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.

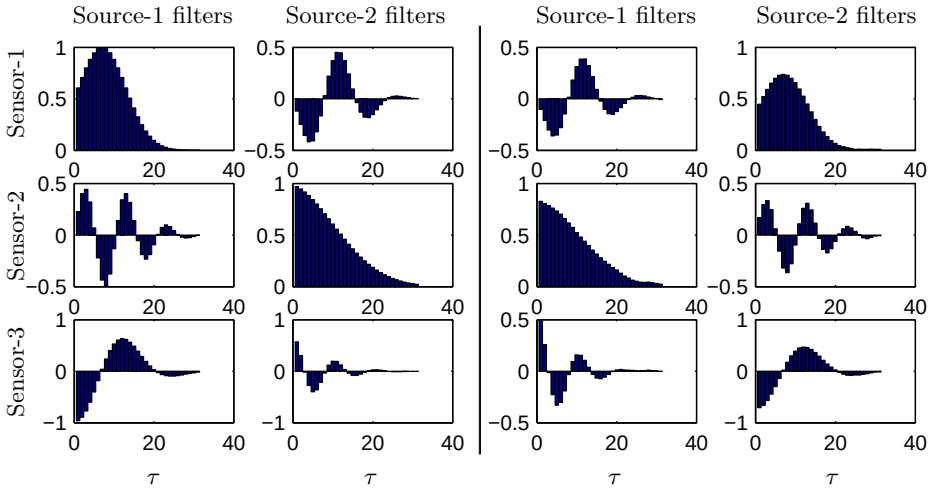

Fig. 1. (left) true mixing filters, (right) estimated mixing filters.

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

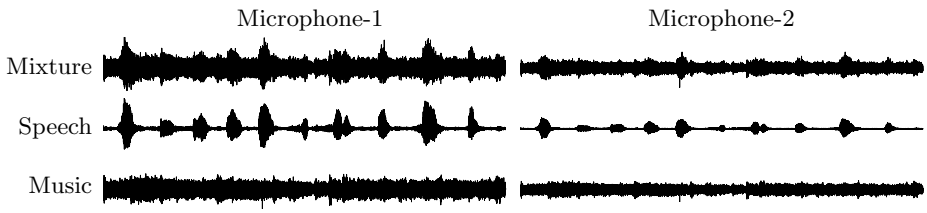

Fig. 2. Separation of real world sound signals. (Top row) The recorded mixture of speech and music. (Middle row) Separated speech reconstructed in the sensor domain. (Bottom row) Separated music reconstructed in the sensor domain.

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

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

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.