MIMO-AR BLIND SOURCE SEPARATION FOR GMM-DISTRIBUTED AND FINITE ALPHABET SIGNALS

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE M.Sc. DEGREE

Composed by:

Tirza Routtenberg

Supervised by: Dr. Joseph Tabrikian

December 2007
Abstract

In this work, two methods for blind separation of independent sources from the output of a convolutive multi-dimensional system are presented. The proposed methods assume a multiple-input multiple-output (MIMO) system, in which a multi-dimensional auto-regressive (AR) model relates the input and the output signals. The proposed algorithms exploit non-Gaussianity of the independent sources by modeling their distribution using the Gaussian mixture model (GMM). The first method considers a more general model for the input signals and it can be applied to problems with arbitrary input signals, such as speech, bio-medical and communication signals. In the second method, additional prior information regarding the input signals statistics is assumed. This method is useful for MIMO communication channel estimation where the input signal constellation is known. In these methods, the sensors distribution parameters and the separation matrix are estimated via the expectation-maximization (EM) and the generalized EM (GEM) algorithms for GMM parameter estimation.

The resulted solution is an extension of an existing technique for single-input single-output (SISO) channel estimation. In addition, for Gaussian-distributed sources, the solution for state transition matrix estimation reduces to the well known Yule-Walker equations for MIMO-AR models. For the noiseless scenario, the second proposed method generalizes an existing blind method for flat-fading channel estimation with known finite alphabet (FA).

The performances of the two convolutive blind source separation (BSS) problems are evaluated and compared to existing BSS techniques, via simulations of synthetic, audio and communication signals. The results show good performances of the proposed meth-
ods in terms of signal-to-interference ratio (SIR), channel impulse response, and mean square error (MSE) of the multi-dimensional AR parameters estimate. It is demonstrated that the proposed algorithm outperforms the well-known multi-dimensional Yule-Walker equations for AR parameter estimation in terms of MSE. In the considered simulations for the second method, it is shown that the obtained symbol error rate (SER) is very close to the SER of the optimal algorithm which assumes perfect channel state information (CSI).
Acknowledgment

I am deeply indebted to my advisor, Dr. Joseph Tabrikian, for his guidance, instruction, assistance, and especially his patience and constant support. Without his help, this work would not be possible.
# Contents

1 Introduction  
1.1 Literature survey  
1.1.1 Blind source separation  
1.1.2 MIMO system identification  
1.1.3 MIMO channel estimation in communication systems  
1.2 Objective  
1.3 Brief summary of the EM and GEM algorithms  
1.4 Thesis layout  

2 MIMO-AR system identification and BSS  
2.1 Problem Formulation  
2.1.1 The MIMO-AR model  
2.1.2 Source distribution model  
2.1.3 Sensors distribution model  
2.2 Unstructured input  
2.3 Structured input  
2.4 Discussion and conclusion
A Mathematical proofs

A.1 Minimization of (2.17) w.r.t. $\{\pi_m\}_{m=1}^M$ ........................................... 64

A.2 Minimization of (2.18) w.r.t. $\{\eta_m\}_{m=1}^M$ ........................................... 65

A.3 Minimization of (2.18) w.r.t. $\{R_m\}_{m=1}^M$ ........................................... 65

A.4 Derivative of (2.18) w.r.t. $A$ .................................................. 66

A.5 Minimization of (2.31) w.r.t. $\{C_m\}_{m=1}^M$ ........................................... 67

A.6 Minimization of $Q_2^{(i+1,j+1)}$ w.r.t. $B$ ........................................... 68

A.7 The partial derivative of (3.8) w.r.t. $A$ .................................................. 70

A.8 The partial derivative of (3.8) w.r.t. $H$ .................................................. 71

A.9 Solution of the system of equations (3.12) - (3.13) ......... 72

B Stability of the MIMO-AR model

B.1 Stability of first-order MIMO-AR model ........................................ 75

B.2 Stability of a $P$-order MIMO-AR model ........................................ 76

C Matrix derivatives

C.1 Definitions .......................................................... 78

C.1.1 Real derivatives .......................................................... 78

C.1.2 Complex derivatives .......................................................... 79

C.2 Useful derivatives .......................................................... 79

D Greedy learning of GMM ........................................ 82
Abbreviations

AIC - Akaike information criterion
AR - Auto-regressive
ARMA - Auto-regressive moving-average
BIC - Bayesian information criterion
BPSK - Binary phase-shift keying
BSS - Blind source separation
CA - Coordinate ascent
CDMA - Code division multiple access
CLD - Complete log-likelihood
CIR - Channel impulse response
CSI - Channel state information
EM - Expectation-maximization algorithm
FA - Finite alphabet
FIR - Finite impulse response
GEM - Generalized EM algorithm
GMM - Gaussian mixture model
HMM - Hidden Markov model
HOS - Higher order statistics
i.i.d. - Independent and identically distributed
ICA Independent component analysis
IIR- Infinite impulse response
ISI - Inter-symbol interference
KL - Kullback-Leibler
LPC - Linear predictive coding
LTI - Linear time-invariant
MA - Moving average
MDL - Minimum description length
MIMO - Multiple-input multiple-output
ML - Maximum likelihood
MSE - Mean-square error
PDF - Probability density function
PSK - Phase-shift keying
QPSK - Quadrature phase-shift keying
RMSE - Root mean-square error
SER - Symbol error rate
SIMO - Single-input multiple-output
SIR - Signal-to-interference ratio
SISO - Single-input single-output
SNR - Signal-to-noise ratio
SVD - Singular value decomposition
WSS - Wide sense stationary
Notations

\((\cdot)^{-1}\) - Matrix inverse
\((\cdot)^T\) - Matrix transpose
\((\cdot)^*\) - Conjugate
\((\cdot)^H\) - Conjugate transpose (Hermitian)
\((\cdot)^{-H}\) - Inverse Hermitian
\(E(\cdot)\) - Expectation
\(\text{DIAG}(\cdot)\) - Diagonal matrix with the same diagonal as its matrix argument
\(|\cdot|\) - Absolute value of a scalar or determinant of a matrix
\(||\cdot||\) - Frobenius norm
\(\Re\) - Real part of a complex number
\(\Im\) - Imaginary part of a complex number
\(\mathbf{I}_L\) - Identity matrix of size \(L\)
\(A_{ij}\) - \((i,j)\)th element of matrix \(A\)
\(n\) - Discrete time index
\(t\) - Continuous time index
\(\boldsymbol{\theta}\) - Set of unknown parameters
\(\hat{\boldsymbol{\theta}}\) - Estimation of \(\boldsymbol{\theta}\)
\(\Omega\) - The unknown parameters space
\(L\) - Number of sources/sensors
\(N\) - Number of samples
\(P\) - MIMO-AR model order
\(A\) - \(L \times LP\) state transition matrix
\( \mathbf{H} \) - \( L \times L \) input mixing matrix

\( \mathbf{B} \) - \( L \times L \) separation matrix

\( \mathbf{b}_l^T \) - The \( l \)th row of the matrix \( \mathbf{B} \)

\( \mathbf{s}_n \) - Source vector at time instance \( n \)

\( \mathbf{x}_n \) - Observation vector at time instance \( n \)

\( x_{n,i} \) - Observation at the at the \( i \)th sensor and time instance \( n \)

\( \mathbf{x}_n^P \) - Past samples vector at time instance \( n \)

\( \mathbf{y}_n \) - The hidden indication vector for time instance \( n \)

\( f_\mathbf{s}(\cdot) \) - Probability density function of the vector \( \mathbf{s} \)

\( f_{\mathbf{x}_n^P}(\mathbf{x}_n^P) \) - The initial conditions probability density function

\( f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \) - Conditional probability density function of \( \mathbf{x} \) given \( \mathbf{y} \)

\( Q_l \) - FA set size of the \( l \)th source

\( \{\zeta_{qi}\}_{q=1}^{Q_l} \) - The FA set for the \( l \)th source
List of Figures

5.1 Performance of the state transition matrix estimation by the structured-GEM and Yule-Walker algorithms (upper), performance of the input mixing matrix estimation by the structured-GEM and “Yule-Walker+FastICA” method (middle), and the corresponding BSS (lower) for GMM-distributed sources. ............................................. 48

5.2 Impulse response of the MIMO-AR system, from source 1 to sensor 1, compared to the estimated response by the structured algorithm for Gaussian-distributed sources and $N = 1000$ samples. ................................. 49

5.3 Scatter plot of the source signals (upper), the mixed signals (middle), and the estimated source signals (lower), for Gaussian-distributed sources with $N = 1000$ samples. .................................................... 50

5.4 Performance of the state transition matrix estimation of the unstructured MIMO-AR-GMM and the multi-dimensional Yule-Walker algorithms with 3 GMM-distributed sources and 2 sensors. ........................................ 51
5.5 Performance of the state transition matrix estimation by the structured-GEM and Yule-Walker algorithms (upper), performance of the input mixing matrix estimation by the structured-GEM and “Yule-Walker+FastICA” algorithms (middle), and the corresponding BSS (lower), for Laplacian-distributed sources. ............................... 52

5.6 Impulse response of the MIMO-AR system, from source 1 to sensor 1, compared to the estimated response by the structured algorithm for Laplacian-distributed sources and $N = 1000$ samples. ............................... 53

5.7 Scatter plot of the source signals (upper), the mixed signals (middle), and the estimated source signals (lower), for Laplacian-distributed sources with $N = 1000$ samples. ............................... 54

5.8 Convergence of the log-likelihood for simulation of 2 GMM-distributed (left) and Laplacian-distributed (right) sources, for $N = 500, 1000, 5000$ samples. ............................... 55

5.9 The MIMO-AR system with speech input signals. ............................... 56

5.10 SER vs. SNR obtained with the proposed channel estimation technique compared with perfect CSI for QPSK and BPSK source signals. ............................... 57

5.11 Scatter plot of the two sources, the mixture in the first sensor, and the equalized signals, for SNR=12dB. ............................... 58

5.12 SER vs. data length obtained with the proposed channel estimation technique compared with perfect CSI for QPSK (higher) and BPSK (lower) source signals, for SNR=8dB. ............................... 59
5.13 Impulse response of the MIMO-AR system, from source 1 to sensor 1, compared to the estimated response by the proposed EM algorithm for two QPSK and BPSK modulated sources and $N = 5000$ samples.
Index terms

MIMO-AR, MIMO system identification, BSS, EM, GEM, convolutive mixtures, GMM, finite-alphabet, maximum-likelihood, blind deconvolution, blind system identification
Chapter 1

Introduction

1.1 Literature survey

1.1.1 Blind source separation

Blind source separation (BSS) has been intensively investigated in the literature in the recent two decades. This problem is important for several applications, such as, communications, biomedical engineering, and blind audio source separation. In general, the sources are assumed to be independent or at least decorrelated. Classical research in this field has intensively dealt with blind separation of instantaneous mixtures where the observed signal samples are obtained by a linear combination of the source samples [1]-[3]. Some known methods for an instantaneous BSS problem are cumulant-based methods, like JADE [2] and FastICA [3]. The FastICA algorithm is based on a fixed-point iteration scheme maximizing the non-Gaussianity of the estimated source signals.

A more challenging problem is multiple-input multiple-output (MIMO) system identification and blind separation of convolutive mixtures where the mixing matrix is frequency-
dependent. There are two common approaches for handling the convolutive BSS problem [4]: frequency-domain [5], [6], and time-domain [7]-[11]. In the frequency-domain approach, the sensor signal model at each frequency bin is similar to the instantaneous mixture case. Therefore, instantaneous BSS techniques can be independently applied for each frequency bin. The main disadvantage of this approach is the frequency permutations problem [6].

Some of the methods in the time-domain approach are based on directly extending instantaneous BSS methods to the convolutive case. For example, in [7], Amari et al. extended the natural gradient search method to derive a set of on-line algorithms for combined multichannel blind source separation and time-domain deconvolution of convolved signal mixtures. An additional natural gradient-based algorithm for convolutive BSS is proposed in [8]. At every iteration of this algorithm, only a subset of the coefficients in the separation system is updated. It is demonstrated that the partial update scheme for a single channel can be extended to the multichannel time-domain convolutive BSS with a slight performance degradation and unchanged computational complexity. Other convolutive BSS methods that extend instantaneous ones are based on minimization of an objective function based on some independence measures. Methods in this category are, for example, the minimal distortion principle [9], minimization of Kullback-Leibler (KL) divergence [10], joint-diagonalization [11] and mutual information based algorithms [12]. The time-domain approaches do not suffer from the permutation problem; however, they are usually not able to achieve good separation for long mixing channels and they have a low convergence rate.

Several research studies have utilized the GMM in solving the BSS problem. For example, Moulines et al. [13] developed an approximate maximum likelihood (ML) method
for instantaneous and convolutive BSS, where the density of each source was modeled by a univariate GMM. The method for the instantaneous case is based on an EM algorithm, which jointly estimates the mixing matrix, the sources distribution parameters and the noise covariance matrix of the additive noise. For the convolutive BSS problem, the algorithm is based on the concept of a hidden Markov model (HMM) estimation. The GMM distribution combined with the auto-regressive (AR) model for the source signals has been used for BSS of instantaneous mixtures [14], where each source is assumed to be a GMM-distributed AR process generated from a temporally independent and identically distributed (i.i.d.) process. In [15], the AR model has been used to separate sources from a multi-dimensional finite impulse response (FIR) mixture.

1.1.2 MIMO system identification

The goal of blind MIMO system identification is to identify the system driven by unobservable inputs based on its outputs. The major part of the research in the literature is involved with identifying single-input single-output (SISO) systems [16] and single-input-multiple-output (SIMO) systems [17]. In most of the recent years research on MIMO systems, the unknown system is assumed to be linear time-invariant (LTI) and restricted to a parametric model, and the unobservable inputs are assumed to be i.i.d. sequences [18]-[19]. For this problem, a wide range of methods have been developed.

Pham and Tong [20] assumed a noiseless multichannel AR model, and offered two iterative procedures for the ML estimator of the unknown parameters, while guaranteeing a stable solution. Hasan and Yahagi [21] proposed a method for the identification of multichannel AR processes contaminated by additive white observation noise, in which a noise-compensated multichannel Yule-Walker method is implemented by two iterative
methods. Gorokhov and Loubaton [22] addressed the blind MIMO-FIR identification problem in the case where the number of inputs is less than the number of outputs. Their method is based on the linear prediction approach, introduced in [23], while the unknown parameters are estimated using the weighted least-squares approach.

Hasan et al. [24] presented a multichannel AR estimation method using a finite set of noisy observations without \textit{a-priori} knowledge about the noise power. Their method is based on solving alternatively a set of nonlinear equations using the iterative Newton-Raphson algorithm in order to estimate the noise variance, and a set of Yule-Walker linear equations in order to estimate the AR parameters.

In some research studies, MIMO problems have been solved by extension of SIMO or SISO approaches, for example, methods based on higher order statistics (HOS) [25], [26]. In [25], third-order cumulants are employed for estimating a multichannel AR model of known orders. Inouye and Hirano [26] addressed the blind identification problem of linear MIMO systems driven by unobservable colored inputs using HOS, particularly the fourth-order cumulants of the outputs, where the unobservable inputs are mutually and temporally colored linear processes.

1.1.3 MIMO channel estimation in communication systems

In many applications involving transmission and reception of a signal simultaneous, estimation of the channel and the transmitted signal are required. Optimal communication receivers should be designed using the channel state information (CSI). Therefore, channel estimation and equalization in a communication system is an important problem.

Depending on the delay spread and the data rate, the channel may be approximately flat fading [27], or frequency-selective [28], [29]. The AR model has been used in various
ways as the basis for the channel models, where the MIMO channel variations is modeled by an AR process. More recently, the Kalman filter [30] and the extended Kalman filter (EKF) [31] have been applied for the channel estimation problem.

MIMO systems in wireless communications have been attracted increasing attention due to their promising capacity gains [32]. MIMO techniques have the advantage of providing high data rate at no extra bandwidth expansion or power consumption. Moreover, the MIMO channel equalizer is not only responsible for removing inter-symbol interference (ISI), but also for removing multiple-access interference (MAI). In order to accomplish this double duty, the equalizer often requires perfect CSI.

Communication channel estimation and tracking has received significant attention over the past two decades. Review on the SISO blind channel estimation problem and methods can be found in [33]. Various algorithms have been considered for MIMO channel estimation. These algorithms can be divided into three categories: training based [34], blind [35] and semi-blind [29], [36], [37], algorithms. Training-based algorithms estimate the channel using knowledge of the symbols in the transmitted signals. Even though these methods are popular in SISO digital communication systems, in MIMO systems a large amount of training symbols is required, so the system bandwidth efficiency is reduced [38]. Blind algorithms estimate the channel based on the properties of the transmitted signals, such as finite alphabet (FA) properties and cyclo-stationarity [39]. Semi-blind channel estimation algorithms provide better performance than the blind algorithms while requiring fewer known symbols than training-based channel estimation algorithms.

In several papers, a wide range of the EM algorithm implementations have been considered for channel estimation and equalization. The EM algorithm has been used for channel estimation [27], [40] and for channel equalization [41], [42]. The EM algorithm
has been used to iteratively find the maximum \textit{a-posteriori} probability (MAP) \[43\], or the ML \[28\] estimators of the channel response.

The constant modulus property of communication signals is widely used for channel estimation and equalization. A related but distinct property is the FA in digital signals. In many channel estimation and signal detection problems, the source symbols are taken from a FA. FA signal estimation arises, for example, in digital communication, in deoxyribonucleic acid (DNA) sequence detection, or in state estimation problems. The FA property of the transmitted signal has been shown to be very useful in channel estimation and blind equalization. A single user FA approach was first proposed in \[44\]. A multiuser detection algorithm assuming channel knowledge had already been proposed with known FA \[45\], \[46\] or unknown FA \[47\]. Belouchrani \[45\] proposed a ML approach to separate i.i.d. sources in an instantaneous mixture with additive Gaussian noise. Maximization of the likelihood function is carried out using the EM algorithm. Two versions of this approach were presented: ML separation and stochastic ML separation. By exploiting the FA property, the scalar ambiguity (inherent in all blind algorithms) is restricted to a unit amplitude and phase values belonging to a finite set, which can be easily resolved. An algorithm for FA signal estimation in flat-fading channels was proposed in \[48\].

1.2 Objective

In this study, the problems of BSS and system identification of MIMO-AR system under the GMM-distributed sources assumption is investigated. There are several works on system identification in MIMO-AR with Gaussian-distributed sources, based on the linear predictive coding (LPC). The novelty of this work is that the sources are assumed to be
GMM-distributed. Moreover, the non-Gaussianity of the sources enables to separate them. In most convolutive BSS problems, the algorithm is assumed to be a FIR type. The MIMO-AR system has an infinite impulse response (IIR) and therefore the separation system can be implemented by FIR system with finite number of coefficients. In the second part of this work, a new algorithm for frequency-selective fading channel estimation with known FA is proposed.

This model can approximate many linear systems with non-Gaussian distributed sources, because with a sufficiently high order, the AR model can approximate moving-average (MA) or auto-regressive moving-average (ARMA) models. In addition, in [49] it is shown that any density can be estimated to any desired degree of approximation, in terms of KL divergence [50], using a finite order GMM. To avoid the inherent permutation problem in frequency-domain methods, the time-domain approach for the convolutive BSS problem is adopted in this work. The disadvantage of the methods presented in this work is their computational complexity, as in most time-domain convolutive BSS techniques.

In Chapter 2, a multichannel AR mixing model is considered, where the sources are assumed to be i.i.d. random processes. We consider both the system identification and the source separation problems. The MIMO-AR parameters and the GMM parameters are jointly estimated via a ML based estimator, implemented by using the generalized expectation-maximization (GEM) method. The proposed algorithm is a generalization of the SISO system identification problem presented in [51]. It is also an extension of the Yule-Walker equations to GMM.

In Chapter 3, we consider MIMO frequency-selective communication systems, where both the sources signals and the channel are unknown. The frequency-selective channel is modeled by a MIMO-AR system. The source signals are taken from a known FA, up
to an unknown complex scaling. The discrete symbols from the FA are modeled by i.i.d.
random GMM-distributed processes, with different means and infinitesimal covariance
matrices. The MIMO-AR parameters and the signal symbols are jointly estimated via
the ML estimator, implemented by using the EM algorithm. The proposed algorithm
blindly estimates the source signals up to discrete unknown phases, which is an inherent
problem in BSS. The performance of the proposed method is compared with existing
methods in terms of symbol error rate (SER), mean-square error (MSE) of the channel
parameter estimate, and signal-to-interference ratio (SIR).

1.3 Brief summary of the EM and GEM algorithms

In this work, the ML estimation is implemented by using the EM and GEM algorithms.
A brief summary of these algorithms is provided in this section. The EM and GEM
algorithms, proposed in [52], are iterative techniques for maximizing the likelihood func-
tion. The GEM algorithm can be used in problems where there is a complete data set
\( Z = (x, y) \), with joint PDF \( f_Z(z) \), where \( x \) and \( y \) are the incomplete data observations
set and the hidden data, respectively. This complete data set is chosen such that the
expectation \( E [ \log f_Z(Z; \theta') | x; \theta'' ] \) can be easily computed for any two parameters sets
\( \theta', \theta'' \), from the unknown parameters space, \( \Omega \). The algorithm includes 2 steps:

- Expectation (E-step): compute the complete log-likelihood via the conditional ex-
  pectation:

\[
U \left( \theta, \hat{\theta}^{(i)} \right) \triangleq E_{y|x} \left[ \log f_x(x; \theta) | x; \hat{\theta}^{(i)} \right] \tag{1.1}
\]

where \( i \) denotes the iteration index.
Maximization (M-step): determine $\hat{\theta}^{(i+1)}$ such that

$$
\hat{\theta}^{(i+1)} = \arg \max_{\theta} U \left( \theta, \hat{\theta}^{(i)} \right)
$$

(1.2)

in the EM algorithm, or such that

$$
U \left( \hat{\theta}^{(i+1)}, \hat{\theta}^{(i)} \right) \geq U \left( \hat{\theta}^{(i)}, \hat{\theta}^{(i)} \right)
$$

(1.3)

in the GEM algorithm. Thus, the GEM algorithm simply increases the complete log-likelihood (CLD) function during the M-step rather than maximizing the function. The GEM algorithm is used in cases where the maximization at the M-step is not tractable.

Under some conditions, this algorithm is guaranteed to converge to (at least) a local maximum of the log-likelihood function [52], [53].

Compared to other algorithms employing numerical optimization techniques, such as, gradient ascent methods and Newton type methods, the EM algorithm has the following advantages:

- The EM algorithm monotonically increases the likelihood at each iteration.

- The EM algorithm often provides intuitive insights to the estimation problem, while the other numerical methods provide no such insight.

- It does not require a selection of step size. Finding the optimum time-dependent step size in the gradient ascent methods is an ad-hoc process. It also does not require finding and inverting the Hessian matrix as required in every iteration of the Newton type methods. Computation of the second order derivatives and manipulations on the Hessian matrix can be very time-consuming and difficult to compute.
1.4 Thesis layout

The thesis is organized as follows. In Chapter 2, the system model is presented and the ML estimation problem is formulated, with and without prior knowledge on the structure of the input signal statistics. In Chapter 3, we consider the problem of BSS and MIMO system identification with known FA. The properties of the proposed algorithms are discussed in Chapter 4. The performances of the proposed algorithms, presented in Chapters 2 and 3, are evaluated via computer simulations and presented in Chapter 5. Finally, in Chapter 6 the main conclusions of this work are drawn and topics for future work are outlined.
Chapter 2

MIMO-AR system identification and BSS

In this chapter, the system model is presented and the ML estimation problem is formulated and implemented via the GEM algorithm for the general case of GMM-distributed input signals. This chapter is organized as follows. Mathematical models for the source and observation signals PDFs are derived in Section 2.1. In Section 2.2, the ML estimator for the state transition matrix is derived under the assumption of unstructured input signal statistics with arbitrary unknown joint PDF of the sources. In Section 2.3, this ML estimator is extended to include the prior knowledge about the structure of the input signal statistics due to statistically independent source assumption.
2.1 Problem Formulation

2.1.1 The MIMO-AR model

Consider the following MIMO-AR model:

\[ x_n = Ax_n^P + Hs_n \quad \forall n = 1, \ldots, N \]  \hspace{1cm} (2.1)

where \( n \) represents the time index, \( s_n \) and \( x_n \) denote the source and observation vectors at time instance \( n \), respectively, and \( x_{n,i} \) represents the \( n \)th sample at the \( i \)th sensor. The past samples vector, \( x_n^P \), is defined as follows

\[ x_n^P = (x_{n-1,1}, \ldots, x_{n-P,1}, \ldots, x_{n-1,L}, \ldots, x_{n-P,L})^T, \]

where \( P \) is the AR order, assumed to be known. \( A \) is an unknown \( L \times LP \) deterministic state transition matrix relating the state vector of the system between time instance \( n \) and time instances \( n-1, \ldots, n-P \). The matrix \( H \) is a full rank unknown \( L \times L \) deterministic input mixing matrix. The number of sensors, \( L \), is assumed to be equal to the number of sources. This model can also be represented in the following form:

\[ \sum_{p=0}^{P} A_p x_{n-p} = Hs_n \quad \forall n = 1, \ldots, N \] \hspace{1cm} (2.2)

where the sequence \( \{x_{n-p}\}_{p=0}^{P} \) includes the received vectors at times \( n, n-1, \ldots, n-P \), and the matrices \( \{A_p\}_{p=0}^{P} \) are the unknown \( L \times L \) AR coefficient matrices with \( A_0 = I_L \).

The state transition matrix, \( A \), in (2.1) and the AR coefficient matrices, \( \{A_p\}_{p=0}^{P} \), in (2.2), are related via

\[ A = \begin{bmatrix} -A_1; \cdots; -A_P \end{bmatrix}. \]

The stability conditions for this model are given in [64] and discussed in Appendix B.

In this work, we are interested in two problems: 1) system identification, i.e., estimation of the state transition matrix \( A \) and input mixing matrix \( H \), and 2) source separation, i.e., estimation of the source signals \( \{s_n\}_{n=1}^{N} \).
2.1.2 Source distribution model

The sources are assumed to be statistically independent, where each source signal is an i.i.d. GMM-distributed sequence. Thus, the PDF of the $l$th source signal at the $n$th time instance is

$$f_{s_n^l}(s; \theta_s) = \sum_{i=1}^{n_l} \phi_{l,i} N(s; \mu_{l,i}, \sigma_{l,i}^2) \quad \forall n = 1, \ldots, N$$ (2.3)

where $\theta_s \triangleq \{\{\mu_{l,i}, \sigma_{l,i}\}_{i=1}^{n_l}\}_{l=1}^L$. The notation $N(\xi; \mu, \Lambda)$ represents a normal distribution function with variable $\xi$, mean $\mu$, and covariance matrix $\Lambda$. The number of Gaussians for the $l$th source is denoted by $n_l$. The mean, variance and the weighting coefficients of the $i$th Gaussian of the $l$th source are denoted by $\mu_{l,i}$, $\sigma_{l,i}^2$, and $\phi_{l,i}$, respectively. The weighting coefficients, $\phi_{l,i}$, satisfy $\sum_{i=1}^{n_l} \phi_{l,i} = 1$, $0 < \phi_{l,i} \leq 1$, $\forall l = 1, \ldots, L$.

Under the assumption of independent source signals, the joint PDF of the sources is modeled by a multivariate GMM with diagonal covariance matrices:

$$f_{s_n}(s_n; \theta_s) = \prod_{l=1}^L f_{s_n^l}(s_n^l; \theta_s) = \sum_{m=1}^M \pi_m N(s; \mu_m, C_m)$$ (2.4)

where $s_n = [s_n^1, \ldots, s_n^L]^T$ and $M = \prod_{l=1}^L n_l$ is the GMM order, assumed to be known. The proposed method can be simply extended to include an unknown GMM order by using BIC [66], MDL [67] or AIC [68]. The index $m$ denotes a single combination of Gaussians from all the sources. The weighting coefficients, $\{\pi_m\}_{m=1}^M$, satisfy $0 < \pi_m \leq 1$ $\forall m = 1, \ldots, M$ and $\sum_{m=1}^M \pi_m = 1$. The vector $\mu_m$ and the diagonal matrix $C_m$ represent the mean and the covariance matrix of the $m$th Gaussian.

2.1.3 Sensors distribution model

Let $x \triangleq [x_1^T, \ldots, x_N^T]^T$ denote the measurements at $N$ time instances. Since a linear transformation of a GMM-distributed random variable is also GMM-distributed, the con-
ditional PDF of $x_n|x_n^P$ is also GMM, and using (2.1) and (2.4), it can be written in the form

$$f_{x_n|x_n^P}(x_n|x_n^P; \theta) = \sum_{m=1}^M \pi_m \mathcal{N}(x_n; \eta_m + Ax_n^P, R_m)$$ (2.5)

where $\theta = \{\{\pi_m, \mu_m, C_m\}_{m=1}^M, A, H\}$, $\theta \in \Omega$ denotes the set of unknown distribution parameters of the observation signals, $\Omega$ is the unknown parameters space, and

$$\eta_m \triangleq H\mu_m, \quad m = 1, \ldots, M, \quad (2.6)$$

$$R_m \triangleq HC_mH^T, \quad m = 1, \ldots, M. \quad (2.7)$$

The sequence $\{x_n\}_{n=1}^N$ is a $P$-order Markov process, and thus, the PDF of $x$ can be expressed as

$$f_x(x; \theta) = f_{x_0} \prod_{n=1}^N f_{x_n|x_n^P}(x_n|x_n^P, \theta)$$ (2.8)

where $f_{x_0}$ is the initial conditions distribution. For sufficiently large $N$, the information on $\theta$ from the initial conditions distribution, $f_{x_0}$, is negligible comparing to the information from the PDF of $\{x_n|x_n^P\}$. Thus, $f_{x_0}$ is omitted in the equations below.

### 2.2 Unstructured input

In this section, an ML estimator for the state transition matrix, $A$, and the GMM parameters, $\{\pi_m, \eta_m, R_m\}_{m=1}^M$, is derived under the assumption of unstructured input signal statistics. Thus, the covariance matrices $\{R_m\}_{m=1}^M$ are assumed to be unstructured, that is, the structure in (2.7) is ignored. Note that the structure $\eta_m \triangleq H\mu_m$ is not informative for a square, invertible matrix, $H$. This implies that the source signals are assumed to be statistically dependent jointly GMM-distributed. It will be shown that for estimation of the state transition matrix, $A$, there is no requirement of statistically independent source.
signals. This model does not enable to directly estimate the input mixing matrix, $H$. However, estimation of $H$ and separation of the input sources is possible by whitening the observation vectors using the estimated transition matrix, $A$, by applying common BSS techniques for instantaneous BSS mixtures.

The ML estimate for the set of unknown parameters under the unstructured model,

$$\hat{\theta}_u = \{\{\pi_m, \eta_m, R_m\}_{m=1}^M, A\},$$

from the observation vector $x$ is given by

$$\hat{\theta}_u = \arg \max_{\theta_u \in \Omega} \log f_x(x; \theta_u). \quad (2.9)$$

According to (2.8) and (2.5), the log-likelihood function under the above model is given by

$$\log f_x(x; \theta_u) = \sum_{n=1}^N \log \sum_{m=1}^M \pi_m N(x_n; \eta_m + Ax_n, R_m). \quad (2.10)$$

Since the above maximization cannot be analytically performed, the GEM algorithm is used.

**The complete data**

The complete data is chosen to be $\mathcal{Z} = (x, y)$ with $y = [y_1^T, \ldots, y_N^T]^T$ [69], [70], where the discrete random hidden indication vectors, $y_n = (y_{n,1}, \ldots, y_{n,M})^T$, $n = 1, \ldots, N$, are distributed according to the following PDF:

$$f_{y_n}(y_n) = \sum_{m=1}^M \pi_m \delta(y_{n,m} - 1) \quad (2.11)$$

where $\delta(\cdot)$ denotes the Dirac’s delta function and

$$y_{n,m} = \begin{cases} 
1 & \text{if } s_n \text{ is generated by the } m\text{th Gaussian} \\
0 & \text{otherwise}
\end{cases} \quad (2.12)$$
According to Bayes’ theorem:

\[
f_{x_n|y_n}(x_n|x_n^P; \theta_u) = E_{y_n} \left[ f_{x_n|y_n, x_n^P}(x_n|y_n, x_n^P; \theta_u) \right] = \sum_{m=1}^{M} \pi_m f_{x_n|y_n, m, x_n^P}(x_n|y_n, m = 1, x_n^P; \theta_u), \tag{2.13}\]

which is identical to (2.5). Note that the left equality is obtained using the statistical independency between \(x_n^P\) and \(y_n\). The logarithm of the joint PDF of \(x, y\) is given by [61]:

\[
\log f_{x,y}(x, y; \theta_u) = \sum_{n=1}^{N} \sum_{m=1}^{M} y_{n,m} \log \left[ \pi_m N(x_n; \eta_m + Ax_n^P, R_m) \right]. \tag{2.14}\]

The steps of the GEM algorithm for estimating \(\theta_u\) are described in the following.

The E-step

According to (1.1) in the introduction, the expectation step in the GEM algorithm is the computation of the conditional expectation of (2.14):

\[
U \left( \theta_u, \hat{\theta}_u^{(i)} \right) = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \left[ \pi_m N(x_n; \eta_m + Ax_n^P, R_m) \right] \tag{2.15}\]

where \(\hat{\theta}_u^{(i)} = \left\{ \hat{\eta}_m^{(i)}, \hat{R}_m^{(i)} \right\}_{m=1}^{M}, \hat{A}^{(i)} \) contains the estimates of \(\theta_u\) from the previous step, the \(i\)th step, and \(\gamma_{n,m}^{(i)} \triangleq E_{y|x} \left[ y_{n,m} \hat{\theta}_u^{(i)} \right] \). Since \(y_{n,m}\) can have only discrete values of 0 and 1, by applying Bayes’ theorem, \(\gamma_{n,m}^{(i)}\) can be calculated in the following manner [61]:

\[
\gamma_{n,m}^{(i)} = \frac{\hat{\pi}_m^{(i)} N(x_n; \hat{\eta}_m^{(i)} + \hat{A}^{(i)} x_n^P, \hat{R}_m^{(i)})}{\sum_{m'=1}^{M} \hat{\pi}_{m'}^{(i)} N(x_n; \hat{\eta}_{m'}^{(i)} + \hat{A}^{(i)} x_n^P, \hat{R}_{m'}^{(i)})}. \tag{2.16}\]

The M-step

The maximization step is performed by maximizing (2.15) w.r.t. \(\theta_u\). This is equivalent to solving the two maximization problems:

\[
\left\{ \hat{\pi}_m^{(i+1)} \right\}_{m=1}^{M} = \arg \max_{\left\{ \pi_m \right\}_{m=1}^{M}} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \pi_m \tag{2.17}\]
under the constraint \( \sum_{m=1}^{M} \pi_m = 1 \), and

\[
\left( \left\{ \hat{\eta}_{m}^{(i+1)}, \hat{R}_{m}^{(i+1)} \right\}_{m=1}^{M}, \hat{A}^{(i+1)} \right) = \arg \max_{\{\eta_m, R_m\}_{m=1}^{M}, A} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \left[ N \left( x_n; \eta_m + Ax_n^P, R_m \right) \right].
\]

(2.18)

In Appendix A.1, it is shown that the maximization in (2.17) can be performed analytically, which yields:

\[
\hat{\pi}_{m}^{(i+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \quad \forall m = 1, \ldots, M.
\]

(2.19)

However, the maximization in (2.18), is not tractable. Thus, in similar to [51], the coordinate ascent (CA) technique [59], [60] was chosen to increase the complete log-likelihood at each iteration. In the CA technique, the maximization is performed by maximizing the objective function w.r.t. a subset of the unknown parameters, while freezing the other unknown parameters. These steps are repeated till a predefined termination criterion is satisfied. This technique converges to a local maximum of the function [59], [60]. The CA technique is utilized to maximize (2.18) w.r.t. the three “coordinates”: \( A, \{\eta_m\}_{m=1}^{M} \) and \( \{R_m\}_{m=1}^{M} \). The derivatives w.r.t. matrix and vector parameters are defined in [55] and can be found in Appendix C. The CA algorithm for the ith GEM iteration is described as follows:

a. Initialization:

\[
\left\{ \eta_{m}^{(i+1,j=0)} = \hat{\eta}_{m}^{(i)}, \hat{R}_{m}^{(i+1,j=0)} = \hat{R}_{m}^{(i)} \right\}_{m=1}^{M}, \hat{A}^{(i+1,j=0)} = \hat{A}^{(i)} \quad \text{where } j \text{ is the CA iteration index, while } i \text{ is the GEM iteration index.}
\]

b. CA iteration

The mean vectors, \( \{\eta_m\}_{m=1}^{M} \), are estimated by maximization of (2.18) w.r.t. \( \{\eta_m\}_{m=1}^{M} \).
and setting \( \mathbf{A} = \hat{\mathbf{A}}^{(i+1,j)} \):

\[
\hat{\eta}_m^{(i+1,j+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(i)} (\mathbf{x}_n - \hat{\mathbf{A}}^{(i+1,j)}\mathbf{x}_n^P)}{\sum_{n=1}^{N} \gamma_{n,m}^{(i)}} \quad \forall m = 1, \ldots, M \quad (2.20)
\]

As shown in Appendix A.4. The covariance matrices, \( \{\mathbf{R}_m\}_{m=1}^{M} \), are estimated by maximization of (2.18) w.r.t. \( \{\mathbf{R}_m\}_{m=1}^{M} \) and setting \( \mathbf{A} = \hat{\mathbf{A}}^{(i+1,j)} \) and \( \eta_m = \hat{\eta}_m^{(i+1,j+1)} \) \( \forall m = 1, \ldots, M \):

\[
\hat{\mathbf{R}}_m^{(i+1,j+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(i)} (\mathbf{x}_n - \eta_m^{(i+1,j+1)} - \hat{\mathbf{A}}^{(i+1,j)}\mathbf{x}_n^P)^T (\mathbf{x}_n - \eta_m^{(i+1,j+1)} - \hat{\mathbf{A}}^{(i+1,j)}\mathbf{x}_n^P)}{\sum_{n=1}^{N} \gamma_{n,m}^{(i)}} \quad \forall m = 1, \ldots, M \quad (2.21)
\]

The proof can be found in Appendix A.3. Maximization of (2.18) w.r.t. \( \mathbf{A} \) and setting \( \{\hat{\eta}_m^{(i+1,j+1)}, \hat{\mathbf{R}}_m^{(i+1,j+1)}\}_{m=1}^{M} \), is achieved via equating the corresponding partial derivatives to zero, which yields (Appendix A.4):

\[
\sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \hat{\mathbf{R}}_m^{(i+1,j+1)}^{-1} (\mathbf{x}_n - \hat{\eta}_m^{(i+1,j+1)}) \mathbf{x}_n^T = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \hat{\mathbf{R}}_m^{(i+1,j+1)}^{-1} \mathbf{A}^{(i+1,j+1)} \mathbf{x}_n \mathbf{x}_n^T \quad (2.22)
\]

By applying the vec operator on both sides of (2.22) and using the Kronecker products property \( vec(\mathbf{A}_1\mathbf{A}_2\mathbf{A}_3) = (\mathbf{A}_3^T \otimes \mathbf{A}_1)vec(\mathbf{A}_2) \), we obtain:

\[
vec(\hat{\mathbf{A}}^{(i+1,j+1)}) = \left[ \sum_{m=1}^{M} \mathbf{C}^{(i)}_{\mathbf{x}\mathbf{x}\mathbf{x},m} \otimes \hat{\mathbf{R}}_m^{(i+1,j+1)} \right]^{-1} vec \left[ \sum_{m=1}^{M} \hat{\mathbf{R}}_m^{(i+1,j+1)} \mathbf{C}^{(i)}_{\mathbf{x}\mathbf{x},m} - \mathbf{C}^{(i)}_{\mathbf{x}\mathbf{x},m} \hat{\eta}_m^{(i+1,j+1)} \right] \quad (2.23)
\]

where

\[
\mathbf{C}^{(i)}_{\mathbf{x}\mathbf{x}\mathbf{x},m} \triangleq \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \mathbf{x}_n \mathbf{x}_n^T \otimes \mathbf{x}_n \quad (2.24)
\]

\[
\mathbf{C}^{(i)}_{\mathbf{x}\mathbf{x},m} \triangleq \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \mathbf{x}_n \mathbf{x}_n^T \quad (2.25)
\]

and

\[
\zeta_{\mathbf{x},m}^{(i)} \triangleq \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \mathbf{x}_n^T \quad (2.26)
\]
For $N \geq M$, the matrices $\hat{R}_{m}^{(i+1,j+1)}$ and $C_{x_{p}x_{p},m}^{(i)}$ are not singular w.p.1., and thus, 
$\sum_{m=1}^{M} C_{x_{p}x_{p},m}^{(i)} \otimes \hat{R}_{m}^{(i+1,j+1)^{-1}}$ is also not singular w.p.1.

These CA steps are iterated till a predefined termination criterion is satisfied, and then we define our parameter updates for the M-step by

\[
\hat{\eta}_{m}^{(i+1)} = \hat{\eta}_{m}^{(i+1,j)}, \quad m = 1, \ldots, M;
\]

\[
\hat{R}_{m}^{(i+1)} = \hat{R}_{m}^{(i+1,j)}, \quad m = 1, \ldots, M;
\]

\[
\hat{A}_{(i+1)} = \hat{A}_{(i+1,j)}.
\]

The expectation and maximization steps are repeated till a predefined termination criterion is satisfied. The termination criterion of the external GEM iterations or the internal CA iterations can be set by thresholding the increscent level of the log-likelihood or the complete log-likelihood, respectively, or by limiting the maximum number of iterations.

In Chapter 4 it is shown that in the case of MIMO-AR system with Gaussian-distributed sources, or equivalently GMM-distributed sources with $M = 1$ and $\mu_1 = 0$, the ML estimate of $A$, presented in (2.23), can be obtained by solving the well known multi-dimensional Yule-Walker equations obtained in the LPC [62], [63]:

\[
\hat{A}_{YW} = \left[ \sum_{n=1}^{N} x_{n} (x_{n}^{P})^{T} \right]^{-1} \left[ \sum_{n=1}^{N} x_{n}^{P} (x_{n}^{P})^{T} \right]^{-1} \tag{2.27}
\]

which is the ML estimator under the Gaussian assumption.

In Chapter 4 it is shown that the above procedure also coincides with the SISO-AR system identification technique, described in [51]. In addition, this system identification technique, described in (2.19)-(2.23), can be used also in the degenerative case, when the number of sensors is lower than the number of sources, as long as the matrix $\sum_{n=1}^{N} x_{n}^{P} x_{n}^{P^{T}}$ is not singular. Note that if the aforementioned matrix is singular - also the Yule-Walker equations can not be solved.
Source separation

As mentioned above, some statistical information regarding the model structure of the input signal is ignored, and thus, the source separation cannot be done directly. However, the estimate of the state transition matrix, $A$, can be used to separate the sources by the following simple procedure.

Using the model structure in (2.1), the whitened observation vector can be defined as

$$z_n = x_n - Ax^P_n = Hs_n.$$  

The source separation problem from the whitened signal is identical to the instantaneous BSS problem. For estimating the input mixing matrix, $H$, and the input sources, an existing instantaneous BSS technique can be implemented on the estimated whitened signal:

$$\hat{z}_n = x_n - \hat{A}x^P_n$$

where $\hat{A}$ is the state transition matrix estimate obtained by the unstructured GEM algorithm. The BSS technique can be simple, like FastICA [3], or more complicated methods that take into consideration the statistical information about the GMM-distributed sources, as in [58]. Therefore, for the state transition matrix estimation task the source signals are not assumed to be statistically independent, but for the BSS stage and the input mixing matrix estimation they need to be statistically independent. The method presented in the next section is derived under the assumption of statistically independent source signals for both the system identification and the BSS.

In summary, the proposed iterative GEM algorithm is summarized in Table 2.2.
Table 2.2: The proposed GEM algorithm for unstructured input

**GEM initialization**

Guess \( \{\hat{\eta}_m^{(0)}, \hat{R}_m^{(0)}\}_{m=1}^M \) and \( \hat{\mathbf{A}}^{(0)} \), and initialize the GEM iterations index, \( i = 0 \).

**GEM iterations**

Repeat until convergence:

1. **E-step:** Compute \( \gamma_{n,m}^{(i)} \) according to (2.16).

2. **M-step:**
   - Calculate \( \hat{\pi}_{m}^{(i+1)} \) according to (2.19).
   - CA initialization: Set \( \{\hat{\eta}_m^{(i+1,0)}, \hat{R}_m^{(i+1,0)}\}_{m=1}^M = \{\hat{\eta}_m^{(i)}, \hat{R}_m^{(i)}\}_{m=1}^M \), \( \hat{\mathbf{A}}^{(i+1,0)} = \hat{\mathbf{A}}^{(i)} \), and initialize the CA iterations index, \( j = 0 \).
   - **CA iterations**
     Repeat until convergence:
     - Calculate \( \{\hat{\eta}_m^{(i+1,j+1)}, \hat{R}_m^{(i+1,j+1)}\}_{m=1}^M \), and \( \hat{\mathbf{A}}^{(i+1,j+1)} \) via (2.20), (2.21) and (2.23), respectively.
     - Set \( j = j + 1 \)
   - Set \( \{\hat{\eta}_m^{(i+1)}, \hat{R}_m^{(i+1)}\}_{m=1}^M = \{\hat{\eta}_m^{(i+1,j+1)}, \hat{R}_m^{(i+1,j+1)}\}_{m=1}^M \), \( \hat{\mathbf{A}}^{(i+1)} = \hat{\mathbf{A}}^{(i+1,j+1)} \), and \( i = i + 1 \).

**Source separation:**

Apply an instantaneous BSS technique on \( \hat{z}_n = \mathbf{x}_n - \hat{\mathbf{A}} \mathbf{x}_n^P \), where \( \hat{\mathbf{A}} \) is the GEM estimate of \( \mathbf{A} \).
2.3 Structured input

In this section, the ML estimator from the previous section is extended to include the prior knowledge about the structure of the input signal statistics, due to the assumption of statistically independent source signals. Thus, the aim of this section is to find the ML estimator of $\theta = \{\{\pi_m, \mu_m, C_m\}_{m=1}^M, A, H\}$ by maximizing (2.8) without ignoring the structures in (2.6), (2.7). The GEM algorithm is utilized again with the CA maximization technique, but now we consider a different set of coordinates. Using the GEM algorithm with the same complete data, $Z = (x, y)$, the structured complete log-likelihood is

$$U(\theta, \hat{\theta}^{(i)}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log[\pi_m N(x_n; H\mu_m + Ax_n, HC_mH^T)].$$

(2.28)

The algorithm in this case is similar to the GEM algorithm presented in the previous section, where the maximization w.r.t. $\{\eta_m, R_m\}_{m=1}^M$ is replaced by the maximization w.r.t. $H$ and $\{\mu_m, C_m\}_{m=1}^M$. Equation (2.28) can be rewritten as

$$U(\theta, \hat{\theta}^{(i)}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log[\pi_m N(x_n; B^{-1}\mu_m + Ax_n, B^{-1}C_mB^{-T})].$$

(2.29)

where $B = H^{-1}$ whereby our assumption is that the matrix $H$ is not singular. For simplicity, we use $B = H^{-1}$ instead of $H$, and the maximization in the M-step is performed w.r.t. $B$ instead of $H$. Using the invariant property of the ML estimator, the ML estimate of $H$, $\hat{H}_{ML}$, is given by the inverse of the ML estimate of $B$, $\hat{B}_{ML}^{-1}$. Thus, the ML estimate of $B$ can be found instead of $H$. The steps of the GEM algorithm for estimating $\theta$ are described in the following.
E-step

The E-step is performed in the same way as in Section 3.2.2:

\[
\gamma_{n,m}^{(i)} = \frac{\gamma_{m,n}^{(i)} \mathcal{N}(x_n; \hat{H}_n^{(i)} \mu_m^{(i)} + \hat{A}^{(i)} x_n^P, \hat{H}_n^{(i)} \hat{C}_m^{(i)} (\hat{H}_n^{(i)})^T)}{\sum_{m=1}^{M} \gamma_{m,n}^{(i)} \mathcal{N}(x_n; \hat{H}_n^{(i)} \mu_m^{(i)} + \hat{A}^{(i)} x_n^P, \hat{H}_n^{(i)} \hat{C}_m^{(i)} (\hat{H}_n^{(i)})^T)}.
\]

(2.30)

M-step

The complete log-likelihood in (2.28), is maximized w.r.t. the new coordinates: \(\{\pi_m\}_{m=1}^{M}\), \(\{C_m\}_{m=1}^{M}\), \(B\) and \(A\). Again, as in (2.17)-(2.18), the maximization is separated into two different problems: The maximization w.r.t. \(\{\pi_m\}_{m=1}^{M}\) is identical to the unstructured case, which results in (2.19). The second maximization is:

\[
\left( \hat{B}^{(i+1)} \right) \left\{ \hat{\mu}_m^{(i+1)}, \hat{C}_m^{(i+1)} \right\}_{m=1}^{M} \hat{A}^{(i+1)} = \arg \max_{B, \{\mu_m, C_m\}_{m=1}^{M}, A} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \left[ \mathcal{N} (x_n; H \mu_m + A x_n^P, B^{-1} C_m B^{-T}) \right]
\]

(2.31)

Since this maximization is intractable, the CA technique is applied also in this case. The CA algorithm for the \(i\)th GEM iteration is described as follows:

a. Initialization:

\[
\hat{B}^{(i+1,j=0)} = \hat{B}^{(i)}, \quad \left\{ \hat{\mu}_m^{(i+1,j=0)} = \mu_m^{(i)}, \hat{C}_m^{(i+1,j=0)} = C_m^{(i)} \right\}_{m=1}^{M}, \quad \hat{A}^{(i+1,j=0)} = \hat{A}^{(i)}, \quad \text{where } j \text{ is the CA iteration index, while } i \text{ is the GEM iteration index.}
\]

b. CA iteration

The diagonal covariance matrices, \(\{C_m\}_{m=1}^{M}\), are estimated by maximization of (2.31) w.r.t. \(\{C_m\}_{m=1}^{M}\) under the constraint of diagonality and setting \(\{\mu_m = \hat{\mu}_m^{(i+1,j)}\}_{m=1}^{M}\), \(B = \hat{B}^{(i+1,j)}\) and \(A = \hat{A}^{(i+1,j)}\). In Appendix A.5 it is shown that this maximization yields

\[
\text{DIAG} \left( \hat{B}^{(i+1,j)}, \hat{R}_m^{(i+1,j)}, \hat{B}^{(i+1,j)^T} \right)
\]

(2.32)
where \( \text{DIAG}(\cdot) \) denotes a diagonal matrix with the same diagonal elements of its argument and

\[
\hat{R}_m^{(i+1,j)} \triangleq \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(i)} \left( \hat{z}_n^{(i+1,j)} - \hat{B}^{(i+1,j)} \mu_m^{(i+1,j)} \right) \left( \hat{z}_n^{(i+1,j)} - \hat{B}^{(i+1,j)} \mu_m^{(i+1,j)} \right)^T}{\sum_{n=1}^{N} \gamma_{n,m}^{(i)}} \quad (2.33)
\]

where

\[
\hat{z}_n^{(i+1,j)} \triangleq x_n - \hat{A}^{(i+1,j)} x^n. \quad (2.34)
\]

Note that in (2.21) \( \hat{R}_m^{(i+1,j+1)} \) is the unstructured estimation of the covariance matrices \( R_m^{(i+1,j+1)} \), while here it is only a notation.

In similar to (2.23), maximization of (2.31) w.r.t. \( A \) and setting \( \{\mu_m = \hat{\mu}_m^{(i+1,j+1)}, C_m = \hat{C}_m^{(i+1,j+1)} \}_{m=1}^M \), and \( B = \hat{B}^{(i+1,j)} \) yields

\[
\text{vec} \left( \hat{A}^{(i+1,j+1)} \right) = \left[ \sum_{m=1}^{M} \hat{C}_m^{(i+1,j+1)} \otimes \hat{B}^{(i+1,j)} \hat{C}_m^{(i+1,j+1)} \hat{B}^{(i+1,j)} \right]^{-1} \text{vec} \left( \sum_{m=1}^{M} \hat{B}^{(i+1,j)} \hat{C}_m^{(i+1,j+1)} (B^{(i+1,j)} C_m^{(i)} - \hat{\mu}_m^{(i+1,j)} \zeta_m^{(i)}) \right) \quad (2.35)
\]

where \( C_m^{(i)}, C_m^{(i)}, \) and \( \zeta_m^{(i)} \) are defined in (2.24), (2.25), and (2.26), respectively.

The separation matrix, \( B \), and the mean vectors, \( \{\mu_m \}_{m=1}^M \), are jointly estimated by maximization of (2.31) w.r.t. \( B \) and \( \{\mu_m \}_{m=1}^M \) together and setting \( A = \hat{A}^{(i+1,j+1)} \) and \( \{C_m = \hat{C}_m^{(i+1,j+1)} \}_{m=1}^M \). It can be seen that these maximizations can be obtained by minimizing

\[
Q^{(i+1,j+1)} = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \left[ \log |2\pi B^{-1} \hat{C}_m^{(i+1,j+1)} B^{-T}| + \left( \hat{z}_n^{(i+1,j+1)} - B^{-1} \mu_m \right)^T B \hat{C}_m^{(i+1,j+1)} B^{-1} \left( \hat{z}_n^{(i+1,j+1)} - B^{-1} \mu_m \right) \right] \quad (2.36)
\]

w.r.t. \( B \) and \( \{\mu_m \}_{m=1}^M \). The objective function in (2.36) is minimized w.r.t. \( \mu_m \) for a
fixed value of $B$ by

$$\hat{\mu}^{(i+1,j+1)}_m = B \sum_{n=1}^{N} \frac{\gamma_{n,m}^{(i)} \hat{z}^{(i+1,j+1)}_{n,m}}{\sum_{n=1}^{N} \gamma_{n,m}^{(i)}}, \quad \forall m = 1, \ldots, M. \quad (2.37)$$

By substitution of (2.37) into (2.36) and using the trace operator, the attained minimum is:

$$Q^{(i+1,j+1)}_2 = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log |2\pi B^{-1} \hat{C}_m^{(i+1,j+1)} B^{-T}| + tr \left( B \hat{C}_m^{(i+1,j+1)-1} B^T \hat{R}_m^{(i+1,j+1)} \right). \quad (2.38)$$

It can be shown [58] that the objective function in (2.38) can be rewritten as

$$Q^{(i+1,j+1)}_2 = \sum_{m=1}^{M} \hat{n}_m^{(i+1)} KL_{norm} \left[ B \hat{R}_m^{(i+1,j+1)} B^T | \hat{C}_m^{(i+1,j+1)} \right] + const \quad (2.39)$$

where $const$ denotes a constant w.r.t. $B$, and $KL_{norm}[\Sigma_1|\Sigma_2]$ is the KL divergence between two zero-mean multivariate normal densities with covariance matrices $\Sigma_1$ and $\Sigma_2$ [50]. The Pythagorean property of the KL divergence [1], [57] implies that (2.39) can be decomposed in the following manner:

$$Q^{(i+1,j+1)}_2 = \sum_{m=1}^{M} \hat{n}_m^{(i+1)} \left( KL_{norm} \left[ B \hat{R}_m^{(i+1,j+1)} B^T | DIAG \left( B \hat{R}_m^{(i+1,j+1)} B^T \right) \right] \right) + \left( KL_{norm} \left[ DIAG \left( B \hat{R}_m^{(i+1,j+1)} B^T \right) | \hat{C}_m^{(i+1,j+1)} \right] \right). \quad (2.40)$$

In Appendix A.6, it is shown that the estimate of the matrix $B$ can be rewritten as:

$$\hat{B}^{(i+1,j+1)} = \arg \min_B Q^{(i+1,j+1)}_2 = \arg \min_B \left[ \frac{1}{2} \sum_{l=1}^{L} b_l^T \Psi_l b_l - \log |B| \right] \quad (2.41)$$

where $b_l^T$ and $\hat{b}_l^T$ are the $l$th rows of the matrices $B$ and $\hat{B}^{(i+1,j)}$, respectively, and

$$\Psi_l \triangleq \sum_{m=1}^{M} \hat{n}_m^{(i+1)} \frac{1}{b_l^T \hat{R}_m^{(i+1,j+1)} b_l} \hat{R}_m^{(i+1,j+1)}. \quad \text{The denominator term is positive w.p.1., because the matrices} \{ \hat{R}_m^{(i+1,j+1)} \}_{m=1}^{M} \text{are positive-definite matrices w.p.1.} \text{Equating the partial derivative of the objective function in (2.41) w.r.t.} b_l \text{to zero, results}$$

$$\hat{B}^{(i+1,j+1)} = \left[ \Psi_1 \hat{b}_1^{(i+1,j+1)}; \Psi_2 \hat{b}_2^{(i+1,j+1)}; \cdots; \Psi_L \hat{b}_L^{(i+1,j+1)} \right]^{-1}. \quad (2.42)$$
In general, for \( L > 2 \) this nonlinear equation has no analytical solution and may be solved iteratively. In the following, an analytic solution of (2.42) for \( L = 2 \) is derived.

**Input mixing matrix estimation for \( L = 2 \)**

For the sake of simplicity, in this subsection the iteration indexes are omitted. Let \( \Psi_{ij} \) and \( \hat{B}_{ij} \) denote the \( ij \)th elements of \( \Psi \) and \( \hat{B} \), respectively. Then, for \( L = 2 \) (2.42) can be expressed as:

\[
\begin{bmatrix}
\hat{B}_{11} & \hat{B}_{12} \\
\hat{B}_{21} & \hat{B}_{22}
\end{bmatrix}
= \begin{bmatrix}
\Psi_{11} & \Psi_{12} \\
\Psi_{21} & \Psi_{22}
\end{bmatrix}
\begin{bmatrix}
\hat{B}_{11} \\
\hat{B}_{12}
\end{bmatrix} + \begin{bmatrix}
\Psi_{21} & \Psi_{22} \\
\Psi_{11} & \Psi_{12}
\end{bmatrix}
\begin{bmatrix}
\hat{B}_{21} \\
\hat{B}_{22}
\end{bmatrix}
\]

(2.43)

or

\[
\frac{1}{|B|}
\begin{bmatrix}
\hat{B}_{22} & -\hat{B}_{12} \\
-\hat{B}_{21} & \hat{B}_{11}
\end{bmatrix}
= \begin{bmatrix}
\Psi_{11}\hat{B}_{11} + \Psi_{12}\hat{B}_{12} & \Psi_{21}\hat{B}_{21} + \Psi_{22}\hat{B}_{22} \\
\Psi_{12}\hat{B}_{11} + \Psi_{11}\hat{B}_{12} & \Psi_{22}\hat{B}_{21} + \Psi_{21}\hat{B}_{22}
\end{bmatrix}.
\]

(2.44)

Let define \( \bar{b} \triangleq \left( \hat{B}_{22}, \hat{B}_{21}, \hat{B}_{12}, \hat{B}_{11} \right)^T \), then, (2.44) can be rearranged in the following form:

\[
\frac{1}{|B|} \bar{b} = \begin{bmatrix}
0 & \tilde{\Psi}_1 \\
\tilde{\Psi}_2 & 0
\end{bmatrix} \bar{b} = \Psi \bar{b}
\]

(2.45)

where \( \tilde{\Psi}_1 \triangleq \begin{bmatrix}
\Psi_{12} & \Psi_{11} \\
-\Psi_{12} & -\Psi_{11}
\end{bmatrix} \) and \( \tilde{\Psi}_2 \triangleq \begin{bmatrix}
-\Psi_{22} & -\Psi_{21} \\
\Psi_{22} & \Psi_{21}
\end{bmatrix} \). Since \( \Psi_1 \) and \( \Psi_2 \) are symmetric, \( \Psi_{121} = \Psi_{112} \) and \( \Psi_{221} = \Psi_{212} \), and thus \( tr(\tilde{\Psi}_1) = tr(\tilde{\Psi}_2) = 0 \). It can be seen that the vector \( \bar{b} \) is an eigenvector of the matrix \( \Psi \) with a corresponding eigenvalue \( \frac{1}{|B|} \).

Solving this eigenvalues-eigenvectors problem yields infinite solutions as described in the next paragraph.

It can be easily verified that the eigenvalues of \( \tilde{\Psi}_1 \) and \( \tilde{\Psi}_2 \) are \( \pm j \sqrt{|\Psi_1|} \) and \( \pm j \sqrt{|\Psi_2|} \), respectively, and the corresponding eigenvectors are proportional to

\[
\nu_1^{(1)} = \left[ \Psi_{11}, \Psi_{12}, j \sqrt{|\Psi_1|} - \Psi_{12} \right]^T
\]
\[ \mathbf{v}_1^{(2)} = \left[ \Psi_{11}, -j \sqrt{\Psi_1} - \Psi_{12} \right]^T \]

for the matrix \( \tilde{\Psi}_1 \), and

\[ \mathbf{v}_2^{(1)} = \left[ -\Psi_{21}, j \sqrt{\Psi_2} + \Psi_{22} \right]^T \]

\[ \mathbf{v}_2^{(2)} = \left[ -\Psi_{21}, -j \sqrt{\Psi_2} + \Psi_{22} \right]^T \]

for the matrix \( \tilde{\Psi}_2 \). It can be shown that the eigenvalues of \( \Psi \) are \( \pm \sqrt{\alpha_1}, \pm \sqrt{\alpha_2} \), where \( \alpha_1 \) and \( \alpha_2 \) are the eigenvalues of the matrix \( \tilde{\Psi}_1 \tilde{\Psi}_2 \), and the four eigenvectors of \( \Psi \) are proportional to \( \mathbf{v} = \left[ \beta_1 \mathbf{v}_1^{(k)}, -j \sqrt{\Psi_1} + \Psi_{12} \right] \), \( k = 1, 2 \), where \( \beta_1 \) and \( \beta_2 \) are arbitrary constants. Thus, there are four possible types of solutions for the matrix \( \mathbf{B} \):

\[ \mathbf{B} = \begin{bmatrix} \beta_2 \left( -j \sqrt{|\Psi_2|} + \Psi_{21} \right) & -\beta_2 \Psi_{21} \\ \beta_1 \left( -j \sqrt{|\Psi_1|} + \Psi_{12} \right) & \beta_1 \Psi_{11} \end{bmatrix}. \tag{2.46} \]

The ambiguity in selection of the constants \( \beta_1 \) and \( \beta_2 \) is the inherent scaling ambiguity in BSS problems. Equation (2.46) is the solution of (2.42), which is part of the M-step in the GEM algorithm for the structured input case.

**Summary of the structured GEM algorithm**

This CA step is iterated till a termination criterion is satisfied, and the parameter updates for the M-step are defined by

\[ \hat{\mu}_{m}^{(i+1)} = \hat{\mu}_{m}^{(i+1,j)}, \quad m = 1, \ldots, M, \]

\[ \hat{C}_{m}^{(i+1)} = \hat{C}_{m}^{(i+1,j)}, \quad m = 1, \ldots, M, \]

\[ \hat{A}^{(i+1)} = \hat{A}^{(i+1,j)}, \]

\[ \hat{B}^{(i+1)} = \hat{B}^{(i+1,j)}, \]

where \( j \) is the index in the last CA iteration. The expectation and maximization steps are repeated till a predefined termination criterion is satisfied. The termination criterion
of the external GEM iteration and the internal CA iteration can be set as in the previous section.

The sources can be separated by \( \hat{s}_n = \hat{B} \left( x_n - \hat{A} \hat{x}_n^P \right) \) where \( \hat{B} \) and \( \hat{A} \) are the GEM estimates. Due to the scaling and permutation ambiguities of the BSS problem, the source signal estimation is up to scaling and permutation of the sources.

The GEM algorithm developed in this section is summarized in Table 2.3.

### 2.4 Discussion and conclusion

A new ML-based algorithms for BSS of convolutive mixtures modeled by MIMO-AR in the time-domain were presented. The algorithms assume GMM distribution of the sources.

In the first method, a GEM algorithm for estimation of the state transition matrix, \( \mathbf{A} \) is derived under the assumption of unstructured input signal statistics. In the second method, this GEM algorithm is extended to estimate both the matrices \( \mathbf{A} \) and \( \mathbf{H} \) for \( L = 2 \). In addition, two ad-hoc algorithms for estimation of \( \mathbf{A} \) and \( \mathbf{H} \), for all \( L \) are proposed.

The proposed method extends the approach presented in [51] for SISO-AR system identification to MIMO-AR systems. In the case of lower number of sources, \( K \), than sensors, \( L \), the set of the covariance matrices \( \{ \mathbf{R}_m \}_{m=1}^{M} \) are singular matrices, because \( \text{rank}(\mathbf{H}) \leq K < L \). In this case, \( \mathbf{R}_m \) can be replaced with \( \mathbf{R}_m = \lim_{\varepsilon \to 0} (\mathbf{H} \mathbf{C}_m \mathbf{H}^T + \varepsilon \mathbf{I}_L) \), and a dimension reduction procedure [58] is required. The model presented here assumed real signals and system, but extension to the complex case is straightforward.
Table 2.3: The proposed GEM algorithm for structured input

GEM initialization

Guess \( \{ \hat{\mu}_m^{(0)}, \hat{C}_m^{(0)} \}_{m=1}^M, \hat{B}^{(0)} \) and \( \hat{A}^{(0)} \), and initialize the GEM iterations index, \( i = 0 \).

GEM iterations

Repeat until convergence:

1. E-step: Compute \( \gamma_{n,m}^{(i)} \) according to (2.16).

2. M-step:
   - Calculate \( \hat{\pi}^{(i+1)}_m \) according to (2.19).
   - CA initialization
     
     Set \( \{ \hat{C}_m^{(i+1,j=0)} = \hat{C}_m^{(i)}, \hat{\mu}_m^{(i+1,j=0)} = \hat{\mu}_m^{(i)} \}_{m=1}^M, \hat{B}^{(i+1,j=0)} = \hat{B}^{(i)} \) and \( \hat{A}^{(i+1,j=0)} = \hat{A}^{(i)} \), and initialize the CA iterations index, \( j = 0 \).
   - Repeat until convergence:
     - Calculate \( \hat{A}^{(i+1,j+1)} \) according to (2.35).
     - Calculate \( \hat{z}_n^{(i+1,j+1)} \) and \( \hat{R}_m^{(i+1,j+1)} \) according to (2.34) and (2.33), respectively.
     - Calculate \( \{ \hat{\mu}_m^{(i+1,j+1)}, \hat{C}_m^{(i+1,j+1)} \}_{m=1}^M \) by substituting \( \hat{z}_n^{(i+1,j+1)} \) and \( \hat{R}_m^{(i+1,j+1)} \) in (2.35), (2.37) and (2.32), respectively.
     - For \( L = 2 \): Compute \( \Psi \) as defined in (2.45), and \( \hat{B}^{(i+1,j+1)} \) as defined in (2.46). For \( L \neq 2 \): Find \( \hat{B}^{(i+1,j+1)} \) numerically by (2.42).
     - Set \( j = j + 1 \)

3. Set \( i = i + 1 \)

Source separation

The separated sources are \( \hat{s}_n = \hat{B} \left( x_n - \hat{A} x_n^P \right) \) where \( \hat{B} \) and \( \hat{A} \) are the GEM estimates for \( B \) and \( A \), respectively.
Chapter 3

Frequency-selective MIMO channel estimation

In this chapter, we consider the problem of separation of source signals taken from a known FA, up to a complex scaling. In many channel estimation and signal detection problems, the source symbols are taken from a FA. The FA property of the transmitted signal has been shown to be very useful in channel estimation and blind equalization in communication. The discrete symbols from the FA are modeled by i.i.d. random GMM-distributed processes, with different means and infinitesimal covariance matrices. The MIMO-AR parameters and the signal symbols are jointly estimated via the ML estimator, implemented by using the EM algorithm. The proposed algorithm blindly estimates the source signal up to discrete unknown phases, which is an inherent problem in BSS.

This chapter is organized as follows. The model is presented in Section 3.1.1. Mathematical models for the source and observation signals PDFs are derived in in Sections 3.1.2 and 3.1.3, respectively. In Section 3.2, the EM algorithm for estimating the state transition matrix and the input mixing matrix is derived.
3.1 Problem formulation

3.1.1 The MIMO-AR model

Consider a baseband equivalent communication system signaling through a stationary frequency-selective fading uplink channel with $L$ transmit and $L$ receive antennas. The $L$ transmit antennas may be divided into $L$ or less users, each one having one or more antennas. The MIMO frequency-selective channel is modeled by a MIMO-AR system, in which the input sources have the same transmission symbol rate. In a multiuser problem, the delays between the different users can be included in the MIMO-AR system. Thus, with no loss of generality (except possible increase in the model order), the sources are assumed to be synchronized. The MIMO-AR channel denotes a composite channel impulse response that includes transmitter and receiver filters, as well as the physical propagation channel and the different delays.

The equivalent baseband system can be modeled as

$$x_n = Ax_n^P + Hs_n \quad \forall n = 1, \ldots, N$$

(3.1)

where $x_n \in \mathbb{C}^{L \times 1}$ denotes the complex envelope of the received data at time instance $n$, $s_n \in \mathbb{C}^{L \times 1}$ is the complex transmit symbol vector at time instance $n$. The matrices $H$ and $A$, the vector $x_n^P$, and the parameters $L$ and $P$ are all defined in the previous chapter. The innovation in this chapter is that the matrices $H$ and $A$ are assumed to be complex-valued and the source signal has known discrete values.
3.1.2 Transmitted signal distribution model

Consider \( L \) source signals drawn from \( L \) known FA sets of size \( Q_l, \ l = 1, \ldots, L \). The \( l \)th source signal at time instance \( n \), satisfies \( s^l_n \in \{\xi^l_1, \ldots, \xi^l_{Q_l}\} \), \( \forall l = 1, \ldots, L \), where \( \{\xi^l_q\}_{q=1}^{Q_l} \) is the FA set for the \( l \)th source. In addition, the transmitted signals are assumed to be statistically independent in time and space, and equiprobable, i.e., \( Pr(s^l_n = \xi^l_q) = \frac{1}{Q_l}, \ \forall q = 1, \ldots, Q_l, \ \forall l = 1, \ldots, L \). The statistical model for the discrete source signal is assumed to be a continuous i.i.d. process with GMM distribution, with infinitesimal identical variances. Thus, the PDF of the \( l \)th source signal at each time instance \( n \), can be modeled by GMM:

\[
    f_{s^l_n}(s^l_n) = \lim_{\varepsilon \to 0} \frac{1}{Q_l} \sum_{q=1}^{Q_l} \mathcal{N}^C(s^l_n; \xi^l_q, \varepsilon^2) \quad \forall n = 1, \ldots, N \quad (3.2)
\]

where the notation \( \mathcal{N}^C(z; \mu, \Lambda) \) represents a circular complex normal distribution function with variable \( z \), mean \( \mu \), and covariance matrix \( \Lambda \). The number of Gaussians in the distribution of each source is given by the alphabet length, \( Q_l \). The expectations are the known alphabet values, \( \{\xi^l_q\}^{Q_l}_{q=1} \). Note that the FA sets of the different sources are not necessarily identical. The infinitesimal variances are denoted by \( \varepsilon^2 \to 0 \). Under the assumption of independent source signals, the joint PDF of the sources is multivariate GMM with diagonal covariance matrices:

\[
    f_{s_n}(s_n) = \prod_{l=1}^{L} f_{s^l_n}(s^l_n) = \lim_{\varepsilon \to 0} \frac{1}{M} \sum_{m=1}^{M} \mathcal{N}^C(s_n; \alpha_m, \varepsilon^2 I_L) \quad \forall n = 1, \ldots, N \quad (3.3)
\]

where \( s_n = [s^1_n, \ldots, s^L_n]^T \). The index \( m \) denotes a single combination of symbols from all the sources. The number of possible combinations is \( M = \prod_{l=1}^{L} Q_l \), which is the GMM order in multivariate GMM distribution in (3.3). The mean vector, \( \alpha_m \), is composed of the FA sets of the different sources, i.e., the \( l \)th argument of \( \alpha_m \) is taken from the FA set \( \{\xi^l_q\}^{Q_l}_{q=1} \).
3.1.3 Received signal distribution model

Let $\mathbf{x} \triangleq [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]^T$ denote the measurements at $N$ time instances. In similar to the previous chapter, according to (2.5) the conditional PDF of $x_n|\mathbf{x}_n^P$ can be written in the form

$$f_{x_n|\mathbf{x}_n^P}(x_n|\mathbf{x}_n^P, \theta) = \lim_{\varepsilon \to 0} \sum_{m=1}^{M} \frac{1}{M} \mathcal{N}^C \left( x_n; \mathbf{A}x_n^P + \mathbf{H}\alpha_m, \varepsilon^2\mathbf{H}\mathbf{H}^H \right)$$

(3.4)

where $\theta = \{\mathbf{A}, \mathbf{H}\}$, $\theta \in \Omega$ denotes the set of unknown distribution parameters of the observation signals, and $\Omega$ is the unknown parameters space. It can be seen that only the GMM expectations carry significant statistical information about $\theta$. The equal covariance matrices are meaningless for channel estimation because they are infinitesimal. This is the reason that the channel estimation for the described MIMO-AR model with unknown FA is impossible.

3.2 Joint channel estimation and symbol detection

3.2.1 Likelihood function for channel estimation

According to (2.8) and (3.4), the ML estimate of $\theta$ from the observation vector $\mathbf{x}$ is:

$$\hat{\theta} = \arg \max_{\theta \in \Omega} \log f_{\mathbf{x}}(\mathbf{x}; \theta)$$

(3.5)

$$\log f_{\mathbf{x}}(\mathbf{x}; \theta) = \sum_{n=1}^{N} \log f_{x_n|\mathbf{x}_n^P}(x_n|x_n^P; \theta)$$

$$= \lim_{\varepsilon \to 0} \sum_{n=1}^{N} \log \sum_{m=1}^{M} \frac{1}{M} \mathcal{N}^C \left( x_n; \mathbf{A}x_n^P + \mathbf{H}\alpha_m, \varepsilon^2\mathbf{H}\mathbf{H}^H \right).$$

(3.6)

Maximization of (3.5) cannot be analytically performed, thus the EM algorithm is used for this problem.
3.2.2 The proposed algorithm

The complete data

The complete data is chosen to be \( Z = (x, y) \) with the same random hidden indication vector, \( y = [y_1^T, \ldots, y_N^T]^T \), from the previous section. In this section, the hidden indication vector, \( y_n \), indicates which symbol from the alphabet is used at time instance \( n \). In similar to (2.14) the logarithm of the joint PDF of \( x, y \) is given by

\[
\log f_{x,y}(x, y; \theta) = \lim_{\varepsilon \to 0} \frac{1}{M} \sum_{n=1}^{N} \sum_{m=1}^{M} y_{n,m} \log \left( \frac{1}{M} \mathcal{N}(x_n; Ax_n^P + H\alpha_m, \varepsilon^2HH^H) \right). \tag{3.7}
\]

The E-step

By substituting of (2.14) into (1.1) one obtains

\[
U(\theta, \hat{\theta}^{(i)}) = \lim_{\varepsilon \to 0} \frac{1}{M} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \left( \frac{1}{M} \mathcal{N}(x_n; Ax_n^P + H\alpha_m, \varepsilon^2HH^H) \right) \tag{3.8}
\]

where \( \gamma_{n,m}^{(i)} \triangleq \mathbb{E}_{y|x} \left[ y_{n,m} | \hat{\theta}^{(i)} \right] \) and \( i \) denotes the iteration index. \( \gamma_{n,m}^{(i)} \) can be calculated by applying the Bayes’ theorem [61]:

\[
\gamma_{n,m}^{(i)} = \lim_{\varepsilon \to 0} \frac{\mathcal{N}(x_n; \hat{A}^{(i)}x_n^P + \hat{H}^{(i)}\alpha_m, \varepsilon^2\hat{H}^{(i)}\hat{H}^{(i)H})}{\sum_{m'=1}^{M} \mathcal{N}(x_n; \hat{A}^{(i)}x_n^P + \hat{H}^{(i)}\alpha_{m'}, \varepsilon^2\hat{H}^{(i)}\hat{H}^{(i)H})} = \lim_{\varepsilon \to 0} \frac{\exp\left(-\frac{1}{2\varepsilon^2} t_{n,m}^{(i)}\right)}{\sum_{m'=1}^{M} \exp\left(-\frac{1}{2\varepsilon^2} t_{n,m'}^{(i)}\right)} \tag{3.9}
\]

where

\[
t_{n,m}^{(i)} \triangleq \left( z_n^{(i)} - \hat{H}^{(i)}\alpha_m \right)^H \left( \hat{H}^{(i)}\hat{H}^{(i)H} \right)^{-1} \left( z_n^{(i)} - \hat{H}^{(i)}\alpha_m \right) = \| \hat{H}^{(i)^{-1}} z_n^{(i)} - \alpha_m \|^2 \tag{3.10}
\]

where \( \hat{z}_n^{(i)} \triangleq x_n - \hat{A}^{(i)}x_n^P \). It can be seen that in the limit \( \varepsilon^2 \to 0 \) one obtains

\[
\gamma_{n,m}^{(i)} = \begin{cases} 
1 & \text{if } t_{n,m}^{(i)} < t_{n,m'}^{(i)} \quad \forall m' = 1, \ldots, M, \quad m' \neq m \\
0 & \text{otherwise} 
\end{cases} \tag{3.11}
\]

Note that if \( m \neq m' \), then \( t_{n,m}^{(i)} \neq t_{n,m'}^{(i)} \) w.p.1., and that \( \sum_{m=1}^{M} \gamma_{n,m}^{(i)} = 1 \). The last property will be useful in the sequel.
The M-step

The maximization step is performed by maximizing (3.8) w.r.t. the state transition matrix, $A$, and the input mixing matrix, $H$.

- State transition matrix estimation:

In Appendix A.7, it is shown that equating the partial derivatives of (3.8) w.r.t. $A$ to zero, yields:

$$ \hat{A}^{(i+1)} = \left[ \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \left( x_n - \hat{H}^{(i+1)} \alpha_m \right) x_n^{PH} \right]^{-1} \left[ \sum_{n=1}^{N} x_n^{P} x_n^{PH} \right]. \quad (3.12) $$

- Input mixing matrix estimation:

In Appendix A.8, it is shown that equating the partial derivatives of (3.8) w.r.t. $H$ to zero and taking the limit $\epsilon \to 0$, yields:

$$ \hat{H}^{(i+1)} = \left( \sum_{n=1}^{N} \hat{z}_n^{(i+1)} \right) \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \alpha_m \hat{z}_n^{(i+1)H} \right)^{-1}. \quad (3.13) $$

It can be seen that in flat-fading channels, i.e. $A = 0$, the input mixing matrix estimation is reduced to the mixing matrix estimation presented in [45] for flat-fading channel estimation without noise.

Finally, in Appendix A.9 it is shown that $\hat{H}^{(i+1)}$ and $\hat{A}^{(i+1)}$ can be found by solving (3.12) and (3.13):

$$ \hat{H}^{(i+1)} = \left( C_{xx} - C_{xxP} C_{xP}^{-1} C_{xP}^{x} \right) \left[ \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \alpha_m x_n^{H} - \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \alpha_m x_n^{PH} \right) C_{xP}^{-1} C_{xP}^{x} \right]^{-1}, \quad (3.14) $$

$$ \hat{A}^{(i+1)} = \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \left( x_n - \hat{H}^{(i+1)} \alpha_m \right) x_n^{PH} \right) C_{xP}^{-1} C_{xP}. \quad (3.15) $$
where

\[ C_{xx} \triangleq \sum_{n=1}^{N} x_n x_n^H , \]

\[ C_{x^P x}^H = C_{xx^P} \triangleq \sum_{n=1}^{N} x_n x_n^P^H , \]

\[ C_{x^P x^P} \triangleq \sum_{n=1}^{N} x_n^P x_n^P^H . \]

The iterative steps, (3.11), (3.14), (3.15), are repeated till a predefined termination criterion is satisfied. The termination criterion can be set by thresholding the increase level of the log-likelihood or by limiting the maximum number of iterations.

**Equalization**

Based on the model in (3.1), the input signals can be simply reconstructed by

\[ \hat{s}_n = \hat{H}^{-1}(x_n - \hat{A} x_n^P) , \]  

(3.16)

where \( \hat{A}, \hat{H} \) are the estimates of \( A, H \), respectively, and signals \( \hat{s}_n \) is the estimated source signals, up to the inherent permutation and scaling ambiguities in BSS problems. For FA data sets, the scaling ambiguity depends on the signal constellation. For example, in the case of binary phase-shift keying (BPSK) and quadrature phase-shift keying (QPSK) modulations, it takes values from \( \{+1, -1\} \) and \( \{1, e^{-j\pi/4}, e^{j\pi/4}, e^{-j3\pi/4}, e^{j3\pi/4}\} \), respectively. In these cases, the scaling ambiguity is confined to discrete phase ambiguity. In some cases, the scaling ambiguity can be resolved, for example, by using a differential encoding of data, but the permutation ambiguity is much more difficult to resolve. The scaling and permutation ambiguity problems can be simultaneously solved using few training symbols.

The algorithm developed in this section is summarized in Table 3.2.
Table 3.2: The proposed EM algorithm for channel estimation and equalization with FA

**EM initialization**

Guess $\hat{\mathbf{A}}^{(0)}$, $\hat{\mathbf{H}}^{(0)}$ and initialize the iteration index $i = 0$.

**The EM algorithm**

Repeat until convergence:

1. E-step: Compute $t_{n,m}^{(i)}$ according to (3.10).
   
   Compute $\gamma_{n,m}^{(i)}$ according to (3.11).

2. M-step: Calculate $\hat{\mathbf{H}}^{(i+1)}$ and $\hat{\mathbf{A}}^{(i+1)}$ according to (3.14) and (3.15), respectively.

**Source separation:**

- Equalization $\hat{s}_n = \hat{\mathbf{H}}^{-1} \left( x_n - \hat{\mathbf{A}} x_{n}^{\ell} \right)$
- Quantization
3.3 Discussion and conclusion

A finite-alphabet based channel estimation technique for frequency-selective fading channel has been developed in this section. The frequency-selective channel is modeled by a MIMO-AR system. The state transition matrix and the input mixing matrix are jointly estimated by applying the EM algorithm. This section demonstrates one of the applications of the convolutive BSS method presented in Chapter 2.

The model presented here assumes known FA. It can be seen that the separation is impossible for the same MIMO-AR model with unknown FA, unless more assumptions are added. These assumptions can be, for example, constant modulus property of the transmitted signals. Nevertheless, the state transition matrix, $A$, can be estimated by the proposed algorithm even with unknown alphabet, where the estimate of $H$ is replaced by the estimate of $H\mathbf{s}_n$.

The proposed method can be implemented also in the case of lower number of sources, $K$, than sensors, $L$. In this case a dimension reduction procedure [58] is needed. Additionally, the common covariance matrix of the Gaussians, $\varepsilon^2 HH^T$, is singular because $\text{rank}(H) \leq K < L$. For using the proposed algorithm, $HH^H$ can be replaced with $\lim_{\delta \to 0}(HH^H + \delta I_L)$. 
Chapter 4

Properties

In this chapter, the properties and the behavior of the algorithms described in Chapters 2 and 3 are discussed.

4.1 Special cases

4.1.1 MIMO-AR system with Gaussian-distributed sources

In this section, it is shown that for zero-mean Gaussian-distributed sources the unstructured estimation of the state transition matrix, $A$, reduces to the well known multidimensional Yule-Walker equations presented in (2.27).

Gaussian-distributed sources can be represented by 1st-order GMM distribution. For i.i.d. stationary Gaussian-distributed sources only the estimation of the state transition matrix, $A$, is possible and the unstructured estimation is utilized. The GEM algorithm in this case is changed to the following:
E-step

Using (2.15)-(2.16), the E-step for \( M = 1 \) is

\[
U \left( \theta_u, \hat{\theta}_u^{(i)} \right) = \sum_{n=1}^{N} \gamma_n^{(i)} \log \left[ \pi_{1, n} \mathcal{N}(x_n; \eta + Ax_n^P, R) \right]
\]

(4.1)

where \( \pi_1 = 1 \) and

\[
\gamma_n^{(i)} = \frac{\hat{\pi}_1^{(i)} \mathcal{N}(x_n; \hat{\eta}_1^{(i)} + \hat{A}(i)x_n^P, \hat{R}(i))}{\hat{\pi}_1^{(i)} \mathcal{N}(x_n; \hat{\eta}_1^{(i)} + \hat{A}(i)x_n^P, \hat{R}(i))} = 1.
\]

(4.2)

Therefore, the complete log-likelihood, defined in (2.28), is unchanged during the GEM algorithm and the algorithm requires only one iteration to converge.

M-step

The state transition matrix estimation in the M-step is given by (2.23). The mean vector is assumed to be a zero vector, therefore \( \eta_1 = H \mu_1 \) is also a zero vector and we can substitute \( \hat{\eta}_{m}^{(i+1,j)} = 0 \) in (2.23). Thus, by substituting \( M = 1 \), (2.23) can be rewritten by:

\[
vec \left( \hat{A}^{(i+1,j+1)} \right) = \left[ C_{x^P x^P, 1}^{(i)} \otimes \hat{R}^{(j)} \right]^{-1} \cdot vec \left[ \hat{R}^{(j)} C_{x x^P, 1}^{(i)} \right]
\]

(4.3)

Using \( \gamma_{n,1} = 1 \), (2.24) and (2.25) are reduced to \( C_{x^P x^P, 1}^{(i)} = \sum_{n=1}^{N} x_n x_n^{PT} \) and \( C_{x x^P, 1}^{(i)} = \sum_{n=1}^{N} x_n x_n^{PT} \), respectively. Using these matrices and the Kronecker products inverse-property \( (A_1 \otimes A_2)^{-1} = A_1^{-1} \otimes A_2^{-1} \), one obtains:

\[
vec \left( \hat{A}^{(i+1,j+1)} \right) = \left[ \left( \sum_{n=1}^{N} x_n x_n^{PT} \right)^{-1} \otimes \hat{R}^{(j)} \right]^{-1} \cdot vec \left[ \hat{R}^{(j)} \sum_{n=1}^{N} x_n x_n^{PT} \right].
\]

(4.4)
If the matrices dimensions are suitable, the Kronecker products property \( \text{vec}(A_1A_2A_3) = (A_3^T \otimes A_1)\text{vec}(A_2) \) can be used:

\[
\text{vec} \left( \hat{A}^{(i+1,j+1)} \right) = \text{vec} \left[ \hat{R}^{(j)} (\sum_{n=1}^{N} x_n x_n^T) (\sum_{n=1}^{N} x_n^P x_n^{PT})^{-T} \right]
\]

\[
= \text{vec} \left[ (\sum_{n=1}^{N} x_n x_n^T) (\sum_{n=1}^{N} x_n^P x_n^{PT})^{-1} \right] .
\] (4.5)

This estimation is independent on the iteration indexes, \( j \) and \( i \), and thus, they can be omitted. Moreover, using the inverse-vec operator (4.5) can be reshaped to matrix form:

\[
\hat{A} = \hat{A}_{YW} = \left[ \sum_{n=1}^{N} x_n x_n^T \right] \left[ \sum_{n=1}^{N} x_n^P x_n^{PT} \right]^{-1}
\] (4.6)

which is the well known multi-dimensional Yule-Walker equations obtained in the linear predictive coding (LPC) [62], [63].

### 4.1.2 SISO-AR system identification

In this section, it is shown that for SISO-AR system identification the proposed method reduces to the approach presented in [51] for SISO-AR system identification with GMM-distributed sources.

For SISO-AR system with single GMM-distributed source, our model can be rewritten as:

\[
x_n = a^T x_n^P + s_n \quad \forall n = 1, \ldots, N
\] (4.7)

where \( n \) represents the time index, \( \{s_n\}_{n=1}^{N} \) is i.i.d. GMM-distributed sequence (termed the driving process or driving noise). The sequence \( \{x_n\}_{n=1}^{N} \) denotes the observation samples, and the past samples vector, \( x_n^P \), is defined as \( x_n^P = (x_{n-1}, \ldots, x_{n-P})^T \) where \( P \) is the AR order. The state transition matrix in this case is reduced to the AR coefficients
vector, \( \mathbf{a} = (a_1, \ldots, a_P)^T \). The GEM algorithm for the SISO case is derived from the unstructured GEM method, presented in Section 2.2.

Using (2.15)-(2.16), the E-step for the SISO case is

\[
U(\theta_u, \hat{\theta}_u^{(i)}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \left[ \pi_m N(x_n; \eta_m + \mathbf{a}^T \mathbf{x}_n^P, \sigma_m^2) \right]
\]

(4.8)

where the mean, variance and the weighting coefficient of the \( m \)th Gaussian of the single source are denoted by \( \eta_m, \sigma_m^2 \), and \( \phi_m \), respectively, and

\[
\gamma_{n,m}^{(i)} = \frac{\hat{\pi}_m^{(i)} N \left( x_n; \hat{\eta}_m^{(i)} + \hat{\mathbf{a}}^{(i)}^T \mathbf{x}_n^P, \hat{\sigma}_m^{(i)^2} \right)}{\sum_{m'=1}^{M} \hat{\pi}_{m'}^{(i)} N \left( x_n; \hat{\eta}_{m'}^{(i)} + \hat{\mathbf{a}}^{(i)}^T \mathbf{x}_n^P, \hat{\sigma}_{m'}^{(i)^2} \right)}.
\]

(4.9)

Using (2.20)-(2.23), the M-step in the GEM algorithm is as follows:

The mixture means, \( \{\eta_m\}_{m=1}^{M} \) are estimated by

\[
\hat{\eta}_m^{(i+1,j+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(i)} \left( x_n - \hat{\mathbf{a}}^{(i+1,j)^T} \mathbf{x}_n^P \right)}{\sum_{n=1}^{N} \gamma_{n,m}^{(i)}}, \quad \forall m = 1, \ldots, M.
\]

(4.10)

The mixture variances, \( \{\sigma_m^2\}_{m=1}^{M} \), are estimated by

\[
\hat{\sigma}_m^{2(i+1,j+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(i)} \left( x_n - \hat{\eta}_m^{(i+1,j+1)} - \hat{\mathbf{a}}^{(i+1,j)^T} \mathbf{x}_n^P \right)^2}{\sum_{n=1}^{N} \gamma_{n,m}^{(i)}}
\]

(4.11)

\( \forall m = 1, \ldots, M \). The state transition matrix estimation, in the SISO case, is reduced to coefficients vector estimation, and (2.22) can be written as

\[
vec \left( \hat{\mathbf{a}}^{(i+1,j+1)} \right) = \left( \sum_{m=1}^{M} \mathbf{C}_{x^P \mathbf{x}^P, \mathbf{m}}^{(i)} \otimes \hat{\sigma}_m^{2(i+1,j+1)-1} \right)^{-1} vec \left[ \sum_{m=1}^{M} \hat{\sigma}_m^{2(i+1,j+1)-1} \left( \mathbf{C}_{\mathbf{x}^P, \mathbf{m}}^{(i)} - \hat{\eta}_m^{(i+1,j+1)} \zeta_{\mathbf{x}^P, \mathbf{m}}^{(i)} \right) \right]
\]

(4.12)

where \( \mathbf{C}_{x^P \mathbf{x}^P, \mathbf{m}}^{(i)} \triangleq \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \mathbf{x}_n^P \mathbf{x}_n^{P^T} \) is a \( P \times P \) matrix, \( \mathbf{C}_{\mathbf{x}^P, \mathbf{m}}^{(i)} \triangleq \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \mathbf{x}_n \mathbf{x}_n^{P^T} \) is a \( 1 \times P \) vector, and \( \zeta_{\mathbf{x}^P, \mathbf{m}}^{(i)} \triangleq \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \mathbf{x}_n^{P^T} \) is also a \( 1 \times P \) vector. It is easy to see that the \( vec \) operator and the Kronecker product are unnecessary. Thus, (4.12) can be rewritten as

\[
\hat{\mathbf{a}}^{(i+1,j+1)} = \left( \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \frac{\mathbf{x}_n^P \mathbf{x}_n^{P^T}}{\hat{\sigma}_m^{2(i+1,j+1)}} \right)^{-1} \left( \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \frac{(x_n - \eta_m) \mathbf{x}_n^{P^T}}{\hat{\sigma}_m^{2(i+1,j+1)}} \right)
\]

(4.13)
The GEM algorithm, presented in equations (4.8)-(4.13), is identical to the approach described in [51]. Thus, the proposed method for unstructured input in Chapter 2 extends the approach presented in [51] for SISO-AR system identification to MIMO-AR systems.

4.2 Convergence

In this section, the convergence properties of the proposed algorithms are investigated and illustrated.

Let define the ML estimate of $\theta$ as $\hat{\theta} = \arg \max_\theta \mathcal{L}(x; \theta)$, $\theta \in \Omega$ where $x$ is the observed data and $\mathcal{L}(\cdot)$ is the log-likelihood function. Convergence of the GEM algorithm is studied in two parts: 1) convergence of $\mathcal{L}(x; \hat{\theta}^{(i)})$, 2) convergence of $\hat{\theta}^{(i)}$, $i = 1, 2, \ldots$. The convergence of the EM (or GEM) algorithm has been studied in depth in [52], [53]. We use earlier results to prove convergence for the proposed algorithms.

Wu in [53] narrowed down the scope of GEM problem to problems that satisfying the following three conditions, named Wu’s regularity conditions:

- $\Omega \in \mathcal{R}^D$.
- $\Omega_{\theta^{(0)}} = \{ \theta \in \Omega; \mathcal{L}(x; \theta) \geq \mathcal{L}(x; \theta^{(0)}) \}$ is compact $\forall \mathcal{L}(x; \theta^{(0)}) > -\infty$.
- $\mathcal{L}(x; \theta)$ is continuous in $\Omega$ and differentiable in the interior of $\Omega$.

These three conditions are all satisfied in our model. Thus, $\{ \mathcal{L}(\theta^{(k)}) \}$ is upper bounded for any $\theta^{(0)} \in \Omega$.

The proof of convergence of the log-likelihood is as follows. Any EM or GEM sequence, $\{\theta^k\}$, increases the likelihood and if $\mathcal{L}(\theta^{(k)})$ is upper bounded, it converges to some $L^*$ [52]. If $Q(\theta, \phi)$ is continuous in both $\theta$ and $\phi$, $L^*$ is a stationary point of $\mathcal{L}$. The continuity
of $Q$ holds for the case of a curved exponential family in general, and in particular in the GMM case. If $\theta^{(k)}$ convergence to some point $\theta^*$, then $\theta^*$ is a stationary point under some continuity conditions. If $||\theta^{(k+1)} - \theta^{(k)}|| \to 0$ as $k \to \infty$ and the set of stationary points with a given $L$ value is discrete, then $\theta^{(k)}$ convergence to a stationary point. For a regular exponential family, this condition is satisfied by the EM or GEM algorithms.

In the CA method, if $Q(\theta, \hat{\theta})$ is continuously differentiable over $\theta$ and the minimum of $Q(\theta, \hat{\theta})$ w.r.t. each coordinate is uniquely attained, then every limit point of $\hat{\theta}^{(i)}$ is a stationary point [60].

In general, if the log-likelihood has several maxima and stationary points, convergence of the EM sequence to either type of points depends on the initialization [53]. The convergence rate of an EM algorithm is inversely related to the Fisher information of its complete-data space [52]. In Chapter 5, the convergence for specific simulations are shown, in order to examine the typical convergence patterns of the log-likelihood for the proposed algorithm.
Chapter 5

Simulation results

In this chapter, the performances of the proposed methods are evaluated via simulations. The simulations for Chapters 2 and 3 are presented in Sections 5.1 and 5.2, respectively.

5.1 MIMO-AR System Identification and BSS

The performances of the proposed system identification and BSS techniques, presented in Chapter 2, are evaluated via simulations with synthetic and speech data. Estimation performance of the state transition matrix $\mathbf{A}$ by the structured estimation is compared to the unstructured estimation, presented in Section 2.2, and to the multi-dimensional Yule-Walker equations, presented in (2.27). The estimation performance of $\mathbf{A}$ is evaluated via the normalized RMSE, defined as $\hat{\mathbf{A}}_{RMSE} = \frac{||\mathbf{\hat{A}} - \hat{\mathbf{A}}||_F}{||\mathbf{A}||_F}$, where $|| \cdot ||_F$ denotes the Frobenius norm and $\mathbf{\hat{A}} = [\mathbf{I}_L; - \mathbf{A}]$ is the extended transition matrix which includes also the zero-order coefficients matrix.

Estimation performance of the input mixing matrix $\mathbf{H}$ by the structured estimation is compared to the separation matrix estimated by the well known instantaneous BSS
method, FastICA [3], applied to the whitened signal where the whitening procedure is based on the estimated $A$ via multi-dimensional Yule-Walker equations. In this chapter, this method is denoted by “Yule-Walker+FastICA”. The estimation performance of $H$ is evaluated via the normalized RMSE, defined as $\hat{H}_{RMSE} = \frac{||H - \hat{H}||_F}{||H||_F}$. The channel estimation performance is also evaluated by comparing the real and estimated system impulse response in the time and frequency domains.

The performance of the BSS algorithm is evaluated in terms of SIR [77], defined as the ratio between the target signal power to the interference signal power: $SIR = 10 \log_{10} \left( \sum_{l=1}^{L} \frac{|s_l^H s_l|^2}{||s_l||^2 ||s_l||^2 - |s_l^H s_l|^2} \right)$ where $s_l = [s_{l,1}, \ldots, s_{l,N}]^T$. The SIR takes into account the fact that, in general, using BSS it is only possible to recover the sources up to a gain factor $\alpha$. It is easy to check that if $\hat{s}_k = \alpha s_k$ the SIR is infinite. In addition, when the estimated source is orthogonal to the true source the SIR is equal to zero. The BSS performances of the following methods is compared:

- The proposed structured GEM method.
- The “Yule-Walker+FastICA” method, defined in the previous section.
- The quadratic higher-order criterion for convolutive BSS method described in [73] with the toolbox available in [74].
- The convolutive method presented in [75].

In the example presented here, the following simple method for initialization of the EM algorithm is used. The matrix $\hat{A}^{(0)}$ is computed using the multi-dimensional Yule-Walker equations (2.27) and the weighting coefficients are all set to $\frac{1}{M}$. The initial structured covariance matrices are set to $\hat{R}_m^{(0)} = S \cdot 2^{(m-M/2)}$, where $S = \text{cov}(x - \hat{A}^{(0)} x_n^p)$ is the
estimated sample covariance matrix of the whitened signal. In the structured algorithm, \( \hat{B}^{(0)} \) is set to be a matrix that jointly diagonalizes the covariance matrices, \( \hat{R}^{(0)}_m \).

### 5.1.1 Simulation results with synthetic data

In all the simulations, the case of \( L = 2 \) sources and sensors was considered. The source signals were synthesized and mixed by

\[
H = \begin{bmatrix}
0.79 & 0.28 \\
-0.55 & 0.72
\end{bmatrix}
\]

An AR process of order \( P = 6 \) was generated by the following state transition matrix

\[
A = 10^{-2} \begin{bmatrix}
0.27 & -0.06 & -0.21 & 0.04 & 5.43 & -17.61 & -16.52 & -2.41 & 13.67 & -31.6 & -24.8 & -11.5 \\
14.51 & -3.06 & 5.27 & -0.91 & 13.85 & 32.51 & 1.14 & 2.92 & 6.17 & -2.85 & 24.3 & -2.3
\end{bmatrix}
\]

The estimation performances were evaluated using 100 Monte-Carlo trials. In the first example, each one of the sources was zero-mean GMM-distributed with order \( M = 3 \). The weighting coefficients and the variances of the GMM-distributions for both sources were set as follows: \( \{\phi_{1,m}\}_{m=1}^3 = (0.2, 0.6, 0.2), \{\mu_{1,m}\}_{m=1}^3 = (0, 0, 0), \sigma_{1,m}^2 = \{3, 30, 300\} \), \( \{\phi_{2,m}\}_{m=1}^3 = (0.55, 0.15, 0.3), \{\mu_{2,m}\}_{m=1}^3 = (0, 0, 0), \sigma_{2,m}^2 = \{10, 100, 1\} \).

In all the simulations, the GMM order was determined using MDL criterion [67], implemented by the greedy EM algorithm for GMM parameter estimation [76]. The greedy EM algorithm is presented in Appendix D. The number of free adjustment parameters in the MDL criterion is obtained by counting the number of degrees of freedom of the space spanned by \( \theta(M) \). The number of free adjustment parameters of the proposed model is:

\[
\begin{align*}
\text{for matrix } H & : L^2 \\
\text{for matrix } A & : L^2 P \\
\text{for weighting coefficients} & : M - 1 \\
\text{for mean vectors and diagonal covariance matrices} & : 2LM
\end{align*}
\]  

(5.1)

Fig. 5.1 shows that the proposed algorithm outperforms the multi-dimensional Yule-Walker, in the state transition matrix estimation. This figure also shows that the proposed algorithm outperforms the “Yule-Walker+FastICA” method in the input mixing matrix.
estimation. The bottom plot in Fig. 5.1 shows the source separation performance. Fig. 5.2 depicts the impulse response of the real MIMO-AR system and of the estimated system. It shows that the proposed algorithm enables good reconstruction of the channel impulse response. Note that the channel impulse response can only be determined up to a complex scaling factor. The scatter plots of the source, observation and estimated source signals are depicted in Fig. 5.3. One can observe that due to the scaling and permutation ambiguities of the BSS problem, the estimated source signals are scaled and permuted.

Figure 5.1: Performance of the state transition matrix estimation by the structured-GEM and Yule-Walker algorithms (upper), performance of the input mixing matrix estimation by the structured-GEM and “Yule-Walker+FastICA” method (middle), and the corresponding BSS (lower) for GMM-distributed sources.
The unstructured estimation can also be used in the degenerative case of more sources than sensors. In order to demonstrate the performance of the proposed unstructured algorithm in this case, it was tested with 2 sensors and 3 zero-mean GMM-distributed sources with parameters \( \{\phi_{1,m}\}_{m=1}^{3} = (0.2, 0.6, 0.2) \), \( \{\mu_{1,m}\}_{m=1}^{3} = (0, 0, 0) \), \( \sigma_{1,m}^2 = \{3, 30, 300\} \), \( \{\phi_{2,m}\}_{m=1}^{3} = (0.55, 0.15, 0.3) \), \( \sigma_{2,m}^2 = \{1, 10, 100\} \), \( \{\mu_{2,m}\}_{m=1}^{3} = (0, 0, 0) \), \( \{\phi_{3,m}\}_{m=1}^{3} = (0.33, 0.33, 0.34) \), \( \{\mu_{3,m}\}_{m=1}^{3} = (0, 0, 0) \), \( \sigma_{3,m}^2 = \{100, 200, 3\} \). An AR process of order \( P = 6 \) was generated by the aforementioned state transition matrix with the input mix-
Figure 5.3: Scatter plot of the source signals (upper), the mixed signals (middle), and the estimated source signals (lower), for Gaussian-distributed sources with $N = 1000$ samples.

The mixing matrix $H = \begin{bmatrix} 0.39 & -0.64 & 0.44 \\ 0.880 & -0.5596 & -0.9499 \end{bmatrix}$. Fig. 5.4 shows, as before, that the proposed algorithm outperforms the multi-dimensional Yule-Walker, in estimation of the state transition matrix.

As mentioned above, the finite-order GMM can closely approximate a large family of PDFs. In order to demonstrate the performance of the proposed algorithm with non-Gaussian signals which are not GMM-distributed, it was tested with Laplacian-distributed sources with scaling parameters 0.5 and 1.5, and with the same MIMO-AR system. Fig. 5.5 shows, as in the previous examples, that the proposed algorithm outperforms the
Figure 5.4: Performance of the state transition matrix estimation of the unstructured MIMO-AR-GMM and the multi-dimensional Yule-Walker algorithms with 3 GMM-distributed sources and 2 sensors.

multi-dimensional Yule-Walker, in estimation of the state transition matrix and results in a higher SIR in the source separation process. The performance of the multi-dimensional Yule-Walker in this scenario is closer to the proposed GEM algorithm, since the Laplacian distribution is closer to the Gaussian distribution for which the multi-dimensional Yule-Walker is optimal. Fig. 5.6 shows that the proposed algorithm enables good reconstruction of the channel response for Laplacian-distributed sources as well. Note that the channel impulse response can be determined up to a complex scaling factor. The scatter plots for this example are depicted in Fig. 5.7.

Fig. 5.8 shows the convergence of the proposed algorithm for the aforementioned scenarios with GMM-distributed and Laplacian-distributed sources for $N = 500$, $N = 1000$, and $N = 5000$ samples.
5.1.2 Simulation results with audio signals

In the second example, the performance of the proposed algorithm was evaluated using speech signals from the ICA’99 synthetic benchmarks [65] database, convolved with the MIMO-AR system from the previous example, to obtain 2 mixed signals. The sampling frequency was $f_s = 22.05 kHz$. The observed mixed signals were segmented into frames of 0.3, 0.5, 1, 1.5, and 2 seconds with 50% overlapping. The separation algorithms were
Figure 5.6: Impulse response of the MIMO-AR system, from source 1 to sensor 1, compared to the estimated response by the structured algorithm for Laplacian-distributed sources and $N = 1000$ samples.

applied to each frame independently and the performances were evaluated based on the SIR averaged over the frames. A good qualitative recovery is confirmed by subjective listening to the recovered audio signals and by the average SIR performance presented in Table 5.1. Note that in this simulation there exists a modeling mismatch between the generated data and the model presented in (2.1). Speech signals are known to be colored, which can be modeled as the output of a SISO-AR system. The overall system for speech signal mixing, can be modeled by the composed system in Fig. 5.9. Therefore,
Figure 5.7: Scatter plot of the source signals (upper), the mixed signals (middle), and the estimated source signals (lower), for Laplacian-distributed sources with $N = 1000$ samples.

Table 5.1: The average SIR for audio sources mixed by MIMO-AR system with different BSS methods

<table>
<thead>
<tr>
<th>Time [sec]</th>
<th>Average SIR [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FastICA</td>
</tr>
<tr>
<td>0.3</td>
<td>7.31</td>
</tr>
<tr>
<td>0.5</td>
<td>8.56</td>
</tr>
<tr>
<td>1</td>
<td>8.94</td>
</tr>
<tr>
<td>1.5</td>
<td>9.15</td>
</tr>
<tr>
<td>2</td>
<td>9.34</td>
</tr>
</tbody>
</table>
Figure 5.8: Convergence of the log-likelihood for simulation of 2 GMM-distributed (left) and Laplacian-distributed (right) sources, for $N = 500, 1000, 5000$ samples.

the proposed GEM method, with sufficiently high model order, estimates the virtual input signals, denoted by $s^0_l \forall l = 1, \ldots, L$ in Fig. 5.9 and not the desired speech signals, $s_l \forall l = 1, \ldots, L$. Thus, in the simulations with audio signals we need an additional stage in which the different source signals, $s_l \forall l = 1, \ldots, L$, are estimated as they are received at a reference sensor, which can be arbitrarily chosen. This procedure is performed by using the channel impulse responses from each source to the reference sensor, and convolving $\hat{s}^0_l \forall l = 1, \ldots, L$ with the appropriate estimated channel impulse responses.
5.2 Frequency-selective MIMO channel estimation

In this section, the performances of the proposed joint channel tracking and equalization algorithm, presented in Chapter 3, is evaluated via simulations. The equalization performance of the system is measured in terms of SER. The SER is computed by comparing the source signals symbols to the symbols in the equalized signal, obtained by (3.16). The SER performance is compared to receivers with perfect CSI. The SER for perfect CSI is computed by comparing the source signals symbols to the symbols in the equalized signal, obtained by substituting $A$ and $H$ in (3.16). The simulations are performed with white additive Gaussian noise, introducing mismatch with the noiseless model presented in (3.1). The SNR is set to be the same for both users.

In the simulations, the case of $L = 2$ sources and sensors was considered with $N = 250$ symbols per user. The frequency-selective fading channel is modeled by a MIMO-AR system of order $P = 15$. The two sources are QPSK and BPSK modulated and the input mixing matrix is randomly set to be $H = \begin{bmatrix} 0.3395 & 0.5193 \\ 0.9387 & -0.2397 \end{bmatrix}$ while the state transition matrix, $A$, is randomly chosen and its stability was verified. In the example presented...
here, the following simple method for initialization of the EM algorithm is used. The matrix $\hat{A}^{(0)}$ was computed using the multi-dimensional Yule-Walker equations, and the matrix $\hat{H}^{(0)}$ is set to be an identity matrix. For performance evaluation, we performed a total of 5000 independent trials.

Fig. 5.10 compares the symbol detection performance using perfect CSI with those using estimates obtained by the proposed algorithm. The figure shows that by using the proposed MIMO channel estimation technique one obtains SER close to the case of perfect CSI. Fig. 5.11 shows the scatter plots of the two sources, one of the noisy mixture signals, and the equalized signals before quantization. Fig. 5.12 compares the symbol detection performance as a function of data lengths using perfect CSI and using the estimates obtained by the proposed algorithm, for $SNR = 8dB$ and AR order $P = 8$.

Figure 5.10: SER vs. SNR obtained with the proposed channel estimation technique compared with perfect CSI for QPSK and BPSK source signals.
Figure 5.11: Scatter plot of the two sources, the mixture in the first sensor, and the equalized signals, for SNR=12dB.

It can be seen that, in this case, for $N \geq 1000$ the estimated CSI convergence to its final value and the SER obtained by the proposed MIMO channel estimation method coincides with the SER obtained with known CSI.

Fig. 5.13 depicts the impulse response of the real MIMO-AR system and of the estimated system. It shows that the proposed algorithm enables good reconstruction of the channel impulse response.
Figure 5.12: SER vs. data length obtained with the proposed channel estimation technique compared with perfect CSI for QPSK (higher) and BPSK (lower) source signals, for SNR=8dB.
Figure 5.13: Impulse response of the MIMO-AR system, from source 1 to sensor 1, compared to the estimated response by the proposed EM algorithm for two QPSK and BPSK modulated sources and $N = 5000$ samples.
Chapter 6

Summary

6.1 Discussion and conclusion

In this thesis, new time-domain ML-based algorithms for BSS of convolutive mixtures, modeled by MIMO-AR, were presented. In Chapter 2, two different approaches that assume GMM distribution of the sources were described, with structured or unstructured input signals. In both methods, the state transition matrix and the distribution parameters are jointly estimated by applying the GEM algorithm. The structured-input method also estimates the input mixing matrix and can be utilized to separate the sources. The proposed unstructured method extends the approach presented in [51] for SISO-AR system identification to MIMO-AR systems. It is shown that for zero-mean Gaussian-distributed sources the unstructured estimation of the state transition matrix, $A$, reduces to the well known multi-dimensional Yule-Walker equations presented in (2.27).

In Chapter 3, a finite-alphabet based channel estimation technique for frequency-selective fading channel has been developed. The state transition matrix and the input mixing matrix are jointly estimated by applying the EM algorithm. The proposed method
reduces to the mixing matrix estimation presented in [45] for flat-fading channel estimation without noise.

The methods were tested via simulations with synthetic and speech data. It is shown that the proposed methods outperform the multi-dimensional Yule-Walker solution for the state transition matrix estimate, and provide good separation performances for convolutive mixtures. For communication channel estimation and equalization, it is shown that for SNR of 8 dB and data length higher than 1000 samples, the estimated CSI by the proposed method coincides with perfect CSI.

6.2 Further research

In this work, the GMM was utilized for blind separation of convolutive noiseless linear mixture of independent sources, where the number of sensors was equal or greater than the number of sources. Therefore, future research may address utilization of GMM for blind source separation, where the number of sources is greater than the number of sensors.

The noiseless MIMO-AR model presented in this work (2.1) may be extended to model with measurement noise. For this purpose, the model should be

\[
x_n = A\mathbf{x}_n^P + \mathbf{Hs}_n \quad n = 1, \ldots, N
\]
\[
\mathbf{u}_n = \mathbf{x}_n + \mathbf{v}_n
\]

where \(\mathbf{v}_n\) is an additive noise vector and \(\mathbf{u}_n\) is the observation vector at time instance \(n\).

As described in Chapter 5, the simulations with speech signals introduce a mismatch with the model presented in (2.1). Therefore, for real audio signal applications, the model has to deal with the composed system described in Fig. 5.9. In addition, the performances of the proposed algorithm should evaluated using real-time sound-separation experiments.
with acoustic source signals recorded in the real room.

The channel estimation and equalization for communication channel, presented in Chapter 3, can be utilized for channel tracking. In this case, the log-likelihood function should be extended to include an a-priori information about the channel change. For example, the channel variation can be modeled as a Markov process. In addition, the proposed algorithm with known FA should be tested with popular communication standards and channel models. This method can also be implemented for known distribution with non-infinitesimal variances. Such a model is useful, for example, for audio signals.
Appendix A

Mathematical proofs

A.1 Minimization of (2.17) w.r.t. \( \{\pi_m\}_{m=1}^M \)

In this appendix, it is shown that (2.17) is minimized as described in (2.19). To find the expression for \( \{\pi_m\}_{m=1}^M \), we introduce the Lagrange multiplier \( \lambda \) with the constraint \( \sum_{m=1}^{M} \pi_m = 1 \).

\[
\frac{\partial}{\partial \pi_m} \left[ \sum_{n=1}^{N} \sum_{m'=1}^{M} \gamma_{n,m'}^{(i)} \log \pi_{m'} + \lambda \left( \sum_{m'=1}^{M} \pi_{m'} - 1 \right) \right] = \sum_{n=1}^{N} \gamma_{n,m}^{(i)} \frac{1}{\hat{\pi}_m} + \lambda = 0
\]

\[
\Rightarrow \sum_{n=1}^{N} \gamma_{n,m}^{(i)} = -\hat{\pi}_m \lambda. \quad (A.1)
\]

Summing both sides over \( M \), yields:

\[
\sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(i)} = - \sum_{m=1}^{M} \hat{\pi}_m \lambda. \quad (A.2)
\]

Since \( \sum_{m=1}^{M} \gamma_{n,m}^{(i)} = 1 \) \( \forall n = 1, \ldots, N \) and \( \sum_{m=1}^{M} \hat{\pi}_m = 1 \), we obtain \( \lambda = -N \), resulting in:

\[
\hat{\pi}_m = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,m}^{(i)}. \quad (A.3)
\]
A.2 Minimization of (2.18) w.r.t. \( \{ \eta_m \}_{m=1}^M \)

In this appendix, \( \hat{\eta}_{m}^{(i+1,j+1)} \) is obtained by minimization of (2.18) w.r.t. the mean vectors, \( \{ \eta_m \}_{m=1}^M \), after substituting \( A = \hat{A}^{(i+1,j)} \) and \( R_m = \hat{R}_m^{(i+1,j)} \), \( \forall m = 1, \ldots, M \), as described in (2.20).

\[
0 = \frac{\partial}{\partial \eta_m} \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}^{(i)} \log \left[ N \left( x_n; \eta_m + \hat{A}^{(i+1,j)}x_n^P, \hat{R}_m^{(i+1,j)} \right) \right] \\
= - \left( \sum_{n=1}^N \frac{\partial}{\partial \eta_m} \left( x_n - \eta_m + \hat{A}^{(i+1,j)}x_n^P \right)^T \hat{R}_m^{(i+1,j)^{-1}} \left( x_n - \eta_m + \hat{A}^{(i+1,j)}x_n^P \right) \right) \\
= \sum_{n=1}^N \left( \eta_m - x_n + \hat{A}^{(i+1,j)}x_n^P \right)^T \left( \hat{R}_m^{(i+1,j)^{-1}} + \hat{R}_m^{(i+1,j)^{-T}} \right) .
\]

where the last equality is obtained using (C.9). Since \( \hat{R}_m^{(i+1,j)} \) is a symmetric matrix, \( \hat{R}_m^{(i+1,j)^{-1}} = \hat{R}_m^{(i+1,j)^{-T}} \), and from (A.4) we obtain

\[
\hat{\eta}_{m}^{(i+1,j+1)} = \frac{\sum_{n=1}^N \gamma_{n,m}^{(i)} \left( x_n - \hat{A}^{(i+1,j)}x_n^P \right)}{\sum_{n=1}^N \gamma_{n,m}^{(i)}}, \quad \forall m = 1, \ldots, M .
\]

A.3 Minimization of (2.18) w.r.t. \( \{ R_m \}_{m=1}^M \)

In this appendix, \( \hat{R}_m^{(i+1,j+1)} \) is obtained by minimization of (2.18) w.r.t. the covariance matrices, \( \{ R_m \}_{m=1}^M \), after substituting \( A = \hat{A}^{(i+1,j)} \) and \( \eta_m = \hat{\eta}_{m}^{(i+1,j+1)} \), \( \forall m = 1, \ldots, M \) as described in (2.21).

\[
0 = \frac{\partial}{\partial R_m} \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}^{(i)} \log \left[ N \left( x_n; \hat{\eta}_{m}^{(i+1,j+1)} + \hat{A}^{(i+1,j)}x_n^P, R_m \right) \right] \\
= - \frac{\partial}{\partial R_m} \log |R_m| - \sum_{n=1}^N \gamma_{n,m}^{(i)} \frac{\partial}{\partial R_m} \hat{z}_{n,m}^{(i+1,j+1)^T} R_m^{-1} \hat{z}_{n,m}^{(i+1,j+1)} \\
= \hat{R}_m^{(i+1,j+1)^{-1}} - \sum_{n=1}^N \gamma_{n,m}^{(i)} \hat{R}_m^{(i+1,j+1)^{-T}} \hat{z}_{n,m}^{(i+1,j+1)^T} \hat{R}_m^{(i+1,j+1)^{-1}} .
\]
where \( \hat{z}_{n,m}^{(i+1,j+1)} \triangleq (x_n - \hat{\eta}_{m}^{(i+1,j+1)} - \hat{A}^{(i+1,j)}x_n^P) \) and the last equality is obtained using (C.10) and (C.29). Since \( \hat{R}_m^{(i+1,j+1)} \) is a symmetric matrix, \( \hat{R}_m^{(i+1,j+1)^{-1}} = \hat{R}_m^{(i+1,j+1)^{-T}} \), and from (A.6) we obtain

\[
\hat{R}_m^{(i+1,j+1)} = \sum_{n=1}^{N} \gamma_{n,m} \frac{(x_n - \hat{\eta}_{m}^{(i+1,j+1)} - \hat{A}^{(i+1,j)}x_n^P)(x_n - \hat{\eta}_{m}^{(i+1,j+1)} - \hat{A}^{(i+1,j)}x_n^P)^T}{\sum_{n=1}^{N} \gamma_{n,m}} \tag{A.7}
\]

\( \forall m = 1, \ldots, M. \)

### A.4 Derivative of (2.18) w.r.t. A

In this appendix, it is shown that by equating the partial derivative of (2.18) w.r.t. the state transition matrix, A, after substituting \( \{\hat{\eta}_m^{(i+1,j+1)}, \hat{R}_m^{(i+1,j+1)^{-1}}\}_{m=1}^{M} \), one obtains (2.22).

\[
0 = \frac{\partial}{\partial A} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \log \left[ N \left( x_n; \hat{\eta}_m^{(i+1,j+1)} + Ax_n^P, \hat{R}_m^{(i+1,j+1)} \right) \right]
\]

\[
= - \sum_{n=1}^{N} \gamma_{n,m} \frac{\partial}{\partial A} \left( x_n - \hat{\eta}_m^{(i+1,j+1)} - Ax_n^P \right)^T \hat{R}_m^{(i+1,j+1)^{-1}} \left( x_n - \hat{\eta}_m^{(i+1,j+1)} - Ax_n^P \right)
\]

\[
= - \sum_{n=1}^{N} \gamma_{n,m} x_n^P \left( x_n - \hat{\eta}_m^{(i+1,j+1)} - Ax_n^P \right)^T \left( \hat{R}_m^{(i+1,j+1)^{-1}} + \hat{R}_m^{(i+1,j+1)^{-T}} \right) \tag{A.8}
\]

where the last equality is obtained using (C.9). Since \( \hat{R}_m^{(i+1,j+1)} \) is a symmetric matrix, \( \hat{R}_m^{(i+1,j+1)^{-1}} = \hat{R}_m^{(i+1,j+1)^{-T}} \), thus, we obtain

\[
\sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \hat{R}_m^{(i+1,j+1)^{-1}} \left( x_n - \hat{\eta}_m^{(i+1,j+1)} \right) x_n^P = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \hat{R}_m^{(i+1,j+1)^{-1}} Ax_n x_n^P. \tag{A.9}
\]
A.5 Minimization of (2.31) w.r.t. \( \{C_m\}^M_{m=1} \)

In this appendix, \( \hat{C}^{(i+1,j+1)}_m \) is obtained by minimization of (2.31) w.r.t. the diagonal covariance matrices, \( \{C_m\}^M_{m=1} \), under the constraint of diagonality, after substituting

\[
\{\mu_m = \hat{\mu}^{(i+1,j)}_m\}^M_{m=1}, \quad B = \hat{B}^{(i+1,j)} \quad \text{and} \quad A = \hat{A}^{(i+1,j)}.
\]

For simplicity, in this proof, the notations \( \hat{A}, \hat{\mu}_m, \) and \( \hat{B} \) are used instead of \( \hat{A}^{(i+1,j)}, \hat{\mu}^{(i+1,j)}_m, \) and \( \hat{B}^{(i+1,j)} \), respectively. The maximization of (2.31) w.r.t. \( \{C_m\}^M_{m=1} \) after the aforementioned substitutions is

\[
\left\{ \hat{C}^{(i+1)} \right\}^M_{m=1} = \arg \max_{\{C_m\}^M_{m=1}} \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}^{(i)} \log \left[ N \left( x_n; \hat{B}^{-1} \hat{\mu}_m + \hat{A} x_n^P \hat{B}^{-1} C_m \hat{B}^{-T} \right) \right]
\]

\[
= \arg \max_{\{C_m\}^M_{m=1}} \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}^{(i)} \left[ \text{const} - \log |C_m| - tr \left( \hat{R}_m^{(i+1,j)} \hat{B}^T C_m^{-1} \hat{B} \right) \right]
\]

\[
= \arg \max_{\{C_m\}^M_{m=1}} \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}^{(i)} \left[ \text{const} - \log |C_m| - tr \left( D_m C_m^{-1} \right) \right],
\]

(A.10)

where \( \hat{R}_m^{(i+1,j)} \) is defined in (2.33) and \( D_m = \hat{B} \hat{R}_m^{(i+1,j)} \hat{B}^T \). The matrices \( \{C_m\}^M_{m=1} \) are diagonal matrix with the diagonal values \( \{\lambda_m, l\}_{l=1}^L \forall l = 1, \ldots, L \). Therefore,

\[
\log |C_m| = \sum_{l=1}^L \log \lambda_{m,l}
\]

(A.11)

and

\[
tr \left( D C_m^{-1} \right) = \sum_{l=1}^L \frac{|D_m|_{ll}}{\lambda_{m,l}}
\]

(A.12)

where \( |D_m|_{ll} \forall l = 1, \ldots, L \) is the \( l \)-th diagonal element of \( D_m \). Substituting (A.11) and (A.12) into (A.10) gives

\[
\hat{\lambda}_{m,l} = \arg \max_{\lambda_{m,l}} \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}^{(i)} \left[ \text{const} - \sum_{l=1}^L \log \lambda_{m,l} - \sum_{l=1}^L \frac{|D_m|_{ll}}{\lambda_{m,l}} \right].
\]

(A.13)

Equating the partial derivative of (A.13) w.r.t. \( \lambda_{m,l} \) to zero, yields

\[
0 = -\frac{1}{\lambda_{m,l}^2} + \frac{|D_m|_{ll}}{\lambda_{m,l}^2}
\]

(A.14)
or
\[ \hat{\lambda}_{m,l} = [D_m]_{ll} . \]  
(A.15)

Rearranging all the estimated diagonal values in matrix form, gives
\[ \hat{C}_{m}^{(i+1,j+1)} = \text{DIAG}(D_m) \]  
(A.16)

and by substituting \( D_m = \hat{B} \hat{R}_m^{(i+1,j)} \hat{B}^T \), the estimated covariance matrix is
\[ \hat{C}_{m}^{(i+1,j+1)} = \text{DIAG}\left(\hat{B} \hat{R}_m^{(i+1,j)} \hat{B}^T\right) \quad \forall m = 1, \ldots, M . \]  
(A.17)

### A.6 Minimization of \( Q_2^{(i+1,j+1)} \) w.r.t. \( \hat{B} \)

In this appendix, it is shown that the minimization
\[ \hat{B}^{(i+1,j+1)} = \arg \min_{\hat{B}} Q_2^{(i+1,j+1)} , \]  
(A.18)

where
\[ Q_2^{(i+1,j+1)} \triangleq \sum_{m=1}^{M} \tilde{\pi}_m^{(i+1)} \left( KL_{\text{norm}} \left[ \hat{B} \hat{R}_m^{(i+1,j+1)} \hat{B}^T | \text{DIAG} \left( \hat{B} \hat{R}_m^{(i+1,j+1)} \hat{B}^T\right) \right] + KL_{\text{norm}} \left[ \text{DIAG} \left( \hat{B} \hat{R}_m^{(i+1,j+1)} \hat{B}^T\right) | \hat{C}_{m}^{(i+1,j+1)} \right] \right) , \]  
(A.19)

is achieved by
\[ \hat{B}^{(i+1,j+1)} = \left[ \Psi_1 b_1; \Psi_2 b_2; \ldots; \Psi_L b_L \right]^{-1} \]  
(A.20)

where \( \Psi_l \triangleq \sum_{m=1}^{M} \tilde{\pi}_m^{(i+1)} \frac{1}{b_l^T \hat{R}_m^{(i+1,j+1)} b_l} \hat{R}_m^{(i+1,j+1)} \) and \( b_l^T \) and \( \hat{b}_l^T \) are the \( l \)th rows of the matrices \( B \) and \( \hat{B}^{(i+1,j)} \), respectively.

The KL divergence of a zero-mean \( L \)-variate normal density with covariance matrix \( \Sigma_1 \) from a zero-mean \( L \)-variate normal density with covariance matrix \( \Sigma_2 \) can be expressed
as [50]: \( KL_{\text{norm}}(\Sigma_1|\Sigma_2) = \frac{1}{2} tr(\Sigma_2^{-1}\Sigma_1) - \frac{1}{2} \log|\Sigma_2^{-1}\Sigma_1| - \frac{1}{2}. \) Then

\[
KL_{\text{norm}} \left( \hat{B}\hat{R}_mB^T|\text{DIAG}(\hat{B}\hat{R}_mB^T) \right) + \frac{1}{2}L = \\
\frac{1}{2} \left[ tr \left( \text{DIAG}(\hat{B}\hat{R}_m^{(i+1,j+1)}B^T)^{-1}\hat{B}\hat{R}_mB^T \right) - \
\log|\text{DIAG}(\hat{B}\hat{R}_mB^T)^{-1}\hat{B}\hat{R}_mB^T| \right] 
\] (A.21)

where for simplicity, in this proof, the notations \( \hat{R}_m \) and \( \hat{B} \) are used instead of \( \hat{R}_m^{(i+1,j+1)} \) and \( \hat{B}^{(i+1,j)} \), respectively. By applying \( tr(\text{DIAG}(G)^{-1}G) = L, \forall G \in \mathcal{R}^L \), (A.21) is reduced to:

\[
KL_{\text{norm}} \left( \hat{B}\hat{R}_mB^T|\text{DIAG}(\hat{B}\hat{R}_mB^T) \right) = \\
\frac{1}{2} \log|\text{DIAG}(\hat{B}\hat{R}_mB^T)| - \frac{1}{2} \log|\hat{B}\hat{R}_mB^T|. 
\] (A.22)

In similar,

\[
KL_{\text{norm}} \left( \text{DIAG}(\hat{B}\hat{R}_mB^T)|\text{DIAG}(\hat{B}\hat{R}_m\hat{B}) \right) + \frac{1}{2}L = \\
\frac{1}{2} \left[ tr \left( \text{DIAG}(\hat{B}\hat{R}_m\hat{B})^{-1}\text{DIAG}(\hat{B}\hat{R}_mB^T) \right) - \
\log| \left( \text{DIAG}(\hat{B}\hat{R}_m\hat{B})^{-1}\text{DIAG}(\hat{B}\hat{R}_mB^T) \right) | \right] 
\] (A.23)

It can be seen that

\[
tr \left( \text{DIAG}(\hat{B}\hat{R}_m\hat{B})^{-1}\text{DIAG}(\hat{B}\hat{R}_mB^T) \right) = \sum_{l=1}^{L} \frac{b_l^T\hat{R}_m b_l}{b_l^T\hat{R}_m b_l} 
\] (A.24)

where \( b_l^T \) and \( \hat{b}_l^T \) are the \( l \)th rows of the matrices \( B \) and \( \hat{B}^{(i+1,j)} \), respectively. Using (A.24), equation (A.23) can be rewritten as

\[
KL_{\text{norm}} \left( \text{DIAG}(\hat{B}\hat{R}_mB^T)|\text{DIAG}(\hat{B}\hat{R}_m\hat{B}) \right) = \\
\frac{1}{2} \left[ \sum_{l=1}^{L} \frac{b_l^T\hat{R}_m b_l}{b_l^T\hat{R}_m b_l} - \log| \left( \text{DIAG}(\hat{B}\hat{R}_m\hat{B})^{-1}\text{DIAG}(\hat{B}\hat{R}_mB^T) \right) | - L \right]. 
\] (A.25)
By substitution of (A.22) and (A.25) into (A.19), one obtains
\[
Q_2^{(i+1,j+1)} = \frac{1}{2} \sum_{m=1}^{M} \tilde{\pi}_m^{(i+1)} \left( \sum_{l=1}^{L} \frac{b_l^T \hat{R}_m b_l}{b_l^T \hat{R}_m b_l} - \log |B \hat{R}_m B^T| - L \right). \tag{A.26}
\]
Let define \(\Psi_l \triangleq \sum_{m=1}^{M} \tilde{\pi}_m^{(i+1)} \frac{1}{b_l^T \hat{R}_m b_l} \hat{R}_m\). Then equation (A.26) can be formulated in the following manner:
\[
Q_2^{(i+1,j+1)} = \frac{1}{2} \sum_{l=1}^{L} b_l^T \Psi_l b_l - \log |B| + \text{const}. \tag{A.27}
\]
By equating the partial derivative of (A.27) w.r.t. \(b_l^T\) to zero one obtains
\[
\frac{\partial Q_2^{(i+1,j+1)}}{\partial b_l^T} = \frac{1}{2} b_l^T (\Psi_l + \Psi_l^T) - [B^{-1}]_l = 0, \quad \forall l = 1, \ldots, L. \tag{A.28}
\]
where \([\cdot]_l\) is the \(k\)th row of its matrix argument. It can be seen that the matrices \(\Psi_l \ l = 1, \ldots, L\) are symmetric, thus (A.28) can be rewritten as
\[
\hat{B}^{(i+1,j+1)^{-1}} = \left[ \Psi_1 b_1; \Psi_2 b_2; \ldots; \Psi_L b_L \right]. \tag{A.29}
\]

A.7 The partial derivative of (3.8) w.r.t. \(A\)

In this appendix, it is shown that equating the partial derivative of (3.8) w.r.t. \(A\) to zero, yields (3.12).

According to (3.8):
\[
U(\theta, \tilde{\theta}^{(i)}) = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \left[ \frac{1}{M} \mathcal{N}^C(x_n; A x_n^P + H \alpha_m; \epsilon^2 HH^H) \right]
\]
\[
= c - \frac{1}{\epsilon^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} (x_n - H \alpha_m - A x_n^P)^H (HH^H)^{-1} (x_n - H \alpha_m - A x_n^P) \tag{A.30}
\]
where \(c\) denotes a constant w.r.t. \(A\). Using (C.21), the complex derivatives w.r.t. \(A\) is
\[
\frac{\partial U(\theta, \tilde{\theta}^{(i)})}{\partial A} = -\frac{1}{\epsilon^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} x_n^P (x_n - H \alpha_m - A x_n^P)^H (HH^H)^{-1}. \tag{A.31}
\]
By equating (A.31) to zero, one obtains
\[ \hat{A} = \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} (x_n - H\alpha_m)x_n^{PH} \right) \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} x_n^{PH} \right)^{-1}. \] (A.32)

Since \( \sum_{m=1}^{M} \gamma_{n,m} = 1 \), (A.32) can be simplified as:
\[ \hat{A} = \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} (x_n - H\alpha_m)x_n^{PH} \right) \left( \sum_{n=1}^{N} x_n^{PH} \right)^{-1}, \] (A.33)
which is identical to (3.12) \( \square \).

**A.8 The partial derivative of (3.8) w.r.t. \( H \)**

In this appendix, it is shown that equating the partial derivative of (3.8) w.r.t. \( H \) to zero, yields (3.13).

According to (3.8):
\[
U \left( \theta, \hat{\theta}^{(i)} \right) = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \log \left[ \frac{1}{M} N^{C} (x_n; A x_n^{P} + H\alpha_m, \epsilon^2 H H^H) \right]
= c - \log |H H^H| - \frac{1}{\epsilon^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} (z_n - H\alpha_m)^H (H H^H)^{-1} \left( \begin{array}{c} z_n - H\alpha_m \end{array} \right)
= c - \log |H H^H| - \frac{1}{\epsilon^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \left[ z_n^H (H H^H)^{-1} z_n - z_n^H (H H^H)^{-1} H\alpha_m \right.
\left. - \alpha_m^H H^H (H H^H)^{-1} z_n + \alpha_m^H H^H (H H^H)^{-1} H\alpha_m \right] \quad (A.34)
\]

where \( c \) denotes a constant w.r.t. \( H \) and \( z_n \triangleq x_n - A x_n^P \). The matrix \( H \) is non-singular by our assumptions, and therefore, \( (H H^H)^{-1} = H^{-H} H^{-1} \). Thus
\[
U \left( \theta, \hat{\theta}^{(i)} \right) = c_2 - \log |H H^H| - \frac{1}{\epsilon^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \left[ z_n^H H^{-H} H^{-1} z_n - 2 Re \{ z_n^H H^{-H} \alpha_m \} \right]
\]

where \( c_2 = c + \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \alpha_m^H \alpha_m \) denotes a constant w.r.t. \( H \). Using the following derivatives:
• According to (C.33), \( \frac{\partial}{\partial H} \log |HH^H| = \frac{\partial}{\partial H} \log |H| + \frac{\partial}{\partial H} \log |H^H| = H^{-1} \),

• According to (C.27), \( \frac{\partial}{\partial H} (z_n^H H^{-H} H^{-1} z_n) = -(H^{-1} z_n z_n^H H^{-H} H^{-1})^T \),

• According to (C.27), \( \frac{\partial}{\partial H} (2\text{Re}\{z_n^H H^{-H} \alpha_m\}) = -H^{-1} z_n \alpha_m^H H^{-1} \),

the derivative of (A.35) w.r.t. \( H \) is

\[
\frac{\partial U(\theta, \hat{\theta}^{(i)})}{\partial H} = -H^{-1} + \frac{1}{\epsilon^2} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma^{(i)}_{n,m} \left[ H^{-1} z_n z_n^H H^{-H} H^{-1} - H^{-1} z_n \alpha_m^H H^{-1} \right]
\] (A.36)

By equating the above derivative to zero, one obtains:

\[
\epsilon^2 \hat{H}^{-1} = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma^{(i)}_{n,m} \left[ H^{-1} z_n z_n^H H^{-H} H^{-1} - H^{-1} z_n \alpha_m^H H^{-1} \right]
\] (A.37)

Using \( \sum_{m=1}^{M} \gamma_{n,m} = 1 \) in the limitation \( \epsilon \to 0 \), (A.37) can be simplified as:

\[
\left( \sum_{n=1}^{N} z_n z_n^H \right) H^{-H} = \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma^{(i)}_{n,m} z_n \alpha_m^H
\] (A.38)

or

\[
\hat{H} = \left( \sum_{n=1}^{N} z_n z_n^H \right)^{-1} \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma^{(i)}_{n,m} \alpha_m^H z_n \right)
\] (A.39)

which is identical to (3.13) □.

**A.9 Solution of the system of equations (3.12) - (3.13)**

In this appendix, it is shown that as a results of (3.12) and (3.13), the matrices \( \hat{H}^{(i+1)} \) and \( \hat{A}^{(i+1)} \) can be obtained by (3.14) and (3.15).

Equation (3.13) implies that

\[
\hat{H}^{(i+1)} = \left( \sum_{n=1}^{N} z_n^{(i+1)} z_n^{(i+1)H} \right) \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma^{(i)}_{n,m} \alpha_m z_n^{(i+1)H} \right)^{-1},
\] (A.40)
or by substituting $\hat{z}_n^{(i+1)} = x_n - \hat{A}^{(i+1)}x_n^P$ and by simple algebra

\[
\begin{align*}
\hat{H}^{(i+1)} & \left( \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}(i) \alpha_m (x_n - \hat{A}^{(i+1)}x_n^P) \right) = \\
&= \sum_{n=1}^N \hat{z}_n^{(i+1)} \hat{z}_n^{(i+1)H} \\
&= \sum_{n=1}^N \left( x_n - \hat{A}^{(i+1)}x_n^P \right) \left( x_n - \hat{A}^{(i+1)}x_n^P \right)^H \\
&= \sum_{n=1}^N \left( C_{xx} - \hat{A}^{(i+1)}C_{xp} - C_{xx}^p \hat{A}^{(i+1)H} + \hat{A}^{(i+1)}C_{x^p x^p} \hat{A}^{(i+1)H} \right)
\end{align*}
\]

(A.41)

where $C_{xx} \triangleq \sum_{n=1}^N x_n x_n^H$, $C_{xp}^H = C_{xp} \triangleq \sum_{n=1}^N x_n x_n^H$, $C_{x^p x^p} \triangleq \sum_{n=1}^N x_n x_n^H$. By (3.12), the state transition matrix estimation is

\[
\hat{A}^{(i+1)} = \left[ \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}(i) \left( x_n - \hat{H}^{(i+1)} \alpha_m \right) x_n^P \right] C_{x^p x^p}^{-1}.
\]

(A.42)

Thus,

\[
\hat{A}^{(i+1)} C_{x^p x^p} = \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}(i) \left( x_n - \hat{H}^{(i+1)} \alpha_m \right) x_n^P.
\]

(A.43)

By substituting (A.43) in (A.41), one obtains

\[
\begin{align*}
\hat{H}^{(i+1)} & \left( \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}(i) \alpha_m (x_n - \hat{A}^{(i+1)}x_n^P) \right) = \\
&= \sum_{n=1}^N C_{xx} - \hat{A}^{(i+1)}C_{xp} - C_{xx}^p \hat{A}^{(i+1)H} + \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}(i) \left( x_n - \hat{H}^{(i+1)} \alpha_m \right) (\hat{A}^{(i+1)}x_n^P)^H \\
&= \sum_{n=1}^N C_{xx} - \hat{A}^{(i+1)}C_{x^p x^p}.
\end{align*}
\]

(A.44)

and by adding $\hat{H}^{(i+1)} \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}(i) \alpha_m (\hat{A}^{(i+1)}x_n^P)^H$ to both sides

\[
\hat{H}^{(i+1)} \left( \sum_{n=1}^N \sum_{m=1}^M \gamma_{n,m}(i) \alpha_m x_n^H \right) = \sum_{n=1}^N C_{xx} - \hat{A}^{(i+1)}C_{x^p x^p}.
\]

(A.45)
By substituting (A.42) in (A.46)

\[
\hat{H}^{(i+1)} \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} \alpha_m x_n^H \right) = \\
N \sum_{n=1}^{N} C_{xx} - \left[ \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m}^{(i)} (x_n - \hat{H}^{(i+1)} \alpha_m) x_n^P \right] C_{x,x}^{-1} C_{x,x} \quad \text{(A.46)}
\]

or

\[
\hat{H}^{(i+1)} = \left( C_{xx} - C_{xx} C_{x,x}^{-1} C_{x,x} \right) \cdot \left[ \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \alpha_m x_n^H - \left( \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{n,m} \alpha_m x_n^P \right) C_{x,x}^{-1} C_{x,x} \right]^{-1} \quad \text{(A.47)}
\]

The state transition matrix estimation is obtained by substituting (A.47) in (3.15).
Appendix B

Stability of the MIMO-AR model

B.1 Stability of first-order MIMO-AR model

Consider a first-order MIMO-AR model with the AR coefficient matrix $A_1$, then:

$$x_n = A_1 x_{n-1} + v_n = A_1^{j+1} x_{n-j-1} + \sum_{i=0}^{j} A_1^i v_{n-i} \tag{B.1}$$

where $v_n$ is a zero-mean white noise. If all the eigenvalues of $A_1$ have modulus less than 1, i.e., they are inside the unit circle, the MIMO-AR process is stable [64]. Hence, the sequence $A_1^i$, $i = 0, 1, \ldots$, is absolutely summable, and the infinite sum $\sum_{i=1}^{\infty} A_1^i v_{n-i}$ exists in mean square. The eigenvalues, roots of $|A_1 - \lambda I_L| = 0$, are the reciprocal of the roots of $|I_L - A_1 z| = 0$. In other words, when $P = 1$, the roots of $|I_L - A_1 z| = 0$ outside the unit circle, i.e., $|z| > 1$, implies that the eigenvalues of $A_1$ be inside the unit circle. Thus, the aforementioned stability condition is equivalent to

$$|I_L - A_1 z| \neq 0 \quad \forall |z| \leq 1 \tag{B.2}$$

where $z$ is a complex number. For the stability discussion it is often convenient to assume that the process has been started at $n = -\infty$. 

75
B.2 Stability of a $P$-order MIMO-AR model

The previous discussion can be extended easily to $P$th-order model, because any $P$th-order MIMO-AR process can be written in a first-order MIMO-AR form. In other words, if $x_n$ is a $P$ order MIMO-AR process defined as in (2.1), then an equivalent $LP$-dimensional 1st-order MIMO-AR model,

$$X_n = AX_{n-1} + V_n,$$  \hspace{1cm} (B.3)

can be defined, where

$$X_n = [x_{n}^{T}, x_{n-1}^{T}, \ldots, x_{n-P+1}^{T}]_{LP \times 1}^{T},$$

$$V_n = [(Hs_{n})^{T}, 0^{T}, \ldots, 0^{T}]_{LP \times 1}^{T},$$

$$A = \begin{bmatrix}
A_1 & A_2 & \cdots & A_{P-1} & A_P \\
I_{L \times L} & 0 & 0 & \cdots & 0 \\
0 & I_{L \times L} & \cdots & 0 & \vdots \\
\vdots & \vdots & \ddots & \ddots & 0 \\
0 & \cdots & \cdots & I_{L \times L} & 0
\end{bmatrix}_{LP \times LP}.$$  \hspace{1cm}

Following the foregoing discussion $X_n$ is stable if the eigenvalues of the companion matrix $A$ are inside the unit circle or, equivalently, if

$$|I_{LP} - Az| \neq 0 \quad \forall |z| \leq 1.$$  \hspace{1cm} (B.4)

It is easy to see that

$$|I_{LP} - Az| = |I_{L} - A_1 z - A_2 z^2 - \cdots - A_P z^P|.$$  \hspace{1cm} (B.5)

Defining the matrix polynomial $A(z) \triangleq I_{L} - A_1 z - A_2 z^2 - \cdots - A_P z^P$, the equation $|A(z)| = 0$ is called the reverse characteristic equation of the MIMO-AR model [64].
Hence, the process (2.1) is **stable** if its reverse characteristic polynomial has no roots in and on the complex unit circle. Formally, the process is stable if

\[ |I_L - A_1 z - A_2 z^2 - \ldots - A_P z^P| \neq 0 \quad \forall |z| \leq 1. \]  

(B.6)
Appendix C

Matrix derivatives

This appendix defines and presents some matrix derivatives used in the derivations in the thesis. It is assumed that the matrix $X$ has no special structure, i.e., that the elements of $X$ are independent (e.g. not symmetric, Toeplitz, positive-definite).

C.1 Definitions

C.1.1 Real derivatives

Let $x$ and $y$ be vectors of orders $n$ and $m$, respectively. Various derivatives can be defined in the following way: If $y = f(x_1, \ldots, x_n)$ is an $(m \times 1)$ vector that depends on the $(n \times 1)$ vector $x$, then the derivative of the vector $y$ w.r.t. the vector $x$ is the following matrix:

$$
\frac{\partial y}{\partial x} = \begin{bmatrix}
\frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\
\frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial y_m}{\partial x_1} & \cdots & \cdots & \frac{\partial y_m}{\partial x_n}
\end{bmatrix}
$$

(C.1)
of order \((m \times n)\) where \(y_1, \ldots, y_m\) and \(x_1, \ldots, x_n\) are the components of \(y\) and \(x\), respectively.

If \(y = f(X)\) is a scalar function of an \((m \times n)\) matrix, \(X = (X_{ij})\), then

\[
\left[ \frac{\partial y}{\partial X} \right]_{ij} = \frac{\partial y}{\partial X_{ji}}
\]

is an \((m \times n)\) matrix of partial derivatives.

**Chain rule for vector differentiation**

Let \(x\) and \(y\) be \((n \times 1)\) and \((m \times 1)\) vectors, respectively, and suppose \(h(y)\) is \((p \times 1)\) and \(g(x)\) is \((m \times 1)\). Then, with \(y = g(x)\),

\[
\frac{\partial h(y)}{\partial x} = \frac{\partial h(g(x))}{\partial x} = \frac{\partial h(y)}{\partial y} \frac{\partial g(x)}{\partial x} \quad (p \times n)
\]

**C.1.2 Complex derivatives**

We define the generalized complex derivatives w.r.t. a complex scalar \(x = x_R + jx_I\) in terms of partial derivatives w.r.t. the real part of \(x_R\) and \(x_I\)

\[
\frac{dg(x)}{dx} = \frac{1}{2} \left( \frac{\partial g(x)}{\partial x_R} - j \frac{\partial g(x)}{\partial x_I} \right)
\]

where \(g(x)\) is a real function of \(x\) and \(j\) is the square root of -1.

**C.2 Useful derivatives**

Let \(x\) be a vector of order \(n\) and \(X\) be a matrix of order \((n \times m)\). According to the above definitions, we obtain the following derivatives:

- **Real values parameters**

\[
\frac{\partial Ax}{\partial x} = A
\]
\[
\frac{\partial x^T Ax}{\partial x} = x^T (A + A^T) \tag{C.6}
\]
\[
\frac{\partial x^T Ax}{\partial x^2} = A + A^T \tag{C.7}
\]
\[
\frac{\partial b^T X a}{\partial X} = ab^T \tag{C.8}
\]
\[
\frac{\partial (Xa + b)^T C (Xa + b)}{\partial X} = a (Xa + b)^T (C + C^T) \tag{C.9}
\]
\[
\frac{\partial a^T X^{-1} b}{\partial X} = -X^{-T} ab^T X^{-T} \tag{C.10}
\]
\[
\frac{\partial tr(X)}{\partial X} = I \tag{C.11}
\]
\[
\frac{\partial tr(AX)}{\partial X} = \frac{\partial tr(XA)}{\partial X} = A^T \tag{C.12}
\]
\[
\frac{\partial tr(X^T AX)}{\partial X} = (A + A^T) X \tag{C.13}
\]
\[
\frac{\partial tr(X^{-1} A)}{\partial X} = - (X^{-1} AX^{-1})^T \tag{C.14}
\]
\[
\frac{\partial tr(AX^{-1} B)}{\partial X} = - (X^{-1} BAX^{-1})^T \tag{C.15}
\]

- Complex values parameters

\[
\frac{\partial x^H b}{\partial x} = 0 \tag{C.16}
\]
\[
\frac{\partial b^H x}{\partial x} = b^H \tag{C.17}
\]
\[
\frac{\partial Ax}{\partial x} = A \tag{C.18}
\]
\[
\frac{\partial x^H Ax}{\partial x} = x^H A \tag{C.19}
\]
\[
\frac{\partial a^H X^H b}{\partial X} = 0 \tag{C.20}
\]

if \( C = C^H \):
\[
\frac{\partial (Xa + b)^H C (Xa + b)}{\partial X} = aa^H X^H C + ab^H C \tag{C.21}
\]
\[
\frac{\partial tr(X)}{\partial X} = I \tag{C.22}
\]
\[
\frac{\partial tr(AX)}{\partial X} = \frac{\partial tr(XA)}{\partial X} = A^T \tag{C.23}
\]
\[
\frac{\partial tr(AX^H)}{\partial X} = \frac{\partial tr(X^H A)}{\partial X} = 0 \tag{C.24}
\]
\[
\frac{\partial \text{tr}(X^HAX)}{\partial X} = A^TX^* \quad \text{(C.25)}
\]
\[
\frac{\partial \text{tr}(X^{-1}A)}{\partial X} = - (X^{-1}AX^{-1})^T \quad \text{(C.26)}
\]
\[
\frac{\partial \text{tr}(AX^{-1}B)}{\partial X} = - (X^{-1}BAX^{-1})^T \quad \text{(C.27)}
\]

where \((\cdot)^*\) is the conjugate of its argument.

Derivatives of a determinant: In the following the determinant operator \(| \cdot |\) apply on square matrices. Some of the expressions below involve inverses: these forms apply only if the quantity being inverted is square and non-singular; alternative forms involving the adjoint, \(ADJ(\cdot)\), do not have the non-singular requirement.

- **real values**
  \[
  \frac{\partial |AXB|}{\partial X} = |AXB|X^{-T} \quad \text{(C.28)}
  \]
  \[
  \frac{\partial \log |X|}{\partial X} = X^{-T} \quad \text{(C.29)}
  \]
  \[
  \frac{\partial |X^TAX|}{\partial X} = |X^TAX| \left( X^TA^T \left( X^TAX \right)^{-1} + X^TA \left( X^TAX \right)^{-1} \right) \quad \text{(C.30)}
  \]
  \[
  \frac{\partial |X^{-1}|}{\partial X} = - |X^{-1}|X^{-T} \quad \text{(C.31)}
  \]

- **Complex values:**
  \[
  \frac{\partial |AXB|}{\partial X} = |AXB|X^{-1} \quad \text{(C.32)}
  \]
  \[
  \frac{\partial \log |X|}{\partial X} = X^{-1} \quad \text{(C.33)}
  \]
  \[
  \frac{\partial |X^HAX|}{\partial X} = |X^HAX|X^{-1} \quad \text{(C.34)}
  \]
Appendix D

Greedy learning of GMM

In this appendix, a brief summary of the greedy approach for the EM and GEM algorithms is presented [76].

In the greedy approach, the mixture components are inserted into the mixture one after the other. In a randomized manner a set of candidate new components is generated. For each of these candidates it find the locally optimal new component by the EM or GEM algorithm. The best local optimum is then inserted into the existing mixture. The greedy algorithm has running time linear in the number of data points and quadratic in the (final) number of mixture components. The algorithm is useful when the optimal number of mixture components is unknown.

Let \( L(x, f_m) = \sum_{n=1}^{N} \log f_m(x_n) \) denote the log-likelihood of the data set \( x \) under the \( m \)-component mixture \( f_m(\cdot) \). The greedy learning procedure consists of the following steps:

1. Compute the maximal log-likelihood function for one-component mixture, \( f_1(\cdot) \). Set \( m = 1 \).
2. Find the optimal new component $\psi(x; \theta^*)$ and the corresponding mixing weight, $\alpha'$:

$$\{\theta', \alpha'\} = \arg \max_{\theta, \alpha} \sum_{n=1}^{N} \log [(1 - \alpha)f_m(x_n) + \alpha\psi(x_n; \theta)]$$  \hspace{1cm} (D.1)

while keeping $f_m(\cdot)$ fixed.

3. Set $f_{m+1}(x) = (1 - \alpha')f_m(x) + \alpha'\psi(x; \theta')$ and $m = m + 1$.

4. Update $f_m(\cdot)$ using EM or GEM algorithm until convergence.

5. If a stopping criterion is met then quit, else, go to step 2.

Step 5 may implement any kind of model selection criterion, for example BIC [66], MDL [67] or AIC [68]. In addition, step 4 may also be implemented by other algorithms than EM or GEM.
Bibliography


