PROBABILISTIC BLIND DECONVOLUTION OF NON-STATIONARY SOURCES

Rasmus Kongsgaard Olsson and Lars Kai Hansen

Informatics and Mathematical Modelling, B321 Technical University of Denmark
DK-2800 Lyngby, Denmark
email: rko@isp.imm.dtu.dk,lkh@imm.dtu.dk

ABSTRACT

We solve a class of blind signal separation problems using a constrained linear Gaussian model. The observed signal is modelled by a convolutive mixture of colored noise signals with additive white noise. We derive a time-domain EM algorithm 'KaBSS' which estimates the source signals, the associated second-order statistics, the mixing filters and the observation noise covariance matrix. KaBSS invokes the Kalman smoother in the E-step to infer the posterior probability of the sources, and one-step lower bound optimization of the mixing filters and noise covariance in the M-step. In line with (Parra and Spence, 2000) the source signals are assumed time variant in order to constrain the solution sufficiently. Experimental results are shown for mixtures of speech signals.

1. INTRODUCTION

Reconstruction of temporally correlated source signals observed through noisy, convolutive mixtures is a fundamental theoretical issue in signal processing and is highly relevant for a number of important signal processing applications including hearing aids, speech processing, and medical imaging. A successful current approach is based on simultaneous diagonalization of multiple estimates of the source cross-correlation matrix [5]. A basic assumption in this work is that the source cross-correlation matrix is time variant. The purpose of the present work is to examine this approach within a probabilistic framework, which in addition to estimation of the mixing system and the source signals will allow us to estimate noise levels and model likelihoods.

We consider a noisy convolutive mixing problem where the sensor input $\mathbf{x}_t$ at time t is given by

$$\mathbf{x}_t = \sum_{k=0}^{L-1} \mathbf{A}_k \mathbf{s}_{t-k} + \mathbf{n}_t. \quad (1)$$

The L matrices $\mathbf{A}_k$ define the delayed mixture and $\mathbf{s}_t$ is a vector of possibly temporally correlated source processes. The noise $\mathbf{n}_t$ is assumed i.i.d. normal. The objective of blind source separation is to estimate the sources, the mixing parameters, and the parameters of the noise distribution.

Most blind deconvolution methods are based on higher-order statistics, see e.g. [4], [1]. However, the approach is proposed by Parra and Spence [5] is based on second order statistics and is attractive for its relative simplicity and implementation, yet excellent performance. The Parra and Spence algorithm is based on estimation of the inverse mixing process which maps measurements to source signals. A heuristic second order correlation function is minimized by the adaptation of the inverse process. The scheme needs multiple correlation measurements to obtain a unique inverse. This can be achieved, e.g., if the source signals are non-stationary or if the correlation functions are measured at time lags less than the correlation length of the source signals.

The main contribution of the present work is to provide an explicit statistical model for the decorrelation of convolutive mixtures of non-stationary signals. As a result, all parameters including mixing filter coefficients, source signal parameters and observation noise covariance are estimated by maximum-likelihood and the *exact* posterior distribution of the sources is obtained. The formulation is rooted in the theory of linear Gaussian models, see e.g., the review by Ghahramani and Roweis in [7]. The so-called Kalman Filter model is a state space model that can be set up to represent convolutive mixings of statistically independent sources added with observation noise. The standard estimation scheme for the Kalman filter model is an EM-algorithm that implements maximum-likelihood (ML) estimation of the parameters and maximum-posterior (MAP) inference of the source signals, see e.g. [3]. The specialization of the Kalman Filter model to convolutive mixtures is covered in section 2 while the adaptation of the model parameters is described in section 3. An experimental evaluation on a speech mixture is presented in section 4.

2. THE MODEL

The Kalman filter model is a generative dynamical state-space model that is typically used to estimate unobserved or hidden variables in dynamical systems, e.g. the velocity of an object whose position we are tracking. The basic Kalman filter model (no control inputs) is defined as

$$\begin{aligned} \mathbf{s}_t &= \mathbf{F}\mathbf{s}_{t-1} + \mathbf{v}_t \\ \mathbf{x}_t &= \mathbf{A}\mathbf{s}_t + \mathbf{n}_t \end{aligned} \quad (2)$$

The observed d_x-dimensional mixture, $\mathbf{x}_t = [x_{1,t}, x_{2,t}, .., x_{d_x,t}]^T$, is obtained from the multiplication of the mixing matrix, $\mathbf{A}$, on $\mathbf{s}_t$, the hidden state. The source innovation noise, $\mathbf{v}_t$, and the evolution matrix, $\mathbf{F}$, drive the sources. The signals are distributed as $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and $\mathbf{s}_1 \sim \mathcal{N}(\mu, \mathbf{\Sigma})$.

By requiring $\mathbf{F}, \mathbf{Q}$ and $\mathbf{\Sigma}$ to be diagonal matrices, equation (2) satisfies the fundamental requirement of

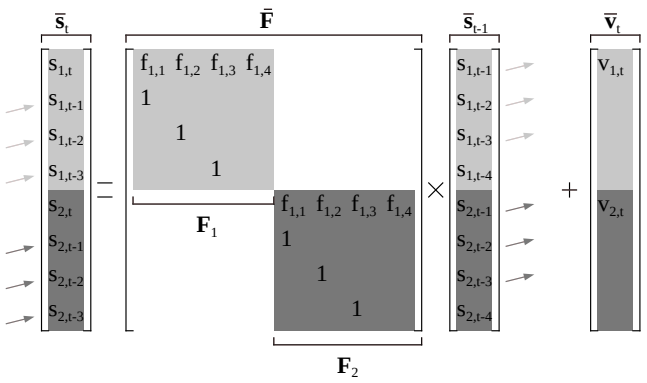

Figure 1: The AR(4) source signal model. The memory of $\mathbf{s}_t$ is updated by discarding $s_{i,t-4}$ and composing new $\mathbf{s}_{1,t}$ and $\mathbf{s}_{2,t}$ using the AR recursion. Blanks signify zeros.

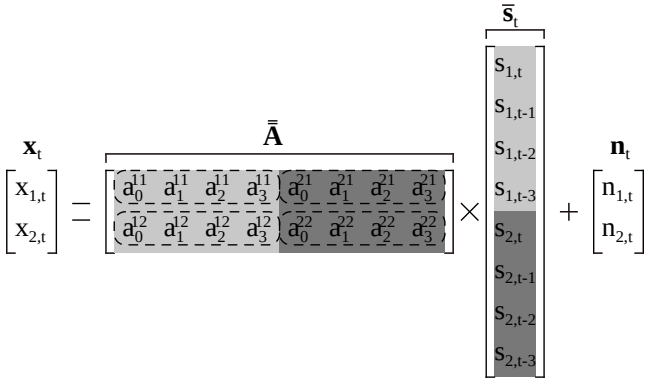

Figure 2: The convolutive mixing model requires a full $\bar{\bar{\mathbf{A}}}$ to be estimated.

any ICA formulation, namely that the sources are statistically independent. Under the diagonal constraint, this source model is identical to an AR(1) random process. In order for the Kalman model to be useful in the context of convolutive ICA for general temporally correlated sources we need to generalize it in two aspects, firstly we will move to higher order AR processes by stacking the state space, secondly we will introduce convolution in the observation model.

2.1 Model generalization

By generalizing (2) to AR(p) source models we can model wider classes of signals, including speech. The AR(p) model for source i is defined as:

$$s_{i,t} = f_{i,1}s_{i,t-1} + f_{i,2}s_{i,t-2} + .. + f_{i,p}s_{i,t-p} + v_{i,t}. \quad (3)$$

In line with e.g. [2], we implement the AR(p) process in the basic Kalman model by stacking the variables and parameters to form the augmented state vector

$$\bar{\mathbf{s}}_t = \left[\begin{array}{cccc} \mathbf{s}_{1,t}^T & \mathbf{s}_{2,t}^T & .. & \mathbf{s}_{d_s,t}^T \end{array} \right]^T$$

where the bar indicates stacking. The 'memory' of the individual sources is now represented in $\mathbf{s}_{i,t}$:

$$\mathbf{s}_{i,t} = \left[\begin{array}{cccc} s_{i,t} & s_{i,t-1} & .. & s_{i,t-p+1} \end{array} \right]^T$$

The stacking procedure consists of including the last p samples of $\mathbf{s}_t$ in $\bar{\mathbf{s}}_t$ and passing the $(p-1)$ most recent of those unchanged to $\bar{\mathbf{s}}_{t+1}$ while obtaining a new $\mathbf{s}_t$ by the AR(p) recursion of equation (3). Figure 1 illustrates the principle for two AR(4) sources. The involved parameter matrices must be constrained in the following way to enforce the independency assumption:

$$\bar{\mathbf{F}} = \left[\begin{array}{cccc} \bar{\mathbf{F}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{F}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{F}}_L \end{array} \right]$$

$$\bar{\mathbf{F}}_i = \left[\begin{array}{ccccc} f_{i,1} & f_{i,2} & \cdots & f_{i,p-1} & f_{i,p} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{array} \right]$$

$$\bar{\mathbf{Q}} = \left[\begin{array}{cccc} \bar{\mathbf{Q}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{Q}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{Q}}_L \end{array} \right]$$

$$(\bar{\mathbf{Q}}_i)_{jj'} = \left\{ \begin{array}{ll} q_i & j = j' = 1 \\ 0 & j \neq 1 \bigvee j' \neq 1 \end{array} \right.$$

Similar definitions apply to $\bar{\mathbf{\Sigma}}$ and $\bar{\mu}$. The generalization of the Kalman Filter model to represent convolutive mixing requires only a slight additional modification of the observation model, augmenting the observation matrix to a full $d_x \times p \times d_s$ matrix of filters,

$$\bar{\bar{\mathbf{A}}} = \left[\begin{array}{cccc} \mathbf{a}_{11}^T & \mathbf{a}_{12}^T & .. & \mathbf{a}_{1d_s}^T \\ \mathbf{a}_{21}^T & \mathbf{a}_{22}^T & .. & \mathbf{a}_{2d_s}^T \\ \mathbf{a}_{d_x1}^T & \mathbf{a}_{d_x2}^T & .. & \mathbf{a}_{d_xd_s}^T \end{array} \right]$$

where $\mathbf{a}_{ij} = [a_{ij,1}, a_{ij,2}, .., a_{ij,L}]^T$ is the length $L(= p)$ impulse response of the signal path between source i and sensor j. Figure 2 illustrates the the convolutive mixing matrix.

It is well-known that deconvolution cannot be performed using *stationary* second order statistics. We therefore follow Parra and Spence and *segment* the signal in windows in which the source signals can be assumed stationary. The overall system then reads

$$\begin{aligned} \bar{\mathbf{s}}_t^n &= \bar{\mathbf{F}}^n \bar{\mathbf{s}}_{t-1}^n + \bar{\mathbf{v}}_t^n \\ \mathbf{x}_t^n &= \bar{\bar{\mathbf{A}}} \bar{\mathbf{s}}_t^n + \mathbf{n}_t^n \end{aligned}$$

where n identify the segment of the observed mixture. A total of N segments are observed. For learning we will assume that during this period the mixing matrices $\bar{\bar{\mathbf{A}}}$ and the observation noise covariance, $\mathbf{R}$ are stationary.

3. LEARNING

A main benefit of having formulated the convolutive ICA problem in terms of a linear Gaussian model is that we can draw upon the extensive literature on parameter learning for such models. The likelihood is defined in abstract form for hidden variables $\mathbf{S}$ and parameters θ

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log \int d\mathbf{S} p(\mathbf{X}, \mathbf{S}|\theta)$$

The generic scheme for maximum likelihood learning of the parameters is the EM algorithm. The EM algorithm introduces a model posterior pdf. $\hat{p}(\cdot)$ for the hidden variables

$$\mathcal{L}(\theta) \geq \mathcal{F}(\theta, \hat{p}) \equiv \mathcal{J}(\theta, \hat{p}) - \mathcal{R}(\hat{p}) \tag{4}$$

where

$$\begin{aligned} \mathcal{J}(\theta, \hat{p}) &\equiv \int d\mathbf{S} \hat{p}(\mathbf{S}) \log p(\mathbf{X}, \mathbf{S}|\theta) \\ \mathcal{R}(\hat{p}) &\equiv \int d\mathbf{S} \hat{p}(\mathbf{S}) \log \hat{p}(\mathbf{S}) \end{aligned}$$

In the E-step we find the conditional source pdf based on the most recent parameter estimate, $\hat{p}(\mathbf{S}) = p(\mathbf{S}|\mathbf{X}, \theta)$. For linear Gaussian models we achieve $\mathcal{F}(\theta, \hat{p}) = \mathcal{L}(\theta)$. The M-step then maximize $\mathcal{J}(\theta, \hat{p})$ wrt. θ. Each combined M and E step cannot decrease $\mathcal{L}(\theta)$.

3.1 E-step

The Markov structure of the Kalman model allows an effective implementation of the E-step referred to as the *Kalman smoother*. This step involves forward-backward recursions and outputs the relevant statistics of the posterior probability $p(\bar{\mathbf{s}}_t|\mathbf{x}_{1:\tau}, \theta)$, and the log-likelihood of the parameters, $\mathcal{L}(\theta)$[1]. The posterior source mean (i.e. the posterior average conditioned on the given segment of observations) is given by

$$\hat{\bar{\mathbf{s}}}_t \equiv \langle \bar{\mathbf{s}}_t \rangle$$

for all t. The relevant second order statistics, i.e. source i autocorrelation and time-lagged autocorrelation, are:

$$\begin{aligned} \mathbf{M}_{i,t} &\equiv \langle \mathbf{s}_{i,t} (\mathbf{s}_{i,t})^T \rangle \\ &\equiv [\ \mathbf{m}_{i,1,t} \quad \mathbf{m}_{i,2,t} \quad .. \quad \mathbf{m}_{i,L,t}\ ]^T \\ \mathbf{M}^1_{i,t} &\equiv \langle \mathbf{s}_{i,t} (\mathbf{s}_{i,t-1})^T \rangle \end{aligned}$$

The block-diagonal autocorrelation matrix for $\bar{\mathbf{s}}_t$ is denoted $\bar{\mathbf{M}}_t$., It contains the individual $\mathbf{M}_{i,t}$, for $i = 1, 2, .., d_s$.

[1]For notational brevity, the segment indexing by n has been omitted in this section.

3.2 M-step

In the M-step, the first term of (4) is maximized with respect to the parameters. This involves the average of the logarithm of the data model wrt. the source posterior from the previous E-step

$$\begin{aligned} \mathcal{J}(\theta, \hat{p}) = -\frac{1}{2} \sum_{n=1}^{N} [& \sum_{i=1}^{d_s} \log \det \mathbf{\Sigma}_i^n + (\tau - 1) \sum_{i=1}^{d_s} \log q_i^n \\ & + \tau \log \det \mathbf{R} + \sum_{i=1}^{d_s} \langle (\mathbf{s}_{i,1}^n - \mu_i^n)^T (\mathbf{\Sigma}_i^n)^{-1} (\mathbf{s}_{i,1}^n - \mu_i^n) \rangle \\ & + \sum_{t=2}^{\tau} \sum_{i=1}^{d_s} \langle \frac{1}{q_i^n} (s_{i,t}^n - (\mathbf{f}_i^n)^T \mathbf{s}_{i,t-1}^n)^2 \rangle \\ & + \sum_{t=1}^{\tau} \langle (\mathbf{x}_t^n - \bar{\bar{\mathbf{A}}} \bar{\mathbf{s}}_t^n)^T \mathbf{R}^{-1} (\mathbf{x}_t^n - \bar{\bar{\mathbf{A}}} \bar{\mathbf{s}}_t^n) \rangle] \end{aligned}$$

where $\mathbf{f}_i^T = [\ f_{i,1} \quad f_{i,2} \quad .. \quad f_{i,p}\ ]$. The derivations are analogous with the formulation of the EM algorithm in [3]. The special constrained structure induced by the independency of the source signals introduces tedious but straight-forward modifications. The segment-wise update equations for the M-step are:

$$\begin{aligned} \mu_{i,\mathbf{new}} &= \hat{\mathbf{s}}_{i,1} \\ \mathbf{\Sigma}_{i,\mathbf{new}} &= \mathbf{M}_{i,1} - \mu_{i,\mathbf{new}} \mu_{i,\mathbf{new}}^T \\ \mathbf{f}_{i,\mathbf{new}}^T &= \Big[\sum_{t=2}^{\tau} (\mathbf{m}_{i,t}^1)^T \Big] \Big[\sum_{t=1}^{\tau} \mathbf{M}_{i,t-1} \Big]^{-1} \\ q_{i,\mathbf{new}} &= \frac{1}{\tau - 1} \Big[\sum_{t=2}^{\tau} m_{i,t} - \mathbf{f}_{i,\mathbf{new}}^T \mathbf{m}_{i,t}^1 \Big] \end{aligned}$$

Reconstruction of $\bar{\mu}_{\mathbf{new}}$, $\bar{\mathbf{\Sigma}}_{\mathbf{new}}$, $\bar{\mathbf{F}}_{\mathbf{new}}$ and $\bar{\mathbf{Q}}_{\mathbf{new}}$ from the above is performed according to the stacking definitions of section 2. The estimators $\bar{\bar{\mathbf{A}}}_{\mathbf{new}}$ and $\mathbf{R}_{\mathbf{new}}$ include the statistics from all observed segments:

$$\begin{aligned} \bar{\bar{\mathbf{A}}}_{\mathbf{new}} &= \Big[\sum_{n=1}^{N} \sum_{t=1}^{\tau} \mathbf{x}_{t,n} (\hat{\bar{\mathbf{s}}}_{t,n})^T \Big] \Big[\sum_{n=1}^{N} \sum_{t=1}^{\tau} \bar{\mathbf{M}}_{t,n} \Big]^{-1} \\ \mathbf{R}_{\mathbf{new}} &= \frac{1}{N\tau} \sum_{n=1}^{N} \sum_{t=1}^{\tau} \mathrm{diag}[\mathbf{x}_{t,n} \mathbf{x}_{t,n}^T - \bar{\bar{\mathbf{A}}}_{\mathbf{new}} \hat{\bar{\mathbf{s}}}_{t,n} \mathbf{x}_{t,n}^T] \end{aligned}$$

We accelerate the EM learning by a relaxation of the lower bound, which amounts to updating the parameters proportionally to an self-adjusting step-size, α, as described in [6]. We refer to the Kalman filter based blind source separation approach as 'KaBSS'.

4. EXPERIMENTS

The proposed algorithm was tested on a binaural convolutive mixture of two speech signals with additive noise in varying signal to noise ratios (SNR). A male speaker generated *both signals* that were recorded at $8kHz$. This is a strong test of the blind separation ability, since the 'spectral overlap' is maximal for a single speaker.

The noise-free mixture was obtained by convolving the source signals with the impulse responses:

$$\bar{\bar{\mathbf{A}}} = \begin{bmatrix} 1 & 0.3 & 0 & 0 & 0 & 0.8 \\ 0 & 0.8 & 0.24 & 1 & 0 & 0 \end{bmatrix}$$

Subsequently, observation noise was added in each sensor channel to construct the desired SNR. Within each experiment, the algorithm was restarted 10 times, each time estimating the parameters from 10 randomly sampled segments of length $\tau = 70$. Based on a test log-likelihood, $\mathcal{L}_{test}(\theta)$, the best estimates of $\bar{\bar{\mathbf{A}}}$ and $\mathbf{R}$ were used to infer the source signals and estimate the source model ($\bar{\mathbf{F}}$ and $\bar{\mathbf{Q}}$). The model parameters were set to $p = 2$ and $L = 3$.

The separation quality was compared with the State-of-the-Art method proposed by Parra and Spence[2][5]. A signal to interference ratio (SIR): SIR $= \frac{P_{11}+P_{22}}{P_{12}+P_{21}}$ is used as comparison metric. P_{ij} is the power of the signal constituting the contribution of the ith original source to the jth source estimate. The normalized cross-correlation function was used to estimate the powers involved. The ambiguity of the source assignment was fixed prior to the SIR calculations. The results are shown in figure 3. Noise-free scenarios excepted, the new method produce better signal-to-interference values peaking at an improvement of 4dB for an SNR of 20dB. It should be noted that the present method is considerably more computational demanding than the reference method.

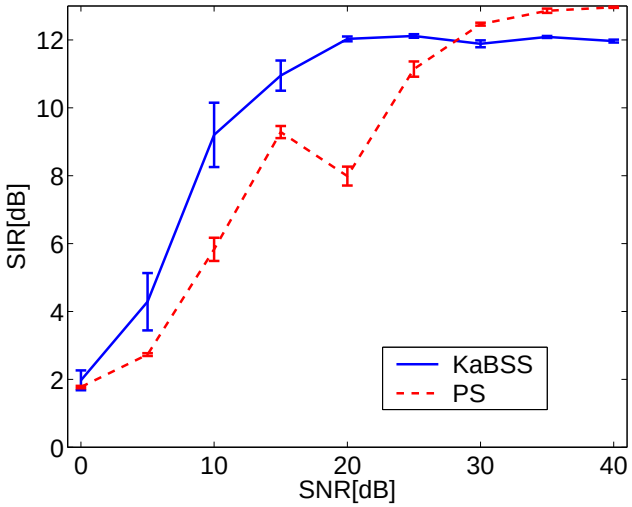

Figure 3: The separation performance for varying SNR of KaBSS and the reference method proposed by Parra and Spence (PS) [5]. The signals are two utterances by the same speaker. Two convolutive mixtures were created with variable strength additive white noise. The SIR measures the crosstalk between the two sources in the source estimates. The error bars represent the standard deviation of the mean for 10 experiments at each SNR.

5. CONCLUSION

Blind source separation of non-stationary signals has been formulated in a principled probabilistic linear Gaussian framework allowing for (exact) MAP-estimation of the sources and ML-estimation of the parameters. The derivation involved augmentation of state-space representation to model higher order AR processes and augmentation of the observation model to represent convolutive mixing. The independency constraint could be implemented exactly in the parameter estimation procedure. The source estimation and the parameter adaptation procedures are based on second-order statistics ensuring robust estimation for many classes of signals. In comparison with other current convolutive ICA models the present setup allows blind separation of noisy mixtures and it can estimate the noise characteristics. Since it is possible to compute the likelihood function on test data it is possible to both use validation sets for model order estimation as well as approximate schemes such as AIC and BIC based model order selection. A simulation study was used to validate the model in comparison with a State-of-the-Art reference method. The simulation consisted in a noisy convolutive mixture of two recordings of the *same* speaker. The simulation indicated that speech signals are described well-enough by the colored noise source model to allow separation. For the given data set, the proposed algorithm outperforms the reference method for a wide range of noise levels. However, the new method is computationally demanding. We expect that significant optimization and computational heuristics can be invoked to simplify the algorithm for real-time applications. Likewise, future work will be devoted to monitor and tune the convergence of the EM algorithm.

[2]See "http://newton.bme.columbia.edu/ lparra/publish/". The hyper-parameters of the reference method were fitted to the given data-set: $T = 1024$, $Q = 6$, $K = 7$ and $N = 5$. It should be noted that the estimated SIR is sensitive to the hyper-parameters.

REFERENCES

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

[2] G. Doblinger. An adaptive Kalman filter for the enhancement of noisy AR signals. In *IEEE Int. Symp. on Circuits and Systems*, volume 5, pages 305–308, 1998.

[3] Z. Ghahramani and G. E. Hinton. Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, Department of Computer Science, University of Toronto, 2 1996.

[4] T.W. Lee, A. J. Bell, and R. H. Lambert. Blind separation of delayed and convolved sources. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9, page 758. The MIT Press, 1997.

[5] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions Speech and Audio Processing*, pages 320–7, 5 2000.

[6] R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani. Optimization with EM and Expectation-Conjugate-Gradient. In *International Conference on Machine Learning*, volume 20, pages 672–679, 2003.

[7] S.Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345, 1999.