

# AMICA: An Adaptive Mixture of Independent Component Analyzers with Shared Components

Jason A. Palmer, Ken Kreutz-Delgado, and Scott Makeig

## Abstract

We derive an asymptotic Newton algorithm for Quasi Maximum Likelihood estimation of the ICA mixture model, using the ordinary gradient and Hessian. The probabilistic mixture framework can accommodate non-stationary environments and arbitrary source densities. We prove asymptotic stability when the source models match the true sources. An application to EEG segmentation is given.

## Index Terms

Independent Component Analysis, Bayesian linear model, mixture model, Newton method, EEG

## I. INTRODUCTION

### A. Related Work

The Gaussian liner model approach is described [1]–[3]. Non-Gaussian sources in the form of Gaussian scale mixtures, in particular Student’s  $t$  distribution, were developed in [4]–[6]. A mixture of Gaussians source model was employed in [7]–[11]. Similar approaches were proposed in [12], [13]. These models generally include noise and involve computationally intensive optimization algorithms. The focus in these models is generally on “variational” methods of automatically determining the number of mixtures in a mixture model during the optimization procedure. There is also overlap between the variational technique used in these methods, and the Gaussian scale mixture approach to representing non-Gaussian densities.

A model similar to that proposed here was presented in [14]. The main distinguishing features of the proposed model are,

J. A. Palmer and S. Makeig are with the Swartz Center for Computational Neuroscience, La Jolla, CA, {jason, scott}@scn.ucsd.edu. K. Kreutz-Delgado is with the ECE Department, Univ. of California San Diego, La Jolla, CA, kreutz@ece.ucsd.edu.

- 1) Mixtures of Gaussian scale mixture sources provide more flexibility than the Gaussian mixture models of [7], [11], or fixed density models used in [14]. Accurate source density modeling is important to take advantage of Newton convergence for the true source model, as well as to distinguish between partially overlapping ICA models by posterior likelihood.
- 2) Implementation of the Amari Newton method described in [15] greatly improving the convergence, particularly in the multiple model case, in which prewhitening is not possible (in general a different whitening matrix will be required for each unknown model.)
- 3) The second derivative source density quantities are converted to first derivative quantities using integration by parts related properties of the score function and Fisher Information Matrix. Again accurate modeling of the source densities makes this conversion possible, and makes it robust in the presence of other (interfering) models.

The proposed model is readily extendable to MAP estimation or Variational Bayes or Ensemble Learning approaches, which put conjugate hyperpriors on the parameters. We are interested primarily in the large sample case, so we do not pursue these extensions here.

The probabilistic framework can also be extended to incorporate Markov dependence of state parameters in the ICA and source mixtures.

We have also extended the model to include mixtures of linear processes [16], where blind deconvolution is treated in a manner similar to [17]–[20], as well as complex ICA [21] and dependent sources [21]–[23]. In all of these contexts the adaptive source densities, asymptotic Newton method, and mixture model features can all be maintained.

## II. ICA MIXTURE MODEL

In the standard linear model, observations  $\mathbf{x}(t) \in \mathbb{R}^m$ ,  $t = 1, \dots, N$ , are modeled as linear combinations of a set of basis vectors  $\mathbf{A} \triangleq [\mathbf{a}_1 \cdots \mathbf{a}_n]$  with random and independent coefficients  $s_i(t)$ ,  $i = 1, \dots, n$ ,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

We assume for simplicity the noiseless case, or that the data has been pre-processed, e.g. by PCA, filtering, etc., to remove noise. The data is assumed however to be non-stationary, so that different linear models may be in effect at different times. Thus for each observation  $\mathbf{x}(t)$ , there is an index  $h_t \in \{1, \dots, M\}$ , with corresponding complete basis set  $\mathbf{A}_{h_t}$  with “center”  $\mathbf{c}_{h_t}$ , and a random vector of zero mean, independent sources  $\mathbf{s}(t) \sim q_{h_t}(\mathbf{s})$ , where,

$$q_h(\mathbf{s}) = \prod_{i=1}^n q_{hi}(s_i)$$

such that,

$$\mathbf{x}(t) = \mathbf{A}_h \mathbf{s}(t) + \mathbf{c}_h$$

with  $h = h_t$ . We shall assume that only one model is active at each time, and that model  $h$  is active with probability  $\gamma_h$ . For simplicity we assume temporal independence of the model indices  $h_t$ ,  $t = 1, \dots, N$ .

Since the model is conditionally linear, the conditional density of the observations is given by,

$$p(\mathbf{x}(t) | h) = |\det \mathbf{W}_h| q_h(\mathbf{W}_h(\mathbf{x}(t) - \mathbf{c}_h))$$

where  $\mathbf{W}_h \triangleq \mathbf{A}_h^{-1}$ .

The sources are taken to be mixtures of (generally *nongaussian*) Gaussian Scale Mixtures (GSMs), as in [24],

$$q_{hi}(s_i(t)) = \sum_{j=1}^m \alpha_{hij} \sqrt{\beta_{hij}} q_{hij}(\sqrt{\beta_{hij}}(s_i(t) - \mu_{hij}); \rho_{hij})$$

where each  $q_{hij}$  is a GSM parameterized by  $\rho_{hij}$ .

Thus the density of the observations  $\mathbf{X} \triangleq \{\mathbf{x}(t)\}$ ,  $t = 1, \dots, N$ , is given by,

$$p(\mathbf{X}; \Theta) = \prod_{t=1}^N \sum_{h=1}^M \gamma_h p(\mathbf{x}(t) | h),$$

$\gamma_h \geq 0$ ,  $\sum_{h=1}^M \gamma_h = 1$ . The parameters to be estimated are,

$$\Theta = \{\mathbf{W}_h, \mathbf{c}_h, \gamma_h, \alpha_{hij}, \mu_{hij}, \beta_{hij}, \rho_{hij}\},$$

$h = 1, \dots, M$ ,  $i = 1, \dots, n$ , and  $j = 1, \dots, m$ .

#### A. Invariances in the model

Besides the accepted invariance to permutation of the component indices, invariance or redundancy in the model also exists in two other respects. The first concerns the model centers,  $\mathbf{c}_h$ , and the source density location parameters  $\mu_{hij}$ . Specifically, we have  $p(\mathbf{X}; \Theta) = p(\mathbf{X}; \Theta')$ ,  $\Theta = \{\dots, \mathbf{c}_h, \mu_{hij}, \dots\}$ ,  $\Theta' = \{\dots, \mathbf{c}'_h, \mu'_{hij}, \dots\}$ , if

$$\mathbf{c}'_h = \mathbf{c}_h + \Delta \mathbf{c}_h, \quad \mu'_{hij} = \mu_{hij} - [\mathbf{W}_h \Delta \mathbf{c}_h]_i, \quad j = 1, \dots, m$$

for any  $\Delta \mathbf{c}_h$ . Putting  $\mathbf{c}'_h = E\{\mathbf{x}(t) | h\}$ , we make the sources  $\mathbf{s}(t)$  zero mean given the model. The zero mean assumption is used in the calculation of the expected Hessian for the Newton algorithm.