Independent Component Analysis in a convoluted world

Mads Dyrholm

Kongens Lyngby 2005
IMM-PHD-2005-158

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Abstract

This thesis is about convolutive ICA with application to EEG. Two methods for convolutive ICA are proposed.

One method, the CICAP algorithm, uses a linear predictor in order to formulate the convolutive ICA problem in two steps: linear deconvolution followed by instantaneous ICA.

The other method, the CICAAR algorithm, generalizes Infomax ICA to include the case of convolutive mixing. One advantage to the CICAAR algorithm is that Bayesian model selection is made possible, and in particular, it is possible to select the optimal order of the filters in a convolutive mixing model. A protocol for detecting the optimal dimensions is proposed, and verified in a simulated data set.

The role of instantaneous ICA in context of EEG is described in physiological terms, and in particular the nature of dipolar ICA components is described. It is showed that instantaneous ICA components of EEG lacks independence when time lags are taken into consideration. The CICAAR algorithm is shown to be able to remove the delayed temporal dependencies in a subset of ICA components, thus making the components "more independent". A general recipe for ICA analysis of EEG is proposed: first decompose the data using instantaneous ICA, then select a physiologically interesting subspace, then remove the delayed temporal dependencies among the instantaneous ICA components by using convolutive ICA. By Bayesian model selection, in a real world EEG data set, it is shown that convolutive ICA is a better model for EEG than instantaneous ICA.

Resumé

Denne afhandling omhandler convolutive ICA med applikation indenfor EEG. To metoder til convolutive ICA er beskrevet.

Den ene metode, CICAP metoden, benytter en lineær prædiktor for at formulere problemet i to skridt: lineær affoldning efterfulgt af instantan ICA.

Den anden metode, CICAAR metoden, generaliserer Infomax ICA til at omfatte foldede miksturer. En fordel ved CICAAR metoden er at Bayesiansk model selektion er mulig, og specielt er det muligt at vælge den optimale længde af filtrene i en foldende mikstur model. En protokol til at finde de optimale dimensioner er foreslået, of verificeret i et simuleret datasæt.

Instantan ICA bliver belyst i forbindelse med EEG, og specielt med henblik på dipolare ICA komponenters opståen af fysiologiske årsager. Det vises at de instantane ICA komponenter ikke er uafhængige hvis man tager forsinkelser i betragtning. Det viser sig at CICAAR metoden kan fjerne disse forsinkede temporale afhængigheder i en delmængde af de instantane ICA komponenter, og således gøre komponenterne "mere uafhængige". En generel opskrift på ICA analyse af EEG bliver foreslået: først dekomponeres data med instantan ICA, dernæst vælges en delmængde af fysiologisk interessante komponenter, dernæst fjernes de forsinkede afhængigheder med convolutive ICA. Ved Bayesiansk model selektion, i et ægte EEG datasæt, bliver det vist at convolutive ICA er en bedre model end instantan ICA.

Preface

This thesis was prepared at Informatics and Mathematical Modelling (IMM), the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

Lyngby, December 2005

Mads Dyrholm

Papers included in the thesis

[21] L. K. Hansen and M. Dyrholm, A prediction matrix approach to convolutive ICA, Proceedings of IEEE Workshop on Neural Networks for Signal Processing XIII, pp. 249-258, 2003

[13] M. Dyrholm and L. K. Hansen, CICAAR: Convolutive ICA with an Auto-Regressive Inverse Model, Independent Component Analysis and Blind Signal Separation, pp. 594-601, 2004

[14] M. Dyrholm, L. K. Hansen, L. Wang, L. Arendt-Nielsen and A. C. Convolutive ICA (c-ICA) captures complex spatio-temporal EEG activity, 10th annual meeting of the organization for human brain mapping, 2004

[16] M. Dyrholm, S. Makeig and L. K. Hansen, Model selection for convolutive ICA with an application to spatio-temporal analysis of EEG, Neural Computation, (submitted) 2005

[15] M. Dyrholm, S. Makeig and L. K. Hansen, Model structure selection in convolutive mixtures, (submitted) 6th International Conference on Independent Component Analysis and Blind Source Separation, 2006

Acknowledgements

TAK!...

♡Nuser♡, Scott, P, Kyllingsbæk, Olsson, Syskind, Terry, LKH, Torben, NHP, Fabricius,... and all who worked at SCCN while I was there,... and everyone at ISP IMM.

This work was funded by the Danish Technical Research Council through the International Center For Biomedical Research.

Contents

Chapter 1

Introduction

Electromagnetic activity from the human brain can be measured by sensitive electrodes positioned on the skin surface of the human head. This measurement technique is known as electroencephalography (EEG), and it opens the possibility of studying ongoing dynamics in a working human brain without having to open the skull that surrounds the brain. For instance, EEG is well known for diagnostics of epilepsy, and for monitoring patients that suffer from epileptic strokes. Other measurement techniques, as for instance functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET), do not require opening the skull either, but EEG has a very high temporal resolution compared to these methods. Furthermore, hardware for EEG data acquisition is cheap and easy to implement and the technique has thus achieved widespread focus in both research and industry.

Historically, time-domain analysis of EEG has mainly been limited to averages of many experimental repeats. However, progress in neurophysiology suggest that brain dynamics are closely connected to dynamic reallocation of attention, to memory related activity, and to self evaluation of consequences of actions. This again suggests that an appropriate analysis of EEG must include separation of independent brain components, and modelling of their dynamic interrelations.

Independent Component Analysis (ICA) is a method for separating signals that occur in an observed mixture, and ICA has become a widespread technique for

removing some of the artifacts that very often contaminate EEG. The reason that ICA works well for this purpose is that the noise is mixed 'instantaneously' with the EEG. This means that there are no echoes or delays in the mixing, and the mathematical model that underlies ICA can therefore be written on the form of a simple General Linear Model. However, limitations of this simple form imply that ICA can not be used to solve a significant classical problem, namely "The Cocktail Party Problem" where sound signals are mixed in a reverberant environment. In EEG there are no echoes as such in the mixture, but there is potentially an interesting analogy from EEG to the Cocktail Party Problem anyway: Different cortical areas might interact in a 'reverberant' way.

It turns out that ICA can be generalized to include so-called 'convolutive mixtures' and can potentially solve The Cocktail Party Problem. The form of ICA that builds around the convolutive mixing model is known as 'convolutive Independent Component Analysis' (convolutive ICA) and is the main focus of this thesis. This thesis will present two original methods for convolutive ICA, and explore the problem theoretically. Furthermore, effort will be put into testing whether convolutive ICA is relevant in EEG.

Chapter 2

Convolutive Mixtures

A convolutive mixture model can be seen as a generalization of the General Linear Model (GLM) in which the 'source' signals (the regressors) are *filtered* before mixing in the data. The filtering is individual for each combination of data dimension $d \in \{1 \dots D\}$ and source dimension $k \in \{1 \dots K\}$, and a noise-free D-dimensional convolutive mixture can thus be written as

$$\mathbf{x}_t = \sum_{\tau=1}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \tag{2.1}$$

where L is the order of the mixing filters, the $L+1$ matrices $\{\mathbf{A}_\tau\}$ are the time-lagged 'mixing matrices', and the N source signal vectors $\mathbf{s}_t$ are of dimension K. When the number of data dimensions equals the number of sources, i.e. when $D = K$, the mixture is 'square'. When $D > K$ the mixture is 'overdetermined'. When $D < K$ the mixture is 'underdetermined'.

This chapter deals with two fundamentally different situations where the mixing matrices are sought estimated: 1) when the sources are known, and 2) when both the mixing matrices and the sources are unknown.

2.1 The multivariate Wiener Filter

If the sources are known, the mixing matrices of a convolutive mixture can be estimated by least-squares estimation, i.e. solving

$$< \mathbf{x}_t \mathbf{s}_{t-\lambda}{}^{\mathrm{T}} > = \sum_{\tau} \mathbf{A}_\tau < \mathbf{s}_{t-\tau} \mathbf{s}_{t-\lambda}{}^{\mathrm{T}} > \tag{2.2}$$

for $\mathbf{A}_\tau$ by matrix inversion. This is a generalization of the Wiener-Hopf equations (see e.g. [48]) for estimating the coefficients of a 'Wiener filter' to the multivariate case. Thus, the 'multivariate Wiener filter equation' (2.2) is the key to estimating the mixing matrices of a convolutive mixture when the sources are known.

2.2 Convolutive ICA

The problem of identifying both the mixing matrices and the source signals from the data, based on the assumption that the source signals are statistically independent, is known as 'convolutive Independent Component Analysis' or 'convolutive ICA'. One common application for convolutive ICA is the problem of acoustic blind source separation (BSS) where sound sources have been mixed in a reverberant environment and are sought separated. 'Instantaneous' ICA is a special case of convolutive ICA where $L = 0$, i.e. not taking signal delays and echoes into account. Hence, instantaneous ICA methods fail to produce satisfactory results for the acoustic BSS problem which has thus been the focus of much convolutive ICA research, see e.g. [29, 44, 2].

2.2.1 Identifiability

Generally in ICA, the ordering of the sources is arbitrary since any reordering would simply imply the same reordering of the columns of each mixing matrix. This ambiguity is known as the 'permutation ambiguity'. Furthermore, an arbitrary linear filter can be applied to any of the sources since the inverse filtering applied to each of the mixing filters for that source would keep the model consistent with the same data. This ambiguity is known as the 'filter ambiguity'.

Assuming a convolutive mixture, correlations in the data are summarized in this

linear system

$$< \mathbf{x}_t \mathbf{x}_{t-\lambda}^{\mathrm{T}} >= \sum_{\tau,\tau'} \mathbf{A}_\tau < \mathbf{s}_{t-\tau} \mathbf{s}_{t-\tau'-\lambda}^{\mathrm{T}} > \mathbf{A}_{\tau'}^{\mathrm{T}} \tag{2.3}$$

For *stationary* sources, all source auto-correlation (and scaling) can be explained by the mixing matrices (due to the filter ambiguity) and the sources can thus be assumed temporally uncorrelated, i.e.

$$< \mathbf{s}_{t-\tau} \mathbf{s}_{t-\tau'-\lambda}^{\mathrm{T}} >= \begin{cases} \mathbf{I} & \text{for} \quad \tau = \tau' + \lambda \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{2.4}$$

Then

$$< \mathbf{x}_t \mathbf{x}_{t-\lambda}{}^{\mathrm{T}} >= \sum_{\tau'} \mathbf{A}_{\tau'+\lambda} \mathbf{A}_{\tau'}^{\mathrm{T}} \tag{2.5}$$

meaning that correlations in the data can be explained entirely by the mixing matrices. But for any orthogonal matrix $\mathbf{Q}$ it holds that

$$< \mathbf{x}_t \mathbf{x}_{t-\lambda}{}^{\mathrm{T}} >= \sum_{\tau'} \mathbf{A}_{\tau'+\lambda} \mathbf{Q} \mathbf{Q}^{\mathrm{T}} \mathbf{A}_{\tau'}^{\mathrm{T}} \tag{2.6}$$

and the mixing matrices can thus only be identified up to an arbitrary column rotation. The conclusion is, for stationary sources, that convolutive ICA is not to be solved using second order statistics only, see also [19, 24].

2.2.2 Invertibility

Formally, when the number of sources does not exceed the dimension of the data, i.e. when $K \leq D$, perfect inversion of a convolutive mixture is obtained through the autoregressive operator

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^+ \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \hat{\mathbf{s}}_{t-\tau} \right) \quad , \; K \leq D \tag{2.7}$$

where $\mathbf{A}_0^+$ denotes Moore-Penrose inverse of $\mathbf{A}_0$. This follows simply from eliminating $\mathbf{s}_t$ in (2.1). For $D = K$, (2.7) is the only perfect inverse of the convolutive mixture, and the recursive structure of (2.7) illustrates an inherent problem in convolutive ICA: — some convolutive mixtures are not invertible — i.e. the sources can not be separated perfectly, as is the case when the recursive filter (2.7) is unstable and $D = K$.

For the sake of stability, the use of IIR filters for unmixing has often been discouraged in convolutive ICA research, see e.g. [29], and most previous methods

for convolutive ICA have formulated the problem instead as one of identifying a FIR unmixing model

$$\hat{\mathbf{s}}_t = \sum_{\lambda=0}^{Q} \mathbf{W}_\lambda \mathbf{x}_{t-\lambda} \tag{2.8}$$

see e.g. [7, 29, 38, 3, 12, 44, 4, 8, 49, 56, 9, 50, 2]. Using such FIR model for unmixing can ensure stable estimation of the sources but will not solve the fundamental problem of perfect inversion of a linear system in cases in which it is not invertible.

Invertibility of a linear system is related to the phase characteristic of the system transfer function. A SISO (single input / single output) system is invertible if and only if the complex zeros of its transfer function are all situated within the unit circle. Such a system is characterized as 'minimum phase' [48]. If the system is not minimum phase, only an approximate, 'regularized' inverse can be sought, see e.g. [22] on techniques for regularizing a system with known coefficients. For MIMO (multiple input / multiple output) systems, the matter is more involved. The stability of (2.7), and hence the invertibility of (2.1), is related to the eigenvalues λ_m of the matrix

$$\begin{bmatrix} -\mathbf{A}_0^+\mathbf{A}_1 & -\mathbf{A}_0^+\mathbf{A}_2 & \dots & -\mathbf{A}_0^+\mathbf{A}_L \\ \mathbf{I} & & & \mathbf{0} \\ & \ddots & & \vdots \\ & & \mathbf{I} & \mathbf{0} \end{bmatrix} \tag{2.9}$$

For $K = D$, a necessary and sufficient condition is that all eigenvalues λ_m of (2.9) are situated within the unit circle, $|\lambda_m| < 1$ [39]. The 'minimum phase' concept can thus be generalized to MIMO systems, where the eigenvalues of (2.9) are the generalized 'poles' of the transfer function of the inverse of the MIMO system. A SISO system being minimum phase implies that no system with the same frequency response can have a smaller phase shift and system delay. Generalizing that concept to MIMO systems, it is possible to get a feeling for what a generalized 'minimum phase' MIMO system must look like. In particular, most energy must occur at the beginning of each filter, and less towards the end. However, not all SISO source-to-sensor paths in the MIMO system need to be minimum phase for the MIMO system as a whole to be generalized 'minimum phase'.

2.2.3 FFT based methods

Convolution in time domain is equivalent to multiplication of Fourier transforms, thus, the convolutive mixture can be written for each frequency f as

$$\tilde{\mathbf{x}}_f = \tilde{\mathbf{A}}_f \tilde{\mathbf{s}}_f \tag{2.10}$$

where $\tilde{\mathbf{x}}_f$, $\tilde{\mathbf{A}}_f$, and $\tilde{\mathbf{s}}_f$ are Fourier transforms of $\mathbf{x}_t$, $\mathbf{A}_t$, and $\mathbf{s}_t$ respectively. This suggests that convolutive ICA can be reduced to solving an instantaneous ICA problem at each frequency. But, an individual permutation problem applies to every instantaneous ICA decomposition, i.e. the ordering of the sources is individual for each frequency, hence reconstruction of the convolutive components involves solving a massive cross-frequency permutation problem. For non-stationary sources, second order statistics can be used to solving the massive cross-frequency permutation problem as in e.g. [44, 2].

Chapter 3

Algorithm I: CICAAR

In this chapter a maximum likelihood algorithm for convolutive ICA is proposed. The 'CICAAR' algorithm is a pure generalization of the Infomax ICA algorithm (the Bell-Sejnowski algorithm [5]) to convolutive mixtures. Infomax ICA is highly regarded in EEG analysis (see e.g. [32, 10]) and the generalization of Infomax ICA to convolutive mixtures is a principled direction for investigating the properties of convolutive ICA in EEG.

3.1 Likelihood for square mixing

The derivation of the likelihood for a square convolutive mixing model takes departure in the following matrix product abbreviation of a general convolutive mixing model:

$$\begin{bmatrix} \mathbf{x}_N \\ \mathbf{x}_{N-1} \\ \vdots \\ \mathbf{x}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{A}_1 & \dots & \mathbf{A}_L & & \\ & \mathbf{A}_0 & \mathbf{A}_1 & \dots & \mathbf{A}_L & \\ & & & \ddots & & \\ & & & & & \mathbf{A}_0 \end{bmatrix} \begin{bmatrix} \mathbf{s}_N \\ \mathbf{s}_{N-1} \\ \vdots \\ \mathbf{s}_1 \end{bmatrix} \tag{3.1}$$

where, from here on, the upper triangular block Toeplitz mixing matrix is denoted by $\mathbf{T}$, the left column vector by $\mathbf{x}$, and the right column vector by $\mathbf{s}$. This

representation allows the likelihood, assuming no noise, to be written

$$l(\{\mathbf{A}_\tau\}) = \int \delta(\mathbf{x} - \mathbf{T}\mathbf{s})\, \mathrm{p}(\mathbf{s})\, \mathrm{d}\,\mathbf{s} \tag{3.2}$$

which evaluates to

$$l(\{\mathbf{A}_\tau\}) = |\det \mathbf{T}|^{-1}\, \mathrm{p}(\mathbf{T}^{-1}\mathbf{x}) \tag{3.3}$$

c.f. section M.2. The determinant of an upper block triangular matrix equals the product of the determinants for each block on the diagonal [40], hence

$$l(\{\mathbf{A}_\tau\}) = |\det \mathbf{A}_0|^{-N}\, \mathrm{p}(\mathbf{T}^{-1}\mathbf{x}) \tag{3.4}$$

By assuming the source signals to be i.i.d., the likelihood is now written

$$l(\{\mathbf{A}_\tau\}) = |\det \mathbf{A}_0|^{-N} \prod_{t=1}^{N} \mathrm{p}(\hat{\mathbf{s}}_t) \tag{3.5}$$

where $\hat{\mathbf{s}}_t$ is the estimate of source vector $\mathbf{s}_t$ from matrix inversion of $\mathbf{T}$. The inverse of $\mathbf{T}$ can be written on operator form as a multivariate AR(L) process

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^{-1} \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \hat{\mathbf{s}}_{t-\tau} \right) \tag{3.6}$$

which follows simply by eliminating $\mathbf{s}_t$ in (2.1). One important property of (3.6) is that it is presented in terms of the model parameters, i.e. the $\mathbf{A}_\tau$'s. The cost-function of the CICAAR algorithm, the *negative log* likelihood, can thus be written in terms of the mixing model parameters

$$-\log l(\{\mathbf{A}_\tau\}) = N \log |\det \mathbf{A}_0| - \sum_{t=1}^{N} \log \mathrm{p}(\hat{\mathbf{s}}_t) \tag{3.7}$$

Thus, the cost-function is calculated by first *un*mixing the sources using (3.6), then measuring (3.7). It is clear that this cost-function reduces to that of standard Infomax ICA [5] when the order L of the convolutive model is set to zero; in that case (3.7) can be estimated using $\hat{\mathbf{s}}_t = \mathbf{A}_0^{-1}\mathbf{x}_t$.

Other authors have proposed the use of IIR filters for separating convolutive mixtures using the maximum likelihood principle. The CICAAR cost-function (3.7) generalizes that of [57] to allow separation of more than only two sources. Furthermore, the auto-regressive inverse (3.6) used in the CICAAR cost-function bears interesting resemblance to that of [7, 6]. Though put in different analytical terms, the inverses used there are equivalent to the CICAAR inverse. However, the unique CICAAR expression (3.6), and its remarkable analytical simplicity, is the key to learning the parameters of the *mixing* model (2.1) directly.

3.1.1 Automatic handling of instability

As described in section 2.2.2, an IIR unmixing process can potentially become unstable. Since (3.6) is IIR, instability must be controlled somehow. Fortunately, the maximum likelihood approach has a built-in regularization that avoids the problem. This can be seen in the likelihood equation (3.5) by noting that although an unstable IIR filter will lead to a divergent source estimate, $\hat{\mathbf{s}}_t$, such large amplitude signals are exponentially penalized under most reasonable source probability density functions (pdf's), e.g. for EEG data $\mathrm{p}(s) = \mathrm{sech}(s)/\pi$, ensuring that unstable solutions are avoided in the evolved solution. Therefore, it may prove safe to use an unconstrained iterative learning scheme to unmix e.g. EEG data. Once the unmixing process has been stably initialized, each learning step will produce model refinements that are stable in the sense of equation (3.6). Even if the system (2.1) is not invertible, meaning no exact stable inverse exists, the maximum-likelihood approach will give a regularized and stable generalized minimum phase solution c.f. section 2.2.2.

3.1.2 Computing the gradient

The tradition with Infomax ICA is to optimize the cost-function w.r.t. to the parameters of the *un*mixing matrix see e.g. [31, 47]. The CICAAR algorithm follows this tradition by optimizing w.r.t. the parameters of the *un*mixing system, i.e. the elements of $\mathbf{A}_0^{-1}$ and the elements of $\mathbf{A}_\tau$. Optimization is gradient based, and the gradient of the cost-function is now presented in two parts. Part one reveals the partial derivatives of the source estimates while part two uses the result from part one to compute the gradient of the cost function.

Part one — Partial derivatives of the unmixed source estimates

The partial derivatives which shall be used in part two are given by

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_0^{-1})_{ij}} = \delta(i-k)\left(\mathbf{x}_t - \sum_{\tau=1}^{L}\mathbf{A}_\tau\hat{\mathbf{s}}_{t-\tau}\right)_j - \left(\mathbf{A}_0^{-1}\sum_{\tau=1}^{L}\mathbf{A}_\tau\frac{\partial\hat{\mathbf{s}}_{t-\tau}}{\partial(\mathbf{A}_0^{-1})_{ij}}\right)_k \tag{3.8}$$

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} = -(\mathbf{A}_0^{-1})_{ki}(\hat{\mathbf{s}}_{t-\tau})_j - \left(\mathbf{A}_0^{-1}\sum_{\tau'=1}^{L}\mathbf{A}_{\tau'}\frac{\partial\hat{\mathbf{s}}_{t-\tau'}}{\partial(\mathbf{A}_\tau)_{ij}}\right)_k \tag{3.9}$$

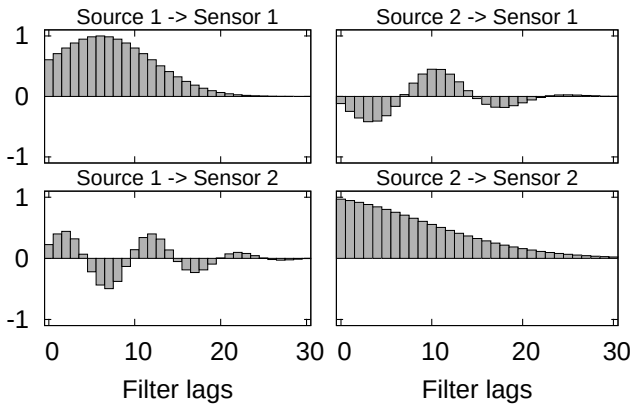


Figure 3.1: Convolutive mixing model of order $L = 30$ for producing a square mixture two sources. This system is well-posed, meaning that the eigenvalues of (2.9) are situated within the unit circle and hence exact inversion is possible through (3.6).

Part two — Gradient of the cost function

The gradient of the cost function with respect to $\mathbf{A}_0^{-1}$ is given by

$$\frac{\partial - \log l(\{\mathbf{A}_\tau\})}{\partial (\mathbf{A}_0^{-1})_{ij}} = -N(\mathbf{A}_0{}^{\mathrm{T}})_{ij} - \sum_{t=1}^{N} \boldsymbol{\psi}_t{}^{\mathrm{T}} \frac{\partial \hat{\mathbf{s}}_t}{\partial (\mathbf{A}_0^{-1})_{ij}} \tag{3.10}$$

where $(\ \boldsymbol{\psi}_t\ )_k = \mathrm{p}'(\ (\hat{\mathbf{s}}_t)_k\ )/\,\mathrm{p}(\ (\hat{\mathbf{s}}_t)_k\ ) = -\tanh(\ (\hat{\mathbf{s}}_t)_k\ )$. The gradient with respect to the other mixing matrices is

$$\frac{\partial - \log l(\{\mathbf{A}\})}{\partial (\mathbf{A}_\tau)_{ij}} = -\sum_{t=1}^{N} \boldsymbol{\psi}_t{}^{\mathrm{T}} \frac{\partial \hat{\mathbf{s}}_t}{\partial (\mathbf{A}_\tau)_{ij}} \tag{3.11}$$

These expressions allow use of general gradient optimization methods. Refer to section T.2 for implementation details.

3.1.3 Example — Joint deconvolution and unmixing

Two sources of length $N = 30000$ are drawn i.i.d. from a Laplace distribution. For visualization purposes the signals are raised to the power of two while preserving the sign. They are then mixed through the square system shown in figure 3.1. The system is well-posed, meaning that perfect inversion is possible

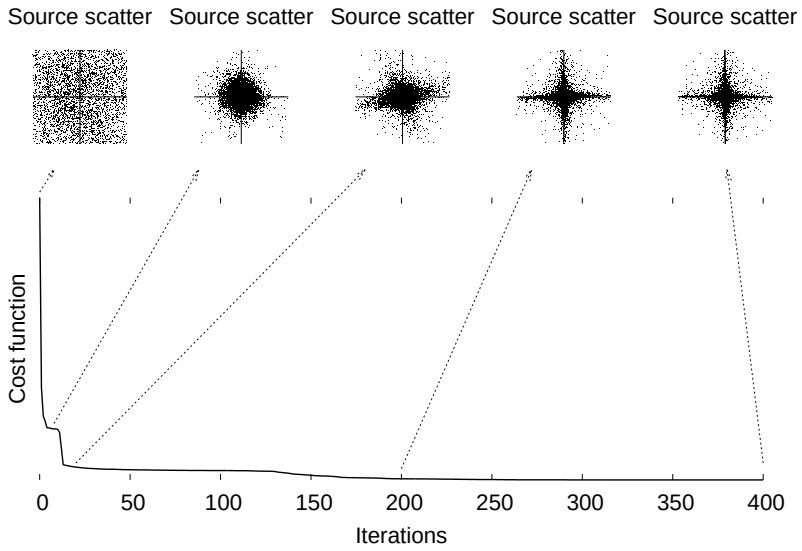


Figure 3.2: Cost function optimization. The scatter plots illustrate the progress of joint deconvolution and unmixing. In the early iterations, the cost function decreases a lot due to the fact that a simple re-scaling of the data will fit the prior much better. Later, deconvolution is responsible for the refinements that produce the star-shape in the scatter plots, and jointly, the refinements align the star-shape along the axes thus unmixing the sources.

as the eigenvalues of (2.9) are situated within the unit circle. Figure 3.2 shows how minimization of the cost function yields joint deconvolution and unmixing. Deconvolution is responsible for producing the star shape in the source scatter plots, and the signals are unmixed by aligning the star shape along the axes.

3.1.4 Modelling auto-correlated sources

The assumption in the likelihood, that source signals are i.i.d., is fundamentally okay for stationary sources since source auto-correlations can be modelled by the mixing model, c.f. the filter ambiguity in section 2.2.1. However, a more economic representation in terms of the number of parameters can be obtained by introducing a model for each of the sources

$$s_k(t) = \sum_{\lambda=0}^{M} h_k(\lambda) z_k(t-\lambda) \tag{3.12}$$

where $z_k(t)$ represents an i.i.d. signal—a *whitened* version of the source signal. This allows a reduction of the value of L, i.e. lowering the number of parameters in the mixing model while still modelling the same amount of temporal dependencies in the data. Note that some authors of FIR unmixing methods have also used source models, e.g. [46, 45, 3].

The negative log likelihood of the model combining (2.1) and (3.12) is given for the square case

$$-\log l(\{\mathbf{A}_\tau\}, \{h_k(\lambda)\}) = N \log|\det \mathbf{A}_0| + N \sum_k \log|h_k(0)| - \sum_{t=1}^{N} \log \mathrm{p}(\tilde{\mathbf{z}}_t) \tag{3.13}$$

where $\tilde{\mathbf{z}}_t$ is a vector of whitened source signal estimates at time t using the AR operator

$$\tilde{z}_k(t) = \left(s_k(t) - \sum_{\lambda=1}^{M} h_k(\lambda) \tilde{z}_k(t-\lambda) \right) / h_k(0) \tag{3.14}$$

which is the inverse of (3.12). Without loss of generality the first coefficient in the filters can be set $h_k(0) = 1$, allowing the negative log likelihood to be written

$$-\log l(\{\mathbf{A}_\tau\}, \{h_k(\lambda)\}) = N \log|\det \mathbf{A}_0| - \sum_{t=1}^{N} \log \mathrm{p}(\hat{\mathbf{z}}_t) \tag{3.15}$$

where

$$\hat{z}_k(t) = \left(\hat{s}_k(t) - \sum_{\lambda=1}^{M} h_k(\lambda) \hat{z}_k(t-\lambda) \right) \tag{3.16}$$

The number of parameters in this model is $D^2(L+1)+DM$, thus if L can be reduced by increasing M instead, a more economic representation is obtained.

3.1.5 Computing the gradient

For notational convenience introduce the following matrix notation instead of (3.16), handling all sources in one matrix equation

$$\hat{\mathbf{z}}_t = \hat{\mathbf{s}}_t - \sum_{\lambda=1}^{M} \mathbf{H}_\lambda \hat{\mathbf{z}}_{t-\lambda} \tag{3.17}$$

where the $\mathbf{H}_\lambda$'s are diagonal matrices defined by $(\mathbf{H}_\lambda)_{ii} = h_i(\lambda)$.

In the following, the algorithm equations are split into three parts; 'part A' and 'part C' are in principle identical to the equations of part one and part two found in section 3.1.2, but with a new 'part B'. If M is set to zero part B does nothing, $\mathbf{z}_t$ equals $\mathbf{s}_t$ and the algorithm reduces to the plain square CICAAR without a source model.

Part A — Partial derivatives of the unmixed source estimates

The partial derivatives which shall be used in part B are given by

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_0^{-1})_{ij}} = \delta(i-k)\left(\mathbf{x}_t - \sum_{\tau=1}^{L}\mathbf{A}_\tau\hat{\mathbf{s}}_{t-\tau}\right)_j - \left(\mathbf{A}_0^{-1}\sum_{\tau=1}^{L}\mathbf{A}_\tau\frac{\partial\hat{\mathbf{s}}_{t-\tau}}{\partial(\mathbf{A}_0^{-1})_{ij}}\right)_k \tag{3.18}$$

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} = -(\mathbf{A}_0^{-1})_{ki}(\hat{\mathbf{s}}_{t-\tau})_j - \left(\mathbf{A}_0^{-1}\sum_{\tau'=1}^{L}\mathbf{A}_{\tau'}\frac{\partial\hat{\mathbf{s}}_{t-\tau'}}{\partial(\mathbf{A}_\tau)_{ij}}\right)_k \tag{3.19}$$

Part B — Partial derivatives of the whitened source estimates

The partial derivatives which shall be used in part C are given by

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{A}_0^{-1})_{ij}} = \frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_0^{-1})_{ij}} - \sum_{\lambda=1}^{M}\mathbf{H}_\lambda\frac{\partial(\hat{\mathbf{z}}_{t-\lambda})_k}{\partial(\mathbf{A}_0^{-1})_{ij}} \tag{3.20}$$

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} = \frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} - \sum_{\lambda=1}^{M}\mathbf{H}_\lambda\frac{\partial(\hat{\mathbf{z}}_{t-\lambda})_k}{\partial(\mathbf{A}_\tau)_{ij}} \tag{3.21}$$

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{H}_\lambda)_{ii}} = -\delta(k-i)(\hat{\mathbf{z}}_{t-\lambda})_i - \left(\sum_{\lambda'=1}^{M} \mathbf{H}_{\lambda'} \frac{\partial \hat{\mathbf{z}}_{t-\lambda'}}{\partial(\mathbf{H}_\lambda)_{ii}}\right)_k \tag{3.22}$$

The work involved in part B is minimal due to the diagonal structure of the $\mathbf{H}_\lambda$ matrices.

Part C — Gradient of the cost-function

The gradient of the cost-function (3.15), using the result in part B, is given by

$$\frac{\partial - \log l}{\partial(\mathbf{A}_0^{-1})_{ij}} = -N(\mathbf{A}_0{}^{\mathrm{T}})_{ij} - \sum_{t=1}^{N} \boldsymbol{\psi}_t{}^{\mathrm{T}} \frac{\partial \hat{\mathbf{z}}_t}{\partial(\mathbf{A}_0^{-1})_{ij}} \tag{3.23}$$

$$\frac{\partial - \log l}{\partial(\mathbf{A}_\tau)_{ij}} = -\sum_{t=1}^{N} \boldsymbol{\psi}_t{}^{\mathrm{T}} \frac{\partial \hat{\mathbf{z}}_t}{\partial(\mathbf{A}_\tau)_{ij}} \tag{3.24}$$

$$\frac{\partial - \log l}{\partial(\mathbf{H}_\lambda)_{ii}} = -\sum_{t=1}^{N} \boldsymbol{\psi}_t{}^{\mathrm{T}} \frac{\partial \hat{\mathbf{z}}_t}{\partial(\mathbf{H}_\lambda)_{ii}} \tag{3.25}$$

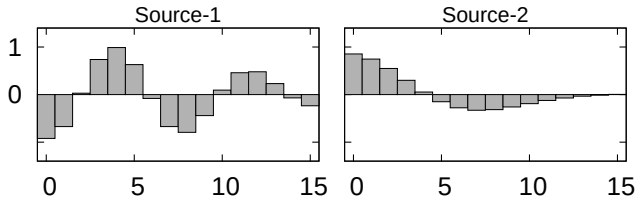
where $(\boldsymbol{\psi}_t)_k = \mathrm{p}'((\hat{z}_t)_k)/\,\mathrm{p}((z_t)_k)$.

3.1.6 Example — The optimal model structure

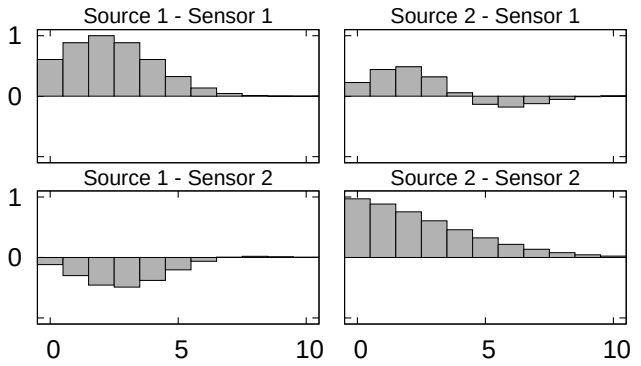
Two source signals are generated by taking two synthetic i.i.d. signals and filtering each of them using the respective filters shown on figure 3.3(a). This generates two independent and auto-correlated signals, and these are then mixed using the square system with $L = 10$ shown on figure 3.3(b). The generating model has thus $(L, M) = (10, 15)$.

First note that the generating model is in itself ambiguous; an arbitrary filter can be applied to a source model filter if the inverse of the arbitrary filter is applied to the respective column of mixing filters. Therefore, to compare results visually, each system of arbitrary dimension (L, M) must be visualized by its equivalent 'mixing only' system which has the dimensions $(L_{\mathrm{eq}}, M_{\mathrm{eq}}) = (L + M, 0)$. The equivalent system is found by convolving the source model filters with each of the filters in the corresponding column in the mixing model.

Figure 3.4 displays such equivalent mixing systems, i.e. where each mixing filter has been convolved with the respective source model filter. Figure 3.4(a) shows

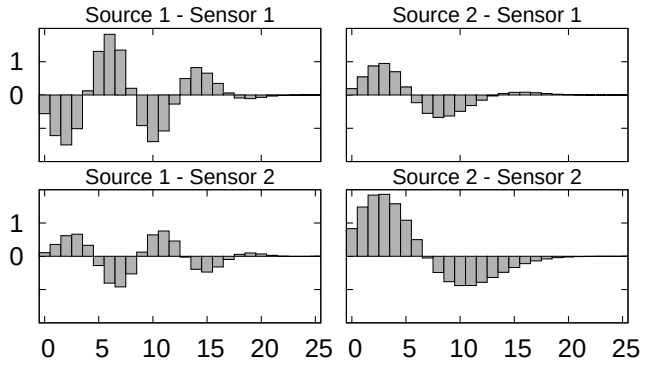


(a) Source model filters, $M = 15$.

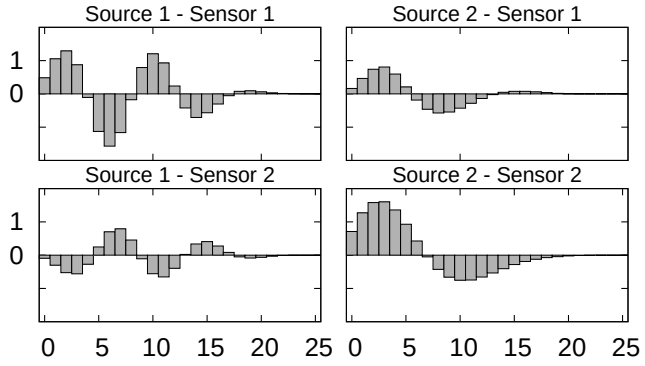


(b) Convolutive mixing system, $L = 10$.

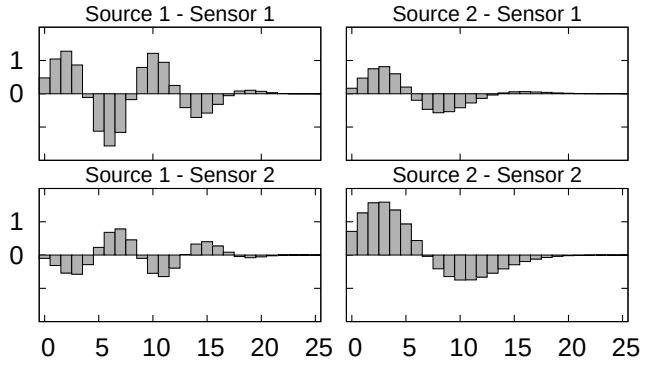
Figure 3.3: Filters for generating synthetic data. First, two i.i.d. signals are filtered through their respective filters shown in (a). Both filters are minimum-phase meaning that they can be perfectly inverted by (3.17). Then, the filtered signals are mixed using a distinct filter for each source-sensor path shown in (b). The mixing system shown in (b) is well-posed meaning that (3.6) is stable.



(a) Generating model $(L, M) = (10, 15)$



(b) Estimated by the algorithm $(L, M) = (10, 15)$



(c) Estimated by algorithm $(L, M) = (5, 20)$

Figure 3.4: Mixing filters convolved with respective source model filters. (a) for the generating model. (b) for an estimated model with the 'true' L and M. Clearly, the algorithm has succesfully identified the situation. (c) for the Bayes optimal model with $(L, M) = (5, 20)$. This is a more economic representation than the generating model, still it clearly resembles the true situation to great accuracy.

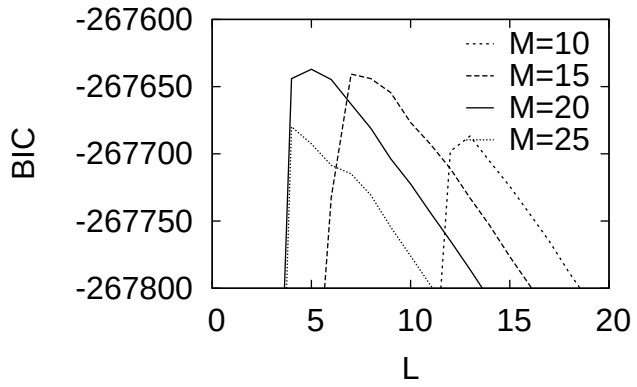


Figure 3.5: The BIC for various combinations of L and M. The true generating model was $(L, M) = (10, 15)$, but here $(L, M) = (5, 20)$ is found optimal. The optimal model has fewer parameters than the true model, still it resembles the response of the true model as is illustrated in figure 3.4.

the equivalent for the true generating model shown in figure 3.3; Figure 3.4(b) shows the equivalent for a run with the algorithm using $N = 300000$ training samples and using the (L, M) of the generating model. The result is perfect up to sign and scaling ICA ambiguities; Figure 3.4(c) shows the equivalent for a run with the algorithm using $N = 100000$ and the Bayes optimal choice of $(L, M) = (5, 20)$ which is found by monitoring Bayes Information Criterion (BIC, [54]), see figure 3.5 and refer to appendix B for a description of BIC in the context of the CICAAR algorithm. In the finite data, BIC has found a model with an equivalent transfer function that resembles that of the generating model (compare figure 3.4(a) with figure 3.4(c)), but using fewer parameters than in the generating model.

This finding is further underlined by studying learning curves, i.e. how does the training set dimension N influence learning. The likelihood evaluated on a test set is used to measure the learning of different models. Three models are now up for comparison; one which is the generating model $(L, M) = (10, 15)$, one $(L, M) = (25, 0)$ which is more complex but fully capable of imitating the first model, and $(L, M) = (5, 20)$ which is the BIC optimal choice. Figure 3.6 shows learning curves of the three models, the test set is $N_{\text{test}} = 300000$ samples. The uniform improvements in generalization of the 'optimal model' further underlines the importance of model selection in the context of convolutive mixing.

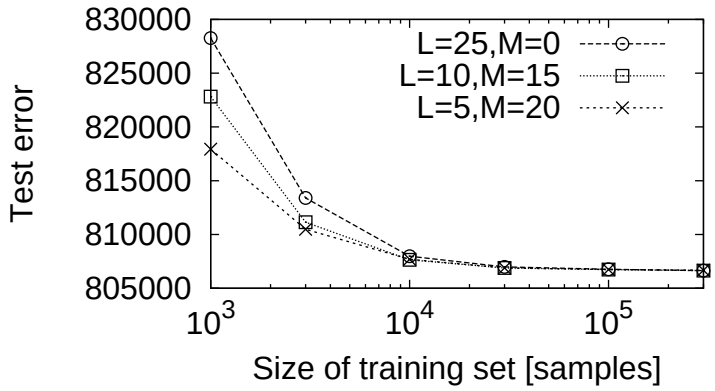


Figure 3.6: Learning curves for three models: The generating model $(L, M) = (10, 15)$, a model with $(L, M) = (25, 0)$ which is more complex but fully capable of 'imitating' the first model, and the model $(L, M) = (5, 20)$ which was found Bayes optimal according to BIC. The generalization error is estimated as the likelihood of a test set ($N_{\text{test}} = 300000$). The uniform improvements in generalization of the 'optimal model' further underlines the importance of model selection in the context of convolutive mixing.

3.2 Protocol for selecting L and M

A simple protocol is now proposed for determining the dimensions (L, M) of the mixing model and source model. First, expand the mixing model L while keeping $M = 0$, and find the optimal L by monitoring BIC. This will model the total temporal dependency structure of the system. From here on the optimal L is termed $L_{\max}$. Next, expand the order M of the source model while keeping $L + M = L_{\max}$; finding the optimal (L, M) by monitoring BIC. This will move as much correlation as possible from the mixing model to the source model.

3.2.1 Example — Detecting a convolutive mixture

This example is designed to illustrate the protocol, and to illustrate the importance of the source model when dealing with the following fundamental question: 'Is there evidence in the data for using convolutive ICA instead of instantaneous ICA?'. Detecting the order of L holds the answer to that question. In the framework of Bayesian model selection, models that are immoderately complex are penalized by the Occam factor, and will therefore only be chosen if there is a relevant need for their complexity. However, this compelling feature can be

disrupted if fundamental assumptions are violated, and the analyst must be extra careful when claiming an answer to a question like the above. One such assumption is involved in the derivation of the likelihood without the source model. The problem is that the likelihood will favor models based not only on achieved independence but on source whiteness as well. A model selection scheme for L which does not take the source auto-correlations into account will therefore be biased upwards because models with a larger value for L can absorb more source auto-correlation than models with lower L values. The cure to this problem is to invoke the source auto-correlation model of section 3.1.4.

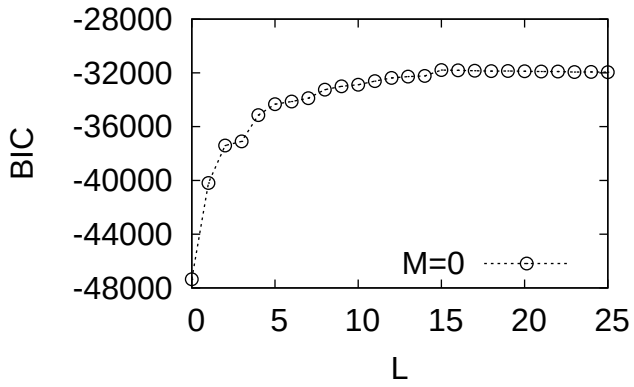
An *instantaneous* mixture is now produced by mixing the two auto-correlated sources from section 3.1.6 with a random matrix. The data thus holds correlations, but the mixing model is instantaneous and there should be no evidence for using convolutive ICA instead of instantaneous ICA.

First step in the protocol is to keep $M = 0$. Figure 3.7(a) shows the result of using Bayesian model selection *without* the source model ($M = 0$). Since the signals are auto-correlated, the model BIC simply increases as function of L up to the maximum which is attained at a value of $L_{\max} = 15$.

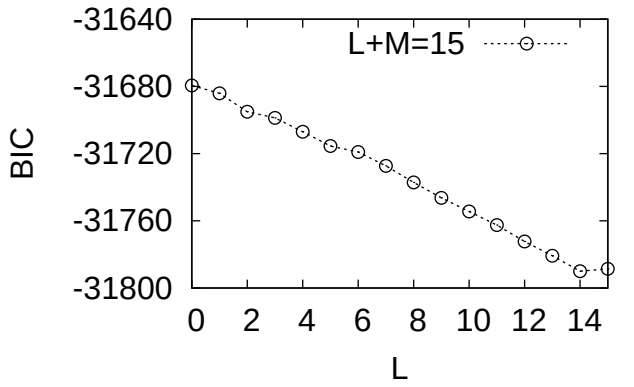
The next step in the protocol is to invoke the source model, increasing M while keeping $L + M = 15$ fixed. Figure 3.7(b) shows that lower L are preferable (because the models has fewer parameters while still explaining the same amount of temporal dependencies in the data). Thus, thanks to the source model, the correct answer is obtained: L should be zero — 'there is *no* evidence of convolutive ICA'!

3.3 Likelihood for overdetermined mixing

The likelihood has yet only been derived for square mixing. However, the overdetermined case, where the number of sources is strictly less than the number of sensors ($K < D$), is often relevant in practice. For instance, current EEG experiments typically involve simultaneous recording from 30 to 100 or more electrodes, forming a high (D) dimensional signal. After signal separation the hope could be to find a relatively small number (K) of independent components. In line with the square CICAAR which was derived for $K = D$, this section describes the 'rectangular' CICAAR which is derived for overdetermined mixing. In the following derivation of the likelihood, it is assumed that the number of convolutive source processes does not exceed the dimension of the data, i.e. $K \leq D$.



(a) $M = 0$



(b) $M + L = 15$

Figure 3.7: (a) the result of using Bayesian model selection without allowing for a filter ($M = 0$). Since the signals are auto-correlated L is detected at a value of $L = 15$. (b) fix $L + M = 15$, and now get the correct answer: $L = 0$ — 'There is no evidence of convolutive ICA' !

Assuming independent and identically distributed (i.i.d.) sources and no noise, the likelihood for the model (2.1) is

$$l(\{\mathbf{A}_\tau\}) = \int \cdots \int \prod_{t=1}^{N} \delta(\mathbf{e}_t)\, \mathrm{p}(\mathbf{s}_t)\, \mathrm{d}\, \mathbf{s}_t \tag{3.26}$$

where

$$\mathbf{e}_t = \mathbf{x}_t - \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \tag{3.27}$$

and $\delta(\mathbf{e}_t)$ is the Dirac delta function.

First, note that only the N'th term under the product operator in (3.26) is a function of $\mathbf{s}_N$. Hence, the $\mathbf{s}_N$-integral may be evaluated first, using (M.7) it yields

$$l(\{\mathbf{A}_\tau\}) = |{\mathbf{A}_0}^{\mathrm{T}} \mathbf{A}_0|^{-1/2} \int \cdots \int \mathrm{p}(\hat{\mathbf{s}}_N) \prod_{t=1}^{N-1} \delta(\mathbf{e}_t)\, \mathrm{p}(\mathbf{s}_t)\, \mathrm{d}\, \mathbf{s}_t \tag{3.28}$$

where the remaining integrals are over all sources except $\mathbf{s}_N$, and

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^{+} \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \mathbf{u}_{t-\tau} \right) \ , \ \mathbf{u}_n \equiv \begin{cases} \mathbf{s}_n \text{ for } n < N \\ \hat{\mathbf{s}}_n \text{ for } n \geq N \end{cases} \tag{3.29}$$

Now, as before, only one of the factors under the product operator in (3.28) is a function of $\mathbf{s}_{N-1}$. Hence, the $\mathbf{s}_{N-1}$-integral can now be evaluated, yielding

$$l(\{\mathbf{A}_\tau\}) = |{\mathbf{A}_0}^{\mathrm{T}} \mathbf{A}_0|^{-1} \int \cdots \int \mathrm{p}(\hat{\mathbf{s}}_N)\, \mathrm{p}(\hat{\mathbf{s}}_{N-1}) \prod_{t=1}^{N-2} \delta(\mathbf{e}_t)\, \mathrm{p}(\mathbf{s}_t)\, \mathrm{d}\, \mathbf{s}_t \tag{3.30}$$

where the remaining integrals are over all sources except $\mathbf{s}_N$ and $\mathbf{s}_{N-1}$, and

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^{+} \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \mathbf{u}_{t-\tau} \right) \ , \ \mathbf{u}_n \equiv \begin{cases} \mathbf{s}_n \text{ for } n < N-1 \\ \hat{\mathbf{s}}_n \text{ for } n \geq N-1 \end{cases} \tag{3.31}$$

By induction, and assuming $\mathbf{s}_n$ is zero for $n < 1$, the result is finally

$$l(\{\mathbf{A}_\tau\}) = |{\mathbf{A}_0}^{\mathrm{T}} \mathbf{A}_0|^{-N/2} \prod_{t=1}^{N} \mathrm{p}(\hat{\mathbf{s}}_t) \tag{3.32}$$

where

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^{+} \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \hat{\mathbf{s}}_{t-\tau} \right) \tag{3.33}$$

Thus, the likelihood is calculated by first *un*mixing the sources using (3.33), then measuring (3.32).

3.3.1 Computing the gradient

The gradient of the cost-function is presented here in two parts. Part one reveals the gradient of the source estimates while part two uses the result of part one to compute the gradient of the negative log likelihood. Differentiation w.r.t. a Moore-Penrose inverse matrix is described in appendix M.

Part one — Partial derivatives of the unmixed source estimates

The partial derivatives which shall be used in part two are given by

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_0^+)_{ij}} = \delta(i-k)\left(\mathbf{x}_t - \sum_{\tau=1}^{L}\mathbf{A}_\tau\hat{\mathbf{s}}_{t-\tau}\right)_j - \left(\mathbf{A}_0^+\sum_{\tau=1}^{L}\mathbf{A}_\tau\frac{\partial\hat{\mathbf{s}}_{t-\tau}}{\partial(\mathbf{A}_0^+)_{ij}}\right)_k \tag{3.34}$$

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} = -(\mathbf{A}_0^+)_{ki}(\hat{\mathbf{s}}_{t-\tau})_j - \left(\mathbf{A}_0^+\sum_{\tau'=1}^{L}\mathbf{A}_{\tau'}\frac{\partial\hat{\mathbf{s}}_{t-\tau'}}{\partial(\mathbf{A}_\tau)_{ij}}\right)_k \tag{3.35}$$

Part two — Gradient of the cost-function

The gradient of the negative log likelihood with respect to $\mathbf{A}_0^+$ is given by

$$\frac{\partial - \log l(\{\mathbf{A}_\tau\})}{\partial(\mathbf{A}_0^+)_{ij}} = -N(\mathbf{A}_0{}^{\mathrm{T}})_{ij} - \sum_{t=1}^{N}\boldsymbol{\psi}_t{}^{\mathrm{T}}\frac{\partial\hat{\mathbf{s}}_t}{\partial(\mathbf{A}_0^+)_{ij}} \tag{3.36}$$

where $(\ \boldsymbol{\psi}_t\ )_k = \mathrm{p}'(\ (\hat{\mathbf{s}}_t)_k\ )/\,\mathrm{p}(\ (\hat{\mathbf{s}}_t)_k\ )$. The gradient with respect to the other mixing matrices is

$$\frac{\partial - \log l(\{\mathbf{A}\})}{\partial(\mathbf{A}_\tau)_{ij}} = -\sum_{t=1}^{N}\boldsymbol{\psi}_t{}^{\mathrm{T}}\frac{\partial\hat{\mathbf{s}}_t}{\partial(\mathbf{A}_\tau)_{ij}} \tag{3.37}$$

3.3.2 The null-space problem

Even though the above derivation is valid for the overdetermined case ($D > K$), the validity of the zero-noise assumption proves vital in this case. The explanation for this can be seen in the definitions of the likelihood (3.32) and unmixing filter (3.33):

- In (3.32), note that rotation in the columnspace of $\mathbf{A}_0$ will not influence the determinant term of the likelihood. From (3.33) note that the estimated source vectors $\hat{\mathbf{s}}_t$ are found by linear mapping through $\mathbf{A}_0^+ : \mathbb{R}^D \mapsto \mathbb{R}^K$. Hence, the source-prior term in (3.32) alone will be responsible for determining a rotation of $\mathbf{A}_0$ that "hides" as much variance as possible in the null-space ($\mathbb{R}^{D-K}$) of $\mathbf{A}_0^+$ in (3.33). In an unconstrained optimization scheme, this side-effect will be untamed and consequently will hide data variance in the null-space of $\mathbf{A}_0^+$ and achieve an artificially high likelihood while relaxing the effort to make the sources independent.

3.4 Practical propositions for overdetermined convolutive ICA

It has just been argued that the rectangular CICAAR suffers from the null-space problem. Three ways of avoiding the null-space problem is now proposed:

1. (Residual cost term) Add a term to the cost function so that the model is punished for not explaining the data.

2. ('Augmented' configuration) Perform the decomposition with K set to D, i.e. attempting to estimate some extra sources.

3. ('Diminished' configuration) Perform the decomposition with D set to K, i.e. on a K-dimensional subspace projection of the data.

The first proposition, adding a residual cost term, would involve even more calculus. That trail stops here. The other two propositions, the augmented and diminished configurations, are more appealing because they are simple and practical approaches that use the square CICAAR. They will be described in the following.

3.4.1 Augmented configuration CICAAR

One solution to the null-space problem could be to parameterize the null-space of $\mathbf{A}_0^+$, or equivalently the orthogonal complement space of $\mathbf{A}_0$. This can be seen as a special case of the algorithm in which $\mathbf{A}_0$ is D-by-D and $\mathbf{A}_\tau$ is D-by-K. With the $D-K$ additional columns of $\mathbf{A}_0$ denoted by $\mathbf{B}$, the model can be

written

$$\mathbf{x}_t = \mathbf{B}\mathbf{v}_t + \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \tag{3.38}$$

where $\mathbf{v}_t$ and $\mathbf{B}$ constitute a low-rank approximation to the noise. The prior p.d.f. on $\mathbf{v}_t$ must be chosen so that large variances in that subspace become improbable. Here the proposed p.d.f. is a Gaussian. Note that (3.38) is a special case of the *square* convolutive mixing model. In this case, attempt to estimate the extra instantaneous noise sources in addition to the convolutive sources. The implementation is thus a special case of the square CICAAR, but with $D - K$ sources being instantaneous.

3.4.2 Diminished configuration CICAAR

An even simpler procedure is to project the data down to K dimensions and then use the regular square case CICAAR on the projection. This will extract K sources, and the overdetermined model can be obtained afterwards by solving the multivariate Wiener filter equation (2.2).

3.4.3 Example — Extracting fewer sources than sensors

In this example the performance of the rectangular CICAAR (suffering from the null-space problem), the augmented, and the diminished configurations are investigated as a function of signal-to-noise ratio (SNR). First, two synthetic i.i.d. source signals $s_1(t)$ and $s_2(t)$ (with $1 \leq t \leq N$ and $N = 30000$) were generated from a Laplace distribution, $s_k(t) \sim p(x) = \frac{1}{2}\exp(-|x|)$ with variance $\mathrm{var}\{s_k(t)\} = 2$. These signals were then mixed using the filters of length $L = 30$ shown in figure 3.8 producing an overdetermined mixture ($D = 3$, $K = 2$). A 3-D i.i.d. Gaussian noise signal $\mathbf{n}_t$ was added to the mixture $\mathbf{x}_t = \sigma\mathbf{n}_t + \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau}$ with a controlled variance σ^2. In the following, the rectangular, augmented, and diminished configurations are compared by how well they estimate the two sources by measuring the correlations between each true source signal, $s_k(t)$, and the best-correlated estimated source, $\hat{s}_{k'}(t)$. Figure 3.9 shows how well the sources were estimated at different SNR levels ($\mathrm{SNR} = 2/\sigma^2$).

Rectangular. All three data channels were decomposed using the rectangular CICAAR and the two true sources estimated. As shown in figure 3.9, the quality of the estimation using this configuration was the worst one out of the

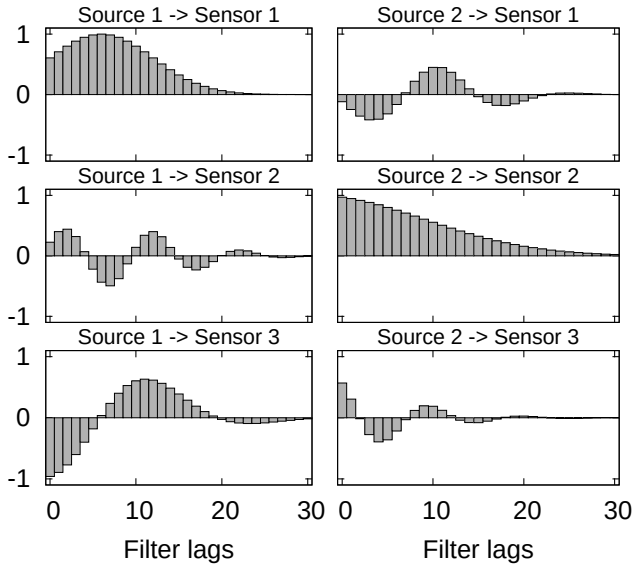


Figure 3.8: An overdetermined mixing system. This system is well-posed, meaning that the eigenvalues of (2.9) are situated within the unit circle and hence an exact and stable inverse exists in the sense of (3.33).

three configurations. But even though the rectangular CICAAR gives the worst source estimates, it has the highest (best) likelihood as is illustrated in figure 3.10. The figure compares the likelihood for the rectangular and augmented configurations since these two are given the exact same data as input.

Augmented. Figure 3.9 shows how well the sources were estimated using this configuration for different SNR levels. For the best estimated source (figure 3.9-A), the augmented configuration gave better estimates than the rectangular or diminished configurations. This was also the case for the second source (figure 3.9-B) at low SNR, but not at high SNR since in this case the 'true' $\mathbf{B}$ of (3.38) was near zero which is improbable under the likelihood. But, in the presence of considerable noise, the best separation was obtained by augmenting the model and extracting, from the D-dimensional mixture, K sources as well as a (rank $D-K$) approximation of the noise.

Diminished. To investigate the possibility of extracting the two sources from a two-dimensional projection of the data, the third 'sensor' was simply removed from the decomposition. Figure 3.9 shows that in the presence of considerable

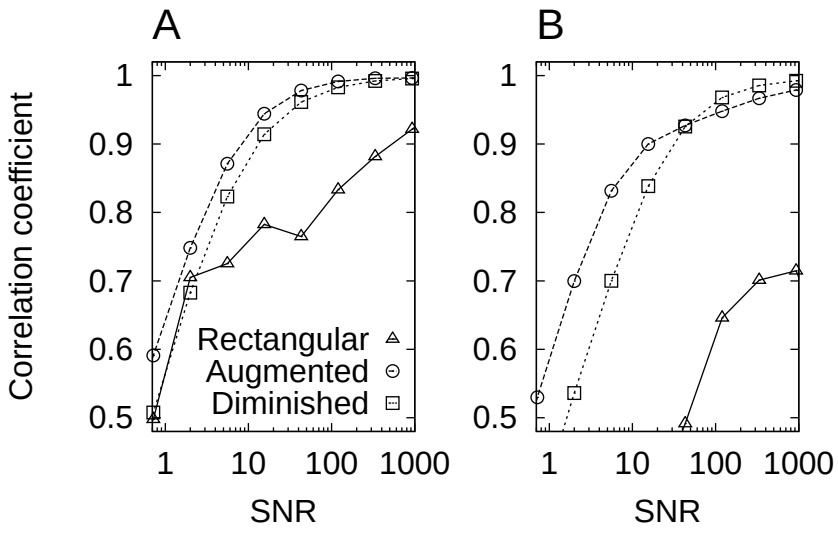


Figure 3.9: Comparison of source separation of the system in figure 3.8 using three CICAAR configurations (Rectangular, Augmented, Diminished). A: Estimates of true source activity: correlations with the best-estimated source. B: Similar correlations for the less well estimated source.

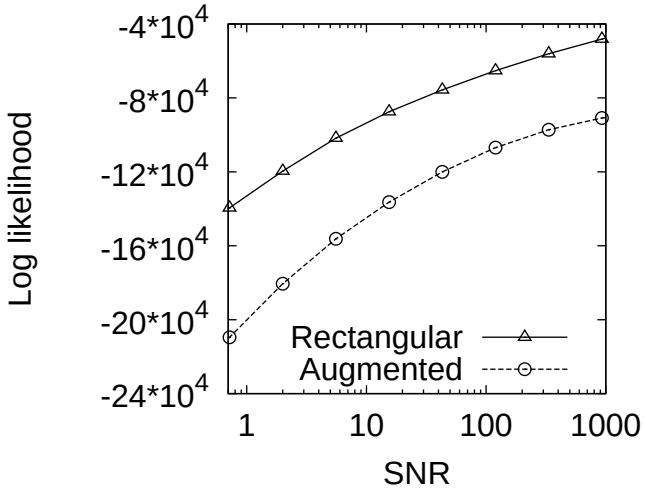


Figure 3.10: The rectangular CICAAR achieves an artificially high likelihood due to the null-space problem.

noise, the separation achieved was not as good as in the augmented configuration. However, the diminished configuration used the lowest number of parameters and hence had the lowest computational complexity, while the separation it achieved was close to that of the augmented configuration. At very high SNR, the diminished configuration was even slightly better than the augmented configuration.

Chapter 4

Algorithm II: CICAP

This chapter describes the CICAP algorithm for convolutive ICA. The derivation uses an approximation allowing the problem to be reduced to simple blind deconvolution based on second-order statistics followed by a linear mapping which can be identified using instantaneous ICA.

4.1 Linear prediction

The derivation takes its departure in assuming the existence of a multi-lag linear predictor of the form

$$\mathbf{x}_{t+\tau} = \sum_{\lambda=0}^{M} \mathbf{W}_{\tau,\lambda} \mathbf{x}_{t-\lambda} + \boldsymbol{\epsilon}_t(\tau) \tag{4.1}$$

where $\mathbf{W}_{\tau,\lambda}$ are the prediction parameter matrices and $\boldsymbol{\epsilon}_t(\tau)$ is the prediction error at prediction horizon τ. Now, in place of $\mathbf{x}_t$ substitute the convolutive model (2.1) to get

$$\sum_{\tau'=0}^{L} \mathbf{A}_{\tau'} \mathbf{s}_{t+\tau-\tau'} = \sum_{\lambda=0}^{M} \mathbf{W}_{\tau,\lambda} \sum_{\tau'=0}^{L} \mathbf{A}_{\tau'} \mathbf{s}_{t-\tau'-\lambda} + \boldsymbol{\epsilon}_t(\tau) \tag{4.2}$$

Now the second-order statistics come into play; multiply by $\mathbf{s}_t^{\mathrm{T}}$ from right and average assuming that the sources are temporally uncorrelated, i.e

$$< \mathbf{s}_{t+\tau}\mathbf{s}_t^{\mathrm{T}} >= \begin{cases} < \mathbf{s}_t\mathbf{s}_t^{\mathrm{T}} > & \text{for } \tau = 0 \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{4.3}$$

Then (4.2) yields

$$\mathbf{A}_\tau < \mathbf{s}_t\mathbf{s}_t^{\mathrm{T}} >= \mathbf{W}_{\tau,0}\mathbf{A}_0 < \mathbf{s}_t\mathbf{s}_t^{\mathrm{T}} > + < \boldsymbol{\epsilon}_t(\tau)\mathbf{s}_t^{\mathrm{T}} > \tag{4.4}$$

Assuming that all sources have non-vanishing variance

$$\mathbf{A}_\tau = \mathbf{W}_{\tau,0}\mathbf{A}_0 + < \boldsymbol{\epsilon}_t(\tau)\mathbf{s}_t^{\mathrm{T}} >< \mathbf{s}_t\mathbf{s}_t^{\mathrm{T}} >^{-1} \tag{4.5}$$

which elegantly expresses the relationship between mixing matrices and linear predictors. Inserting (4.5) into the convolutive mixing equation (2.1) yields

$$\mathbf{x}_t = \sum_{\tau=0}^{L} \left(\mathbf{W}_{\tau,0}\mathbf{A}_0 + < \boldsymbol{\epsilon}_t(\tau)\mathbf{s}_t^{\mathrm{T}} >< \mathbf{s}_t\mathbf{s}_t^{\mathrm{T}} >^{-1}\right) \mathbf{s}_{t-\tau} \tag{4.6}$$

Thus, the problem has been reduced to identifying the zero-lag mixing matrix $\mathbf{A}_0$, and the source variances, and the correlation sequences between prediction errors and sources. Further note that $\mathbf{W}_{0,0} = \mathbf{I}$ and $\boldsymbol{\epsilon}_t(0) = \mathbf{0}$, hence

$$\mathbf{x}_t = \mathbf{A}_0\mathbf{s}_t + \sum_{\tau=1}^{L} \left(\mathbf{W}_{\tau,0}\mathbf{A}_0 + < \boldsymbol{\epsilon}_t(\tau)\mathbf{s}_t^{\mathrm{T}} >< \mathbf{s}_t\mathbf{s}_t^{\mathrm{T}} >^{-1}\right) \mathbf{s}_{t-\tau} \tag{4.7}$$

4.2 Prediction error approximation

The 'prediction error approximation' is now invoked,

$$< \boldsymbol{\epsilon}_t(\tau)\mathbf{s}_t^{\mathrm{T}} > \approx \mathbf{0} \tag{4.8}$$

see also figure 4.2. Now, (4.6) can be rewritten

$$\mathbf{x}_t = \sum_{\tau=0}^{L} \mathbf{W}_{\tau,0}\mathbf{A}_0\mathbf{s}_{t-\tau} \equiv \sum_{\tau=0}^{L} \mathbf{W}_{\tau,0}\mathbf{u}_{t-\tau} \tag{4.9}$$

i.e. a convolutive mixture of the $\mathbf{u}_t$'s with *known* mixing matrices. Solving (4.9) for $\mathbf{u}_t$ yields a classical MIMO (multiple input multiple output) deconvolution problem. As in the CICAAR, by eliminating $\mathbf{u}_t$ in (4.9) the 'naive' deconvolution filter is obtained, i.e.

$$\hat{\mathbf{u}}_t \equiv \mathbf{A}_0\mathbf{s}_t = \mathbf{W}_{0,0}^{-1}\left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{W}_{\tau,0}\hat{\mathbf{u}}_{t-\tau}\right) \tag{4.10}$$

which can potentially become unstable. A suggestion for regularized deconvolution with the Conjugate Gradients Least Squares algorithm (see e.g. [55, 22]) is given in the implementation which is described in section 4.3. Anyhow, assuming that $\hat{\mathbf{u}}_t$ is obtained by the MIMO deconvolution, the problem is now reduced to an instantaneous ICA problem since

$$\hat{\mathbf{u}}_t = \mathbf{A}_0 \mathbf{s}_t \tag{4.11}$$

thus the zero-lag mixing matrix $\mathbf{A}_0$ can be estimated using an appropriate algorithm for that sort of problem.

Finally — with the prediction error approximation — (4.5) can be rewritten

$$\mathbf{A}_\tau = \mathbf{W}_{\tau,0} \mathbf{A}_0 \tag{4.12}$$

suggesting that the remaining mixing matrices could be generated from the zero-lag mixing matrix using the linear predictor. In section 4.3, however, an alternative method is put forth because of practical reasons which will be mentioned there.

4.3 Implementation

This section describes the steps to convolutive ICA with the CICAP algorithm in more detail.

4.3.1 Step 1 — Estimating the linear predictor

The prediction matrices are estimated using least-squares, i.e. solving

$$< \mathbf{x}_{t+\tau} \mathbf{x}_{t-\delta}^{\mathrm{T}} > = \sum_{\lambda=0}^{M} \hat{\mathbf{W}}_{\tau,\lambda} < \mathbf{x}_{t-\lambda} \mathbf{x}_{t-\delta}^{\mathrm{T}} > \tag{4.13}$$

for $\hat{\mathbf{W}}_{\tau,\lambda}$ by matrix inversion. The linear system is not huge and can be solved using a direct method such as the Matlab 'backslash' operator. The correlation matrices in (4.13) are measured respecting data epoch boundaries.

4.3.2 Step 2 — Regularized deconvolution

Solving (4.9) for $\hat{\mathbf{u}}_t$ is equivalent to solving the potentially huge linear system

$$\mathbf{x} = \mathbf{T}\mathbf{u} \tag{4.14}$$

where $\mathbf{x}$ is a stacked column vector of observations $\mathbf{x}_t$ and $\mathbf{u}$ is a stacked column vector of righthand-side vectors $\mathbf{u}_t$ and

$$\mathbf{T} = \begin{bmatrix} \mathbf{W}_{0,0} & \mathbf{W}_{1,0} & \cdots & \mathbf{W}_{L,0} & & \\ & \mathbf{W}_{0,0} & \mathbf{W}_{1,0} & \cdots & \mathbf{W}_{L,0} & \\ & & & \ddots & & \\ & & & & & \mathbf{W}_{0,0} \end{bmatrix} \tag{4.15}$$

which is in likely to be ill-posed, meaning that an exact solution is impossible to obtain in practice, as the norm of the solution diverges, see e.g. [22].

The problem of solving ill-posed inverse problems can be approached by optimization of the lagrangian function (see e.g. [18, 22])

$$\min_{\mathbf{u}} \|\mathbf{T}\mathbf{u} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{u}\|_2^2 \tag{4.16}$$

the normal equation formulation of this problem is

$$(\mathbf{T}^{\mathrm{T}}\mathbf{T} + \lambda \mathbf{I})\mathbf{u} = \mathbf{T}^{\mathrm{T}}\mathbf{x} \tag{4.17}$$

and the solution to that problem is known as the 'Tikhonov' regularized solution. Since the system is probably huge, an iterative method for solving it is chosen. The Tikhonov solution can in principle be obtained through the Singular Value Decomposition (SVD) [22], and recent progress in the field of huge SVD's suggests that the problem could be solved using an off-the-shelve toolbox such as [23]. However, at the time of implementation another method was chosen, namely the Conjugate Gradients Least Squares (CGLS) algorithm for solving linear systems in the least squares sense. The CGLS, with early stopping, has regularization properties similar to Tikhnov regularization [22].

CGLS with early stopping

Translated from [22] and [55] to the present context, the CGLS algorithm for deconvolution is implemented like this:

1. Initialize
   (a) Initial guess $\mathbf{u}^{(0)} = \mathbf{0}$
   (b) The residual vector for the least squares problem $\mathbf{r}^{(0)} = \mathbf{x} - \mathbf{T}\mathbf{u}^{(0)} = \mathbf{x}$
   (c) The residual for the normal equations $\mathbf{d}^{(0)} = \mathbf{T}^{\mathrm{T}}\mathbf{r}^{(0)}$
2. Iterate, k denotes iteration number.

(a) $\alpha_k = \|\mathbf{T}^{\mathrm{T}}\mathbf{r}^{(k-1)}\|_2^2 / \|\mathbf{T}\mathbf{d}^{(k-1)}\|_2^2$

(b) $\mathbf{u}^{(k)} = \mathbf{u}^{(k-1)} + \alpha_k \mathbf{d}^{(k-1)}$

(c) $\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} - \alpha_k \mathbf{T}\mathbf{d}^{(k-1)}$

(d) $\beta_k = \|\mathbf{T}^{\mathrm{T}}\mathbf{r}^{(k)}\|_2^2 / \|\mathbf{T}^{\mathrm{T}}\mathbf{r}^{(k-1)}\|_2^2$

(e) $\mathbf{d}^{(k)} = \mathbf{T}^{\mathrm{T}}\mathbf{r}^{(k)} + \beta_k \mathbf{d}^{(k-1)}$

3. Stop when the number of iterations k exceeds some maximum $k_{\max}$, or when the residual drops below a certain threshold, $\|\mathbf{r}^{(k)}\|_2 < \xi \|\mathbf{r}^{(0)}\|_2$ [55].

In each iteration two matrix-vector products are performed; one matrix-vector product is $\mathbf{T}\mathbf{d}^{(k-1)}$ which is performed by the forward sweeping filter

$$(\mathbf{T}\mathbf{d}^{(k-1)})_t = \sum_{\tau=0}^{L} \mathbf{W}_{\tau,0} \mathbf{d}_{t-\tau}^{(k-1)} \tag{4.18}$$

and the other matrix-vector product is $\mathbf{T}^{\mathrm{T}}\mathbf{r}^{(k)}$ which is performed by the reverse sweeping filter

$$(\mathbf{T}\mathbf{r}^{(k)})_t = \sum_{\tau=0}^{L} \mathbf{W}_{\tau,0} \mathbf{r}_{t+\tau}^{(k)} \tag{4.19}$$

The CGLS algorithm is thus easily implemented but perhaps less easy to understand. Refer to [55] for a good introduction. A short intuitive explanation of the CGLS is: It uses multiplication with $\mathbf{T}^{\mathrm{T}}$ and $\mathbf{T}$, then updates the estimated solution in each iteration. The larger eigenvalues of $\mathbf{T}$ are thus mainly contributing to the solution in the early iterations, while the smaller eigenvalues need more iterations for their contribution to take effect in the solution. By limiting the number of iterations (early stopping) the solution is thus mainly flavored by the larger eigenvalues of $\mathbf{T}$, thus the stopping threshold ξ (see the CGLS above) effectively works as a regularization parameter.

4.3.3 Step 3 — Instantaneous ICA

In this step, the source signals are extracted from an instantaneous mixture. Choosing an algorithm for doing so involves the typical considerations about source kurtosis etc. Here, the Infomax algorithm [5] is chosen in the context of EEG, see also [32].

4.3.4 Step 4 — Re-estimating the mixing matrices

Since $\hat{\mathbf{u}}$ in general has to be is estimated using regularization, the generator equation (4.12) is not valid in practice. Instead, the mixing matrices can be estimated by solving the multivariate Wiener filter equation (2.2).

4.4 Example — Extracting two stationary sources from a well-posed mixture

The CICAP algorithm is now illustrated on a synthetic mixture of two source signals. Each source signal is generated by first drawing a signal of length $N = 30000$ i.i.d. from a Laplace distribution with variance 2, then raising each sample to the power of three. The source signals are then mixed using the square system shown on figure 4.1(a) generating the 2D mixture which is now subject to analysis with the CICAP algorithm.

The linear predictor matrices are then estimated as described above. To investigate the validity of the prediction error approximation, the prediction error is measured for every τ

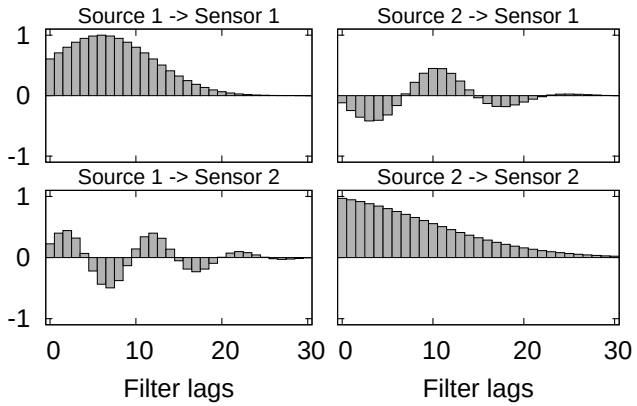
$$\boldsymbol{\epsilon}_t(\tau) = \mathbf{x}_{t+\tau} - \sum_{\lambda=0}^{M} \mathbf{W}_{\tau,\lambda}\mathbf{x}_{t-\lambda} \tag{4.20}$$

and then the relative prediction error

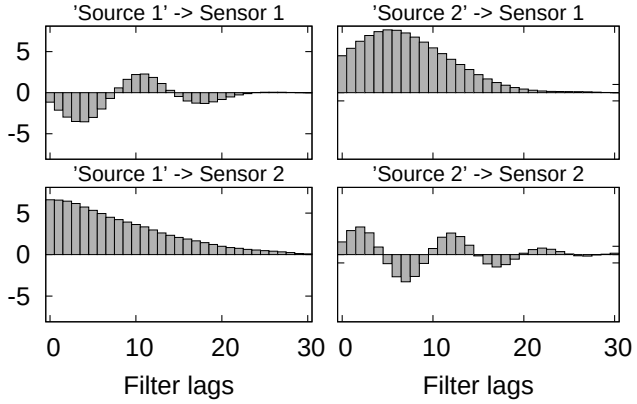
$$e(\tau) = \frac{\sum_{t=1}^{N} \boldsymbol{\epsilon}_t(\tau)^{\mathrm{T}}\boldsymbol{\epsilon}_t(\tau)}{\sum_{t=1}^{N} \mathbf{x}_t{}^{\mathrm{T}}\mathbf{x}_t} \tag{4.21}$$

As expected, the relative prediction error increases as a function of the prediction horizon τ as shown on figure 4.2. The correlation coefficient between the prediction error in one data channel and one of the sources is shown along with the prediction error in figure 4.2. The coefficient stays bounded as the prediction error increases, suggesting that the prediction error is uncorrelated with the sources. Thus, in this case the prediction error approximation turned out to be valid.

The next step in the CICAP algorithm is to deconvolve the data using CGLS. The regularization parameter was set to $\xi = 0.01$ and CGLS converged in 42 iterations. A scatter plot of the data channels and the deconvolved data is shown on figure 4.3(a) and figure 4.3(b) respectively.



(a) True mixing system. This system is well-posed, meaning that the eigenvalues of (2.9) are situated within the unit circle and hence exact inversion is possible through (3.6).



(b) estimated mixing system

Figure 4.1: CICAP estimation of the mixing system. The estimate is perfect up to scaling and permutation.

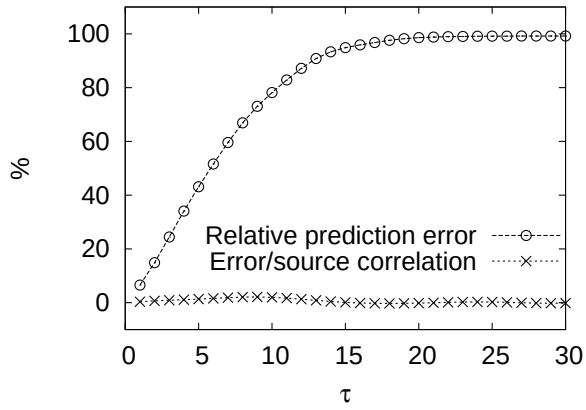
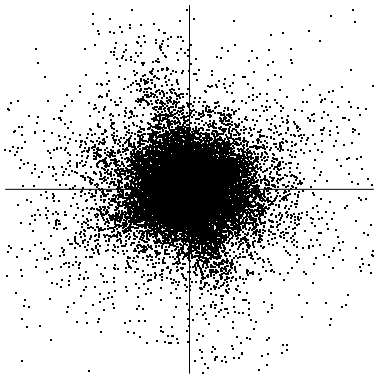


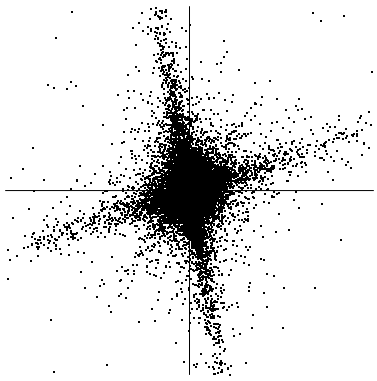
Figure 4.2: Validating the prediction error approximation. The correlation coefficient between the prediction error in one data channel and one of the sources is shown along with the prediction error. The relative prediction error increases as a function of prediction horizon τ, but the error/source correlation stays bounded. This suggests that the prediction error is uncorrelated with the sources, thus, in this case the prediction error approximation turned out to be valid.

The next step is to use ICA to estimate a linear mapping of the deconvolved data which will generate the independent sources. A scatter plot of the estimated sources is shown in figure 4.3(c).

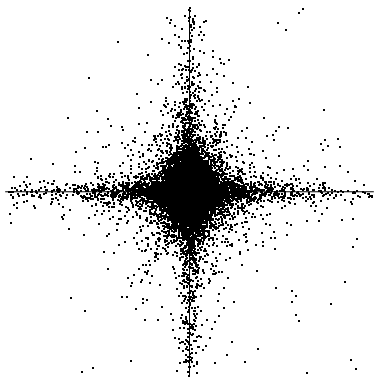
Finally, the mixing matrices are estimated using the method described in section 4.3.4 and the result is shown in figure 4.1(b) which is very similar to the true mixing system except for a scaling/permutation.



(a) scatter plot of the data channels



(b) scatter plot of the deconvolved data



(c) scatter plot of the unmixed sources

Figure 4.3: The CICAP algorithm solves the problem by deconvolution (a) → (b); Then unmixing is achieved by Infomax ICA (b) → (c).

Chapter 5

Comparative evaluation of algorithms

Comparison of algorithms for convolutive ICA is a research topic with many challenging aspects in itself. Most research in this field has addressed the acoustic problem of separating sound sources in a reverberant environment. In acoustic situations, the aim is typically to extract source signals from the mixture such that the extracted signals do not interfere with each other in the audible sense. Different interference measures has been proposed and typically these measures involve benchmark data sets including the 'true' sources for comparison, see e.g. [53, 1, 58].

The algorithm evaluations that are about to be made in this chapter are not about optimal separation of speech and music sources from data. To separate sources perfectly from acoustic data would require the model order to be so large that entire room reverberations could be explained by the model. The purpose of this chapter is instead to give a fair treatment of some fundamental properties of the CICAAR and CICAP algorithms. These will be highlighted in mutual contrast, and in contrast to another algorithm namely the algorithm proposed by Lucas Parra and Clay Spence in [44]. The 'Parra' algorithm represents a state-of-the-art algorithm for acoustic blind source separation, and in the remainder of this chapter it is chosen as representative for FFT based algorithms with FIR unmixing in general. The implementation of

the Parra algorithm was kindly provided by Stefan Harmeling and is available at `http://ida.first.gmd.de/~harmeli/download/download_convbss.html`

5.1 Non-stationary audio

The Parra algorithm is known to work efficiently with acoustic data including speech sources. Therefore the first choice of data for the purpose of algorithm comparison is this:

- A 16kHz signal recorded indoor by two microphones. The two sound sources in the room was a male speaker counting from one to ten and a loud music source respectively. The microphones and the sources were located in the corners of a square. The signal is kindly provided by Dr. T-W. Lee, and is identical to the one used in [29].

The true sources are unknown, thus other means must be used to assessing the quality of separation...

5.1.1 A quality measure for unknown sources

The goal in convolutive ICA is (now stating it simply...) to decompose the data into sources while removing temporal dependencies between the sources up to lags of order L. The separation achieved in a convolutive ICA decomposition can therefore be assessed by measuring the temporal dependencies between sources up to lags of order L. A necessary (but not sufficient) condition for temporal independence up to lags of order L is uncorrelatedness up to lags of order L.

Define the 'crosstalk of prediction order R matrix'

$$(\mathbf{C}_R)_{ij} = \frac{\text{var } \hat{s}_i^j}{\text{var } \hat{s}_i^i} \quad \text{where} \quad \hat{s}_i^j(t) = \sum_{\tau=0}^{R} \hat{w}_\tau^{i,j} \hat{s}_j(t-\tau) \tag{5.1}$$

where $\{\hat{w}_\tau^{i,j}\}$ are estimated by solving the univariate case of the Wiener filter equation (2.2), i.e. minimizing

$$\{\hat{w}_\tau^{i,j}\} = \arg\min \sum_t [\ \hat{s}_i^j(t) - \hat{s}_i^i(t)\ ]^2 \tag{5.2}$$

Uncorrelatedness of the sources up to lags of order L would imply the crosstalk of prediction order $R = L$ matrix being the identity. The crosstalk matrix is

related to the statistics of Granger causality tests, see e.g. [19]. The 'crosstalk' measure is now defined as the maximal off-diagonal element in the crosstalk matrix, i.e.

$$\text{Crosstalk}_R \equiv \max_{i \neq j} (\mathbf{C}_R)_{ij} \tag{5.3}$$

The crosstalk says how much variance of one source can be explained linearly from the history of order R of another source — the maximal value over exclusive combinations of sources. Thus for any convolutive ICA decomposition, if the order (L) of the model is known, the quality of achieved separation can be assessed from the separated sources directly. A necessary condition for perfect separation in a convolutive mixture is the crosstalk of order $R = L$ being zero. For the non-stationary audio data used here, the crosstalk measure can thus be used for metering the performance of the CICAAR, the CICAP, and the Parra algorithms.

5.1.2 Assessing the implicit model order

On the other hand, the convolutive model order (L) is not always defined. For instance, the Parra algorithm defines the order (Q) of the *un*mixing FIR system instead. As a first step to assessing the implicit model order given some data and a set of separated sources, the convolutive model is estimated by solving the multivariate Wiener filter equation (2.2) for some L. To check that the order L of the multivariate Wiener filter is large enough, the residual of data channel d is measured

$$r_d(t) = x_d(t) - \sum_{\tau=0}^{L} \sum_k (\hat{\mathbf{A}}_\tau)_{dk} \hat{s}_k(t-\tau) \tag{5.4}$$

where $x_d(t)$ is the data in channel d, $\hat{s}_k(t)$ is estimated source number k. Finally the 'leftovers' measure is hereby defined as the largest relative channel residual, for the model of order L

$$\text{Leftovers}_L \equiv \max_d \frac{\text{var } r_d}{\text{var } x_d} \tag{5.5}$$

Thus, estimating the implicit convolutive model for a given set of separated sources involves estimation of the mixing matrices using the multivariate Wiener filter equation (2.2), and checking of the model order by measuring the leftovers.

5.1.3 Evaluating CICAAR for $L = 50$

The CICAAR algorithm was applied to the data using $L = 50$. There are no other tunable parameters for the CICAAR algorithm. The resulting source

crosstalk and leftovers was

- $\text{Crosstalk}_{50} = 1.7\%$ @ $\text{Leftovers}_{50} = 0\%$

meaning that 1.7% of the variance of one source could be explained by the history of order 50 from the other source. For the square CICAAR algorithm the leftovers is always zero meaning that the sources and the model explain the data with zero residual.

5.1.4 Evaluating CICAP for $L = 50$

The CICAP algorithm was applied to the data using $L = 50$. For the CICAP algorithm there is another parameter that needs to be addressed here:

ξ — the regularization parameter, the higher the more regularization.

Figure 5.1(a) shows the leftovers that was measured for various values of ξ. If ξ is large (e.g. close to 1), it means that the CICAP algorithm is extremely regularized, and the figure shows how the solution did not explain the data when the algorithm was too regularized. Figure 5.1(b) shows the crosstalk for various values of ξ. Clearly, regularization was necessary as can be seen from the steep rise in the crosstalk for low values of ξ (little regularization). At $\xi = 0.00046$ the values

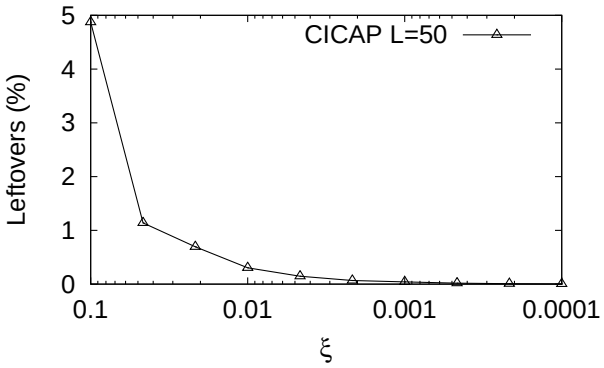
- $\text{Crosstalk}_{50} = 10.2\%$ @ $\text{Leftovers}_{50} = 0.02\%$

seemed to be optimal with the lowest crosstalk while explaining the data to great accuracy.

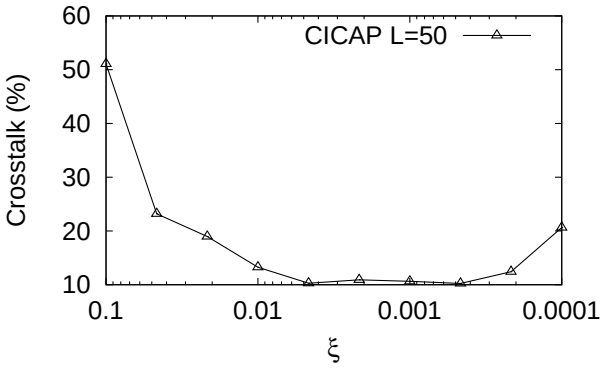
5.1.5 Evaluating Parra for $L \approx 50$

The model order is undefined for the Parra algorithm. Other tunable parameters to be addressed in this experiment were

T — the length of the FFT.

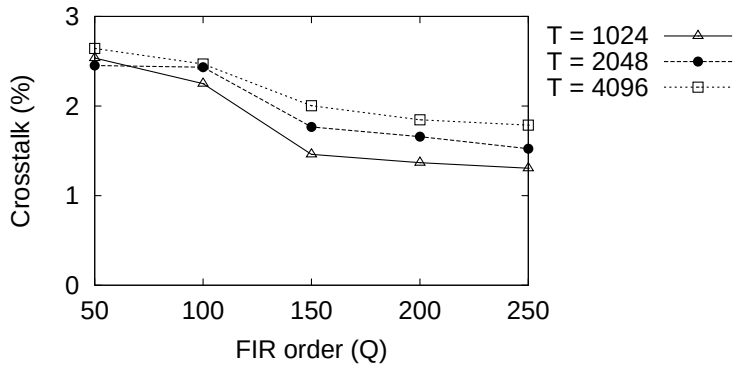


(a)

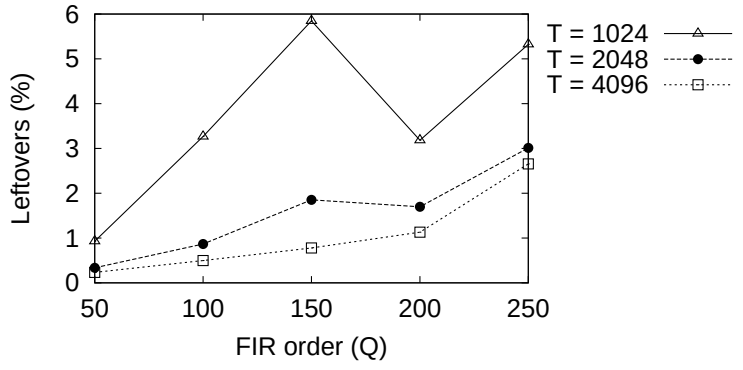


(b)

Figure 5.1: Evaluation of the CICAP algorithm at various values of the regularization parameter ξ; (a) Too much regularization results in a large model residual; (b) Too little regularization results in an increased crosstalk.



(a)



(b)

Figure 5.2: Crosstalk and leftovers of the Parra algorithm, for different combinations of FFT length T and FIR unmixing order Q.

Q — the order of the FIR unmixing system.

The maximum number of iterations was set so high that the effective stopping criterion was machine precision convergence in the cost function of the algorithm. Figure 5.2(a) shows the crosstalk measure assuming an implicit model order of $L = 50$ and for different combinations of FFT length T and FIR unmixing order Q. Figure 5.2(b) shows the leftovers correspondingly. The best separation (lowest crosstalk) was obtained with $T = 1024$ and $Q = 250$ obtaining

- $\text{Crosstalk}_{50} = 1.3\%$ @ $\text{Leftovers}_{50} = 5.3\%$

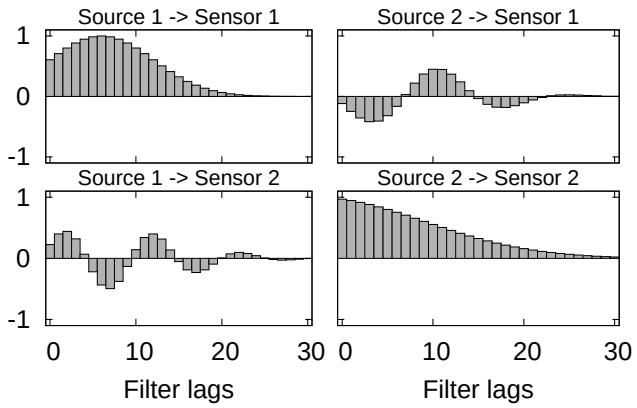


Figure 5.3: True mixing filters. This system is well-posed, meaning that the eigenvalues of (2.9) are situated within the unit circle and hence exact inversion is possible through (3.6).

However, the high leftovers value indicates that the implicit model order does not fit within the assumed model order of $L = 50$. The implicit model order of the algorithm in this instance must be somewhat larger than $L = 50$, thus comparison with the CICAAR and CICAP algorithms (for $L = 50$) would be unfair. Instead, a threshold value is now set for what is an acceptable leftovers; models with a higher leftovers than that is then rejected on the basis of not sticking to the assumed model order (L). Here, the threshold is arbitrarily set to 1%; then the best instance of the Parra algorithm was with $T = 4096$ and $Q = 150$ obtaining

- $\text{Crosstalk}_{50} = 2.0\%$ @ $\text{Leftovers}_{50} = 0.78\%$

5.2 Stationary white noise mixture

A mixture of stationary sources is now generated. Two sources are drawn i.i.d. from a Laplace distribution and mixed through the system shown on figure 5.3. The mixing system is well-posed.

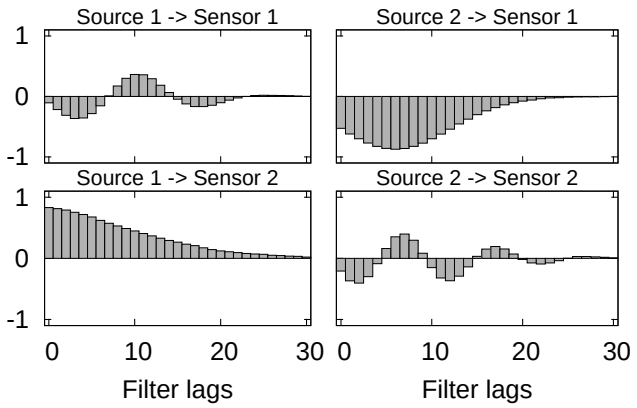


Figure 5.4: Estimated by the CICAAR algorithm. The CICAAR algorithm estimated the convolutive model with great accuracy.

5.2.1 Evaluating CICAAR, CICAP, Parra

The CICAAR algorithm estimated the convolutive model with great accuracy, see figure 5.4.

The CICAP algorithm estimated the convolutive model with great accuracy, see figure 5.5. The regularization parameter was set to a very small value of $\xi = 0.0001$ because the mixture was known to be well-posed.

The Parra algorithm failed to estimate the convolutive model, as could be expected due to its assumption that sources are non-stationary (as is the case for speech signals). Combinations of T and Q were tried on a grid of every combination of $Q \in [50, 100, 150, 200, 250]$ and $T \in [128, 256, 512, 1024, 2048, 4096]$. The maximum number of iterations was set so high that the effective stopping criterion was machine precision convergence in the cost function of the algorithm. An estimate is shown for $T = 256$ and $Q = 100$ in figure 5.6. The estimate shown there is typical for the experiment, and is effectively equivalent to a unit matrix mixing system.

5.3 Summary

The evaluations in this chapter was about removing temporal dependencies up to lags of an order L. To summarize the evaluation...

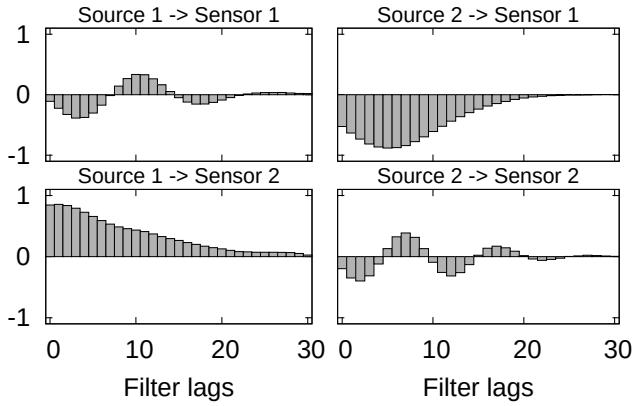


Figure 5.5: Estimated by the CICAP algorithm. The CICAP algorithm estimated the convolutive model with great accuracy, see figure 5.5. The regularization parameter was set to a very small value of $\xi = 0.0001$ because the mixture was known to be well-posed.

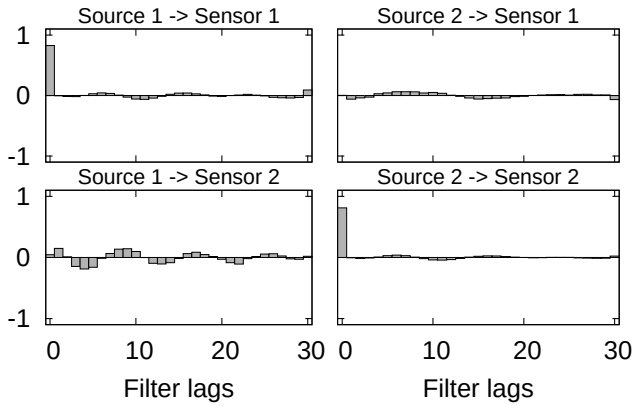


Figure 5.6: Estimated by the Parra algorithm. The estimate is shown for $T = 256$ and $Q = 100$, but many combinations of T and Q were tried. The estimate shown here is typical for the experiment, and is effectively equivalent to a unit matrix mixing system.

CICAAR In both non-stationary and stationary data, the CICAAR algorithm performed well without tuning any parameters.

CICAP In the stationary data, the performance of the CICAP algorithm was close to that of the CICAAR. In the non-stationary data however, the CICAP algorithm was inferior. The reason can be that the linear predictors are contaminated to some degree by the non-stationary correlations in the signal, and thus tampering with the algorithm in the deconvolution step.

Parra In the non-stationary data, the performance of the Parra algorithm was close to that of the CICAAR algorithm. Comparison was made possible by assessing the model order (L) though a measure of model data residual. In the stationary data, the Parra algorithm failed to produce a useful result.

Chapter 6

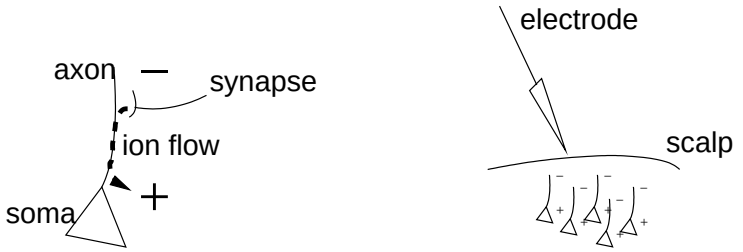
EEG physiology and ICA

This chapter defines what EEG is and deals with what is currently understood about ICA in EEG — not the complete reference, but a foundation for later chapters. The physiological statements made herein are based on [27, 28] (refer for further reading about EEG and physiology). Features of EEG that are relevant to understanding ICA decomposition of EEG are addressed here.

Figure 6.2, figure 6.3 and figure 6.4(a) were produced using the EEGLAB toolbox for Matlab, see [10], with the DIPFIT plug-in for dipole fitting and visualization by Robert Oostenveld, see also [52].

6.1 Dipoles — The physiological basis of EEG

Most neurons in the surface of the brain (Cerebral Cortex) are 'pyramidal cells'. A pyramidal cell has a body (the soma) and a single long nerve fiber (the axon) extending away from the body. The axon conducts electrical communication to and from the soma. When a cell receives 'excitatory stimulation' from other cells and reaches a certain threshold, the cell undergoes depolarization creating an 'action potential'. The action potential causes a flow of positively charged ions along the axon and generates a dipole with orientation along the axon as



(a) Single cell

(b) Many cells firing in synchrony

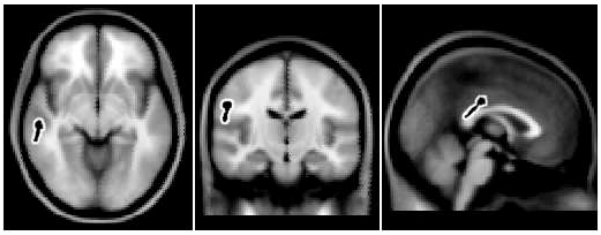
Figure 6.1: Each pyramidal cell generates an electromagnetic dipole when it fires. The activity of many cells firing in synchrony can reach the electrode.

illustrated in figure 6.1(a). The electroencephalogram (EEG) is the recording of brain activity using electrodes on the scalp. However, the potential generated by a single neuron is not strong enough to be picked up by an extra-cranial electrode. Instead, EEG electrodes can pick up potentials generated by larger groups of neurons firing in synchrony as illustrated in figure 6.1(b). On a local scale in Cerebral Cortex, pyramidal cells are very well aligned and highly connected, and the necessary synchrony is often in place. This makes it possible to pick up brain activity with EEG.

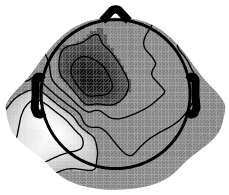
Cerebral Cortex is highly wrinkled with ridges (gurus) and fissure (sulcus). Therefore, a dipole can have any orientation relative to the scalp depending on where it sits in the brain and on whether it sits in a gyrus or in a sulcus. The topography of the voltage potentials generated by a dipole which is perpendicular to the scalp surface is unipolar, and the topography for a dipole which is not perpendicular to the scalp surface is in general bipolar. Figure 6.2 illustrates these two different dipole situations. In figure 6.2(a) a dipole is situated in a sulcus, and its orientation is parallel to the local scalp surface. The resulting topography is shown in figure 6.2(b). In figure 6.2(c) another dipole is situated on a gyrus, and its orientation is perpendicular to the local scalp surface. The resulting topography is shown in figure 6.2(d).

6.1.1 Topographic convention

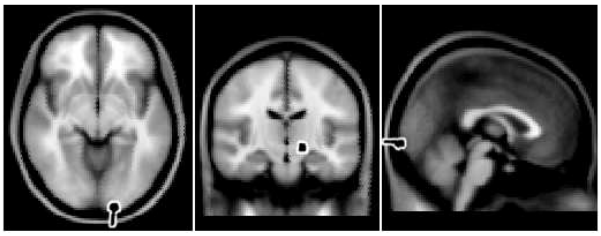
EEG is recorded using a finite number of electrodes. Figure 6.3 shows the individual positioning of 124 electrodes projected onto a cartoon head. Electrodes positioned 'below equator' on the real head are drawn outside the cartoon head. For visualization, scalp topography values are interpolated from the measure-



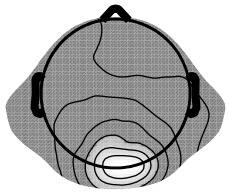
(a) Dipole near the left side of the skull.



(b) Scalp topography for the dipole in (a).

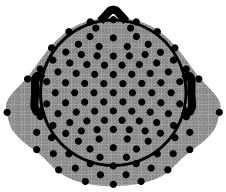


(c) Dipole situated in the back of the head. The orientation is perpendicular to the skull.

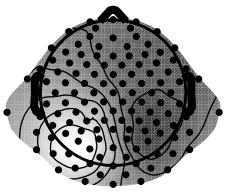


(d) Scalp topography for perpendicular dipole in (c).

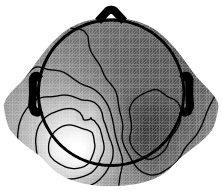
Figure 6.2: Two situations where a dipole is close to the skull. The local orientation of the dipole makes a big difference in the local scalp topography of the dipole.



(a) electrode positions



(b) electrode positions and interpolated topography



(c) interpolated topography

Figure 6.3: Placement of 124 electrodes on the scalp and visualization of interpolated topography. Electrodes positioned 'below equator' are drawn outside the cartoon head.

ments of surrounding electrodes. Brighter pixels represent higher numerical values than dimmer pixels.

In practice the EEG measurement readouts are relative to some electrical reference. Electromagnetic interference from the surrounding world is minimized by choosing a reference which is close to the EEG electrodes and also kept spatially fixed with the EEG electrodes. In a typical recording setup, all EEG electrodes share a common reference which could for instance be securely attached to an earlobe on the subject. EEG topographies are sensitive to the choice of electrical reference but can be re-referenced to another electrode, or to a group of electrodes, by linear mapping of the obtained recording. Several reference systems has been invented, but for the remainder of this thesis the choice of reference will be the 'average reference' which is the instant average potential of all electrodes. The reason for this choice is that the topographic response of a dipole will be fairly non-sensitive to its spatial orientation relative to the electrical reference.

6.2 Instantaneous ICA — A physiologically meaningful basis for EEG

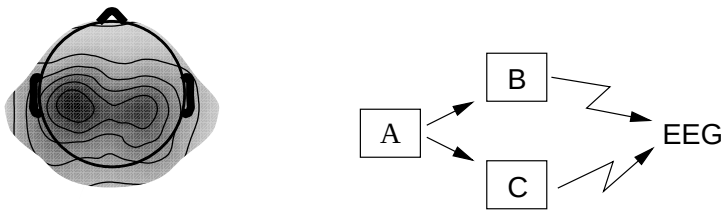
Denote by D the number of electrodes, and by $\mathbf{x}_t$ the D-dimensional vector of electrode potentials measured at time t.

First assume a simplistic world. In this world there is a number of dipoles inside the brain which constitute brain activity, a number of dipoles outside the brain constituting noise interference, and no other electromagnetic activity in this world. All dipoles are spatially fixed. Since the generated potentials from different dipoles are linearly and instantly added in the electrodes, the following linear model is valid

$$\mathbf{x}_t = \mathbf{G}\mathbf{y}_t \tag{6.1}$$

where $\mathbf{y}_t$ is a K-dimensional vector representing the amplitude of each dipole at time t, and $\mathbf{G}$ is a D-by-K matrix where each column represents the topography of its respective dipole configuration. In this world, assuming that the number of dipoles K does not exceed the number of electrodes D, it would be possible to isolate the contribution of each dipole to the electrodes by inverting $\mathbf{G}$. Thus, if the dipoles acted independently, Independent Component Analysis (ICA) would be a valid technique for estimating $\mathbf{G}$ in this simplistic world.

In practice it turns out that the dipoles act independently to some extend. As mentioned in section 6.1 dipolar activity is due to the high connectedness on a



(a) (b)

Figure 6.4: A bimodal dipolar ICA component. (a) spatial projection; (b) functional explanation: A implies B and C so that B and C are somewhat synchronized. To some extend B and C act together as a bimodal dipole configuration, which has been learned by ICA.

local scale in the Cerebral Cortex. On a longer scale however, dipoles are more loosely connected, and with significant communication delays. The implication is that ICA decomposition of EEG most often includes a number of components having 'dipolar' spatial projections to the sensors, see e.g. [36, 35, 26, 11, 33, 43]. In fact, the topographies shown in figure 6.2(b) and figure 6.2(d) were found by running ICA on an EEG data set[1]. These components are 'unimodal' dipolar components, and the dipole configurations in figure 6.2(a) and figure 6.2(c) were actually found by fitting a single dipole to the topographic maps, see [52, 60, 61].

Occasionally, components can exhibit 'bimodal' dipolar spatial projection. E.g. the spatial projection shown in figure 6.4(a) must be explained by (at least) two dipoles. The explanation for this must be that two dipoles at different locations in the brain act partially in synchrony as illustrated in figure 6.4(b) where the functional abstracts 'B' and 'C' are both responding to 'A' and act in synchrony.

6.2.1 Removing interferences by projection

The use of ICA for identifying and removing interference activity in EEG has been widely studied since the work of Makeig et al. in 1996 [32]. The work then was based on the standard Infomax ICA algorithm (from 1995 [5]) for sources with positive kurtosis, and Infomax ICA has turned out to be an effective tool for removing e.g. eye-blink interferences from EEG. This is simply done by removing artifact components from the ICA decomposition, hence obtaining a

[1]The data set was a 126 electrode recording during a pain experiment which was conducted at Center for Sensory-Motor Interaction, Aalborg University, Denmark.

projection of the data which is then clean from the artifacts. Other interferences such as 50Hz line noise with negative kurtosis can similarly be removed by using an 'extended ICA' algorithm which is again the Infomax ICA algorithm but allowing for sources with negative kurtosis [25, 30, 26].

6.2.2 Incurable artifacts

In the real world, electrodes come off, the electrode-to-skin impedance varies over time, or the subject wiggles his ears. The point is that (6.1) often breaks down in practice. In order to obtain a clean and useful ICA decomposition, epochs of data with such incurable artifacts are typically rejected from the data before ICA. Epoch rejection is typically performed by visual inspection of the data time series, or by simple automatic heuristics based on e.g. short time power, see e.g. [11].

Chapter 7

Convolutive ICA in EEG

In chapter 6 it was discussed how ICA works as a tool for finding a physiologically meaningful basis for EEG, i.e. finding independent components (ICs) with dipolar spatial projections. This chapter deals with the use of convolutive ICA on a subspace of such dipolar ICs in order to make them temporally independent up to lags of order L. Figure 7.1 illustrates that idea: The EEG data are decomposed into a mixing matrix ($\mathbf{A}^{\mathrm{ICA}}$) and 'activations' (the IC time series); A subspace is then chosen by picking out a few of the ICs, and the activations belonging to that subspace are then subject to convolutive ICA decomposition.

A key finding will is put forth in this chapter: It is shown that convolutive ICA is relevant for EEG. This finding is based on Bayesian model selection in a real EEG data set, and further elaborated by analysis of a convolutive decomposition in both time- and frequency domain.

7.1 Case study

The data set for this case study was 20 minutes of a 71-channel human EEG recording. The data set was obtained from the Schwartz Center for Computa-

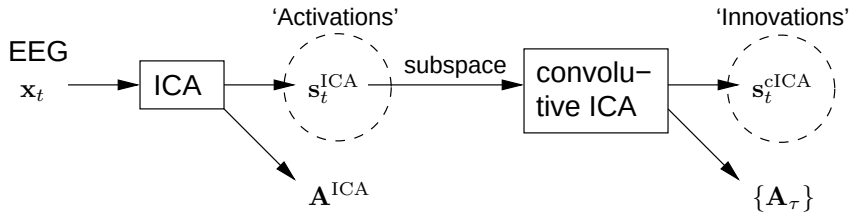


Figure 7.1: The idea of applying convolutive ICA to an ICA subspace of EEG data. The EEG data are decomposed into a mixing matrix ($\mathbf{A}^{\mathrm{ICA}}$) and activations. A subspace is then chosen by picking out a few of the ICs, and the activations belonging to that subspace are then subject to convolutive ICA decomposition. Resulting from the convolutive ICA decomposition are CCs, i.e. the estimated mixing filters $\{\mathbf{A}_\tau\}$ and the innovations.

tional Neuroscience[1], University of California, San Diego, USA. The electrode locations were digitized and their locations were as shown in figure 7.2. The subject was working on a 'letter two-back with feedback' memory task:

- Letters (A, B or C) were presented to the subject with a constant rate of two letters per three seconds. Each time a letter was presented, the task was to know whether that letter was the same as the letter that preceded the preceding letter (two-back). The subject would press a button to indicate his solution. If the answer was wrong, auditory feedback was given — a buzz — and the subject would then know that an error had been made. The feedback was given at a fixed latency of one second relative to the presented letter.

20 epochs were recorded, each of a duration of one minute, the sampling rate was 250Hz.

7.1.1 An ICA subspace

The recording was decomposed using extended Infomax ICA (see [25, 30, 26], c.f. section 6.2.1) into 71 independent components, i.e. assuming the model

$$\mathbf{x}_t = \mathbf{A}^{\mathrm{ICA}}\mathbf{s}_t^{\mathrm{ICA}} \tag{7.1}$$

For numeric convenience the data was (low-pass filtered and) downsampled to

[1] `http://sccn.ucsd.edu/`

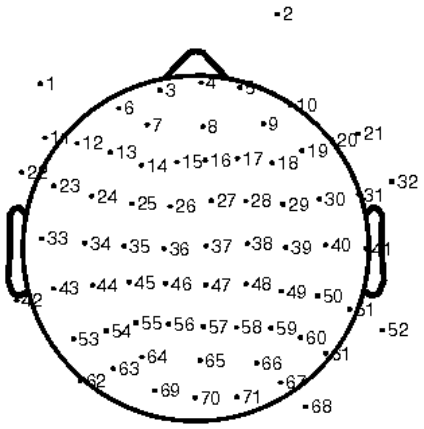


Figure 7.2: The 71 electrodes were digitized, their locations were as shown here.

50Hz sampling rate after filtering between 1 and 250 Hz using a 0.2Hz transition band FIR filter with zero phase shift.

Five of the resulting independent components (ICs) were selected for further analysis. These components were chosen because they showed event-related activity following the subject button presses [34]. Their scalp maps (from the relevant five columns of $\mathbf{A}^{\mathrm{ICA}}$) are shown on the left margin of figure 7.3. The five ERP-images[2] right next to the scalp maps in figure 7.3 are ERP-images of the ICA activations. The vertical line indicates when the feedback occurs in each trial, and the sigmoid curve indicates the latency of the button press. All trials have been sorted by the button press latency. Clearly, the activations for the five chosen ICs show activity following the button press, but the response is different between the five ICs.

The crosstalk matrix[3] of order $R = 0$ for the five ICs is shown in figure 7.4. The off-diagonal elements were very small, as expected, because the activations had been made maximally independent by the Infomax ICA algorithm. However, there were delayed temporal dependencies between the activations; this was evident from the crosstalk matrix of order $R = 10$, shown in figure 7.5, where some of the off diagonal elements were significantly larger than those in figure 7.4.

[2] See section E.2 for a description of ERP-images.
[3] See section 5.1.1 for the definition of the crosstalk matrix.

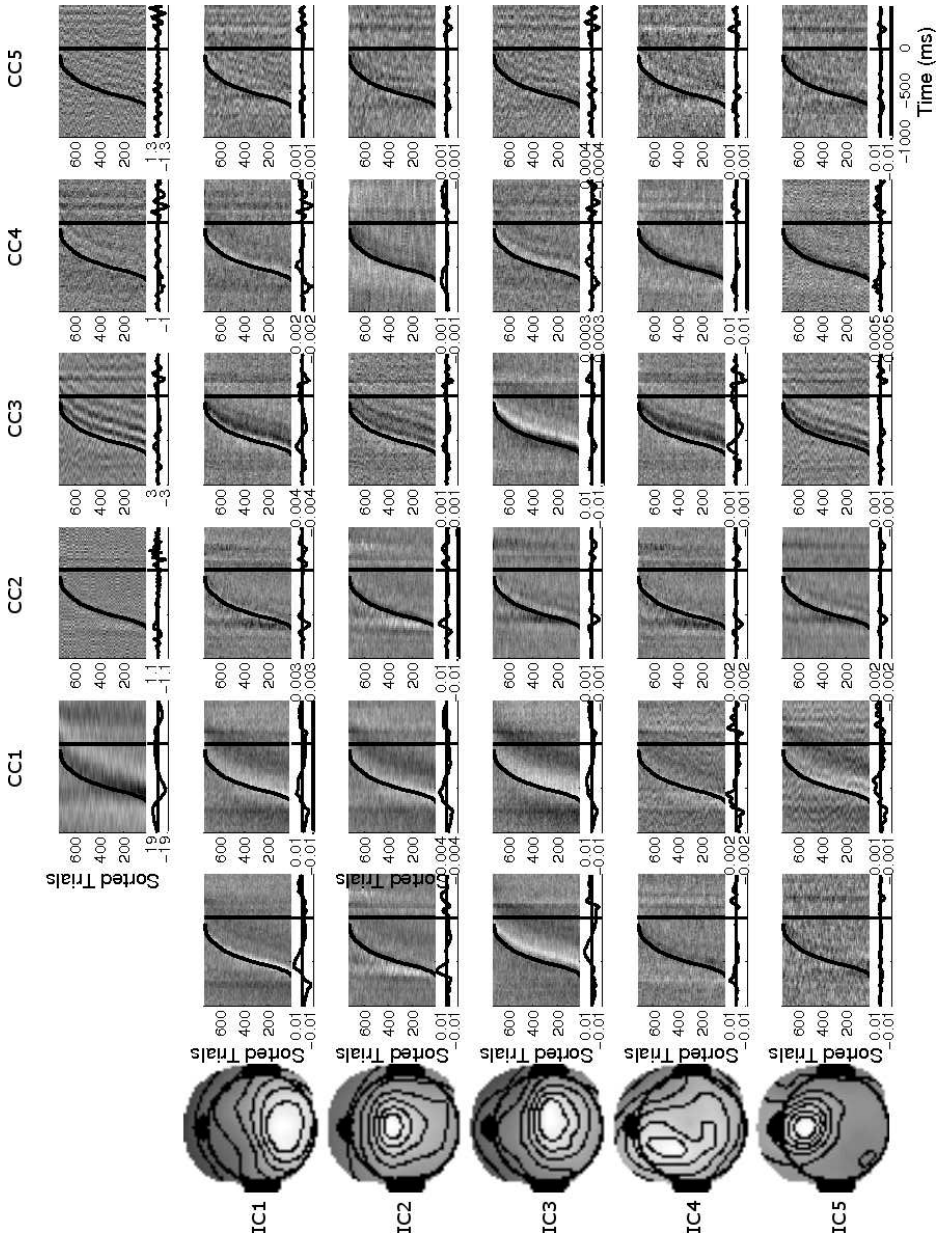


Figure 7.3: Erp images for the IC activations (on the left margin, towards this caption), for the CC innovations (top margin), and for the CC contributions to the IC activations (lower right 5×5). The scalp maps for the five ICs are shown along on the left margin. The color map scaling is individual for each ERP image. The vertical line indicates when the feedback occurs in each trial, and the sigmoid curve indicates the latency of the button press. All trials have been sorted by the button press latency.

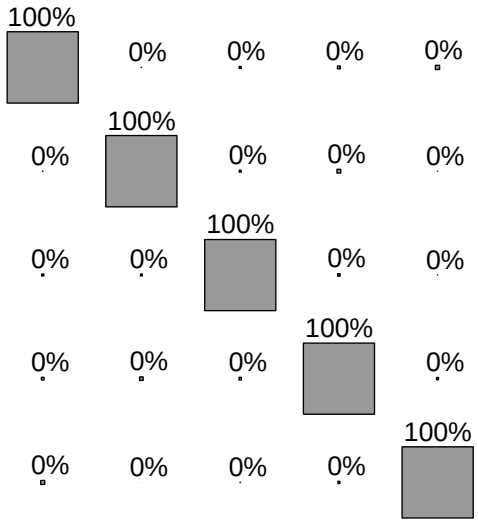


Figure 7.4: Crosstalk matrix of order $R = 0$, measured from the ICA activations. The squares have areas proportional to the respective crosstalk matrix elements. The percentage written above the squares have been rounded. The off-diagonal elements were very small, as expected, because the activations had been made maximally independent by the Infomax ICA algorithm.

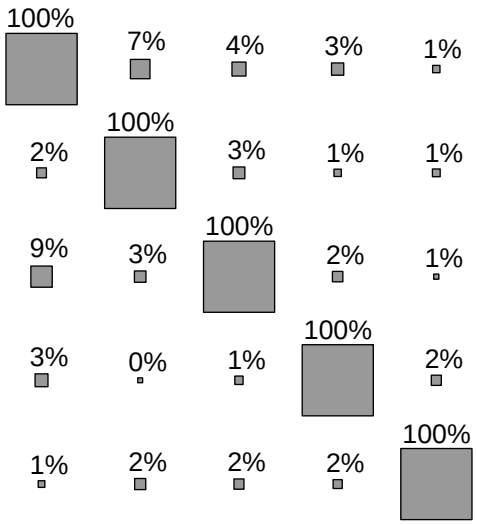


Figure 7.5: Crosstalk matrix of order $R = 10$, measured from the ICA activations. The squares have areas proportional to the respective crosstalk matrix elements. The percentage written above the squares have been rounded. Here some of the off diagonal elements were significantly larger than zero indicating that there were temporal dependencies between the activations when lags up to order $R = 10$ were considered.

7.1.2 Detecting the optimal convolutive model

To investigate whether convolutive ICA was relevant for the delayed temporal dependencies, convolutive ICA decomposition was applied to the five component activation time series, i.e. assuming the model

$$\mathbf{s}_t^{\mathrm{ICA}} = \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau}^{\mathrm{cICA}} \tag{7.2}$$

As described in section 3.2 a convolutive mixture can be detected using the CICAAR algorithm with the following protocol:

- First, increasing the order of the convolutive model L (keeping $M = 0$) while monitoring the BIC. To produce error bars, jackknife resampling was used [17]; i.e. for each value of L, 20 runs with the algorithm were performed, one for each jackknifed epoch, thus the data in each run consisted of the 19 remaining epochs. Figure 7.6A shows the jackknifed BIC. Clearly, the BIC was at least $L_{\max} = 40$, meaning that some correlations in the data extended to at least 800 ms.
- Next, the range of possible source model filters was swept, i.e. values of M while keeping $L + M = 40$. Figure 7.6B shows that the optimal convolutive model order L was significantly larger than zero, and hence convolutive ICA was indeed relevant. The optimal jackknifed BIC was $(L, M) = (10, 30)$.

This is indeed a key result — it can be concluded that convolutive ICA is relevant for EEG, and the temporal dependencies between the different activations are mainly to be modelled up to 10 lags ($\sim$ 200ms).

7.1.3 Exploring the optimal model, $(L, M) = (10, 30)$

The innovations were estimated using the AR inverse (3.6). Figure 7.7 shows the resulting percent of variance of the contributions from each of the innovations to each of the activations. Before plotting, the order of the five CCs was arranged so that the diagonal elements in the shown matrix were dominant. As the large diagonal contributions in figure 7.7 show, each CC dominated one IC. However, there were clearly significant off-diagonal contributions as well, indicating that each CC had captured some interaction between the ICs. Figure 7.8 shows the 5×5 matrix of learned convolutive mixing filters. The dominant diagonal filters are shown in one-third scale in figure 7.8.

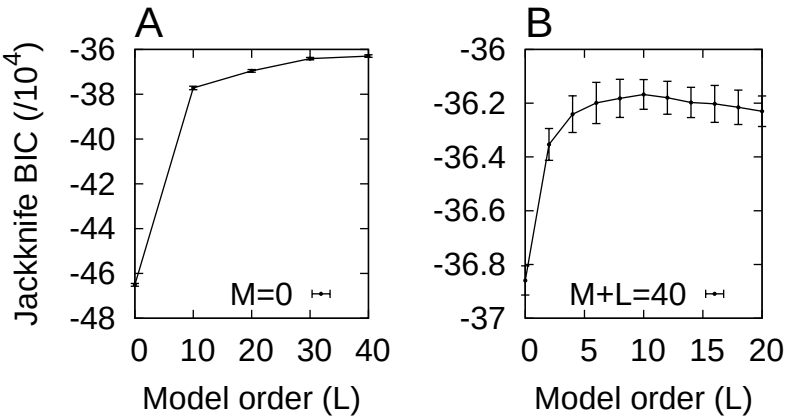


Figure 7.6: Jackknife estimated BIC for convolutive ICA, and jackknifed errorbars. (A) Clearly, the BIC was at least $L_{\max} = 40$, meaning that some correlations in the data extended to at least 800 ms. (B) The optimal convolutive model order L was significantly larger than zero, and hence convolutive ICA was indeed relevant. The optimal jackknifed BIC was $(L, M) = (10, 30)$.

Figure 7.9 shows the crosstalk matrix of order $R = 10$ for the innovations. The vanishing off-diagonal elements indicate that the convolutive ICA model of order $L = 10$ has successfully removed temporal dependencies up to order 10 between the innovations. This can be compared directly to the crosstalk matrix of order $R = 10$ measured for instantaneous ICA in figure 7.5; clearly convolutive ICA has successfully removed some temporal dependencies that were present in the activations to begin with. This finding was further elaborated. Figure 7.10 shows the crosstalk[4] for various prediction orders R for the IC activations and for the CC innovations. As expected from the previous results, as the prediction order R increased, the crosstalk of the instantaneous IC activations increased. E.g. for one of the IC activations, 9% of the variance could be explained by linear prediction from the previous 10 time points (200 ms) of another IC. For the CC innovations, however, the crosstalk in figure 7.10 remained low as the prediction order increased, indicating that convolutive ICA in fact deconvolved delayed correlations present in the EEG subspace data. For prediction orders greater than $R = 10$, the crosstalk for the CC innovations also increased because the convolutive model with $L = 10$ had only removed temporal dependencies up to order 10.

The five ERP-images on the top margin of figure 7.3 are made from the five CC innovations. The 5×5 matrix of ERP-images in the lower right area of figure

[4]See section 5.1.1 for the definition of crosstalk.

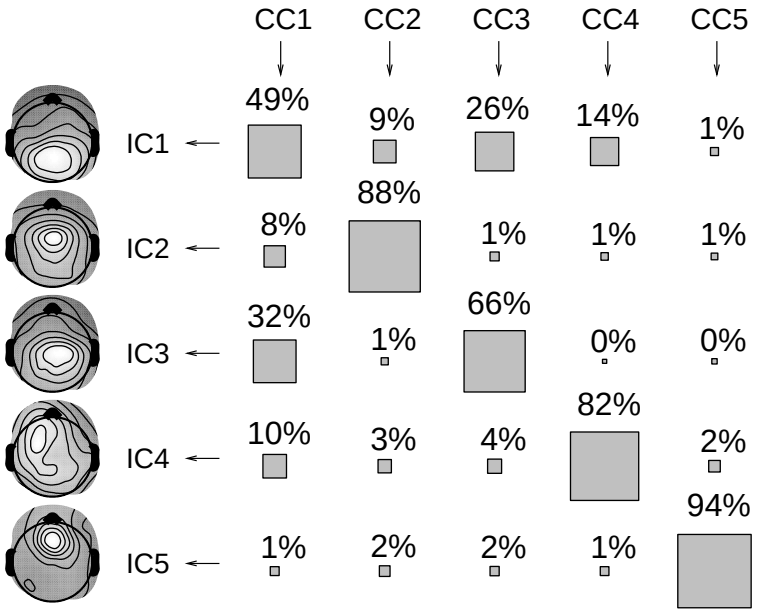


Figure 7.7: Percent variance of the IC activations accounted for by the five derived CCs through the learned convolutive model. The IC scalp maps on the left are shown for interest. Contributions arranged on the diagonal are dominant. Squares represent the (rounded) percent variance of the IC activation time series accounted for by each CC through the convolutive model. Significant off-diagonal elements indicate that each CCs describes some interaction between the IC activations.

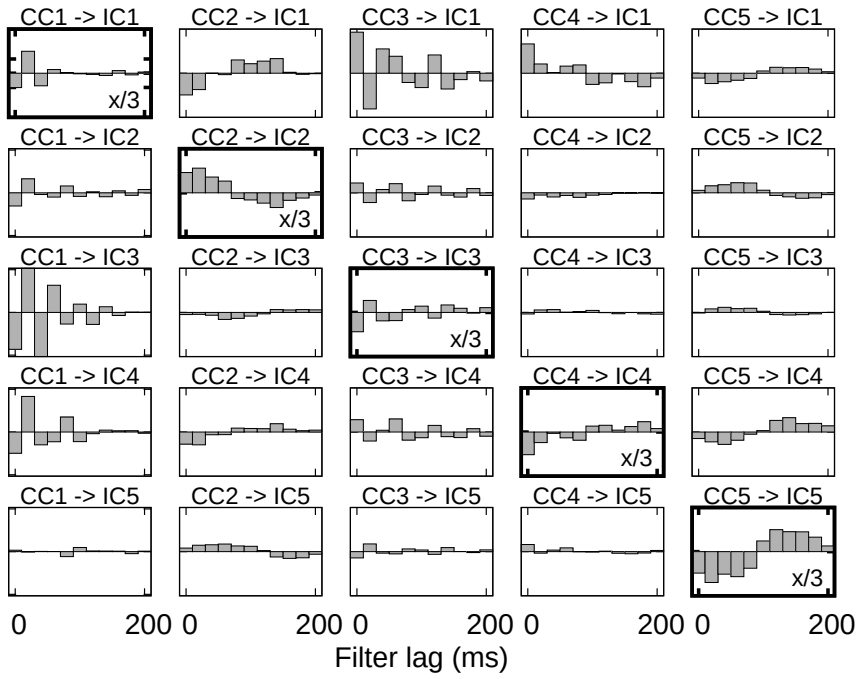


Figure 7.8: Convolutive mixing filters. Learned by convolutive ICA using the CICAAR algorithm $(L, M) = (10, 30)$. The dominant diagonal filters are shown in one-third scale.

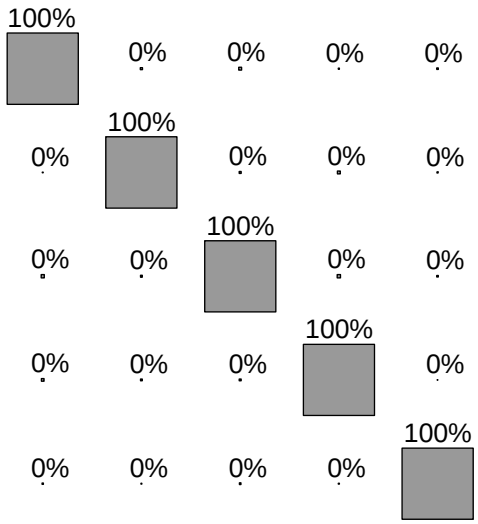


Figure 7.9: Crosstalk matrix of order $R = 10$ measured for the innovations. The squares have areas proportional to the respective crosstalk matrix elements. The percentage written above the squares have been rounded. Clearly convolutive ICA has successfully removed some temporal dependencies that were present in the activations to begin with (compare to figure 7.5).

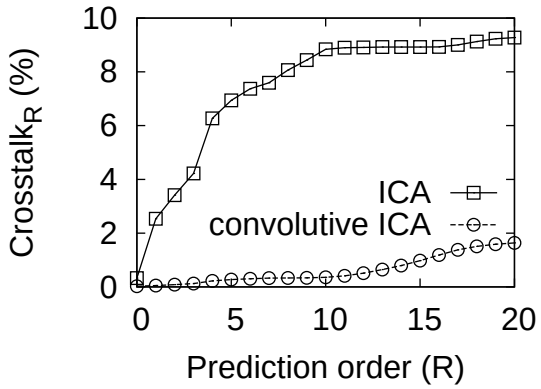


Figure 7.10: Crosstalk for different prediction orders. The crosstalk was measured for instantaneous ICA activations and for convolutive ICA innovations. As the prediction order R increased, the crosstalk of the instantaneous IC activations increased. E.g. for one of the activations, 9% of the variance could be explained by linear prediction from the previous 10 time points (200 ms) of another IC. For the innovations, however, the crosstalk remained low as the prediction order increased, indicating that convolutive ICA had in fact unmixed the delayed temporal dependencies present between the activations. For prediction orders greater than $R = 10$, the crosstalk for the convolutive ICA innovations also increased because the convolutive model with $L = 10$ had only removed temporal dependencies up to order 10.

7.3 are made from the CC contributions to each IC through the model. Note how each individual CC has a different contribution to the different ICs.

Frequency domain

Figure 7.11 shows the power spectral contributions of the most contributing CCs to the five ICs. The most contributing CCs are found in accordance with figure 7.7. Note that the broad alpha[5] band spectral peak in IC1 (uppermost panel in figure 7.11) around 10Hz has been split between CC1 and CC3. In the middle panel, note the distinct spectral contributions of CC1 and CC3 to the double alpha peak in the IC3 spectrum. In line with the time domain analysis, the CCs made different spectral contributions to the ICs. For example, CC1 made different power spectral density contributions to IC1, IC3 and IC4.

Figure 7.12 shows the power spectra for the five innovations; clearly, the innovations had distinct non-white power spectra. The color of each innovation was mainly due to the source model as can be seen from *whitening* the innovations by using the AR inverse of the source color model, i.e. (3.16); the whitened innovation power spectra are shown in figure 7.13. From figure 7.13 it seems that the source model order ($M = 30$) was not completely enough to model the long autocorrelations (low frequencies) in the innovations. This finding stems with the fact that the BIC in figure 7.6A had not yet peaked at $L + M = 40$. So, the model order of $L + M = 40$ was not enough to model all temporal dependencies in the data. However, this was not the point of introducing the source model. The point was to be able to detect a convolutive *mixture*, and the finding that convolutive ICA was relevant instead of instantaneous ICA still remains.

[5] See section E.1 for spectral properties of EEG.

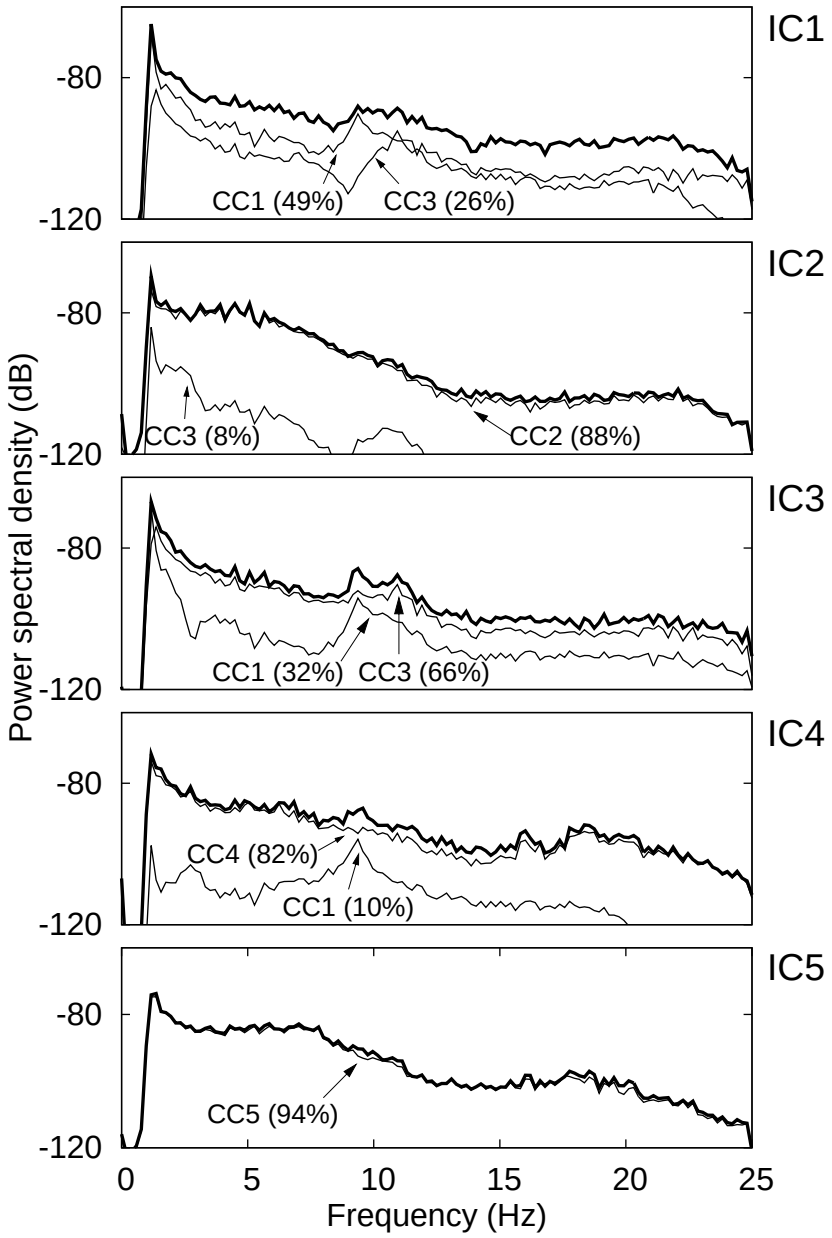


Figure 7.11: Spectral contributions (in thin traces) of the most contributing CCs to the five ICs (in bold traces), in accordance with figure 7.7. The broad alpha band spectral peak in IC1 (uppermost panel) around 10Hz has been split between CC1 and CC3. Similarly, in the middle panel, there are distinct spectral contributions of CC1 and CC3 to the double alpha peak in the IC3 spectrum. Note also how CC1 made different power spectral density contributions to IC1, IC3 and IC4.

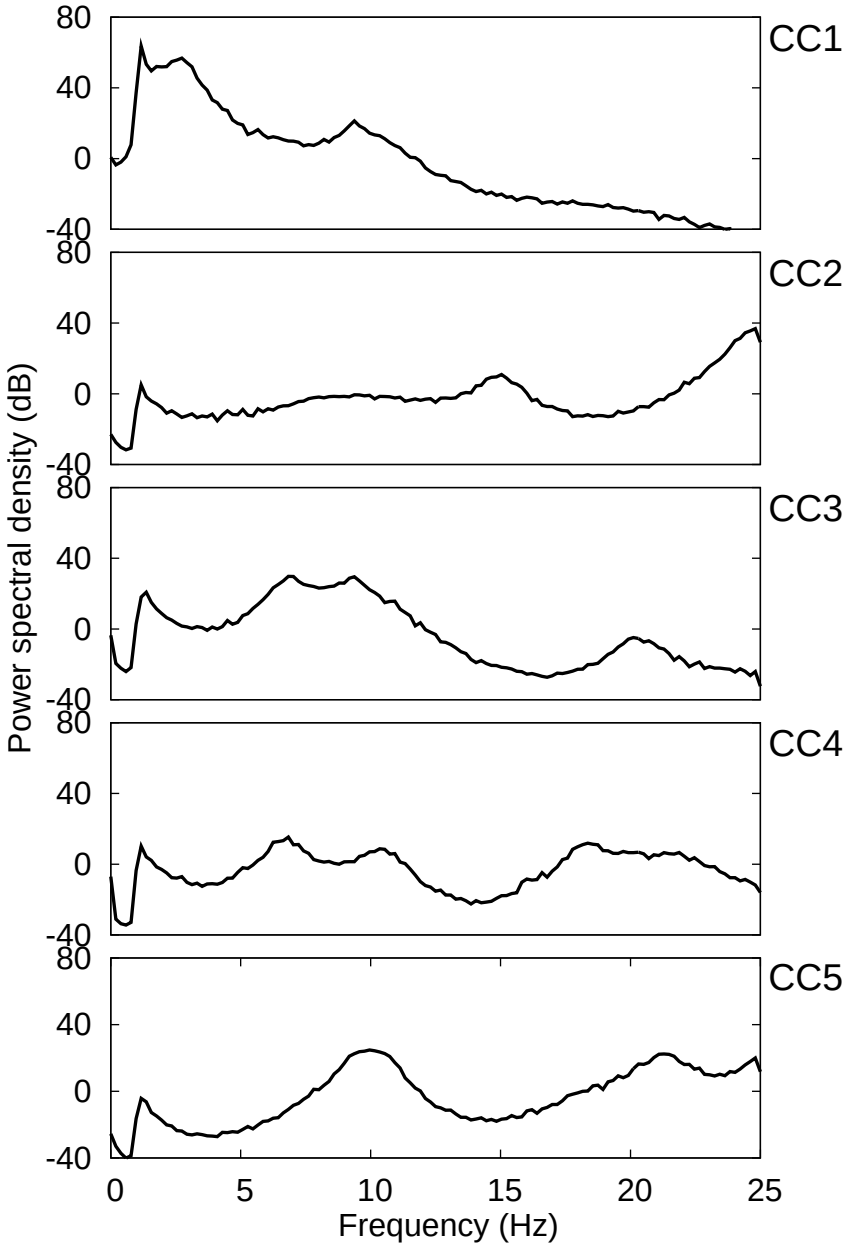


Figure 7.12: Power spectra for the five innovations. They are clearly non-white. The color of each innovation was mainly due to the source model as can be seen from the whitened innovations whose power spectra are shown in figure 7.13.

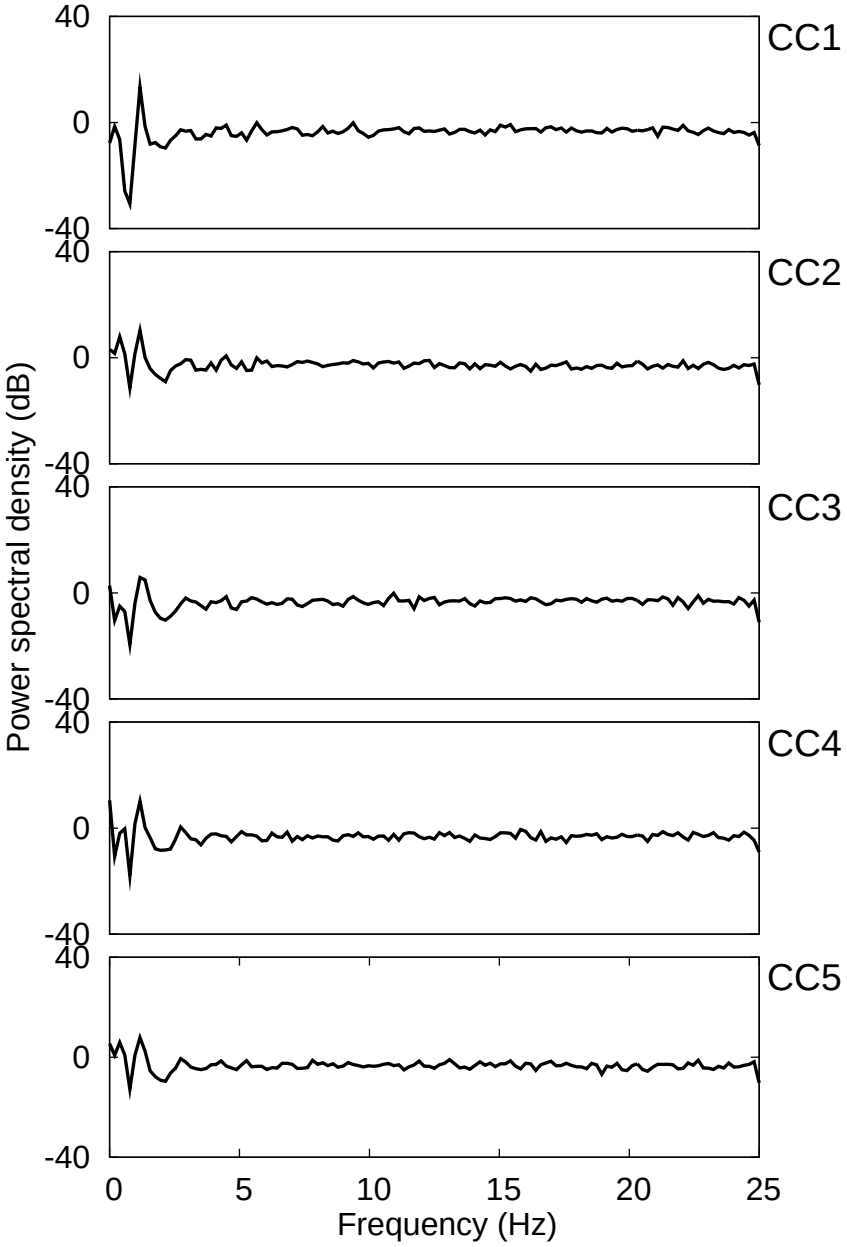


Figure 7.13: Power spectra for the five *whitened* innovations. It seems that the source model order ($M = 30$) was not completely enough to model the long autocorrelations (low frequencies) in the innovations.

Chapter 8

Conclusion

This thesis was about convolutive ICA, and about investigating whether convolutive ICA would be a relevant model for EEG.

Two original methods for convolutive ICA were proposed, namely the CICAAR and the CICAP algorithms.

The CICAP algorithm assumed the existence of a linear predictor in order to formulate the convolutive ICA problem in two steps: linear deconvolution followed by instantaneous ICA. The derivation of the algorithm gave great insight into the nature of the convolutive ICA problem, but the CICAP algorithm had problems in a test situation with non-stationary data. A possible explanation for this deficiency was that the linear predictor, which was estimated from the data, was under the influence of non-stationary correlations in the data. Thus, the CICAP algorithm was disrupted in the deconvolution step where an inverted form of the linear predictor was to be applied to the data.

The CICAAR algorithm was based on a derivation of the likelihood function, involving a multivariate auto-regressive inverse filter which was kept stable by a density declaration on the sources. The algorithm was a direct generalization of Infomax ICA to include the case of convolutive mixing, and hence was a natural choice for investigating the outcome of convolutive ICA decomposition of EEG data in contrast to Infomax ICA decomposition. The likelihood function was

also derived for the overdetermined case, but it turned out that the algorithm then suffered from the null-space problem. Two practical cures to the null-space problem were investigated, namely the augmented and the diminished approaches (both utilized the square case of the CICAAR), and they turned out to be practically valid.

One advantage to the CICAAR algorithm was that Bayesian model selection was possible, and in particular, it was possible to select the optimal order of the filters in the convolutive mixing model. To be able to reject convolution correctly in favor for instantaneous ICA instead, a FIR model for the source auto-correlations was introduced in the CICAAR algorithm. A protocol for detecting the optimal dimensions of the model was proposed, and it was shown by simulation that the protocol successfully rejected convolution in an instantaneous mixture.

The role of instantaneous ICA in context of EEG was described in physiological terms, and in particular the nature of dipolar ICA components was described. A proposed measure, the crosstalk measure (which was related to measures of Granger causality), showed that the instantaneous ICA components lacked independence when time lags were taken into consideration. It was shown that the CICAAR algorithm could be used to remove the delayed temporal dependencies in a subset of ICA components, thus making the components "more independent". A general recipe for ICA analysis of EEG was proposed: first decompose the data using instantaneous ICA, then select a physiologically interesting subspace, then remove the delayed temporal dependencies among the instantaneous ICA components by using convolutive ICA. This recipe turned out computationally feasible with the CICAAR algorithm while yielding results that were easy to interpret. Finally, by careful Bayesian model selection it was shown that convolutive ICA was a better model for EEG than instantaneous ICA.

Appendix B

Bayes Information Criterion (BIC)

Let $\mathcal{M}$ represent a specific choice of model. The Bayes Information Criterion (BIC) is given by [54]

$$\log \mathrm{p}(\mathcal{M}|data) \approx \log \mathrm{p}(data|\boldsymbol{\theta}_0, \mathcal{M}) - \frac{\dim \boldsymbol{\theta}}{2} \log N \tag{B.1}$$

where $\dim \boldsymbol{\theta}$ represents the number of parameters in the model, $\boldsymbol{\theta}_0$ are the maximum likelihood parameters, N is the number of samples in the data set.

The number of parameters for different configurations of the CICAAR algorithm is outlined in the following table

CICAAR configuration	$\dim \boldsymbol{\theta}$
Square	$D^2(L+1) + DM$
Overdetermined (Augmented)	$D^2 + DK_cL + DM$
Overdetermined (Diminished)	$K^2(L+1) + KM + \dim_{\mathrm{PCA}}$

where K_c is the number of convolutive components, and $\dim_{\mathrm{PCA}}$ is the number of parameters used when PCA is used for the projection, ($\dim_{\mathrm{PCA}}$ is given in [20]).

Appendix E

EEG primer and event-related transforms

E.1 Spectral properties of EEG

Historically, much EEG research has been presented in terms of signal power at empirical frequency bands. Figure E.1 shows a typical power spectrum of an EEG channel recording. The peak at 50Hz comes from electromagnetic 'line-

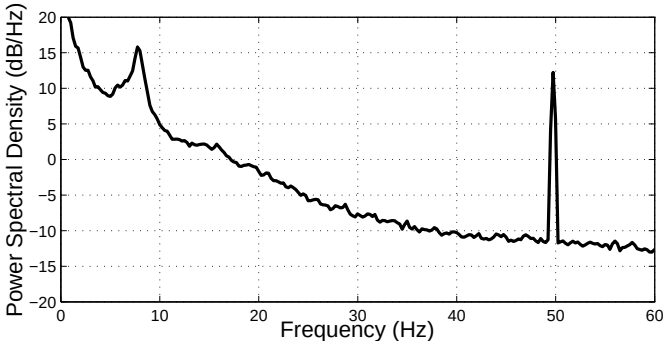
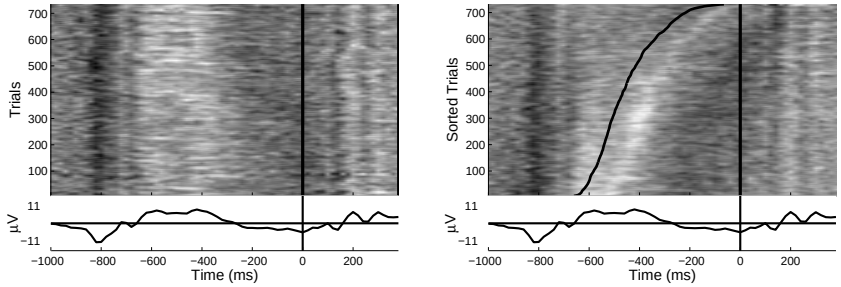


Figure E.1: A typical EEG powerspectrum.



(a) ERP image and average ERP (b) Sorted ERP image and average ERP

Figure E.2: The ERP image is a nice way to display ERP variability while enhancing structure. In this image, averaging was performed in bins of 10 trials which gives a structure-enhancing smoothing effect.

noise' and is not of any interest. Most of the brain-related power lies below 30Hz. The peak at 8Hz is in the so-called 'Alpha' frequency band. The Alpha band expands from about 8Hz to about 12Hz and is characteristic of a relaxed and alert state of consciousness. Other frequency bands, which are important in the mapping of EEG and mind, have been identified empirically. Their precise frequencies might vary slightly from presentation to presentation, but to name a few: 'Delta' (0 to 4Hz), 'Theta' (4Hz to 8Hz), 'Alpha' (8Hz to 12Hz), 'Beta' (12Hz to 30Hz), 'Gamma' (30Hz to 80Hz).

E.2 The ERP image

One property that makes EEG attractive is its temporal resolution which is very high compared to other brain imaging techniques such as functional Magnetic Resonance Imaging (fMRI). EEG can potentially tell us precisely *when* something happened in the brain. EEG activity from cognitive processing which is directly related to a certain event is called the 'Event Related Potential' (ERP). The amplitude of the ERP is in general so small, compared to the ongoing EEG and noise, that the tradition has been to analyze the ERP by aligning and averaging many trials. Noise and other activity, which is random with respect to the event, will be attenuated by the averaging while the reproducible activity will be enhanced in the average ERP. However, the problem with averaging is that brain activity which does not reproduce itself exactly the same way in each trial, so-called 'induced' activity, will disappear or be distorted by the averaging.

Makeig et al. [37] has proposed the 'ERP image' which generalizes the simple

ERP averaging. A pixel, say at position (x, y) in an ERP image, is the signal intensity of trial x at time-latency y relative to the beginning of the trial. The structure in the image is enhanced further by vertical smoothing, i.e. binning of the trials and averaging within the bins. Figure E.2(a) shows an ERP image and the associated average ERP. The activity is clearly variant from trial to trial, and the average ERP is a rough summary of this activity. A particularly useful feature of the ERP image is that it can potentially reveal some interplay between the event used for aligning and another event. An example of this is seen in figure E.2(b) where the trials have been sorted according to the timing of another event and the sigmoidal curve displays the timing of that other event.

E.3 Coherence

A simple measure for the interrelation between cortical areas is obtained by measuring the cross correlation between signals from relevantly sited electrodes. The areas of interest might interact with a typical delay, and furthermore, a better measure should not be affected by the variance of the signals. Hence the cross-correlation function coefficient

$$\hat{c}_{xy}(\tau) = \hat{\gamma}_{xy}(\tau)/\sqrt{\hat{\mathrm{var}}(x)\hat{\mathrm{var}}(y)} \tag{E.1}$$

is the natural choice, where $\gamma_{xy}(\tau) =< x(t)y(t-\tau) >$ is the cross correlation function between signals $x(t)$ and $y(t)$. As certain frequencies in the EEG seem to have certain functional relevance another measure, namely 'coherence', is popular in EEG analysis (see e.g. [42, 59]). Coherence is defined by

$$\hat{C}_{xy}(f) = |\hat{G}_{xy}(f)|/\sqrt{\hat{G}_{xx}(f)\hat{G}_{yy}(f)} \tag{E.2}$$

where $G_{xy}(f)$ denotes cross power spectra that must be estimated by epoch averaging [51, 42]. The coherence gives a measure of the linear dependence between two signals as a function of frequency. Coherence can also be examined at different frequencies for the two sites, i.e.

$$\hat{C}_{xy}(f_1, f_2) = |\hat{G}_{xy}(f_1, f_2)|/\sqrt{\hat{G}_{xx}(f_1)\hat{G}_{yy}(f_2)} \tag{E.3}$$

see e.g. [59].

E.4 Inter-trial coherence (ITC)

Inter-trial coherence (ITC) is a measure for the synchronization between a signal and an event-indicator function. In particular, inter-trial phase coherence

(ITPC) is a measure of phase synchronization, at a given frequency f, as a function of event-relative time t. It is defined by

$$\mathrm{ITPC}(f,t) = \frac{1}{N}\sum_{n=1}^{N}\frac{F_n(f,t)}{|F_n(f,t)|} \tag{E.4}$$

where $F_n(f,t)$ is a (complex) time-frequency transform of signal trial n. ITPC is also known as the 'phase-locking factor' [37, 10]. Similarly, another ITC measure is the inter-trial linear coherence $\mathrm{ITLC} = \sum_n F_n(f,t)/\sqrt{N\sum_{n'}|F_{n'}(f,t)|^2}$.

Appendix M

Matrix Results

M.1 Derivatives involving the pseudo inverse

A wrt. Moore-Penrose pseudo inverse

$$\frac{\partial A^{\mathrm{T}}}{\partial (A^{+})_{ij}} = A^{\mathrm{T}} A (EL)_{ij} - A^{\mathrm{T}} (EL)_{ji} A^{\mathrm{T}} - A^{\mathrm{T}} A (EL)_{ij} A A^{+} \quad , \quad A \in \mathbb{R}^{M \times N}, M \geq N \tag{M.1}$$

and as a special case when A is square

$$\frac{\partial A^{\mathrm{T}}}{\partial (A^{-1})_{ij}} = -A^{\mathrm{T}} (EL)_{ji} A^{\mathrm{T}} \quad , \quad A \in \mathbb{R}^{M \times M} \tag{M.2}$$

proof

First, define $B = A^+$, i.e. also, $A = B^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}$. Then

$$\begin{aligned}
\frac{\partial A^{\mathrm{T}}}{\partial (A^+)_{ij}} =& \frac{\partial (BB^{\mathrm{T}})^{-1}B}{\partial (B)_{ij}} \\
=& (BB^{\mathrm{T}})^{-1}\frac{\partial B}{\partial (B)_{ij}} + \frac{\partial (BB^{\mathrm{T}})^{-1}}{\partial (B)_{ij}}B \\
=& (BB^{\mathrm{T}})^{-1}(EL)_{ij} - (BB^{\mathrm{T}})^{-1}\left[B\frac{\partial B^{\mathrm{T}}}{\partial (B)_{ij}} + \frac{\partial B}{\partial (B_{ij})}B^{\mathrm{T}}\right](BB^{\mathrm{T}})^{-1}B \\
=& (BB^{\mathrm{T}})^{-1}(EL)_{ij} - (BB^{\mathrm{T}})^{-1}\left[B(EL)_{ji} + (EL)_{ij}B^{\mathrm{T}}\right](BB^{\mathrm{T}})^{-1}B \\
=& A^{\mathrm{T}}A(EL)_{ij} - A^{\mathrm{T}}(EL)_{ji}A^{\mathrm{T}} - A^{\mathrm{T}}A(EL)_{ij}AA^+
\end{aligned}$$

alternative proof of (M.2)

$$\begin{aligned}
\frac{\partial A^{\mathrm{T}}(A^{\mathrm{T}})^{-1}}{\partial x} =& 0 \\
\frac{\partial A^{\mathrm{T}}}{\partial x}(A^{\mathrm{T}})^{-1} + A^{\mathrm{T}}\frac{\partial (A^{\mathrm{T}})^{-1}}{\partial x} =& 0 \\
\frac{\partial A^{\mathrm{T}}}{\partial x}(A^{\mathrm{T}})^{-1} =& -A^{\mathrm{T}}\frac{\partial (A^{\mathrm{T}})^{-1}}{\partial x} \\
\frac{\partial A^{\mathrm{T}}}{\partial x} =& -A^{\mathrm{T}}\frac{\partial (A^{\mathrm{T}})^{-1}}{\partial x}A^{\mathrm{T}}
\end{aligned}$$

log determinant of square wrt. A

$$\frac{\partial \log |A^{\mathrm{T}}A|}{\partial (A)_{mn}} = 2(A^{\mathrm{T}+})_{mn} \quad , \quad A \in \mathbb{R}^{M\times N}, M \geq N \tag{M.3}$$

proof

$$
\begin{aligned}
&= \mathrm{Tr}\left(\frac{\partial \log|A^{\mathrm{T}}A|}{\partial A^{\mathrm{T}}A}\frac{\partial (A^{\mathrm{T}}A)^{\mathrm{T}}}{\partial (A)_{mn}}\right)\\
&= \mathrm{Tr}\left((A^{\mathrm{T}}A)'^{-1}\frac{\partial (A^{\mathrm{T}}A)^{\mathrm{T}}}{\partial (A)_{mn}}\right)\\
&= \mathrm{Tr}\left((A^{\mathrm{T}}A)^{-1}\frac{\partial A^{\mathrm{T}}A}{\partial (A)_{mn}}\right)\\
&= \mathrm{Tr}\left((A^{\mathrm{T}}A)^{-1}\left[\frac{\partial A^{\mathrm{T}}}{\partial (A)_{mn}}A + A^{\mathrm{T}}\frac{\partial A}{\partial (A)_{mn}}\right]\right)\\
&= \sum_{j=1}^{N}(A^{\mathrm{T}}A)^{-1}_{jn}A_{mj} + \sum_{j=1}^{N}(A^{\mathrm{T}}A)^{-1}_{nj}A^{\mathrm{T}}_{jm}\\
&= 2\sum_{j=1}^{N}(A^{\mathrm{T}}A)^{-1}_{nj}A^{\mathrm{T}}_{jm}\\
&= 2\left[(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\right]_{nm}\\
&= 2(A^{+})_{nm}\\
&= 2(A^{\mathrm{T}+})_{mn}
\end{aligned}
$$

log determinant of square wrt. Moore-Penrose pseudo inverse

$$
\frac{\partial \log|A^{\mathrm{T}}A|}{\partial (A^{+})_{ij}} = -2(A^{\mathrm{T}})_{ij} \quad , \quad A \in \mathbb{R}^{M\times N}, M \geq N \tag{M.4}
$$

proof

Using (M.3) and the chain rule

$$\begin{aligned}
\frac{\partial \log|A^{\mathrm{T}}A|}{\partial (A^{+})_{ij}} =& 2\,\mathrm{Tr}\left((A^{\mathrm{T}})^{+}\frac{\partial A^{\mathrm{T}}}{\partial (A^{+})_{ij}}\right)\\
&\text{(define } B = A^{+})\\
=& 2\,\mathrm{Tr}\left(B^{\mathrm{T}}\frac{\partial A^{\mathrm{T}}}{\partial (A^{+})_{ij}}\right)\\
&\text{(using (M.1))}\\
=& 2\,\mathrm{Tr}\left(B^{\mathrm{T}}((BB^{\mathrm{T}})^{-1}(EL)_{ij} - (BB^{\mathrm{T}})^{-1}\left[B(EL)_{ji} + (EL)_{ij}B^{\mathrm{T}}\right](BB^{\mathrm{T}})^{-1}B)\right)\\
=& 2\,\mathrm{Tr}\left(B^{\mathrm{T}}((BB^{\mathrm{T}})^{-1}(EL)_{ij}\right) - 2\,\mathrm{Tr}\left(B^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}\left[B(EL)_{ji} + (EL)_{ij}B^{\mathrm{T}}\right](BB^{\mathrm{T}})^{-1}B\right)\\
=& 2\,\mathrm{Tr}\left(A(EL)_{ij}\right) - 2\,\mathrm{Tr}\left(A\left[B(EL)_{ji} + (EL)_{ij}B^{\mathrm{T}}\right](BB^{\mathrm{T}})^{-1}B\right)\\
=& 2\,\mathrm{Tr}\left(A(EL)_{ij}\right) - 2\,\mathrm{Tr}\left(\left[B(EL)_{ji} + (EL)_{ij}B^{\mathrm{T}}\right](BB^{\mathrm{T}})^{-1}\right)\\
=& 2\,\mathrm{Tr}\left(A(EL)_{ij}\right) - 2\,\mathrm{Tr}\left(B(EL)_{ji}(BB^{\mathrm{T}})^{-1}\right) - 2\,\mathrm{Tr}\left((EL)_{ij}B^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}\right)\\
=& 2\,\mathrm{Tr}\left(A(EL)_{ij}\right) - 2\,\mathrm{Tr}\left(A(EL)_{ij}\right) - 2\,\mathrm{Tr}\left((EL)_{ij}B^{\mathrm{T}}(BB^{\mathrm{T}})^{-1}\right)\\
=& -2\,\mathrm{Tr}\left((EL)_{ij}A\right)\\
=& -2(A^{\mathrm{T}})_{ij}
\end{aligned}$$

M.2 Integrals involving Dirac delta function

Scalar

$$\int p(s)\delta(vs - x)ds = \frac{1}{|v|}p(x/v) \tag{M.5}$$

proof

The delta function is defined by the tractable form

$$\int \delta(u - x/v)p(u)du = p(x/v).$$

To get the integral (M.5) to the tractable form use the transformation which satisfies $v\phi(u) - x = u - x/v$, namely,

$$\phi(u) = u/v - x/v^2 - x/v.$$

Then, transforming the integral and plugging in the Jacobian we get

$$\begin{aligned}\int p(s)\delta(vs - x)ds &= \int \left|\frac{\partial\phi(u)}{\partial u}\right| \delta(u - x/v)p(u)du \\ &= \frac{1}{|v|}p(x/v)\end{aligned}$$

Mixing matrix

$$\int p(s)\delta(As - x)ds = |\det A|^{-1}p(A^{-1}x) \tag{M.6}$$

proof

The delta function is defined by the tractable form

$$\int \delta(u - A^{-1}x)p(u)du = p(A^{-1}x).$$

To get the integral (M.6) to the tractable form use the transformation which satisfies $A\phi(u) - x = u - A^{-1}x$, namely,

$$\phi(u) = A^{-1}u - A^{-1}(A^{-1}x - x).$$

Then, transforming the integral and plugging in the Jacobian we get

$$\begin{aligned}\int p(s)\delta(As - x)ds &= \int \frac{\partial\phi(u)}{\partial u}\delta(u - A^{-1}x)p(u)du \\ &= |\det(A)|^{-1}p(A^{-1}x)\end{aligned}$$

Undercomplete mixing matrix

For $A \in \mathbb{R}^{M\times N}, M \geq N$ we find

$$\int p(s)\delta(x - As)ds = \begin{cases} |A^{\mathrm{T}}A|^{-1/2}p(A^{+}x) & , x = AA^{+}x \\ 0 & , \text{otherwise} \end{cases} \tag{M.7}$$

proof

We shall make use of the mapping $x \mapsto (x_\perp \ , \ x_\parallel) = (U_\perp^{\mathrm{T}} x \ , \ U_\parallel^{\mathrm{T}} x)$, i.e. $U_\parallel^{\mathrm{T}} = (A^{\mathrm{T}}A)^{-1/2}A^{\mathrm{T}}$, such that $U_\parallel U_\parallel^{\mathrm{T}} A = A$.

$$
\begin{aligned}
&= \int p(s)\delta(x_\perp)\delta(x_\parallel - (As)_\parallel)ds \\
&= \delta(x_\perp) \int p(s)\delta(x_\parallel - (As)_\parallel)ds \\
&= \delta(x_\perp) \int p(s)\delta((A^{\mathrm{T}}A)^{-1/2}A^{\mathrm{T}}x - (A^{\mathrm{T}}A)^{-1/2}A^{\mathrm{T}}As)ds \\
&\quad (\ldots \text{integral transformation} \ldots) \\
&= \delta(x_\perp) \int |\det(A^{\mathrm{T}}A)|^{-1/2} p(u)\delta((A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}x - u)du \\
&= \delta(x_\perp)|\det(A^{\mathrm{T}}A)|^{-1/2} p((A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}x) \\
&= \delta(x_\perp)|\det(A^{\mathrm{T}}A)|^{-1/2} p(A^+x) \\
&= \delta(x - AA^+x)|\det(A^{\mathrm{T}}A)|^{-1/2} p(A^+x) \\
&= \begin{cases} |A^{\mathrm{T}}A|^{-1/2} p(A^+x) & , x = AA^+x \\ 0 & , \text{otherwise} \end{cases}
\end{aligned}
$$

Appendix P

Publications

P.1 M. Dyrholm, S. Makeig and L. K. Hansen, Model selection for convolutive ICA with an application to spatio-temporal analysis of EEG, Neural Computation

[16] M. Dyrholm, S. Makeig and L. K. Hansen, Model selection for convolutive ICA with an application to spatio-temporal analysis of EEG, Neural Computation, (submitted) 2005

Model selection for convolutive ICA with an application to spatio-temporal analysis of EEG

Mads Dyrholm, Scott Makeig and Lars Kai Hansen

November 8, 2005

Abstract We present a new algorithm for maximum likelihood convolutive independent component analysis (ICA) in which sources are unmixed using stable auto-regressive filters determined implicitly by estimating a convolutive model of the mixing process. By introducing a convolutive mixing model for the sources we show how the order of the filters in the convolutive model can be correctly detected using Bayesian model selection. We demonstrate a framework for deconvolving a subspace of independent components in electroencephalography (EEG). Initial results suggest that in some cases convolutive mixing may be a more realistic model for EEG signals than the instantaneous ICA model.

1 Introduction

Motivated by the EEG signal's complex temporal dynamics we are interested in convolutive independent component analysis (cICA), which in its most basic form concerns reconstruction of $L+1$ mixing matrices $\mathbf{A}_\tau$ and N source signal vectors ('innovations'), $\mathbf{s}_t$, of dimension K, combining to form an observed D-dimensional linear convolutive mixture

$$\mathbf{x}_t = \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \tag{1}$$

That is, cICA models the observed data $\mathbf{x}$ as produced by K source processes whose time courses are first convolved with fixed, finite-length time filters and then summed in the D sensors. This allows a single source signal to be expressed in the different sensors with variable delays and frequency characteristics.

One common application for this model is the acoustic blind source separation problem in which sound sources are mixed in a reverberant environment. Simple ICA methods not taking signal delays into account fail to produce satisfactory results for this problem, which has thus been the focus of much cICA research (e.g., [Lee et al., 1997b; Parra et al., 1998; Sun and Douglas, 2001; Mitianoudis and Davies, 2003; Anemüller and Kollmeier, 2003]).

For analysis of human electroencephalographic (EEG) signals recorded from the scalp, ICA has already proven to be a valuable tool for detecting and enhancing relevant 'source' subspace brain signals while suppressing irrelevant 'noise' and artifacts such as those produced by muscle activity and eye blinks [Makeig et al., 1996; Jung et al., 2000; Delorme and Makeig, 2004]. In conventional ICA each independent component (IC) is represented as a spatially *static* projection of cortical source activity to the sensors. Results of static ICA decomposition are generally compatible with a view of EEG source signals as originating in spatially static cortical domains within which local field potential fluctuations are partially synchronized [Makeig et al., 2000; Jung et al., 2001; Delorme et al., 2002; Makeig et al., 2004a; Onton et al., 2005]. Modelling EEG data as consisting of convolutive as well as static independent processes allow a richer palette for source modeling, possibly leading to more complete signal independence.

In this paper we present a new cICA decomposition method that, unlike most previous work in the area, operates entirely in the time-domain. One advantage of the time-domain approach is that it avoids the need to window the data and hence avoids the need for manual tuning of window length and tapering. Although tuning a wavelet or DFT (discrete fourier transform) domain approach is

possible in many acoustic situations in which 'gold standard' performance measures (e.g., listening tests) are available, no such 'gold standard' of success is available in the case of human EEG. Also, time domain deconvolution is not restricted to one frequency band at a time, and thus can avoid the difficult process of piecing together deconvolutions computed separately at different frequencies [Anemüller et al., 2003].

The new scheme also makes no assumptions about 'non-stationarity' of the source signals, a key assumption in several successful cICA methods (see e.g. [Parra and Spence, 2000; Rahbar et al., 2002]) whose relevance to EEG is unclear. Previous time-domain and DFT-domain methods have formulated the problem as one of finding a finite impulse response (FIR) filter that *un*mixes as in (2) below [Belouchrani et al., 1997; Choi and Cichocki, 1997; Moulines et al., 1997; Lee et al., 1997a; Attias and Schreiner, 1998; Parra et al., 1998; Deligne and Gopinath, 2002; Douglas et al., 1999; Comon et al., 2001; Sun and Douglas, 2001; Rahbar and Reilly, 2001; Rahbar et al., 2002; Baumann et al., 2001; Anemüller and Kollmeier, 2003]

$$\hat{\mathbf{s}}_t = \sum_{\lambda} \mathbf{W}_\lambda \mathbf{x}_{t-\lambda} \tag{2}$$

However, the inverse of the mixing FIR filter modeled in (1) is, in general, an infinite impulse response (IIR) filter. We thus expect that FIR based unmixing will require estimation of extended or potentially infinite length unmixing filters. Our method, by contrast, finds such an unmixing *IIR* filter implicitly in terms of the *mixing* model parameters, i.e. the $\mathbf{A}_\tau$'s in (1), isolating $\mathbf{s}_t$ in (1) as

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^{\#} \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \hat{\mathbf{s}}_{t-\tau} \right) \tag{3}$$

where $\mathbf{A}_0^{\#}$ denotes Moore-Penrose inverse of $\mathbf{A}_0$. Another advantage of this parametrization is that the $\mathbf{A}_\tau$'s allow a separated source signal to be easily back-projected into the original sensor domain.

Other authors have proposed the use of IIR filters for separating convolutive

mixtures using the maximum likelihood principle. The unmixing IIR filter (3) generalizes that of [Torkkola, 1996] to allow separation of more than only two sources. Furthermore, it bears interesting resemblance to that of [Choi and Cichocki, 1997; Choi et al., 1999]. Though put in different analytical terms, the inverses used there are equivalent to the unmixing IIR (3). However, the unique expression (3), and its remarkable analytical simplicity, is the key to learning the parameters of the *mixing* model (1) directly.

2 Learning the mixing model parameters

Statistically motivated maximum likelihood approaches for cICA have been proposed ([Torkkola, 1996; Pearlmutter and Parra, 1997; Parra et al., 1997; Moulines et al., 1997; Attias and Schreiner, 1998; Deligne and Gopinath, 2002; Choi et al., 1999; Dyrholm and Hansen, 2004]) and are attractive for a number of reasons. First, they force a declaration of statistical assumptions—in particular the assumed distribution of the source signals. Secondly, a maximum likelihood solution is asymptotically optimal given the assumed observation model and the prior choices for the 'hidden' variables.

Assuming independent and identically distributed (i.i.d.) sources and no noise, the likelihood of the parameters in (1) given the data is

$$\mathrm{p}(\mathbf{X}|\{\mathbf{A}_\tau\}) = \int \cdots \int \prod_{t=1}^{N} \delta(\mathbf{e}_t)\, \mathrm{p}(\mathbf{s}_t)\, \mathrm{d}\, \mathbf{s}_t \tag{4}$$

where

$$\mathbf{e}_t = \mathbf{x}_t - \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \tag{5}$$

and $\delta(\mathbf{e}_t)$ is the Dirac delta function.

In the following derivation, we assume that the number of convolutive source processes K does not exceed the dimension D of the data. First, we note that only the N'th term under the product operator in (4) is a function of $\mathbf{s}_N$. Hence,

the $\mathbf{s}_N$-integral may be evaluated first, yielding

$$\mathrm{p}(\mathbf{X}|\{\mathbf{A}_\tau\}) = |\mathbf{A}_0^T\mathbf{A}_0|^{-1/2} \int \cdots \int \mathrm{p}(\hat{\mathbf{s}}_N) \prod_{t=1}^{N-1} \delta(\mathbf{e}_t)\, \mathrm{p}(\mathbf{s}_t)\, \mathrm{d}\,\mathbf{s}_t \tag{6}$$

where integration is over all sources except $\mathbf{s}_N$, and

$$\hat{\mathbf{s}}_N = \mathbf{A}_0^{\#} \left(\mathbf{x}_N - \sum_{\tau=1}^{L} \mathbf{A}_\tau \mathbf{s}_{N-\tau} \right) \tag{7}$$

Now, as before, only one of the factors under the product operator in (6) is a function of $\mathbf{s}_{N-1}$. Hence, the $\mathbf{s}_{N-1}$-integral can now be evaluated, yielding

$$\mathrm{p}(\mathbf{X}|\{\mathbf{A}_\tau\}) = |\mathbf{A}_0^T\mathbf{A}_0|^{-1} \int \cdots \int \mathrm{p}(\hat{\mathbf{s}}_N)\, \mathrm{p}(\hat{\mathbf{s}}_{N-1}) \prod_{t=1}^{N-2} \delta(\mathbf{e}_t)\, \mathrm{p}(\mathbf{s}_t)\, \mathrm{d}\,\mathbf{s}_t \tag{8}$$

where integration is over all sources except $\mathbf{s}_N$ and $\mathbf{s}_{N-1}$, and

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^{\#} \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \mathbf{u}_{t-\tau} \right) \ , \ \mathbf{u}_n = \begin{cases} \mathbf{s}_n \text{ for } n < N-1 \\ \hat{\mathbf{s}}_n \text{ for } n \geq N-1 \end{cases} \tag{9}$$

By induction, and assuming $\mathbf{s}_n$ is zero for $n < 1$, we get

$$\mathrm{p}(\mathbf{X}|\{\mathbf{A}_\tau\}) = |\mathbf{A}_0^T\mathbf{A}_0|^{-N/2} \prod_{t=1}^{N} \mathrm{p}(\hat{\mathbf{s}}_t) \tag{10}$$

where

$$\hat{\mathbf{s}}_t = \mathbf{A}_0^{\#} \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \hat{\mathbf{s}}_{t-\tau} \right) \tag{11}$$

Thus, the likelihood is calculated by first *un*mixing the sources using (11), then measuring (10). It is clear that the algorithm reduces to standard Infomax ICA [Bell and Sejnowski, 1995] when the length of the convolutional filters L is set to zero and $D = K$; in that case (10) can be estimated using $\hat{\mathbf{s}}_t = \mathbf{A}_0^{-1}\mathbf{x}_t$.

2.1 Model source declaration ensures stable un-mixing

Because of inherent instability concerns, the use of IIR filters for unmixing has often been discouraged [Lee et al., 1997a]. Using FIR unmixing filters could certainly ensure stability but would not solve the fundamental problem of inverting

a linear system in cases in which it is not invertible. Invertibility of a linear system is related to the phase characteristic of the system transfer function. A SISO (single input / single output) system is invertible if and only if the complex zeros of its transfer function are all situated within the unit circle. Such a system is characterized as 'minimum phase'. If the system is not minimum phase, only an approximate, 'regularized' inverse can be sought. (See [Hansen, 2002] on techniques for regularizing a system with known coefficients).

For MIMO (multiple input / multiple output) systems, the matter is more involved. The stability of (11), and hence the invertibility of (1), is related to the eigenvalues λ_m of the matrix

$$\tilde{\mathbf{A}} = \begin{bmatrix} -\mathbf{A}_0^{\#}\mathbf{A}_1 & -\mathbf{A}_0^{\#}\mathbf{A}_2 & \dots & -\mathbf{A}_0^{\#}\mathbf{A}_L \\ \mathbf{I} & & & \mathbf{0} \\ & \ddots & & \vdots \\ & & \mathbf{I} & \mathbf{0} \end{bmatrix} \tag{12}$$

For $K = D$, a necessary and sufficient condition is that all eigenvalues λ_m of $\tilde{\mathbf{A}}$ are situated within the unit circle, $|\lambda_m| < 1$ [Neumaier and Schneider, 2001]. We can generalize the 'minimum phase' concept to MIMO systems if we think of the λ_m's as quasi 'poles' of the inverse MIMO transfer function. A SISO system being minimum phase implies that no system with the same frequency response can have a smaller phase shift and system delay.

Generalizing that concept to MIMO systems, we can get a feeling for what a quasi 'minimum phase' MIMO system must look like. In particular, most energy must occur at the beginning of each filter, and less towards the end. However, not all SISO source-to-sensor paths in the MIMO system need be minimum phase for the MIMO system as a whole to be quasi 'minimum phase'.

Certainly, unmixing data using FIR filters is regularized in the sense that their joint impulse response is of finite duration, whereas IIR filter impulse responses may potentially become unstable. Fortunately, the maximum likelihood

approach has a built-in regularization that avoids this problem [Dyrholm and Hansen, 2004]. This can be seen in the likelihood equation (10) by noting that although an unstable IIR filter will lead to a divergent source estimate, $\hat{\mathbf{s}}_t$, such large amplitude signals are exponentially penalized under most reasonable source probability density functions (pdf's), e.g. for EEG data $\mathrm{p}(s) = \mathrm{sech}(s)/\pi$, ensuring that unstable solutions are avoided in the evolved solution.

If so, it may prove safe to use an unconstrained iterative learning scheme to unmix EEG data. Once the unmixing process has been stably initialized, each learning step will produce model refinements that are stable in the sense of equation (11). Even if the system (1) we are trying to unmix is not invertible, meaning no exact stable inverse exists, the maximum-likelihood approach will give a regularized and stable quasi 'minimum phase' solution.

2.2 Gradients and optimization

The cost-function of the algorithm is the *negative log* likelihood

$$\mathcal{L}(\{A_\tau\}) = \frac{N}{2} \log |\det \mathbf{A}_0^T \mathbf{A}_0| - \sum_{t=1}^{N} \log \mathrm{p}(\hat{\mathbf{s}}_t) \tag{13}$$

The gradient of the cost-function is presented here in two steps. Step one reveals the partial derivatives of the source estimates while step two uses the step one results in a chain rule to compute the gradient of the cost-function (see also [Dyrholm and Hansen, 2004])

Step one — Partial derivatives of the unmixed source estimates

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_0^{\#})_{ij}} = \delta(i-k) \left(\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{A}_\tau \hat{\mathbf{s}}_{t-\tau} \right)_j - \left(\mathbf{A}_0^{\#} \sum_{\tau=1}^{L} \mathbf{A}_\tau \frac{\partial \hat{\mathbf{s}}_{t-\tau}}{\partial(\mathbf{A}_0^{\#})_{ij}} \right)_k \tag{14}$$

and $(\ \boldsymbol{\psi}_t\ )_k = p'(\ (\hat{\mathbf{s}}_t)_k\ )/p(\ (\hat{\mathbf{s}}_t)_k\ )$.

$$\frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} = -(\mathbf{A}_0^{\#})_{ki} (\hat{\mathbf{s}}_{t-\tau})_j - \left(\mathbf{A}_0^{\#} \sum_{\tau'=1}^{L} \mathbf{A}_{\tau'} \frac{\partial \hat{\mathbf{s}}_{t-\tau'}}{\partial(\mathbf{A}_\tau)_{ij}} \right)_k \tag{15}$$

Step two — Gradient of the cost-function The gradient of the cost-function with respect to $\mathbf{A}_0^{\#}$ is given by

$$\frac{\partial \mathcal{L}(\{\mathbf{A}_\tau\})}{\partial (\mathbf{A}_0^{\#})_{ij}} = -N(\mathbf{A}_0^T)_{ij} - \sum_{t=1}^{N} \boldsymbol{\psi}_t^T \frac{\partial \hat{\mathbf{s}}_t}{\partial (\mathbf{A}_0^{\#})_{ij}} \tag{16}$$

and the gradient with respect to to the other mixing matrices is

$$\frac{\partial \mathcal{L}(\{\mathbf{A}\})}{\partial (\mathbf{A}_\tau)_{ij}} = -\sum_{t=1}^{N} \boldsymbol{\psi}_t^T \frac{\partial \hat{\mathbf{s}}_t}{\partial (\mathbf{A}_\tau)_{ij}} \tag{17}$$

These expressions allow use of general gradient optimization methods, a stable starting point being $\mathbf{A}_\tau = 0$ (for $\tau \neq 0$) with arbitrary $\mathbf{A}_0$. In the experiments reported below, we have used a Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for optimization. See [Cardoso and Pham, 2004] for a relevant discussion and [Nielsen, 2000] for a reference to the precise implementation we used.

3 Three approaches to overdetermined cICA

Current EEG experiments typically involve simultaneous recording from 30 to 100 or more electrodes, forming a high (D) dimensional signal. After signal separation we hope to find a relatively small number (K) of independent components. Hence we are interested in studying the so-called 'overdetermined' problem ($K < D$). There are at least three different approaches to performing overdetermined cICA:

1. (Rectangular) Perform the decomposition with $D > K$.
2. (Augmented) Perform the decomposition with K set to D, i.e. attempting to estimate some extra sources.
3. (Diminished) Perform the decomposition with D equal to K, i.e. on a K-dimensional subspace projection of the data.

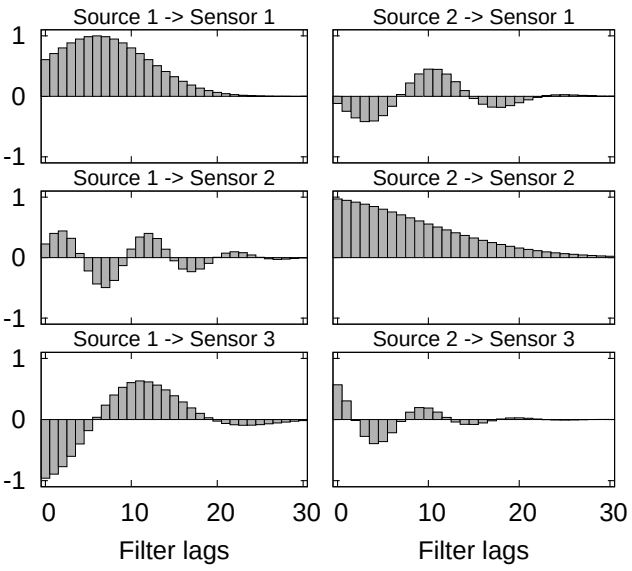


Figure 1: A synthetic MIMO mixing system. Here, two sources were convolutively mixed at three sensors. The 'poles' of the mixture (as defined in section 2.1) are all situated within the unit circle, hence an exact and stable inverse exists in the sense of (11).

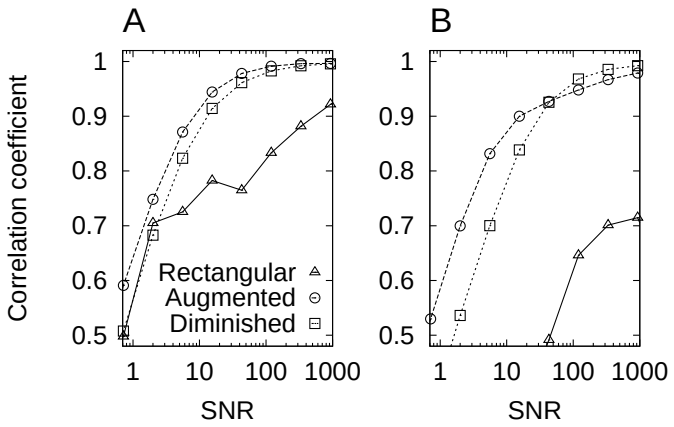


Figure 2: Comparison of source separation of the system in Fig. 1 using three cICA approaches (Rectangular, Augmented, Diminished). A: Estimates of true source activity: correlations with the best-estimated source. B: Similar correlations for the less well estimated source.

We compared the performance of these three approaches experimentally as a function of signal-to-noise ratio (SNR). First, we created a synthetic mixture, two i.i.d source signals $s_1(t)$ and $s_2(t)$ (with $1 \leq t \leq N$ and $N = 30000$) generated from a laplacian distribution, $s_k(t) \sim p(x) = \frac{1}{2}\exp(-|x|)$ with variance $\mathrm{Var}\{s_k(t)\} = 2$. These signals were mixed using the filters of length $L = 30$ shown in Figure 1 producing an overdetermined 3-D mixture ($D = 3$, $K = 2$). A 3-D i.i.d. Gaussian noise signal $\mathbf{n}_t$ was added to the mixture $\mathbf{x}_t = \sigma\mathbf{n}_t + \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau}$ with a controlled variance σ^2.

Next, we investigated how well the three analysis approaches estimated the two sources by measuring the correlations between each true source innovation, $s_k(t)$, and the best-correlated estimated source, $\hat{s}_{k'}(t)$.

Approach 1 (Rectangular). Here, all three data channels were decomposed and the two true sources estimated. Figure 2 shows how well the sources were estimated at different SNR levels. The quality of the estimation dropped dra-

matically as SNR decreased. Even though our derivation (Section 2) is valid for the overdetermined case ($D > K$), the validity of the zero-noise assumption proves vital in this case. The explanation for this can be seen in the definitions of the likelihood (10) and unmixing filter (11).

In (10), any rotation on the columns of $\mathbf{A}_0$ will not influence the determinant term of the likelihood. From (11) we note that the estimated source vectors $\hat{\mathbf{s}}_t$ are found by linear mapping through $\mathbf{A}_0^{\#} : \mathbb{R}^D \mapsto \mathbb{R}^K$. Hence, the source-prior term in (10) alone will be responsible for determining a rotation of $\mathbf{A}_0$ that hides as much variance as possible in the nullspace ($\mathbb{R}^{D-K}$) of $\mathbf{A}_0^{\#}$ in (11). In an unconstrained optimization scheme, this side-effect will be untamed and consequently will hide source variance in the nullspace of $\mathbf{A}_0^{\#}$ and achieve an artificially high likelihood while relaxing the effort to make the sources independent.

Approach 2 (Augmented). One solution to the problem with the Rectangular approach above could be to parameterize the nullspace of $\mathbf{A}_0^{\#}$, or equivalently the orthogonal complement space of $\mathbf{A}_0$. This can be seen as a special case of the algorithm in which $\mathbf{A}_0$ is D-by-D and $\mathbf{A}_\tau$ is D-by-K. With the $D-K$ additional columns of $\mathbf{A}_0$ denoted by $\mathbf{B}$, the model can be written

$$\mathbf{x}_t = \mathbf{B}\mathbf{v}_t + \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \tag{18}$$

where $\mathbf{v}_t$ and $\mathbf{B}$ constitute a low-rank approximation to the noise. Hence, we declare a Gaussian prior p.d.f. on $\mathbf{v}_t$. Note that (18) is a special case of the convolutive model (1). In this case, we attempt to estimate the third (noise) source in addition to the two convolutive sources.

Figure 2 shows how well the sources are estimated using this approach for different SNR levels. For the best estimated source (Fig. 2-A), the Augmented approach gave better estimates than the Rectangular or Diminished approaches. This was also the case for the second source (Fig. 2-B) at low SNR, but not at high SNR since in this case the 'true' $\mathbf{B}$ was near zero and became improbable

under the likelihood model.

Approach 3 (Diminished). Finally, we investigated the possibility of extracting the two sources from a two-dimensional projection of the data. Here, we simply excluded the third 'sensor' from the decomposition. Figure 2 shows that even in the presence of considerable noise, the separation achieved was not as good as in the Augmented approach. However, the Diminished approach used the lowest number of parameters and hence had the lowest comutational complexity. Furthermore, it lacked the peculiarities of the Augmented approach at high SNR. Finally we note that once the Diminished model has been learned, an estimate of the Rectangular model can be obtained by solving

$$< \mathbf{x}_t \mathbf{s}_{t-\lambda}^T > = \sum_\tau \mathbf{A}_\tau < \mathbf{s}_{t-\tau} \mathbf{s}_{t-\lambda}^T > \tag{19}$$

for $\mathbf{A}_\tau$ by regular matrix inversion using the estimated sources and $< \cdot >= \frac{1}{N} \sum_{1=1}^{N}$.

Summary of the three approaches. In the presence of considerable noise, the best separation was obtained by augmenting the model and extracting, from the D-dimensional mixture, K sources as well as a (rank $D-K$) approximation of the noise. However, the Diminished approach had the advantage of lower computational complexity, while the separation it achieved was close to that of the Augmented approach. At very high SNR, the Diminished approach was even slightly better than the Augmented approach. The Rectangular approach, meanwhile, had difficulties and should not be considered for use in practice as the presence of some channel noise may be assumed.

4 Detecting a convolutive mixture

Model selection is a fundamental issue of interest, in particular, detecting the order of L can tell us whether the convolutive mixing model is a better model

than the simpler instantaneous mixing model of standard ICA methods. In the framework of Bayesian model selection, models that are immoderately complex are penalized by the Occam factor, and will therefore only be chosen if there is a relevant need for their complexity. However, this compelling feature can be disrupted if fundamental assumptions are violated. One such assumption was involved in our derivation of the likelihood, in which we assumed that the sources are iid, i.e. not auto-correlated. The problem with this assumption is that the likelihood will favor models based not only on achieved independence but on source whiteness as well. A model selection scheme for L which does not take the source auto-correlations into account will therefore be biased upwards because models with a larger value for L can absorb more source auto-correlation than models with lower L values. To cure this problem, we introduce a model for each of the sources

$$s_k(t) = \sum_{\lambda=0}^{M} h_k(\lambda) z_k(t-\lambda) \tag{20}$$

where $z_k(t)$ represents an i.i.d. signal—a whitened version of the source signal. Introducing the K source filters of order M allows us to reduce the value of L, i.e. lowering the number of parameters in the model while achieving uniformly better learning for limited data [Dyrholm et al., 2005].

We note that some authors of FIR unmixing methods have also used source models, e.g. [Pearlmutter and Parra, 1997; Parra et al., 1997; Attias and Schreiner, 1998].

4.1 Learning source auto-correlation

The negative log likelihood for the model combining (1) and (20) is given by

$$\mathcal{L} = N \log |\det \mathbf{A}_0| + N \sum_k \log |h_k(0)| - \sum_{t=1}^{N} \log \mathrm{p}(\hat{\mathbf{z}}_t) \tag{21}$$

where $\hat{\mathbf{z}}_t$ is a vector of whitened source signal estimates at time t using an operator that represents the inverse of (20), and we assume $\mathbf{A}_0$ to be square as

in the Diminished and Augmented approaches above. We can without loss of generality set $h_k(0) = 1$, then

$$\mathcal{L} = N \log |\det \mathbf{A}_0| - \sum_{t=1}^{N} \log \mathrm{p}(\hat{\mathbf{z}}_t) \tag{22}$$

For notational convenience we introduce the following matrix notation instead of (20), bundling all sources in one matrix equation

$$\mathbf{s}_t = \sum_{\lambda=0}^{M} \mathbf{H}_\lambda \mathbf{z}_{t-\lambda} \tag{23}$$

where the $\mathbf{H}_\lambda$'s are diagonal matrices defined by $(\mathbf{H}_\lambda)_{ii} = h_i(\lambda)$.

To derive an algorithm for learning the source auto-correlations in addition to the mixing model we modify the equations found in Section 2.2; inserting a third, Source model step (see below) between the two steps found there, i.e. substituting $\hat{\mathbf{z}}_t$ for $\hat{\mathbf{s}}_t$ in step two.

Source model step The inverse source coloring operator is given by

$$\hat{\mathbf{z}}_t = \hat{\mathbf{s}}_t - \sum_{\lambda=1}^{M} \mathbf{H}_\lambda \hat{\mathbf{z}}_{t-\lambda} \tag{24}$$

and the partial derivatives, which we shall use in step two, are given by

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{A}_0^{-1})_{ij}} = \frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_0^{-1})_{ij}} - \sum_{\lambda=1}^{M} \mathbf{H}_\lambda \frac{\partial(\hat{\mathbf{z}}_{t-\lambda})_k}{\partial(\mathbf{A}_0^{-1})_{ij}} \tag{25}$$

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} = \frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{A}_\tau)_{ij}} - \sum_{\lambda=1}^{M} \mathbf{H}_\lambda \frac{\partial(\hat{\mathbf{z}}_{t-\lambda})_k}{\partial(\mathbf{A}_\tau)_{ij}} \tag{26}$$

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{H}_\lambda)_{ii}} = -\delta(k-i)(\hat{\mathbf{z}}_{t-\lambda})_i - \left(\sum_{\lambda'=1}^{M} \mathbf{H}_{\lambda'} \frac{\partial \hat{\mathbf{z}}_{t-\lambda'}}{\partial(\mathbf{H}_\lambda)_{ii}} \right)_k \tag{27}$$

Step two modified — Gradient of the cost-function The gradient of the cost-function with respect to $\mathbf{A}_0^{\#}$ with the source model invoked is given by

$$\frac{\partial \mathcal{L}(\{\mathbf{A}_\tau\})}{\partial(\mathbf{A}_0^{\#})_{ij}} = -N(\mathbf{A}_0^T)_{ij} - \sum_{t=1}^{N} \boldsymbol{\psi}_t^T \frac{\partial \hat{\mathbf{z}}_t}{\partial(\mathbf{A}_0^{\#})_{ij}} \tag{28}$$

and the gradient with respect to to the other mixing matrices is

$$\frac{\partial \mathcal{L}(\{\mathbf{A}\})}{\partial (\mathbf{A}_\tau)_{ij}} = -\sum_{t=1}^{N} \boldsymbol{\psi}_t^T \frac{\partial \hat{\mathbf{z}}_t}{\partial (\mathbf{A}_\tau)_{ij}} \tag{29}$$

4.2 Protocol for detecting L

We propose a simple protocol for determining the dimensions (L, M) of the convolutional and source filters. First, expand the convolution without an autofilter ($M = 0$). This will model the total temporal dependency structure of the system $L_{\max}$. The optimal dimension is found by monitoring the Bayes Information Criterion (BIC) [Schwarz, 1978]

$$\log \mathrm{p}(\mathcal{M}|\mathbf{X}) \approx \log \mathrm{p}(\mathbf{X}|\boldsymbol{\theta}_0, \mathcal{M}) - \frac{\dim \boldsymbol{\theta}}{2} \log N \tag{30}$$

where $\mathcal{M}$ represents a specific choice of model structure (L, M), $\boldsymbol{\theta}$ represents the parameters in the model, $\boldsymbol{\theta}_0$ are the maximum likelihood parameters, and N is the size of the data set (number of samples).

Next, keep the temporal dependency constant, $(L + M) = L_{\max}$, while expanding the length of the source autofilters M, again monitoring the BIC to determine the optimal choice of $L = L_{\max} - M$.

4.3 Example: Correctly rejecting cICA of an instantaneous mixture

We will now illustrate the importance of the source model and the validity of the protocol for detecting L when dealing with the following fundamental question: Do we learn anything by using convolutive ICA instead of instantaneous ICA? Or, put in another way, Should L be larger than zero?

To produce an instantaneous mixture we now generate two random signals from a Laplace distribution, filter them through filters of order 15 shown in Figure 3, and mix the two filtered sources using an arbitrary square mixing

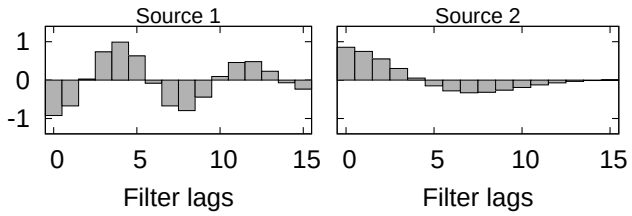


Figure 3: These filters are used to produce autocorrelated sources ($M = 15$).

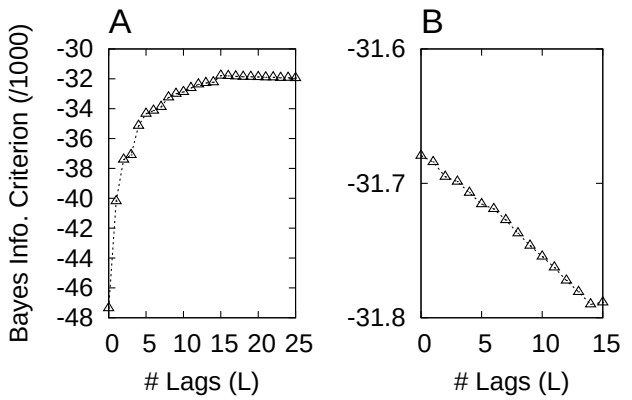


Figure 4: A: The result of using Bayesian model selection without allowing for an autofilter (M = 0). Since the signals are non-white, the validity of L is unquestioned even at 15 lags ($L = 15$). B: We fix $L + M = 15$, and now get the correct answer, that model information is largest for $L = 0$, meaning there is no evidence of convolutive mixing.

matrix. Figure 4A shows the result of using Bayesian model selection for this mixture without allowing for a filter ($M = 0$). This corresponds to model selection in a conventional convolutive model. Since the signals are non-white, L is detected and the model BIC simply increases as function of L up to the maximum, here stopped at $L = 15$. Next, (Fig. 4B) we fix $L + M = 15$. Models with a larger L have at least the same capability as models with lower L, though models with lower L are preferable because they have fewer parameters. By adding the source model, we get the correct answer in this case: These data contain no evidence of convolutive mixing.

5 Deconvolving an EEG ICA subspace

We will now show by example how cICA can be used to separate the delayed influences of statically defined ICA components on each other, thereby achieving a larger degree of independence in the convolutive component time courses. The procedure described here can be seen as a Diminished approach in which we extract K convolutive components from the D-dimensional data by deconvolving a K-dimensional subspace projection of the data. In [Dyrholm et al., 2004] we used a subspace from Principal Component Analysis (PCA), but as our experiment will show, using ICA for that projection has the benefit that the subspace can be chosen e.g. for physiological interest.

As a first test of this approach, we applied convolutive decomposition to 20 minutes of a 71-channel human EEG recording (20 epochs of 1 minute duration), downsampled for numeric convenience to a 50-Hz sampling rate after filtering between 1 and 25 Hz with phase-indifferent FIR filters. First, the recorded (channels-by-times) data matrix ($\mathbf{X}$) was decomposed using extended Infomax ICA [Bell and Sejnowski, 1995; Makeig et al., 1996; Jung et al., 1998; Lee et al., 1999; Jung et al., 2001] into 71 maximally independent components whose ('activation') time series were contained in (components-by-times) ma-

trix $\mathbf{S}^{\mathrm{ICA}}$ and whose ('scalp map') projections to the sensors were specified in (channels-by-components) mixing matrix $\mathbf{A}^{\mathrm{ICA}}$, assuming instantaneous linear mixing $\mathbf{X} = \mathbf{A}^{\mathrm{ICA}}\mathbf{S}^{\mathrm{ICA}}$.

Five of the resulting independent components (ICs) were selected for further analysis on the basis of event-related coherence results that showed a transient partial collapse of component independence following the subject button presses [Makeig et al., 2004b]. Their scalp maps from the relevant five columns of $\mathbf{A}^{\mathrm{ICA}}$ are shown on the left margin of Figure 7. Next, cICA decomposition was applied to the five component activation time series (relevant five rows of $\mathbf{S}^{\mathrm{ICA}}$), assuming the model

$$\mathbf{s}_t^{\mathrm{ICA}} = \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau}^{\mathrm{cICA}} \tag{31}$$

As a qualified guess of the order L, we applied the approach to estimating L outlined in Section 4.2 above to the EEG subspace data. First, we increased the order of the convolutive model L (keeping $M = 0$) while monitoring the BIC. To produce error bars, we used jackknife resampling [Efron and Tibshirani, 1993]; i.e. for each value of L, 20 runs with the algorithm were performed, one for each jackknifed epoch, thus the data in each run consisted of the 19 remaining epochs. Figure 5A shows the mean jackknifed BIC. Clearly, the BIC, without an autofilter included, was at least $L_{\max} = 40$, since some correlations in the data extended to at least 800 ms. Next, we swept the range of possible source model filters M, keeping $L + M = 40$. Figure 5B shows that $L = 10$, corresponding to a filter length of 200 ms, proved optimal.

Figures 6 shows the 5×5 matrix of learned convolutive kernels. Before plotting, we arranged the order of the five output CCs so that the diagonal ($CC_i \rightarrow IC_i$) kernels, shown in one-third scale in Fig. 6, were dominant.

Figure 7 shows the resulting percent of variance of the contributions from each of the CC innovations to each of the IC activations. As the large diagonal contributions in Figure 7 show, each *convolutive* CCj dominated one *spatially*

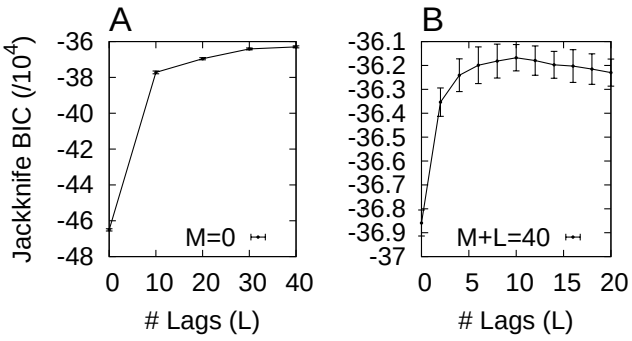


Figure 5: Using the protocol for detecting the order of L for EEG. A: There are correlations over at least 40 lags in the data. This corresponds to 800ms. B: By introducing the source model it turns out that L should only be on the order of 10 corresponding to 200 ms.

static IC (ICj). However, there were clearly significant off-diagonal contributions as well, indicating that spatiotemporal relationships between the static ICA components was captured by the cICA model.

To explore the robustness of this result further, we tested for the presence of delayed correlations, first between the static IC activations ($s_{k'}^{\mathrm{ICA}}(t)$) and then between the learned CC innovations ($s_k^{\mathrm{cICA}}(t)$). Figure 8 shows, for the most predictable IC and CC, the percent of their time course variances that was accounted for by linear prediction from the past history (of order r) of the largest contributing remaining ICs or CCs, respectively.

As expected from the cICA results, as the prediction order (r) increased, the predictability of the static ICA component activation also increased. For the ICA component activation, 9% of the variance could be explained by linear prediction from the previous 10 time points (200 ms) of another ICA component. The static ICA component time courses were nearly 'independent' only in the sense of zero-order prediction ($r = 0$), as expected from their derivation. Their lack of independence at other lags is compatible with the cICA results. For

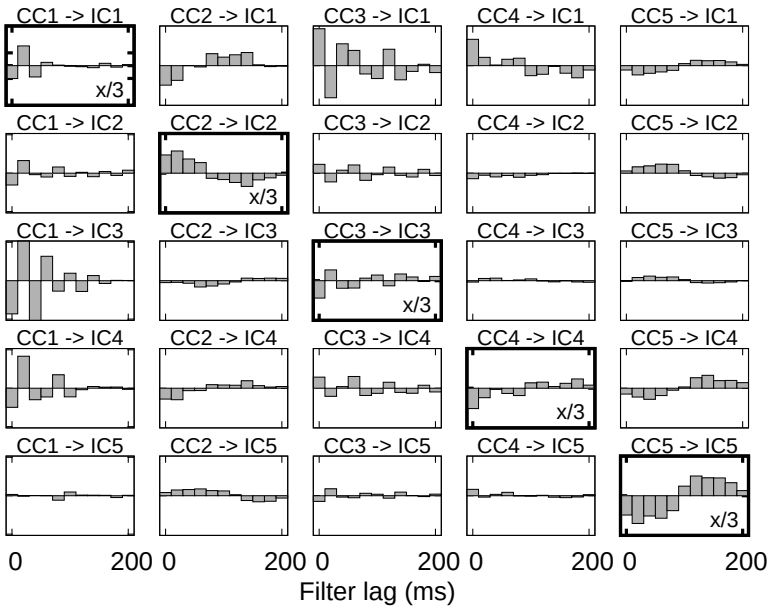


Figure 6: Kernels of the five derived convolutive ICA components (CCs), arranged (in columns) in order of their respectively contributions to the five static ICA components (ICs) (rows). Each CC made a dominant contribution to one IC; these were ordered so as to appear on the diagonal. Scaling of the diagonal kernels is one third that of the off-diagonal kernels.

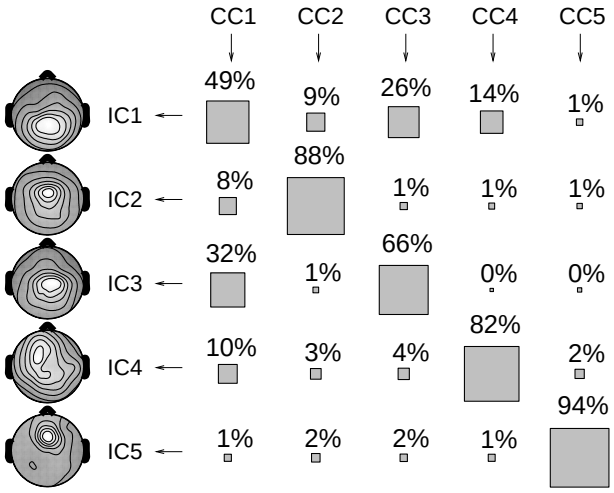


Figure 7: Percent variance of five static ICA components (ICs) accounted for by the five derived convolutive components (CCs). The IC scalp maps on the left are shown for interest. Contributions arranged on the diagonal are dominant. Squares represent the (rounded) percent variance of the IC activation time series accounted for by each CC. Significant off-diagonal elements indicate the presence of significant delayed spatiotemporal interactions between the static IC activations.

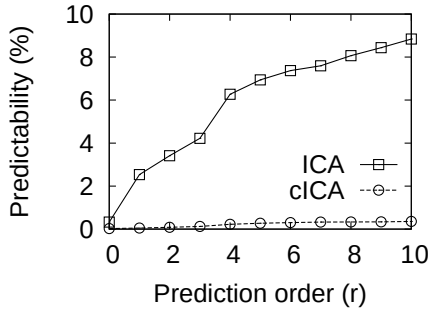


Figure 8: Predictability of the most predictable ICA component activation (IC3) and cICA component innovation (CC5) from the most predictive other IC and CC component, respectively (IC1, CC3).

the CC innovation, however, the predictability in Figure 8 remained low as r increased, indicating that cICA in fact deconvolved delayed correlations present in the EEG subspace data.

Figure 9 shows the power spectral densities for each of the IC activations (in bold traces) along with the two CCs (in thin traces) that, in accordance with Figure 7, contributed the most to the respective IC (c.f. Figure 7). Note that the broad alpha band spectral peak in IC1 (uppermost panel in Figure 9) around 10Hz has been split between CC1 and CC3. In the middle panel, note the distinct spectral contributions of CC1 and CC3 to the double alpha peak in the IC3 spectrum. As expected, the CCs made different spectral contributions to the IC time courses. For example, CC1 made different power spectral density contributions to IC1, IC3 and IC4.

6 Discussion

In general, the usefulness of any blind decomposition method applied to biological time series data is most likely relative to the fit between the assumptions of the algorithm and the underlying physiology and biophysics. Therefore it is important to consider the physiological basis of the delayed interactions between statically-defined independent component time courses we observed here, and the possible physiological significance of the derived convolutive component filters and time courses.

These results have at least two possible interpretations. First, static ICA decomposition in this case may have found a maximally-independent basis of a five or more dimensional subspace of spatially dynamic EEG processes. This explanation could be sensible if the five IC source areas were adjacent or overlapping, compatible with patterns of continuous spatial current flow across a single cortical region. However, in this case simple inverse source modeling using equivalent dipole modes (not shown) suggested that the five IC scalp maps

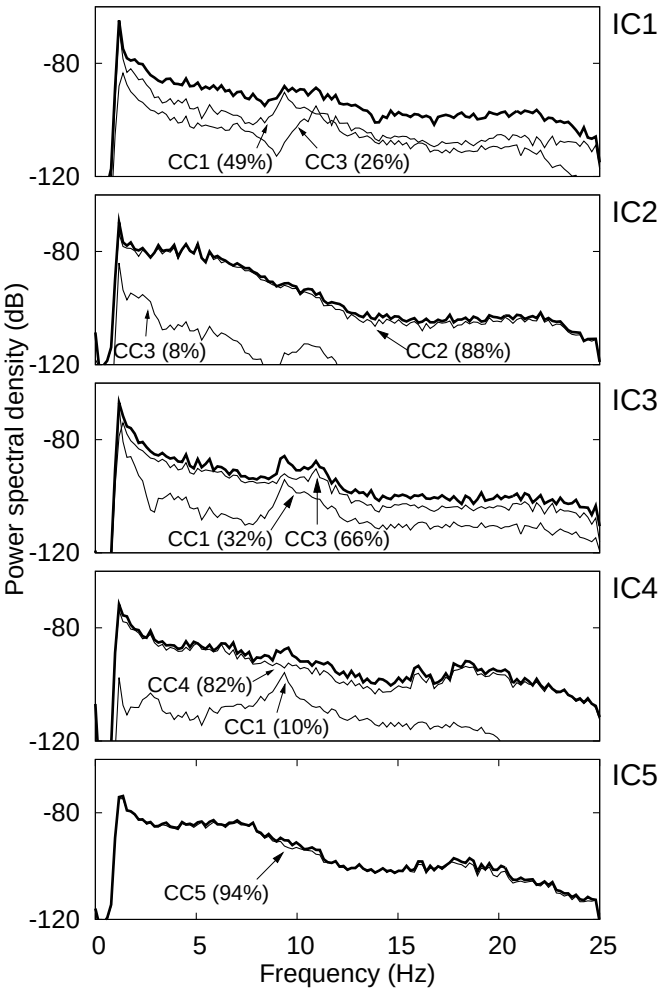


Figure 9: Power spectrum of the most powerful CC contributions to the five ICs.

used here might be associated with source activities generated in fairly well separated cortical territories. The physiological explanation for the observed lagged interactions between them thus might depend on delayed influences produced by neural spike-mediated communications from other cortical areas. These spike-mediated influences might not themselves produce far-field EEG signals at the scalp, but might add to the coherent source field oscillations occurring in the target source domain.

In this model, each cICA kernel would represent a local delayed EEG response in one ICA source area induced by cICA activity in another ICA source area. The cICA components then represent the local oscillatory (and/or other) EEG signal originating within each spatially separate ICA source domain, shorn of the delayed oscillatory influences arriving from other, distant cortical EEG source areas. Whatever the ultimate biological interpretation, the convolutional ICA data model presented here suggests that further study of delayed interactions between distinct EEG activities may be useful for modeling network dynamics underlying motor planning, attentional dynamics, and other cognitive processes that are known to involve simultaneous dynamic changes in multiple cortical regions [Makeig et al., 2002, 2004b].

Applied to these EEG data static ICA gave 15–20 components that were of physiological interest according to their spatial projections or activation time series, although we were not able to practically deconvolve more than five sources here because of numeric complexity. Open questions, therefore, are to identify independent component subspaces of interest for cICA decomposition and/or to explore the efficiency of performing cICA on larger computer clusters. In future, convolutive ICA might also be applied usefully to other types of biomedical time series data that involve stereotyped source movements, thus presenting problems for static ICA decomposition. These might include electrocardiographic (ECG) and brain hemodynamic measures such as diffusion tensor imaging (DTI)

[Anemüller et al., 2004].

References

Anemüller, J., Duann, J.-R., Sejnowski, T. J., and Makeig, S. (2004). Unraveling spatio-temporal dynamics in fmri recordings using complex ica. In Puntonet, C. G. and Prieto, A., editors, *Independent Component Analysis and Blind Signal Separation*, pages 1103–1110, Granada, Spain.

Anemüller, J. and Kollmeier, B. (2003). Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach. *IEEE Transactions on Speech and Audio Processing*, 39(1-2):79–95.

Anemüller, J., Sejnowski, T., and Makeig, S. (2003). Complex independent component analysis of frequency-domain eeg data. *Neural Networks*, 16:1313–1325.

Attias, H. and Schreiner, C. E. (1998). Blind source separation and deconvolution: the dynamic component analysis algorithm. *Neural Computation*, 10(6):1373–1424.

Baumann, W., Kohler, B.-U., K., D., and Orglmeister, R. (2001). Real time separation of convolutive mixtures. In Lee, T.-W., Jung, T.-P., Makeig, S., and Sejnowski, T. J., editors, *3'rd International Conference on Independent Component Analysis and Blind Signal Separation.*, pages 65–69, San Diego, CA, USA.

Bell, A. and Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159.

Belouchrani, A., Meraim, K. A., Cardoso, J.-F., and Moulines, É. (1997). A

blind source separation technique based on second order statistics. *IEEE Transacarions on Signal Processing*, 42:434–444.

Cardoso, J.-F. and Pham, D.-T. (2004). Optimization issues in noisy gaussian ica. In Puntonet, C. G. and Prieto, A., editors, *Independent Component Analysis and Blind Signal Separation*, pages 41–48, Granada, Spain.

Choi, S. and Cichocki, A. (1997). Blind signal deconvolution by spatio- temporal decorrelation and demixing. In Principe, J., Gile, L., Morgan, N., and Wilson, E., editors, *Neural Networks for Signal Processing*, pages 426–435, Amelia Island, CA, USA.

Choi, S., ichi Amari, S., Cichocki, A., and wen Liu, R. (1999). Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels. In *Independent Component Analysis and Blind Signal Separation*, pages 371–376, Aussois, France.

Comon, P., Moreau, E., and Rota, L. (2001). Blind separation of convolutive mixtures: A constrast based joint diagonalization appraoch. In Lee, T.-W., Jung, T.-P., Makeig, S., and Sejnowski, T. J., editors, *Independent Component Analysis and Blind Source Separation*, pages 686–691, San Diego, CA, USA.

Deligne, S. and Gopinath, R. (2002). An em algorithm for convolutive independent component analysis. *Neurocomputing*, 49:187–211.

Delorme, A. and Makeig, S. (2004). Eeglab: an open source toolbox for analysis of single-trial eeg dynamics. *Journal of Neuroscience Methods*, 134:9–21.

Delorme, A., Makeig, S., and Sejnowski, T. J. (2002). From single-trial eeg to brain area dynamics. *Neurocomputing*, 44-46:1057–1064.

Douglas, S. C., Cichocki, A., and Amari, S. (1999). Self-whitening algorithms for adaptive equalization and deconvolution. *IEEE Transactions on Signal Processing*, 47:1161–1165.

Dyrholm, M. and Hansen, L. K. (2004). CICAAR: Convolutive ICA with an auto-regressive inverse model. In Puntonet, C. G. and Prieto, A., editors, *Independent Component Analysis and Blind Signal Separation*, pages 594–601, Granada, Spain.

Dyrholm, M., Hansen, L. K., Wang, L., Arendt-Nielsen, L., and Chen, A. C. (2004). Convolutive ICA (c-ICA) captures complex spatio-temporal EEG activity. In *10th annual meeting of the organization for human brain mapping*.

Dyrholm, M., Makeig, S., and Hansen, L. K. (2005). Model structure selection in convolutive mixtures. In *(submitted) Independent Component Analysis and Blind Signal Separation*.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Hansen, P. C. (2002). Deconvolution and regularization with toeplitz matrices. *Numerical Algorithms*, 29:323–378.

Jung, T.-P., Humphries, C., Lee, T.-W., Makeig, S., McKeown, M. J., Iragui, V., and Sejnowski, T. J. (1998). Extended ICA removes artifacts from electroencephalographic recordings. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems*.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., and Sejnowski, T. J. (2000). *Psychophysiology*, 37:163–78.

Jung, T.-P., Makeig, S., McKeown, M. J., Bell, A., Lee, T.-W., and Sejnowski, T. J. (2001). Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–22.

Lee, T.-W., Bell, A. J., and Lambert, R. H. (1997a). Blind separation of delayed and convolved sources. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, pages 758–764.

Lee, T.-W., Bell, A. J., and Orglmeister, R. (1997b). Blind source separation of real world signals. In *International Conference Neural Networks*, pages 2129–2135, Houston, TX, USA.

Lee, T.-W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):417–441.

Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. (1996). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8:145–151.

Makeig, S., Debener, S., Onton, J., and Delorme, A. (2004a). Mining event-related brain dynamics. *Trends in Cognitive Science*, 8(5):204–210.

Makeig, S., Delorme, A., Westerfield, M., Townsend, J., Courchense, E., and Sejnowski, T. (2004b). Electroencephalographic brain dynamics following visual targets requiring manual responses. *PLoS Biology*.

Makeig, S., Enghoff, S., Jung, T.-P., and Sejnowski, T. J. (2000). A natural basis for efficient brain-actuated control. *IEEE Trans. Rehab. Eng.*, 8:208–211.

Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., and Sejnowski, T. J. (2002). Dynamic brain sources of visual evoked responses. *Science*, 295(5555):690–694.

Mitianoudis, N. and Davies, M. (2003). Audio source separation of convolutive mixtures. *IEEE Transactions on Speech and Audio Processing*, 11(5):489–497.

Moulines, É., Cardoso, J.-F., and Gassiat, E. (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3617–3620, Munich, Germany.

Neumaier, A. and Schneider, T. (2001). Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):27–57.

Nielsen, H. B. (2000). Ucminf - an algorithm for unconstrained, nonlinear optimization. Technical Report IMM-REP-2000-19, Department of Mathematical Modelling, Technical University of Denmark.

Onton, J., Delorme, A., and Makeig, S. (2005). Frontal midline eeg dynamics during working memory. *NeuroImage*, 27:342–356.

Parra, L. and Spence, C. (2000). Convolutive blind source separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8:320–327.

Parra, L., Spence, C., and Vries, B. (1997). Convolutive source separation and signal modeling with ml. In *International Symposium on Intelligent Systems*, Reggio Calbria, Italy.

Parra, L., Spence, C., and Vries, B. D. (1998). Convolutive blind source separation based on multiple decorrelation. In *Neural Networks for Signal Processing*, pages 23–32, Cambridge, UK.

Pearlmutter, B. A. and Parra, L. C. (1997). Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, pages 613–619.

Rahbar, K. and Reilly, J. (2001). Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2745–2748, Salk Lake City, Utah, USA.

Rahbar, K., Reilly, J. P., and Manton, J. H. (2002). A frequency domain approach to blind identification of mimo fir systems driven by quasi-stationary signals. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1717–1720, Orlando, Florida, USA.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Sun, X. and Douglas, S. (2001). A natural gradient convolutive blind source separation algorithms for speech mixtures. In Lee, T.-W., Jung, T.-P., Makeig, S., and Sejnowski, T. J., editors, *Independent Component Analysis and Blind Source Separation*, pages 59–64, San Diego, CA, USA.

Torkkola, K. (1996). Blind separation of convolved sources based on information maximization. In *Neural Networks for Signal Processing*, pages 423–432, Kyoto, Japan.

P.2 M. Dyrholm, S. Makeig and L. K. Hansen, Model Structure Selection in Convolutive Mixtures, ICA2006

[15] M. Dyrholm, S. Makeig and L. K. Hansen, Model structure selection in convolutive mixtures, (submitted) 6th International Conference on Independent Component Analysis and Blind Source Separation, 2006

Model structure selection in convolutive mixtures

Mads Dyrholm[1], Scott Makeig[2] and Lars Kai Hansen[1]

[1]Informatics and Mathematical Modelling
Technical University of Denmark, 2800 Lyngby, Denmark
mad,lkh@imm.dtu.dk
[2]Swartz Center for Computational Neuroscience
University of California, San Diego 0961, La Jolla CA 92093-0961
scott@sccn.ucsd.edu

Abstract. The CICAAR algorithm (convolutive independent component analysis with an auto-regressive inverse model) allows separation of white (i.i.d) source signals from convolutive mixtures. We introduce a source color model as a simple extension to the CICAAR which allows for a more parsimoneous representation in many practical mixtures. The new filter-CICAAR allows Bayesian model selection and can help answer questions like: 'Are we actually dealing with a convolutive mixture?'. We try to answer this question for EEG data.

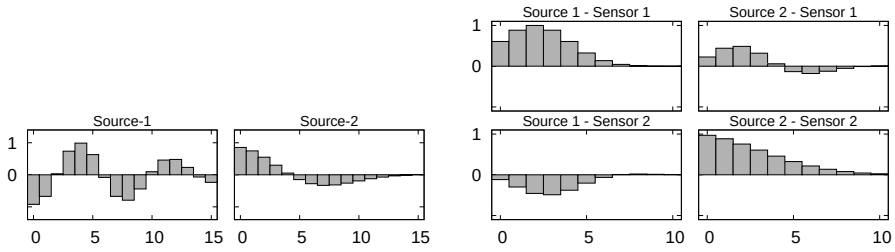
1 Introduction

Convolutive ICA (CICA) is a topic of high current interest and several schemes are now available for recovering mixing matrices and sources signals from convolutive mixtures, see e.g., [4]. Convolutive models are more complex than conventional instantaneous models, hence, the issue of model optimization is important. Convolutive ICA in its basic form concerns reconstruction of the L+1 mixing matrices $\mathbf{A}_\tau$ and the N source signal vectors $\mathbf{s}_t$ of dimension K, from a D-dimensional convolutive mixture

$$\mathbf{x}_t = \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau} \tag{1}$$

Here we focus, for simplicity, on the case where the number of sources equals the number of sensors, $D = K$.

We have earlier proposed the CICAAR approach for convolutive ICA [3] as a generalization of Infomax [2] to convolutive mixtures. The CICAAR exploits the relatively simple structure of the un-mixing system resulting when the inverse mixing is represented as an autoregressive process. In the original derivation we were forced to assume white (i.d.d) sources, i.e., that all temporal correlation in the mixture signals appeared through the convolutive mixing process. A more economic representation is obtained, however, if we explicitly introduce filters to represent possible auto-correlation of sources. This added degree of freedom also



(a) Source color filters, $M = 15$. (b) Convolutive mixing system, $L = 10$.

Fig. 1. Filters for generating synthetic data. First, two i.i.d. signals are colored through their respective filters (a). Then, the colored signals are convolutively mixed using a distinct filter for each source-sensor path (b).

carries another benefit, it allows for optimizing the model structure: How much correlation should be accounted for by the source filters, and how much should be accounted for by the convolutive mixture? Explicit source auto-correlation modeling using filtered white noise has been proposed earlier by several authors, see e.g., [1, 7, 8].

2 Modelling convolutive ICA with auto-correlated sources

We introduce a model for each of the sources

$$s_k(t) = \sum_{\lambda=0}^{M} h_k(\lambda) z_k(t-\lambda) \tag{2}$$

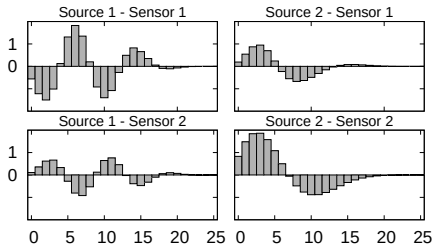
where $z_k(t)$ represents a whitened version of the source signal. The negative log likelihood for the model combining (1) and (2) is given by

$$\mathcal{L} = N \log |\det \mathbf{A}_0| + N \sum_k \log |h_k(0)| - \sum_{t=1}^{N} \log \mathrm{p}(\hat{\mathbf{z}}_t) \tag{3}$$

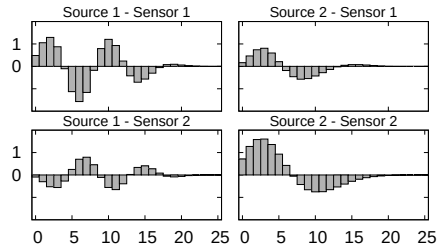
where $\hat{\mathbf{z}}_t$ is a vector of whitened source signal estimates at time t using an operator that represents the inverse of (2). We can without loss of generality set $h_k(0) = 1$, then

$$\mathcal{L} = N \log |\det \mathbf{A}_0| - \sum_{t=1}^{N} \log \mathrm{p}(\hat{\mathbf{z}}_t) \tag{4}$$

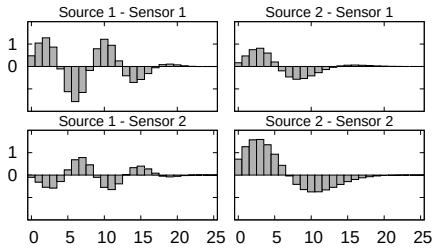
The number of parameters in this model is $D^2(L+1) + DM$, and it can thus be minimized if M is increased so as to explain the source auto-correlations allowing L to be reduced in return. An algorithm for convolutive ICA which includes the source model can be derived by making a relative straight forward modification to the equations of the CICAAR algorithm found in [3], see appendix A.



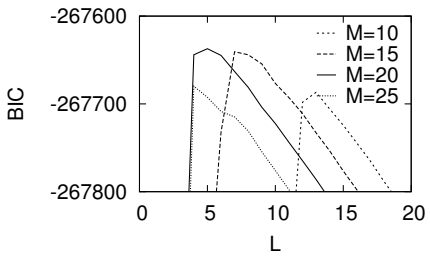
(a) Generating model $(L, M) = (10, 15)$



(b) Estimated by algorithm $(L, M) = (10, 15)$



(c) Estimated by algorithm $(L, M) = (5, 20)$



(d) BIC optimal is $(L, M) = (5, 20)$.

Fig. 2. Mixing filters convolved with respective color filters. (a) for the generating model. (b) for an estimated model with the 'true' L and M. (c) for the Bayes optimal model with $(L, M) = (5, 20)$. (d) shows the BIC for various models, and (L,M)=(5,20) is found optimal.

3 Model Selection Protocol

Let $\mathcal{M}$ represent a specific choice of model structure (L, M). The Bayes Information Criterion (BIC) is given by $\log \mathrm{p}(\mathcal{M}|\mathbf{X}) \approx \log \mathrm{p}(\mathbf{X}|\boldsymbol{\theta}_0, \mathcal{M}) - \frac{\dim \boldsymbol{\theta}}{2} \log N$ where $\dim \boldsymbol{\theta}$ is the number of parameters in the model, and $\boldsymbol{\theta}_0$ are the maximum likelihood parameters [9].

We propose a simple protocol for the dimensions (L, M) of the convolutional- and source-filters. First, expand the convolution length L without a source model (i.e. keeping $M = 0$). This will model the total temporal dependency structure of the system. The optimal L, denote it $L_{\max}$, is found by monitoring BIC. Next, expand the dimensions M of the source model filters while keeping the temporal dependency constant, i.e. keeping $(L + M) = L_{\max}$.

3.1 Simulation example

The first experiment is designed to illustrate the protocol for determining the dimensions of the convolution and the source filters. We create a 2×2 system with known source filters $M = 15$ and known convolution $L = 10 \ldots$

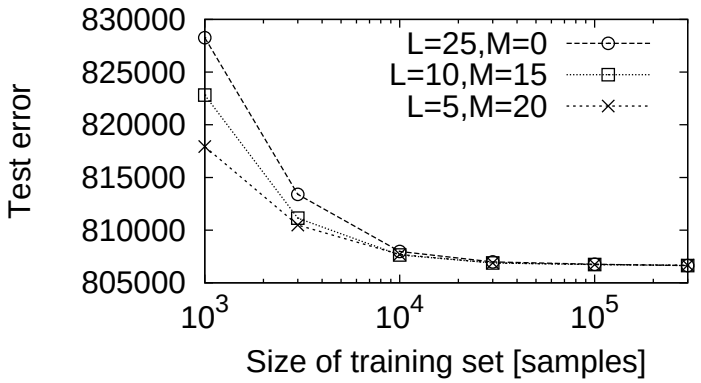


Fig. 3. Learning curves for three models: The generating model $(L, M) = (10, 15)$, a model with $(L, M) = (25, 0)$ which is more complex but fully capable of 'imitating' the first model, and the model $(L, M) = (5, 20)$ which was found Bayes optimal according to BIC. The generalization error is estimated as the likelihood of a test set ($N_{\text{test}} = 300000$). The uniform improvements in generalization of the 'optimal model' further underlines the importance of model selection in the context

Data — Two signals are generated by filtering temporally white signals using the filters shown on Figure-1(a). The signals are then mixed using the $2 \times 2 \times 10$ system shown on Figure-1(b). The generating model has thus $(L, M) = (10, 15)$.

Result — First we note, the model is in itself ambiguous; an arbitrary filter can be applied to a color filter if the inverse filter is applied to the respective column of mixing filters. Therefore, to compare results we inspect the system as a whole, i.e. source color convolved with a column of mixing filters.

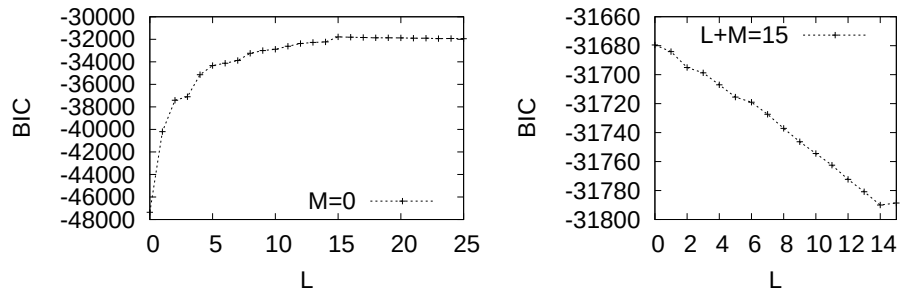
Figure-2 displays convolutive mixing systems where each mixing channel has been convolved with the respective color filter; (a) for the true generating model; (b) a run with the algorithm using $N = 300000$ training samples and using the (L, M) of the generating model. The result is perfect up to sign and scaling ICA ambiguities; (c) shows a run with the algorithm using $N = 100000$ and the Bayes optimal choice of $(L, M) = (5, 20)$ c.f. (d), in the finite data the protocol has found a parsimonious model with similar overall transfer function. We first study the learning curves, i.e., how does the training set dimension N, influence learning. We use the likelihood evaluated on a test set to measure the learning of different models. We now compare learning curves for three models; one which is the generating model $(L, M) = (10, 15)$, one $(L, M) = (25, 0)$ which is more complex but fully capable of imitating the first model, and $(L, M) = (5, 20)$ which is optimal according to BIC. Figure-3 shows learning curves of the three models, the test set is $N_{\text{test}} = 300000$ samples. The uniform improvements in generalization of the 'optimal model' further underlines the importance of model selection in the context of convolutive mixing.

3.2 Rejecting convolution in an instantaneous mixture

We will now illustrate the importance of the source color filters when dealing with the following fundamental question: 'Do we learn anything by using Convolutive ICA instead of instantaneous ICA?'—or put in another way: 'should L be larger than zero?'.

Data — To produce an instantaneous mixture we now mix the two colored sources from before using a random matrix.

Result — Figure-4(a) shows the result of using Bayesian model selection without allowing for a filter ($M = 0$). This corresponds to model selection in a conventional convolutive model. Since the signals are non-white L is detected and the model BIC simply increases as function of L up to the maximum which is attained at a value of $L = 15$. Next, in Figure-4(b) we fix $L + M = 15$. Models with a greater L have at least the same capability as a model with a lower L; but as expected lower L are preferable because the models has fewer parameters. Thus, thanks to the filters, we now get the correct answer: 'There is no evidence of convolutive ICA'.



(a) $M = 0$ (b) $M + L = 15$

Fig. 4. (a) the result of using Bayesian model selection without allowing for a filter ($M = 0$). Since the signals are non-white L is detected at a value of $L = 15$. (b) we fix $L + M = 15$, and now get the correct answer: $L = 0$ — 'There is no evidence of convolutive ICA'.

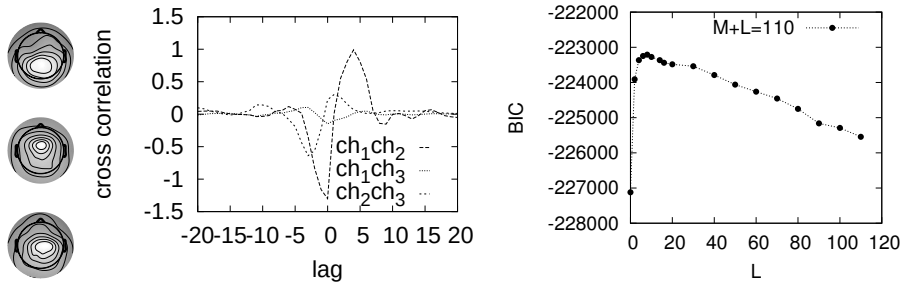
4 Is convolutive ICA relevant for EEG?

The EEG signals from the entire brain superimpose onto every EEG electrode instantaneously; there are no delays or echoes, hence, the mixing of the electromagnetic activity is definitely not a convolutive process. However, the question is whether the convolutive mixing model is relevant as a model for the brain

activity itself. It is well known that EEG activity exhibits rich spatio-temporal dynamics and that different tasks of the brain combine different regions in different frequency bands, and so, we expect the Bayes optimal model to potentially include some convolutive mixing $L > 0$.

Data — 20 minutes of a 71-channel human EEG recording downsampled to a 50-Hz sampling rate after filtering between 1 and 25 Hz with phase-indifferent FIR filters. First, the recorded (channels-by-times) data matrix ($\mathbf{X}$) was decomposed using extended infomax ICA [2, 5] into 71 maximally independent components whose ('activation') time series were contained in (components-by-times) matrix $\mathbf{S}^{\mathrm{ICA}}$ and whose ('scalp map') projections to the sensors were specified in (channels-by-components) mixing matrix $\mathbf{A}^{\mathrm{ICA}}$, assuming instantaneous linear mixing $\mathbf{X} = \mathbf{A}^{\mathrm{ICA}}\mathbf{S}^{\mathrm{ICA}}$. Three of the resulting independent components were selected for further analysis on the basis of event-related coherence results that showed a transient partial collapse of component independence following the subject button presses [6]. Their scalp maps (the relevant three columns of $\mathbf{A}^{\mathrm{ICA}}$) are shown on Figure 5(a).

Convolutive ICA analysis — Next, convolutive ICA decomposition was applied to the three component activation time series (relevant three rows of $\mathbf{S}^{\mathrm{ICA}}$) which we shall refer to as channels ch_1, ch_2 and ch_3. Following our proposed protocol, we find $L_{\max} = 110$, then $L = 9$ as shown on Figure-5(c) — so, we are in fact dealing with a convolutive mixture. Figure-5(b) shows, for one of the resulting convolutive ICA components, cross correlation functions between its contribution to the channels (with each a scalp map associated). Clearly, there are delayed correlation between the different brain regions, and this is not possible to model with an instantaneous ICA model, hence the need for convolutive mixing.



(a) (b) (c) We find $L = 9$

Fig. 5. (a) Scalp maps for the three ICA components. (b) for one of the resulting convolutive ICA components, cross correlation functions between its contribution to the channels. (c) Finding $L = 9$ for the EEG data.

5 Conclusion

We have incorporated filters for modelling possible source auto-correlations into an existing algorithm for convolutive ICA. We have proposed a protocol for determining the dimension L of a convolutive mixture utilizing the filters. We have shown that convolutive ICA is relevant for real EEG data.

Appendix A: Source modeling with the CICAAR algorithm

For notational convenience we introduce the following matrix notation instead of (2), handling all sources in one matrix equation

$$\mathbf{s}_t = \sum_{\lambda=0}^{M} \mathbf{H}_\lambda \mathbf{z}_{t-\lambda} \tag{5}$$

where the $\mathbf{H}_\lambda$'s are diagonal matrices defined by $(\mathbf{H}_\lambda)_{ii} = h_i(\lambda)$.

Given a current estimate of the mixing matrices $\mathbf{A}_\tau$ and the source filter coefficients $h_k(\lambda)$, First apply equation 7 of [3] to obtain $\hat{\mathbf{s}}_t$. Then apply the inverse source coloring operator

$$\hat{\mathbf{z}}_t = \hat{\mathbf{s}}_t - \sum_{\lambda=1}^{M} \mathbf{H}_\lambda \hat{\mathbf{z}}_{t-\lambda} \tag{6}$$

which must replace $\hat{\mathbf{s}}_t$ in [3] (in equations 6,8,9 and 11). This involves the following partial derivatives which in turn uses the result from [3] (from equations 7,10,12)

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{B}_\tau)_{ij}} = \frac{\partial(\hat{\mathbf{s}}_t)_k}{\partial(\mathbf{B}_\tau)_{ij}} - \sum_{\lambda=1}^{M} \mathbf{H}_\lambda \frac{\partial(\hat{\mathbf{z}}_{t-\lambda})_k}{\partial(\mathbf{B}_\tau)_{ij}} \tag{7}$$

where $\mathbf{B}_\tau = \mathbf{A}_\tau$ for $\tau > 0$ and $\mathbf{B}_0 = \mathbf{A}_0^{-1}$. Furthermore

$$\frac{\partial(\hat{\mathbf{z}}_t)_k}{\partial(\mathbf{H}_\lambda)_{ii}} = -\delta(k-i)(\hat{\mathbf{z}}_{t-\lambda})_i - \left(\sum_{\lambda'=1}^{M} \mathbf{H}_{\lambda'} \frac{\partial \hat{\mathbf{z}}_{t-\lambda'}}{\partial(\mathbf{H}_\lambda)_{ii}} \right)_k \tag{8}$$

The work involved in this plug-in is minimal due to the diagonal structure of the $\mathbf{H}_\lambda$ matrices. Finally,

$$\frac{\partial \mathcal{L}}{\partial(\mathbf{H}_\lambda)_{ii}} = -\sum_{t=1}^{N} \boldsymbol{\psi}_t^T \frac{\partial \hat{\mathbf{z}}_t}{\partial(\mathbf{H}_\lambda)_{ii}} \tag{9}$$

where $(\boldsymbol{\psi}_t)_k = p'((\hat{z}_t)_k)/p((\hat{z}_t)_k)$.

References

1. Hagai Attias and C. E. Schreiner. Blind source separation and deconvolution: the dynamic component analysis algorithm. *Neural Computation*, 10(6):1373–1424, 1998.
2. Tony Bell and Terrence Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
3. M. Dyrholm and L. K. Hansen. CICAAR: Convolutive ICA with an auto-regressive inverse model. In Carlos G. Puntonet and Alberto Prieto, editors, *Independent Component Analysis and Blind Signal Separation*, volume 3195, pages 594–601, sep 2004.
4. A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons., 2001.
5. Scott Makeig, Anthony J. Bell, Tzyy-Ping Jung, and Terrence J. Sejnowski. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8:145–151, 1996.
6. Scott Makeig, Arnaud Delorme, Marissa Westerfield, Tzyy-Ping Jung, Jeanne Townsend, Eric Courchesne, and Terrence J. Sejnowski. Electroencephalographic brain dynamics following manually responded visual targets. *PLoS Biology*, 2004.
7. L. Parra, C. Spence, and B. Vries. Convolutive source separation and signal modeling with ml. In *International Symposioum on Inteligent Systems*, 1997.
8. Barak A. Pearlmutter and Lucas C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 613. The MIT Press, 1997.
9. G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

P.3 M. Dyrholm and L. K. Hansen, CICAAR: Convolutive ICA with an Auto-Regressive Inverse Model, ICA2004

[13] M. Dyrholm and L. K. Hansen, CICAAR: Convolutive ICA with an Auto-Regressive Inverse Model, Independent Component Analysis and Blind Signal Separation, pp. 594-601, 2004

CICAAR: Convolutive ICA with an Auto-Regressive Inverse Model

Mads Dyrholm and Lars Kai Hansen

Informatics and Mathematical Modelling
Technical University of Denmark
2800 Kgs. Lyngby, Denmark

Abstract. We invoke an auto-regressive IIR inverse model for convolutive ICA and derive expressions for the likelihood and its gradient. We argue that optimization will give a stable inverse. When there are more sensors than sources the mixing model parameters are estimated in a second step by least squares estimation. We demonstrate the method on synthetic data and finally separate speech and music in a real room recording.

1 Introduction

Independent component analysis (ICA) of convolutive mixtures is a key problem in signal processing, the problem is important in speech processing and numerous other applications including medical, visual, and industrial signal processing, see, e.g., [1–5] . Convolutive ICA in its basic form concerns reconstruction of the $L+1$ mixing matrices A_τ and the N source signal vectors s_t of dimension K, from a D-dimensional convolutive mixture,

$$x_t = \sum_\tau A_\tau s_{t-\tau}. \tag{1}$$

We will assume L so large that all correlations in the process x can be 'explained' by the mixing process, and the source signal vectors are assumed temporally independent: $p(\{s_t\}) = \prod_{t=1}^N p(s_t)$. This is motivated by the observation that source signal auto-correlations can not be identified without additional a priori information [1]. This is most apparent in the frequency domain $A_\omega s_\omega$. A non-zero 'filter' $h(\omega)$ can be multiplied on a given source if $1/h(\omega)$ is applied to the corresponding column of the set of Fourier transformed mixing matrices A_ω.

Statistically motivated maximum likelihood schemes have been proposed, see e.g. [1,6–8]. The likelihood approach is attractive for a number of reasons. First, it forces a declaration of the statistical assumptions—in particular the a priori distribution of the source signals, secondly, the maximum likelihood solution is asymptotically optimal given the assumed observation model and the prior choices for the 'hidden' variables.

IIR representations of an inverse model have been proposed in e.g. [9,10]. In this paper we will invoke an auto-regressive IIR inverse model. This involves a

linear recursive filter for estimation of the source signal and a non-linear recursive filter for maximum likelihood estimation of the mixing matrices. Our derivation formally allows the number of sensors to be greater than the number of sources.

2 Estimating the sources through a stable inverse

Let us define x, A, and s such that $x = As$ is a matrix product abbreviation of the convolutive mixture

$$\begin{bmatrix} x_N \\ x_{N-1} \\ \vdots \\ x_1 \end{bmatrix} = \begin{bmatrix} A_0 & A_1 & \dots & A_L & & \\ & A_0 & A_1 & \dots & A_L & \\ & & & \ddots & & \\ & & & & & A_0 \end{bmatrix} \begin{bmatrix} s_N \\ s_{N-1} \\ \vdots \\ s_1 \end{bmatrix} \tag{2}$$

which allows the likelihood to be written $p(x|\{A_\tau\}) = \int \delta(x - As)p(s)ds$.

2.1 Square case likelihood

In the square case, $D = K$, the likelihood integral evaluates to

$$p(x|\{A_\tau\}) = |\det A|^{-1} p(A^{-1}x). \tag{3}$$

Since A is upper block triangular we obtain $p(x|\{A_\tau\}) = |\det A_0|^{-N} p(A^{-1}x)$, furthermore, assuming i.i.d. source signals we finally get

$$p(\{x_t\}|\{A_\tau\}) = |\det A_0|^{-N} \prod_{t=1}^{N} p((A^{-1}x)_t). \tag{4}$$

The inverse operation $A^{-1}x$ is the multivariate AR(L) process

$$\tilde{s}_t = A_0^{-1} x_t - A_0^{-1} \sum_{\tau=1}^{L} A_\tau \tilde{s}_{t-\tau} \tag{5}$$

which follows simply by eliminating s_t in (1). In terms of (5) we now rewrite the negative log likelihood

$$\mathcal{L}(\{A_\tau\}) = N \log |\det A_0| - \sum_{t=1}^{N} \log p(\tilde{s}_t) \ , \ K = D. \tag{6}$$

2.2 Overdetermined case likelihood

When $D > K$ there are many inverse operations $A^{-1} : \mathbb{R}^D \mapsto \mathbb{R}^K$ which satisfy $A^{-1}A = I$. In this work we base the source estimates $\hat{s}_t$ on a particular choice of inverse operation, i.e. we define $\hat{s} = A^{-1}x$ by the multivariate AR(L) process

$$\hat{s}_t = A_0^{\#} x_t - A_0^{\#} \sum_{\tau=1}^{L} A_\tau \hat{s}_{t-\tau}, \tag{7}$$

where $A_0^{\#}$ denotes Moore-Penrose generalized inverse. The process (7) is inverse in the sense $A^{-1}A = I$ which means that when it is configured with the true mixing matrices it allows perfect reconstruction of the sources. Evoking (7) the likelihood integral can be evaluated to

$$\mathcal{L}(\{A_\tau\}) = \frac{N}{2}\log|\det A_0^T A_0| - \sum_{t=1}^{N}\log p(\hat{s}_t)\ ,\ K \leq D. \tag{8}$$

The derivation of (8) is deferred to Sec. A for aesthetic reason, but note that (8) is based on our particular choice of inverse (7). For $K = D$ we note that (7) and (8) are identical to (5) and (6) respectively.

2.3 Optimization yields a stable inverse

In praxis, convolution system matrices such as A are often found to be poorly conditioned and hence the inverse problem $\hat{s} = A^{-1}x$ sensitive to noise, see e.g. [11]. The extreme case for the inverse is it being *unstable* and sensitive to machine precision rounding errors. Fortunately, the maximum likelihood approach has a built-in regularization against this problem. This is seen from the likelihood noting that an ill-conditioned estimator $\{\hat{A}_\tau\}$ will lead to a divergent source estimate $\hat{s}_t$; but such large amplitude signals are exponentially penalized under the source pdf's typically used in ICA ($p(s) = \mathrm{sech}(s)/\pi$). Therefore, our proposition is that it is 'safe' to use an iterative learning scheme for optimizing (8) because once it has been initialized with a well-conditioned convolution matrix A a learning decrease in (8) will lead to further refinements $\{\hat{A}_\tau\}$ which are stable in the context of equation (7). If no exact stable inverse exists the Maximum-Likelihood approach will give us a regularized estimator.

We propose here to use a gradient optimization technique. The gradient of the negative log likelihood w.r.t. $A_0^{\#}$ is given by

$$\frac{\partial\mathcal{L}(\{A\})}{\partial(A_0^{\#})_{ij}} = -N(A_0^T)_{ij} - \sum_{t=1}^{N}\psi^T(\hat{s}_t)\frac{\partial\hat{s}_t}{\partial(A_0^{\#})_{ij}} \tag{9}$$

where

$$\frac{\partial(\hat{s}_t)_k}{\partial(A_0^{\#})_{ij}} = \delta(i-k)\left(x_t - \sum_{\tau=1}^{L}A_\tau\hat{s}_{t-\tau}\right)_j - \left(A_0^{\#}\sum_{\tau=1}^{L}A_\tau\frac{\partial\hat{s}_{t-\tau}}{\partial(A_0^{\#})_{ij}}\right)_k \tag{10}$$

and $(\ \psi(\hat{s}_t)\ )_k = p'(\ (\hat{s}_t)_k\ )/p(\ (s_t)_k\ )$. The gradient w.r.t. to the other mixing matrices is given by

$$\frac{\partial\mathcal{L}(\{A\})}{\partial(A_\tau)_{ij}} = -\sum_{t=1}^{N}\psi^T(\hat{s}_t)\frac{\partial\hat{s}_t}{\partial(A_\tau)_{ij}} \tag{11}$$

where

$$\frac{\partial(\hat{s}_t)_k}{\partial(A_\tau)_{ij}} = -(A_0^{\#})_{ki}(\hat{s}_{t-\tau})_j - \left(A_0^{\#}\sum_{\tau'=1}^{L}A_{\tau'}\frac{\partial\hat{s}_{t-\tau'}}{\partial(A_\tau)_{ij}}\right)_k \tag{12}$$

These expressions allow for general gradient optimization schemes. A starting point for the algorithm is A_0 being random numbers and $A_\tau = 0$ for $\tau \neq 0$ — a stable initialization according to (7).

2.4 Re-estimating the mixing filters

When the dimension of x_t is strictly greater than the number of sources, $D > K$, the mixing matrices which figure as parameters for the learning process can not be taken as mixing filter estimates because $AA^{-1} \neq I \Rightarrow \hat{A}\hat{s} \neq x$. Instead we here propose to estimate the mixing filters by least-squares. Multiplying (1) with $s_{t-\lambda}^T$ from right and taking the expectation we obtain the following normal equations

$$< x_t s_{t-\lambda}^T > = \sum_\tau A_\tau < s_{t-\tau} s_{t-\lambda}^T > \tag{13}$$

which is solved for A_τ by regular matrix inversion using the estimated sources and $< \cdot >= \frac{1}{N}\sum_{1=1}^{N}$. This system is unlikely to be ill conditioned because the sources are typically uncorrelated mutually and temporally.

2.5 Dimensionality reduction

For lowering the training complexity we here propose to use a K-dimensional subspace representation of the data $y_t = U_K^T x_t$ where $U_K \in \mathbb{R}^{D\times K}$ is a projection. We can write a regular convolutive mixture where the number of sensors is now equal to K,

$$y_t = \sum_{\tau=0}^{L} B_\tau s_{t-\tau} \ , \ B_\tau = U_K^T A_\tau, \tag{14}$$

and note that the sources are unaltered by the projection. This means that we should be able to recover the sources from the projection using the square case of our algorithm. Once the sources have been estimated the D-by-K mixing matrices $\{A\tau\}$ are estimated c.f. Sec 2.4.

3 Experiments

3.1 Simulation data

We now illustrate the algorithm on a three-dimensional convolutive mixture of two sources, i.e. $D = 3$, $K = 2$. The true mixing filters are shown in the left panel of Fig.1 and set to decay within 30 lags, i.e. $L = 30$. The source signals, $N = 30000$, are both drawn from a Laplace distribution. 5000 consecutive samples is zeroed out from one of the sources, say 'Source-1'. Results are then evaluated from the estimated Source-1 by measuring the interference power P_i in the period where the true Source-1 is silent. We here define the Signal to Interference Ratio (SIR) $P_\mathrm{s}/P_\mathrm{i}$, where P_s is the signal power which is estimated in a period where both sources are active.

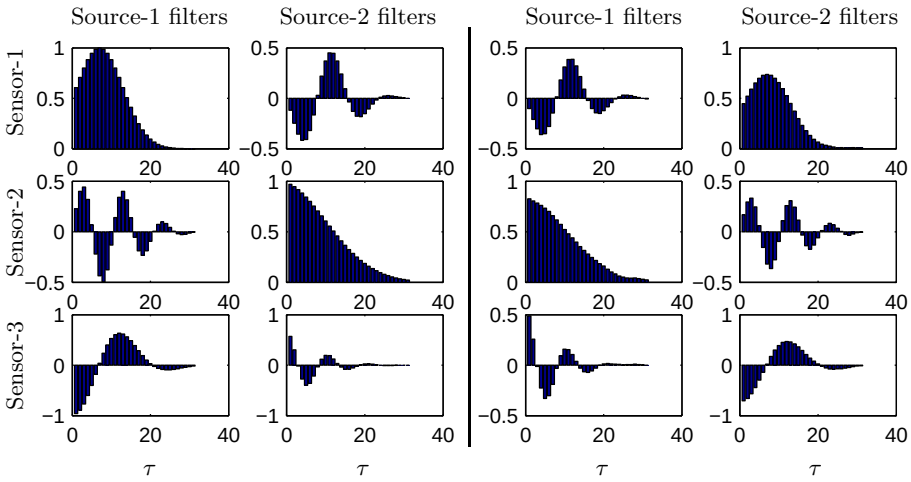


Fig. 1. (left) true mixing filters, (right) estimated mixing filters.

The data is projected onto the two major principal components and the sources $\hat{s}_t$ are estimated c.f. Sec. 2.5. The optimization scheme is Newton steps, i.e. updating $\{\hat{A}_\tau\}$ by $-H^{-1}g$ where g is the gradient vector and H^{-1} is the inverse Hessian which is estimated using the outer product approximation update per sample (see e.g. [12, page 153]). Convergence detected in 124 iterations. Obtained SIR = 19.3dB. The corresponding mixing filters estimated by (13) are then used as a starting guess for the general overdetermined algorithm using the original three-dimensional data as input. Convergence detected in 20 iterations. Obtained SIR = 34.2dB. Then we use (13) to estimate the corresponding mixing filters and the result is displayed in the right pane of Fig. 1.

3.2 Real audio recording

We now apply the proposed method to a 16kHz signal which was recorded indoor by two microphones and produced by a male speaker counting one-ten and a loud music source respectively. The microphones and the sources were located in the corners of a square. The signal is kindly provided by Dr. T-W. Lee, and is identical to the one used in [13]. We choose the number of mixing matrices $L = 50$. This time we use a BFGS Quasi-Newton optimization scheme (see e.g. [12, page 288]) convergence is reached in 490 iterations.

As noted, the source signals can only be recovered up to an arbitrary filter and we experience indeed a whitening effect on the sources. In [13] a low-pass filter was applied to overcome the whitening effect, hence, to make the sources 'sound more real'. In our presentation, because we have the forward model parameters, we reconstruct the microphone signals separately as they would sound if the other source was shut. This is simply achieved by propagating the given source signal through the estimated mixing model. Fig. 2 shows the recorded mixture

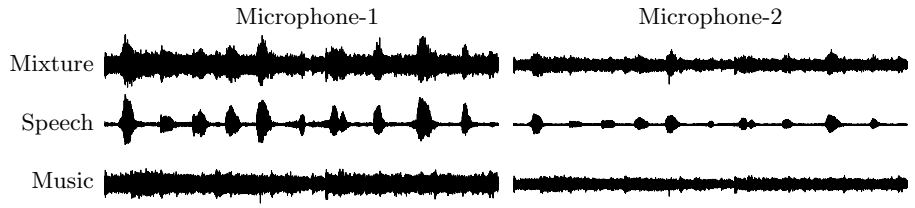


Fig. 2. Separation of real world sound signals. (Top row) The recorded mixture of speech and music. (Middle row) Separated speech reconstructed in the sensor domain. (Bottom row) Separated music reconstructed in the sensor domain.

along with the results of separation. For listening test and further analysis we have placed the resulting audio files at URL `http://www.imm.dtu.dk/~mad/cicaar/sound.html`. Again we evaluate the result by SIR; the interference power P_{i} as the mean power in ten manually segmented intervals in which the speaker is silent, and the signal power P_{s} is similarly estimated as the mean power in ten manually intervals where the speaker is clearly audible (and subtracting off the interference power). The SIR of the proposed algorithm and using the parameters described is SIR = 12.42 dB. The algorithm proposed by Parra and Spence [2] represents a state-of-the-art alternative for evaluation of performance. In the following table we give SIR's for the Parra-Spence algorithm using the implementation kindly provided by Stefan Harmeling[1] based on window lengths (N) and for three different numbers of un-mixing matrices (Q):

SIR (dB)	$Q = 50$	$Q = 100$	$Q = 200$
$N = 512$	11.9	11.8	12.3
$N = 1024$	12.0	12.2	12.5
$N = 2048$	11.9	12.0	12.3

The table indicates that in order to obtain a separation performance similar to that of the proposed algorithm the Parra-Spence inverse filter Q needs to be somewhat larger than the length of the IIR filter $L = 50$ we have used. Future quantitative studies are needed to substantiate this finding invoking a wider variety of signals and interferences.

4 Conclusion

We have proposed a maximum-likelihood approach to convolutive ICA in which an auto-regressive inverse model is put in terms of the forward model parameters. The algorithm leads to a stable (possibly regularized) inverse and formally allows the number of sensors to be greater than the number of sources. Our experiment shows good performance in a real world situation. In general, for *perfect* separation a stable un-regularized inverse must exist. An initial delay,

[1] `http://ida.first.gmd.de/~harmeli/download/download_convbss.html`

e.g., is not minimum phase and no causal inverse exist. On the other hand, in that case, the source can simply be delayed and thus remove the initial delay in the filter — exploiting the filter ambiguity. Such manoeuvre will in some cases make a real room impulse response minimum phase [14].

A Derivation of the likelihood in the overdetermined case

We shall make use of the following definition: $\hat{s}_t(s_{t-1}, s_{t-2}, \ldots, s_{t-L}) \equiv A_0^{\#} x_t - A_0^{\#} \sum_{\tau=1}^{L} A_\tau s_{t-\tau}$. We can write the likelihood

$$p(X|\{A_\tau\}) = \int_{s_1} \int_{s_2} \cdots \left(\int_{s_N} p(s_N)\delta(f_N) ds_N \right) \prod_{t=1}^{N-1} p(s_t)\delta(f_t) ds_1 \ldots ds_{N-1}. \tag{15}$$

where $f_t \equiv x_t - \sum_{\tau=0}^{L} A_\tau s_{t-\tau}$. The first step in this derivation is to marginalize out s_N, using

$$\int_{s_N} p(s_N)\delta(f_N) ds_N = |A_0^T A_0|^{-1/2} p(\ \hat{s}_N^{(1)}\ ) \tag{16}$$

where $\hat{s}_N^{(1)} = \hat{s}_N(s_{N-1}, \ldots, s_{N-L})$. Then we can rewrite the likelihood with one integral evaluated, i.e.

$$p(X|\{A_\tau\}) = |A_0^T A_0|^{-1/2} \int_{s_1} \int_{s_2} \cdots \int_{s_{N-1}} p(\ \hat{s}_N^{(1)}\ ) \prod_{t=1}^{N-1} p(s_t)\delta(f_t) ds_1 \ldots ds_{N-1}. \tag{17}$$

Following the same idea to marginalize out s_{N-1} now using

$$\int_{s_{N-1}} p(\ \hat{s}_N^{(1)}\ ) p(s_{N-1})\delta(f_{N-1}) ds_{N-1} = |A_0^T A_0|^{-1/2} p(\ \hat{s}_N^{(2)}\ ) p(\ \hat{s}_{N-1}^{(1)}\ ) \tag{18}$$

where $\begin{cases} \hat{s}_{N-1}^{(1)} & = \hat{s}_{N-1}(s_{N-2}, s_{N-3,} \ldots, s_{N-1-L}) \\ \hat{s}_N^{(2)} & = \hat{s}_N(\hat{s}_{N-1}^{(1)}, s_{N-2}, \ldots, s_{N-L}) \end{cases}$. Then we can write the likelihood with two integrals evaluated

$$\begin{aligned} p(X|\{A_\tau\}) = |A_0^T A_0|^{-2/2} \int_{s_1} \int_{s_2} \cdots \\ \int_{s_{N-2}} p(\ \hat{s}_N^{(2)}\ ) p(\ \hat{s}_{N-1}^{(1)}\ ) \prod_{t=1}^{N-2} p(s_t)\delta(f_t) ds_1 \ldots ds_{N-2}. \end{aligned} \tag{19}$$

By repeating this procedure to evaluate all integrals we eventually get

$$p(X|\{A_\tau\}) = |A_0^T A_0|^{-N/2} \prod_{t=1}^{N} p(\ \hat{s}_t^{(t)}\ ) \ , \quad \begin{cases} \hat{s}_1^{(1)} = \hat{s}_1(s_0, s_{-1}, \ldots, s_{1-L}) \\ \hat{s}_2^{(2)} = \hat{s}_2(\hat{s}_1^{(1)}, s_0, \ldots, s_{2-L}) \\ \hat{s}_3^{(3)} = \hat{s}_3(\hat{s}_2^{(2)}, \hat{s}_1^{(1)}, \ldots, s_{3-L}) \\ \vdots \\ \hat{s}_t^{(t)} = \hat{s}_t(\hat{s}_{t-1}^{(t-1)}, \hat{s}_{t-2}^{(t-2)}, \ldots, \hat{s}_{t-L}^{(t-L)}) \end{cases} \tag{20}$$

Assuming s_t zero for $t \leq 0$ we finally get

$$p(X|\{A_\tau\}) = |A_0^T A_0|^{-N/2} \prod_{t=1}^{N} p(\hat{s}_t) \ , \ \hat{s}_t = \hat{s}_t(\hat{s}_{t-1}, \hat{s}_{t-2}, \ldots, \hat{s}_{t-L}). \qquad (21)$$

References

1. Hagai Attias and C. E. Schreiner, "Blind source separation and deconvolution: the dynamic component analysis algorithm," *Neural Computation*, vol. 10, no. 6, pp. 1373–1424, 1998.
2. L. Parra, C. Spence, and B. De Vries, "Convolutive blind source separation based on multiple decorrelation," in *IEEE Workshop on Neural Networks and Signal Processing, Cambridge, UK, September 1998*, 1998, pp. 23–32.
3. Kamran Rahbar, James P. Reilly, and Jonathan H. Manton, "A frequency domain approach to blind identification of mimo fir systems driven by quasi-stationary signals," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 1717–1720.
4. Jörn Anemüller and Birger Kollmeier, "Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach," *IEEE transactions on Speech and Audio processing*, vol. 39, no. 1-2, pp. 79–95, 2003.
5. Mitianoudis N. and Davies M., "Audio source separation of convolutive mixtures," *IEEE transactions on Speech and Audio processing*, vol. 11:5, pp. 489–497, 2003.
6. Eric Moulines, Jean-Francois Cardoso, and Elizabeth Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. ICASSP'97 Munich*, 1997, pp. 3617–3620.
7. Sabine Deligne and Ramesh Gopinath, "An em algorithm for convolutive independent component analysis," *Neurocomputing*, vol. 49, pp. 187–211, 2002.
8. Seungjin Choi, Sun ichi Amari, Andrezej Cichocki, and Ruey wen Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, Aussois, France, January 11–15 1999, pp. 371–376.
9. K. Torkkola, "Blind separation of convolved sources based on information maximization," in *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, September 4-6 1996, pp. 423–432.
10. S. Choi and A. Cichocki, "Blind signal deconvolution by spatio- temporal decorrelation and demixing," in *Neural Networks for Signal Processing, Proc. of the 1997 IEEE Workshop (NNSP-97), IEEE Press, N.Y. 1997*, 1997, pp. 426–435.
11. Per Christian Hansen, "Deconvolution and regularization with toeplitz matrices," *Numerical Algorithms*, vol. 29, pp. 323–378, 2002.
12. Christopher M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Inc., 1995.
13. Te-Won Lee, Anthony J. Bell, and Russell H. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems*, Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, Eds. 1997, vol. 9, p. 758, The MIT Press.
14. Stephen T. Neely and Jont B. Allen, "Invertibility of a room impulse response," *Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, July 1979.

P.4 M. Dyrholm, L. K. Hansen, L. Wang, L. Arendt-Nielsen and A. C. Chen, Convolutive ICA (c-ICA) captures complex spatio-temporal EEG activity, HBM2004

[13] M. Dyrholm and L. K. Hansen and L. Wang and L. Arendt-Nielsen and A. C. Chen, Convolutive ICA (c-ICA) captures complex spatio-temporal EEG activity, 10th annual meeting of the organization for human brain mapping, 2004

Convolutive ICA (c-ICA) captures complex spatio-temporal EEG activity.

Mads Dyrholm, Lars Kai Hansen, Li Wang, Lars Arendt-Nielsen*, Andrew CN Chen**

Informatics and Mathematical Modelling, Technology University of Denmark, Denmark
*Human Brain Mapping and Cortical Imaging Laboratory, Aalborg University, Denmark

[Background]

Independent Component Analysis (ICA) is a useful tool for removing electroencephalographic (EEG) artifacts such as eye-blink or eye-movement. Artifact activity that is spatially-separable and temporally independent from other EEG activity will, in a successful ICA decomposition, appear in a separate component. The ICA method is advocated because the obtained artifact components can be excluded from the EEG by a linear projection. Hence it is possible to clean EEG in its full length without loosing contaminated data segments. However, this approach still requires an expert judgment to determine which of the obtained ICA components are wanted or unwanted. In this work we show how Convolutive ICA (c-ICA) can capture more complex spatio-temporal behavior in a single component than is possible with conventional ICA. This creates components with more realistic temporal structure and furthermore assists the component inspection procedure by reducing the number of components to inspect. Convolutive ICA of EEG data has been studied by Makeig et al (2002,2003) in the complex frequency domain, here we apply a temporal un-mixing c-ICA approach which does not require windowing or frequency based representation of data.

[Methods]

The data used for the analysis was a 124 channel EEG recorded at 204.8Hz sampling rate. Electric pulses were generated at approximately 2Hz and applied to the subjects little-finger as stimulus. An eighty seconds long recording was obtained with approximately 150 stimulation epochs. DC components and slow drift were eliminated from each channel separately by high-pass filtering with a 0.2Hz transition-band around 1Hz cutting frequency. Five principal component features were extracted from the resulting data matrix for convolutive independent analysis (fig. 1).

ICA algorithm: Maximum-Likelihood instantaneous ICA (Bell & Sejnowski, 1995). **Convolutive ICA algorithm:** Maximum-Likelihood (Dyrholm & Hansen, 2003). The number of convolutive lags was set to fifty samples (0.25 sec).

[Results]

The ICA and c-ICA algorithms each resulted in five components. We illustrate the difference between the two ICA approaches by analysis of the components with the maximum correlation with the stimulus delivery. In Fig. 2 and 3 we show time series for the conventional and c-ICA for the five spatial variance components. The conventional ICA time series all follow a stereotypical time-course, hence appear as being completely time synchronized. While the c-ICA time series show non-trivial delay structure between the five spatial patterns, hence, can give rise to time variant scalp contours of activity. This is an important advantage for c-ICA because it directly, within a single component can capture delayed correlations across the features and locations. In Fig. 4 we show the cross-correlation between time series associated with two of the spatial variance features. The cross-correlation function shows two off-center peaks characteristic of two symmetrically delayed signal components. The conventional ICA algorithm captures only the "average" behavior, while the c-ICA component captures the delayed presence of one of these components.

[Conclusion]

Convolutive ICA (c-ICA) offers a more flexible representation with non-trivial temporal structure of the component time series, highly relevant for EEG analysis.

*<**Acknowledgement:** supported by the Danish Technological Council >*

fig. 1:

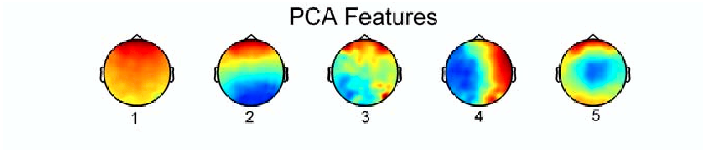

PCA Features
1
2
3
4
5

fig. 2:

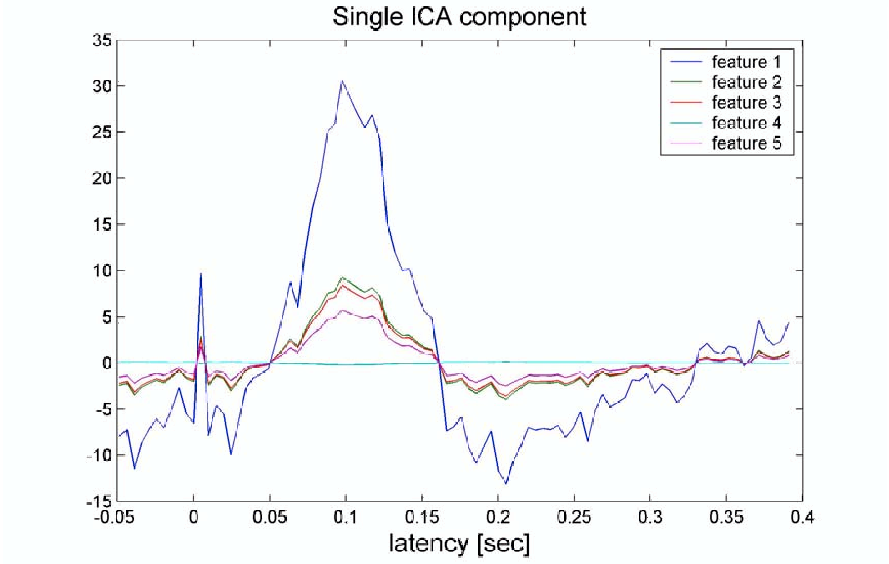

Single ICA component
feature 1
feature 2
feature 3
feature 4
feature 5
35
30
25
20
15
10
5
0
-5
-10
-15
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
latency [sec]

fig. 3:

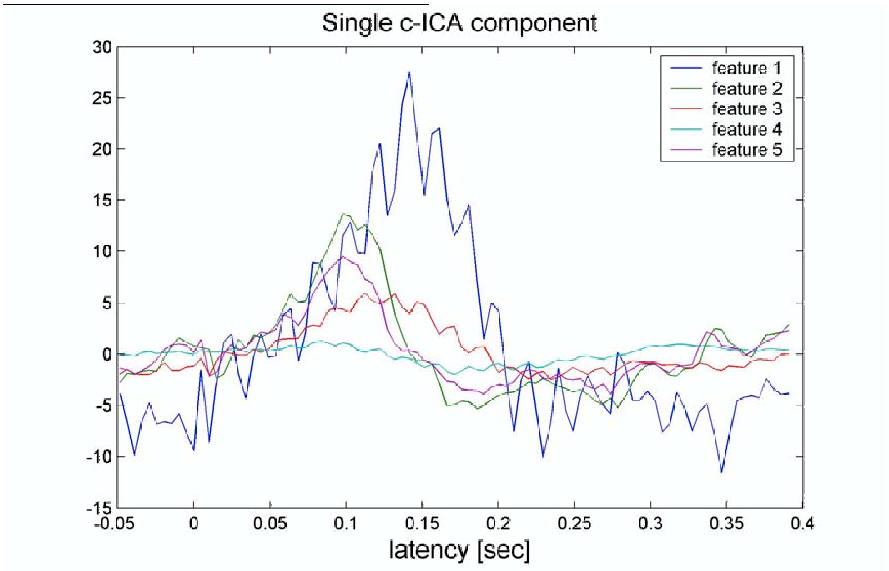
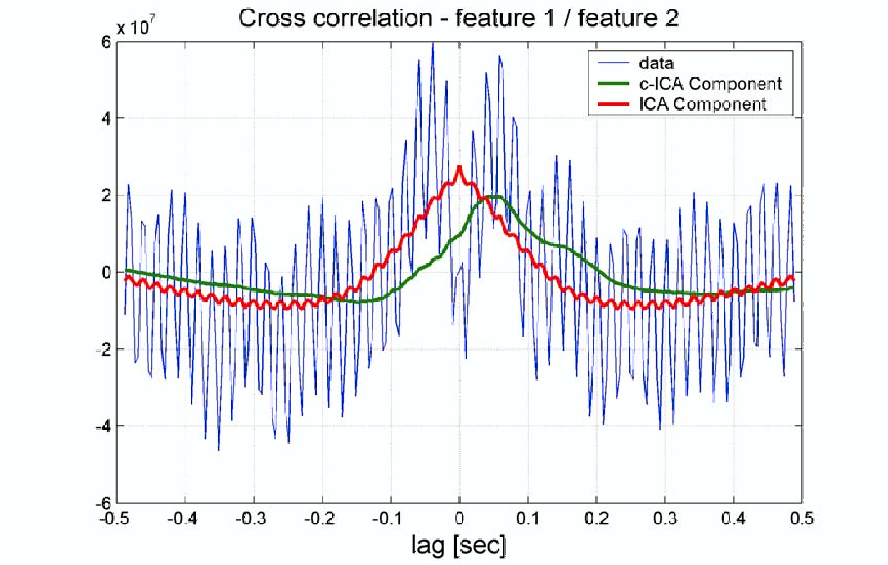

Single c-ICA component
feature 1
feature 2
feature 3
feature 4
feature 5
latency [sec]

fig. 4:


Cross correlation - feature 1 / feature 2
x 10^7
data
c-ICA Component
ICA Component
lag [sec]

P.5 L. K. Hansen and M. Dyrholm, A prediction matrix approach to convolutive ICA, NNSP2003

[21] L. K. Hansen and M. Dyrholm, A prediction matrix approach to convolutive ICA, Proceedings of IEEE Workshop on Neural Networks for Signal Processing XIII, pp. 249-258, 2003

A PREDICTION MATRIX APPROACH TO CONVOLUTIVE ICA

Lars Kai Hansen and Mads Dyrholm
Informatics and Mathematical Modelling
Technical University of Denmark B321
DK-2800 Lyngby, DENMARK
lkh@imm.dtu.dk mad@imm.dtu.dk

Abstract. ***A linear prediction approach reduces convolutive independent component analysis (ICA) to the following three steps: Solution of a set of multivariate linear prediction problems, a linear multivariate deconvolution problem with known matrix coefficients, and finally solution of a conventional instantaneous mixing ICA problem.***

CONVOLUTIVE MIXING

Independent component analysis (ICA) of convolutive mixtures is a problem of considerable current interest in neural computation, say for modeling speech processing and furthermore has numerous applications in signal processing, see, e.g., [1, 2, 3, 4, 5].

Convolutive ICA in its simplest form concerns reconstruction of the $L+1$ mixing matrices and the T source signal vectors from a D-dimensional convolutive mixture,

$$\mathbf{x}_t = \sum_{\tau=0}^{L} \mathbf{A}_\tau \mathbf{s}_{t-\tau}, \quad t = 1, ..., T. \tag{1}$$

The K-dimensional source signal vectors are assumed temporally independent: $p(\{\mathbf{s}_t\}) = \prod_{t=1}^{T} p(\mathbf{s}_t)$. We will assume L is so large that the correlations in the process $\mathbf{x}$ can be explained by the mixing matrices. In fact, as noted by [1], possible auto-correlations of the source signals can not be identified without additional a priori information. In order to see this, note that in the frequency domain the convolution becomes a product of Fourier transforms

$$\mathbf{x}_\omega = \mathbf{A}_\omega \mathbf{s}_\omega, \tag{2}$$

hence, any non-zero 'filter' $h(\omega)$ can be multiplied on a given source if $1/h(\omega)$ is applied to the corresponding column of $\mathbf{A}_\omega$. Another observation is that

for stationary Gaussian white noise sources, the sufficient statistic $\langle \mathbf{x}_t \mathbf{x}_{t+\delta}^\top \rangle$ does not allow full recovery of the mixing matrices since

$$\langle \mathbf{x}_t \mathbf{x}_{t+\delta}^\top \rangle = \sum_\tau \mathbf{A}_\tau \mathbf{A}_{\tau+\delta}^\top, \tag{3}$$

which is invariant to common rotation $\mathbf{A}_\tau \to \mathbf{A}_\tau \mathbf{U}$ of all mixing matrices.

Most earlier approaches to convolutive ICA are based on frequency domain estimation using (2). This leads to a set of conventional 'instantaneous' ICA problems, one for each frequency, and is hampered by a massive permutation problem which can be tamed by adding a prior source 'smoothness' information or other more elaborate schemes [2, 3, 4, 5, 6]. Another line of work is based on optimization of certain 'independency measures', information maximization or other heuristics, see e.g., [7, 8, 9, 10, 11, 12, 13].

Statistically motivated maximum likelihood schemes have been proposed, typically leading to high-dimensional optimizations w.r.t. to all elements of all mixing matrices, see e.g. [14, 1, 15]. The aim in this paper is to invoke a few simple approximations and use these and straightforward linear algebra to reduce the problem to a conventional ICA problem. We will avoid the frequency domain representation all together, hence, we will not further address the frequency component permutation problem.

TEMPORAL UN-MIXING

We will present a temporal un-mixing procedure in which the key new ingredient is the use of prediction matrices, hence, this step is first illustrated on the well-understood problem of *instantaneous* ICA of temporally correlated sources.

Consider the instantaneous mixing system

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t. \tag{4}$$

For simplicity we will consider square mixing so that $D = K$, i.e., $\mathbf{A}$ is a $K \times K$ matrix with real elements. Let the prediction matrix $\mathbf{W}_\tau$ be the best linear predictor of the series $\mathbf{x}$

$$\mathbf{x}_{t+\tau} = \mathbf{W}_\tau \mathbf{x}_t + \boldsymbol{\epsilon}_{t+\tau}. \tag{5}$$

Now right multiply (5) by the transposed source vector $\mathbf{s}_t^\top$ and average w.r.t. the source distribution. If we assume $\langle \boldsymbol{\epsilon}_{t+\tau} \mathbf{s}_t^\top \rangle \approx \mathbf{0}$, we obtain

$$\begin{aligned} \mathbf{A}\langle \mathbf{s}_{t+\tau} \mathbf{s}_t^\top \rangle &= \mathbf{W}_\tau \mathbf{A} \langle \mathbf{s}_t \mathbf{s}_t^\top \rangle, \\ \mathbf{A}\mathbf{C}_\tau &= \mathbf{W}_\tau \mathbf{A}\mathbf{C}_0, \end{aligned} \tag{6}$$

where the matrices $\mathbf{C}_0$ and $\mathbf{C}_\tau$ are diagonal because the sources are independent and constant in time by stationarity. From (6) we learn that the mixing matrix $\mathbf{A}$ is the matrix formed by the eigenvectors of the prediction matrix.

The eigenvalues in the diagonal of the matrix $\mathbf{C}_\tau \mathbf{C}_0^{-1}$ are normalized auto-correlation values of the given source at the lag τ. Equations (5-6) form an alternative route to the so-called Molgedey-Schuster algorithm, see [16, 17]. This algorithm is a quick (closed form) ICA approach, for mixing problems with time-correlated sources and where there are values of τ for which the sources have different normalized auto-correlations, see e.g., [18] for a more detailed discussion and multi-media applications.

Next we will show how the prediction matrix method can be use to simplify the convolutive mixing problem. The linear prediction approach is first generalized to a multi-lag linear predictor of the form,

$$\mathbf{x}_{t+\tau} = \sum_{\lambda=0}^{M} \mathbf{W}_{\tau,\lambda} \mathbf{x}_{t-\lambda} + \epsilon_{t-\tau}. \tag{7}$$

Substituting the convolutive process (1) we find

$$\sum_{\tau'=0}^{L} \mathbf{A}_{\tau'} \mathbf{s}_{t+\tau-\tau'} = \sum_{\lambda=0}^{M} \mathbf{W}_{\tau,\lambda} \sum_{\tau'=0}^{L} \mathbf{A}_{\tau'} \mathbf{s}_{t-\tau'-\lambda}. \tag{8}$$

As above we multiply (8) by $\mathbf{s}_t^\top$ and average w.r.t. the source distribution now assuming, as discussed above, that the sources are *temporally uncorrelated*: $\langle \mathbf{s}_{t+\tau} \mathbf{s}_t^\top \rangle = \mathbf{C}_0 \delta_{\tau,0}$, to get

$$\mathbf{A}_\tau \mathbf{C}_0 = \mathbf{W}_{\tau,0} \mathbf{A}_0 \mathbf{C}_0, \tag{9}$$

Furthermore, assuming that all sources have non-vanishing variance we can divide by the diagonal source covariance matrix $\mathbf{C}_0$ to arrive at the result,

$$\mathbf{A}_\tau = \mathbf{W}_{\tau,0} \mathbf{A}_0. \tag{10}$$

Hence, the existence of the linear predictor (7) implies that the delayed mixing matrices are generated from the 'zero lag' mixing matrix by the prediction matrices.

We estimate the prediction matrices $\mathbf{W}_{\tau,\lambda}$ by least squares. For each value of τ separately we obtain a coupled set of equations,

$$\langle \mathbf{x}_{t+\tau} \mathbf{x}_{t-\delta}^\top \rangle = \sum_{\lambda=0}^{M} \widehat{\mathbf{W}}_{\tau,\lambda} \langle \mathbf{x}_{t-\lambda} \mathbf{x}_{t-\delta}^\top \rangle, \tag{11}$$

with the expectations estimated from the measured time series $\mathbf{x}_t$ by $\langle \ldots \rangle \approx \frac{1}{T} \sum_t (\ldots)$. The linear equations in (11) are easily solved for $\widehat{\mathbf{W}}_{\tau,\lambda}$ by matrix inversion. For each value of τ we will eventually need the set of $L+1$ matrices $\mathbf{W}_{\tau,0}$, c.f., (10). Note that the coupling to the other prediction matrices (for a given τ) in (11) makes $\mathbf{W}_{\tau,0}$ different from the matrix obtained by making a linear prediction in (7) with $M = 0$.

The generator property (10) is next used to simplify the convolutive mixing problem. First rewrite (1)

$$\mathbf{x}_t = \sum_{\tau=0}^{L} \mathbf{W}_{\tau,0}\mathbf{A}_0\mathbf{s}_{t-\tau} \equiv \sum_{\tau=0}^{L} \mathbf{W}_{\tau,0}\mathbf{u}_{t-\tau}. \tag{12}$$

The signals $\mathbf{u}_t = \mathbf{A}_0\mathbf{s}_t$ form an uncorrelated series as they are proportional to the source series $\mathbf{s}_t$.

We have already estimated the prediction matrices from measured data, hence, (12) is a standard linear MIMO system with *known* matrix coefficients $\mathbf{W}_{\tau,0}$, and can be solved by a variety of methods producing an estimate of the time series $\widehat{\mathbf{u}}_t$, $t = 1, ..., T$. In this work we use the simple recursive filter

$$\widehat{\mathbf{u}}_t = \mathbf{W}_{0,0}^{-1}\mathbf{x}_t - \sum_{\tau=1}^{L} \mathbf{W}_{0,0}^{-1}\mathbf{W}_{\tau,0}\widehat{\mathbf{u}}_{t-\tau}. \tag{13}$$

This filter may become unstable, in such case a more robust *regularized* estimator can be invoked, e.g., substituting

$$\mathbf{W}_{0,0}^{-1} \;\rightarrow\; \left(\kappa\mathbf{1} + \mathbf{W}_{0,0}^{\top}\mathbf{W}_{0,0}\right)^{-1}\mathbf{W}_{0,0}^{\top}, \tag{14}$$

in (13). The remaining problem is to estimate $\mathbf{A}_0$ and the source signals $\mathbf{s}_t$ from the series

$$\widehat{\mathbf{u}}_t = \mathbf{A}_0\mathbf{s}_t. \tag{15}$$

This is a conventional ICA problem with temporally independent source signals and can be solved by any of the standard approaches. If the distribution of the source signals have positive kurtosis, as appropriate for, e.g., speech signals we can use the Infomax approach of Bell and Sejnowski [19].

Solving the problem (15) we obtain $\widehat{\mathbf{A}}_0$ and $\widehat{\mathbf{s}}_t$, using (10) we can then generate the matrices $\widehat{\mathbf{A}}_\tau$ using the $\widehat{\mathbf{W}}_{\tau,0}$'s, hence concluding our recipe for solving the convolutive mixing problem (1).

SIMULATION EXAMPLE

We illustrate the viability of the prediction based approach by a small simulation example.

A $D = K = 2$ convolutive mixture was created by first designing a set of 2×2 mixing matrices ($L = 30$). These were next applied as in (1) to a i.i.d. random source signal ($T = 30000$). The distribution of the source signals was made non-gaussian, with positive kurtosis, by the transformation $s = \mathrm{sign}(u) * |u|^2$ where $u \sim \mathcal{N}(0, 1)$. The source and the mixed signals can be seen in Figure 1, while the mixing matrices are shown for reference in Figure 7.

In Figures 2-3 we first illustrate the excellent quality of the linear model in (7).

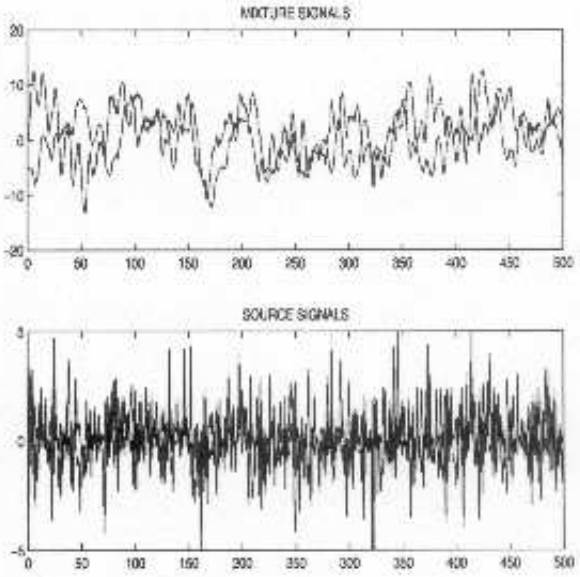


Figure 1: Simulation experiment involving $L = 30$ square mixing matrices ($D = 2$) convolved with i.d.d. long-tailed source signals. In the upper panel we show a short segment of the two convolved signals ($\mathbf{x}_t$), in the lower panel we show the corresponding segments of the two sources ($\mathbf{s}_t$). The mixing matrices are shown in Figure 7.

Figure 2 shows scatter plots of the prediction error ($\boldsymbol{\epsilon}_{t+\tau}$, $\tau = 3$) vs. the source signal ($\mathbf{s}_t$). It is important for the generator relation (10) that these time series are roughly uncorrelated. In Figure 3 we have further quantified this relation as function of the prediction horizon (τ). As expected, the predictions become more and more noisy as we increase τ, i.e., the relative power in $\boldsymbol{\epsilon}_t$ increases, however, more important is it that the correlation between the source signal and the error remains limited, supporting relation (10).

Next we investigate the quality of the prediction matrix estimates. The ratios $\mathbf{A}_\tau \mathbf{A}_0^{-1}$ were computed with the 'true' matrices used in the simulation. In Figure 4 we compare these matrices with the matrices estimated from data, the match is good and the other four channels are of similar quality (data not shown).

The MIMO problem is solved using (13). The relative reconstruction error was small ($\langle\langle \mathbf{x}_t - \widehat{\mathbf{x}}_t)^2\rangle / \langle(\mathbf{x}_t)^2\rangle < 10^{-6}$). Using our in-house implementation of the Bell and Sejnowski algorithm[1], instantaneous ICA was applied to the resulting time series $\widehat{\mathbf{u}}_t$. The estimated sources are compared with the 'true' sources in Figure 5 and the consistency is remarkable.

Using the reconstructed $\mathbf{A}_0$ we estimated the remaining matrices $\mathbf{A}_\tau$ us-

[1]MatLab toolbox available from `www.imm.dtu.dk/cisp`

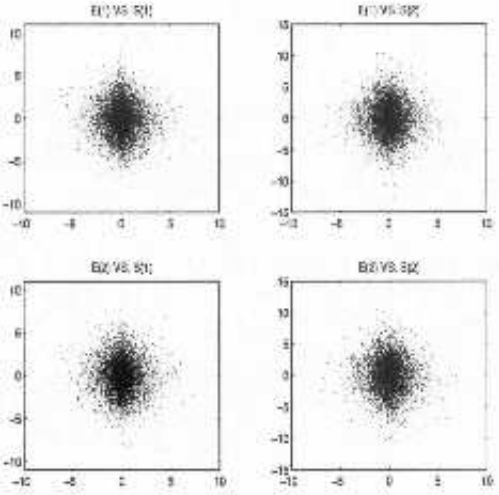


Figure 2: Simulation as in Figure 1. The scatter plots illustrate the dependency between $\epsilon_{t-\tau}$ ($\tau = 3$) and s_t, c.f. (7).

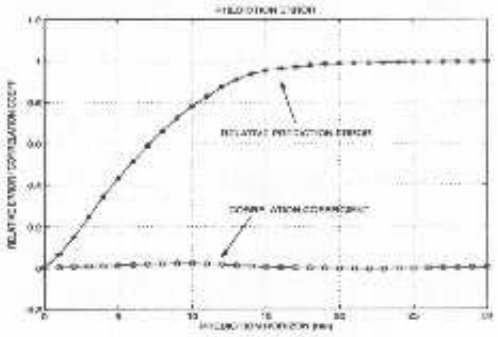


Figure 3: Simulation experiment as in Figure 1. We show the relative prediction error (the mean square error normalized by the signal variance) as function of the prediction horizon (τ), and the correlation coefficients between $\epsilon_{1,1+\tau}$ and $s_{2,t}$. While the predictions become increasingly random, the correlation coefficients stay in the range $-0.05 - 0.05$, ensuring that the error in (10) is bounded.

ing (10). The matrix elements $\mathbf{A}_{1,2,\tau}$ are compared in Figure 6 with the corresponding element of matrices found by generation using the true $\mathbf{A}_0$. Apart from the absolute amplitude, these elements are in good agreement, indicating that the approach has quite successfully solved the convolutive mixing problem.

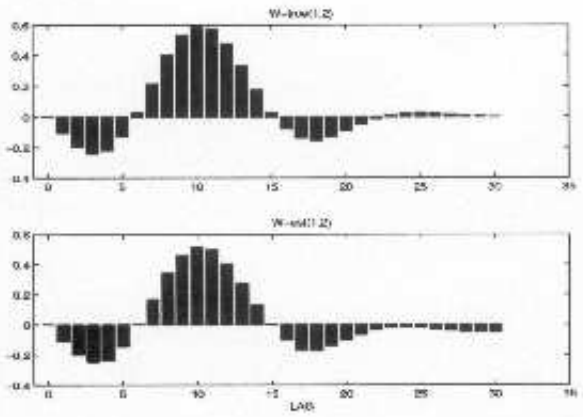


Figure 4: Simulation experiment as in Figure 1. The estimated W-matrices (using (13)) compared favorably with the 'true' matrices $\mathbf{W}_{\tau,0} \equiv \mathbf{A}_\tau \mathbf{A}_0^{-1}$.

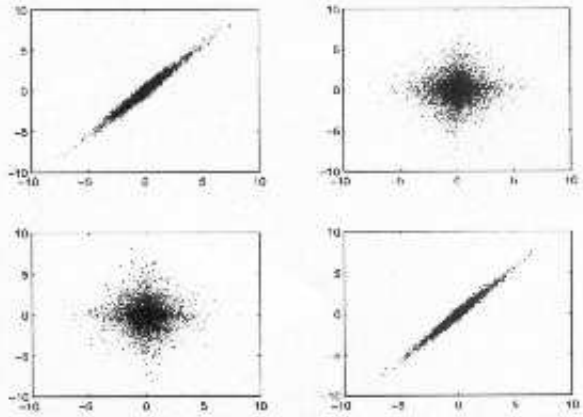


Figure 5: Simulation experiment as in Figure 1. We plot the true sources of the simulation experiment vs. the reconstructed sources. The sign and the ordering of the reconstructed sources have been modified for clarity. The reconstructed sources are well aligned with the true sources, this is highly non-trivial for convolutive mixtures.

CONCLUSION

We have proposed a linear prediction approach to the convolutive ICA problem. Within a linear prediction assumption and linear algebra, the problem is reduced to the following three steps: Solving a set of multivariate linear prediction problems, solving a linear multivariate deconvolution problem with known matrix coefficients, and finally solving a conventional instantaneous mixing ICA problem. A small simulation example showed that the

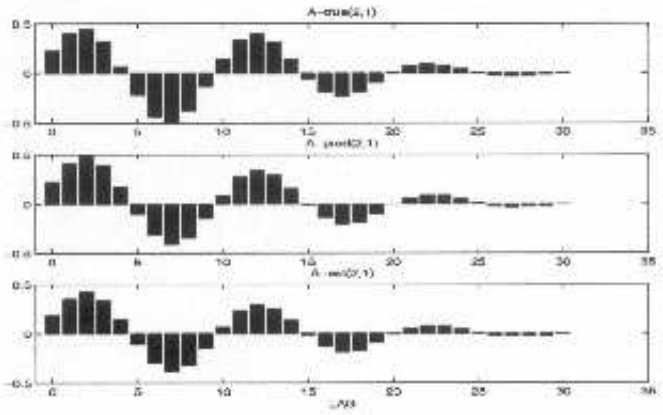


Figure 6: Simulation experiment as in Figure 1. The recovered matrix elements $\mathbf{A}_{2,1,\tau}$ (bottom row) are compared with true matrices (upper row) and the matrices obtained by multiplying the prediction matrices $\widehat{\mathbf{W}}_{\tau,0}$ on the true $\mathbf{A}_0$-matrix (middle row).

approach is able to accurately estimate the mixing matrices and the source signals. We are currently trying to identify proper conditions for the linear prediction assumption and also to invoke more robust schemes for solving the MIMO problem.

Acknowledgments

LKH thanks Scott Makeig of the Swartz Center for Neuroimaging for hosting a visit summer 2002, where this work was initiated. We thank Jan Larsen, Ole Winther and Scott Makeig for stimulating ICA discussions. This work is supported by the Danish Technical Research Council (STVF) through the International Center for Biomedical Research.

References

[1] H. Attias and C. E. Schreiner, "Blind source separation and deconvolution: the dynamic component analysis algorithm," *Neural Computation*, vol. 10, no. 6, pp. 1373–1424, 1998.

[2] L. Parra, C. Spence, and B. D. Vries, "Convolutive blind source separation based on multiple decorrelation," in *IEEE Workshop on Neural Networks and Signal Processing, Cambridge, UK, September 1998*, pp. 23–32, 1998.

[3] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources," *IEEE Trans. on Speech and Audio Processing*, vol. 8,

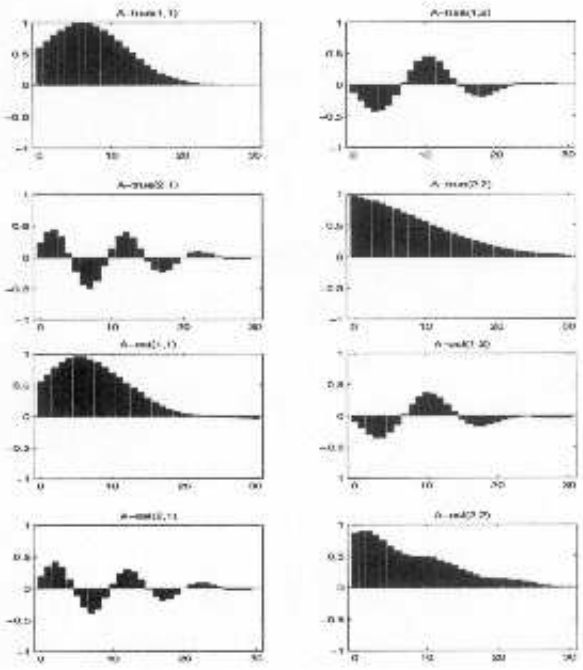


Figure 7: Simulation experiment as in Figure 1. The complete set of recovered matrix elements $\widehat{\mathbf{A}}_\tau$ (bottom four panels) are compared with the true matrix elements $\mathbf{A}_\tau$ (upper four panels).

pp. 320–327, 2000.

[4] K. Rahbar and J. Reilly, "Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2745–2748, 2001.

[5] K. Rahbar, J. P. Reilly, and J. H. Manton, "A frequency domain approach to blind identification of mimo fir systems driven by quasi-stationary signals," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1717–1720, 2002.

[6] W. Baumann, B.-U. Kohler, D. Kolossa, and R. Orglmeister, "Real time separation of convolutive mixtures," in *3'rd International Conference on ICA. Eds. T-W. Lee et al.*, pp. 65–69, San Diego December 2001 2001.

[7] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, pp. 423–432, September 4-6 1996.

[8] A. Belouchrani, K. A. Meraim, J.-F. Cardoso, and Éric Moulines, "A blind source separation technique based on second order statistics," *IEEE Trans. on Signal Processing*, vol. 42, pp. 434–444, 1997.

[9] S. Choi and A. Cichocki, "Blind signal deconvolution by spatio- temporal decorrelation and demixing," in *Neural Networks for Signal Processing. Proc. of the 1997 IEEE Workshop (NNSP-97), IEEE Press, N.Y. 1997*, pp. 426–435, 1997.

[10] S. Douglas, A. Cichocki, and S. Amari, "Self-whitening algorithms for adaptive equalization and deconvolution," *IEEE Trans. on Signal Processing*, vol. 47, pp. 1161–1165, 1999.

[11] R. Cristescu, T. Ristaniemi, J. Joutsensalo, and J. Karhunen, "Blind separation of convolved mixtures for cdma systems," in *Proc. of the X European Signal Processing Conference (EUSIPCO 2000), Tampere, Finland*, pp. 619–622, 2000.

[12] P. Comon, E. Moreau, and L. Rota, "Blind separation of convolutive mixtures: A constrast based joint diagonalization appraoch," in *3'rd International Conference on ICA. Eds. T-W. Lee et al.*, pp. 686–691, San Diego December 2001 2001.

[13] X. Sun and S. Douglas, "A natural gradient convolutive blind source separation algorithms for speech mixtures," in *3'rd International Conference on ICA. Eds. T-W. Lee et al.*, pp. 59–64, San Diego December 2001 2001.

[14] E. Moulines, J.-F. Cardoso, and E. Gassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *Proc. ICASSP'97 Munich*, pp. 3617–3620, 1997.

[15] S. Deligne and R. Gopinath, "An em algorithm for convolutive independent component analysis," p. To appear, 2002.

[16] L. Molgedey and H. Schuster, "Separation of a mixture of independent signals using time delayed correlation," *Physical Review Letters*, vol. 72, pp. 3634–3637, 1994.

[17] A. S. Lukic, M. N. Wernick, L. K. Hansen, and S. C. Strother, "An ica algorithm for analyzing multiple data sets," in *IEEE 2002 Int. Conf. on Image Processing (ICIP-2002)* (M. T. et al., ed.), IEEE, 2002.

[18] L. K. Hansen, J. Larsen, and T. Kolenda, "On independent component analysis for multimedia signals," in *Multimedia Image and Video Processing, CRC Press* (S. K. L. Guan and J. Larsen, eds.), pp. 175–199, 2000.

[19] T. Bell and T. Sejnowski, "An information maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

Appendix T

Toolbox implementation notes

T.1 Functions in the toolbox

Function name	Description	Matlab	Fortran90
cicaar	CICAAR algorithm.	x	x
cicaarmpi	CICAAR algorithm for multiple CPUs running in parallel.		x
cicaarwrite	Export matlab array for use with the binary CICAAR implementation.	x	
cicaarread	Import result from the binary CICAAR implementation.	x	
cicap	CICAP algorithm.	x	
convestmix	Estimate the multivariate Wiener filter.	x	
convmix	Produce a convolutive mixture.	x	
convis	Calculate the generalized poles of a convolutive mixing system	x	
crosstalk	Measure the crosstalk matrix.	x	

T.2 Pointers towards a CICAAR computer implementation

The implementation of the CICAAR algorithm used in this thesis uses the BFGS optimization routine of [41] for gradient based optimization which was also used in the ICA:DTU implementation of Infomax ICA available at `http://mole.imm.dtu.dk/toolbox/ica/`.

A stable starting point is $\mathbf{A}_\tau = \mathbf{0}$ (for $\tau \neq 0$) with arbitrary $\mathbf{A}_0$.

In each iteration, the parameter refinements might result in an unstable inverse used for estimation of the likelihood. In such cases, the negative log likelihood is likely to be represented as IEEE `inf` in the computer. When that happens, IEEE `inf` is replaced with an arbitrary (very) high number due to the line-search procedure of the BFGS optimizer. Also, when the auto-regressive inverse is detected to diverge, the gradient is unnecessary and is not calculated. Thus, the infeasible iterations consume considerably less computational power than the feasible iterations.

Bibliography

[1] J. Anemüller. Across-frequency processing in convolutive blind source separation. Phd dissertation, Universität Oldenburg, 2001.

[2] J. Anemüller and B. Kollmeier. Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach. *IEEE transactions on Speech and Audio processing*, 39(1-2):79–95, 2003.

[3] H. Attias and C. E. Schreiner. Blind source separation and deconvolution: the dynamic component analysis algorithm. *Neural Computation*, 10(6):1373–1424, 1998.

[4] W. Baumann, B.-U. Kohler, D. Kolossa, and R. Orglmeister. Real time separation of convolutive mixtures. In *3'rd International Conference on ICA. Eds. T-W. Lee et al.*, pages 65–69, San Diego December 2001 2001.

[5] A. Bell and T. J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[6] S. Choi, S.-I. Amari, A. Cichocki, and R.-W. Liu. Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels. In *International Workshop on Independent Component Analysis and Blind Signal Separation (ICA'99)*, pages 371–376, Aussois, France, January 11–15 1999.

[7] S. Choi and A. Cichocki. Blind signal deconvolution by spatio- temporal decorrelation and demixing. In *Neural Networks for Signal Processing, Proc. of the 1997 IEEE Workshop (NNSP-97), IEEE Press, N.Y. 1997*, pages 426–435, 1997.

[8] P. Comon, E. Moreau, and L. Rota. Blind separation of convolutive mixtures: A constrast based joint diagonalization appraoch. In *3'rd International Conference on ICA. Eds. T-W. Lee et al.*, pages 686–691, San Diego December 2001 2001.

[9] S. Deligne and R. Gopinath. An em algorithm for convolutive independent component analysis. *Neurocomputing*, 49:187–211, 2002.

[10] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *J Neurosci Methods*, 134(1):9–21, Mar 2004.

[11] A. Delorme, S. Makeig, and T. J. Sejnowski. From single-trial eeg to brain area dynamics. *Neurocomputing*, 44-46:1057–1064, 2002.

[12] S. C. Douglas, A. Cichocki, and S. Amari. Self-whitening algorithms for adaptive equalization and deconvolution. *IEEE Trans. on Signal Processing*, 47:1161–1165, 1999.

[13] M. Dyrholm and L. K. Hansen. CICAAR: Convolutive ICA with an autoregressive inverse model. In Carlos G. Puntonet and Alberto Prieto, editors, *Independent Component Analysis and Blind Signal Separation*, pages 594–601, sep 2004.

[14] M. Dyrholm, L. K. Hansen, L. Wang, L. Arendt-Nielsen, and A. C. Chen. Convolutive ICA (c-ICA) captures complex spatio-temporal EEG activity. In *10th annual meeting of the organization for human brain mapping*, 2004.

[15] M. Dyrholm, S. Makeig, and L. K. Hansen. Model structure selection in convolutive mixtures. In *(submitted) Independent Component Analysis and Blind Signal Separation*, 2006.

[16] M. Dyrholm, S. Makeig, and L. K. Hansen. Model selection for convolutive ica with an application to spatio-temporal analysis of eeg. *Neural Computation*, (submitted) 2005.

[17] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

[18] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Baltimore, MD, USA, third edition, 1996.

[19] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, Aug. 1969.

[20] L. K. Hansen and J. Larsen. Unsupervised learning and generalization. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, pages 25–30, Washington DC, june 1996.

[21] L.K. Hansen and M. Dyrholm. A prediction matrix approach to convolutive ica. In C. Molina, T. Adali, J. Larsen, M. Van Hulle, S. Douglas, and J. Rouat, editors, *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XIII*, pages 249–258, 2003.

[22] P. C. Hansen. Deconvolution and regularization with toeplitz matrices. *Numerical Algorithms*, 29:323–378, 2002.

[23] Esben Høgh-Rasmussen. *BBTools – a Matlab Toolbox for Black-Box Computations.* Neurobiology Research Unit, Copenhagen University Hospital, 2005. Available from `http://nru.dk/software/bbtools/`.

[24] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley, 2001.

[25] T.-P. Jung, C. Humphries, T.-W. Lee, S. Makeig, M. J. McKeown, V. Iragui, and T. J. Sejnowski. Extended ICA removes artifacts from electroencephalographic recordings. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, 1998.

[26] T.-P. Jung, S. Makeig, M. J. McKeown, A. Bell, T.-W. Lee, and T. J. Sejnowski. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–22, 2001.

[27] E. R. Kandel, J.H. Schwartz, and T. M. Jessell. *Principles of Neural Science.* Mc Graw Hill, 2000.

[28] B. Kolb and I. Q. Whishaw. *Fundamentals of Human Neuropsychology.* Series of Books in PsychologySeries of Books in Psychology. Worth Publishers, Incorporated, 5 edition, march 2003.

[29] T.-W. Lee, A. J. Bell, and R. H. Lambert. Blind separation of delayed and convolved sources. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 758. The MIT Press, 1997.

[30] T.-W. Lee, M. Girolami, and T. J. Sejnowski. Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, 11(2):417–441, 1999.

[31] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003. Available from `http://www.inference.phy.cam.ac.uk/mackay/itila/`.

[32] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 145–151, 1996.

[33] S. Makeig, S. Debener, J. Onton, and A. Delorme. Mining event-related brain dynamics. *Trends in Cognitive Science*, 8(5):204–210, 2004.

[34] S. Makeig, A. Delorme, M. Westerfield, J. Townsend, E. Courchense, and T. Sejnowski. Electroencephalographic brain dynamics following visual targets requiring manual responses. *PLoS Biology*, 2004.

[35] S. Makeig, S. Enghoff, T.-P. Jung, and T. J. Sejnowski. A natural basis for efficient brain-actuated control. *IEEE Trans. Rehab. Eng.*, 8:208–211, 2000.

[36] S. Makeig, T.-P. Jung, A. J. Bell, D. Ghahremani, and T. J. Sejnowski. Blind separation of auditory event-related brain responses into independent components. *Proc. Natl. Acad. Sci.*, 94:10979–10984, 1997.

[37] S. Makeig, M. Westerfield, T.-P. Jung, S. Enghoff, J. Townsend, E. Courchesne, and T. J. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295(5555):690–694, 2002.

[38] E. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. ICASSP'97 Munich*, pages 3617–3620, 1997.

[39] Arnold Neumaier and Tapio Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):27–57, 2001.

[40] H. B. Nielsen. *Numerisk Lineær Algebra*. IMM - DTU, 2 edition, 1996.

[41] Hans Brun Nielsen. Ucminf - an algorithm for unconstrained, nonlinear optimization. Technical Report IMM-REP-2000-19, Department of Mathematical Modelling, Technical University of Denmark., december 2000.

[42] P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch. Eeg coherency i: Statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalogr Clin Neurophysiol.*, 103(5):499–515, Nov 1997.

[43] J. Onton, A. Delorme, and S. Makeig. Frontal midline eeg dynamics during working memory. *NeuroImage*, 27:342–356, 2005.

[44] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, pages 320–327, May 2000. — an implementation is available at `http://ida.first.gmd.de/ harmeli/download/download_convbss.html`.

[45] L. Parra, C. Spence, and B. Vries. Convolutive source separation and signal modeling with ml. In *International Symposioum on Inteligent Systems*, 1997.

[46] B. A. Pearlmutter and L. C. Parra. Maximum likelihood blind source separation: A context-sensitive generalization of ICA. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 613. The MIT Press, 1997.

[47] M. S. Pedersen, J. Larsen, and U. Kjems. On the difference between updating the mixing matrix and updating the separation matrix. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP'05)*, volume V, pages 297–300, Philadelphia, PA, USA, mar 2005.

[48] John G. Proakis and Dimitris G. Manolakis. *Digital signal processing (3rd ed.): principles, algorithms, and applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[49] K. Rahbar and J. Reilly. Blind source separation of convolved sources by joint approximate diagonalization of cross-spectral density matrices. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2745–2748, 2001.

[50] K. Rahbar, J. P. Reilly, and J. H. Manton. A frequency domain approach to blind identification of mimo fir systems driven by quasi-stationary signals. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1717–1720, 2002.

[51] R. B. Randall. *Frequency Analysis*. 3 edition, 1987.

[52] M. Scherg. Fundamentals of dipole source potential analysis. volume 6 of *Advances in Audiology*, pages 40–69, Karger, Basel, 1990.

[53] D. Schobben, K. Torkkola, and P. Smaragdis. Evaluation of blind signal separation methods. In *Proceedings Int. Workshop Independent Component Analysis and Blind Signal Separation, Aussois, France.*, 1999.

[54] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[55] Jonathan R Shewchuk. An introduction to the conjugate gradient method without the agonizing pain. Technical report, Pittsburgh, PA, USA, 1994.

[56] X. Sun and S. Douglas. A natural gradient convolutive blind source separation algorithms for speech mixtures. In *3'rd International Conference on ICA. Eds. T-W. Lee et al.*, pages 59–64, San Diego December 2001 2001.

[57] K. Torkkola. Blind separation of convolved sources based on information maximization. In *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, pages 423–432, Sep. 1996.

[58] E. Vincent, C. Févotte, and R. Gribonval. Performance measurement in blind audio source separation. to appear in IEEE Trans. Speech and Audio Processing 2005.

[59] A. von Stein, C. Chiang, and P. Konig. Top-down processing mediated by interareal synchronization. *PNAS*, 97(26):14748–14753, 2000.

[60] D. M. Weinstein and C. R. Johnson. Effects of geometric uncertainty on the inverse EEG problem. In R.L. Barbour, M.J. Carvlin, and M.A. Fiddy, editors, *Computational, Experimental, and Numerical Methods for Solving Ill-Posed Inverse Imaging Problems:Medical and Nonmedical Applications*, volume 3171 of *SPIE '97*, pages 138–145. SPIE, 1997.

[61] L. Zhukov, D. M. Weinstein, and C. R. Johnson. Independent component analysis for EEG source localization in realistic head models. In *Third International Conference on Inverse Problems in Engineering*, 1999.