# Preface

The work leading to this Master's thesis was carried out at Informatics and Mathematical Modelling between the August 4, 2003 and February 2, 2004. Professor Lars Kai Hansen supervised the project.


**Thanks**

First and foremost, I thank Lars Kai Hansen for his professionalism and dedicated effort to motivate the students. Furthermore, I wish to thank Daniel Jacobsen, Frederik Brink, Mikkel Putzek, Slimane Bazou and Thomas Stolz for sharing office with me and for discussing various issues as well as providing ideas for the project. Mads Dyrholm and Tue Lehn-Schiøler shared knowledge and proofread sections of the report. So did Daniel Jacobsen, Frederik Brink and Mikkel Putzek. Thanks to Anke for doing the dishes.


Lyngby, February 2, 2004

Rasmus Kongsgaard Olsson

# Abstract

This thesis focuses on Blind source separation (BSS), which is the problem of finding hidden source signals in observed mixtures given no or little knowledge about the sources and the mixtures. Based on the well-performing, yet heuristically based, algorithm of Parra and Spence, 2000, a probabilistic model is formulated for the BSS problem. A time-domain EM algorithm 'KaBSS' is derived which estimates the source signals, the associated second-order statistics, the mixing filters and the observation noise covariance matrix. In line with the literature, it is found that the estimated quantities are unique within the model only if the sources can be assumed non-stationary and contain sufficient time-variation. Furthermore, the statistical framework is exploited in order to assess the correct model order: the number of sources within the mixture can be determined using the socalled Bayes Information Criterion (BIC). Monte Carlo simulations as well as experimental results for mixtures of speech signals are documented and compared to results obtained by the algorithm of Parra and Spence.

*Keywords:* Blind source separation, Independent component analysis, non-stationary sources, EM.

# Resumé

Denne afhandlings emne er blind signalseparation (BSS), der drejer sig om at estimere skjulte kildesignaler i observerede blandinger på basis af ringe eller ingen viden om kildesignaler og blandinger. En probabilistisk model for BSS-problemet formuleres med afsæt i Parra og Spences (2000) højt-ydende, men heuristisk funderede algoritme. På baggrund af modellen udledes 'KaBSS', en EM-algoritme, der estimerer kildesignalerne og deres 2. ordensstatistik, blandingsfiltrene og observationsstøjens kovarians. I overensstemmelse med litteraturen findes det, at de estimerede størrelser kun er unikke indenfor modellen, hvis en antagelse om kildernes ikke-stationaritet er rimelig, og hvis kilderne er tilstrækkelig tidsvariante. Ydermere udnyttes den statistiske ramme til at vurdere den korrekte modelorden: Antallet af kilder i blandingen fastslås ved at benytte det såkaldte Bayes Information Criterion (BIC). Såvel Monte Carlo simulationer som eksperimentelle resultater for blandinger af talesignaler dokumenteres og sammenlignes med resultater, opnået via Parra og Spences algoritme.

# Nomenclature

Below follow the most used symbols and abbreviations. Scalars, vectors and matrices appear as : $y$, $\mathbf{y}$ and $\mathbf{Y}$.

| | |
|---|---|
| $p$ | The order of an autoregressive random process |
| $L$ | The filter length of the source signal channels |
| $d_s$ | The number of sources |
| $d_x$ | The number of sensors |
| $N$ | The number of segments |
| $\mathbf{x}_t$ | Multivariate sensor signal at time $t$ |
| $\mathbf{n}_t$ | Multivariate sensor noise signal at time $t$ |
| $\tau$ | The number samples in a segment |
| $\mathbf{s}_t$ | Multivariate source signal at time $t$ |
| $s_{i,t}$ | Source signal $i$ at time $t$ |
| $\bar{\mathbf{s}}_t$ | Multivariate source signal at time $t$. Stacked for use in the model concerning the mixing of AR(p) random processes |
| $\mathbf{v}_t$ | Multivariate source innovation noise signal at time $t$ |
| $\bar{\mathbf{v}}_t$ | Multivariate source innovation noise signal at time $t$. Stacked for use in the model concerning the mixing of AR(p) random processes |
| $v_{i,t}$ | Innovation noise signal of source $i$ at time $t$ |
| $\theta$ | Set of all parameters |
| $\mathbf{A}$ | The mixing matrix of the model concerning the instantaneous mixing of AR(1) random processes |
| $\bar{\mathbf{A}}$ | The mixing matrix of the model concerning the instantaneous mixing of AR(p) random processes |
| $\bar{\bar{\mathbf{A}}}$ | The mixing matrix of the model concerning the convolutive mixing of AR(p) random processes |
| $\mathbf{R}$ | The sensor noise covariance matrix |

| | |
|---|---|
| $\mathbf{F}$ | The evolution matrix of the AR(1) random process sources |
| $\bar{\mathbf{F}}$ | The evolution matrix of the AR(p) random process sources |
| $\bar{\mathbf{f}}_i$ | The AR parameters of source $i$ |
| $\mathbf{Q}$ | The innovation noise covariance of the AR(1) random process sources |
| $\bar{\mathbf{Q}}$ | The innovation noise covariance of AR(p)random process sources |
| $q_i$ | The innovation noise variance of source $i$ |
| $\mathcal{L}(\theta)$ | The log-likelihood function of the parameter vector $\theta$ |
| $\alpha$ | Adaptation rate of the stepsize $\eta$ |
| $\eta$ | Step-size of the AEM algorithm |
| KaBSS | Kalman blind source separation |
| BSS | Blind source separation or Blind signal separation |
| ICA | Independent component analysis |
| EM | Expectation-maximization |
| AEM | Adaptive overrelaxed expectation-maximization |
| BIC | Bayes information criterion |

# Contents

# Chapter 1

# Introduction

Blind source separation (BBS) is the problem of recovering the hidden source signals from a number of different mixture signals, which are observed at sensors. The term *blind* refers to the fact that the mixing process and the source signals are unknown. An example that occurs to most humans is the so-called cocktail party problem: *during a social event, extract a single voice from the composition of chatter and other noises that reach the ears.* The cocktail party problem is especially relevant due to its applicability in hearing aids and speech recognition. It is a problem that is solved in difficult noise settings by the human auditory system. No algorithm has ever come close to that.

The general BSS problem comes at many levels of difficulty differentiated by various mixing processes and the numbers of sources and sensors. However, two *linear* mixing functions completely dominate the literature. They have names referring to the fact that often sources are best described as time-series, or signals. The first and simplest of those is the instantaneous mixing:

$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \mathbf{n}_t$$

where the sensor vector, $\mathbf{x}_t$, at time $t$ results from a matrix multiplication of the mixing matrix, $\mathbf{A}$, with the source signal, $\mathbf{s}_t$, added with observation noise, $\mathbf{n}_t$. The dimensions of vectors $\mathbf{x}_t$ and $\mathbf{s}_t$ correspond to the number of sensors and sources, $d_x$ and $d_x$, respectively. A more challenging task arises when the sources have been mixed convolutively, i.e. a convolution sum involving the source signals is required for the description of data:

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

where $\mathbf{A}_k$ is a matrix of *filters* and contains $L$ times the number of parameters in $\mathbf{A}$.

The instantaneous mixing model is insufficient for many real-life problems, since physical signals, like sound waves, propagate with finite velocity. Hence the signals arrive at different times at different sensors, requiring a model that can handle delays. Figure 1.1 illustrates a situation where a signal travels a number of paths to reach a sensor. Different attenuations and delays result.

Blind source separation algorithms function in mainly two ways: One family of methods exploits that the probability density function of $\mathbf{x}_t$ bear traits of

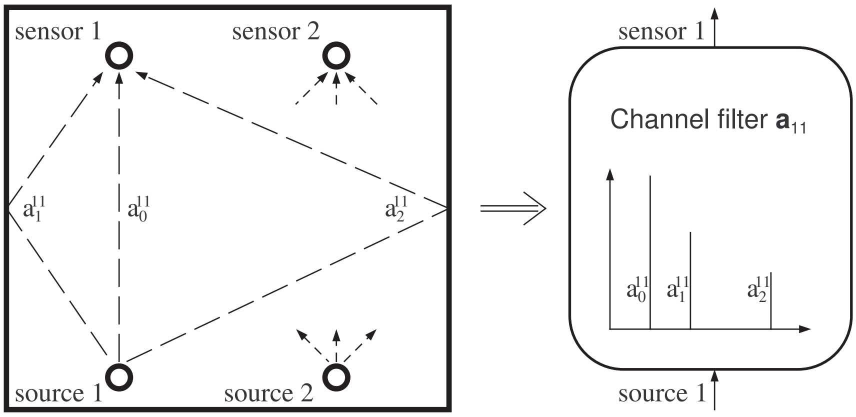


Figure 1.1: The convolutive mixing problem. On the left: the signal from source 1 travels to sensor 1 on a number of paths associated with different delays and attenuations. On the right: the impulse response of the linear channel filter models the signals paths. The signal at the sensor is obtained by convolving the source signal with the filter.

the (non-Gaussian) sparse or dense distribution of $\mathbf{s}_t$. Independent component analysis (ICA) estimates $\mathbf{A}$ or the equivalent inverse model $\mathbf{W}$ using those higher-order statistics, see e.g. [1] and [2]. This is mainly a spatial approach, in that the temporal correlation of the individual sources is ignored.

Another group of algorithms identify the sources based on their temporal distribution. A pioneer in this field is Molgedey's and Schuster's decorrelation algorithm [3] for instantaneous noise-free mixings. The mixing matrix, $\mathbf{A}$, is found by computing the time-lagged second-order statistics of $\mathbf{x}_t$, i.e. $\mathbf{R}_x(\tau) = \sum_t \mathbf{x}_t \mathbf{x}_{t-\tau}^T$, at lags $\tau = 0, \tau_0$ and diagonalizing it by the solving of the resulting eigenvalue problem. Second-order statistics are implicitly assumed a sufficient descriptor for $\mathbf{x}_t$ and $\mathbf{s}_t$. While computationally efficient, the algorithm is limited in a number of ways: it addresses primarily noise-free, quadratic, i.e. the number of sources and observation channels are equal, mixtures of stationary source signals.

The direct application of the decorrelation technique to the convolutive problem does not provide solutions that are unique. By explicitly assuming the non-stationarity of the sources and measuring the second-order statistics at different times, sufficient constraints are imposed on $\mathbf{A}$. A number of authors have exploited this fact: Parra and Spence [4] provide well-performing off-line and on-line algorithms for the noisy convolutive mixture problem based on decorrelation in the frequency domain. Matsuoko, Ohya and Kawamoto also work along those lines, see [5] In [6], the problem is solved in the time-domain. Higher-order statistics and temporal methods converge in a vast number contributions, e.g. [7] and [8].

Common to many of the contributions mentioned is the technique of transforming the convolutive problem into an instantaneous problem by the means of a discrete Fourier transform (DFT). Replacing convolution with multiplication is attractive but comes at a cost, see section 2.4.3. Recently, Anemüller and Kollmeier have elaborated this approach by considering correlation across bands in the spectrogram, see [9].

The main contribution of this work is to provide an explicit probabilistic model and its associated estimators for the decorrelation of convolutive mixtures of non-stationary signals. The algorithm, which is termed 'KaBSS', estimates all parameters including mixing filter coefficients, source signal parameters and observation noise covariance and the posterior distribution of the sources conditioned on the observations by employing an Expectation Maximization (EM) scheme. Parts of this work have been submitted for publication, see appendix C and [10].

A formulation of the convolutive problem in the general framework of Gaussian linear models, well reviewed by Ghahramani and Roweis in [11], serves as a starting point for the derivation of the algorithm. The Kalman Filter model is a special case that can be made serve the purpose of modelling the instantaneous or convolutive mixings of statistically independent sources added with observation noise. The natural estimation scheme for the Kalman filter model is the EM-algorithm which iteratively employs maximum-likelihood (ML) estimation of the parameters and maximum-posterior (MAP) inference of the source signals, see e.g. [12]. The log-likelihood of the parameters is computed exactly, which can be used to determine the correct model order, e.g. the number of sources. In conclusion, the thesis has the following focus.

> **Problem statement**
> Based on the decorrelation algorithms [3] and [4] devise a statistical model for the blind source separation of instantaneous and convolutive mixtures. Investigate the identifiability of the parameters and derive the estimators for the algorithm. Explore the conditions that allow for artificially generated and real mixtures to be separated by the algorithm. Exploit the advantages that are associated with being probabilistic, such as estimating the noise levels and determining the model order. Give suggestions to promising paths of future research.

The specialization of the Kalman filter model to non-stationary convolutive mixtures is covered in chapter 2, while the learning in this particular model is described in chapter 3. Monte Carlo simulations and experiments with real speech data are documented in chapter 4.

# Chapter 2

# The model

The following chapter will address the formulation of a model that fits the specifications of a general blind source separation problem. To begin with, we note that half the model is already specified by either the instantaneous or the convolutive mixing defined in the introductory chapter. This part of the model we term the *observation model*, and both mixing functions are addressed in the following.

What is left to modify is the *source model*. The literature contains a rich variety of suggestions how to best describe the sources. However, one assumption is common to the vast majority of contributions, namely the statistical independence of the sources:

$$p(s_{1,t}, s_{2,t}, .., s_{d_s,t}) = \prod_{i=1}^{d_s} p(s_{i,t}) \tag{2.1}$$

where $s_{i,t}$ is the $i$th element of the source vector $\mathbf{s}_t$. It has already been mentioned that some methods exploit the temporal correlation of each of the sources, which is an approach that will be elaborated here.

## 2.1 AR(1)

To begin with, we suggest a first-order linear autoregressive process:

$$\mathbf{s}_t = \mathbf{F}\mathbf{s}_{t-1} + \mathbf{v}_t$$

where $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ is the source innovation noise. The innovation noise covariance, $\mathbf{Q}$, and the evolution matrix $\mathbf{F}$ are assumed diagonal in order to abide to equation 2.1. Furthermore, stability is ensured by $|(F)_{ii}| < 1$, for all $i$. The AR process is started from $\mathbf{s}_1 \sim \mathcal{N}(\mu, \mathbf{\Sigma})$.

It is now noted that the above recursion fits into the general framework of Gaussian linear models, lately popularized by Roweis and Ghahramani, [11]. A special case is the Kalman filter model, which consists of a dynamical continuous state-space model and an observation model identical to the instantaneous mixing model. In other words, the Kalman filter model (no inputs) fits our purposes perfectly:

$$\begin{aligned} \mathbf{s}_t &= \mathbf{F}\mathbf{s}_{t-1} + \mathbf{v}_t \\ \mathbf{x}_t &= \mathbf{A}\mathbf{s}_t + \mathbf{n}_t \end{aligned} \tag{2.2}$$

The observed $d_x$-channel mixture, $\mathbf{x}_t$, results from the multiplication of the mixing matrix, $\mathbf{A}$, on $\mathbf{s}_t$. The observation noise is distributed as $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, where $\mathbf{R}$ is assumed diagonal for simplicity.

In conclusion, by requiring $\mathbf{F}, \mathbf{Q}$ and $\mathbf{\Sigma}$ to be diagonal matrices, the Kalman filter model satisfies the fundamental requirement of any ICA formulation, namely that the sources are statistically independent. The underlying source model is an AR(1) process.

## 2.2 AR(p) sources

Limiting the source model to a 1st order AR random process may prevent achieving results with real world signals. By generalizing equation 2.2 to employ AR($p$) models for the sources, many classes of signals, including speech, are expected to be well described - at least on a small time-scale. Speech is examined closer in sections 2.5 and 4.1. The general AR(p) model for source $i$ is defined as follows:

$$s_{i,t} = f_{i,1}s_{i,t-1} + f_{i,2}s_{i,t-2} + .. + f_{i,p}s_{i,t-p} + v_{i,t} \tag{2.3}$$

where $v_{i,t} \sim \mathcal{N}(0, q_i)$.

A common 'trick' in dealing with Kalman filters allows for the inclusion of the above regression into the model, e.g. see [13]. The technique is based on the stacking of variables and parameters in order to maintain a memory of past samples of $\mathbf{s}_t$. At the same time, restrictions on the format of the matrices of the model are enforced to maintain the source independency assumption of (2.1). The stacked source vector is defined as follows:

$$\bar{\mathbf{s}}_t = \begin{bmatrix} \mathbf{s}_{1,t}^T & \mathbf{s}_{2,t}^T & \cdots & \mathbf{s}_{d_s,t}^T \end{bmatrix}^T \tag{2.4}$$

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

$$\mathbf{s}_{i,t} = \begin{bmatrix} s_{i,t} & s_{i,t-1} & \cdots & s_{i,t-p+1} \end{bmatrix}^T \tag{2.5}$$

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 (2.3). An example is shown in figure 2.1 that illustrates the principle for two AR(4) sources. The involved parameter matrices must be constrained in the following way to enforce the independency

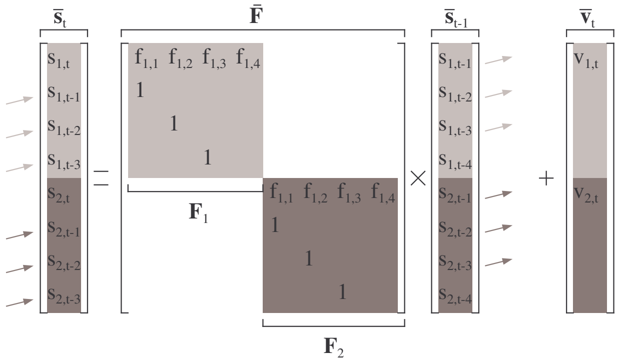


Figure 2.1: The AR(4) source signal model. The memory of $\mathbf{s}_t$ is updated by discarding $s_{i,t-4}$ and composing new $s_{1,t}$ and $s_{2,t}$ using the AR recursion. Blank spaces signify zeros.

assumption and produce the AR processes:

$$
\begin{aligned}
\bar{\mathbf{F}} &= \begin{bmatrix} \bar{\mathbf{F}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{F}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{F}}_L \end{bmatrix} \\
\bar{\mathbf{F}}_i &= \begin{bmatrix} f_{i,1} & f_{i,2} & \cdots & f_{i,p-1} & f_{i,p} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \\
\bar{\mathbf{Q}} &= \begin{bmatrix} \bar{\mathbf{Q}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{Q}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{Q}}_L \end{bmatrix} \\
(\bar{\mathbf{Q}}_i)_{jj'} &= \begin{cases} q_i & j = j' = 1 \\ 0 & j \neq 1 \bigvee j' \neq 1 \end{cases}
\end{aligned}
$$

Similar definitions apply to $\Sigma$ and $\mu$. Since only the $s_{i,t}$'s are relevant to the instantaneous mixing, the delayed versions are discarded by inserting $(d_s-1)\times d_x$ dimensional matrices of zeros into $\mathbf{A}$:

$$
\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{0} & \mathbf{a}_2 & \mathbf{0} & .. & \mathbf{a}_L & \mathbf{0} \end{bmatrix}
$$

where the $\mathbf{a}_i$ is the $i$'th column of $\mathbf{A}$. Figure 2.2 illustrates this concept.

In conclusion, the basic Kalman Filter model formulation, describing the instantaneous mixture of AR(1) processes has been augmented to incorporate

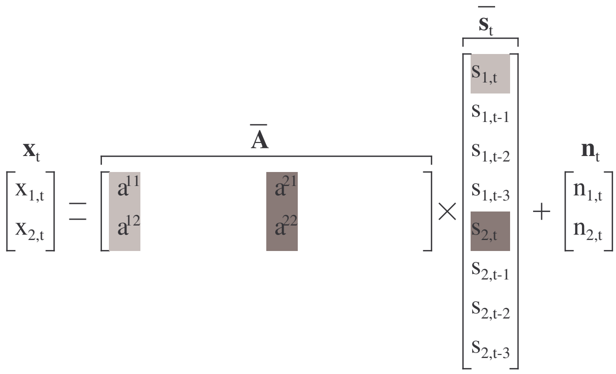


Figure 2.2: The instantaneous mixing model. Only the most recent sample from each source is included in the mixing. All other are excluded by the zeros of $\bar{\mathbf{A}}$. Blank spaces signify zeros.

AR(p) processes:

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

## 2.3 Convolutive mixing of AR(p) sources

It has already been argued that instantaneous mixing is inadequate for the solving of e.g. a real cocktail party problem. Conveniently, a further generalization of the Kalman Filter model to convolutive mixing requires only a slight modification of the observation model, namely an 'upgrade' of $\bar{\mathbf{A}}$ to a full $d_x \times (p \times d_s)$ matrix of mixing filters:

$$\bar{\bar{\mathbf{A}}} = \begin{bmatrix} \mathbf{a}_{11}^T & \mathbf{a}_{12}^T & .. & \mathbf{a}_{1d_s}^T \\ \mathbf{a}_{21}^T & \mathbf{a}_{22}^T & .. & \mathbf{a}_{2d_s}^T \\ \mathbf{a}_{d_x1}^T & \mathbf{a}_{d_x2}^T & .. & \mathbf{a}_{d_xd_s}^T \end{bmatrix} \tag{2.6}$$

where $\mathbf{a}_{ij} = [a_{ij,1}, a_{ij,2}, .., a_{ij,L}]^T$ is the impulse response of the signal path between source $i$ and sensor $j$, length $L$. Figure 2.3 illustrates this principle.

It will be argued in details in section 2.4.2 that the sources cannot be recovered from a stationary convolutive mixing based solely on the second-order statistics of the sources. Inspired by [4], we instead assume that the sources are generally non-stationary in the long term and only stationary on the short term. This leads to a partition of all signals involved into $N$ *segments* that each contains $\tau$ consecutive samples of the original signals. The resulting convolutive mixing of AR(p) sources is still within the framework of the Kalman Filter model:

$$\begin{aligned}\bar{\mathbf{s}}_t^n &= \bar{\mathbf{F}}^n\bar{\mathbf{s}}_{t-1}^n + \bar{\mathbf{v}}_t^n \\ \mathbf{x}_t^n &= \bar{\bar{\mathbf{A}}}\bar{\mathbf{s}}_t^n + \mathbf{n}_t^n\end{aligned} \tag{2.7}$$

where $n = 1, 2, .., N$ is the segment index. The source model parameters, $\bar{\mathbf{F}}^n$, $\bar{\mathbf{Q}}^n$, $\bar{\mathbf{\Sigma}}^n$ and $\bar{\mu}^n$, are assumed segment-local. The channel, $\bar{\bar{\mathbf{A}}}$, and the observation noise covariance, $\mathbf{R}$, are assumed stationary over segments, on the other hand.

Hua et al. [14] prove that a stationary mixing (of colored noise sources) can be separated if the spectra of the sources are non-overlapping. Kawamoto et al. [15] develop an algorithm based on this principle. Wan et. al [16] uses an extended Kalman filter to model speech and noise in a noise-reduction setup. However, only a single observation channel is considered.

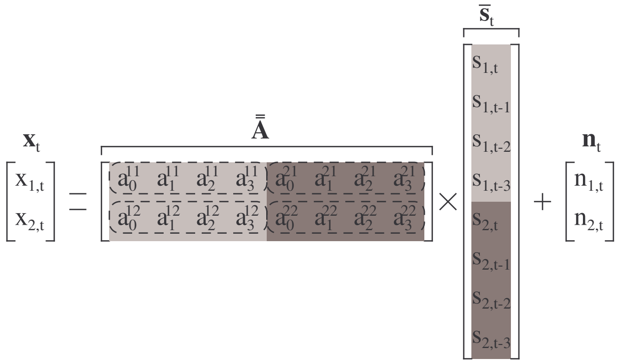


Figure 2.3: The convolutive mixing model requires a full $\bar{\bar{\mathbf{A}}}$ to be estimated.

## 2.4 Identifiability

Before proceeding to the learning of the parameters and the inference of the source signals, it is investigated to which degree the source signal and parameters are unique within the model. It will be uncovered that certain parameters may scale, rotate and permute arbitrarily, something that is often a tolerated evil in learning. Some of the ambiguities are already well-known, e.g. the permutation problem where source estimator 1 might estimate true source 2.

As a foundation for the following arguments we will argue that all signals involved are normally distributed. The individual source signals are jointly distributed according to a multivariate Gaussian distribution, because any $s_{i,t}$ can be expressed in terms of a linear operation on multivariate Gaussians, $s_{i,t-1}$ and $v_{i,t}$. The argument is then applied recursively until the initial source conditions are reached. A similar line of thought can be applied to $\mathbf{x}_t$. If zero mean is assumed, the time-lagged covariance $\mathbf{C}_x(\tau)$ is a sufficient statistic for the observed data, meaning that the identification of parameters and source signal must be based on this quantity alone. In the following, it is implicitly assumed that $\mathbf{C}_x(\tau)$ can be accurately estimated.

The distinct cases of the demixing of instantaneously mixed AR(1) processes and the demixing of convolutively mixed AR(p) processes are treated in the following sections.

### 2.4.1 Instantaneous mixing of AR(1) sources

All information about the sources and the mixing that can possibly be extracted is contained in $\mathbf{C}_x(\tau)$. To derive an expression for $\mathbf{C}_x(\tau)$, the fact is used that the observation noise $\mathbf{n}_t$ and the source signal $\mathbf{s}_t$ are statistically independent:

$$\mathbf{C}_x(\tau) \equiv \langle \mathbf{x}_t \mathbf{x}_{t+\tau}^T \rangle = \mathbf{A}\mathbf{C}_s(\tau)\mathbf{A}^T + \delta_\tau \mathbf{R} \tag{2.8}$$

where $\mathbf{C}_s(\tau) \equiv \langle \mathbf{s}_t \mathbf{s}_{t+\tau}^T \rangle$ and $\mathbf{C}_n(0) \equiv \langle \mathbf{n}_t \mathbf{n}_t^T \rangle = \mathbf{R}$. Averages, $\langle \cdot \rangle$, are performed over the relevant distributions and the stationarity of all processes is assumed.

In order to obtain an expression for $\mathbf{C}_s(\tau)$, the AR(1) process assumption on the source signals is invoked. To begin with, $\mathbf{C}_s(0)$ is determined by evaluating $\langle \mathbf{s}_t \mathbf{s}_t^T \rangle$:

$$\begin{aligned} \mathbf{C}_s(0) \equiv \langle \mathbf{s}_t \mathbf{s}_t^T \rangle &= \mathbf{F}\langle \mathbf{s}_{t-1} \mathbf{s}_{t-1}^T \rangle \mathbf{F}^T + \langle \mathbf{v}_t \mathbf{v}_t^T \rangle & (2.9) \\ &= \mathbf{F}\mathbf{C}_s(0)\mathbf{F}^T + \mathbf{Q} & (2.10) \end{aligned}$$

It is now exploited that $\mathbf{F}$ and $\mathbf{C}_s(0)$ are diagonal matrices, hence symmetric and commutative to multiplication:

$$\mathbf{C}_s(0) = \mathbf{Q}(\mathbf{I} - \mathbf{F}^2)^{-1} \tag{2.11}$$

In order to continue the derivation, a fortunate property of the autocorrelation function of AR-processes is exploited:

$$\begin{aligned} \mathbf{C}_s(\tau+1) \equiv \langle \mathbf{s}_t \mathbf{s}_{t-(\tau+1)}^T \rangle &= \mathbf{F}\langle \mathbf{s}_{t-1} \mathbf{s}_{t-(\tau+1)}^T \rangle \\ &= \mathbf{F}\mathbf{C}_s(\tau) & (2.12) \end{aligned}$$

Combining (2.11) and (2.12) with (2.8) yields the sought expression for $\mathbf{C}_x(\tau)$:

$$\mathbf{C}_x(\tau) = \mathbf{A}\mathbf{Q}\mathbf{F}^\tau(\mathbf{I} - \mathbf{F}^2)^{-1}\mathbf{A}^T + \delta_\tau \mathbf{R} \tag{2.13}$$

From the above, we see that an arbitrary permutation of the diagonal elements of $\mathbf{F}$, i.e. switching the order of the elements, can be 'undone' by a corresponding permutation of the diagonal elements of $\mathbf{Q}$ and the columns of $\mathbf{A}$. Furthermore $\mathbf{A}$ and $\mathbf{Q}^2$ scale inversely.

An alternative route to wisdom can be taken by examining the effect of the insertion of an invertible linear transformation $\mathbf{P}$ in the model, (2.2). We premultiply $\mathbf{s}_t$ by $\mathbf{P}$ and right-multiply $\mathbf{P}^{-1}$ on $\mathbf{A}$:

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

A linearly transformed source model is then obtained:

$$\tilde{\mathbf{s}}_t = \tilde{\mathbf{F}}\tilde{\mathbf{s}}_{t-1} + \tilde{\mathbf{v}}_t$$

with the definitions:

$$\begin{aligned} \tilde{\mathbf{s}}_t &\equiv \mathbf{P}\mathbf{s}_t \\ \tilde{\mathbf{F}} &\equiv \mathbf{P}\mathbf{F}\mathbf{P}^{-1} \\ \tilde{\mathbf{v}}_t &\equiv \mathbf{P}\mathbf{v}_t \end{aligned} \tag{2.14}$$

From definition 2.14 we get the following matrix equation:

$$\mathbf{PF} = \tilde{\mathbf{F}}\mathbf{P}$$

The above is treated element-wise and exploiting that $\tilde{\mathbf{F}}$ and $\mathbf{F}$ are assumed diagonal. For all $i, j$:

$$\begin{aligned} (\mathbf{PF})_{ij} &= (\tilde{\mathbf{F}}\mathbf{P})_{ij} \Leftrightarrow \\ f_{ii}p_{ij} &= p_{ij}\tilde{f}_{jj} \Leftrightarrow \\ \tilde{f}_{jj} = f_{ii} &\bigvee p_{ij} = 0 \end{aligned}$$

Constraints on $\mathbf{P}$ can now be deduced: Each row and column of $\mathbf{P}$ must contain at least one non-zero element, since $\mathbf{P}$ is required to be invertible. Further limitation is achieved by assuming that $f_{ii} \neq f_{jj}$, for all $i \neq j$. Under this assumption, a row or column of $\mathbf{P}$ must not contain more than one non-zero element. Thus $\mathbf{P}$ is a permutation and scaling matrix. As a result, we can only identify the sources up to scaled versions that are ordered arbitrarily.

Furthermore, in agreement with [3], it was found that the sources need to have different autocorrelations.

Having proved that the possible *linear* transformations leave the parameters unique up to scaling and permutation, we still need to investigate the more general effect of a *non-linear* ambiguity. By considering the worst case scenario of monaural signal separation, where only one observation channel is available, a lower bound on the performance of the algorithm is obtained.

**Case-study: Monaural separation.**

Since the scaling and permutation ambiguities have already been addressed, the uniqueness of the parameters in the vicinity of the true parameters is investigated. We hope to find no local invariance, other than the one established between $\mathbf{A}$ and $\mathbf{Q}$. We first observe that the single-channel version of (2.13) simplifies as:

$$\begin{aligned} c_x(\tau) &= \mathbf{aQF}^{\tau}(\mathbf{I} - \mathbf{F}^2)^{-1}\mathbf{a}^T + \delta_\tau r \\ &= \sum_{i=1}^{d_s} u_i \frac{f_i^\tau}{1 - f_i^2} + \delta_\tau r \end{aligned} \tag{2.15}$$

where a reparametrization, $u_i = a_i^2 q_i$, was introduced.

The general solution to the local uniqueness question is answered by proving the local bilinearity of the function in 2.15, i.e. verifying that an observation of $c_x(\tau)$ maps to a number of discrete points in parameter-space. A pragmatic approach is taken here due to the obvious problem of inverting the given function. The function in question is defined:

$$g(\tau, \mathbf{u}, \mathbf{f}) \equiv \sum_{i=1}^{d_s} u_i \frac{f_i^\tau}{1 - f_i^2} + \delta_\tau r \tag{2.16}$$

In order to determine whether or not the seemingly large number of free parameters, $r$ and $u_i, f_i$ for $i = 1..d_s$, are identifiable from $c_x(\tau)$, a first order Taylor approximation of $g(\tau > 0, \mathbf{u}, \mathbf{f})$ is developed around the true parameters $\mathbf{u}^*$, $\mathbf{f}^*$:

$$\hat{g}(\tau, \mathbf{u}^* + \Delta\mathbf{u}, \mathbf{f}^* + \Delta\mathbf{f}) \equiv$$
$$g(\tau, \mathbf{u}^*, \mathbf{f}^*) + \sum_{i=1}^{L} [g_{u_i}(\tau, \mathbf{u}^*, \mathbf{f}^*)\Delta u_i + g_{f_i}(\tau, \mathbf{u}^*, \mathbf{f}^*)\Delta f_i]$$

where the partial derivatives involved (evaluated at $\tau, \mathbf{u}^*, \mathbf{f}^*$) are:

$$\begin{aligned}
g_{u_i}(\tau, \mathbf{u}^*, \mathbf{f}^*) &= \frac{f_i^{*\tau}}{1 - f_i^{*2}} \\
g_{f_i}(\tau, \mathbf{u}^*, \mathbf{f}^*) &= u_i^* \frac{f_i^{*\tau}(f_i^{*-1} + f_i^*)}{(1 - f_i^{*2})^2} \\
&= u_i^* \frac{(f_i^{*-1} + f_i^*)}{1 - f_i^{*2}} g_{u_i}(\tau, \mathbf{u}^*, \mathbf{f}^*) \qquad (2.17)
\end{aligned}$$

A linear relationship was found to exist between the partial derivatives. The first-order Taylor polynomial describes $c_x(\tau)$ in the vicinity of the true parameters. It will now be examined if other parametrizations of $\hat{g}(\cdot)$ than the true one $(\tau, \mathbf{u}^*, \mathbf{f}^*)$ yield the true correlation, $g(\tau, \mathbf{u}^*, \mathbf{f}^*)$. For that purpose, the approximation, $\hat{g}(\tau, \mathbf{u}^* + \Delta\mathbf{u}, \mathbf{f}^* + \Delta\mathbf{f})$, is set equal to $g(\tau, \mathbf{u}^*, \mathbf{f}^*)$ deducing a solution for $\Delta\mathbf{u}$ and $\Delta\mathbf{f}$:

$$\begin{aligned}
g(\tau, \mathbf{u}^*, \mathbf{f}^*) &= g(\tau, \mathbf{u}^*, \mathbf{f}^*) + \sum_{i=1}^{L} [g_{u_i}(\tau, \mathbf{u}^*, \mathbf{f}^*)\Delta u_i + g_{f_i}(\tau, \mathbf{u}^*, \mathbf{f}^*)\Delta f_i] \Leftrightarrow \\
0 &= \sum_{i=1}^{L} g_{u_i}(\tau, \mathbf{u}^*, \mathbf{f}^*)[\Delta u_i + u_i^* \frac{(f_i^{*-1} + f_i^*)}{1 - f_i^{*2}} \Delta f_i] \Leftrightarrow \\
0 &= \sum_{i=1}^{L} g_{u_i}(\tau, \mathbf{u}^*, \mathbf{f}^*)\Delta v_i \qquad (2.18)
\end{aligned}$$

where the following reparametrization was introduced based on equation 2.17:

$$\Delta v_i \equiv \Delta u_i + u_i^* \frac{(f_i^{*-1} + f_i^*)}{1 - f_i^{*2}} \Delta f_i$$

Having obtained a condition on the parameters in terms of $\Delta v_i$, a homogenous matrix equation is obtained by the aggregation of the scalar equations 2.18 corresponding to $\tau = 1..d_s$:

$$\mathbf{0} = \mathbf{J}\Delta\mathbf{v} \qquad (2.19)$$

where

$$\mathbf{J} = \begin{bmatrix}
\frac{1}{1-f_1^{*2}} & \frac{1}{1-f_2^{*2}} & \cdots & \frac{1}{1-f_L^{*2}} \\
\frac{f_1^*}{1-f_1^{*2}} & \frac{f_2^*}{1-f_2^{*2}} & \cdots & \frac{f_L^*}{1-f_l^{*2}} \\
\vdots & \vdots & & \vdots \\
\frac{f_1^{*L}}{1-f_1^{*2}} & \frac{f_2^{*L}}{1-f_2^{*2}} & \cdots & \frac{f_L^{*L}}{1-f_l^{*2}}
\end{bmatrix}$$

We are now left with determining the solution space of the matrix equation 2.19. Under the assumption that the sources have different autocorrelation, $f_i \neq f_j$

for all $i, j$, the columns of $\mathbf{J}$ are linearly independent. Therefore, $\mathbf{J}$ is a full rank matrix and the only solution is:

$$\Delta\mathbf{v} = \mathbf{0} \Leftrightarrow \Delta u_i = u_i^* \frac{(f_i^{*-1} + f_i^*)}{1 - f_i^{*2}} \Delta f_i, \forall i \tag{2.20}$$

The first-order Taylor approximation has thus revealed that the only solution ambiguity in the vicinity of the true parameters involves a linear scaling between $f_i$ and $u_i$. However, that remaining uncertainty can discarded by regarding the true $c_x(\tau)$, equation (2.15). Obviously, no linear scaling can exist between $f_i$ and $u_i$, and the proof is complete: the parameters are locally unique.

### 2.4.2 Convolutive mixing of AR(p) sources

It has been proven by e.g. Parra et al. [4] and others that blind source separation algorithms based on the second-order statistics of stationary mixtures do not solve the convolutive mixing problem. Any cost-function that is based on second-order statistics, or the equivalent power density spectrum, turns out to be invariant to an arbitrary filtering of the signal path filter, $\bar{\bar{A}}$. To see why, we compute the (true) power spectra and cross power spectra of $\mathbf{x}_t$:

$$\begin{aligned} \mathbf{\Gamma}_x(\omega) &= \mathbf{A}(\omega)\mathbf{\Lambda}_s(\omega)\mathbf{A}^H(\omega) \\ &= |\mathbf{A}(\omega)|^2\mathbf{\Lambda}_s(\omega) \end{aligned} \tag{2.21}$$

where the Wiener-Khintchine theorem was used to obtain the power spectrum from the autocorrelation function of a stationary signal. The observation noise was omitted for simplicity in the derivations and the notation. The cross power spectra of the sources vanish due to the independency assumption, hence $\mathbf{\Lambda}_s(\omega)$ is diagonal. In principle, an arbitrary (diagonal) filter matrix, $\mathbf{U}(\omega)$, and its inverse can now be inserted into equation 2.21:

$$\begin{aligned} \mathbf{\Gamma}_x(\omega) &= \mathbf{A}(\omega)\mathbf{U}^{-1}(\omega)\mathbf{U}(\omega)\mathbf{\Lambda}_s(\omega)\mathbf{U}^H(\omega)(\mathbf{U}^H(\omega))^{-1}\mathbf{A}^H(\omega) \\ &= |\tilde{\mathbf{A}}(\omega)|^2\tilde{\mathbf{\Lambda}}_s(\omega) \end{aligned}$$

with the definitions:

$$\begin{aligned} \tilde{\mathbf{A}}(\omega) &\equiv \mathbf{A}(\omega)\mathbf{U}^{-1}(\omega) \\ \tilde{\mathbf{\Lambda}}_s(\omega) &\equiv \mathbf{U}(\omega)\mathbf{\Lambda}_s(\omega)\mathbf{U}^H(\omega) \end{aligned}$$

The resulting undesired filtering, however, is restricted by the parametrization of $\mathbf{A}(\omega)$ and $\mathbf{\Lambda}_s(\omega)$. For instance, the source model usually has limited flexibility and cannot represent any arbitrary filtering.

Lucas Parra and Clay Spence [4] attempt to eliminate the remaining non-exclusiveness by explicitly assuming the non-stationarity of the source signals. By measuring the second-order statistics in $N$ segments where the signals are assumed short-term stationary, and at the same time assuming the signal channels, $\bar{\bar{A}}$, to be long-term stationary, an additional number of conditions are introduced that have the potential to constrain the solution enough to be unique, except for the usual scaling and permutation.

A small constructed example illustrates the principle: let the sources be broadband noise with a single spectral peak that moves around between segments of signal. The peaks do not overlap between segments. We restrict the source model to an AR(2) process. As a result, at most 1 spectral peak can be represented per segment.

In the first segment, the peak can be modelled by either $\tilde{\mathbf{A}}(\omega)$, $\tilde{\mathbf{\Lambda}}_s(\omega)$ or a combination. If the peak was modelled by $\tilde{\mathbf{A}}(\omega)$ we would mistakenly estimate $\mathbf{s}_t$ as white noise.

If we add a second segment with a different peak location, the point of the non-stationarity requirement becomes clear: $\tilde{\mathbf{A}}(\omega)$ cannot model any of the spectral peaks, because it would pollute the source estimate of the other segment, adding an extra peak to the spectrum. If a more flexible model was selected for the source, the unwanted peak could be cancelled by notch filtering the original source. In conclusion:

> Provided the source model is sufficiently restricted, the varying features of the source spectra can only be represented by the sources themselves, not by the channel filters. As a result, convolutive demixing based on second order statistics is aided by a measure of spectral dynamism.

Speech signals belong to a class of signals that undoubtedly possess a time-varying spectrum. Moreover, speech is often modelled by rather constrained all-pole models. These features suggest that convolutive mixtures of speech may be separated. All-pole models for speech are treated in section 2.5.

### 2.4.3 Different permutations over frequency

A common trait of a number of algorithms, e.g. [9], [8], [4] and [7], is the transformation of the time-domain convolutive problem to an equivalent frequency-domain instantaneous problem.

A frequency domain representation of the convolutive problem was obtained in equation 2.21. In principle, the diagonal represents an independent instantaneous problem for each frequency, $\omega$. For each of those problems, the scaling and the source ordering are in general not the same.

However, as stated by [4], the representation of the source spectra/autocorrelation functions involved are parameterized by relatively few parameters. Thus, the source spectra are limited to smooth envelopes, which in turn prevent the switching assignment of sources to estimates. An experiment documented in chapter 4 demonstrates the effect of undesired permutations over frequency.

## 2.5 All-pole models for speech

A popular speech model will be reviewed below as well as its connection to the proposed Kalman filter model. As a special case, the representational capability and the limitations of an AR(2) model will be treated due to its analytical tractability.

In many practical applications, e.g. linear predictive coding (LPC), speech signals are treated under the implicit assumption of non-stationarity. It is clear, however, that speech contains structure beyond the short-term spectrum , e.g. syllables, words and sentences. Hidden Markov models (HMM) are often used to describe the transitions between the different 'states' of speech, see, e.g., Rabiner's speech recognizer [17]. Therefore, the conventional assumption of the non-stationarity of speech is a crude approximation. For instance, in a high-level interpretation of speech generation, the 'parameters' of a human being are probably stationary during the uttering of a sentence. Hence, in a general sense, speech is stationary for the duration of the sentence. We are in the business of separating signals, however, and might be satisfied with a partial speech model that is descriptive enough for our purposes.

While the physical examination of the human voice apparatus obviously is fundamental to speech models, the details of the flesh and blood remain largely irrelevant to our discussion and will be used only superficially for stating the model. The details in this regard can be found in [18]. Before advancing to the formulation of the model, we define a *phoneme* as the smallest unit of speech that conveys meaning, e.g. vowels and consonants. The speech signal if often regarded as stationary on the time scale of a phoneme. Vowels, which claim a special interest in the literature due to their importance in the perception of speech, typically last from 40ms to 400ms, see [18] p. 115.

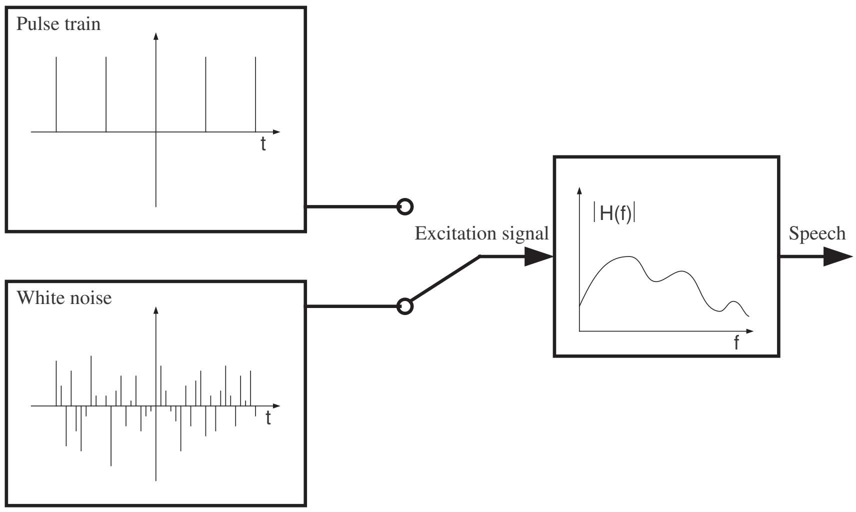


Figure 2.4: A popular speech model: An excitation signal that can either be a a periodic broad band signal or a white noise signal is filtered by the vocal tract filter. Multiple configurations of the filter are possible, leading to a large number of differently sounding phonemes.

The most popular model resulting from the analysis of speech is shown in figure 2.4. Briefly stated, air from the lungs help the glottis produce an excitation signal that is filtered by the vocal tract. A wide variety of phonemes can be produced by varying the type of excitation signal and the spectral envelope of the filter. Specifically, the glottis can produce two different excitation signals corresponding to the two basic classes of phonemes, *voiced* and *unvoiced* phonemes:

Broadly speaking, a voiced phoneme is produced by the glottis emitting a broad band periodic signal, e.g. a pulse train. The vocal tract filter shapes the excitation signal by amplifying and attenuating the signal in certain frequency bands. The so-called *formants* define the spectral peaks of this filter and almost uniquely characterize the different vowels, which can be verified by plotting the location of the two most prominent peaks in a formants plot, see figure 2.5.

Unvoiced phonemes, on the other hand, result from the glottis emitting a white noise signal. Again, the vocal tract shapes the excitation signal in various ways in order to produce a wide array of sounds. The consonants of the English language belong to this class of phonemes.

Although not suggested by human anatomy, the vocal tract filter is often modelled by all-pole filters. The main reason is that efficient estimation is available for such a transfer function, e.g. the solution of the normal equations. The Kalman filter model employs an AR(p) random process as a model for the sources. In agrement with the above formulation of speech, the transfer function of this model is all-pole, which is seen by z-transforming the time-domain source signal representation:

$$\begin{aligned} s_t &= f_1 s_{t-1} + f_2 s_{t-2} + \ldots + s_{t-p} + v_t \Leftrightarrow \\ H(z) &= \frac{S(z)}{V(z)} = \frac{1}{1 - f_1 z^{-1} - f_2 z^{-2} - \ldots - f_p z^{-p}} \end{aligned} \tag{2.22}$$

However, the source excitation signal is white Gaussian noise, which is only partially in accordance with the above speech model: the periodic excitation is *not* represented in KaBSS. As a consequence, some model bias must be expected. In his PhD thesis [19], Preben Kidmose advocates the use of long-tailed excitation noise distributions and demonstrates their superiority. Such models, however, are not easily implemented in KaBBS. A special case of the AR(p) model is

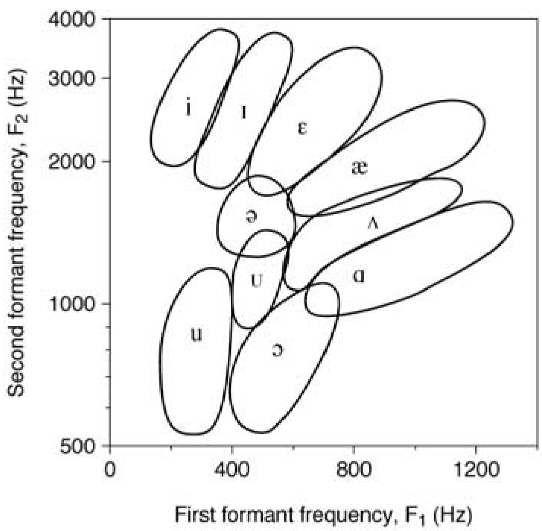


Figure 2.5: Mean location of the two first formants of American English vowels. The international phonetic alphabet (IPA) identify the phonemes. Great variation occurs between speakers. Reprinted from a tutorial of the National Center for Voice and Speech [20].

now reviewed:

### 2.5.1 The AR(2) model.

Based loosely on the deliberations of [21], the analysis sets out by specializing the transfer function of equation 2.22 to the case $p = 2$:

$$H_2(z) = \frac{S_2(z)}{V(z)} = \frac{1}{1 - f_1 z^{-1} - f_2 z^{-2}} \tag{2.23}$$

The rational transfer function is rewritten in terms of only positive powers of $z$, and subsequently factored:

$$\begin{aligned} H_2(z) &= \frac{z^2}{z^2 - f_1 z^1 - f_2} \\ &= \frac{z^2}{(z - z_1)(z - z_2)} \end{aligned}$$

Given the assumption that the coefficients, $f_i$, are real, the poles must be *complex-conjugate*, $z_1 = z_2^*$, or real. In the case of complex-conjugate poles the above can be expressed as:

$$H_2(z) = \frac{z^2}{(z - r\exp[-j\omega_0])(z - r\exp[j\omega_0])} \tag{2.24}$$

The spectral focus of $H_2(z)$ is concentrated at $\omega_0$. It is known from linear time-invariant (LTI) systems theory that in order for the system to be stable, the poles must reside inside the unit circle, i.e. $r < 1$. The filter gets more peaked as $r$ approaches 1.

In the case of real poles, the amplification of the filter is situated at either $\omega = 0$ or $\omega = \pi$, or both. Figure 2.6 displays the possible spectra.

In order to generate AR(2) random processes with a specific peakedness and pole placement, a white Gaussian noise signal is then filtered through $H_2(z)$. The coefficients of the filter are identified from the expansion of equation 2.24:

$$\begin{aligned} H_2(z) &= \frac{z^2}{z^2 - zr(\exp[-j\omega_0] + \exp[-j\omega_0]) + r^2} \\ &= \frac{z^2}{z^2 - z2r\cos(\omega_0) + r^2} \end{aligned}$$

From the above, it is deduced that $f_1 = 2r\cos(\omega_0)$ and $f_2 = -r^2$.

The incorporation of the simple AR(2) model in KaBBS serves the purpose of constraining the source power spectrum to a degree, where permutation over frequency and model ambiguities vanish. On the other hand, only a single spectral peak can be represented. In a vowel-modelling scenario, all three formants would be described by the same filter pole.

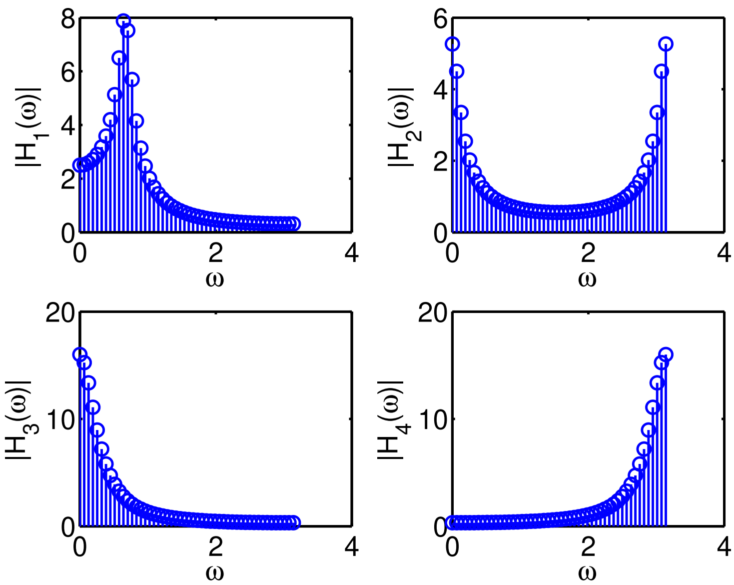


Figure 2.6: Power spectra of the four different types of AR(2) processes. The poles are either complex-conjugate (top-left) , real and distinct (top-right), or real and equal (bottom).

# Chapter 3

# Learning

In Blind Source Separation two connected learning tasks must be addressed based on the observed mixture. One is to recover the unknown source signals, or hidden variables. The second is to estimate the parameters of the mixing process, e.g. the mixing matrix $\mathbf{A}$ of an instantaneous mix or the matrix of linear filters, $\bar{\bar{\mathbf{A}}}$, associated with a convolutive mixing. There is a fine distinction between parameters and hidden variables. Whereas the uncertainty of the former decreases with observed data size, the uncertainty of the sources does not. The number of source samples scale with the number of observed, noisy samples. We term the learning of parameters and hidden variables, *estimation* and *inference*, respectively.

In the preceding chapter, the BSS problem was formulated within a very general class of Gaussian linear models. Therefore, existing algorithms need only to be tailored to the special constrained form commanded by the source independency assumption. The following review is based on [11] and [22]. The derivations lead to the estimators and Kalman smoother which define KaBSS.

A standard method for learning in models with hidden variables is the Expectation Maximization (EM) algorithm. It comprises two basic steps that alternately infers the hidden variables and estimates the parameters while keeping the other fixed. At no iteration of the algorithm can the likelihood decrease. EM is an iterative scheme for maximum-likelihood (ML) parameter estimation, in which the task is to find the parameters that make the model most likely given the observed data.

The simple one-step ML estimation cannot be performed. In abstract form, the likelihood function of the parameters, $\theta$, given the observed signal and the assumed data model $p(\mathbf{X}, \mathbf{S}|\theta)$ appears as:

$$\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) = \log \int d\mathbf{S} p(\mathbf{X}, \mathbf{S}|\theta) \tag{3.1}$$

$$\tag{3.2}$$

where $\mathcal{L}(\theta)$ resulted from the marginalization of the sources. Maximization wrt. to $\theta$ is hard, because the integral is not available in closed form.

Instead, local lower bounds of equation 3.1 are optimized iteratively, guaranteing at each step that $\mathcal{L}(\theta)$ cannot decrease. The below derivations will justify this rationale. We start by taking the logarithm and applying Jensen's inequality, which places a lower bound on a convex function (see [23]):

$$\begin{aligned}
\mathcal{L}(\theta) &= \log \int d\mathbf{S} p(\mathbf{X}, \mathbf{S}|\theta) \\
&= \log \int d\mathbf{S} \hat{p}(\mathbf{S}) \frac{p(\mathbf{X}, \mathbf{S}|\theta)}{\hat{p}(\mathbf{S})} \\
&\geq \mathcal{F}(\hat{p}, \theta)
\end{aligned}$$

where the lower bound function, also known from physics as the negative free energy, is defined as

$$\mathcal{F}(\hat{p}, \theta) \equiv \int d\mathbf{S} \hat{p}(\mathbf{S}) \log \frac{p(\mathbf{X}, \mathbf{S}|\theta)}{\hat{p}(\mathbf{S})} \tag{3.3}$$

The above is true for any choice of pdf for $\hat{p}(\cdot)$. Two terms result from the expansion of equation (3.3) of which only the first depends on the parameters:

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

The two steps of EM can now be defined in terms of the E-step maximization of $\mathcal{F}(\hat{p}, \theta)$ wrt. $q(\cdot)$ and the M-step maximization of $\mathcal{J}(\hat{p}, \theta)$ wrt. $\theta$. It is now argued that these two operations in combination never decrease $\mathcal{L}(\hat{p}, \theta)$:

The M-step optimizes $\mathcal{J}(\hat{p}, \theta)$ (wrt. $\theta$), which is a term of $\mathcal{F}(\hat{p}, \theta)$, not affecting the other term, $\mathcal{R}(\hat{p})$. Therefore the M-step optimizes $\mathcal{F}(\hat{p}, \theta)$ wrt. $\theta$. Subsequently, the E-step maximizes $\mathcal{F}(\hat{p}, \theta)$ wrt. $\hat{p}(\cdot)$ by setting $q(\mathbf{S}) = p(\mathbf{S}|\mathbf{X}, \theta)$, as will be proven shortly. Conveniently, this maximum coincides with $\mathcal{F}(\hat{p}, \theta)$ attaining equality with $\mathcal{L}(\theta)$. Combining the facts: the $\mathcal{F}(\hat{p}, \theta)$ is equal to $\mathcal{L}(\theta)$ before the M-step and then optimized wrt $\theta$. As a result, $\mathcal{L}(\theta)$ cannot decrease, since $\mathcal{F}(\hat{p}, \theta)$ bounds it from below.

By inserting into equation 3.3, it is easily proven that the lower bound touches $\mathcal{L}(\theta)$ for $\hat{p}(\mathbf{S}) = p(\mathbf{S}|\mathbf{X}, \theta)$:

$$\begin{aligned}
\mathcal{F}(\hat{p}, \theta)_{|\hat{p}=p(\mathbf{S}|\mathbf{X},\theta)} &= \int d\mathbf{S} p(\mathbf{S}|\mathbf{X}, \theta) \log \frac{p(\mathbf{X}, \mathbf{S}|\theta)}{p(\mathbf{S}|\mathbf{X}, \theta)} \\
&= \int d\mathbf{S} p(\mathbf{S}|\mathbf{X}, \theta) \log p(\mathbf{X}|\theta) \\
&= \log p(\mathbf{X}|\theta) \\
&= \mathcal{L}(\theta)
\end{aligned}$$

The following two sections will address the E and M steps, respectively.

## 3.1 E-step

It was established that the E-step should obtain the source-posterior, $p(\mathbf{S}|\mathbf{X}, \theta)$. Otherwise, the optimization of $\mathcal{F}(\hat{p}, \theta)$ would not be equivalent to the optimization of $\mathcal{L}(\theta)$. As a result of the particular choice of model, i.e. the Kalman filter model, the *Kalman smoother* will provide the desired posterior distribution. Nothing new is added to the Kalman smoother, which is a standard method of the literature, in this work and only the main traits will be repeated here. The basic anatomy of the Kalman smoother comprises two basic elements: the *forward* recursion, which is equivalent to the popular Kalman filter, and the *backward* recursion.

The distinction between filter and smoother lies in the limited observations available to the filter, which can only condition on past and present observations. The smoother conditions on *all* observations:

$$p(\mathbf{s}_t|\mathbf{x}^\tau) \quad \forall t \leq \tau$$

where the notation $\mathbf{x}^\tau = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_\tau\}$ was used. In this section no parameters or variables will be marked with bars, since the general Kalman smoother is used. Furthermore, the segment index, $n$, is omitted for notational simplicity. Sections 3.1.1 and 3.1.2 describe the forward and backward recursion, respectively.

### 3.1.1 The forward recursion

The Kalman filter equations that make up the first recursion fulfill two tasks. The first and most important is to obtain the filtered sources, which are required for the subsequent backward pass and the second one is the efficient computation of the log-likelihood.

The filter equations outputs the moments of the source-posterior conditioned on past and present observations:

$$p(\mathbf{s}_t|\mathbf{x}^t) \quad \forall t \leq \tau$$

Again we used the notation $\mathbf{x}^t = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_t\}$. Henceforth, superscripts have this meaning (except for the segment index, $n$). An efficient recursive sequence of computation is obtained by applying Bayes' theorem to the sought distribution:

$$p(\mathbf{s}_t|\mathbf{x}^t) = \frac{p(\mathbf{x}_t|\mathbf{s}_t)p(\mathbf{s}_t|\mathbf{x}^{t-1})}{p(\mathbf{x}_t|\mathbf{x}^{t-1})} \tag{3.4}$$

Since all variables involved are (jointly) normally distributed, second-order statistics, i.e mean and covariance, are sufficiently describing the component distributions. Distributions $p(\mathbf{s}_t|\mathbf{x}^{t-1})$ and $p(\mathbf{x}_t|\mathbf{x}^{t-1})$ are available recursively, and $p(\mathbf{x}_t|\mathbf{s}_t)$ is the observation model. The forward recursion therefore simplifies to recursive operations on the first and second-order moments of the component distributions. The resulting update equations for the Kalman filter are:

$$\begin{aligned}
\hat{\mathbf{s}}_t^{t-1} &= \mathbf{F}\hat{\mathbf{s}}_{t-1}^{t-1} && (3.5)\\
\mathbf{P}_t^{t-1} &= \mathbf{F}\mathbf{P}_{t-1}^{t-1}\mathbf{F}^T + \mathbf{Q} && (3.6)\\
\mathbf{K}_t &= \mathbf{P}_t^{t-1}\mathbf{A}^T(\mathbf{R} + \mathbf{A}\mathbf{P}_t^{t-1}\mathbf{A}^T)^{-1} && (3.7)\\
\mathbf{P}_t^t &= (\mathbf{I} - \mathbf{K}_t\mathbf{A})\mathbf{P}_t^{t-1} && (3.8)\\
\hat{\mathbf{s}}_t^t &= \hat{\mathbf{s}}_t^{t-1} + \mathbf{K}_t(\mathbf{x}_t - \mathbf{A}\hat{\mathbf{s}}_t^{t-1}) && (3.9)
\end{aligned}$$

The recursion was initialized by $\hat{\mathbf{s}}_1^0 = \mu$ and $\mathbf{P}_1^0 = \mathbf{\Sigma}$. Equations 3.5 and 3.6 update the moments of the one-step source predictor. Equation 3.7 updates the

Kalman gain, which weighs prediction against observation. Equations 3.8 and 3.9 update the moments of the source posterior, by including one additional sample at a time.

The computation of the log-likelihood, $\mathcal{L}(\theta)$, can be embedded in the forward recursion. This is seen by decomposing $\mathcal{L}(\theta)$:

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{n=1}^{N} \log[p(\mathbf{x}_n^\tau|\theta)] \\ &= \sum_{n=1}^{N}\sum_{t=2}^{\tau} \log[p(\mathbf{x}_{t,n}|\mathbf{x}_n^{t-1},\theta)] + \sum_{n=1}^{N} \log[p(\mathbf{x}_{1,n}|\theta)]\end{aligned}$$

It is clear that the observation predictor distributions, $p(\mathbf{x}_{t,n}|\mathbf{x}_n^{t-1},\theta)$, are related to the source predictor moments of update equations 3.5 and 3.6. In fact, the moments of $p(\mathbf{x}_{t,n}|\mathbf{x}_n^{t-1},\theta)$ can be computed directly from the source predictor moments:

$$\begin{aligned}\hat{\mathbf{x}}_t^{t-1} &= \mathbf{A}\hat{\mathbf{s}}_t^{t-1} \\ \mathbf{H}_t^{t-1} &= \mathbf{R} + \mathbf{A}\mathbf{P}_t^{t-1}\mathbf{A}^T\end{aligned}$$

The moments of $p(\mathbf{x}_{1,n}|\theta)$ are:

$$\begin{aligned}\hat{\mathbf{x}}_1^0 &= \mathbf{A}\mu \\ \mathbf{H}_1^0 &= \mathbf{R} + \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\end{aligned}$$

The computed moments are passed on to the backward recursion:

### 3.1.2 The backward recursion

Although the derivation of the backward recursion is based on identical considerations to those that led to the forward recursion, it is slightly more analytically intensive. For details see e.g. [22] or the original paper by Rauch. The backward recursion consists of equations 3.10-3.12:

$$\mathbf{J}_{t-1} = \mathbf{P}_{t-1}^{t-1}\mathbf{F}^T[\mathbf{P}_t^{t-1}]^{-1} \tag{3.10}$$

$$\mathbf{P}_{t-1}^\tau = \mathbf{P}_{t-1}^{t-1} + \mathbf{J}_{t-1}[\mathbf{P}_t^\tau - \mathbf{P}_t^{t-1}]\mathbf{J}_{t-1}^T \tag{3.11}$$

$$\hat{\mathbf{s}}_{t-1}^\tau = \hat{\mathbf{s}}_{t-1}^{t-1} + \mathbf{J}_{t-1}[\hat{\mathbf{s}}_t^\tau - \hat{\mathbf{s}}_t^{t-1}] \tag{3.12}$$

where the recursion is initialized by $\hat{\mathbf{s}}_\tau^\tau$ and $\mathbf{P}_\tau^\tau$ of the forward recursion. Additionally, the *lag-one covariance* is required for the M-step:

$$\mathbf{P}_{t-1,t-2}^\tau = \mathbf{P}_{t-1}^{t-1}\mathbf{J}_{t-2}^T + \mathbf{J}_{t-1}[\mathbf{P}_{t,t-1}^\tau - \mathbf{F}\mathbf{P}_{t-1}^{t-1}]\mathbf{J}_{t-2}^T$$

where the recursion is initialized as

$$\mathbf{P}_{\tau,\tau-1}^\tau = [\mathbf{I} - \mathbf{K}_\tau\mathbf{A}]\mathbf{F}\mathbf{P}_{\tau-1}^{\tau-1}$$

Before advancing to the actual parameter estimation of section 3.2, the source posterior moments required for the estimation are reviewed. The segment index, $n$, is reintroduced for forward compatibility.

**Source posterior moments**

The direct outputs of the Kalman smoother are $\hat{\bar{\mathbf{s}}}_t^{\tau,n}$, $\mathbf{P}_t^{\tau,n}$ and $\mathbf{P}_{t,t-1}^{\tau,n}$, i.e. the mean, covariance and lag-one covariance of the source posterior. It is assumed in the following that the posterior mean is conditioned on the whole segment. Therefore, we can write $\hat{\bar{\mathbf{s}}}_t^n \equiv \hat{\bar{\mathbf{s}}}_t^{\tau,n}$. Autocorrelation functions are obtained from the covariances of $p(\bar{\mathbf{s}}^{\tau,n}|\bar{\mathbf{x}}^{\tau,n},\theta)$:

$$\begin{aligned}\bar{\mathbf{M}}_t^n &= \mathbf{P}_t^{\tau,n} + \hat{\bar{\mathbf{s}}}_t^n(\hat{\bar{\mathbf{s}}}_t^n)^T \\ \bar{\mathbf{M}}_t^{1,n} &= \mathbf{P}_{t,t-1}^{\tau,n} + \hat{\bar{\mathbf{s}}}_t^n(\hat{\bar{\mathbf{s}}}_{t-1}^n)^T\end{aligned}$$

The special constrained format of the model parameters, $\theta$, is reflected in the moments:

$$\begin{aligned}\hat{\bar{\mathbf{s}}}_t^n &= \begin{bmatrix} (\hat{\mathbf{s}}_{1,t}^n)^T & (\hat{\mathbf{s}}_{2,t}^n)^T & \dots & (\hat{\mathbf{s}}_{d_s,t}^n)^T \end{bmatrix}^T \\ \bar{\mathbf{M}}_t^n &= \begin{bmatrix} \mathbf{M}_{1,t}^n & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{2,t}^n & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M}_{d_s,t}^n \end{bmatrix}\end{aligned}$$

where $\hat{\mathbf{s}}_{i,t}^n$ and $\mathbf{M}_{i,t}^n$ are the posterior mean and autocorrelation of source $i$. Individual source moments are given by:

$$\begin{aligned}\hat{\mathbf{s}}_{i,t}^n &\equiv \langle \mathbf{s}_{i,t}^n \rangle \\ \mathbf{M}_{i,t}^n &\equiv \langle \mathbf{s}_{i,t}^{\tau,n}(\mathbf{s}_{i,t}^{\tau,n})^T \rangle \\ &\equiv \begin{bmatrix} \mathbf{m}_{i,1,t}^n & \mathbf{m}_{i,2,t}^n & .. & \mathbf{m}_{i,L,t}^n \end{bmatrix}^T\end{aligned}$$

The structure of the stacked lag-one covariance is similar to that of $\bar{\mathbf{M}}_t^n$. The time-lagged autocorrelation for source $i$ is:

$$\mathbf{M}_{i,t}^{1,n} \equiv \langle \mathbf{s}_{i,t}^{\tau,n}(\mathbf{s}_{i,t-1}^{\tau,n})^T \rangle$$

For the derivation of the parameter estimators of the *instantaneous* mixing AR(p) model, a memoryless, sparse source signal vector is introduced in order to ease notation:

$$\ddot{\mathbf{s}}_t^n = \begin{bmatrix} s_{1,t}^n & s_{2,t}^n & .. & s_{d_s,t}^n \end{bmatrix}^T \tag{3.13}$$

The corresponding sparse moments are:

$$\begin{aligned}\hat{\ddot{\mathbf{s}}}_t^n &= \langle \ddot{\mathbf{s}}_t^n \rangle \\ \ddot{\mathbf{M}}_t^n &= \langle \ddot{\mathbf{s}}_t^n(\ddot{\mathbf{s}}_t^n)^T \rangle\end{aligned}$$

They are obtained from the stacked moments by indexing the relevant matrix elements.

For the special case of instantaneous mixing of AR(1) processes, the bars on the moment symbols are omitted.

## 3.2 M-step

The purpose of the M-step is to maximize the lower bound function, $\mathcal{F}(\hat{p},\theta)$, with respect to $\theta$. It was proven in the introduction to this chapter that the optimization of the lower bound function was equivalent to moving $\mathcal{L}(\theta)$ in a non-decreasing direction. The only relevant ($\theta$-dependant) term of $\mathcal{L}(\theta)$, i.e. $\mathcal{J}(\hat{p},\theta)$ was given the following definition:

$$\mathcal{J}(\hat{p},\theta) \quad \equiv \quad \int d\mathbf{S} q(\mathbf{S}) \log p(\mathbf{X},\mathbf{S}|\theta)$$

The next step is to express $\log p(\mathbf{X},\mathbf{S}|\theta)$ in terms of the parameters. Before taking the logarithm, we factor the Kalman filter model using the definitions of joint and conditional probability,

$$\begin{aligned} p(\mathbf{S},\mathbf{X}|\theta) \quad &= \quad \prod_{n=1}^{N} \{ \prod_{i=1}^{d_s} p(\mathbf{s}_{i,1}^n | \mu_i^n, \mathbf{\Sigma}_i^n) \} \\ &\times \quad \{ \prod_{t=2}^{\tau} \prod_{i=1}^{d_s} p(\mathbf{s}_{i,t}^n | \mathbf{s}_{i,t-1}^n, \mathbf{F}_i^n, \mathbf{Q}_i^n) \} \\ &\times \quad \{ \prod_{t=1}^{\tau} p(\mathbf{x}_t^n | \bar{\mathbf{s}}_t^n, \bar{\bar{\mathbf{A}}}, \mathbf{R}) \} \end{aligned}$$

The term distributions describe the white noise of the source and observation model. The natural logarithm is applied to the above:

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

Finally the log-model is averaged over the source posterior, $p(\mathbf{S},\mathbf{X}|\theta)$:

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

A utility function has now been obtained that can immediately be optimized wrt. the parameters $\bar{\bar{\mathbf{A}}}$, $\mathbf{R}$, $\bar{\mathbf{\Sigma}}$, $\bar{\mu}$, $\bar{\mathbf{F}}$ and $\bar{\mathbf{Q}}$. The derivations are in complete analogy with [22] except that only partial parameter matrices are estimated, due to the source independency assumption. Fortunately, it is analytically possible in all cases to 1) compute the partial derivatives of the cost function wrt. the parameters, 2) equate them to zero and 3) solve for the ML parameter estimators. The estimators will be derived in the most general case in section 3.2.1, i.e. for convolutive mixings. Subsequently, they are specialized to instantaneous mixings, to cases where $p < L$ and to instantaneous mixings of AR(1) processes in sections 3.2.2, 3.2.3 and 3.2.4, respectively.

### 3.2.1 Estimators

A number of standard derivatives from matrix calculus are used in the following. The following, which is due to Roweis [24], is valid for matrices with functionally independent elements:

$$|\mathbf{Y}| = \frac{1}{|\mathbf{Y}^{-1}|} \tag{3.14}$$

$$\frac{\partial \log |\mathbf{Y}|}{\partial \mathbf{Y}} = (\mathbf{Y}^{-1})^T \tag{3.15}$$

$$\frac{\partial \mathbf{a}^T\mathbf{y}}{\partial \mathbf{y}} = \frac{\partial \mathbf{y}^T\mathbf{a}}{\partial \mathbf{y}} = \mathbf{a} \tag{3.16}$$

$$\frac{\partial \mathbf{a}^T\mathbf{Y}\mathbf{b}}{\partial \mathbf{Y}} = \mathbf{a}\mathbf{b}^T \tag{3.17}$$

$$\frac{\partial \mathbf{a}^T\mathbf{Y}^T\mathbf{b}}{\partial \mathbf{Y}} = \mathbf{b}\mathbf{a}^T \tag{3.18}$$

$$\frac{\partial \mathbf{a}^T\mathbf{Y}^T\mathbf{C}\mathbf{Y}\mathbf{b}}{\partial \mathbf{Y}} = \mathbf{C}^T\mathbf{Y}\mathbf{a}\mathbf{b}^T + \mathbf{C}\mathbf{Y}\mathbf{b}\mathbf{a}^T \tag{3.19}$$

Below follow the tedious derivations, which are pivotal to this work:

$\mu$

We start out by deriving the estimator for $\mu_i$, the initial mean of source $i$ in segment $n$, using rule 3.16:

$$\frac{\partial \mathcal{J}(\theta, \hat{p})}{\partial \mu_i^n} = -\frac{1}{2}[2(\mathbf{\Sigma}_i^n)^{-1}\mu_i^n - (\mathbf{\Sigma}_i^n)^{-1}\langle \mathbf{s}_{i,1}^n \rangle]$$

In order to locate the maximum wrt. $\mu_i$, the partial derivative is equated to zero:

$$\mu_{i,\mathbf{new}}^n = \hat{\mathbf{s}}_{i,1}^n$$

$\mathbf{\Sigma}$

The next estimator to derive is the corresponding covariance, $\mathbf{\Sigma}_{i,\mathbf{new}}^n$. The partial derivative is computed:

$$\begin{aligned}\frac{\partial \mathcal{J}(\theta, \hat{p})}{\partial (\mathbf{\Sigma}_i^n)^{-1}} &= -\frac{1}{2}\frac{\partial}{\partial (\mathbf{\Sigma}_i^n)^{-1}}[-\log\det(\mathbf{\Sigma}_i^n)^{-1} + \langle (\mathbf{s}_{i,1}^n - \mu_i^n)^T(\mathbf{\Sigma}_i^n)^{-1}(\mathbf{s}_{i,1}^n - \mu_i^n)\rangle] \\ &= \frac{1}{2}\mathbf{\Sigma}_i^n - \frac{1}{2}\langle (\mathbf{s}_{i,1}^n - \mu_i^n)(\mathbf{s}_{i,1}^n - \mu_i^n)^T\rangle\end{aligned}$$

The rules 3.14 and 3.15 were applied. We solve for the estimator:

$$\begin{aligned}
\mathbf{\Sigma}^n_{i,\mathbf{new}} &= \mathbf{M}^n_{i,1} + (\mu^n_{i,\mathbf{new}})^T \mu^n_{i,\mathbf{new}} - 2(\mu^n_{i,\mathbf{new}})^T \mu^n_{i,\mathbf{new}} \\
&= \mathbf{M}^n_{i,1} - (\mu^n_{i,\mathbf{new}})^T \mu^n_{i,\mathbf{new}}
\end{aligned}$$

**f**

The estimator for the model parameters of source $i$ in segment $n$ is now derived: The partial derivative wrt. $\mathbf{f}_i$ is computed using the chain rule:

$$\frac{\partial \mathcal{J}(\theta,\hat{p})}{\partial \mathbf{f}^n_i} = -\frac{1}{q^n_i}\sum_{t=2}^{\tau}\langle \mathbf{s}^n_{i,t-1}(s^n_{i,t} - (\mathbf{f}^n_i)^T(\mathbf{s}^n_{i,t-1}))\rangle$$

The estimator is solved for by equating to zero:

$$\begin{aligned}
\mathbf{f}^n_{\mathbf{new},i} &= [\sum_{t=2}^{\tau}\langle \mathbf{s}^n_{i,t-1}(\mathbf{s}^n_{i,t-1})^T\rangle]^{-1}\sum_{t=2}^{\tau}\langle \mathbf{s}^n_{i,t-1}s^n_{i,t}\rangle \\
&= [\sum_{t=1}^{\tau-1}\mathbf{M}^n_{i,t}]^{-1}\sum_{t=2}^{\tau}\mathbf{m}^n_{i,1,t,t-1}
\end{aligned}$$

**q**

The partial derivative wrt. $q^n_i$ is computed:

$$\begin{aligned}
\frac{\partial \mathcal{J}(\theta,\hat{p})}{\partial q^n_i} &= -\frac{(\tau-1)}{2q^n_i} + \frac{1}{2(q^n_i)^2}\sum_{t=2}^{\tau}\langle (s^n_{i,t} - (\mathbf{f}^n_i)^T\mathbf{s}^n_{i,t-1})^2\rangle \\
&= -\frac{(\tau-1)}{2q^n_i} + \frac{1}{2(q^n_i)^2}\Big[\sum_{t=2}^{\tau}\langle (s^n_{i,t})^2\rangle + (\mathbf{f}^n_i)^T[\sum_{t=2}^{\tau}\langle \mathbf{s}^n_{i,t-1}(\mathbf{s}^n_{i,t-1})^T\rangle]\mathbf{f}^n_i \\
&+ \sum_{t=2}^{\tau}[-2(\mathbf{f}^n_i)^T\langle s^n_{i,t}\mathbf{s}^n_{i,t-1}\rangle]\Big]
\end{aligned}$$

The estimator emerges from equating the partial derivative to zero and setting $\mathbf{f}^n_i = \mathbf{f}^n_{i,\mathbf{new}}$, since we are only interested in the solution that simultaneously minimize $\mathcal{J}(\theta,\hat{p})$ wrt. $\mathbf{f}^n_i$:

$$\begin{aligned}
\frac{1}{2(q^n_{i,\mathbf{new}})^2}\sum_{t=2}^{\tau}\langle (s^n_{i,t})^2\rangle &= \frac{(\tau-1)}{2q^n_{i,\mathbf{new}}} + \frac{1}{2(q^n_{i,\mathbf{new}})^2}\sum_{t=2}^{\tau}(\mathbf{f}^n_i)^T\langle s^n_{i,t}\mathbf{s}^n_{i,t-1}\rangle \Leftrightarrow \\
q^n_{i,\mathbf{new}} &= \frac{1}{\tau-1}\sum_{t=2}^{\tau}[(\mathbf{M}^n_{i,t})_{11} - (\mathbf{f}^n_{i,\mathbf{new}})^T\mathbf{m}^n_{i,1,t,t-1}]
\end{aligned}$$

**A**

The estimators for the observation model, $\bar{\bar{\mathbf{A}}}$ and $\mathbf{R}$, are identical to those of the literature on parameter estimation in Gaussian linear models, see e.g. [22]. They will, however, be derived below. The partial derivative of $\mathcal{J}(\theta,\hat{p})$ wrt. $\bar{\bar{\mathbf{A}}}$ is:

$$\frac{\partial \mathcal{J}}{\partial \bar{\bar{\mathbf{A}}}} = -\frac{1}{2}\sum_{n=1}^{N}[2\mathbf{R}^{-1}\bar{\bar{\mathbf{A}}}\sum_{t=1}^{\tau}\langle (\bar{\mathbf{s}}^n_t)^T\bar{\mathbf{s}}^n_t\rangle - 2\mathbf{R}^{-1}\sum_{t=1}^{\tau}[\mathbf{x}^n_t\langle (\bar{\mathbf{s}}^n_t)^T\rangle]]$$

where the rules 3.17, 3.18 and 3.19 were used. The partial derivative is equated to zero and the estimator is solved for:

$$\mathbf{A_{new}} = [\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathbf{x}_t^n\hat{\bar{\mathbf{s}}}_t^T][\sum_{n=1}^{N}\sum_{t=1}^{\tau}\bar{\mathbf{M}}_t^n]^{-1} \tag{3.20}$$

**R**

Finally, the partial derivative wrt. $\mathbf{R}^{-1}$ is computed under the diagonality assumption on $\mathbf{R}$:

$$\begin{aligned}
\frac{\partial\mathcal{J}(\theta,\hat{p})}{\partial\,\mathrm{diag}[\mathbf{R}^{-1}]} &= \mathrm{diag}[\frac{N\tau}{2}\mathbf{R} - \frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{\tau}\langle(\mathbf{x}_t^n - \bar{\bar{\mathbf{A}}}\bar{\mathbf{s}}_t^n)(\mathbf{x}_t^n - \bar{\bar{\mathbf{A}}}\bar{\mathbf{s}}_t^n)^T\rangle] \\
&= \mathrm{diag}[\frac{N\tau}{2}\mathbf{R} - \frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{\tau}(\mathbf{x}_t^n)^T\mathbf{x}_t^n - \frac{1}{2}\mathbf{A}[\sum_{n=1}^{N}\sum_{t=1}^{\tau}\langle\bar{\mathbf{s}}_t^n(\bar{\mathbf{s}}_t^n)^T\rangle]\mathbf{A}^T \\
&+ \frac{1}{2}[\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathbf{x}_t^n(\bar{\mathbf{s}}_t^n)^T]\mathbf{A}^T + \frac{1}{2}\mathbf{A}\sum_{n=1}^{N}\sum_{t=1}^{\tau}\bar{\mathbf{s}}_t^n(\mathbf{x}_t^n)^T]
\end{aligned}$$

We now use the expression for $\bar{\bar{\mathbf{A}}}_{\mathbf{new}}$:

$$\mathbf{R_{new}} = \frac{1}{N\tau}\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathrm{diag}[\mathbf{x}_t^n(\mathbf{x}_t^n)^T - \bar{\bar{\mathbf{A}}}_{\mathbf{new}}\hat{\bar{\mathbf{s}}}_t^n(\mathbf{x}_t^n)^T] \tag{3.21}$$

The estimators, $\bar{\mu}_{i,\mathbf{new}}$, $\bar{\mathbf{\Sigma}}_{i,\mathbf{new}}$, $\bar{\mathbf{F}}_{i,\mathbf{new}}$ and $\bar{\mathbf{Q}}_{i,\mathbf{new}}$, which correspond to individual source models must be invoked for all $i$ in order to construct the total source model, $\bar{\mu}_{\mathbf{new}}$, $\bar{\mathbf{\Sigma}}_{\mathbf{new}}$, $\bar{\mathbf{F}}_{\mathbf{new}}$ and $\bar{\mathbf{Q}}_{\mathbf{new}}$, in accordance with the model definitions of chapter 2.

### 3.2.2 Specialization to instantaneous mixing of AR(p) sources

It is not desirable to estimate a convolutive model in cases where the mixing process is sufficiently described by an instantaneous mixing matrix. Therefore the general estimators are specialized to a constrained form in accordance with the instantaneous mixing model. The first step is to rewrite $\mathcal{J}(\theta,\hat{p})$ slightly for the purposes, so that it reflects an instantaneous mixing model:

$$\mathcal{J}(\mathbf{A}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{\tau}\langle(\mathbf{x}_t^n - \mathbf{A}\ddot{\mathbf{s}}_t^n)^T\mathbf{R}^{-1}(\mathbf{x}_t^n - \mathbf{A}\ddot{\mathbf{s}}_t^n)\rangle$$

where the sparse source vector, $\ddot{\mathbf{s}}_t^n$, was defined in equation 3.13. The estimator for $\mathbf{A}$ is derived in complete analogy with that of section 3.2.1. The partial derivative is:

$$\frac{\partial\mathcal{J}(\mathbf{A})}{\partial\mathbf{A}} = -\frac{1}{2}\sum_{n=1}^{N}[2\mathbf{R}^{-1}\mathbf{A}\sum_{t=1}^{\tau}\langle(\ddot{\mathbf{s}}_t^n)^T\ddot{\mathbf{s}}_t^n\rangle - 2\mathbf{R}^{-1}\sum_{t=1}^{\tau}[\mathbf{x}_t^n\langle(\ddot{\mathbf{s}}_t^n)^T\rangle]]$$

$\mathbf{A}$ is solved for:

$$\mathbf{A_{new}} = [\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathbf{x}_t^n(\hat{\ddot{\mathbf{s}}}_t^n)T][\sum_{n=1}^{N}\sum_{t=1}^{\tau}\ddot{\mathbf{M}}_t^n]^{-1} \tag{3.22}$$

Finally, $\bar{\mathbf{A}}$ is reconstructed by inserting matrices of zeros between the columns. This special constrained format is defined and described in chapter 2.

### 3.2.3 Specialization to low-order source models

It was established in chapter 2, that a low-order AR, or constrained, source model is less prone to be invariant to unwanted rotations. Therefore an estimator is derived that allow for $p < L$, whereas the general estimator assumes that $p = L$. From the definition of $\mathcal{J}(\theta, \hat{p})$, it is clear the constrained estimator can be obtained by a truncation of the AR parameter vector:

$$\tilde{\mathbf{f}}_i^n = [\ (\mathbf{f}_i^n)_1 \quad (\mathbf{f}_i^n)_2 \quad \cdots (\mathbf{f}_i^n)_p \quad 0 \quad \cdots \quad 0\ ]^T$$

The relevant part of $\mathcal{J}(\theta, \hat{p})$ is retained in:

$$\mathcal{J}(\tilde{\mathbf{f}}_i^n) = \langle \frac{1}{q_i^n} \sum_{t=2}^{\tau} (s_{i,t}^n - (\tilde{\mathbf{f}}_i^n)^T \mathbf{s}_{i,t-1}^n)^2 \rangle$$

The derivations of the estimator are analogous to the general case and are omitted here. The result is:

$$\tilde{\mathbf{f}}_{\mathbf{new},i}^n = \Big[\ [\{\textstyle\sum_{t=1}^{\tau-1} \tilde{\mathbf{M}}_{i,t}^n\}^{-1} \sum_{t=2}^{\tau} \tilde{\mathbf{m}}_{i,t}^{1,n}]^T \quad 0 \quad \cdots \quad 0\ \Big]^T$$

where $\tilde{\mathbf{M}}_{i,t}^n$ and $\tilde{\mathbf{m}}_{i,t}^{1,n}$ are the truncated autocorrelation functions.

### 3.2.4 Specialization to instantaneous mixing of AR(1) sources

Straightforward specialization of the obtained estimators for the convolutive AR(p) model yields:

$$\mu_{\mathbf{new}}^n = \hat{\mathbf{s}}_1^n \tag{3.23}$$

$$\mathbf{\Sigma}_{\mathbf{new}}^n = \mathrm{diag}[\mathbf{M}_1^n] - \mathrm{diag}[\mu_{\mathbf{new}}^n (\mu_{\mathbf{new}}^n)^T] \tag{3.24}$$

$$\mathbf{F}_{\mathbf{new}}^n = \Big[\sum_{t=2}^{\tau} \mathrm{diag}[\mathbf{M}_t^{1,n}]\Big]\Big[\sum_{t=1}^{\tau} \mathrm{diag}[\mathbf{M}_{t-1}^n]\Big]^{-1} \tag{3.25}$$

$$\mathbf{Q}_{\mathbf{new}}^n = \frac{1}{\tau - 1}\Big[\sum_{t=2}^{\tau} \mathrm{diag}[\mathbf{M}_t^n] - \mathbf{F}_{\mathbf{new}}^n \,\mathrm{diag}[\mathbf{M}_t^{1,n}]\Big] \tag{3.26}$$

$$\mathbf{A}_{\mathbf{new}} = \Big[\sum_{t=1}^{\tau} \mathbf{x}_t^n (\hat{\mathbf{s}}_t^n)^T\Big]\Big[\sum_{t=1}^{\tau} \mathbf{M}_t^n\Big]^{-1} \tag{3.27}$$

$$\mathbf{R}_{\mathbf{new}} = \frac{1}{\tau} \sum_{t=1}^{\tau} \mathrm{diag}[\mathbf{x}_t^n (\mathbf{x}_t^n)^T - \mathbf{A}_{\mathbf{new}} \hat{\mathbf{s}}_t^n (\mathbf{x}_t^n)^T] \tag{3.28}$$

En passant, it is noted that although the direct specialization of the more general estimators seems obvious considering the present stage of knowledge, it was not the actual sequence of derivation. Instead, the estimators were derived in the order of the model definitions and generalizations adhered to in chapter 2. The procedure was to derive partial derivatives of $\mathcal{J}$ wrt. the individual elements of the parameter matrices and then deduce the total vector/matrix estimator.

### 3.2.5 Normalization

At each M-step one particular solution is chosen from the continuum of equivalent scalings of $\mathbf{A}$ and $\mathbf{Q}$, namely the one where $||[\mathbf{a}_{1j}^T \mathbf{a}_{2j}^T..\mathbf{a}_{d_xj}^T]^T|| = 1$, that is the Euclidian norm of each row of filters in $\bar{\bar{\mathbf{A}}}$ is normalized to 1.

## 3.3 BIC computation

As a result of the probabilistic formulation of the BSS problem and the analytical tractability of the model, the log-likelihood of the parameters, $\mathcal{L}(\theta)$, can be computed *exactly*. A major advantage is the ability to compute directly the Bayes information criterion (BIC). The fundamental idea is that model order control is desirable. In accordance with the principle of Occam's Razor (see [23]), the smallest model possible that explains the data 'satisfactorily' should be selected. In this work, the number of hidden sources, $d_s$, will be determined. For that purpose, we would like to marginalize the parameters, $\theta$ and obtain the *model*-likelihood, or *evidence* of the model:

$$p(\mathbf{X}|d_s) = \int d\theta p(\mathbf{X}|\theta, d_s)p(\theta) \tag{3.29}$$

Unfortunately, the above integral cannot be performed. Instead, the following approximation, which is valid for large $N\tau$, is used:

$$p(\mathbf{X}|d_s) \approx p(\mathbf{X}|\theta_{ML}, d_s) - \frac{|\theta|}{2}\log(N\tau) \tag{3.30}$$

where $\theta_{ML}$ is the maximum-likelihood estimate of the parameters. The latter term penalizes high-order models through the number of free parameters, $|\theta|$. The approximation was obtained by assuming a large data volume ($N\tau$) and imposing a quadratic model on $p(\mathbf{X}|\theta, d_s)$. In the EM scheme, it is hard to obtain $\theta_{ML}$ *exactly*, since only near-convergence is achieved.

Bayes' theorem can then be employed to compare the different model order hypotheses. A flat prior is assumed for $d_s$ in order to deduce the hypothesis posterior:

$$p(d_s|\mathbf{X}) = \frac{p(\mathbf{X}|d_s)}{\sum_{k=1}^{K} p(\mathbf{X}|d_s = k)} \tag{3.31}$$

This section relies on the BIC review in [25]. In chapter 4, experiments will be carried out that show the identification of the number of source from convolutive mixtures.

## 3.4 Adaptive Overrelaxed EM

For 'difficult' problems, standard EM optimization converges slowly or maybe not at all within reasonable time. Only limited time is available in real life applications, and often we can only hope to achieve near-convergence. Therefore it makes sense to try to get closer to the local minimum using an algorithm that converges faster.

Inspired by Salakhutdinov and Roweis [26], a simple optimization scheme is implemented for the speeding up of learning. At the heart of the amended algorithm is the augmented parameter update equation:

$$\theta_{t+1} = \theta_t + \eta(\theta_{t+1}^{EM} - \theta_t) \tag{3.32}$$

Equation 3.32 can be interpreted as a linear extrapolation of the usual M-step update of $\theta$. The learning rate is (optimistically) increased automatically on each step:

$$\eta_{\mathbf{new}} = \alpha \cdot \eta_{\mathbf{old}} \tag{3.33}$$

At each iteration it is checked that the algorithm does not overstep, in which case the usual M-step update is used to optimize $\theta$ and the learning rate is reset to $\eta = 1$. In this case, the standard EM-step results, as can be verified in equation 3.32.

# Chapter 4

# Experiments

The statistical model and the EM learning scheme presented lend itself to the solving of a number of problems from across the spectrum. First and foremost, the separation of an observed mixture into the original source signals is of special interest and will be the focus of most of the following experiments. It is investigated to which degree the separation quality depends on factors such as noise conditions, the data size and the spectral diversity within and between the sources. The performance of the proposed algorithm is furthermore compared with the algorithm of Parra and Spence [4].

Secondarily, the complexity of the model in terms of the number of sources, is addressed. Given two observation channels, the number of sources is determined. The mentioned problems of interest are investigated in 3 data domains of increasing difficulty:

Section 4.2: Initial experimentation verifies that the proposed algorithm successfully separates artificially generated mixtures that fit the model perfectly. For that purpose, realizations of AR-processes are simulated and mixed under varying noise conditions.

Section 4.3: Artificial mixtures of real speech signals are used as a stepping stone before advancing to real room mixtures. Microphone recordings of individual audio sources are filtered through known linear filters and added with noise to construct the sensor signals.

Section 4.4: Real recordings of speech mixtures is used to benchmark KaBSS. The performance of the algorithm is tested on publicly available signals, which have been used as reference signals by a number of authors, e.g. [9] and [4].

Issues concerning the algorithm, such as the frequency permutation problem and the convergence properties, are addressed in sections 4.5 and 4.6, respectively.

A speech recording is analyzed as a preliminary exercise in order to confirm the validity of the non-stationary AR model for a specific class of data, namely speech, see section 4.1.

## 4.1 Speech analysis

The main objective of the following is to obtain hands-on confirmation that the chosen non-stationary model is reasonable for speech as was conjectured in section 2.4.2.

### 4.1.1 Analysis of a speech recording

To illustrate the various features of speech, including the assumed wide-sense non-stationarity, an utterance of the word 'signals' by a male was recorded at $F_s = 8kHz$. Figure 4.1 plots the signal in the time domain. By zooming in on the various parts of the recording, the different waveforms, corresponding to different excitation and formant location, are revealed. The waveforms can be viewed in figure 4.2, although the phenomena are exposed more clearly in the spectral domain representation of the signal that is seen in figure 4.3.

The spectral variation over time caused by the different excitation signals and vocal tract filtering is clearly visible. For instance, during the periods of 's' and 'l', the spectrum is similar to that of broad band noise, while the harmonics stemming from the periodic excitation are clearly discernable during the periods of the vowels 'i' and 'a'.

Furthermore, the vowels are identifiable by the location of the *formants*, which are the points of amplification of the vocal tract filter. For instance, the 'i' is predominantly located in the frequencies lower than 600 and between 2000 and 3500. This observation is in concordance with the formants plot in figure 2.5. Contrarily, the 'a' is located solely in the hundreds.

The spectral time-variance observed in the recording, e.g. between the voiced and the unvoiced speech and the variation among the vowels, is a factor that works in favor of KaBSS and any algorithm based on the non-stationarity of the sources. It was established in section 2.4.2 that the time-variance of the autocorrelation/spectrum was a condition for the uniqueness of $\bar{\bar{\mathbf{A}}}$ within the model. The present analysis confirms common knowledge about speech that such spectral variety exists.

### 4.1.2 The AR(2) model for speech

In order to assess the bias of the simple AR(2) model when applied to speech, a small experiment was carried out: a microphone recording of a male speech signal was obtained at a sample rate of $F_s = 8000kHz$ and segmented into windows consisting of $\tau = 160$ consecutive samples. The spectrogram can be viewed in figure 4.4. Subsequently, AR(2) models were adapted to the segments using the autocorrelation method. The corresponding power spectrogram was obtained by z-transforming the time-domain recursion, which yields the transfer function of the linear system, $H(z)$, and evaluating it on the unit circle as described in section 2.5.1. This has to be multiplied with the estimated innovation noise power:

$$P_s(f) \quad = \quad |H(f)|^2\sigma_v^2$$

The resulting single-pole power spectrum estimate is shown in figure 4.5. Notably, the harmonics of the original signal are absent, as they cannot be contained in the model. Furthermore, a number of interesting segments were se-

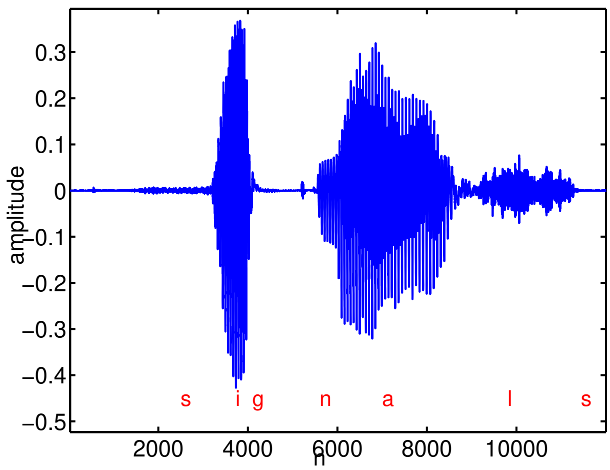


Figure 4.1: Amplitude plot of the utterance of the word 'signals'. The letters on the plot indicate the timing of word components 's', 'i', 'g', 'n', 'a', 'l', 's'. The signal was sampled at $F_s = 8kHz$.

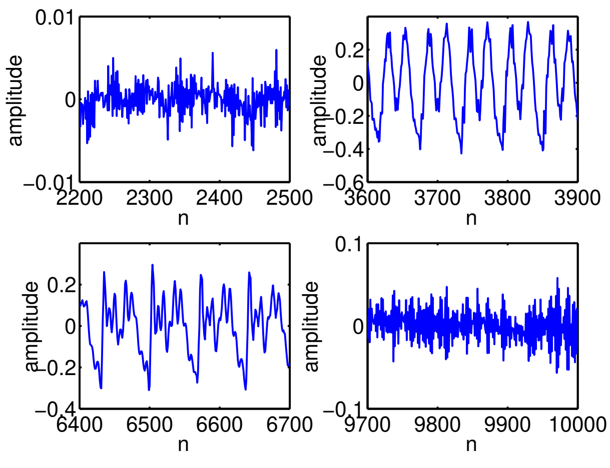


Figure 4.2: Amplitude plots of various waveforms stemming from the signal of figure 4.1. From top left to lower right, the following phonemes are represented: 's', 'i', 'a', 'l'. While 's' and 'l' are both similar to white noise, the vowels, 'i' and 'a', are clearly harmonic and distinguishable by their different harmonics fingerprints.

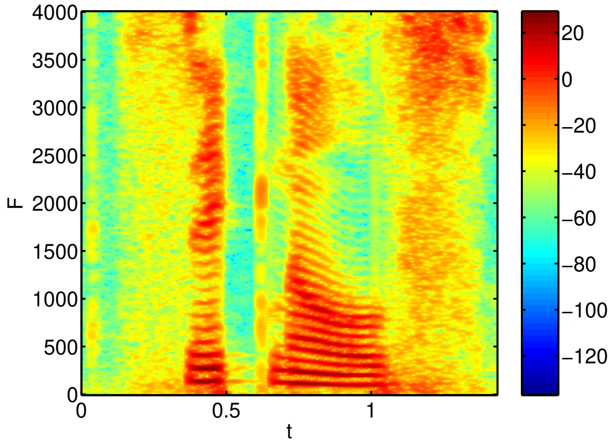


Figure 4.3: Spectrogram of the recorded ($F_s = 8kHz$) utterance of the word 'signals'. The spectra are estimated from overlapping frames, length 512, and displayed in $dB$. Several features of speech are clearly visible, e.g. voiced/unvoiced speech, formants, pulse excitation, non-stationarity and amplitude modulation.

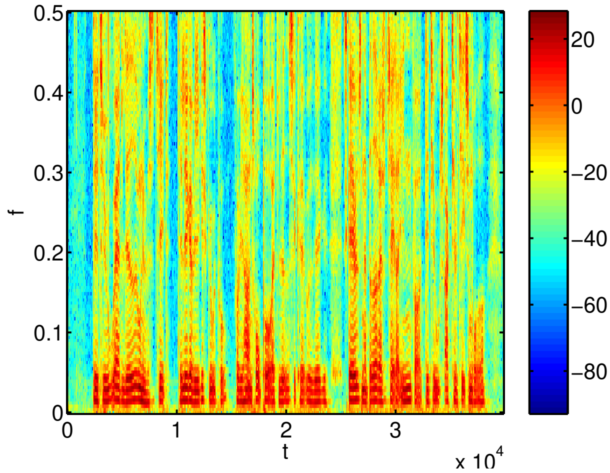


Figure 4.4: Spectrogram (dB) of the original male speech signal sampled at $F_s = 8000kHz$. The spectrogram was produced using a window function and overlapping frames.

lected for analysis, namely those in which the pole of $H(z)$ was estimated to be complex-conjugate. The pole locations, $\omega_0$, and peakedness, $r$, were computed from $H(z)$ in accordance with the descriptions in section 2.5.1. They are graphed as a time-function in figure 4.6. The pole-location, $\omega_0$, varies little over segments, whereas the peakedness seems less stationary.

It can be said that the AR(2) is a rough model for speech as only one spectral peak can be described. In contrast, it is known that vowels are characterized by 2-3 formants, or spectral peaks. In addition, the harmonics cannot be modelled either. This might spell trouble for KaBSS, since the simpler AR models are preferred in order to avoid the arbitrary rotation of the source estimate discussed in section 2.4.2.

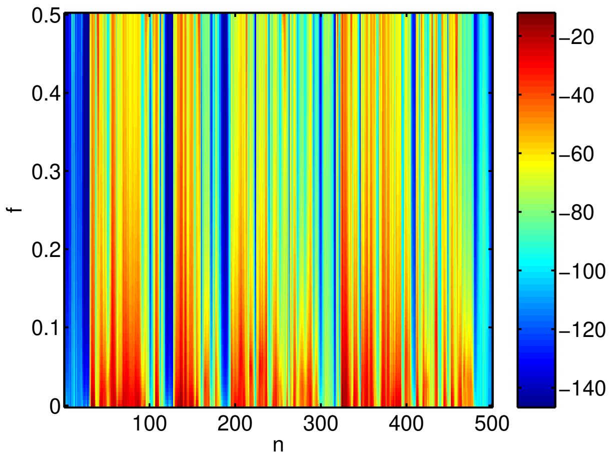


Figure 4.5: Estimated power spectrogram (dB) of a male speech signal ($F_s = 8kHz$) assuming an AR(2) model. The autocorrelation method was used with a frame length of $\tau = 160$.

## 4.2 Artificial data

The purpose of experimenting with non-real data is at least twofold. Firstly, verification is obtained that the algorithm actually converges under optimal conditions, i.e. when data is generated according to the model assumptions. Secondly, the behavior of the algorithm under different conditions can be investigated in a strictly controlled and well-described environment. However, in order for the experiments with artificially generated signals to have any practical value, the generated signals must in some way resemble real world signals. For that purpose, a small study of speech signals was carried out in section 4.1. We learned that low-frequent AR(2) processes emulate some of the characteristics of speech, leading to the employment of this class of signals in the following experiments with artificial data.

It will not be attempted to replicate the reverberation filters of the real world. They are often very long, e.g. that of a cathedral, in order to accommodate the delays of numerous echoes. We will concentrate on representing anechoic-like

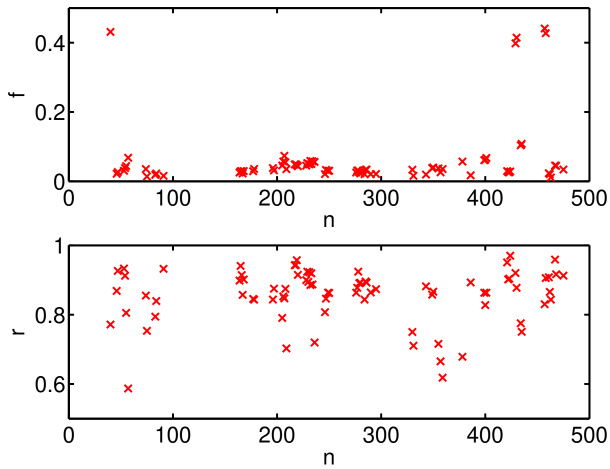


Figure 4.6: The piecewise modelling of a speech signal by an AR(2) process using the autocorrelation method. Only those windows are shown, where transfer functions with complex-conjugate roots resulted from the estimation, and where the signal power exceeded a background noise threshold. The upper panel shows the pole location parameter, $f_0 = \frac{\omega_0}{2\pi}$, in each data window plotted on the normalized frequency axis. The lower panel shows the pole peakedness parameter, $r$, a measure of the peakedness of the spectrum on a scale going from 0 to 1, flat and peaked, respectively.

| | $\omega_0$ | |
|---|---|---|
| $n$ | source 1 | source 2 |
| 1 | 0.8600 | 0.3194 |
| 2 | 0.3597 | 0.9130 |

Table 4.1: The spectral location of the poles of source signals 1 and 2 in segments $n = 1$ and $n = 2$. Frequencies are normalized in the range $[0; \pi]$.

conditions with short filters.

### 4.2.1 Learning in quadratic mixtures

As will be demonstrated shortly, KaBSS works well for quadratic separation problems involving short filter lengths, i.e. $d_s = d_x$ and $L$ small. It turns out that source signals and parameters are recovered with little error up to the well-known permutation and scaling. For the following experiment, low-frequent realizations of AR(2)random processes were generated along the lines of section 2.5.1. Two segments were generated, $N = 2$, consisting of $\tau = 200$ consecutive samples. The pole location can be viewed in table 4.2.1. It is noted that the spectral overlap of the sources is rather limited. The peakedness parameter was set to $r = 0.9$. The sources were filtered trough $\bar{\bar{A}}$, where a filter length of $L = 2$ was selected. The filter coefficients were drawn from univariate zero-mean normal distributions (variance 1):

$$\bar{\bar{\mathbf{A}}} = \begin{bmatrix} -0.3401 & 0.6293 & & -0.0197 & -0.4774 \\ -0.5885 & 0.3768 & & -0.2988 & -0.8261 \end{bmatrix} \tag{4.1}$$

Subsequently Gaussian diagonal observation noise ($\mathbf{R}$ diagonal) was added in both channels to construct an SNR of 20dB:

$$\begin{aligned} P_{n_i} &= \tfrac{P_{x_i}}{10^{SNR/10}} \\ n_{i,t} &\sim \mathcal{N}(0, P_{n_i}) \qquad i = 1, 2 \end{aligned} \tag{4.2}$$

For this purpose, the power of the mixtures was estimated. Figures 4.7 and 4.8 display the true and estimated source signals along with the mixtures and the observation noise. Figure 4.9 shows the non-decreasing learning curve of $\mathcal{L}(\theta)$. It is verified in figures 4.10, 4.11 and 4.12 that the parameter estimates converge to the true values. Prior to the plotting the arbitrary scaling and permutation of the source and parameter estimates were fixed.

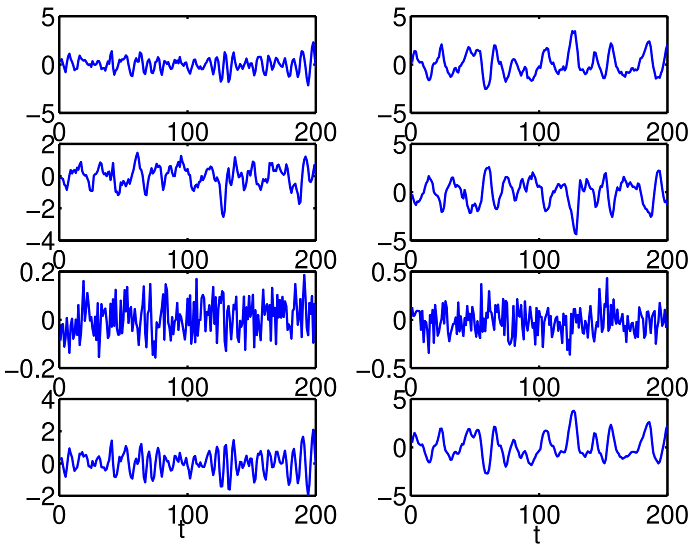


Figure 4.7: The signals of segment 1 of the quadratic mixture. From top to bottom: 1) The true source signals (1 & 2). 2) The noise-free convolutive mixings (1 & 2). 3) The observation noise (channel 1 & 2) . 4) The inferred sources (1 & 2).

With this small demonstration, a success was recorded for KaBSS. Sources and parameters were all recovered within approximately 100 iterations.

### 4.2.2 Monaural signal separation.

The extraction of more than one source signal from a single observation channel attracts its share of attention within the BSS/ICA communities, see e.g. Roweis' one-microphone separator [27]. Although not the main focus of this

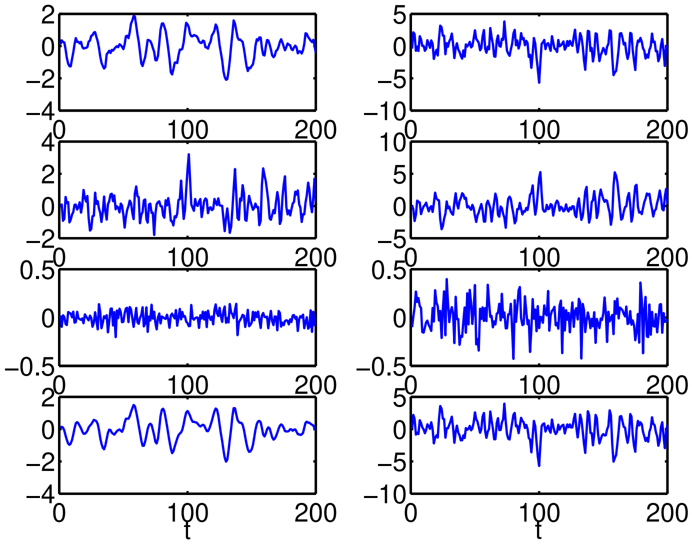


Figure 4.8: Segment 2 of the quadratic mixture. Same key as segment 1.

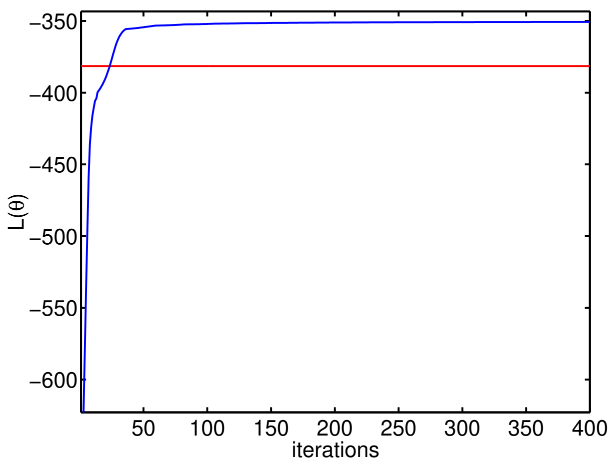


Figure 4.9: Learning in a quadratic mixture: the log-likelihood of the parameters, $\mathcal{L}(\theta)$, never decreases and eventually supersedes the log-likelihood of the true parameters $\mathcal{L}(\theta_{true})$, the horizontal line. This expresses fitting to a limited volume of data.

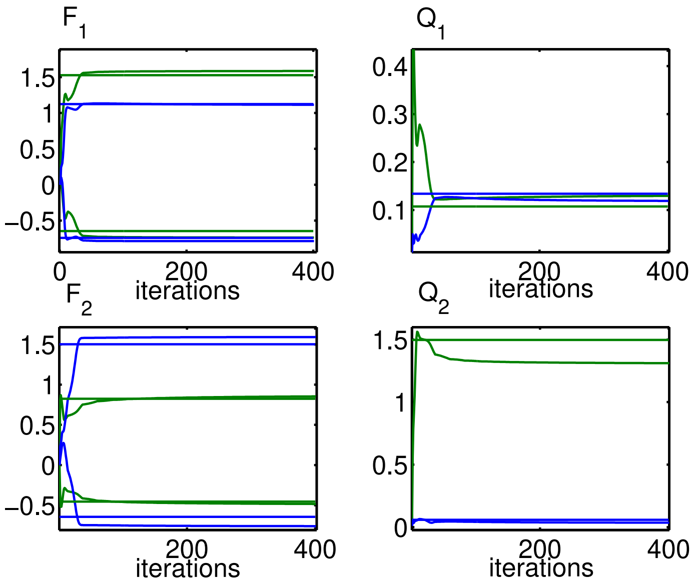


Figure 4.10: Learning in a quadratic mixture: converging estimates of source model parameters, $\bar{\mathbf{F}}$ and $\bar{\mathbf{Q}}$. Parameters of segments 1 and 2 in top and bottom subplots, respectively. Colors blue and green correspond to sources 1 and 2. Straight horizontal lines mark true values.

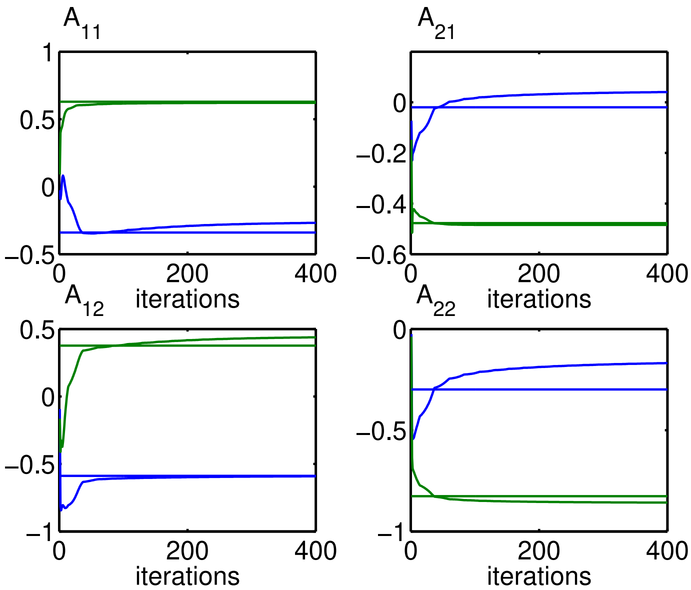


Figure 4.11: Learning in a quadratic mixture: estimate of $\bar{\bar{\mathbf{A}}}$ as a function of EM iterations. The four subplots each hold one filter. Colors blue and green refer to coefficients 1 and 2. Straight horizontal lines mark true values.

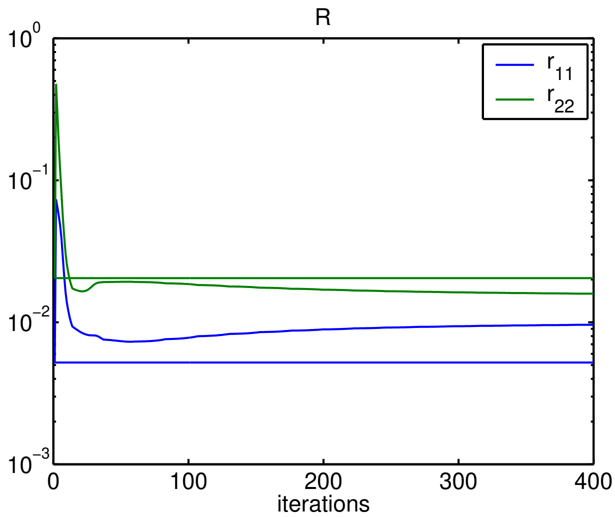


Figure 4.12: Learning in a quadratic mixture: Converging estimates of $\mathbf{R}$. The blue and green colors correspond to the observation noise variance in channels 1 and 2, respectively. Some inaccuracy of the estimators is noted. Straight horizontal lines mark true values.

work, a simple experiment was carried out that involves a monaural, instantaneous mixing of 2 narrow-band AR(2) sources with added white Gaussian observation noise. The resulting mixture consisted of $\tau = 100$ samples. One source was low-frequent ($\omega_0 = \frac{\pi}{10}$), while the other was high-frequent ($\omega_0 = \pi - \frac{\pi}{20}$). The asymmetry of the poles was chosen in order to guarantee a measure of generality.

The convolutive model is not strictly required in monaural signal separation, since its main motivation in, e.g., binaural problems is the potentially different time-shifting in different channels. As a consequence, the mixing was performed by simply adding the two sources, leading to the following true 'filter' matrix:

$$\bar{\mathbf{A}} = \begin{bmatrix} 1 & 0 & & 1 & 0 \\ 1 & 0 & & 1 & 0 \end{bmatrix} \tag{4.3}$$

No mixing matrix was estimated in this experiment. White Gaussian noise was added to construct an SNR of 30dB as described in the previous section.

The original source signals, the mixture, the added noise and the source estimates can be viewed in figure 4.13. The learning curve is shown in figure 4.14. The convergence of $\mathbf{R}$, $\bar{\mathbf{Q}}$ and $\bar{\mathbf{F}}$ are graphed in figures 4.15, 4.16 and 4.17.

The separation of two frequency-distinct signals is not ground-breaking, since it could be done by e.g. bandpass filtering. However, the approach taken here is essentially *blind*. Given the information that the mixture is composed of two sources , their spectral location and spread are recovered along with MAP-estimates of the source signals. Furthermore the noise level is well estimated. Future experiments should be devoted to investigate mixtures of spectrally overlapping sources.

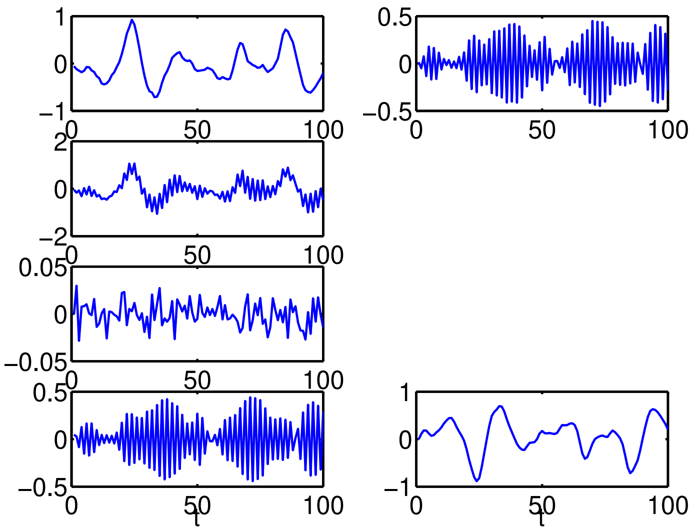


Figure 4.13: Monaural blind source separation, from top to bottom: The original sources 1 and 2, the mixture, the observation noise and the estimated sources 1 and. Note the permutation and (negative) scaling of $\hat{\mathbf{s}}_t$.

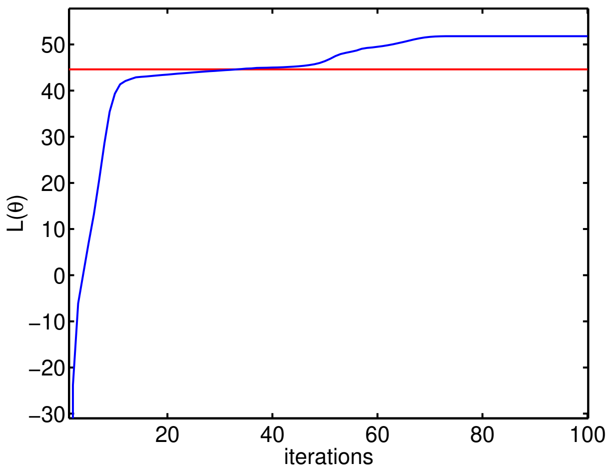


Figure 4.14: The learning curve for the monaural problem in terms of $\mathcal{L}(\theta)$ Something 'new' appears to be learned after 50 iterations. The log-likelihood of the true parameters is marked by the horizontal line.

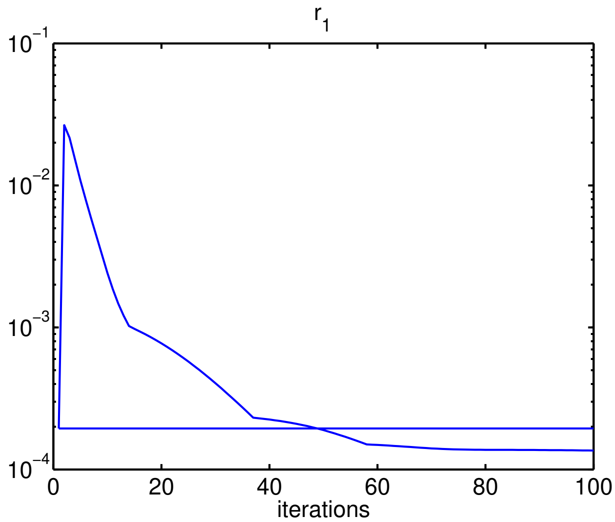


Figure 4.15: The convergence of the noise variance estimate for the monaural problem. The horizontal line marks the true value.

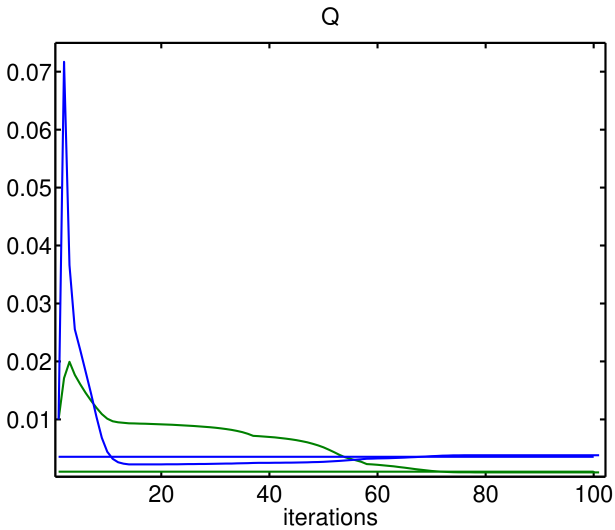


Figure 4.16: The convergence of the estimates of the innovation noise variances. Colors blue and green signify source 1 and 2. The source estimate assignment to numbers was reversed to fix a permutation of sources. The horizontal lines mark the true values.

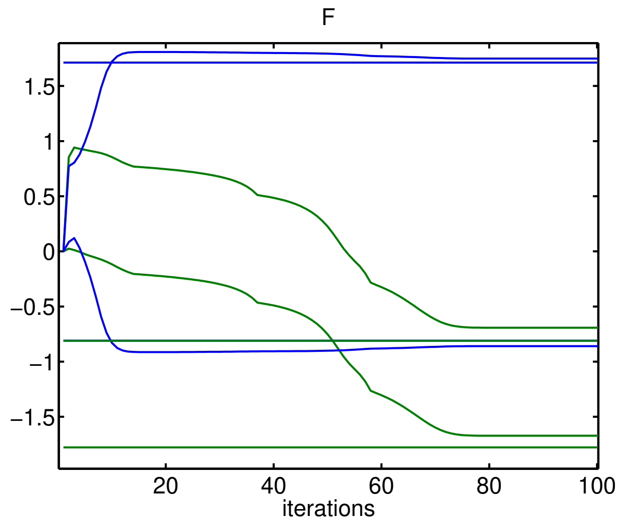


Figure 4.17: The convergence of the estimates of the AR-parameters. Colors blue and green correspond to sources 1 and 2. A permutation of sources was fixed by switching the assignment of numbers to sources. The horizontal lines mark the true values.

### 4.2.3 Model order

It will now be demonstrated that KaBSS can be used to determine the number of sources. A major advantage of having formulated a statistical model is that, e.g., the Bayes Information Criterion (BIC) can be employed to obtain an approximate model-posterior, $p(d_s|\mathbf{X})$. The reasoning is described in section 3.3.

#### Data

The source signals were generated as realizations of narrowband AR(2) processes. The length of the signals was $\tau = 100$. The peakedness parameter, which was discussed in section 2.5.1, was set to $r = 0.9$. Three classes of mixtures were created of varying spectral diversity, by drawing the pole locations, $\omega_0$, from uniform distributions, $\mathcal{U}(\cdot)$, reflecting the desired spectral variation.

1. High spectral diversity. $\omega_0 \sim \mathcal{U}(\frac{\pi}{10}, \frac{17\pi}{20})$
2. Medium spectral diversity. $\omega_0 \sim \mathcal{U}(\frac{\pi}{10}, \frac{3\pi}{5})$
3. Low spectral diversity. $\omega_0 \sim \mathcal{U}(\frac{\pi}{10}, \frac{7\pi}{20})$

When similar pole locations in a single segment resulted from the data construction phase, the pole locations were automatically redrawn from the distributions. A histogram of the sample distribution of $\omega_0$ is shown in figure 4.18.

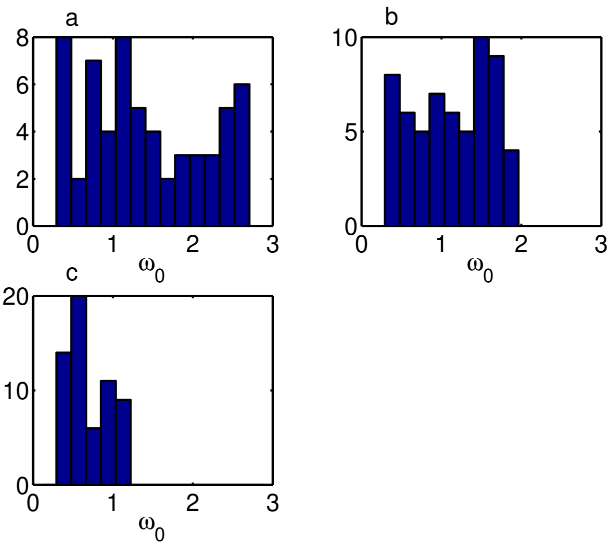


Figure 4.18: Data generation for the experiments concerning the determination of the number of sources: the distribution of the poles of the sources in the experiment with a) high spectral diversity, b) medium spectral diversity and c) low spectral diversity. The corresponding intervals are a) $\frac{\pi}{10} < \omega < \frac{17\pi}{20}$, b) $\frac{\pi}{10} < \omega < \frac{3\pi}{5}$ and c) $\frac{\pi}{10} < \omega < \frac{7\pi}{20}$.

Depending on the number of sources in the particular experiment, the source signals were filtered through the relevant filters, which are inspired by [6]:

$$\begin{aligned}
\begin{bmatrix} \mathbf{a}_{11}^T \\ \mathbf{a}_{12}^T \end{bmatrix} &= \begin{bmatrix} 1 & 0.35 & -0.2 & 0 & 0 \\ 0 & 0 & 0.7 & -0.2 & 0.15 \end{bmatrix} \\
\begin{bmatrix} \mathbf{a}_{21}^T \\ \mathbf{a}_{22}^T \end{bmatrix} &= \begin{bmatrix} 0 & 0 & -0.5 & -0.3 & 0.2 \\ 1.3 & 0.6 & 0.3 & 0 & 0 \end{bmatrix} \\
\begin{bmatrix} \mathbf{a}_{31}^T \\ \mathbf{a}_{32}^T \end{bmatrix} &= \begin{bmatrix} 0.8 & 0.6 & 0.1 & 0 & 0 \\ 0 & 0.7 & 0.9 & 0.05 & 0 \end{bmatrix} \\
\begin{bmatrix} \mathbf{a}_{41}^T \\ \mathbf{a}_{42}^T \end{bmatrix} &= \begin{bmatrix} 0 & 0 & 0.8 & -0.3 & 0.2 \\ 0 & 1 & -0.4 & 0.1 & 0 \end{bmatrix}
\end{aligned}$$

A total of $N = 5$ segments of mixture was generated and added with white Gaussian noise at $SNR = 25dB$, as prescribed previously in this chapter.

**Simulation setup**

In order to determine the most likely model order hypothesis, the model posterior $p(d_s|\mathbf{X})$ must be approximated. For this purpose, the evidences of a range of models are computed according to the following procedure:

1. The parameters of the source model ($\bar{\mathbf{F}}$, $\bar{\mathbf{Q}}$) and of the observation model ($\bar{\bar{\mathbf{A}}}$, $\mathbf{R}$) are initialized by assigning zeros to the former and non-zero values to the latter.

2. The EM-algorithm estimates all parameters ($\bar{\mathbf{F}}$, $\bar{\mathbf{Q}}$, $\bar{\bar{\mathbf{A}}}$, $\mathbf{R}$) and computes the log-likelihood, $\mathcal{L}(\theta|d_s)$. KaBSS is allowed 160 iterations to converge.

3. Repeat 8 times the steps 1-2 and select the highest $\mathcal{L}(\theta|d_s)$.

4. Repeat the steps 1-3 for various model orders, $d_s = 1, 2, 3, 4, 5$, obtaining the various $\mathcal{L}(\theta|d_s)$.

5. Compute the BIC likelihood, $p(\mathbf{X}|d_s)$, and the resulting posterior, $p(d_s|\mathbf{X})$ according to the formulae of section 3.3. The number of free parameters are:

$$|\theta| \quad = \quad d_s \times [\underbrace{(p+1) \times N}_{\bar{\mathbf{F}},\bar{\mathbf{Q}}} + \underbrace{L \times d_x}_{\bar{\bar{\mathbf{A}}}}] + \underbrace{d_x}_{\mathbf{R}} - 1.$$

The above was repeated 6 times for each of the spectral diversity classes that were mentioned in section 4.2.3 and for an array of true model orders, $d_s^{true} = 2, 3, 4$. An example of the resulting $p(d_s|\mathbf{X})$ can be viewed in figure 4.19.

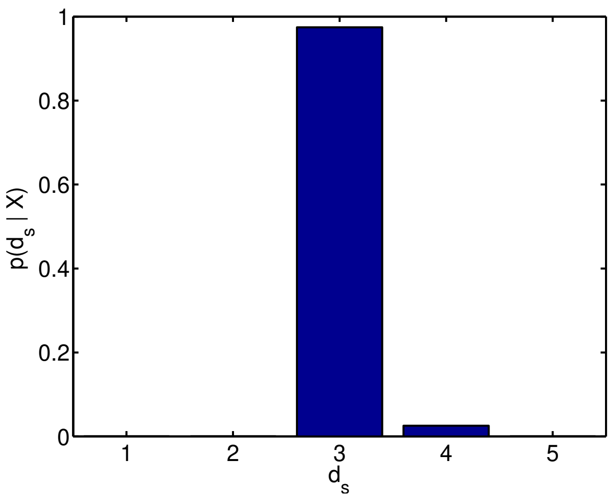


Figure 4.19: Example of a posterior probability distribution resulting from the BIC approximation. The MAP estimate of the number of sources, $d_s$, is $\hat{d}_s = 3$, since this is the single most likely hypothesis.

In table 4.2.3, the fraction of the experiments that yielded an estimate of $d_s$ in agrement with $d_{s,true}$ is listed for the different diversity classes. As anticipated, the model order was easier to estimate, when the various sources distinguished themselves in the frequency domain. It is a remarkable result that 3 sources in 2 sensors can be correctly assessed as such. One might speculate that less spectral variation could be accepted if the number of segments was increased. Future experimentation is to confirm such an hypothesis.

| spectral diversity | True $d_s$ 2 | 3 | 4 |
|---|---|---|---|
| high | 3/3 | 3/3 | 1/3 |
| medium | 6/6 | 1/6 | 0/6 |
| low | 6/6 | 0/6 | 0/6 |

Table 4.2: The fraction of correct classifications into the model orders specified by the column labels. The different rows correspond to the varying spectral diversity of the source signals. Only 3 simulations were carried out for the 'high' category, elsewhere 6 simulations.

## 4.3 Mixtures of audio signals

In the following, experiments are carried out with authentic audio recordings that are mixed artificially. The performance of the proposed algorithm is tested in various noise settings, for different data sizes both in terms of the number of segments, $N$, and the segment length, $\tau$, and for different orders of the AR-processes. The performance is assessed by measuring the cross-talk in terms of the signal to interference ratio (SIR), a metric that is described in appendix A.

### 4.3.1 Noise-dependency

In order to investigate the behavior of the algorithm in noisy conditions, real recordings of speech were mixed through artificial channel filters and added with varying levels of noise. Albeit much shorter than real room impulse functions (that can be seconds long), the designed filters contain delays. Hence the obtained mixing is fundamentally different from instantaneous mixing and resembles that experienced in an anechoic chamber.

The speech signals were recorded at $8kHz$ from one male speaker and 2 active periods of length $T = 5s$ were chosen for the experiment. Signal path filters of length $L = 3$ were used to artificially mix the 2 source signals into the 2-dimensional observable mixture. As a consequence of the problem dimensions, 4 signal paths result. They are contained in $\bar{\bar{\mathbf{A}}}$:

$$\bar{\bar{\mathbf{A}}} = \begin{bmatrix} 1 & 0.3 & 0 & & 0 & 0 & 0.8 \\ 0 & 0.8 & 0.24 & & 1 & 0 & 0 \end{bmatrix} \tag{4.4}$$

Observation noise was added in each sensor channel to construct the desired signal to noise ratio:

$$\begin{aligned} P_{n_i} &= \tfrac{P_{x_i}}{10^{SNR/10}} \\ n_{i,t} &\sim \mathcal{N}(0, P_{n_i}) \qquad\qquad i = 1, 2 \end{aligned} \tag{4.5}$$

In the following experiments, a segment of data consists of $\tau = 50$ subsequent samples. The sequence of experimentation followed the recipe presented below:

1. All data was segmented. A test set of $M_{test}$ segments is randomly sampled from the data. The remainder of the segments are assigned to the training pool, $M_{pool}$.

2. From $M_{pool}$, $M_{train}$ segments are randomly sampled for training.

3. The parameters of the source model ($\bar{\mathbf{F}}$, $\bar{\mathbf{Q}}$) and of the observation model ($\bar{\bar{\mathbf{A}}}$, $\mathbf{R}$) are initialized by assigning zeros to the former and non-zero values to the latter.

4. The EM-algorithm estimates all parameters ($\bar{\mathbf{F}}$, $\bar{\mathbf{Q}}$, $\bar{\bar{\mathbf{A}}}$, $\mathbf{R}$)

5. In order to test the estimates of $\bar{\bar{\mathbf{A}}}$ and $\mathbf{R}$ obtained on step (4), the algorithm estimates $\bar{\mathbf{F}}$ and $\bar{\mathbf{Q}}$ (not $\bar{\bar{\mathbf{A}}}$, $\mathbf{R}$) obtaining the test log-likelihood, $\mathcal{L}_{test}(\theta)$, in the process. Repeat from (2) the first $k$ times at this point.

6. Select the estimates of $\bar{\bar{\mathbf{A}}}$ and $\mathbf{R}$ that yielded the highest $\mathcal{L}_{test}(\theta)$.

7. As a final source signal inference step, use the best estimate of $\bar{\bar{\mathbf{A}}}$ and $\mathbf{R}$ and estimate $\bar{\mathbf{F}}$ and $\bar{\mathbf{Q}}$. The sources are inferred in the process. Overlapping windows are used in this stage.

8. Measure the SIR as described in appendix A.

The model orders were set to $d_s = 2$, $p = 2$ and $L = 3$. Summarizing, the algorithm was restarted $k = 10$ times within each experiment, training and testing the observation model ($\bar{\bar{\mathbf{A}}}$ and $\mathbf{R}$) on separate data sets. The size of the test set was $M_{test} = 100$ (segments), and for each restart the size of the training set was $M_{train} = 10$ (segments). A fixed number of iterations was allotted for the convergence of parameters that was aided by the adaptive step-size scheme of [26]. The learning rate was set to $\alpha = 1.2$. The numbers of iterations for the different stages of training, testing and final inference were:

$$\begin{aligned} I_{train} &= 300 \\ I_{test} &= 15 \\ I_{total} &= 15 \end{aligned}$$

The experiment described above was repeated 10 times for varying SNR. A reference method, the blind source separation algorithm of Parra and Spence (PS), [4], was tried on the same data. The hyper parameters of this algorithm were set to $T = 1024$, $Q = 6$, $K = 7$ and $N = 5$ after a search for the best fit in terms of SIR, see appendix A. Figure 4.20 shows the performance of the two algorithms for varying SNR. In very noisy conditions, KaBSS clearly outperforms PS. However, the separation ability of KaBSS collapses for SNR$\geq$ 30. This finding, which is crucial to the exploitation of the potential of KaBSS, leads to experimentation with added regularization noise.

### Noise regularization

A simple simulated annealing scheme was implemented to 'fix' the apparent convergence problems in noise-free situations. Prior to invoking KaBSS as described above, noise is added at a specified SNR. The noisy mixture is used to obtain a starting guess on the parameters. Subsequently, the algorithm is run on the 'clean' mixture using the parameter estimates obtained on the noisy mixture. This approach proved successful as can be viewed in figure 4.21. Evidently, a wider range of SNR's can be treated by KaBSS. Figure 4.22 demonstrates that the noise-level is correctly estimated for all SNR. These results are significant, as they suggest the applicability of KaBSS to real problems. However, more

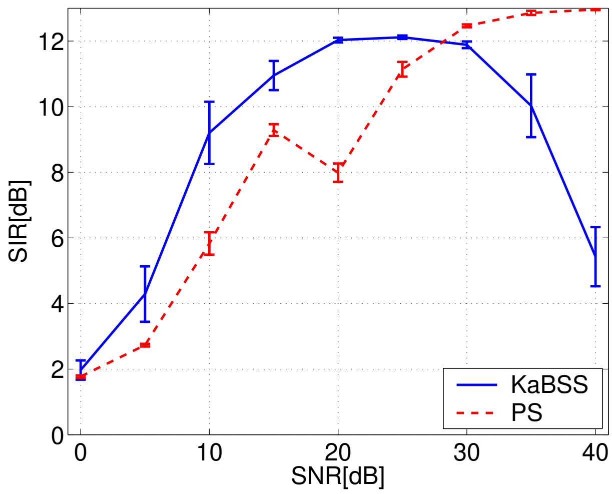


Figure 4.20: The separation performance for varying SNR of KaBSS (solid) and the reference method proposed by Parra and Spence (PS - dotted) [4]. The signals are two utterances by the same speaker. Two convolutive mixtures were created with variable strength additive white noise. The SIR measures the 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.

hyper-parameters have entered the equation, as the power of the added noise must be specified as must the number of annealing steps. In other words, the realm of meta-heuristics has been entered.

### Longer filters

The experiment setup was repeated for slightly larger filter lengths, i.e. $L = 8$:

$$\begin{bmatrix} \mathbf{a}_{11}^T \\ \mathbf{a}_{12}^T \end{bmatrix} = \begin{bmatrix} 0.3679 & 0.1353 & 0.0498 & 0.0183 & 0.0067 & 0.0025 & 0 & 0 \\ 0 & 0.3311 & 0.1218 & 0.0448 & 0.0165 & 0.0061 & 0.0022 & 0 \end{bmatrix}$$
$$\begin{bmatrix} \mathbf{a}_{21}^T \\ \mathbf{a}_{22}^T \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0.2943 & 0.1083 & 0.0398 & 0.0147 & 0.0054 & 0.0020 \\ 0.3311 & 0.1218 & 0.0448 & 0.0165 & 0.0061 & 0.0022 & 0 & 0 \end{bmatrix}$$

The filter coefficients were generated from the truncated and delayed exponential function, $\exp[-n+n_0]$, and therefore has all-pole-like filter characteristics. The SIR was measured for varying SNR on mixtures resulting from this filter. A setup identical to the one with the short filters was selected, and with noise regularization turned on. Figure 4.23 shows the results which indicate that separation can succeed for longer filters. However, the designed exponential filters are still in another category than real-world filters. Considerable degradation of the separation performance resulted from using longer mixing filters. The artificial filter used by Kidmose in [19] of length $L \approx 15$, was tried in the same experiment setup. A SIR of $9.4 \pm 0.6dB$ resulted from 10 repetitions, but a listening test and inspection of the signals revealed that some segments were

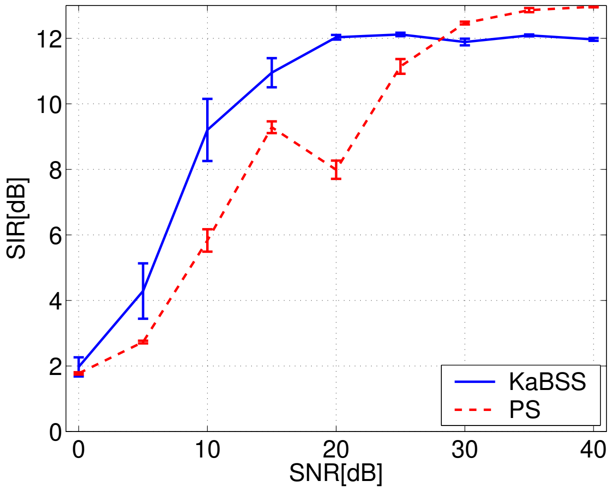


Figure 4.21: Separation performance with added regularization noise. Solid and dotted lines correspond to the SIR of KaBSS and PS, respectively. Otherwise same key as in figure 4.20.

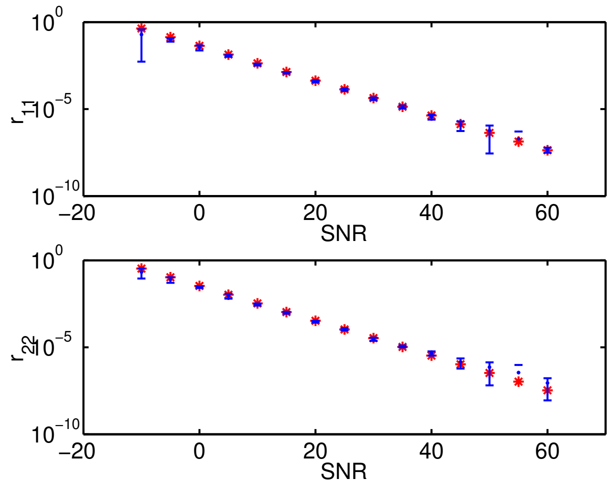


Figure 4.22: True (*) and estimated noise variances in channels 1 and 2. The small noise variances cannot be accurately estimated.

under the influence of unstable AR processes. Many of the runs, however, were successful.

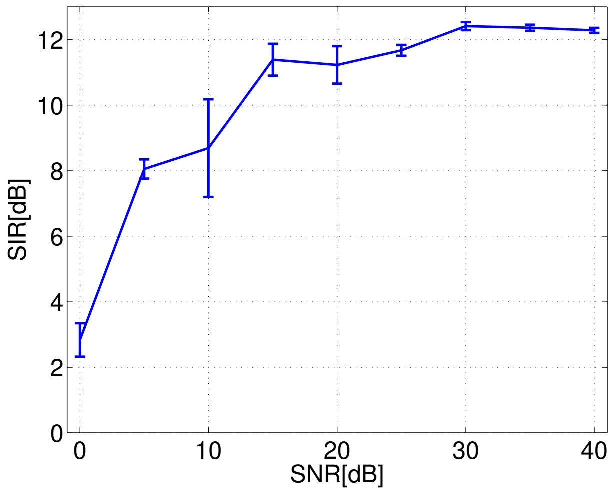


Figure 4.23: The measured SIR performance of KaBSS on the male-male problem with medium length filters.

### 4.3.2 Dependency on data size

In order to assess how much data was required in order for KaBSS to separate the signals, an experiment was undertaken that was similar to the one in section 4.3.1. The SIR was measured for varying $M_{train}$ and $\tau$, the number of training segments and the number of samples in a segment, respectively. The experiment was repeated 10 times for each $M_{train}$ and $\tau$ with the SNR fixed at 30dB.

Figure 4.24 clearly demonstrates that training on a single segment cannot lead to successful separation. For this particular problem, it seems that approx. 10 segments are required. Figure 4.25 show the segment length required for accurate estimation and inference of the sources. As expected, too short segments prove inadequate for estimating the source statistics of relatively low-frequent speech signals. We would expect performance to deteriorate for very large $\tau$ as speech is only wide-sense stationary in periods of up to $\sim 400ms$. This however, was excessively time consuming to prove empirically.

### 4.3.3 Dependency on AR model order

Another experiment similar to that of section 4.3.1 was carried out. Instead of varying the SNR, the model order, $p$, was investigated for a range of values. The SNR was fixed at 20dB. The experiment was repeated 5 times for each $p = [1, 2, 3, 4, 5, 6, 7, 8]$. Figure 4.26 displays the obtained SIR performance. Although the variance of the mean is relatively high, due to the limited number of experiments, a trend emerges from the measurements: simpler source models are to be preferred. This observation is in line with the deliberations of sections 2.4.2 and 2.4.3, which suggest the use of simple source models.

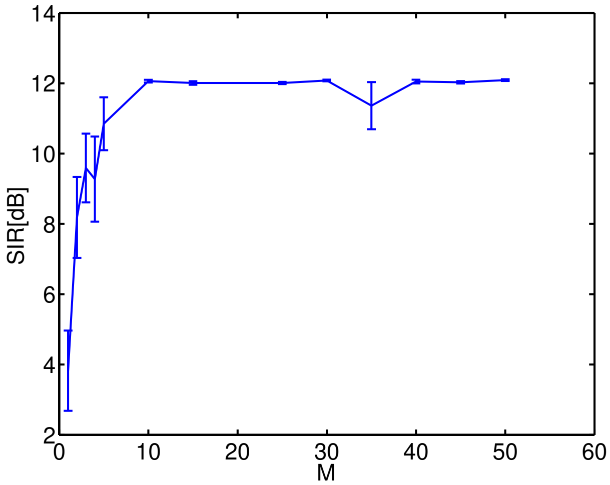


Figure 4.24: The separation performance for varying number of training segments, $M_{train}$. The optimal SIR is obtained for $M \geq 10$. A single segment is clearly inadequate. The segment length was fixed at $\tau = 50$.

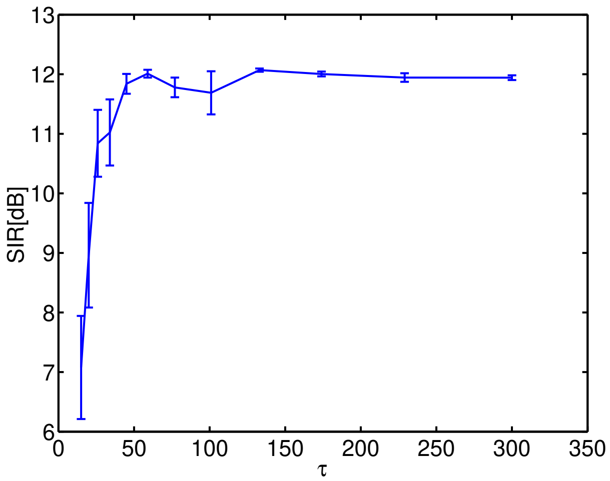


Figure 4.25: The separation performance for varying segment length. It seems that a segment length of $\tau = 70$ is required for optimal separation, given $M = 5$.

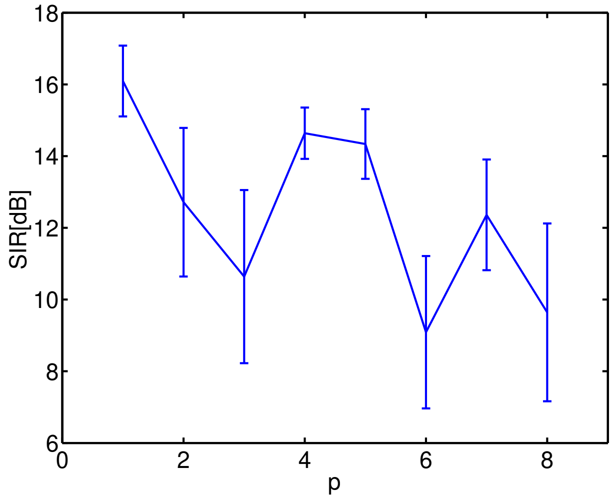


Figure 4.26: SIR for various orders of AR models for the sources. Two male speech signals were artificially mixed, sampled at $F_s = 8000$.

| | | KaBBS | PS |
|---|---|---|---|
| Mixture 1 | $\widehat{SIR}$ | 45.7 | 12.1 |
| | SE($\widehat{SIR}$) | 4.10 | 0.153 |
| Mixture 2 | $\widehat{SIR}$ | 34.8 | 10.5 |
| | SE($\widehat{SIR}$) | 2.77 | 0.0739 |

Table 4.3: The separation performance measured in SIR for the proposed algorithm (KaBSS) and the reference method (PS). The mean and the standard deviation of the mean are given for two mixtures of one male voice and one female voice. A total of 10 runs was performed for each mixture and algorithm.

### 4.3.4 Dependency on spectral diversity

It is well-known that spectral diversity is an important parameter in blind source separation problems. For instance, the separation of a female voice from a male voice obviously is easier than separating same gender voices. An experiment duplicating the setup of section 4.3.1 verifies this simple fact. All hyper parameters of both algorithms were left unchanged from previous experiments. Here only, the SNR was fixed at $20dB$. Male and female voices were recorded at $8kHz$ and subsequently mixed through the filters. Five seconds of mixture was generated. The male recordings originated from the same male speaker that was used in the experiments of section 4.3.1.

Table 4.3 reports the SIR results for two different mixtures. They suggest that KaBSS benefits greatly from the spectral variation of the male/female mixture. Compared to the male/male mixture, an SIR improvement of at least $22dB$ was measured for both the mixtures. The algorithm of Parra and Spence did not seem to benefit as greatly. This might be due to high sensitivity to signal and hyper parameters (which were fitted to the experiment of section

4.3.1).

Researchers, Hua and Tugnait, [14], have proved that distinct source power spectra are mandatory for the separation of colored noise sources based on second-order statistics. They considered only stationary signals. An algorithm, based on theoretic foundation laid out by these authors, was recently published by Kawamoto and Inouye, [15].

## 4.4 Real-life data

Now follow attempts to separate mixtures that were actually measured in the real world. In order to facilitate the comparison with other methods, previously analyzed data sets were acquired. Two such sets are a mixture of speech and music, and a mixture of two male voices. They are dealt with in sections 4.4.1 and 4.4.2, respectively.

### 4.4.1 Speech in music noise

According Te-Won Lee, see [28], the mixture was acquired as follows: Two source signals, one speech and one pop music, were played through two loudspeakers in a real room. A male speaker saying the words 'one', 'two', 'three', etc. generated the speech signal. The mixture was captured by two microphones. Loudspeakers and microphones were placed in the corners of a $60^2 cm^2$ square. The signal was downsampled from $F_s = 16kHz$ to $F_s = 8kHz$ in order to reduce the computational burden. The filter length and the order of the AR processes were set to $L = 5$ and $p = 2$ (however $p = 5$ for the final source inference stage). The segment length was set to $\tau = 160$. The multistart scheme that was described in section 4.3.1 was also used for this experiment.

The unmixed signals are unavailable and the separation quality can only be judged based on a listening test and the prior knowledge about the signals. For instance, it would be expected that the speech signal is close to zero between the utterances of 'one', 'two' etc.

Figure 4.27 shows the estimated source signals. Obviously, the counting sequence has been extracted from the mixture. A listening test confirms that partial separation has been obtained in that the speech and music now dominate each their channel. Furthermore, spectrograms of the estimated sources are provided in figures 4.28 and 4.29. As a feature of KaBSS, local AR models of the source signals are estimated. These models can be combined into a model spectrogram, see section 4.1. Figures 4.30 and 4.31 show these source model spectrograms. The estimated innovation noise variances for the sources, $q_i^n$, are graphed in figure 4.32 over segments. The active periods of the speech signal are convincingly 'detected'. Evidently, KaBSS could serve as a multichannel voice activity detector (VAD).

The estimated signal path filters can be found in figure 4.33. Some variation was found to exist between runs of the algorithm, although the separation quality remained relatively constant. The filters of the indirect channels exhibit delays of a single sample, approximately corresponding to a difference in

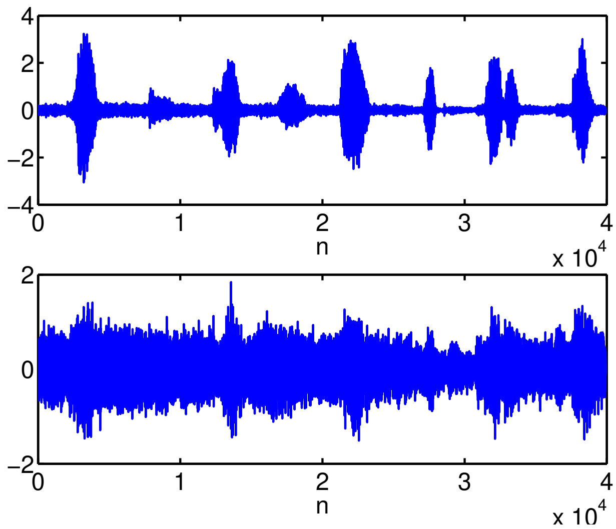


Figure 4.27: The separation of a mixture of speech (top) and music (bottom). The MAP estimates of the source signals. The speech signal is a counting sequence: 'one', 'two', etc. The music signal is pop music.

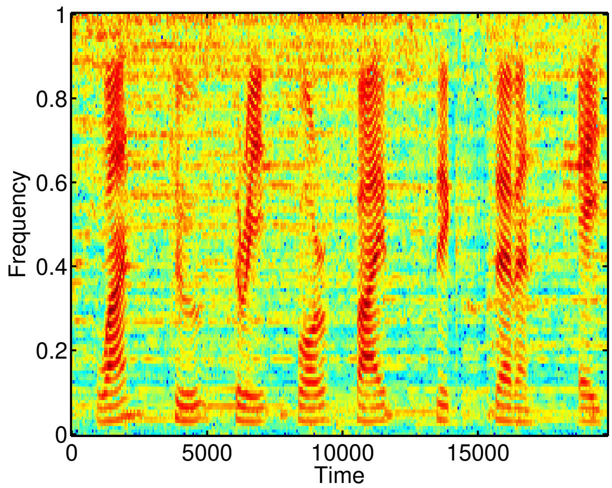


Figure 4.28: Spectrogram (dB) of MAP estimated speech from a real mixture. The features of speech are clearly visible along with remnants of the music. The spectrogram was produced using a Hanning window and overlapping frames.

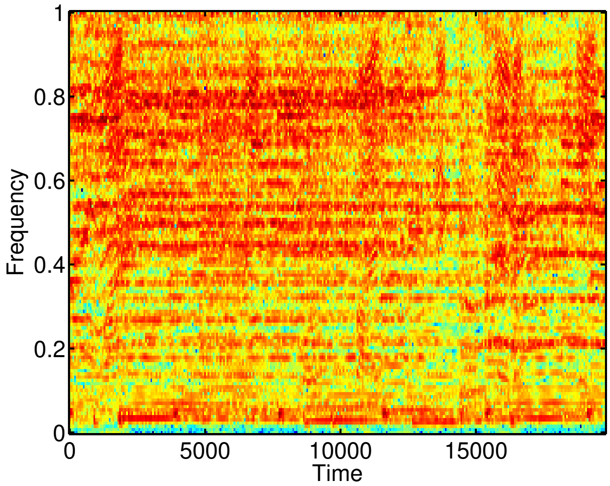


Figure 4.29: Spectrogram (dB) of the music estimate (MAP) resulting from applying KaBSS to a real mixture of speech and music. The sound picture is dominated by the music, although the speech signal has not been completely removed.

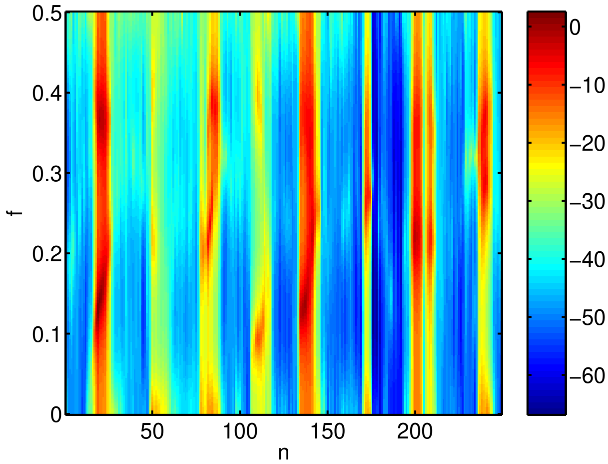


Figure 4.30: The learned model of the speech signal. In each segment, the frequency response of the AR process, $H(f)$, was multiplied with the estimated innovation noise covariance.

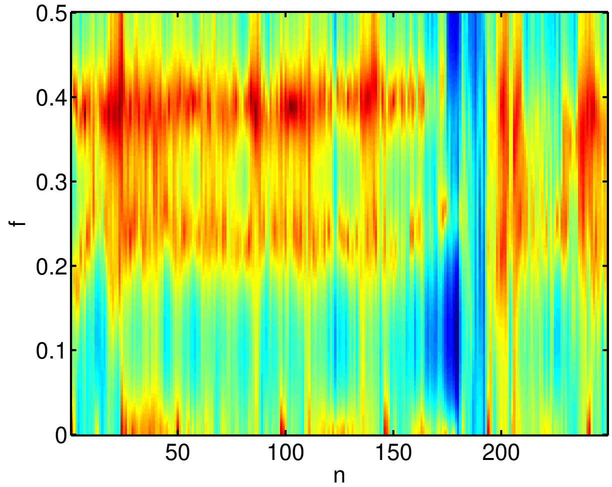


Figure 4.31: The learned model of the music signal.

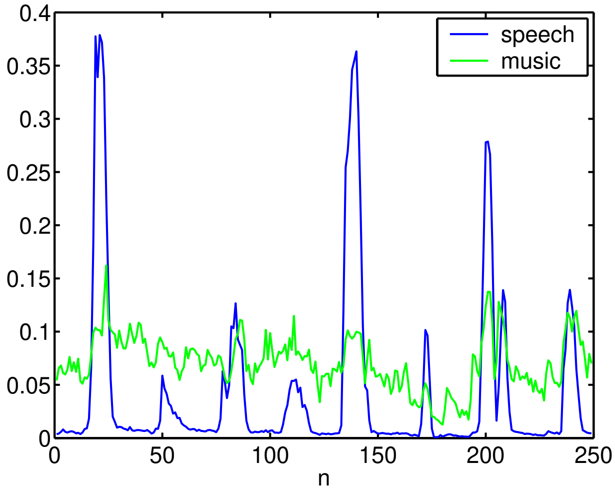


Figure 4.32: The estimated innovation noise variances of sources 1 and 2, $q_i^n$, graphed as a function of segments, $n$. The utterances of 'one', 'two', etc. clearly stand out from the background noise.

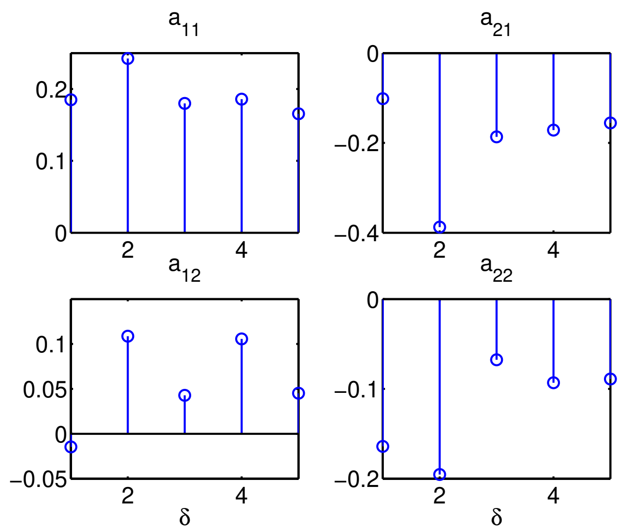


Figure 4.33: The estimated channel filters of the speech/music mixture. $\mathbf{A}_{ij}$ is the filter between source $i$ and sensor $j$. A delay of approximately one sample exists between the direct and indirect channels.

travelled distance of:

$$\begin{aligned}\Delta s &\approx v_{sound}\Delta t \\ &= \frac{300\frac{m}{s}}{8000\frac{1}{s}} \\ &\approx 4cm\end{aligned}$$

where $v_{sound} = 300\frac{m}{s}$ is the speed of sound in office conditions. This distance is certainly shorter than expected, provided the given knowledge of the experiment setup. However, Yellin and Weinstein, who originally provided Te-Won Lee with the data, mention mixtures of speech and music in [29] recorded under circumstances that could have produced a delay of 4cm. A listening test was performed in order to judge the estimated sources. Although the quality of the separation is inferior to that achieved by Te-Won Lee et al. in [7] and by Anemüller et al. in [9] as presented on the authors' web sites, the results are still encouraging further experimentation.

### 4.4.2 Males counting in Spanish and English

This mixture, also made available by Te-Won Lee, consists of two male speakers simultaneously counting in English and Spanish. The speakers and the two microphones that recorded the mixture were arranged in a square, side lengths $60cm$, in a real room. The original signal was downsampled from $F_s = 16kHz$ to $F_s = 8kHz$. KaBSS was invoked with the same hyper parameters as in section 4.4.1. Again, the quality of the separation is slightly inferior to that of Te-Won Lee. However, the Spanish and English utterances clearly dominate one channel each.

## 4.5 The frequency permutation problem

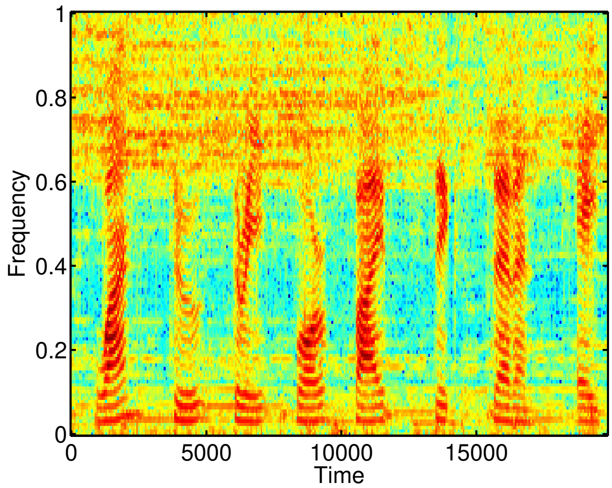


Figure 4.34: Spectrogram displaying the permutation over frequency in source estimate 1. The high frequencies are dominated by the music, whereas the low frequencies essentially contain an estimate of the speech.

During the experimentation with the real data sets of sections 4.4.1 and 4.4.2, the occurrence of the problem of permutation over frequency, mentioned in section 2.4.3, was observed. The phenomenon was primarily noticed for flexible source models, i.e. a high order of the AR processes. This is completely in line with the analysis of section 2.4.3, where it was stated that constraining the source model is mandatory in order to achieve uniqueness of the sources and the parameters within the model. The observation of more data should alleviate the problem, since more spectral variation would probably occur, allowing for more complex source models.

## 4.6 Convergence issues

It was proven that the EM-algorithm (see section 3) converges to a local minimum and that the log-likelihood never decreases in the process. However, the number of iterations required to reach near-convergence is an unknown factor, for some problems it could be very large, see e.g. [26].

In order to speed up the convergence, the adaptive overrelaxed EM (AEM) optimization scheme of section 3.4 was implemented. The hyper-parameter, $\alpha$, determines how fast the step-size, $\eta$, increases. If $\alpha = 1$, then $\eta$ remains 1, which generates a conventional EM step. In other words, $\alpha = 1$ 'disables' AEM.

An experiment was carried out based on an artificially generated convolutive mixture. Different step-sizes, $\alpha$, were tried in order to evaluate the relative importance of AEM. Figure 4.36 shows the learning curves which indicate that standard EM converges critically slower than AEM for any choice of $\alpha$. The advantage of using the AEM scheme comes at a small cost: only a small per-

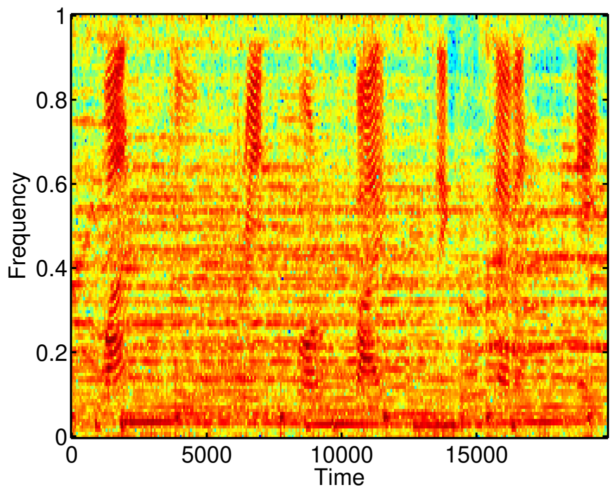


Figure 4.35: Spectrogram displaying the permutation over frequency in source estimate 2. The situation from figure 4.34 reversed.

centage of the updates are rejected, which happens when the update does not lead to an increasing log-likelihood.

In spite of having improved the convergence properties of KaBSS by the adaptive step-size update, non-convergence does still happen. The occurrence of this phenomenon is positively identified when the learning curve never supersedes the log-likelihood of the true parameters, $\mathcal{L}(\theta_{true})$. No local minima exist since $\exp[\mathcal{L}(\theta)]$ is Gaussian, and the convergence problems could be caused by the covariance matrix of $\exp[\mathcal{L}(\theta)]$ being associated with a large eigenvalue spread. Therefore, KaBSS is very much depended on a good starting guess of the parameters. If the true sources are known, the SIR can be used to discriminate between the good and bad solutions. Knowledge of the true sources, however, is not in general the case. Instead, we hypothesize that the good solutions can be chosen based on either: 1) the test log-likelihood, $\mathcal{L}_{test}(\theta)$, 2) the training log-likelihood, $\mathcal{L}_{train}(\theta)$, itself.

In order to assess the problem, an artificial mixture of speech signals was generated along the lines of the experiments in section 4.3 with male and female speech. Signals and filters were selected to be the same. Training segments were repeatedly sampled (randomly) from a training pool of signal segments, and the model was fitted to the data. The observation model was then evaluated on the test segments. As a result, both training and test log-likelihoods were obtained. Subsequently the SIR was estimated, since the original sources are available.

In order to 'verify' hypothesis 1, $\mathcal{L}_{test}(\theta)$ was plotted against the estimated SIR. A clear positive correlation results, indicating that $\mathcal{L}_{test}(\theta)$ does indeed discriminate between bad and good solutions.

Figure 4.38 shows the scatter plot of $\mathcal{L}_{train}(\theta)$ versus $\mathcal{L}_{test}(\theta)$ for different samplings of training segments. The conclusion is that $\mathcal{L}_{train}(\theta)$ and $\mathcal{L}_{test}(\theta)$ are positively correlated in most cases - that the model often generalizes well. An exception occurs when the sampled segments are not representative of the audio recording as a whole. Then, $\mathcal{L}_{train}(\theta)$ and $\mathcal{L}_{test}(\theta)$ may correlate negatively.

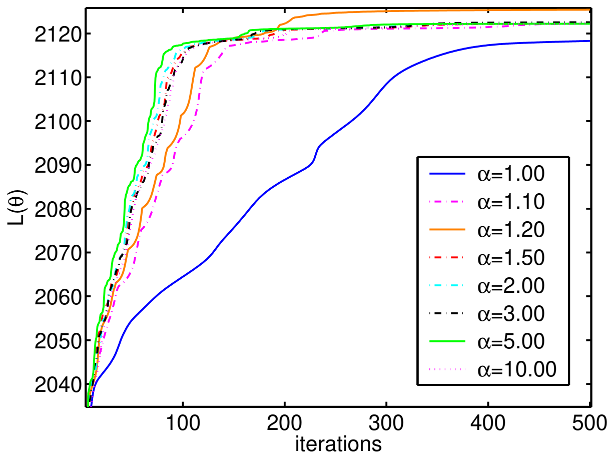


Figure 4.36: Learning curves for the EM ($\alpha = 1$) and AEM estimators in terms of the log-likelihood, $\mathcal{L}(\theta)$. A convolutive mixture of AR(2) processes was used to benchmark the algorithms.

Since limited time is available for running the algorithm, the number of training segments has to be limited, which in turn cause the negative correlation of the training and test log-likelihoods to occur.

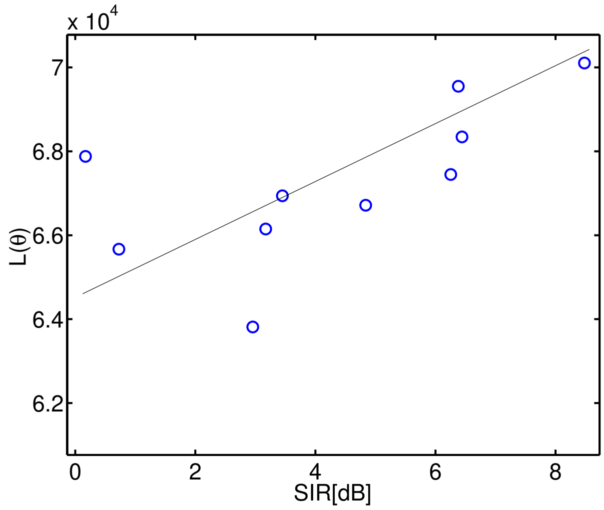


Figure 4.37: The test likelihood, $\mathcal{L}_{test}(\theta)$, plotted against the SIR for a male/male convolutive mixture with added noise.

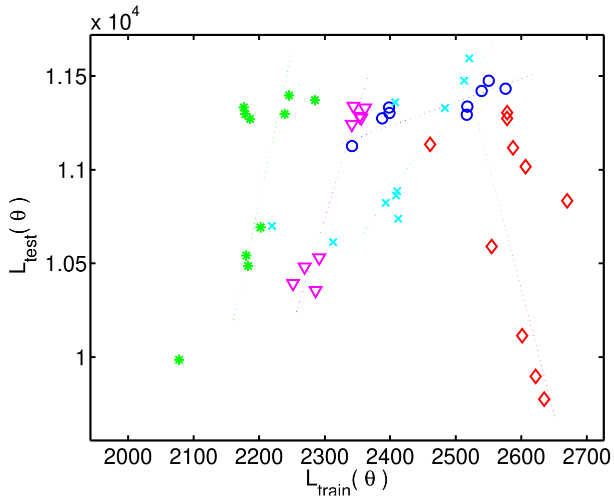


Figure 4.38: Training and test likelihoods for the male/male convolutive problem. Each color corresponds to a new selection of $M_{train} = 10$ training segments. The parameters estimated during the training phase were tested on the same $M_{test} = 100$ test segments.

# Chapter 5

# Discussion

Some efforts were invested into the comparative study of KaBSS and the algorithm of Parra and Spence. A few remarks should be attached to the analysis: only the relative importance of the different a priori assumptions in a given data domain should be judged and *not* which algorithm is the 'better'. Different algorithms are useful for different applications. The experiments show that KaBSS benefits from its noise model and AR source prior when those assumptions are appropriate. In the face of a large data volume, the benefits might vanish.

The computational cost of KaBSS has not been given much attention. However, the separation of 5 seconds of mixture, sampled at 8kHz, took a few hours on a state-of-the-art computer (2.5GHz). The algorithm of Parra and Spence spent in the order of 10 seconds to separate the signals. In order to locate the bottlenecks, Matlab's profiler was invoked. In agreement with theoretical analysis, the critical part of KaBSS was located to the Kalman smoother. The forward-backward recursions perform a matrix inversion which costs $\mathcal{O}([d_S \times L]^3)$. At each iteration of the EM algorithm, this operation is performed for each sample and for each segment. As a result, the computational cost of KaBSS is in the order of $\mathcal{O}(N \times \tau \times [d_S \times L]^3)$, i.e. it scales linearly with data size. This holds provided that the number of required iterations do not increase with the data size. However, long filters cannot be handled well. The algorithm seems to be suitable for under-complete problems with more sensors than sources. Many types of images could be modelled and analyzed in this framework.

## 5.1 Outlook

The author's ideas to improve KaBSS are presented below:

- The applicability of KaBSS in other data domains should be investigated, given the fact that the source model is highly general. As mentioned before, it would be prudent to investigate problems wherein the number of sensors are greater than the number of sources, i.e. undercomplete problems. Image data is a prominent example.

- When a signal is segmented into windows, effects such as spectral leakage and loss of spectral resolution result. Consequently, the spectrum of the signal becomes distorted by these effects. Only the rectangular window function, which does nothing to prevent these issues, was used. A future upgrade of KaBSS should implement a better window function, e.g. the Hanning window, which alleviates the spectral leakage problem.

- In many applications, it is desired that the channel filter models nothing but a single delay and attenuation. The current algorithm estimates as $L$ filter coefficients, when it might have been more appropriate to estimate only a few parameters of a flexible channel filter model.

- The high-level description of the assumed 'non-stationarity' is another model amendment. Hidden Markov models are often used to model the time-variance of speech, and could potentially be used to explain the transitions between the switching AR models.

- The log-likelihood of the parameters is computed in a forward recursive fashion. It is possible that its gradient with respect to the parameters can also be computed recursively. The obtained gradient could then be used in a gradient-based optimizer. A literature study and/or theoretical analysis will answer this question.

- Provided the above recursive gradient could be computed, a stochastic gradient algorithm in line with LMS could be implemented for real-time applications.

- Attention was early in the project diverted from instantaneous mixture problems towards the more challenging convolutive mixtures. Preliminary experimentation with KaBSS in the 'instantaneous' mode suggested that KaBSS could serve well as probabilistic extension of the decorrelation algorithm of Molgedey and Schuster, [3].

- The noise regularization scheme is as of now a heuristic at work. Future work should advance theoretical understanding.

- Minor code/numerical issues remain. In particular, the simultaneous setting of $\alpha \neq 1$ and estimation of $\mu$ and $\mathbf{\Sigma}$ has proved unstable, eventually causing the likelihood to decrease. Therefore, the estimation of $\mu$ and $\mathbf{\Sigma}$ was turned off during the experiments.

## 5.2 Conclusion

The Kalman Blind Source Separator (KaBSS) for convolutive mixtures of signals is obviously the principal result of this work, see appendix C. It separates a number of benchmark mixtures of speech that were measured in real rooms. Furthermore, a comparative study was carried out on various artificial noisy mixtures of speech signals. It indicated that KaBSS is useful in bad noise conditions. Benefits of being probabilistic were reaped, such as determining the number of sources from the log-likelihood of the parameters. The Bayes Information Criterion was employed for this purpose.

KaBSS emerged from the probabilistic formulation of the algorithm of Parra and Spence, [4]. An expectation-maximization (EM) scheme was derived for the estimation of the parameters. The actual estimators adhere to the independency of the sources.

Also, the conditions under which the algorithm does work and does not work were investigated both in terms of theoretical work and empirical study. For instance, it was found out that the parameters of a *sum* of AR(1) processes are unique up to scaling and permutation. For this to hold, the sources need to have different autocorrelations. It was also argued that the sources need to be wide-sense non-stationary and model-constrained. Empirical verification followed from the experiments. More, the Monte Carlo runs and experiments made it clear that the algorithm exhibits poor convergence properties in noise-free conditions. However, experiments showed that these situations could be handled by using regularization noise.

Finally, new ideas were presented that could potentially turn into implementable innovations.

# Bibliography


[1] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[2] J. Cardoso, "Blind signal separation: statistical principles," *IEEE, Blind identification and estimation*, 1998.

[3] L. Molgedey and G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

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

[5] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary sources," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[6] B. S. Krongold and D. L. Jones, "Blind source separation of nonstationary convolutively mixed signals," in *Proceedings of the 10th IEEE SSAP Workshop*, 2000, pp. 53–57.

[7] T.W. Lee, A. J. Bell, and R. H. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. 1997, vol. 9, p. 758, The MIT Press.

[8] 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.

[9] B. Kollmeier J. Anemüller, "Amplitude modulation decorrelation for convolutive blind source separation," in *Second international workshop on independent component analysis and blind signal separation*, 2000, pp. 215–220.

[10] R. K. Olsson and L. K. Hansen, "Probabilistic deconvolution of nonstationary sources," in *European Signal Processing Conference (EUSIPCO)*, 2004, submitted.

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

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

[13] G. Doblinger, "An adaptive Kalman filter for the enhancement of noisy AR signals," in *IEEE Int. Symp. on Circuits and Systems*, 1998, vol. 5, pp. 305–308.

[14] J. Tugnait Y. Hua, "Blind identifiability of fir-mimo systems with colored input using second order statistics," in *IEEE Signal Process. Lett.*, 2000, vol. 7, pp. 348–350.

[15] M. Kawamoto and Y. Inouye, "Blind deconvolution of MIMO-FIR systems with colored inputs using second-order statistics," *IEICE trans. fundamentals*, vol. 3, no. E86-A, 3 2003.

[16] E. Wan and A. Nelson, "Neural dual extended kalman filtering: applications in speech enhancement and monaural blind signal separation," in *IEEE Neural Networks for Signal Processing Workshop*, 1997.

[17] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *IEEE*, 1989, vol. 77.

[18] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, Prentice Hall, 1993.

[19] P. Kidmose, *Blind separation of heavy tail signals*, Ph.D. thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, 2001.

[20] National Center for Voice and Speech, "www.ncvs.org," .

[21] J. G. Proakis and D. G. Manolakis, *Digital signal processing; principles, algorithms and applications*, Prentice Hall, 1996.

[22] M. Welling, "Classnotes: The Kalman filter," 2000.

[23] C. Bishop, *Neural networks for pattern recognition*, Oxford University Press, 1995.

[24] S. Roweis, "Matrix identities," 1999.

[25] P. A. d. F. R. Højen-Sørensen, Ole Winther, and Lars Kai Hansen, "Analysis of functional neuroimages using ica with adaptive binary sources," *Neurocomputing*, , no. 49, 2002.

[26] R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani, "Optimization with EM and Expectation-Conjugate-Gradient," in *International Conference on Machine Learning*, 2003, vol. 20, pp. 672–679.

[27] Sam T. Roweis, "One microphone source separation," in *NIPS*, 2000, pp. 793–799.

[28] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," 1998.

[29] D. Yellin and E. Weinstein, "Multichannel signal separation: Methods and analysis," in *IEEE Transactions on signal processing*, 1996, vol. 44.

[30] S.M. Kay, *Statistical signal processing*, Prentice Hall, 1993.

# Appendix A

# Quality measures

In the following, it will be discussed how to evaluate the inferred sources against the known true sources. In broad terms, we are interested in how closely the estimates approximate the true sources. The computation of the mean square error (MSE) is one naive approach that would fail, because the convolution and deconvolution processes inadvertently may cause the estimate to be a time-shifted version of the original. In this case, a speech signal and its time-shifted replica may produce a high MSE, while a listening test will not reveal any difference.

The block diagram in figure A.1 represents the total system that takes the original sources as inputs and outputs the source estimates. In between, the sources are mixed, exposed to observation noise and processed by a separation system. The total system, $\mathbf{H}(f)$, is a time varying filter that characterizes the various paths of the source signals as they are transformed into the final estimates.

Viewing the mixing and demixing systems as a one, it is interesting to ask which fraction of signal power went the right way from source signal to source signal estimate as opposed to the signal parts that cross over and corrupt the estimates of other signal sources. In other words, we want to quantify the amount of *cross-talk* defined as the ratio between the power in the *direct channels* and the power in the *cross channels*.

$$SIR = \frac{\sum_{k\in K} P_k}{\sum_{m\in M} P_m}$$

where $K$ and $M$ denote the sets of direct and cross signal paths, respectively, and $P_k$, $P_m$ the powers of the signals contribution to the estimates. It can be written as:

$$SIR = \frac{\sum_{k\in K}\sum_{\omega} |H_k(\omega)|^2 |S_k(\omega)|^2}{\sum_{m\in M}\sum_{\omega} |H_m(\omega)|^2 |S_m(\omega)|^2}$$

When the BSS algorithm is a backward system, i.e. $\mathbf{W}$ is estimated and the sources are inferred by the filtering of $\mathbf{x}_t$ through $\mathbf{W}$, the total system is readily available as $\mathbf{H}(\omega) = \mathbf{A}(\omega)\mathbf{W}(\omega)$. The present algorithm, however, is a forward system and no $\mathbf{W}(\omega)$ is estimated.

Under the simplifying assumption that $\mathbf{H}(\omega)$ is a pure-delay filter, the SIR can be computed as:

$$SIR = \frac{\sum_{k \in K} \max_{\delta} r_k(\delta)}{\sum_{m \in M} \max_{\delta'} r_m(\delta')} \tag{A.1}$$

where the unbiased estimate of the normalized cross-correlation of the signal attributed to channel $i$ is used:

$$\begin{aligned} r_i(\delta) &= \frac{1}{T \cdot \alpha} \sum_{\tau} s_{i,\tau} s_{i,\tau-\delta} \\ \alpha &= \sqrt{P_s \cdot P_{\hat{s}}} \end{aligned}$$

The normalization by $\alpha$ is obtained by estimating the powers of the original and estimated sources. Although only strictly correct when $H(f)$ is a delay-only filter, the expression given in equation A.1 remains a good approximation for high SNR. A full linear approach to the estimation of the channel powers would include the system identification of $H(f)$ by e.g. least squares.

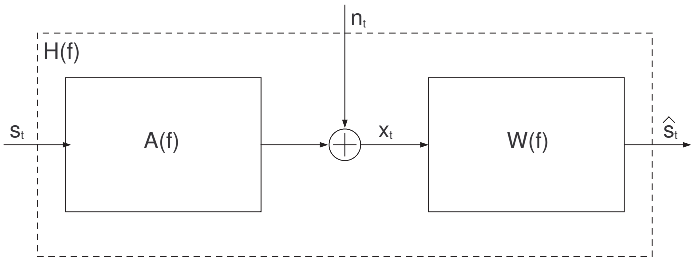


Figure A.1: A blind source separation model.

# Appendix B

# Source code and data

The supplied CD-ROM, which can be obtained from the author, contains source code and data. The content is:

### Matlab

- `AR2`: Analysis of AR(2) random processes.
- `BIC`: Determination of the number of sources in an artificial mixture.
- `generate_conv_mix`: Generation of a mixture from two sources.
- `KaBSS`: The algorithm and test scripts.
- `molgedey`: The author's implementation of the Molgedey-Schuster decorrelation algorithm.
- `monaural_demo`: Demonstration of the difficulty of monaural ICA.
- `parra`: Parra and Spence's algorithm and test scripts.
- `parra_limits`: The fitting of the Parra-Spence algorithm to a mixture.

### Results

The `.mat` files of the experiments:

- `BIC`
- `male_female`
- `SNR`
- `spanish_english`
- `speech_music`

# Appendix C

# Publication

The following pages contain the part of the work that was submitted for publication, see [10].

# 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

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

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

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

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

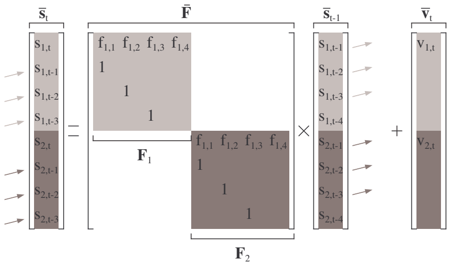

Figure 1: The AR(4) source signal model. The memory of $\mathbf{s}_t$ is updated by discarding $s_{i,t-4}$ and composing new $\mathbf{s}_{1,t}$ and $\mathbf{s}_{2,t}$ using the AR recursion. Blanks signify zeros.

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

$$\bar{\mathbf{s}}_t = \begin{bmatrix} \mathbf{s}_{1,t}^T & \mathbf{s}_{2,t}^T & .. & \mathbf{s}_{d_s,t}^T \end{bmatrix}^T$$

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

$$\mathbf{s}_{i,t} = \begin{bmatrix} s_{i,t} & s_{i,t-1} & .. & s_{i,t-p+1} \end{bmatrix}^T$$

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:

$$\begin{aligned}
\bar{\mathbf{F}} &= \begin{bmatrix} \bar{\mathbf{F}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{F}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{F}}_L \end{bmatrix} \\
\bar{\mathbf{F}}_i &= \begin{bmatrix} f_{i,1} & f_{i,2} & \cdots & f_{i,p-1} & f_{i,p} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \\
\bar{\mathbf{Q}} &= \begin{bmatrix} \bar{\mathbf{Q}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{Q}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{Q}}_L \end{bmatrix} \\
(\bar{\mathbf{Q}}_i)_{jj'} &= \begin{cases} q_i & j = j' = 1 \\ 0 & j \neq 1 \bigvee j' \neq 1 \end{cases}
\end{aligned}$$

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,

$$\bar{\bar{\mathbf{A}}} = \begin{bmatrix} \mathbf{a}_{11}^T & \mathbf{a}_{12}^T & .. & \mathbf{a}_{1d_s}^T \\ \mathbf{a}_{21}^T & \mathbf{a}_{22}^T & .. & \mathbf{a}_{2d_s}^T \\ \mathbf{a}_{d_x1}^T & \mathbf{a}_{d_x2}^T & .. & \mathbf{a}_{d_xd_s}^T \end{bmatrix}$$

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

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

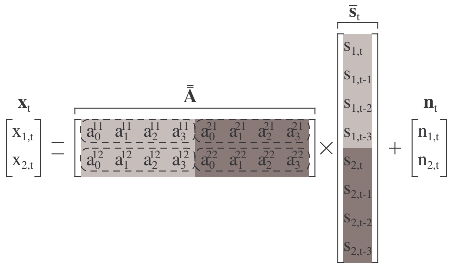

Figure 2: The convolutive mixing model requires a full $\bar{\bar{\mathbf{A}}}$ to be estimated.

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 $\theta$

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

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. $\theta$. 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

$$\begin{aligned} \mathcal{J}(\theta, \hat{p}) = -\frac{1}{2} \sum_{n=1}^{N} [ \sum_{i=1}^{d_s} \log \det \mathbf{\Sigma}^n_i + (\tau - 1) \sum_{i=1}^{d_s} \log q^n_i \\ + \tau \log \det \mathbf{R} + \sum_{i=1}^{d_s} \langle (\mathbf{s}^n_{i,1} - \mu^n_i)^T (\mathbf{\Sigma}^n_i)^{-1} (\mathbf{s}^n_{i,1} - \mu^n_i) \rangle \\ + \sum_{t=2}^{\tau} \sum_{i=1}^{d_s} \langle \frac{1}{q^n_i} (s^n_{i,t} - (\mathbf{f}^n_i)^T \mathbf{s}^n_{i,t-1})^2 \rangle \\ + \sum_{t=1}^{\tau} \langle (\mathbf{x}^n_t - \bar{\bar{\mathbf{A}}} \bar{\mathbf{s}}^n_t)^T \mathbf{R}^{-1} (\mathbf{x}^n_t - \bar{\bar{\mathbf{A}}} \bar{\mathbf{s}}^n_t) \rangle ] \end{aligned}$$

where $\mathbf{f}^T_i = [\ f_{i,1} \quad f_{i,2} \quad .. \quad f_{i,p}\ ]$. 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:

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

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:

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

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, $\alpha$, 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 $i$th original source to the $j$th 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.

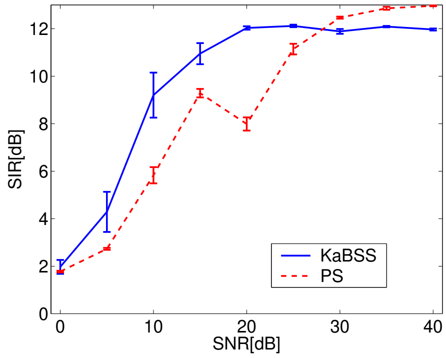


Figure 3: The separation performance for varying SNR of KaBSS and the reference method proposed by Parra and Spence (PS) [5]. The signals are two utterances by the same speaker. Two convolutive mixtures were created with variable strength additive white noise. The SIR measures the crosstalk between the 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.