t) = e^{\alpha(t)} \sin \beta(t), \quad (6)$$

i.e., each signal corresponded to a sinusoid with amplitude $e^{\alpha(t)}$. Each $\alpha(t)$ and $\beta(t)$ were sampled from following distributions:

$$\alpha(0) \sim N\left(0, e^{2v_{\alpha(0)}}\right) \quad (7)$$
$$\alpha(t) \sim N\left(\alpha(t-1), e^{2v_{\alpha}}\right) \quad (8)$$
$$\beta(0) \sim N\left(0, e^{2v_{\beta(0)}}\right) \quad (9)$$
$$\beta(t) \sim N\left(\beta(t-1) + \mu, e^{2v_{\beta}}\right), \quad (10)$$

with parameters: $v_{\alpha(0)} = 1, v_{\beta(0)} = 1, v_{\alpha} = -2.5, v_{\beta} = -2, \mu = 2\pi f/f_s$, where f is the average frequency of the sinusoid and $f_s = 200$ Hz is the sampling frequency. Frequencies were randomly chosen from uniform distributions between $(8, 12)$, $(13, 15)$ and $(19, 21)$ Hz. These three original signals can be seen in Fig. 2.

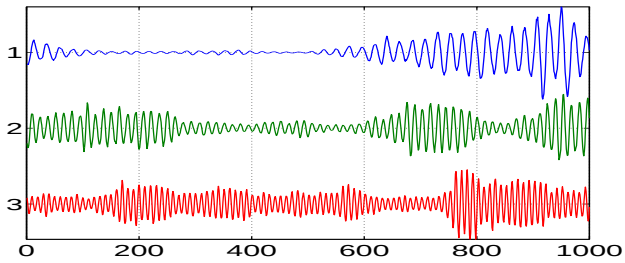

Fig. 2. *Three artificially generated modulated sinusoids.*

Three linear mixtures were generated from the original signals. The weights of the mixing matrix were randomly

picked from uniform distributions between $(-1, 1)$. White Gaussian noise with variance $\delta_n = 0.5$ was added to each mixture. These mixtures are depicted in Fig. 3.

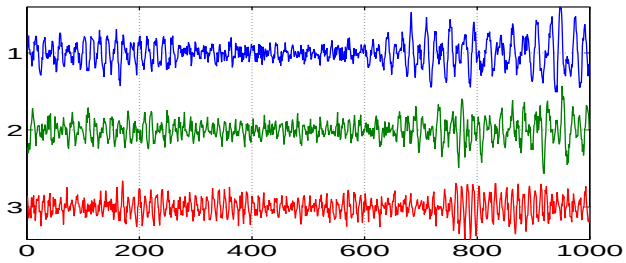

Fig. 3. *Three mixtures of artificially generated modulated sinusoids.*

DFA factors were initialized using the scheme described in section 3.2 with predefined filters with pass-bands of 8–12, 13–15 and 19–21 Hz cut-off frequencies. Weights of the linear mapping were initialized to the maximum likelihood estimates. Results of learning after 2000 iterations of DFA are shown in Fig. 4.

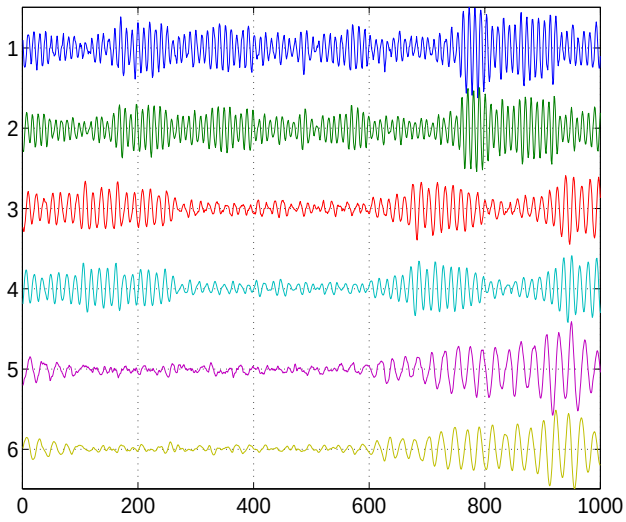

Fig. 4. *Results of learning after 2000 iterations of DFA.*.

The learned linear mapping is shown as Hinton graph in Fig.5. Note that the weights from the even factors to the observations are practically zero as expected (see end of Sec. 3.2).

4.2. Rhythmic MEG data

The usability of DFA in practice was tested in spontaneous rhythmic MEG measurements of a female subject. Of the possible 122 channels and several minutes of recording, a

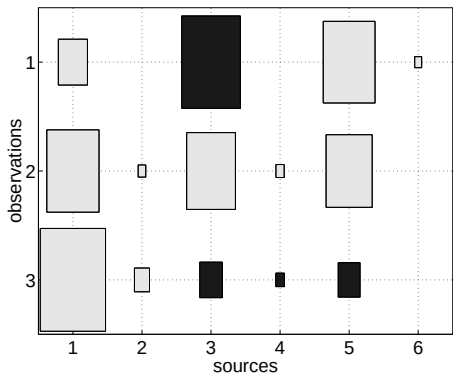

Fig. 5. *Means of the weights of the linear mapping after learning in artificial case. Dark colour correspond to negative values, lighter colour to positive values.*

period of 10 seconds and a selection of 12 MEG channels was used (for complete information on the measuring device, as well as the MEG itself see, *e.g.*, [8]). The data was sampled at $f_s = 200$ Hz, and high-pass filtered with cut-off frequency of 1 Hz. This data is shown in Fig. 6. During the second half of the measurement period, the subject was asked to close her eyes, resulting in the appearance of clear α-rhythm in various sensors. This can be seen, *e.g.*, in the short-time Fourier transform of the first channel (Fig. 7), where also the 50 Hz power line is visible.

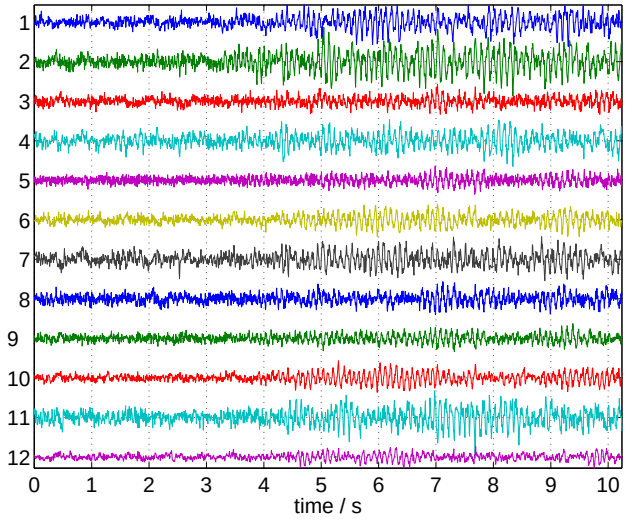

Fig. 6. *Short fragment of MEG recording over twelve channels.*

The factors were initialized using the manual initialization scheme presented in sec. 3.2, with predefined filters with pass-bands of 1–8, 8–12, and 49–51 Hz. These choices were made based on the power spectrum of the data. Six

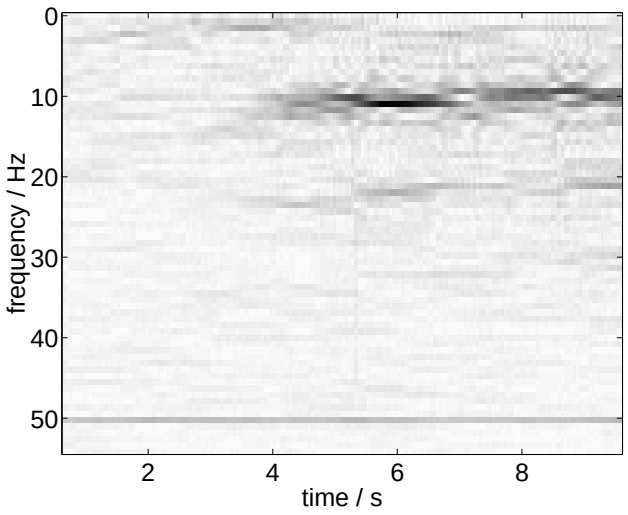

Fig. 7. *Short time Fourier transform of the first channel of MEG recordings. α-rhythm is visible in the latter half of the channel.*

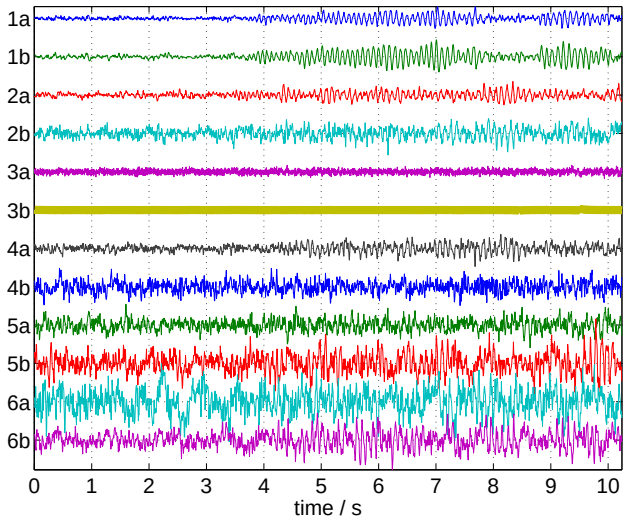

Fig. 8. *Factors learned from MEG recordings.*

blocks of two sources were learned using DFA. Means of their time courses are shown in Fig. 8 and their short-time Fourier transforms in Fig. 9. The first two factors are very clear α-rhythms, starting around 4 seconds. The first one corresponds to the component associated with the observations, whereas the second is the $\pi/2$ delayed version (see table 1 for mean squared weights of each factor in the linear mapping).

	1st	2nd	3rd	4th	5th	6th
a	0.12	0.13	0.05	0.06	0.01	0.01
b	0.00	0.01	0.02	0.01	0.02	0.02

Table 1. *Mean squared weights of the linear observation mapping times the variances of the sources in MEG case. Columns correspond to different blocks and rows to factors in blocks.*

A careful look at the 2a and 2b factors, and their respective spectrograms, tells us that a different α-rhythm is as well present in the measurements. Although the frequency content of these signals is very close to the one of the previous components, the time activations differ significantly. In fact, the main activation of factor 2a lies between the 8th and 9th seconds. This result opens good prospects for the use of DFA in identification, differentiation and characterization of rhythmic brain activity.

Another important information that we can extract from Figs. 8 and 9, is that a frequency component, around 20 Hz, faintly present in the original data, has been successfully recovered by factor 4a. In fact, even the weights of the linear mapping (see Table 1) give this component as an important one. Note that no frequency range around 20 Hz has been used during the initialization stage, showing that DFA, although privileging the frequency contents that have been used, can learn other dynamics present in the data.

Factors 3a and 3b capture both phases of the 50 Hz power line.

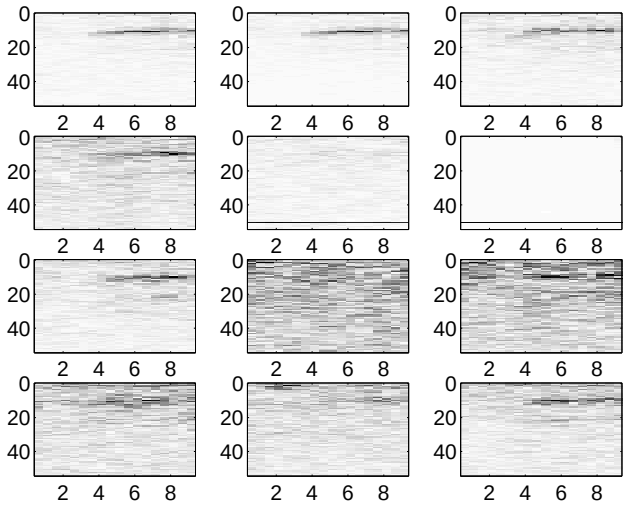

Fig. 9. *Short-time Fourier transforms of the factors learned from MEG recordings. The topmost row correspond to the factors 1–3 etc.*

5. DISCUSSION AND CONCLUSION

As in many other BSS approaches, DFA assumes that the set of observations $\mathbf{x}(t)$ are linear mixtures of underlying hidden factors $\mathbf{s}(t)$. Unlike most approaches, it assumes that each realization of the states can be nonlinearly modeled from their previous realizations. Using a very general and efficient Bayesian approach, the algorithm is capable of uncovering the hidden states. Furthermore, and due to its generative nature, the models found by DFA can easily be used to draw predictions from the observations. In fact, the state of each factor, at a given time instant, determines its state at the successive time instant.

DFA performs an explicit modeling of the factors. The dynamical part of DFA can determine, at each realization t of the observation $\mathbf{x}(t)$, which portion corresponds to the underlying estimated factors, $\mathbf{s}(t)$, and which corresponds to noise, $\mathbf{n}(t)$. DFA can therefore deal better with noisy signals than algorithms that are uncapable of such explicit modeling, even though they may be using implicitly some temporal information.

Ensemble learning favors simpler and smoother models. Furthermore, the predictions are made over a collection of models. The probability of overfitting to a particularly strong, but very peaky, posterior is therefore very small. Because DFA uses ensemble learning to estimate the models for the factors, these are less prone to overlearning, and perform automatically some form of regularization.

Due to its modular nature, it is as well easy to foresee future updates of the present DFA structure, to accomodate further algorithmic developments. As an example, some model of the nonstationarity of the data could be estimated.

In this paper we have given two illustrations of the results one can obtain when using DFA on rhythmic data. The first, in a "controlled environment", has let us understand the mechanisms of its functioning. The second, using MEG data, patented its potential application to real-world recordings.

Further research, covering a wider range of sensors, will allow us to make a more serious use of DFA in the determination and separation of spontaneous brain oscillations. Then it will be possible to better map the location of the brain areas involved in the production of such oscillation. Even though DFA is a computationally demanding technique, it is expected that it will provide, in the near future, physicians and brain researchers with a very nice tool for the analysis of brain rhythms.

6. REFERENCES

[1] Tzyy-Ping Jung, Scott Makig, Te-Won Lee, Martin J. McKeown, Blen Brown, Anthony J. Bell, and Terrence J. Sejnowski, "Independent component analysis of biomedical signals," in *workshop on Independent Component Analysis and Blind Source Separation of Signals (ICA'99)*, Helsinki, Finland, 2000, pp. 633–644.

[2] Ricardo Vigário and Erkki Oja, "Independence: a new criterion for the analysis of the electromagnetic fields in the global brain?," *Neural Networks*, vol. 13, pp. 891–907, 2000.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, Wiley, 1st edition, 2001.

[4] A. Ziehe and K.-R. Müller, "TDSEP — an effective algorithm for blind separation using time structure," in *Proc. int. conf. at neural networks (ICANN'98)*, Skövde, Sweden, 1998, pp. 675–680.

[5] G. Hinton and D. van Camp, "Keeping neural networks simple by minimizing the description length of the weights," in *Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, Santa Cruz, CA, USA, 1993, pp. 5–13.

[6] H. Lappalainen and J. Miskin, "Ensemble learning," in *Advances in Independent Component Analysis*, M. Girolami, Ed., pp. 75–92. Springer, Berlin, 2000.

[7] Ernst Niedermeyer and Fernando Lopes da Silva, Eds., *Electroencephalography. Basic principles, clinical applications, and related fields*, Baltimore: Williams & Wilkins, 1993.

[8] M. Hämäläinen, R. Hari, R.J. Ilmoniemi, J. Knuutila, and O.V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of Modern Physics*, vol. 65, pp. 413–497, 1993.

[9] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE transactions on biomedical engineering*, vol. 47, no. 5, pp. 589–593, 2000.

[10] Jaakko Särelä and Ricardo Vigário, "The problem of overlearning in high-order ICA approaches :analysis and solutions," in *Proceedings of the international workshop on artificial neural networks (IWANN'01)*, Granada,Spain, 2001, pp. 818–825.

[11] S. Haykin, *Neural Networks – A Comprehensive Foundation, 2nd ed.*, Prentice-Hall, 1998.

[12] H. Valpola and J. Karhunen, "An unsupervised ensemble learning method for nonlinear dynamic state-space models," 2001, Submitted to a journal.

[13] H. Lappalainen and A. Honkela, "Bayesian nonlinear independent component analysis by multi-layer perceptrons," in *Advances in Independent Component Analysis*, M. Girolami, Ed., pp. 93–121. Springer-Verlag, 2000.

[14] H. Valpola, "Unsupervised learning of nonlinear dynamic state-space models," Tech. Rep. A59, Lab of Computer and Information Science, Helsinki University of Technology, Finland, 2000.