UNIVERSITY OF NICE - SOPHIA ANTIPOLIS

DOCTORAL SCHOOL STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

P H D T H E S I S

to obtain the title of

PhD of Science

of the University of Nice - Sophia Antipolis

Specialty : Automatics and Signal Processing

Defended by

Siouar BENSAID

Some Contributions to Joint Optimal Filtering and Parameter Estimation with Application to Monaural Speech Separation

Thesis Supervisor: Dirk SLOCK

prepared at EURECOM Sophia Antipolis, Mobile Communications Department

defended on Juin 6, 2014

Jury :

Reviewers :	Pr. Karim Abed-Meraim	-	University of Orleans
	Pr. Søren Holdt Jensen	-	Aalborg University
Examiners :	Pr. Luc Deneire	-	University of Nice
	Pr. Régine Le Bouquin-Jeannes	-	University of Rennes 1
	Dr. Nicholas Evans	-	EURECOM
Advisor:	Pr. Dirk Slock	-	EURECOM

UNIVERSITE DE NICE-SOPHIA ANTIPOLIS

ECOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

T H E S E

pour l'obtention du grade de

Docteur en Sciences

de l'Université de Nice-Sophia Antipolis

Mention : Automatique et Traitement du Signal

présentée et soutenue par

Siouar BENSAID

Quelques Contributions au Filtrage Optimal avec l'Estimation de Paramètres et Application à la Séparation de la Parole Mono-Capteur

Thèse dirigée par: Dirk SLOCK

préparée à EURECOM Sophia Antipolis, Département Communications Mobiles

soutenue le 6 juin 2014

Jury :

Rapporteurs :	Pr. Karim Abed-Meraim	-	Université d'Orleans
	Pr. Søren Holdt Jensen	-	Université Aalborg
Examinateurs:	Pr. Luc Deneire	-	Université de Nice
	Pr. Régine Le Bouquin-Jeannes	-	Université de Rennes 1
	Dr. Nicholas Evans	-	EURECOM
Encadrant:	Pr. Dirk Slock	-	EURECOM

Abstract

This thesis is composed of two main parts. In the first part, we deal with the monomicrophone speech separation problem. In general, The problem of blind source separation (BSS) arises in a variety of fields in speech processing such as speech enhancement, speakers diarization and identification. Several methods for BSS consider several observations of the same recording. Single microphone analysis is the worst underdetermined case, but, it is also the more realistic one. In this thesis, we propose two main monomicrophone speech separation algorithms. The first algorithm is developed in the time domain. The autoregressive (AR) structure and the periodic signature of voiced speech signal are modeled. Two state space models, linear and bilinear, with unknown parameters are derived. The expectation maximization (EM) terminology is used for each state space model to derive a variety of algorithm that estimates jointly the unknown parameters and the sources. A second variety of algorithm for the bilinear state space model is derived where an extra fixed-interval smoothing step is introduced. Simulations show the performance of the different algorithm to achieve separation and establish a comparison between the three varieties.

The second monomicrophone algorithm is derived in the frequency domain. Unlike the methods proposed in the time domain, in the frequency domain algorithm we estimate the parameters and the sources separately. We keep the same speech model used with temporal methods (quasiperiodic sources with AR spectral envelope). We focus on the use of windows in the frequency domain processing of data for the purpose of spectral parameters estimation. Classical frequency domain asymptotics replace linear convolution by circulant convolution leading to approximation errors. We show how the introduction of windows can lead to slightly more complex frequency domain techniques, replacing diagonal matrices by banded matrices, but with controlled approximation error. We focus on the estimation of zero mean Gaussian data with a parametric spectrum model and show the equivalence of three approximation/estimation criteria: Itakura-Saito distance (ISD), Gaussian maximum likelihood (GML) and optimally weighted covariance matrix (OWCM).

In the second part of the thesis, we consider the problem of linear MMSE (LMMSE) estimation (such as Wiener and Kalman filtering) in the presence of a number of unknown parameters in the second-order statistics that need to be estimated also. This well-known joint filtering and parameters estimation problem has numerous applications. It is a hybrid estimation problem in which the signal estimated by linear filtering is random, and the unknown parameters are deterministic. As the signal is random, it can also be eliminated (marginalized), allowing parameters estimation from the marginal distribution of the data. An intriguing question is then the relative performance of joint vs. marginalized parameters estimation. In this part, we consider jointly Gaussian signal and data and we first provide contributions to Cramér-Rao bound (CRB)s. We characterize the difference between the hybrid information matrix (HIM) and the classical marginalized Fisher information matrix (FIM) on the one hand, and between the FIM (with CRB

asymptotically attained by the maximum likelihood (ML)) and the popular modified FIM, modified FIM (MFIM), the inverse of modified CRB which is a loose bound. We then investigate three iterative (alternating optimization) joint estimation approaches: Alternating maximum a posteriori (MAP) for signal and ML for parameters alternating MAP/ML (AMAPML), which in spite of a better HIM suffers from inconsistent parameters bias, EM which converges to (marginalized) ML (but with AMAPML signal estimate), and variational Bayes (VB) which yields an improved signal estimate with the parameters estimate asymptotically becoming ML.

Résumé

Nous traitons le sujet de l'estimation conjointe des signaux aléatoires dépendant de paramètres déterministes et inconnus. En première partie, nous abordons le sujet du côté applicatif en proposant deux algorithmes itératifs de séparation de la parole voisée mono-capteur. Dans le premier algorithme, nous utilisons le modèle autorégressif de la parole qui décrit les corrélations court et long termes (quasi-périodique) pour formuler deux modèles d'Etat dépendant de paramètres inconnus linéairement et bilinéairement. EM-Kalman est ainsi utilisé pour estimer conjointement les sources et les paramètres. Nous dérivons trois variétés de cet algorithme. La première variété applique EM-Kalman sur le modèle linéaire en se limitant à un lissage à délai fixe (un délai de 1 échantillon), cette limitation est due à la singularité de la matrice de covariance d'erreur de prédiction. Les deux autres variétés consistent à appliquer EM-Kalman au modèle bilinéaire en utilisant le lissage RTS dans la première et le lissage à délai fixe (délai de 1 échantillon aussi) dans la deuxième. Les simulations avec signaux synthétiques confirment la supériorité de la deuxième variété (modèle bilinéaire avec lissage RTS) sur le reste grâce à l'étape lissage RTS. Les deux autres méthodes montrent des performances très proches. Avec les signaux réels, la méthode à base du modèle linéaire a montré des performances proches du modèle bilinéaire avec lissage RTS et mieux que celui avec lissage à délai fixe surtout en termes de SIR.

Dans le deuxième, nous proposons une méthode fréquentielle pour le même modèle de la parole utilisé avant mais cette fois nous estimons les sources et les paramètres séparément. Les observations sont découpées à l'aide d'un fenêtrage bien conçu pour assurer une reconstruction parfaite des sources après. Les paramètres (de l'enveloppe spectrale) sont estimés en maximisant le critère GML appliqué au spectre des données. Dans les méthodes asymptotiques qui utilisent des fenêtres infiniment longues, la matrice de covariance du spectre est modélisée par une matrice diagonale (spectre paramétrique en diagonal). Dans le cas d'une fenêtre à longueur finie, nous modélisons plus correctement la matrice de covariance du spectre en tenant compte de l'effet de ce fenêtrage qui introduit des corrélations entre les fréquences au voisinage du lobe principal de la fenêtre. Dans ce cas, la matrice de covariance devient une matrice bande. Après estimation des paramètres, le filtre de Wiener est utilisé pour estimer les sources par fenêtre. Finalement, les sources complètes sont reconstruites en utilisant la technique d'overlap-add.

En deuxième partie, nous abordons l'estimation conjointe d'un point de vue plus théorique en s'interrogeant sur les performances relatives de l'estimation conjointe par rapport à l'estimation séparée d'une manière générale. Nous considérons le cas conjointement Gaussien (observations et variables cachées) et trois méthodes itératives d'estimation conjointe: MAP en alternance avec ML, biaisé même asymptotiquement pour les paramètres, EM qui converge asymptotiquement vers ML et VB que nous prouvons converger asymptotiquement vers la solution ML pour les paramètres déterministes.

Acknowledgements

First of all, I want to thank Professors Karim Abed-Meraim and Jensen Holdt Søren for agreeing to be reviewers of my thesis. I also want to thank Professors Luc Deneire and Régine Le Bouquin-Jeannes, and Doctor Nicholas Evans for agreeing to be examiners in my Jury. I thank my supervisor Dirk Slock for his assistance.

I present special thanks to my parents for their endless care, prayers and big efforts to support me in my long road; to my sister Kobla and my brother Yassine for their encouragements and love.

I want to thank my friend Achraf Essafi for the big help and support he offered to me and without which I could not finish until the end. I learned much from you, I learned how to face my fears and overcome them, and how to get the positive attitude toward problems. I learned a new theory in life, the theory of "becoolism" or how to be happy whatever were the circumstances around. Thank you for allowing me to play "Bubble Mania" for hours in your ipad, and for the tasty Kebabs we shared in Juan Les Pins on the beach at the sunset.

I want to thank my friend Kaoutar for her support in my difficult times especially during the last period of my PhD. I thank her for the high-level discussions we exchanged about political issues, arts and cinema, and of course I can't forget the delicious Moroccan dishes she cooked or brought from Morocco. Thank you Kaoutar for reading and correcting my thesis chapters.

I want to thank my office-mate Arun for the positive energy he was spreading in the office, for his helpful discussions from time to time in my work, for the generous invitations he offered to make me taste homemade Indian tasty dishes, for the interesting philosophic discussions about spiritual life.

I want to thank my friends Fatma, Houda, Hajer, Wael, Bilel, Aymen, Bassem, Riadh, Hocine and Maha for the cheerful moments we shared inside and outside EURECOM. I really enjoyed our dinners together. Thank you Wael for the discussions and quarrels we got, I learned much from it. I want to thank Antony for his interesting cooperation at the beginning of my PhD especially that it was (unfortunately) the only cooperation I could enjoy during my whole PhD. I also thank my office-mates Miltos, Xinping and Rajeev for making our office very favorable for hard work (I know I was the noisiest one in the office).

I can't forget to thank people in IT service and administration in EURECOM especially Alain Gagnon and Gwenaêlle Lestir for their kindness. I also want to thank Régine Saelens from EDSTIC for her kindness and support.

Abbreviations

ACRB asymptotic Cramér-Rao bound

AMAPML alternating MAP/ML

AR autoregressive

ASA auditory scene analysis

BB Barankin bound

BCRB Bayesian Cramér-Rao bound

BIM Bayesian information matrix

BSE blind source extraction

BSS blind source separation

CASA computational auditory scene analysis

CCRB conditional Cramér-Rao bound

CE conjugate-exponential

CELP code-excited linear prediction

CRB Cramér-Rao bound

DFT discrete Fourier transform

DTFT discrete-time Fourier transform

EKF extended Kalman filter

EM expectation maximization

EMCB extended Miller-Chang bound

FIM Fisher information matrix

FIR finite impulse response

GML Gaussian maximum likelihood

GMM Gaussian mixture model

HCRB hybrid Cramér-Rao bound

HIM hybrid information matrix

HMM hidden Markov model

ICA independent component analysis

IIR infinite impulse response

ISD Itakura-Saito distance

JCRB joint Cramér-Rao bound

KLD Kullback-Leibler divergence

LMMSE linear MMSE

LPC linear prediction coding

LSE least square error

LTP long term prediction

LU lower upper

MAP maximum a posteriori

MCB Miller-Chang bound

MCRB modified Cramér-Rao bound

MFIM modified FIM

ML maximum likelihood

MMSE minimum mean square error

MSE mean square error

NCCF normalized cross-correlation function

NLS nonlinear least square

NMF non-negative matrix factorization

OWCM optimally weighted covariance matrix

OWSM optimally weighted spectrum matrix

pdf probability density function

PESQ perceptual evaluation of speech quality

PR perfect reconstruction

PSD power spectral density

QEKF quadratic extended Kalman filter

RAPT robust algorithm for pitch tracking

RLS recursive least squares

RPEM recursive prediction error method

RTS Rauch-Tung-Striebel

SAGE space-alternating generalized EM

SAR sources-to-artifacts ratio

SDR source-to-distortion ratio

SIR source-to-interferences ratio

SNR source to noise ratio

SOEKF second order extended Kalman filter

SOELMMSE second order extended LMMSE

SSR signal to signal ratio

STFT short time Fourier transform

STP short term prediction

UDL upper diagonal lower

VB variational Bayes

VBEM variational Bayes for EM

VQ vector quantization

Contents

Chapter 1

Introduction

1.1 Mono-microphone source separation

Noise and interference suppression is a fundamental problem that signal processing community strive to resolve. Particularly in audio processing, having "clean" sounds is very crucial into many applications such as speech recognition, speech decoding, music transcription, etc. Unfortunately, in real life, noise/interference-free environment does not exist. In general, recorded sequences consist of a mixture of several sounds, some specified as target signals and the rest as noise depending on the application. for example in music transcription, noise will be defined as all the instruments played except the target one. In speech enhancement, the aim is to extract or "enhance" one target speech while considering all the rest as noise, whereas in speech separation, we aim to extract more than one source.

In the first part of this thesis, we tackle the speech separation issue. Many factors affect the recorded mixture model, namely, microphones directivities and spacing, static or moving sources, open or closed space, etc. For example, in a closed room, we call the echoes and reverberation phenomena [Melia & Rickard 2007, Mukai *et al.* 2001], the sound wave is reflected by walls and possible objects present in the room which results in what we call echoes and reverberations. The short delays refer to the echoes and the long delays to reverberation. The reverberation time is exactly measured as the delay when the echoes become 60 dB less than the original sound. Hence, the recorded sound results in a superposition of delayed and attenuated versions of the original sound wave. In this case, the mixture is described as convolutive. In specific contexts, these phenomena can be neglected and the mixture is simplified to instantaneous mixture where the sources contribute instantaneously to the mixture (each source is only weighted by a multiplicative constant).

When the sources and the mixture model are both unknown, the problem is ill-posed and need more assumptions to become resolved. The most famous assumptions used in speech separation is mutual independence like in independent component analysis (ICA) [Lee *et al.* 1999]. In addition to these assumptions and when the number of microphones is greater than (or equal to) the number of sources (overdetermined mixture), the problem becomes well-determined. We need only to

estimate the inverse of the mixture function and the sources are deduced immediately. When the number of microphones is less than the number of sources (underdetermined case), the problem becomes more difficult and more assumptions are needed such as sparsity in time and frequency, periodicity or assuming a certain *a priori* information to the sources. The most difficult case is definitely the monomicrophone speech separation, that is why generally the studied model is very simplified (direct sum of sources) but still realistic in some contexts. In this thesis, we will focus on the study of this model.

1.2 Joint estimation performance

Providing accurate estimations of unknown wanted signals/parameters is an important omnipresent task in signal processing where the ultimate goal is to construct optimal (in the mean square error (MSE) sense) and feasible estimators. A great part of estimation problems can be described generally by the scheme of using observations estimating latent random variables and deterministic parameters. There are two main strategies in literature, whether we estimate jointly the random variables and the parameters by optimizing a joint criterium, or we marginalize the random variables to get a new criterium that depends only on the deterministic parameters, then the estimate of parameters resulting from the new criterium is plugged in the joint criterium in order to estimate the random variables. An other issue is when we transform the deterministic parameters to random ones by supposing a certain *a priori* knowledge for them. The idea of randomizing the deterministic parameters intends to add a new element in the estimation process, besides to the estimator mean, which consists in the estimation error. The latter is also used in the random variable estimation and we expect it to improve their estimation. An example of terminology that applies this idea is the variational Bayes (VB) [Beal 2003].

New varieties of the well-known Kalman filter were proposed to solve the joint estimation problem, namely, extended Kalman filter (EKF), second order extended Kalman filter (SOEKF), expectation maximization (EM)-Kalman, VB-Kalman,etc. The plurality of the proposed solutions raises an important question about the superiority of one method relative to the other. In this thesis, we focus on comparing the joint vs. marginalized estimation techniques in terms of Cramér-Rao bound (CRB)s, we also compare the EM (deterministic parameters) and VB (random parameters) in asymptotic regime.

1.3 Organization and contributions

In Chapter 2, we introduce the problem of monomicrophone speech separation as a special case (and the most difficult) of a general problem which is source separation. We then review some applications of this issue in speech recognition ad hearing-aid to justify our motivation to treat it. Since separation algorithms are applied to the speech signal, an important point was to introduce the different proposals for speech modeling starting from the deterministic model to the bayesian one based on the speech features namely harmonicity and time correlations. Besides, these features will play an important role on algorithms design as cues for separation. Afterward, a survey of the different algorithm is done. The classification of the different methods is very subjective, yet we followed the most common classification in the literature which is the source-driven vs the model-driven methods. Where in the first category, the separation is based on specific cues inferred in the observation signals while in the second a presumed *a priori* model is set to the sources and a training phase is necessary to learn it. The advantages and drawbacks of each category are discussed too. We finally approach an important issue in this topic which is the performance evaluation and comparison. In fact, in the scientific community, two main evaluation strategies are used: an objective strategy based on preset mathematical criteria and which provides a global quality description, a subjective strategy based on human decision where the results are heard and assessed by audience on a specific data base.

In Chapter 3, we present our first contribution in the monaural speech separation topic. A time domain separation approach is proposed. A joint speech model based on short term prediction (STP) and long term prediction (LTP) is used to describe the short term and long term correlations of the signal. This model, well-known in the speech coding framework, is exploited to formulate two state space models (linear and bilinear) parameterized by the autoregressive (AR) coefficients of sources. Since the involved signals are Gaussian, source estimation is achieved using the EM-Kalman algorithm for each state space model, where in the E-step the unknown parameters are updated using the state (sources) statistics computed in the M-step using Kalman. An important advantage of this algorithm is that we override the problems of sources reconstruction encountered in the frequency domain based methods. A part of this work is published in [Bensaid *et al.* 2010b].

In Chapter 4, we present our second contribution in the monaural speech separation topic. We switch to a frequency domain algorithm. The proposed algorithm is processing per frame. The same joint model introduced in chapter 3 is used here too to formulate the parametric spectrum of sources. Under a finite length window constraint, the equivalence of three cost functions that are Gaussian maximum

likelihood (GML), Itakura-Saito distance (ISD) and optimally weighted covariance matrix (OWCM) is analysed. The window effect on the covariance spectrum matrix is also studied by taking into consideration the cross correlation between the different frequencies. The GML criterium is used to derive the estimates of the unknown parameters in an iterative fashion. The estimation of the source Spectra is achieved by a Wiener filtering and finally the overlap-add technique is used for their reconstruction in time .

In Chapter 5, we move to a new topic inspired by the two previous chapters. We compare the joint estimation and separate estimation of jointly-Gaussian signals in terms of lower bounds. In chapters 3 and 4, we dealt with the scenario of estimating random variables (sources) and deterministic parameters (joint speech model parameters) using observation data, in other words, we are in a hybrid estimation scenario. The benchmark of performance in the separate estimation problem (the deterministic parameter estimation using the observations) is fixed using the CRB. In the hybrid case, the benchmark is set by the hybrid Cramér-Rao bound (HCRB). We propose to study the difference between the two bounds in order to understand the effect of the random parameters on the deterministic ones. The difference between the CRB and the modified Cramér-Rao bound (MCRB) is also studied. This work is published in [Bensaid & Slock 2013].

In Chapter 6, we aim to compare two famous iterative maximum likelihood (ML) algorithms that are EM and variational Bayes for EM (VBEM). VBEM differs from the terminology of EM by assuming *a priori* knowledge to the deterministic parameters. This assumption results the appearance of a new element in the estimation process which is the parameters estimation error. The latter is used with the estimate in the estimation of the random variables. In EM, we are only limited to use the mean value of the parameters estimator for this task. Based on this observation, we expect that the VB performs better than EM that is known to converge to the ML solution. We prove that asymptotically, the VB converges to the ML solution. This work is published in [Bensaid & Slock 2013].

Finally, in Chapter 7, we present the conclusions and the perspectives. We highlight the key issues tackled in this thesis, we also point out the limitations of our work which will be the starting point of new perspectives, such us the high dependence of the separation algorithms on pitches tracking accuracy, and the probable superiority of VB to EMin sub-asymptotic conditions.

List of publications

- Bensaid, Siouar, Schutz, Antony, Slock, Dirk T M , "Monomicrophone blind audio source separation using EM-Kalman filters and short+long term AR

modeling ", Asilomar 2009, 43rd Asilomar Conference on Signals Systems and Computers, November 1-4, 2009, Asilomar, California, USA , pp 343 - 345;

- Bensaid, Siouar, Slock, Dirk T M , "Single microphone blind audio source separation using short+long term AR modeling ", IWAENC 2010, International Workshop on Acoustic Echo and Noise Control, August 30-September 2nd, 2010, Tel Aviv, Israel;

- Bensaid, Siouar, Schutz, Antony; Slock, Dirk T M , "Single Microphone Blind Audio Source Separation Using EM-Kalman Filter and Short+Long Term AR Modeling", LVA ICA 2010, 9th International Conference on Latent Variable Analysis and Signal Separation, September 27-30, Saint-Malo, France / Also published in LNCS, 2010, Vol 6365/2010 , pp 106-113;

- Bensaid, Siouar;Slock, Dirk T M , "Comparison of Varieties of Kalman Filtering Algorithms Applied to Single Microphone Blind Audio Source Separation", Asilomar 2011, 45rd Asilomar Conference on Signals Systems and Computers, November 6-9, 2011, Asilomar, California, USA;

- Bensaid, Siouar, Slock, Dirk T M , "Blind Audio Source Separation Exploiting Periodicity and Spectral Envelopes", IWAENC 2012, International Workshop on Acoustic Signal Enhancement, September 4-6, 2012, Aachen, Germany;

- Bensaid, Siouar, Slock, Dirk T M , "On the performance of joint linear minimum mean squared error (LMMSE) filtering and parameter estimation",SPAWC 2013, 14th IEEE International Workshop on Signal Processing Advances in Wireless Communications, June 16-19, 2013, Darmstadt, Germany.

Chapter 2

Monaural speech separation

2.1 Introduction

Source separation is a general problem that may be encountered in different domains, for instance in telecommunications, audio and biomedical processing [Comon & Jutten 2010]. The proposed solutions are very influenced by the problem context, namely, the number of observations compared to the dimension of the input data, the nature of sources, the nature of the propagation space, etc. In this thesis, we are mainly interested in speech signal. Therefore, it is important to understand the features of this physical signal and see how it is mathematically modeled . Definitely, the speech model plays a key role in the separation algorithm design. In this chapter, we start with introducing the source separation problem in general and the specific case the thesis will be limited to. Then we give an overview of the different applications of Monaural speech separation to justify our motivation to this issue. Afterward, we make a brief analysis of speech signal characteristics. Finally, we develop a review about the existing solutions and conclude.

2.2 Problem statement

Suppose we have K speakers $(s_{1,t}, s_{2,t}, \cdots, s_{K,t})$ talking at the same time and their speech recorded by M-dimensional microphone array denoted by $\boldsymbol{y}_t = [y_{1,t}, y_{2,t} \cdots, y_{M,t}]^T$. Here t denotes time. The speech separation task consists in recovering the speech of every speaker separately. Generally, the measurements are interpreted as a mixture of the present sources, corrupted by additive noise denoted $\boldsymbol{n}_t = [n_{1,t}, n_{2,t} \cdots, n_{M,t}]^T$ where $n_{m,t}$ is the noise corrupting the m^{th} microphone

$$\boldsymbol{y}_t \;=\; \boldsymbol{h}(\boldsymbol{s}_t) \;+\; \boldsymbol{n}_t \tag{2.1}$$

Where $\boldsymbol{s}_t$ is the K-dimensional source vector $[s_{1,t}, s_{2,t} \cdots, s_{K,t}]^T$. When M is greater or equal to the number of sources N, the mixture is said overdetermined and the sources are recovered by estimating the inverse of the mixture function (or pseudoinverse if singular). When the number of observations is less than the number of input sources, the problem becomes underdetermined and it requires more assumptions and information about the sources to get solved. The extreme and most

difficult case of the underdetermined mixture is the monaural case where only one observation is available to recover the K sources, hence no space diversity can be exploited. In addition to the number of observations, the characterization of the mixture function $\boldsymbol{h}(.)$ depends tightly on the propagation space (enclosed/opened space, the dimensions of the propagation room,etc.), reverberation and echoic effects may occur. In this case, the mixture will be the linear convolution between the propagation channel and the sources. If we neglect the previous effects, the problem is simplified to the instantaneous mixture

$$\boldsymbol{y}_t \;=\; \boldsymbol{h}^T \boldsymbol{s}_t \;+\; \boldsymbol{n}_t \tag{2.2}$$

In the monaural case and in order to simplify the problem, the channel effect is neglected completely and the measurements are simply interpreted as a direct sum of the sources corrupted by additive noise

$$y_t \;=\; \sum_{k=1}^{K} s_{k,t} \;+\; n_t \tag{2.3}$$

Here, y_t is a scalar. Though this model may look simplistic, it is not far from some realistic cases and many of the scientific community used it in their work [Mowlaee 2010, Virtanen 2006a, Benaroya 2003]. This model will be also our main focus in this first part of the thesis.

2.3 Applications of monaural speech separation

In real life applications, the speech separation is often a necessary pre-processing step that aims to prepare the signals before starting the main process. For example, speech separation is very important to succeed Automatic Speech Recognition (ASR)[Cooke *et al.* 2010, Li *et al.* 2010, Shao *et al.* 2010]. In [Cooke *et al.* 2010], the authors provide a comparison between different monaural speech separation algorithms used in ASR when a background talker is speaking the same sentences of the target talker. They prove by simulations that model-based approaches [Virtanen 2006b, Weiss & Ellis 2010] perform better than CASA approaches [Shao *et al.* 2010] in this task. In [Li *et al.* 2010], authors use jointly MAX-VQ and CASA approaches to improve speaker identification and speech recognition robustness. An other important application of speech separation is the hearing aid and speech coding [H. Viste & G. Evangelista 2001]. The human auditive system has naturally the ability to resolve signals and focus on the wanted target. Unfortunately, this process becomes difficult for hearing-impaired persons especially in the presence of background noise and cocktail-party like situations. Speech separation can be also used for speech enhancement purpose [Schmidt *et al.* 2007]. Actually,

speech enhancement can be considered as a particular case of speech separation where one of the sources we want to separate is the additive noise. In general, speech enhancement algorithms rely on the noise characteristics which are different from the one of the enhanced source, but the speech separation methods generalize to the case where the noise is also an interferant speech source. Another application of speech separation is in the speaker diarization task [Anguera Miro *et al.* 2012] which consists in determining "'who spoke when?"' by portioning the main audio stream into segments according to the speaker identity. Diarization becomes more complicated when the speakers talk simultaneously. A preprocessing step of speech separation is definitely very helpful to simplify the segmentation task.

2.4 Speech signal modeling

The speech signal is a sound wave that owns specific features. These features will definitely provide more useful information to the separation process. When we plot a time domain speech sequence and its corresponding spectrogram (see figure Fig. 2.1), we first notice strong time correlation between the neighboring samples and the quasi-periodicity of some parts. In fact, the speech signal is generally divided into voiced and unvoiced parts. These features have inspired several mathematical models that try to describe as much accurate as possible the physical wave.

Sinusoidal modeling

A very well known suggested model is the sinusoidal model that is described as

$$s_t \;=\; \sum_{l=1}^{L} a_l \cos(2\pi f_l t + \phi_l) \;+\; e_t \tag{2.4}$$

where a_l, f_l and ϕ_l are respectively the amplitude, frequency and phase of the l^{th} sinusoid, L is the order of the model, e_t is the residual noise. The sinusoidal sum models the harmonic structure of speech signal resulting from the vocal cords vibration. The frequencies $\{f_l\}_{l=1..L}$ are the prominent peaks in the signal spectrum. The residual e_t contains the excitation power and represents all what is not harmonic in the signal. In many cases, the sinusoidal model is used in the complex form where the cos function is replaced by the complex exponential function. The sinusoidal model is considerably used in the problem of speech separation [Quatieri & Danisewicz 1990, Mowlaee *et al.* 2011a, Chazan *et al.* 1993] and multipitch tracking [sbø ll Christensen *et al.* 2008, Christensen *et al.* 2009] where the main objective is to estimate accurately the model parameters (amplitudes, frequencies, phases,etc.). In [sbø ll Christensen *et al.* 2008], the nonlinear least

square (NLS), MUSIC and Capon estimators are applied to estimate the fundamental frequencies and the complex amplitudes. Whereas, in [Chazan *et al.* 1993], an expectation maximization (EM)-based iterative approach is used to find the maximum likelihood (ML) estimate of pitches. In [Mowlaee *et al.* 2011a], the sinusoidal model is used in monaural speech separation context, the parameters of the model are considered as codevectors estimated from vector quantization (VQ) codebooks trained from clean speech of the original speakers. The sinusoidal model is also used in music to model instruments sounds [Virtanen & Klapuri 2000, Virtanen & Klapuri 2001].

Autoregressive modeling

An other model proposed for speech is the autoregressive (AR) model where the speech signal s_t is considered as an AR(p) process [Atal & Hanauer 1971]

$$s_t = \sum_{l=1}^{p} a_l s_{t-l} + e_t \tag{2.5}$$

where p is the model order and $\{a_l\}_{l=1..p}$ are the AR coefficients that describe the spectral envelope shape, e_t is the random innovation process. When e_t is described as white noise, the model order p must be large enough to cover at least one pitch period. Yet, this model is not really optimized since in practice, only the first ten coefficients are significant, plus the one corresponding to one pitch delay [Chu 2003]. The innovation process is alternatively modeled by an AR model where only the coefficient corresponding to the pitch delay is nonzero (in the voiced case). this is denoted by the joint model. The AR model is successfully used in speech coding, mainly in linear prediction coding (LPC) where the signal is coded by its AR coefficients and excitation [Atal & Remde 1982, Erkelens 1996]. A very famous LPC algorithm is the code-excited linear prediction (CELP) coder introduced in [Schroeder & Atal 1985]. The AR model was also used in mono-microphone and multi-microphone speech enhancement [Weinstein *et al.* 1994, Gannot *et al.* 1997] where the coefficients and innovation power are estimated in an EM-Kalman framework. In speech separation, this model was used by the blind source separation (BSS) algorithms that exploit the temporal structure of sources to achieve separation such as in [Smith *et al.* 2005, Cichocki *et al.* 2000] where the temporal prediction error is minimized to estimate the AR coefficients and the mixing matrix in the overdetermined case. This model will be treated more in details in chapter 3.

Gaussian mixture modeling

In a probabilistic framework, a speech signal can be modeled with a Gaussian mixture model (GMM). This model considers that a source signal s_t is generated by a

set of Gaussian subsources $\{s_t^i\}_{i=1..I}$, where I is the number of components present in the set. In speech, we generally use it to express the probability density function (pdf) of the short time Fourier transform (STFT) where each subsource helps to model a specific behavior of the speech spectrum

$$p(S(t,f)) \;=\; \sum_{i=1}^{I} w_i \, \mathcal{N}(S(t,f)|\mu_i, \Sigma_i) \tag{2.6}$$

where $S(t,f)$ is the STFT of s_t, the multiplying coefficient w_i represents the probability that the i^{th} subsource is active, the term $\mathcal{N}(S(t,f)|\mu_i, \Sigma_i)$ is the distribution of $S(t,f)$ when the subsource s_t^i is active which is Normal of mean μ_i and covariance Σ_i. The GMM presentation is praised for its generality and ability to model a large number of probable spectrum shapes, and therefore captures the non-stationarity of the signal which made it used in automatic speech recognition [Shao *et al.* 2010]. It is also used in speech separation methods that suppose a prior knowledge about the sources [Radfar & Dansereau 2007, Ozerov *et al.* 2007, Benaroya *et al.* 2006, Benaroya & Bimbot 2003] and where a training step is needed to learn the pdfs parameters (means and covariances) from clean speech.

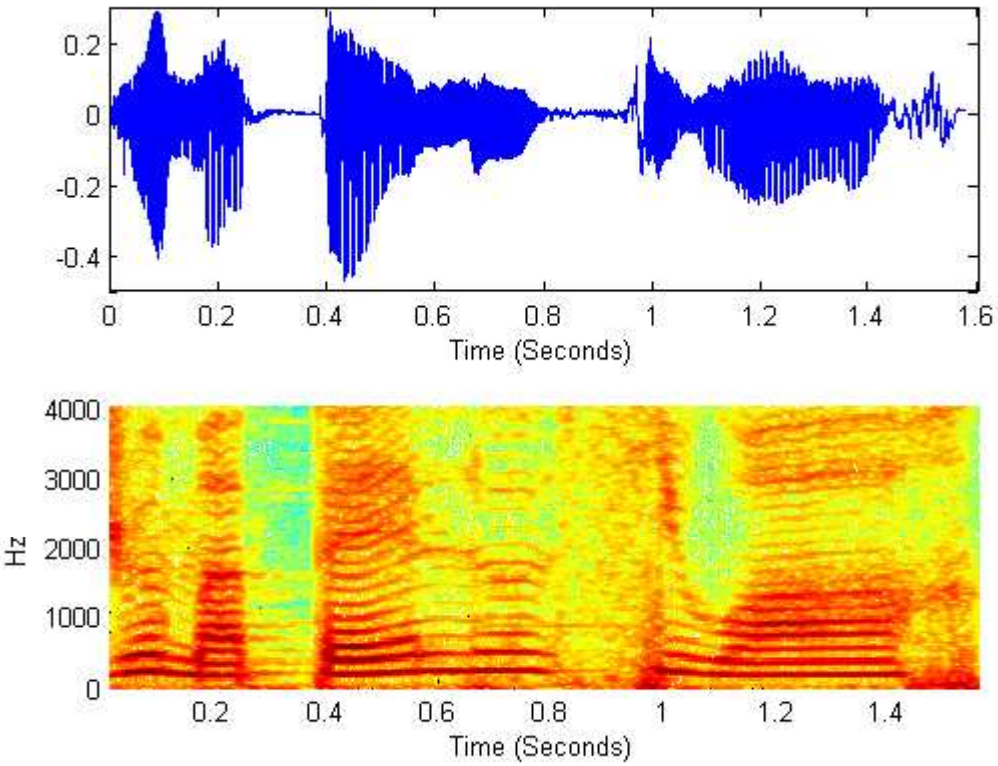

Figure 2.1: *Time domain female speech (top figure) and its corresponding spectrogram (bottom figure)*

2.5 Review of the existing solutions

Since mono-microphone speech separation has started interesting the scientific community, many techniques were developed adopting different point of view of the problem. Sometimes techniques overlap therefore any suggested classification of them in the literature will be very subjective and depends on the authors' points of view. Yet, the most agreed opinion is to consider two main broad categories: source-driven methods and model-driven methods. In the following, we will analyze each group in more details.

2.5.1 Source-driven methods

In source-driven methods, no *a priori* knowledge about speakers is used, the extraction of speech is done directly on the mixture by inferring some specific cues [Bregman 1990], namely, harmonic concordance, synchronous changes of onset and offset, amplitude modulation,etc. The most well-known source-driven method is the CASA approach [Hu & Wang 2004, Shao *et al.* 2010, Li *et al.* 2010, Quatieri & Danisewicz 1990, Wang 2005].

CASA

computational auditory scene analysis (CASA) method is inspired from the natural human processing to acoustic signals or what we term exactly auditory scene analysis (ASA)[Bregman 1990]. CASA processing is done in two steps: segmentation and grouping. In segmentation step, the signal is transformed into time-frequency domain and segmented to units or segments where each one corresponds to one single speaker. The segmentation is based on periodicity for voiced speech, multi-scale onset and offset analysis for unvoiced speech [Shao *et al.* 2010]. In the grouping step, segments showing similarities are grouped in the same stream. There are two types of grouping: simultaneous grouping and sequential grouping. In the simultaneous grouping, for each fixed time t whence the name simultaneous, segments are grouped across frequency. It is especially voiced segments which are grouped here since they are characterized with periodicity. After this step, we end up with simultaneous streams where each one corresponds to one speaker. Then these streams are organized across time in the sequential grouping to produce the final streams of each speaker in the mixture. The output of the CASA process consists of two time-frequency binary masks (if there are two speakers in the mixtures) where 1 means the presence of the target and 0 otherwise. In [Wang 2005], the author proves that the ideal binary mask of a given target is a major computational goal of CASA. The ideal binary mask is a binary matrix where the $1-$valued coefficient means

the target power is stronger than interference and 0 otherwise and the CASA-based methods attempt to estimate this binary mask.

In [Quatieri & Danisewicz 1990], authors propose two methods based on sinusoidal modeling of voiced speech. The speech signal is modeled as the sum of sine waves with time-varying amplitudes and phases which are the parameters to be estimated. The first approach is the extension of the analysis/synthesis system to the case of two speakers. The parameters of higher intensity speaker are estimated first using high-resolution spectral analysis and peak-picking where then only largest peaks are kept. After synthesizing the speech using sine-wave matching and interpolation, the lower intensity speaker is deduced by subtracting the first estimated from the mixture. In the second approach, the sine waves' parameters are estimated by minimizing the least-squares error between the mixture and the parametric model. The performance of these approaches are very sensitive to the quality of pitches estimation. When there is *a priori* information about pitches, good separation is achieved down to $-16dB$ speech-to-speech ratio, otherwise good results are only obtained when the two sources have close powers.

In [Virtanen & Klapuri 2000], the authors use the sinusoidal modeling and the analysis/synthesis process to achieve separation. Frequency and amplitude continuity over frames and harmonic concordance are measured using heuristic distances. Unlike the classical CASA systems, authors in [Hu & Wang 2004] suggest different mechanisms to segregate resolved and unresolved harmonics of voiced speech. Unresolved harmonics can be segregated based on the auditory filter bank responses used in front of the CASA process since they present special behaviour in their presence (strongly amplitude-modulated responses and important fluctuations of the responses' envelopes around the fundamental frequency of target speaker).

2.5.2 Wiener filtering

Wiener filtering is a classical method used in speech enhancement [Meyer & Simmer 1997, Ephraim 1992, Chen *et al.* 2006] where it provides the minimum mean square error (MMSE) estimate of a target stationary source. It is always applied on the STFT of speech signals in order to exploit their short time stationarity. Wiener filtering is also used in monaural speech separation where, in the case of estimating two speakers s_1 and s_2 from the observation y, the STFT estimates of both sources are expressed as:

$$\hat{S}^1(t,f) = \frac{\sigma_1^2(f)}{\sigma_1^2(f)+\sigma_2^2(f)}Y(t,f) \tag{2.7}$$

$$\hat{S}^2(t,f) = \frac{\sigma_2^2(f)}{\sigma_1^2(f)+\sigma_2^2(f)}Y(t,f) \tag{2.8}$$

where $Y(t,f)$ is the STFT of y. $\sigma_1^2(f)$ and $\sigma_2^2(f)$ are the power spectral density (PSD) of their corresponding sources. It is noteworthy that the wiener filter can be interpreted as a time-frequency masking where the mask coefficients are the quantities weighting $Y(t,f)$ in (2.7). In general, the Wiener based separation algorithms differ from each other in the proposed model for sources and the estimation/learning approaches of its corresponding parameters. In [Benaroya *et al.* 2003], Benaroya et al. generalize Wiener filtering to estimate locally stationary non Gaussian signals. They propose a parametric model where each source is decomposed to the sum of Gaussian processes multiplied by slowly time varying amplitude factors. For each source, the PSDs of the Gaussian processes are estimated from a codebook learned heuristically in the training phase. The slowly varying amplitudes are estimated in a sparse non negative decomposition method. In [Benaroya *et al.* 2006, Benaroya & Bimbot 2003], the STFTs of sources are modeled with GMM. The wiener filtering in (2.7) is modified to a time varying Wiener filter and the STFT estimates of both sources become

$$\hat{S}^1(t,f) = \sum_{k_1,k_2} \gamma_{k_1,k_2}(t) \frac{\sigma_{k_1}^2(f)}{\sigma_{k_1}^2(f)+\sigma_{k_2}^2(f)} Y(t,f) \tag{2.9}$$

$$\hat{S}^2(t,f) = \sum_{k_1,k_2} \gamma_{k_1,k_2}(t) \frac{\sigma_{k_2}^2(f)}{\sigma_{k_1}^2(f)+\sigma_{k_2}^2(f)} Y(t,f) \tag{2.10}$$

where k_1 and k_2 are components from the Gaussian mixture of $S^1(t,f)$ and $S^2(t,f)$, $\sigma_{k_1}^2(f)$ and $\sigma_{k_2}^2(f)$ are their corresponding PSDs and $\gamma_{k_1,k_2}(t)$ is the probability of k_1 and k_2 to be active at time t. The parameters of the model which are the PSD and the prior weight of each component are estimated using the EM algorithm in the training phase.

2.5.3 BSS-based methods

BSS techniques showed successful results in solving source separation problems, especially in the determined and overdetermined cases. The most well-known BSS technique is definitely the independent component analysis (ICA) [Comon 1994]. ICA aims to find a linear transformation $\boldsymbol{W}$ that maximizes the independence of the output components of vector $\boldsymbol{y}$ where $\boldsymbol{y} = \boldsymbol{W} \times \boldsymbol{x}$ and $\boldsymbol{x}$ is the observation vector. Since only observations are used, it is necessary to make some assumptions on sources.

A1: the involved sources should be mutually statistically independent.

A2: at most one source can have Gaussian distribution.

The sources are estimated up to a multiplicative constant. ICA and its derivatives have been widely used in speech separation [Makino *et al.* 2007], especially to

generate codebooks or features for a single channel of data [Davies & James 2007]. For instance, in [jin Jang *et al.* 2003, Jang *et al.* 2003, Jang & Lee 2002], Jang et al. consider each source signal as generated by linear superpositions of weighted basis functions, the time varying weighting coefficients have generalized Gaussian pdfs. The generalized Gaussian ICA learning algorithm [Lee & Lewicki 2000] is applied for each source in the training step in order to estimate the basis functions and the parameters of the generalized Gaussian pdfs in order to use them as prior information in the main separation algorithm. In the single channel case, ICA can be efficiently used as long as the spectral supports of sources are disjoint, otherwise, the separation becomes hard.

2.5.4 NMF-based methods

The non-negative matrix factorization (NMF) is an emerging technique in BSS [Lee & Seung 1996, Lee & Seung 2000]. It is based on approximating a matrix $\boldsymbol{X}$ (the known data) as the product of two matrices, $\boldsymbol{W}$ and $\boldsymbol{H}$, with forcing the constraint that all matrices are non-negative:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \quad s.t \quad \boldsymbol{W},\ \boldsymbol{H} \geq 0 \tag{2.11}$$

The approximation in (2.11) is formulated into an optimization problem where $\boldsymbol{W}$ and $\boldsymbol{H}$ are estimated by minimizing (or maximizing) a cost function $\mathcal{D}(.)$

$$(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{H}}) = \arg \min_{\boldsymbol{W},\ \boldsymbol{H}\geq 0} \mathcal{D}(\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{H}) \tag{2.12}$$

The choice of the cost function $\mathcal{D}(.)$ depends on the application context and the prospective *a priori* information assumed for the data $\boldsymbol{X}$ and the parameters $\boldsymbol{W}$ and $\boldsymbol{H}$. Several cost functions was used such as ML, Itakura-Saito distance (ISD), Kullback-Leibler divergence (KLD), least square error (LSE), etc. In audio processing, the NMF decomposition was applied in several applications including feature extraction and music transcription [Smaragdis & Brown 2003, Abdallah & Plumbley 2004, Vincent *et al.* 2008, Bertin *et al.* 2010], sound classification [Cho *et al.* 2003, Cho & Choi 2005, Benetos *et al.* 2006, Holzapfel & Stylianou 2008], and source separation [Schmidt & Olsson 2006a, Virtanen 2007a, Schmidt & Olsson 2006b, Parry & Essa 2007]. The matrix $\boldsymbol{X}$ represents the STFT magnitude (the spectrogram) of the recorded audio sequence. The two resulting factors from the NMF decomposition of $\boldsymbol{X}$ are interpreted as the following : (i) the columns of matrix $\boldsymbol{W}$ represent the basis functions for the spectra (ii) the rows of matrix $\boldsymbol{H}$ represent the time weights corresponding to the spectral basis.

In [Smaragdis 2004], the conventional NMF algorithm is extended to convolutional NMF where components have temporal structure. In [Raj & Smaragdis 2005],

the authors propose a probabilistic framework to model the spectrogram with a mixture of multinomial distributions over frequency bins. In [Schmidt & Olsson 2006a], Schmidt et al. use sparse NMF to factorize the Mel spectrum magnitude of the recorded data into dictionary matrix and sparse code matrix. Sparse NMF ensures that the resulting decomposition is sparse more than the conventional NMF since it imposes one more penalty on the coefficients of the code matrix (L_1 norm penalty). A training phase is used to learn the overcomplete dictionaries of speakers, then the code matrix is updated while keeping the dictionary matrix fixed. In [Schmidt & Olsson 2007], another supervised method is presented based on linear regression where the magnitude time-frequency representations of the sources are this time estimated as a linear regression on features (the time weights matrix) derived from the mixture. In [Virtanen 2007b], Virtanen uses NMF jointly with temporal continuity and sparseness criteria to achieve monaural music separation. He formulates a cost function which consists of the weighted sum of three terms : (*i*) A reconstruction error term using the divergence function introduced in [Lee & Seung 2000], (*ii*) a temporal continuity criterion term that penalizes important changes between adjacent frames, and finally (*iii*) a sparseness cost function term where high coefficients in the time weights matrix are penalized.

2.5.5 Model-driven methods

Model-driven methods relies on *a priori* knowledge about the underlying speakers. Early works were done by Roweis in [Roweis 2000, Roweis 2003]. In [Roweis 2003], Roweis introduces a refiltering (masking) approach and a factorial-VQ modeling to separate the sources by an estimated spectral mask. He uses the fact that the log spectrogram of a mixture of speakers can be approximated to the element-wise maximum of the log spectrograms of the individual speakers. This approximation is motivated by speech sparsity in time/frequency domain. A training phase is necessary to learn the codebooks from clean speech of speakers individually. The MAX-log approximation was then widely used [Reddy & Raj 2004, Reyes-gomez *et al.* 2004]. Several probabilistic models were proposed such as GMM [Reddy & Raj 2004, Kristjansson *et al.* 2004, Ozerov *et al.* 2007], hidden Markov model (HMM) [Reyes-gomez *et al.* 2004] and VQ [Ellis & Weiss 2006].

Different approaches were introduced to estimate the spectral components such as the MMSE [Reddy & Raj 2004, Kristjansson *et al.* 2004, Radfar & Dansereau 2007, Radfar *et al.* 2010] and maximum a posteriori (MAP) [Ozerov *et al.* 2007] and ML [jin Jang *et al.* 2003]. In [Reddy & Raj 2007, Radfar & Dansereau 2007], the binary spectral mask (hard mask) was extended to the soft mask whose coefficients correspond to the posterior probability of the component given the data. In [Jang *et al.* 2003], a time domain approach is presented,

each source signal is modeled as a weighted sum of time-domain basis functions, the weighting coefficients are modeled by generalized Gaussian distributions. The basis functions and the pdfs of the parameters of the weighting coefficients in each source signal are learned *a priori* from a training data set using generalized Gaussian ICA learning algorithm.

In the previous approaches, a difficulty arises in practice from the fact that the source models tend to perform poorly in realistic cases, as there is generally a mismatch between the models and the actual properties of the sources in the mixture. Moreover, an accurate training able to capture the majority of sources features requires a large number of Gaussian functions. In [Ozerov *et al.* 2007], authors deal with the problem of training probabilistic models namely GMM and propose an alternative solution consisting in training the GMMs parameters from the underlying mixture in an adaptive fashion. Since it is applied to voice/music separation in popular songs, the algorithm exploits the sequences where the singer voice is absent to train the parameters of music distributions.

2.5.6 Source separation performance evaluation

The evaluation of the source separation algorithms is a very important issue. In [Vincent *et al.* 2003], Vincent et al. pointed that the evaluation depends tightly on the application/the user expectations (hearing aid, audio scene reconstruction,etc.). In some applications, we need only to estimate the mixture matrix whereas in others the sources should be estimated too. In general, two different categories of evaluation criteria are used in literature.

Objective evaluation

In a source separation algorithm, three different types of noise may degrade the quality of estimated sources: the observation noise, the presence of other interfering sources and the artifacts induced by the used method (musical noise). In [Vincent *et al.* 2006], for each extracted source, four numerical performance criteria were formulated in terms of energy ratios which are: the source-to-distortion ratio (SDR)

$$SDR = 10\log_{10}\frac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{interf}+\boldsymbol{e}_{noise}+\boldsymbol{e}_{artif}\|^2} \tag{2.13}$$

the source-to-interferences ratio (SIR)

$$SIR = 10\log_{10}\frac{\|\boldsymbol{s}_{target}\|^2}{\|\boldsymbol{e}_{interf}\|^2} \tag{2.14}$$

the source to noise ratio (SNR)

$$SNR = 10\log_{10}\frac{\|\boldsymbol{s}_{target}+\boldsymbol{e}_{interf}\|^2}{\|\boldsymbol{e}_{noise}\|^2} \tag{2.15}$$

and the sources-to-artifacts ratio (SAR)

$$SAR = 10\log_{10}\frac{\|\boldsymbol{s}_{target}+\boldsymbol{e}_{interf}+\boldsymbol{e}_{noise}\|^2}{\|\boldsymbol{e}_{artif}\|^2} \tag{2.16}$$

Here, $\boldsymbol{s}_{target}$ denotes the orthogonal projection of the estimated source on the subspace spanned by the original source, $\boldsymbol{e}_{noise}$, $\boldsymbol{e}_{interf}$ and $\boldsymbol{e}_{artif}$ denote the errors induced by the sensor noise, interferes from other sources and the algorithm artifacts respectively. These criteria provide a global assessment of the separation performance when signals are stationary across time. Since audio signals are time-varying, or more precisely, stationary across a short time, they are sliced into short windowed stationary sequences and the previous criteria are computed for each window resulting in "'local"' performance description of the separation method. The advantage of the measures defined in (2.13)-(2.16) is that they provide more accurate and precise description about the real cause of interference instead of giving global and vague description, moreover, they are valid whatever was the type of source distortion. Another important evaluation measure is the perceptual evaluation of speech quality (PESQ) [Rix *et al.* 2001] which is the closest objective measure to the subjective evaluation.

Subjective evaluation

Subjective evaluation is mainly based on listening tests achieved by a volunteer audience that hear the signals test and comment on the quality of the separation. An example of subjective evaluation is the *MUSHRA* protocol proposed in [Emiya *et al.* 2011] where audience are asked to assess the global quality compared to the reference for each test signal and then assess the quality in terms of preservation of the target source in each test signal in terms of suppression of other sources in each test signal and in terms of absence of additional artificial noise in each test signal.

2.6 conclusion

In this chapter, we presented the problem of monaural source separation. Then, we introduced the different applications in audio that show the importance of speech separation algorithms. Afterward, since we are interested in speech signals here, we gave brief description of speech signal models in literature. Next, we tried to make

an overview of the existing monaural source separation algorithms that we classified globally in two categories, the source driven methods and the model driven methods. At the end, we pointed out the issue of source separation method evaluation and presented some examples of performance measures used by the scientific community. In the next chapter, we will introduce our first contribution in this issue.

Chapter 3

Monaural speech separation using EM-Kalman and joint speech model

3.1 Introduction

Monaural speech separation is an important issue in audio processing. It helps solving "'the cocktail party problem"' where each speaker needs to be retrieved independently. Several works exploit the temporal structure of speech signal to help separation. In literature, three categories can be listed : The first exploits only the short term correlation in speech signal and models it with a short term autoregressive (AR) process [Cichocki & Thawonmas 2000]. A second category models the quasi-periodicity of speech by introducing the fundamental frequency (or pitch) in the analysis [Barros & Cichocki 2001, Tordini & Piazza 2002]. Finally, few works combine the two aspects [Smith *et al.* 2005]. In this chapter, we propose a time domain method that is classified in the last category. Our case is more difficult, since only a single sensor is used. Therefore, the proposed model of speech propagation is rather simplified (the observation is the instantaneous sum of sources). Nevertheless, this model is still relevant in several scenarios. Using some mathematical manipulations, two varieties of state space models with unknown parameters are derived from the joint model of speech where the sources are carried in the state vector. In the first variety, the state vector is linear on the unknown parameters, whereas in the second variety, the state vector is bilinear on the unknown parameters. Since the involved signals are Gaussian, Kalman filtering is used in the expectation maximization (EM) algorithm (in the "'Expectation step"') in order to estimate the required statistics of the state vector. Moreover, a second version is derived for the bilinear state space model where a fixed-interval smoothing step is achieved in order to improve the state (sources) estimates. In the simulation part, we compare the three derived algorithms using different blind source separation (BSS) evaluation criteria and finally conclude.

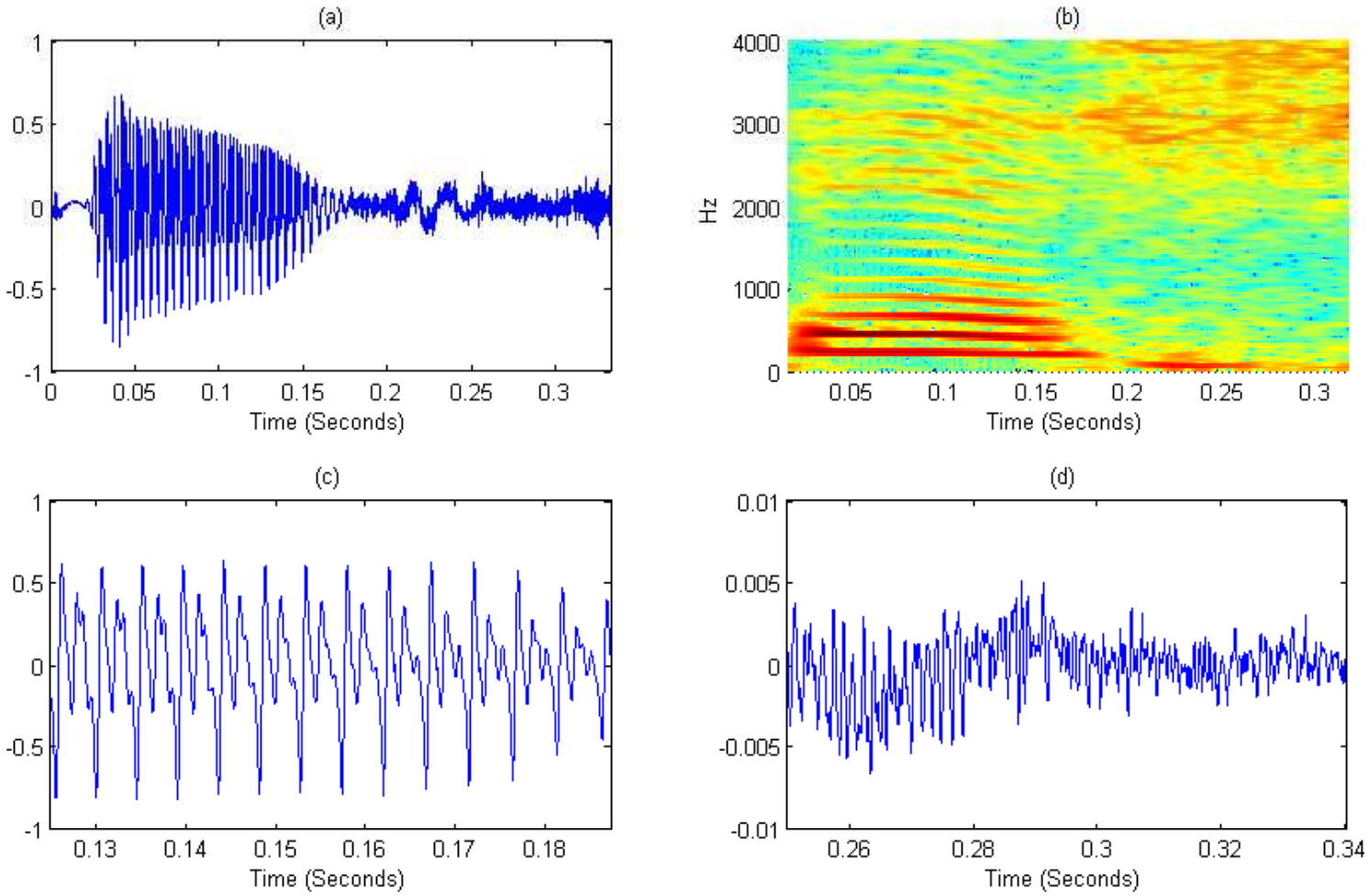

Figure 3.1: *(a) time plot of the word* dish*, (b) spectrogram of the word* dish*, (c) zoom on the voiced part* di *and (d) zoom on the unvoiced part* sh

3.2 State space model formulation

In 2.4, we cited the AR process among the models used to describe the speech sounds. This model was successfully used in speech coding framework [Sambur & Jayant 1976, Atal & Hanauer 1971, Chu 2003]. When speaking, two different kind of sounds may be produced : voiced and unvoiced sounds. Voiced sounds are produced when the vocal tract is excited by series of quasi-periodic pulses generated by the vocal cords. The frequency of pulses generation (pitch) varies according to the gender. For males the interval is of 50 to 250 Hz, whereas for female it ranges from 120 to 500 Hz [Chu 2003]. An example of voiced utterances is the vowels. Vocal sounds represent the harmonic and most energetic part of a speech sequence in contrast to the unvoiced sounds. The latter result from a turbulent air flow passing through the vocal tract when constricted. This situation occurs for example when whispering or pronouncing consonants such as p, t, k. Unvoiced sounds have more random shape and do not present any harmonic feature. An example illustrating the two kinds of sounds is presented in figure Fig. 3.1. In terms of filtering theory, the speech production mechanism can be modeled by the cascade of two infinite impulse response (IIR) filters: the formant synthesis filter that models the impulse response of the vocal tract and the pitch synthesis filter that models the vocal cords activity (see figure Fig. 3.2). The filtering equations will be

$$s_t = -\sum_{l=1}^{p} a_l \, s_{t-l} + \tilde{s}_t \tag{3.1}$$

$$\tilde{s}_t = -b \, \tilde{s}_{t-T} + e_t \tag{3.2}$$

The speech signal s_t is an AR(p) process (p-order AR process) and $\{a_l\}_{l=1..p}$ are the short term prediction (STP) coefficients that describes the strong correlation between the nearby samples. The intermediate process $\tilde{s}_t$ (called also the STP error) reflects the periodicity of s_t. In the voiced scenario, $\tilde{s}_t$ is the output of the feedback comb filter of scaling factor the long term prediction (LTP) coefficient $-b$ ($-1 < b \leq 0$) and delay the pitch period T, excited by the white Gaussian noise e_t, called the LTP error. The more $-b$ tends to 1 the more harmonic the signal is. In the unvoiced scenario, b is equal to zero, hence, the white noise is directly filtrated by the formant synthesis filter. When the pitch T is integer, one coefficient is sufficient to describe the long-term behavior such as in (3.2). Nevertheless, it is not always the case with real speech signals. When the pitch is not integer, the LTP model is described using two taps around the pitch-lag as it follows

$$\tilde{s}_t = -\beta_1 \, \tilde{s}_{t-\lceil T \rceil+1} - \beta_2 \, \tilde{s}_{t-\lceil T \rceil} + e_t \tag{3.3}$$

where $\beta_1 < 0$ and $\beta_2 < 0$. A sufficient condition of stability of this model is when the two taps are constrained to $|\beta_1| + |\beta_2| < 1$

[Ramachandran & Kabal 1987]. In some works, we may generalize the LTP model to more than two taps filter [Ramachandran & Kabal 1989, Kabal & Ramachandran 1989, Giacobello *et al.* 2009a]. Yet, the LTP filter is always sparser than the STP filter.

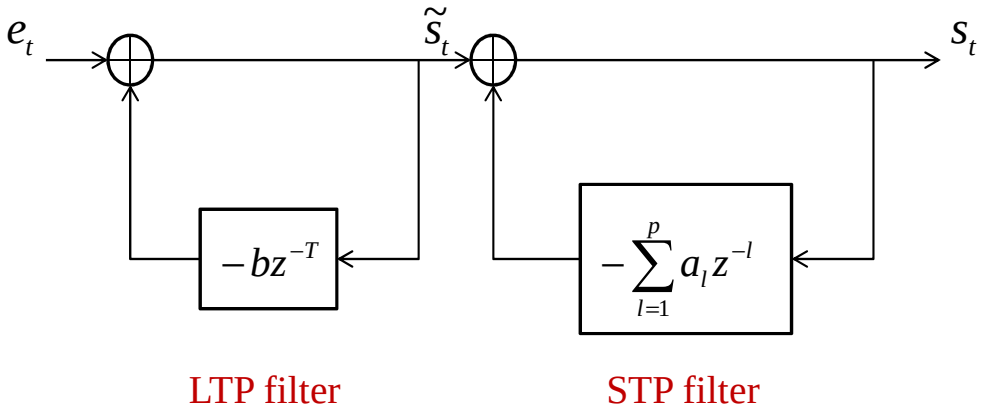

Figure 3.2: *STP and LTP filters cascade for speech production*

The idea of LTP filter was developed in speech coding to circumvent situations where the AR process has a high order p. In fact, in order to capture the whole statistical information of the speech signal, the AR order should be at least equal to the pitch period which may exceed 50 taps in terms of AR coefficients and make computation cumbersome. Moreover, it has been observed in [Chu 2004] that the main contribution to the prediction gain comes from the first 8 to 12 first taps plus the coefficient at the pitch period. Therefore, there is no need to consider the intermediate taps and the model (3.1)-(3.3) offers more compact form with less parameters. As written in (3.1)-(3.3), the filters are invariant in time. Yet, since in reality the speech signal is stationarity piecewise, this model is reliable within the processing frame. The purpose from the previous part was to introduce the parametric model we are going to use for speech sources hereafter. We remind our main problem of estimating K mixed sources from scalar observation y_t. Each source is modeled by the joint STP+LTP model. Then, the probelm is fully described by the equations system

$$y_t = \sum_{k=1}^{K} s_{k,t} \; + \; n_t \tag{3.4}$$

$$s_{k,t} = -\sum_{l=1}^{p_k} a_{k,l} \; s_{k,t-l} \; + \; \tilde{s}_{k,t} \qquad k = 1 \cdots K \tag{3.5}$$

$$\tilde{s}_{k,t} = -\beta_{k,1} \; \tilde{s}_{k,t-\lceil T_k \rceil + 1} \; - \beta_{k,2} \; \tilde{s}_{k,t-\lceil T_k \rceil} \; + \; e_{k,t} \qquad k = 1 \cdots K \tag{3.6}$$

Here, the index k points to the k^{th} source. The LTP errors $\{e_{k,t}\}_{k=1..K}$ are statis-

tically independent zero mean Gaussian processes with variances $\{\rho_k\}_{k=1..K}$. The observation noise n_t is a white zero mean Gaussian process with variance σ_n^2, statistically independent of the innovations $\{e_{k,t}\}_{k\,=\,1..K}$. The idea is to use (3.4)-(3.6) to formulate a state space model where (3.4) is used to form the observation equation and the state space evolution is formed using (3.5) and (3.6).

3.2.1 Linear state space model

Let's denote by $\boldsymbol{x}_{k,t}$ the $(N + p_k + 2)$-dimensional substate vector constructed by the successive last $(p_k + 2)$ samples of $s_{k,t}$ and the last N samples of $\tilde{s}_{k,t}$ where N should be large enough to capture the LTP coefficients of the largest pitch period present in the mixture

$$\boldsymbol{x}_{k,t} = [s_{k,t} \cdots s_{k,t-p_k-1} \mid \tilde{s}_{k,t} \cdots \tilde{s}_{k,t-\lceil T_k \rceil+1}\ \tilde{s}_{k,t-\lceil T_k \rceil} \cdots \tilde{s}_{k,t-N+1}]^T \quad k = 1...K \tag{3.7}$$

The substate vector in (3.7) can be expressed in terms of $\boldsymbol{x}_{k,t-1}$ as it follows

$$\boldsymbol{x}_{k,t} = \boldsymbol{F}_k\ \boldsymbol{x}_{k,t-1}\ +\ \boldsymbol{g}_k\ e_{k,t}, \quad k = 1...K \tag{3.8}$$

where $\boldsymbol{g}_k$ is the $(N + p_k + 2)$-dimensional column vector defined as $\boldsymbol{g}_k = [\ 1\ 0 \cdots 0 \mid 1\ 0\ \cdots\cdots\ 0]^T$. The second nonzero component in $\boldsymbol{g}_k$ is at the $(p_k + 3)^{th}$ position. The $(N + p_k + 2) \times (N + p_k + 2)$ matrix $\boldsymbol{F}_k$ has the following structure

$$\boldsymbol{F}_k\ =\ \begin{bmatrix} \boldsymbol{F}_{11,k} & \boldsymbol{F}_{12,k} \\ \boldsymbol{0} & \boldsymbol{F}_{22,k} \end{bmatrix}$$

where the $(p_k + 2) \times (p_k + 2)$ matrix $\boldsymbol{F}_{11,k}$, the $(p_k + 2) \times N$ matrix $\boldsymbol{F}_{12,k}$ and the $N \times N$ matrix $\boldsymbol{F}_{22,k}$ are given by

$$\boldsymbol{F}_{11,k}\ =\ \begin{bmatrix} -a_{k,1} & -a_{k,2} & \cdots & -a_{k,p_k} & 0 & 0 \\ & & & & & \vdots \\ & & I_{(p_k+1)} & & & \vdots \\ & & & & & \vdots \\ & & & & & 0 \end{bmatrix}$$

$$\boldsymbol{F}_{12,k}\ =\ \begin{bmatrix} 0 & \cdots & -\beta_{k,1} & -\beta_{k,2} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\boldsymbol{F}_{22,k}\ =\ \begin{bmatrix} 0 & \cdots & -\beta_{k,1} & -\beta_{k,2} & 0 & \cdots & 0 \\ & & & & & & \vdots \\ & & & I_{(N-1)} & & & \vdots \\ & & & & & & \vdots \\ & & & & & & 0 \end{bmatrix}$$

In the matrices $\boldsymbol{F}_{12,k}$ and $\boldsymbol{F}_{22,k}$, the coefficients $-\beta_{k,1}$ and $-\beta_{k,2}$ are situated in the $\lfloor T_k \rfloor^{th}$ and $\lceil T_k \rceil^{th}$ columns respectively. Since K sources are present in the mixture, the final state is formed by concatenating the K subspace vectors in (3.7) into one vector $\boldsymbol{x}_t = [\boldsymbol{x}_{1,t}^T \ \boldsymbol{x}_{2,t}^T \ \cdots \ \boldsymbol{x}_{K,t}^T]^T$ which results in the time update equation (3.9). Moreover, by reformulating the expression of (3.4) in function of $\boldsymbol{x}_t$, we obtain the final observation equation (3.10). The final state space model is then deduced

$$\boldsymbol{x}_t = \boldsymbol{F}\ \boldsymbol{x}_{t-1} + \boldsymbol{G}\ \boldsymbol{e}_t \tag{3.9}$$

$$y_t = \boldsymbol{h}^T \boldsymbol{x}_t + n_t \tag{3.10}$$

where

- the K-dimensional column vector $\boldsymbol{e}_t$ results from the concatenation of the K LTP errors, thus, is zero mean Gaussian of covariance matrix the $K \times K$ diagonal matrix $\boldsymbol{Q} = \text{diag}\,(\rho_1, \cdots, \rho_K)$.
- $\boldsymbol{F}$ is the $\sum_{k=1}^{K}(p_k + N + 2) \times \sum_{k=1}^{K}(p_k + N + 2)$ block diagonal matrix given by $\boldsymbol{F} = \text{blockdiag}\,(\boldsymbol{F}_1, \cdots, \boldsymbol{F}_K)$.
- $\boldsymbol{G}$ is the $\sum_{k=1}^{K}(p_k + N + 2) \times K$ matrix given by $\boldsymbol{G} = \text{blockdiag}\,(\boldsymbol{g}_1, \cdots, \boldsymbol{g}_K)$.
- $\boldsymbol{h}$ is the $(\sum_{k=1}^{K}(p_k + N + 2))$-dimensional column vector given by $\boldsymbol{h} = \mathbb{1}_K \otimes \boldsymbol{u}_1$ where $\mathbb{1}_K$ is the all-ones K-dimensional vector and $\boldsymbol{u}_1$ is the first vector in the standard basis for the $(N + p_k + 2)$-dimensional space.

3.2.2 Bilinear state space model

In the bilinear state space model, the substate vector is only composed of successive samples of $s_{k,t}$ without using the STP error. Yet the number of samples has to be high enough to capture the long-term effect. In other words, the STP model in (3.6) is re-expressed using the sources samples as it follows

$$\begin{aligned} s_{k,t} = -\sum_{l=1}^{p_k} a_{k,l}\ s_{k,t-l} \ - \beta_{k,1} \sum_{l=0}^{p_k} a_{k,l}\ s_{k,t-l-\lceil T_k \rceil+1} \\ - \beta_{k,2} \sum_{l=0}^{p_k} a_{k,l}\ s_{k,t-l-\lceil T_k \rceil} + e_{k,t} \end{aligned} \tag{3.11}$$

where $a_{k,0} = 1$ for all k. In this case, we define the N-dimensional substate vector $\boldsymbol{x}_{k,t}$ as it follows

$$\begin{aligned} \boldsymbol{x}_{k,t} = [s_{k,t}\ s_{k,t-1} \cdots s_{k,t-\lceil T_k \rceil+1}\ s_{k,t-\lceil T_k \rceil} \cdots s_{k,t-p-\lceil T_k \rceil+1} \\ s_{k,t-p-\lceil T_k \rceil} \cdots s_{k,t-N+1}]^T \quad k = 1...K \end{aligned} \tag{3.12}$$

Compared to the linear state space model, the substate matrix $\boldsymbol{F}_k$ is modified to the form

$$\boldsymbol{F}_k = \begin{bmatrix} & \boldsymbol{f}_{k,1}^T & 0 \\ & & \vdots \\ & I_{(N-1)} & \vdots \\ & & 0 \end{bmatrix} \tag{3.13}$$

where $\boldsymbol{f}_{k,1}$ is the N-dimensional vector defined as

$$\begin{aligned} \boldsymbol{f}_{k,1} = [&-a_{k,1} \ \cdots \ -a_{k,p} \ 0 \cdots 0 \ -\beta_{k,1} \quad -\beta_{k,1}a_{k,1} - \beta_{k,2} \\ &-\beta_{k,1}a_{k,2} - \beta_{k,2}a_{k,1} \cdots - \beta_{k,1}a_{k,p} - \beta_{k,2}a_{k,p-1} \quad -\beta_{k,2}a_{k,p} \ 0 \cdots 0]^T \end{aligned} \tag{3.14}$$

The parameters of the state space model in (3.9-3.10) are then modified to

- $\boldsymbol{F}$ is the $KN \times KN$ block diagonal matrix given by $\boldsymbol{F} = \text{blockdiag}\,(\boldsymbol{F}_1, \cdots, \boldsymbol{F}_K)$.
- $\boldsymbol{G}$ is the $KN \times K$ matrix given by $\boldsymbol{G} = \text{blockdiag}\,(\boldsymbol{g}, \cdots, \boldsymbol{g})$ where $\boldsymbol{g}$ is the N-dimensional vector $[1 \ 0 \cdots 0]^T$.
- $\boldsymbol{h}$ is the KN-dimensional column vector given by $\boldsymbol{h} = \mathbb{1}_K \otimes \boldsymbol{g}$.

Considering the shape of $\boldsymbol{F}_k$ (consequently $\boldsymbol{F}$), the state vector is no longer linear on the parameters (more precisely on the STP+LTP coefficients), it becomes rather bilinear whence the name of the state space model.

The straightforward classic solution to estimate the state vector is to use Kalman filter [Kalman 1960]. Yet, the linear and bilinear dynamic systems presented beforehand depend on the STP+LTP parameters that are recapitulated in the variable $\boldsymbol{\theta}$

$$\begin{aligned} \boldsymbol{\theta} &= [\theta_0, \boldsymbol{\theta}_1^T, \cdots, \boldsymbol{\theta}_K^T]^T \\ \boldsymbol{\theta}_k &= [\boldsymbol{a}_k^T, \boldsymbol{\beta}_k^T, T_k, \rho_k]^T \ , \quad k = 1...K \\ \boldsymbol{a}_k &= [a_{k,1}, \cdots, a_{k,p_k}]^T \ , \quad k = 1...K \\ \boldsymbol{\beta}_k &= [\beta_{k,1}, \beta_{k,2}]^T \ , \quad k = 1...K \\ \theta_0 &= \sigma_n^2 \end{aligned} \tag{3.15}$$

Obviously, the parameters vector $\boldsymbol{\theta}$ is unknown. Therefore, a joint estimation framework is needed to estimate both the sources and the parameters. The joint estimation of latent variables (the random unknown parameters) and unknown deterministic parameters using only observation is a well-known problem in statistical signal processing. A famous solution to this problem is the EM-Kalman algorithm [Couvreur & Bresler 1995, Gao *et al.* 2003, Feder & Weinstein 1988], that is proved to converge to the maximum likelihood (ML) solution [Dempster *et al.* 1977]. In the next section, we introduce this algorithm and develop it for our specific model.

3.3 EM-Kalman Filter

The EM algorithm was introduced the first time by Dempster et al. in [Dempster *et al.* 1977]. It is an iterative algorithm that aims to estimate the ML solution of a deterministic parameter $\boldsymbol{\theta}$ from the observed data $\boldsymbol{y}$ (incomplete data) in presence of random hidden variables $\boldsymbol{x}$. The set $\{\boldsymbol{y}, \boldsymbol{x}\}$ represents the complete data. In general,EM is used when the derivation of the ML estimator of $\boldsymbol{\theta}$ is cumbersome or intractable. The injection of the hidden variables $\boldsymbol{x}$ helps to alleviate computations and provides an alternative way to approach the ML solution asymptotically [Dempster *et al.* 1977]. If we denote by $\mathcal{F}$ the cost function

$$\mathcal{F}(q(\boldsymbol{x}), \boldsymbol{\theta}) \triangleq \int q(\boldsymbol{x}) \ln \frac{f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})}{q(\boldsymbol{x})} d\boldsymbol{x} \tag{3.16}$$

Where $q(\boldsymbol{x})$ is a free distribution of the hidden variables $\boldsymbol{x}$. The cost function in (3.16) is a lower bound on the likelihood function $f(\boldsymbol{y}|\boldsymbol{\theta})$ and EM consists on maximizing it relative to $q(\boldsymbol{x})$ and $\boldsymbol{\theta}$ in two main steps. Let's suppose the algorithm executed i iterations, at the $(i+1)^{th}$ iteration, EM alternates between two steps.

E-step:

In this step, $\mathcal{F}$ is maximized relative to the distribution $q(\boldsymbol{x})$

$$q(\boldsymbol{x})^{i+1} \leftarrow \arg\max_{q(\boldsymbol{x})} \mathcal{F}(q(\boldsymbol{x}), \boldsymbol{\theta}^i) \tag{3.17}$$

Where $\boldsymbol{\theta}^i$ is the estimate of $\boldsymbol{\theta}$ at iteration i. It turns out that the maximum is obtained for $q(\boldsymbol{x})^{i+1} = f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^i)$ which is the posterior distribution of the hidden variables given the observations. When the involved signals are Gaussian, the inference of $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^i)$ is reduced to the mean and covariance estimation which is realized using the Kalman filter (for online processing) or smoother (for offline processing).

M-step:

In this step, $\mathcal{F}$ is maximized relative to the parameter $\boldsymbol{\theta}$

$$\begin{aligned}
\boldsymbol{\theta}^{i+1} &\leftarrow \arg\max_{\boldsymbol{\theta}} \mathcal{F}(q(\boldsymbol{x})^{(i+1)}, \boldsymbol{\theta}) && (3.18)\\
&\leftarrow \arg\max_{\boldsymbol{\theta}} \langle \ln f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) \rangle_{f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^i)} && (3.19)\\
&\leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^i) && (3.20)
\end{aligned}$$

The second line results from replacing $q(\boldsymbol{x})^{(i+1)}$ by its expression from the *E-step* where $\langle . \rangle_{f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^i)}$ denotes $E_{f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^i)}[.]$. Let's suppose we have M samples

$\{y_1, y_2, \cdots, y_M\}$ of data stream that we note $y_{1:M}$ and $\boldsymbol{x}_{1:M}$, M state vectors, the cost function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^i)$ is expressed as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^i) \quad = \quad E_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)}[\ln f(y_{1:M}, \boldsymbol{x}_{1:M}|\boldsymbol{\theta})] \tag{3.21}$$

In what follows, we develop the EM-Kalman for both linear and bilinear models.

3.3.1 EM-Kalman for linear state space model

Considering the state space model in (3.9)-(3.10) and after some mathematical computations, the cost function (3.21) is expressed

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^i) = & -\frac{M}{2}\ln\sigma_n^2 - \frac{1}{2\sigma_n^2}\left(Y - 2\boldsymbol{h}^T\boldsymbol{R}^i_{y_t\boldsymbol{x}_t} + \boldsymbol{h}^T\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t}\boldsymbol{h}\right) \\ & - \sum_{k=1}^{K}\frac{M}{2}\ln\rho_k - \sum_{k=1}^{K}\frac{\rho_k}{2}\sum_{t=1}^{M}\Big\langle\Big(s_{k,t} + \sum_{l=1}^{p_k} a_{k,l}\ s_{k,t-l} \\ & + \beta_{k,1}\ \tilde{s}_{k,t-\lceil T_k\rceil+1}\ + \beta_{k,2}\ \tilde{s}_{k,t-\lceil T_k\rceil}\Big)^2 + \ln f(\boldsymbol{x}_0)\Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \end{aligned} \tag{3.22}$$

where Y, $\boldsymbol{R}^i_{y_t\boldsymbol{x}_t}$ and $\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t}$ denote the following amounts

$$Y \quad \triangleq \quad \sum_{t=1}^{M} y_t^2 \tag{3.23}$$

$$\boldsymbol{R}^i_{y_t\boldsymbol{x}_t} \quad \triangleq \quad \sum_{t=1}^{M} y_t\ \langle\boldsymbol{x}_t\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \tag{3.24}$$

$$\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t} \quad \triangleq \quad \sum_{t=1}^{M} \left\langle\boldsymbol{x}_t\boldsymbol{x}_t^T\right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \tag{3.25}$$

The *M-step* will consist on maximizing (3.22) w.r.t. the STP and LTP coefficients $\{\boldsymbol{a}_k\}_{k=1:K}$ and $\{\boldsymbol{\beta}_k\}_{k=1:K}$, the powers of the LTP errors $\{\rho_k\}_{k=1:K}$ and the observation noise power σ_n^2.

Estimation of the LTP coefficients

The estimate of the LTP coefficients vector $\boldsymbol{\beta}_k$ of source k from (3.22) is deduced by computing more precisely the gradient of the reduced cost function

$$Q_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}^i) = \sum_{t=1}^{M}\left\langle\left(s_{k,t}+\sum_{l=1}^{p_k} a_{k,l}\ s_{k,t-l}+\beta_{k,1}\ \tilde{s}_{k,t-\lceil T_k\rceil+1}\ +\beta_{k,2}\ \tilde{s}_{k,t-\lceil T_k\rceil}\right)^2\right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \tag{3.26}$$

By replacing the expression $s_{k,t} + \sum_{l=1}^{p_k} a_{k,l}\ s_{k,t-l}$ with $\tilde{s}_{k,t}$ in (3.26), the reduced cost function $Q_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}^i)$ becomes only dependent on $\boldsymbol{\beta}_k$. The estimate of $\boldsymbol{\beta}_k$ at iteration $(i+1)$ is then expressed as

$$\hat{\boldsymbol{\beta}}_k^{\,i+1} = -\left(\boldsymbol{R}^i_{\tilde{s},k}\right)^{-1}\ \boldsymbol{r}^i_{\tilde{s},k} \quad k = 1\cdots K \tag{3.27}$$

where the 2×2-dimensional matrix $\boldsymbol{R}^i_{\tilde{s},k}$ and the 2-dimensional vector $\boldsymbol{r}^i_{\tilde{s},k}$ are defined as it follows

$$\boldsymbol{R}^i_{\tilde{s},k} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k\rceil+1} \\ \tilde{s}_{k,t-\lceil T_k\rceil} \end{bmatrix} \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k\rceil+1} \\ \tilde{s}_{k,t-\lceil T_k\rceil} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.28}$$

$$\boldsymbol{r}^i_{\tilde{s},k} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k\rceil+1} \\ \tilde{s}_{k,t-\lceil T_k\rceil} \end{bmatrix} \tilde{s}_{k,t} \right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.29}$$

The result in (3.27) represents the Yule-Walker equations derived for the long-term AR process $\tilde{s}_{k,t}$.

Estimation of the STP coefficients

The STP coefficients vector $\boldsymbol{a}_k$ of source k is also deduced using $Q_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}^i)$. The estimate of $\boldsymbol{a}_k$ at iteration $(i+1)$ is expressed as

$$\hat{\boldsymbol{a}_k}^{i+1} = -\left(\boldsymbol{R}^i_{s,k}\right)^{-1} \left(\boldsymbol{r}^i_{s,k} + \boldsymbol{R}^i_{s\tilde{s},k}\hat{\boldsymbol{\beta}_k}^{i+1}\right) \quad k = 1\cdots K \tag{3.30}$$

where the $p\times p$-dimensional matrix $\boldsymbol{R}^i_{s,k}$, the p-dimensional vector $\boldsymbol{r}^i_{s,k}$ and the $p\times 2$-dimensional matrix $\boldsymbol{R}^i_{s\tilde{s},k}$ are defined as it follows

$$\boldsymbol{R}^i_{s,k} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} s_{k,t-1} \\ \vdots \\ s_{k,t-p} \end{bmatrix} \begin{bmatrix} s_{k,t-1} \\ \vdots \\ s_{k,t-p} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.31}$$

$$\boldsymbol{r}^i_{s,k} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} s_{k,t-1} \\ \vdots \\ s_{k,t-p} \end{bmatrix} s_{k,t} \right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.32}$$

$$\boldsymbol{R}^i_{s\tilde{s},k} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} s_{k,t-1} \\ \vdots \\ s_{k,t-p} \end{bmatrix} \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k\rceil+1} \\ \tilde{s}_{k,t-\lceil T_k\rceil} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.33}$$

Estimation of the LTP error and observation noise variances

Since ρ_k is the innovation power of the AR process $\tilde{s}_{k,t}$, its estimation can be simply deduced as the inverse of the first component of the matrix $\left(\frac{1}{M}\ \tilde{\boldsymbol{R}}^i_k\right)^{-1}$

$$\hat{\rho}^{i+1}_k = \left[\left(\frac{1}{M}\ \tilde{\boldsymbol{R}}^i_k\right)^{-1}\right]^{-1}_{11} \quad k = 1\cdots K \tag{3.34}$$

where

$$\tilde{\boldsymbol{R}}^i_k = \sum_{t=1}^{M} \left\langle \begin{bmatrix} \tilde{s}_{k,t} \\ \tilde{s}_{k,t-\lceil T_k\rceil+1} \\ \tilde{s}_{k,t-\lceil T_k\rceil} \end{bmatrix} \begin{bmatrix} \tilde{s}_{k,t} \\ \tilde{s}_{k,t-\lceil T_k\rceil+1} \\ \tilde{s}_{k,t-\lceil T_k\rceil} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.35}$$

The observation noise $\hat{\sigma_n^2}^{i+1}$ is deduced from the gradient of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^i)$ as the following

$$\hat{\sigma_n^2}^{i+1} = \frac{1}{M}\left(Y - 2\boldsymbol{h}^T\boldsymbol{R}^i_{y_t\boldsymbol{x}_t} + \boldsymbol{h}^T\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t}\boldsymbol{h}\right) \tag{3.36}$$

All the quantities used to estimate $\boldsymbol{a}_k$, $\boldsymbol{\beta}_k$ and ρ_k form the elements of the $(p_k + 4) \times (p_k + 4)$-dimensional matrix $\breve{\boldsymbol{R}}^i_k$ defined as

$$\breve{\boldsymbol{R}}^i_k = \sum_{t=1}^{M} \left\langle \breve{\boldsymbol{x}}_{k,t}\breve{\boldsymbol{x}}^T_{k,t} \right\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1 \cdots K \tag{3.37}$$

where $\breve{\boldsymbol{x}}_{k,t}$ is the $(p_k + 4)$-dimensional partial state vector

$$\breve{\boldsymbol{x}}_{k,t} = [s_{k,t}, \cdots, s_{k,t-p}, \tilde{s}_{k,t}, \tilde{s}_{k,t-\lceil T_k \rceil+1}, \tilde{s}_{k,t-\lceil T_k \rceil}]^T \quad k = 1 \cdots K \tag{3.38}$$

$\breve{\boldsymbol{R}}^i_k$ (and $\breve{\boldsymbol{x}}_{k,t}$) can be selected from $\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t}$ (and $\boldsymbol{x}_{k,t}$) using an appropriate selection matrix $\boldsymbol{S}_k$.

In the *E-step*, we need to compute the required statistics (3.23), (3.24) and (3.25). For the two latter, we use the Rauch-Tung-Striebel (RTS) smoother [Rauch *et al.* 1965] which is the most common form of the fixed-interval smoother. Let's denote by $\hat{\boldsymbol{x}}_{t|M}$ the RTS state estimate and $\boldsymbol{P}_{t|M}$ its corresponding covariance matrix, then the required statistics are expressed as

$$\boldsymbol{R}^i_{y_t\boldsymbol{x}_t} = \sum_{t=1}^{M} y_t\hat{\boldsymbol{x}}_{t|M} \tag{3.39}$$

$$\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t} = \sum_{t=1}^{M} \boldsymbol{P}_{t|M} + \hat{\boldsymbol{x}}_{t|M}\hat{\boldsymbol{x}}^T_{t|M} \tag{3.40}$$

Stop condition

The EM algorithm is proved to increase monotonically to at least a local maximum of the likelihood function $f(\boldsymbol{y}|\boldsymbol{\theta})$ [Dempster *et al.* 1977]. Therefore, the algorithm is considered converged when the produced likelihoods are stable during sequential iterations. In practical, in order to handle the numerical imprecision, a halting condition consists in measuring the relative change of the log-likelihood [Shumway & Stoffer 1982]. If we denote by $\mathcal{L}^i$ the log-likelihood function produced at the i^{th} iteration and expressed as

$$\mathcal{L}^i = -\frac{1}{2}\sum_{t=1}^{M} \ln\left|\boldsymbol{h}^T\boldsymbol{P}_{t|M}\boldsymbol{h} + \hat{\sigma_n^2}^i\right| - \frac{1}{2}\sum_{t=1}^{M} \frac{\left\|y_t - \boldsymbol{h}^T\hat{\boldsymbol{x}}_{t|M}\right\|^2}{\boldsymbol{h}^T\boldsymbol{P}_{t|M}\boldsymbol{h} + \hat{\sigma_n^2}^i} \tag{3.41}$$

The relative measure of the log-likelihood variations, used as stop condition is expressed as

$$\text{Stop if } \frac{\mathcal{L}^{i+1} - \mathcal{L}^i}{\frac{1}{2}\left|\mathcal{L}^{i+1} + \mathcal{L}^i + \varepsilon\right|} < c \tag{3.42}$$

where ε is a small term added to keep the condition well-behaved, c is a fixed threshold that is determined experimentally. In general c is chosen in the interval $[10^{-4}, 10^{-5}]$.

The full iterative algorithm is summarized in algorithm. 1.

Algorithm 1 Iterative EM-Kalman with RTS smoothing

1: Initialize the parameters $\left\{\hat{\boldsymbol{\beta}}_k^0\right\}_{k=1..K}$, $\{\hat{\boldsymbol{a}}_k^0\}_{k=1..K}$, $\{\hat{\rho}_k^0\}_{k=1..K}$, $\hat{\sigma_n^2}^0$, $\left\{\hat{T}_k\right\}_{k=1..K}$, and Kalman filter with $\hat{\boldsymbol{x}}_{0|0}$ and $\boldsymbol{P}_{0|0}$.

2: While the stop condition not true and $i < i_{max}$ do

3: Run Kalman filter, for $t = 1$ to M

$$\hat{\boldsymbol{x}}_{t|t-1} = \widehat{\boldsymbol{F}}^i \hat{\boldsymbol{x}}_{t-1|t-1} \tag{3.43}$$

$$\boldsymbol{P}_{t|t-1} = \widehat{\boldsymbol{F}}^i \boldsymbol{P}_{t-1|t-1} \left(\widehat{\boldsymbol{F}}^i\right)^T + \boldsymbol{G}\widehat{\boldsymbol{Q}}^i \boldsymbol{G}^T \tag{3.44}$$

$$\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{h}(\boldsymbol{h}^T \boldsymbol{P}_{t|t-1}\boldsymbol{h} + \hat{\sigma_n^2}^i)^{-1} \tag{3.45}$$

$$\hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t(y_t - \boldsymbol{h}^T \hat{\boldsymbol{x}}_{t|t-1}) \tag{3.46}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t \boldsymbol{h}^T \boldsymbol{P}_{t|t-1} \tag{3.47}$$

4: Initialize RTS smoother with $\hat{\boldsymbol{x}}_{M|M}$ and $\boldsymbol{P}_{M|M}$.

5: Run RTS smoother, for $t = M - 1$ to 0

$$\boldsymbol{K}_{rts,t} = \boldsymbol{P}_{t|t} \widehat{\boldsymbol{F}}^i \boldsymbol{P}_{t+1|t}^{-1} \tag{3.48}$$

$$\boldsymbol{P}_{t|M} = \boldsymbol{P}_{t|t} - \boldsymbol{K}_{rts,t} \left(\boldsymbol{P}_{t+1|t} - \boldsymbol{P}_{t+1|M}\right) \boldsymbol{K}_{rts,t}^T \tag{3.49}$$

$$\hat{\boldsymbol{x}}_{t|M} = \hat{\boldsymbol{x}}_{t|t} + \boldsymbol{K}_{rts,t} \left(\hat{\boldsymbol{x}}_{t+1|M} - \hat{\boldsymbol{x}}_{t+1|t}\right) \tag{3.50}$$

6: Compute $\boldsymbol{R}^i_{y_t \boldsymbol{x}_t}$ and $\boldsymbol{R}^i_{\boldsymbol{x}_t \boldsymbol{x}_t}$ using (3.39) and (3.40) respectively.

7: Update the parameters $\left\{\hat{\boldsymbol{\beta}}_k^{i+1}\right\}_{k=1..K}$, $\{\hat{\boldsymbol{a}}_k^{i+1}\}_{k=1..K}$, $\{\hat{\rho}_k^{i+1}\}_{k=1..K}$ and $\hat{\sigma_n^2}^{i+1}$ using (3.27),(3.30), (3.34) and (3.36) respectively.

8: If the stop condition is fulfilled or the iteration value $i > i_{max}$, then stop the algorithm, otherwise, back to the Kalman filtering step.

The RTS smoother provide an improved estimate of the sources carried in the smoothed state vector $\hat{\boldsymbol{x}}_{t|M}$ though it increases the complexity of the algorithm (we make two sweeps of data in each iteration). In our state space model, the prediction error matrix $\boldsymbol{P}_{t|t-1}$ is singular. In fact, notice that in $\boldsymbol{F}_{k,11}$ and $\boldsymbol{F}_{k,22}$, the last columns are all-zeros, which makes the rank of the substate matrix $\boldsymbol{F}_k$ deficient by 2. Consequently, the state matrix $\boldsymbol{F}$ is rank-deficient by $2K$. In (3.44), the additive term $\boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}^T$ is of rank K which does not correct the singularity of $\boldsymbol{P}_{t|t-1}$. In the smoothing part, the computation of the RTS gain in $\boldsymbol{K}_{rts,t}$ requires the inversion of the prediction error covariance matrix, which results in ill-conditioned gain matrix $\boldsymbol{K}_{rts,t}$.

This fact yields to drop the RTS smoothing step and be restricted to the Kalman filter only. Hence, $f(\boldsymbol{x}|y_{1:M}, \boldsymbol{\theta}^i)$ is replaced by $f(\boldsymbol{x}|y_{1:t}, \boldsymbol{\theta}^i)$ in all the previous derivations ($\hat{\boldsymbol{x}}_{t|M}$ and $\boldsymbol{P}_{t|M}$ are replaced by the Kalman filter estimator $\hat{\boldsymbol{x}}_{t|t}$ and its covariance matrix $\boldsymbol{P}_{t|t}$ respectively). Nevertheless, the filtering here is done in a tricky way so that a fixed-lag smoothing operation (lag = 1) is achieved at the same time. In fact, due to the special structure of the state vector $\boldsymbol{x}_t$, it is possible to select from it the delayed partial state $\breve{\boldsymbol{x}}_{k,t-1}$ using an appropriate different selection matrix $\boldsymbol{S}'_k$. The 1-lag smoothing justifies the use of $(p_k + 2)$ consecutive samples of s_k in $\boldsymbol{x}_{k,t}$ instead of $(p_k + 1)$ samples. The number of STP error samples present in $\boldsymbol{x}_{k,t}$, N, has to be at least equal to $(\max(\lceil T_k \rceil) + 2)$ in order to guarantee the presence of $\breve{\boldsymbol{x}}_{k,t-1}$ in $\boldsymbol{x}_t$ for all k. Then, the smoothed quantities needed for the parameters estimation form the elements of the matrix $\breve{\boldsymbol{R}}_k^{i\,\prime}$

$$\breve{\boldsymbol{R}}_k^{i\,\prime} = \sum_{t=1}^{M} \left\langle \breve{\boldsymbol{x}}_{k,t-1} \breve{\boldsymbol{x}}_{k,t-1}^T \right\rangle_{f(\boldsymbol{x}|y_{1:t},\boldsymbol{\theta}^i)} \tag{3.51}$$

$$= \sum_{t=1}^{M} \boldsymbol{S}'_k \left\langle \boldsymbol{x}_{k,t} \boldsymbol{x}_{k,t}^T \right\rangle_{f(\boldsymbol{x}|y_{1:t},\boldsymbol{\theta}^i)} \boldsymbol{S}_k'^T \tag{3.52}$$

$$= \boldsymbol{S}'_k \boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t} \boldsymbol{S}_k'^T \tag{3.53}$$

where $\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t}$, as well as $\boldsymbol{R}^i_{y_t\boldsymbol{x}_t}$, are re-expressed as

$$\boldsymbol{R}^i_{y_t\boldsymbol{x}_t} = \sum_{t=1}^{M} y_t \hat{\boldsymbol{x}}_{t|t} \tag{3.54}$$

$$\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t} = \sum_{t=1}^{M} \boldsymbol{P}_{t|t} + \hat{\boldsymbol{x}}_{t|t} \hat{\boldsymbol{x}}_{t|t}^T \tag{3.55}$$

Notice the 1-lag smoothing effect from (3.51) to (3.52).The choice of the lag is justified with the fact that the state vector $\boldsymbol{x}_t$ is an AR process of order 1, hence, 1-lag smoothing is considered sufficient to improve sources estimation. The smoothed

elements selected from $\breve{\boldsymbol{R}}_k^{i\,\prime}$ become

$$\boldsymbol{R}_{\tilde{s},k}^{i} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k \rceil} \\ \tilde{s}_{k,t-\lceil T_k \rceil-1} \end{bmatrix} \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k \rceil} \\ \tilde{s}_{k,t-\lceil T_k \rceil-1} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:t},\boldsymbol{\theta}^i)} \quad k = 1 \cdots K \tag{3.56}$$

$$\boldsymbol{r}_{\tilde{s},k}^{i} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k \rceil} \\ \tilde{s}_{k,t-\lceil T_k \rceil-1} \end{bmatrix} \tilde{s}_{k,t-1} \right\rangle_{f(\boldsymbol{x}_t|y_{1:t},\boldsymbol{\theta}^i)} \quad k = 1 \cdots K \tag{3.57}$$

$$\boldsymbol{R}_{s,k}^{i} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} s_{k,t-2} \\ \vdots \\ s_{k,t-p-1} \end{bmatrix} \begin{bmatrix} s_{k,t-2} \\ \vdots \\ s_{k,t-p-1} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:t},\boldsymbol{\theta}^i)} \quad k = 1 \cdots K \tag{3.58}$$

$$\boldsymbol{r}_{s,k}^{i} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} s_{k,t-2} \\ \vdots \\ s_{k,t-p-1} \end{bmatrix} s_{k,t-1} \right\rangle_{f(\boldsymbol{x}_t|y_{1:t},\boldsymbol{\theta}^i)} \quad k = 1 \cdots K \tag{3.59}$$

$$\boldsymbol{R}_{s\tilde{s},k}^{i} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} s_{k,t-2} \\ \vdots \\ s_{k,t-p-1} \end{bmatrix} \begin{bmatrix} \tilde{s}_{k,t-\lceil T_k \rceil} \\ \tilde{s}_{k,t-\lceil T_k \rceil-1} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:t},\boldsymbol{\theta}^i)} \quad k = 1 \cdots K \tag{3.60}$$

$$\tilde{\boldsymbol{R}}_{k}^{i} = \sum_{t=1}^{M} \left\langle \begin{bmatrix} \tilde{s}_{k,t-1} \\ \tilde{s}_{k,t-\lceil T_k \rceil} \\ \tilde{s}_{k,t-\lceil T_k \rceil-1} \end{bmatrix} \begin{bmatrix} \tilde{s}_{k,t-1} \\ \tilde{s}_{k,t-\lceil T_k \rceil} \\ \tilde{s}_{k,t-\lceil T_k \rceil-1} \end{bmatrix}^T \right\rangle_{f(\boldsymbol{x}_t|y_{1:t},\boldsymbol{\theta}^i)} \quad k = 1 \cdots K \tag{3.61}$$

The full iterative algorithm is summarized in algorithm. 2.

Algorithm 2 Iterative EM-Kalman with fixed-lag smoothing

1: Initialize the parameters $\left\{\hat{\boldsymbol{\beta}}_k^0\right\}_{k=1..K}$, $\left\{\hat{\boldsymbol{a}}_k^0\right\}_{k=1..K}$, $\left\{\hat{\rho}_k^0\right\}_{k=1..K}$, $\hat{\sigma_n^2}^0$, $\left\{\hat{T}_k\right\}_{k=1..K}$, and Kalman filter with $\hat{\boldsymbol{x}}_{0|0}$ and $\boldsymbol{P}_{0|0}$.

2: While the stop condition not true and $i < i_{max}$ do

3: Run Kalman filter for $t = 1$ to M to compute $\hat{\boldsymbol{x}}_{t|t}$ and $\boldsymbol{P}_{t|t}$

4: Compute $\boldsymbol{R}_{y_t\boldsymbol{x}_t}^i$ and $\boldsymbol{R}_{\boldsymbol{x}_t\boldsymbol{x}_t}^i$ using (3.54) and (3.55) respectively.

5: Update the parameters $\left\{\hat{\boldsymbol{\beta}}_k^{i+1}\right\}_{k=1..K}$, $\left\{\hat{\boldsymbol{a}}_k^{i+1}\right\}_{k=1..K}$, $\left\{\hat{\rho}_k^{i+1}\right\}_{k=1..K}$ and $\hat{\sigma_n^2}^{i+1}$ using (3.27),(3.30), (3.34) and (3.36) respectively, computed with the elements in (3.56-3.61).

6: If the stop condition is fulfilled or the iteration value $i > i_{max}$, then stop the algorithm, otherwise, back to the Kalman filtering step.

3.3.2 EM-Kalman for bilinear state space model

In terms of cost function, the difference between the linear and bilinear cases is limited to the reduced cost function $Q_k(\boldsymbol{\theta}|\boldsymbol{\theta}^i)$ which becomes only function of the sources samples in the bilinear case

$$Q_k(\boldsymbol{\theta}_k|\boldsymbol{\theta}^i) = \sum_{t=1}^{M} \Big\langle \Big(s_{k,t} + \sum_{l=1}^{p_k} a_{k,l}\ s_{k,t-l} \ + \beta_{k,1} \sum_{l=0}^{p_k} a_{k,l}\ s_{k,t-l-\lceil T_k \rceil+1} + \beta_{k,2} \sum_{l=0}^{p_k} a_{k,l}\ s_{k,t-l-\lceil T_k \rceil} \Big)^2 \Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \tag{3.62}$$

Therefore, the estimate $\hat{\sigma_n^2}^{i+1}$ is identical to the linear case. The rest of parameters are estimated following the same methodology like in the linear case. We introduce some notations that are needed afterwards. We denote by $\boldsymbol{s}_{k,t-j,p_k-l}$ and $\hat{A}^i$ the (p_k+1)-dimensional column vectors and by $\tilde{s}^i_{k,t}$

$$\boldsymbol{s}_{k,t-j,p_k-l} = [s_{k,t-j}, s_{k,t-j-1}, \cdots, s_{k,t-j-p_k+l}]^T \qquad k = 1 \cdots K \tag{3.63}$$

$$\hat{A}^i_k = \begin{bmatrix} 1 & \hat{\boldsymbol{a}}_k^{i\,T} \end{bmatrix}^T \qquad k = 1 \cdots K \tag{3.64}$$

$$\tilde{s}^i_{k,t-j} = \hat{A}^i_k \boldsymbol{s}_{k,t-j,p_k-l} \qquad k = 1 \cdots K \tag{3.65}$$

Suppose the estimates of $\hat{\boldsymbol{\beta}}^i_k$ is available from the previous iteration, then at iteration $(i+1)$, the estimate $\hat{\boldsymbol{\beta}}^{i+1}_k$ is expressed as in (3.27) but with different forms of $\boldsymbol{R}^i_{\tilde{s},k}$ and $\boldsymbol{r}^i_{\tilde{s},k}$ which become in the bilinear case

$$\boldsymbol{R}^i_{\tilde{s},k} = \sum_{t=1}^{M} \Big\langle \begin{bmatrix} \tilde{s}^i_{k,t-\lceil T_k \rceil+1} \\ \tilde{s}^i_{k,t-\lceil T_k \rceil} \end{bmatrix} \begin{bmatrix} \tilde{s}^i_{k,t-\lceil T_k \rceil+1} \\ \tilde{s}^i_{k,t-\lceil T_k \rceil} \end{bmatrix}^T \Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \qquad k = 1 \cdots K \tag{3.66}$$

$$\boldsymbol{r}^i_{\tilde{s},k} = \sum_{t=1}^{M} \Big\langle \begin{bmatrix} \tilde{s}^i_{k,t-\lceil T_k \rceil+1} \\ \tilde{s}^i_{k,t-\lceil T_k \rceil} \end{bmatrix} \tilde{s}^i_{k,t} \Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \qquad k = 1 \cdots K \tag{3.67}$$

The expressions of $\boldsymbol{R}^i_{\tilde{s},k}$ and $\boldsymbol{r}^i_{\tilde{s},k}$ are equivalent to the ones used in the linear case with the difference that in the linear case, the samples of the STP error are directly available in the state vector while in the bilinear case we have to produce them using the STP filter estimate $\hat{\boldsymbol{a}}^i_k$. This latter is also estimated by optimization of (3.62)

which results in the following expression

$$\hat{\boldsymbol{a}_k}^{i+1} = -\Big(\sum_{t=1}^{M}\Big\langle\Big(\boldsymbol{s}_{k,t-1,p_k-1} + \begin{bmatrix}\boldsymbol{s}^T_{k,t-\lceil T_k\rceil,p_k-1}\\ \boldsymbol{s}^T_{k,t-\lceil T_k\rceil-1,p_k-1}\end{bmatrix}^T \hat{\boldsymbol{\beta}}_k^{i+1}\Big)\Big(\boldsymbol{s}_{k,t-1,p_k-1} + \begin{bmatrix}\boldsymbol{s}^T_{k,t-\lceil T_k\rceil,p_k-1}\\ \boldsymbol{s}^T_{k,t-\lceil T_k\rceil-1,p_k-1}\end{bmatrix}^T \hat{\boldsymbol{\beta}}_k^{i+1}\Big)^T\Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)}\Big)^{-1}\sum_{t=1}^{M}\Big\langle\Big(\boldsymbol{s}_{k,t-1,p_k-1} + \begin{bmatrix}\boldsymbol{s}^T_{k,t-\lceil T_k\rceil,p_k-1}\\ \boldsymbol{s}^T_{k,t-\lceil T_k\rceil-1,p_k-1}\end{bmatrix}^T \hat{\boldsymbol{\beta}}_k^{i+1}\Big)\Big(s_{k,t} + \begin{bmatrix}s_{k,t-\lceil T_k\rceil+1}\\ s_{k,t-\lceil T_k\rceil}\end{bmatrix}^T \hat{\boldsymbol{\beta}}_k^{i+1}\Big)\Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.68}$$

Finally, the estimate of $\hat{\rho}_k^{i+1}$ is deduced in an equivalent fashion as in (3.34) with the difference that the STP error samples and its correlations are not directly available in the state vector estimate and its covariance, then the STP filter estimate $\hat{\boldsymbol{a}}_k^{i+1}$ is used to afford them.

$$\hat{\rho}_k^{i+1} = \left[\Big(\frac{1}{M}\sum_{t=1}^{M}\Big\langle\begin{bmatrix}\tilde{s}^{i+1}_{k,t}\\ \tilde{s}^{i+1}_{k,t-\lceil T_k\rceil+1}\\ \tilde{s}^{i+1}_{k,t-\lceil T_k\rceil}\end{bmatrix}\begin{bmatrix}\tilde{s}^{i+1}_{k,t}\\ \tilde{s}^{i+1}_{k,t-\lceil T_k\rceil+1}\\ \tilde{s}^{i+1}_{k,t-\lceil T_k\rceil}\end{bmatrix}^T\Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)}\Big)^{-1}\right]_{11}^{-1} \quad k = 1\cdots K \tag{3.69}$$

In the bilinear case, all the statistics that are needed to compute the parameters estimates are available in the $(2p_k+3)\times(2p_k+3)$-dimensional matrix $\breve{\boldsymbol{R}}_k^i$ defined as

$$\breve{\boldsymbol{R}}^i_{k,bilin} = \sum_{t=1}^{M}\Big\langle\breve{\boldsymbol{x}}^{bilin}_{k,t}\breve{\boldsymbol{x}}^{bilin\,T}_{k,t}\Big\rangle_{f(\boldsymbol{x}_t|y_{1:M},\boldsymbol{\theta}^i)} \quad k = 1\cdots K \tag{3.70}$$

where $\breve{\boldsymbol{x}}^{bilin}_{k,t}$ is the $(2p_k+3)$-dimensional partial state vector

$$\breve{\boldsymbol{x}}^{bilin}_{k,t} = [\boldsymbol{s}^T_{k,t,p_k}, \boldsymbol{s}^T_{k,t\lceil T_k\rceil+1,p_k}, s_{k,t-\lceil T_k\rceil-p}]^T \quad k = 1\cdots K \tag{3.71}$$

$\breve{\boldsymbol{R}}^i_{k,bilin}$ (and $\breve{\boldsymbol{x}}^{bilin}_{k,t}$) can be selected from $\boldsymbol{R}^i_{\boldsymbol{x}_t\boldsymbol{x}_t}$ (and $\boldsymbol{x}_{k,t}$) of the bilinear state space model using an appropriate selection matrix $\boldsymbol{S}_{k,bilin}$.

Unlike the linear model where the prediction error is singular, in the bilinear model, $\boldsymbol{P}_{t|t-1}$ is full-rank. In fact, the sub-state matrix in (3.13) is only rank-deficient by 1 (last column of all zeros) which makes $\boldsymbol{F}$ rank-deficient by K. This deficiency is rectified by the additive term $\boldsymbol{GQG}^T$ of rank K. Therefore, the inverse of $\boldsymbol{P}_{t|t-1}$ exists and the computation of $\boldsymbol{K}_{rts,t}$ becomes possible. A fixed-lag smoothing version in the bilinear model is also possible since the delayed partial state $\breve{\boldsymbol{x}}^{bilin}_{k,t-1}$ can be extracted from the bilinear sub-state vector $\boldsymbol{x}_{k,t}$. The initialization and the stop condition of the algorithm in the linear model are also valid in the bilinear model.

3.3.3 Multipitch estimation

EM-Kalman permits to estimate all the components of $\boldsymbol{\theta}$ except to the pitch periods $\{T_k\}_{k=1..K}$. It is because that the cost function (3.22) depends "'implicitly"' on them. It is even important to know their values in advance so that the state matrix $\boldsymbol{F}$ can be constructed accurately. Therefore, their estimations is achieved independently.

Multipitch estimation is an issue in itself that was studied exhaustively in literature. In fact, it plays an important role in several audio applications such as speech enhancement [Jansson & Stoica 1999], music transcription [Cemgil 2004], audio compression [Christensen & Holdt Jensen 2006, Lindblom 2005, Rodbro *et al.* 2003] and audio source separation [Chazan *et al.* 1993, Kostek 2004]. Many algorithms were proposed[Gold & Rabiner 1969, Hess 2008, Hess 1983, Christensen *et al.* 2009]. In many times, the proposed approaches consist on a generalization of a monopitch estimation algorithm to the multipitch case. They can be classified mainly in two categories, parametric and non-parametric approaches. In the first category, a parametric model is assumed for speech signal and based on it, the pitch is estimated. The most famous parametric model is the sinusoidal model [Christensen *et al.* 2009]. In the parametric category, there are three classes of methods. The first class is the statistical methods where pitches are estimated using ML, maximum a posteriori (MAP) and EM estimators [sbø ll Christensen *et al.* 2008]. These methods are efficient asymptotically, thus, need many samples to converge. The second class is the filtering methods where observations are filtered by picking samples positioned at harmonics, then pitches are estimated by maximizing the filter's output power. An example of used filters is Comb filtering [Nehorai & Porat 1986] and the optimal filter designs in [Christensen & Jakobsson 2010, Christensen *et al.* 2008]. The last class is the subspace methods where famous methods such as MUSIC and ESPRIT are used [Christensen *et al.* 2006, Roy & Kailath 1989]. In the second non-parametric category, there are algorithms based on the auto-correlation function maximization[Rabiner 1977], the averaged magnitude difference function minimization[Ross *et al.* 1974] and the averaged squared difference function minimization[Noll 1967, Abeysekera 2004].

In our work, it is very important to have a good (even a very good) estimate of the pitches since they present a strong feature of separability between sources. In simulations, we assume we have the separate sources and use them to estimate the pitches using a robust monopitch estimator. This assumption is not totally unrealistic since in a realistic speech sequence, the present speakers are not talking simultaneously all the time. Thus, it is possible to extract sequences where only one speaker is talking and use them to estimate the corresponding pitch. We use

the robust algorithm for pitch tracking (RAPT) [Talkin 1995] which is a monopitch estimator based on the maximization of the normalized cross-correlation function (NCCF) of the given source. The data is processed in overlapped frames where for a given frame i, the NCCF of source k at lag l, $\Phi_{i,l}^k$, is expressed as it follows

$$\Phi_{i,l}^k = \frac{\sum_{j=im}^{im+N_f-1} s_{k,j} s_{k,j+l}}{\sqrt{\sum_{j=im}^{im+N_f-1} s_{k,j}^2 \sum_{j=im+l}^{im+l+N_f-1} s_{k,j}^2}} \tag{3.72}$$

where N_f and m are respectively the processing frame size and the hop size (time jump) from one frame to the next (in samples). $\Phi_{i,l}^k$ is maximized (tends to be close to 1) when the lag l coincides with multiples of the target pitch period.

3.4 Simulations

The simulation part is carried in two steps. First, we apply the proposed algorithms to artificial signals generated according to the joint model in (3.5)-(3.6). Second, the algorithms are applied to real signals. The number of speakers is assumed known and fixed to $K = 2$. The STP orders for both sources are fixed to $p_1 = p_2 = 10$. The sampling frequency F_s is fixed to $8KHz$.

A very important issue in EM is initialization. When the cost function is multimodal, it depends on the initial value of parameters that the algorithm will converge to a global or local maximum. In the case when the cost function is monomodal, initial value will influence the number of iterations needed to reach the global maximum. In our simulations, we aim to analyze the behavior of the algorithms close to and after convergence. Therefore, we initialize the sources parameters $\{\boldsymbol{a}_k, \boldsymbol{\beta}_k, \rho_k\}_{k=1:K}$ to values not too far from the "'true"' values. We intend by "'true"', the values used to produce the sources in simulations with synthetic signals. When the signals are real, the "'true"' values refer to the one estimated directly from the clean sources which is once again not completely unrealistic if we can afford sequences where only one source is active. For Kalman filter, the initial values of the state vector and its error covariance matrix are set to $\hat{\boldsymbol{x}}_{0|0} = \boldsymbol{0}$ and $\boldsymbol{P}_{0|0} = \text{diag}\,(\rho_1, \cdots, \rho_1, \rho_2, \cdots, \rho_2)$ respectively.

3.4.1 Synthetic signals

In order to synthesize meaningful signals (close to the reality), real voiced speech signals are used to produce the STP coefficients which are set as it follows

$$\boldsymbol{a}_1 = \big[-1.5020, 1.7380, -2.0290, 1.7890, -1.3760, 1.2550, -0.6930, 0.3760, \\ -0.0800, 0.0330\big]^T$$

and

$$\boldsymbol{a}_2 = \big[0.2153, 0.2153, -0.0176, 0.0806, 0.0127, 0.1569, -0.0218, 0.2146, \\ -0.7225, -0.3034\big]^T$$

The pitch periods are set to $T_1 = 7.2ms$ and $T_2 = 4ms$ which results in fundamental frequencies $F_1 = 139.86Hz$ and $F_2 = 250Hz$. Converted in samples, T_1 and T_2 results in fractional pitch $T_1 = 57.2$ and integer pitch $T_2 = 32$ respectively which permits to simulate the two models of the STP error in (3.3) and (3.2). In order to emphasize the harmonic feature, the LTP coefficients is fixed close to 1 for the integer pitch period case ($\beta_{2,1} = 0.99$). For the fractional pitch period of speaker 1, the two LTP coefficients $\beta_{1,1}$ and $\beta_{1,2}$ are interpreted as the real LTP coefficient of the fractional pitch that we denote b_1, splitted between $\lceil T_1 \rceil + 1$ and $\lceil T_1 \rceil$ according to the contribution of each one in T_1 which is deduced from the fraction value. Then, the coefficients $\beta_{1,1}$ and $\beta_{1,2}$ are respectively computed as $\alpha_1 b_1$ and $(1-\alpha_1)b_1$ where $\alpha_1 = T_1 - \lceil T_1 \rceil$. Notice that using this interpretation of $\beta_{1,1}$ and $\beta_{1,2}$ respects the stability condition announced before ($\beta_{1,1}+\beta_{1,2} = b_1 < 1$). In order to highlight the harmonic feature again, the real LTP coefficient is set to $b_1 = 0.99$ which results in $\beta_{1,1} = 0.792$ and $\beta_{1,2} = 0.198$. The powers are fixed so that the signal to signal ratio (SSR) is close to $0dB$. Thus, the innovative powers are set to $\rho_1 = 1$ and $\rho_2 = 2$. The speech sequence length is set to 384 ms. The mixture is created artificially by adding the two sources in addition to the additive white Gaussian observation noise. The produced sources and their mixture are plotted in time in figure Fig. 3.3 and their corresponding spectra in figure Fig. 3.4 for an input source to noise ratio (SNR) of 20 dB. The produced sources are infinitely stationary and voiced sounds.

In this section, simulations consist of two parts. In the first part, we study and compare the performance of the proposed algorithms in terms of linear MMSE (LMMSE) estimation quality. In this case, the parameters are supposed known beforehand, only the sources are estimated. This simulation gives an idea about the best performance the algorithms can achieve (an upper bound). The input SNR is varied in the following interval $[-25\ dB, 30\ dB]$. Performances are assessed using the objective criteria explained in 2.5.6. The obtained results are presented in

figures Fig. 3.5,Fig. 3.6 and Fig. 3.7. In the figures, the LMMSE estimation using the linear state space model is denoted by **filt_lin**, the ones using the bilinear state space model with and without RTS smoothing are denoted **filt_bilin** and **filt_bilin_RTS** respectively. In this scenario limited to source estimation, we expect that both the linear and bilinear models will have the same performance. The **filt_bilin_RTS** will produce better results due to the supplementary step of fixed-interval smoothing. These expectations are confirmed with the results obtained in Fig. 3.5,Fig. 3.6 and Fig. 3.7. Thus, the only benefit to use the bilinear model is that it allows the possibility to perform the RTS smoothing which offers an improvement that reaches to 5 dB at high SNR in terms of source-to-interferences ratio (SIR).

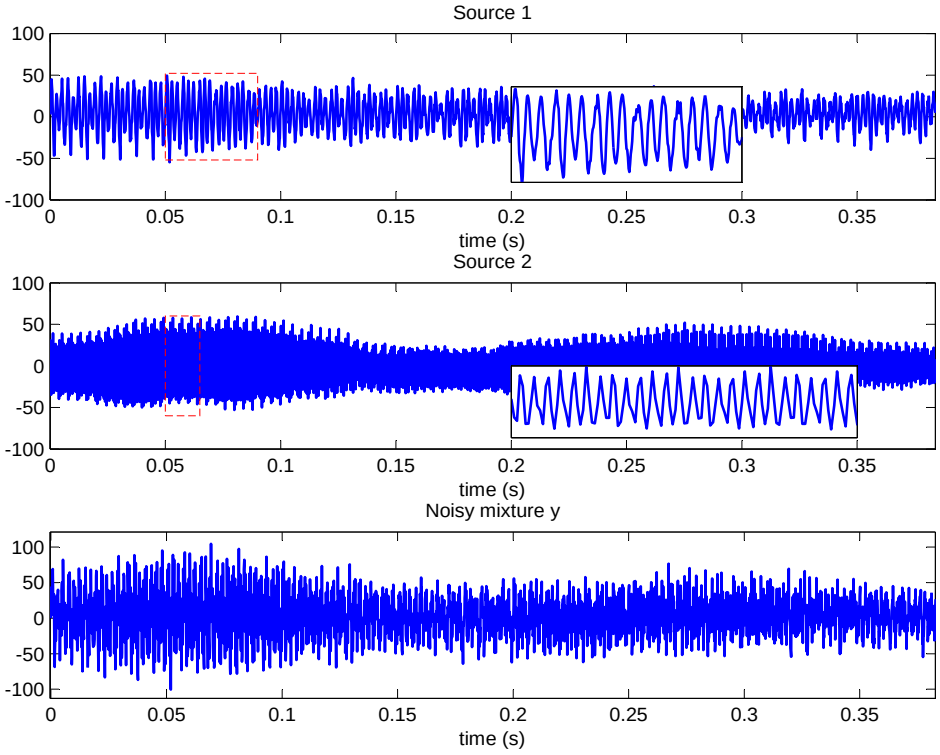

Figure 3.3: *Time evolution of sources 1 and 2, and observations with zoom on the sources.*

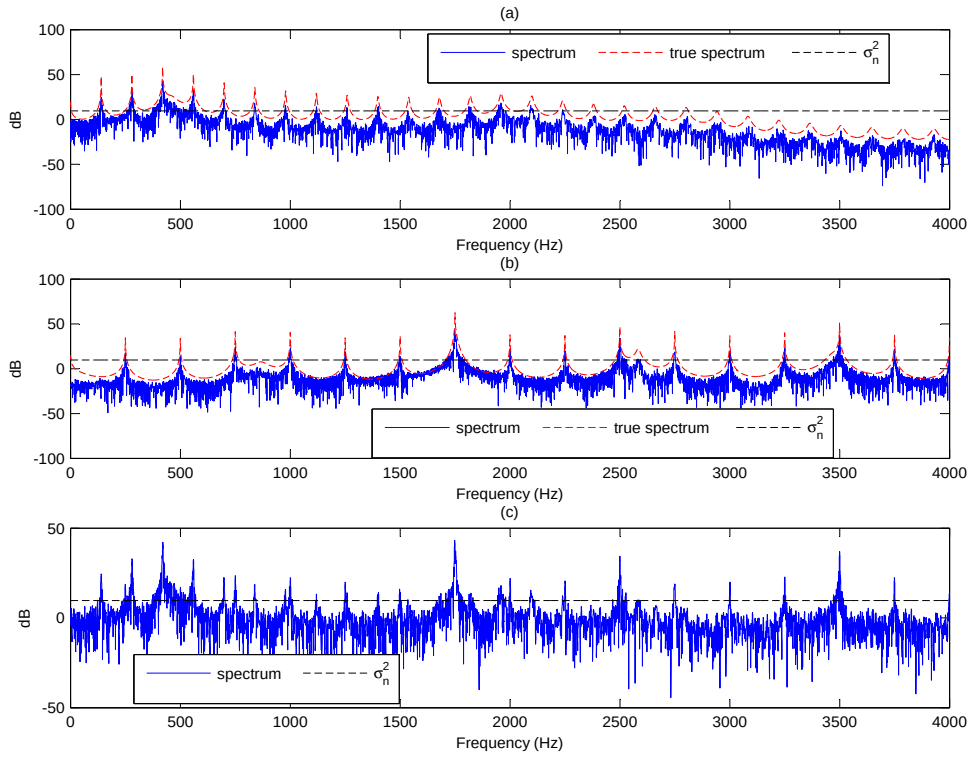

Figure 3.4: *spectrum and true spectrum of sources 1 (a) and 2 (b), and observation spectrum (c).*

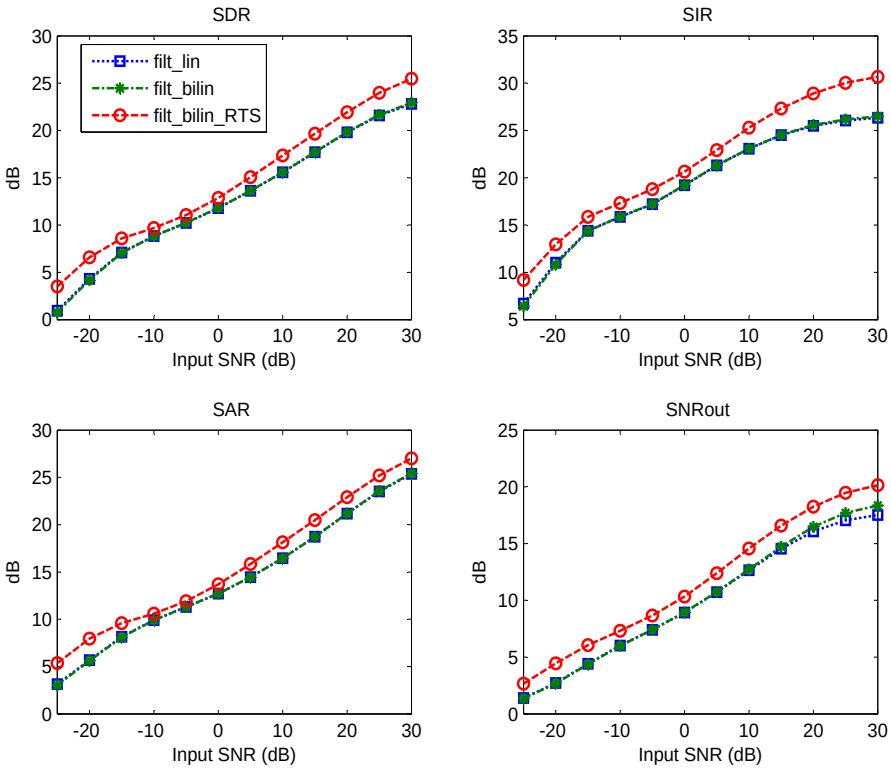

Figure 3.5: *Performance criteria of Kalman filtering for source 1.*

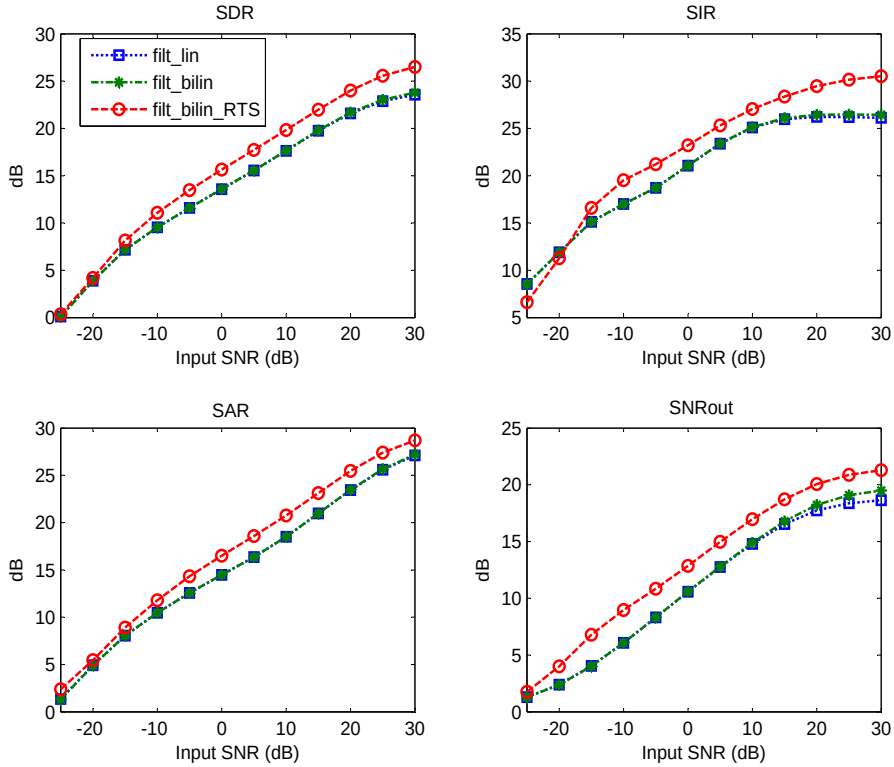

Figure 3.6: *Performance criteria of Kalman filtering for source 2.*

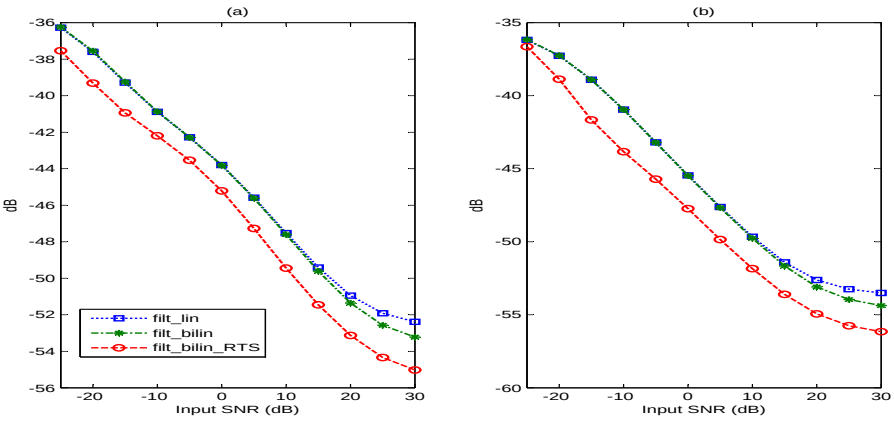

Figure 3.7: *MSE of Kalman filtering for sources 1 (a) and 2 (b).*

In the second part, the parameters are assumed unknown and estimated along with the sources. Since the synthetic signals are purely voiced sounds produced

by constant pitch periods. these latter are perfectly estimated by the used mono-pitch estimator. The obtained results are presented in figures Fig. 3.8,Fig. 3.9 and Fig. 3.10. EM-Kalman using the linear state space model is denoted by **estim_lin**, the ones using the bilinear state space model with and without RTS smoothing are denoted **estim_bilin** and **estim_bilin_RTS** respectively. Compared to the results in the first part, the behavior of the three algorithms does not change a lot except that at high SNR the performance of **estim_bilin_RTS** in terms of mean square error (MSE) and output SNR becomes closer to the two other algorithms.

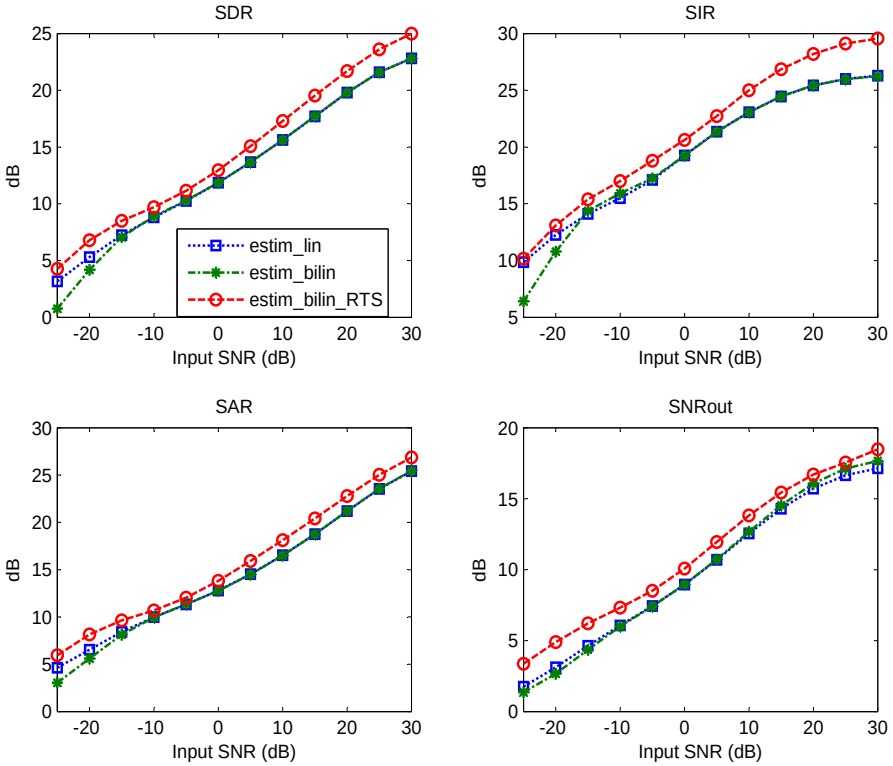

Figure 3.8: *Performance criteria of EM-Kalman estimation for source 1.*

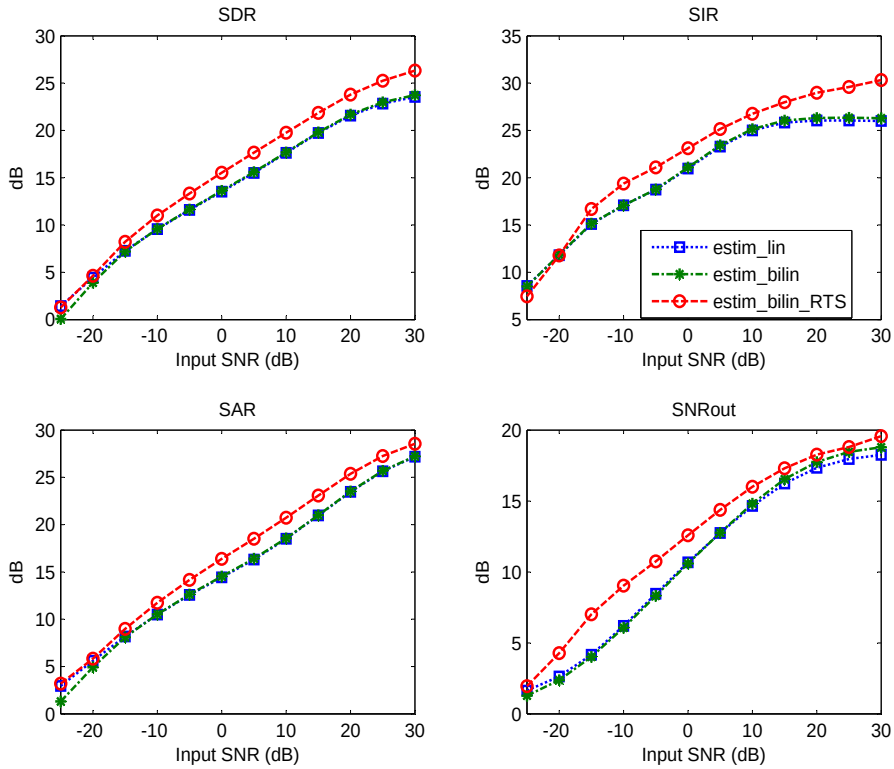

Figure 3.9: *Performance criteria of EM-Kalman estimation for source 2.*

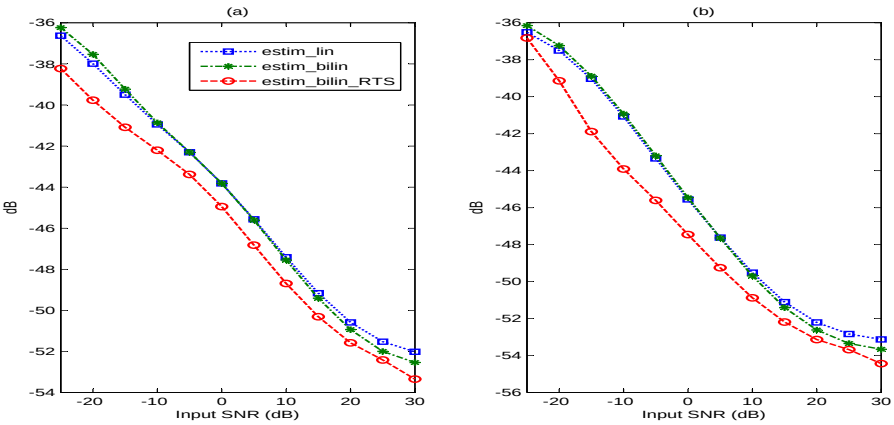

Figure 3.10: *MSE of EM-Kalman estimation for sources 1 (a) and 2 (b).*

3.4.2 Real signals

In this section, we apply the algorithms to real signals selected from the GRID database [Cooke *et al.* 2006] used in the making of the 1^{st} Speech Separation Challenge. It consists of a large audiovisual sentence corpus of 1000 sentences spoken by each of 34 talkers (18 males, 16 females). We select the utterance "'*lwax8s*"' of the male speaker 14 and the utterance "'*bgwf7n*"' of the female speaker 22. The selected speakers were downsampled from 25 kHz to 8 kHz. We apply the algorithms on a strongly voiced sequence where the pitch periods fluctuations are very small for both speakers. The observation noise is added artificially.The selected voiced sequence (of length 120.1 ms) is highlighted in red in figure Fig. 3.11 where the SNR is of 14.7 dB and the SSR is of 3.4 dB.

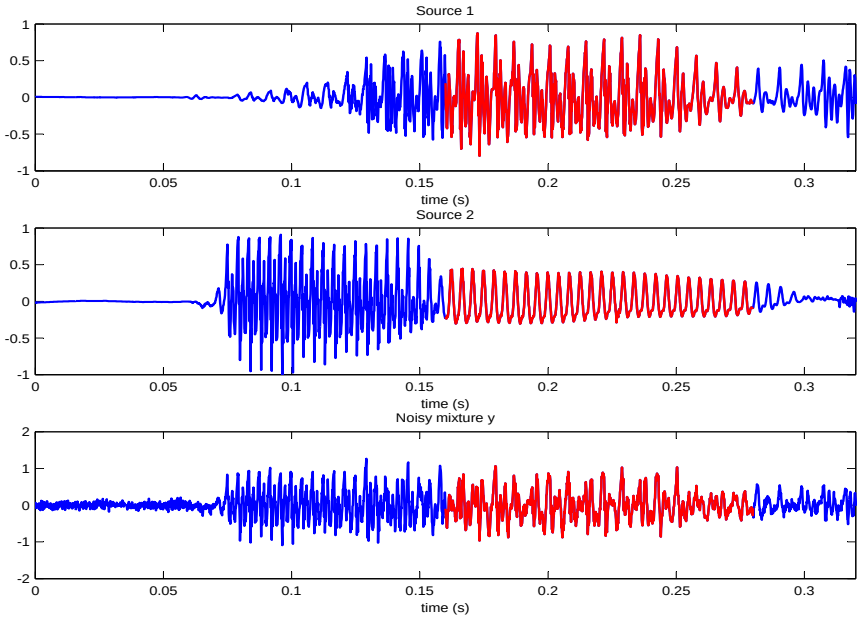

Figure 3.11: *Real speech sequence*

In table 3.4.2, we display the corresponding estimation results (in dB). We notice that, in terms of MSE and output SNR, **estim_lin** performs slightly better than **estim_bilin** and close to **estim_bilin_RTS**. In terms of artifacts noise, the three algorithm are equivalent. In terms of source-to-distortion ratio (SDR) and SIR, **estim_lin** and **estim_bilin_RTS** extract both sources with similar qualities, unlike **estim_bilin** that privileges the dominant (in terms of SSR) source. In global, **estim_bilin_RTS** presents the best performance, but it is also the most expensive in terms of computation complexity which may privileges **estim_lin** in

some applications.

source 1	SDR	SIR	SAR	SNRout	MSE
estim_lin	12.0403	12.8418	21.6209	5.4016	-33.4702
estim_bilin	13.3886	14.4546	20.1642	3.5819	-31.6505
estim_bilin_RTS	13.6264	14.6589	20.5177	5.8391	-33.9077
source 2	SDR	SIR	SAR	SNRout	MSE
estim_lin	10.4169	13.0641	24.8942	7.1179	-35.1865
estim_bilin	7.3035	7.4124	24.0916	4.1803	-32.2489
estim_bilin_RTS	14.2068	14.6905	24.1230	9.1890	-37.2576

Table 3.1: *Results of EM-Kalman for real signals estimation*

3.5 Conclusion

In this chapter we use the iterative EM-Kalman algorithm for speech source separation problem. The model takes into account short and long-term correlations of speech signals by using the famous speech joint model. Three algorithms are derived using two state space models formulated from the joint speech model. The traditional 1-lag smoothing step is included into the algorithms where only Kalman filtering is used., and it is no more an additional step. Simulations show the equivalence of the linear model and the bilinear one without RTS smoothing in synthetic signals. In simulation with real signals, the linear model base algorithm shows better performance in handling the case where the SSR is not close to 0 dB along with the bilinear model based algorithm with RTS smoothing. Yet, this performance depends a lot on the multipitch estimation quality. An error on tracking the pitches may induce the performance decreasing drastically.

Chapter 4

Monaural speech separation exploiting periodicity and spectral envelopes

4.1 Introduction

Audio signal quasi-periodicity and spectral information have been widely exploited to perform speech enhancement. In fact, in [Nehorai & Porat 1986], Nehorai et al. propose a sinusoidal model based algorithm for enhancement of speech corrupted by additive white Gaussian noise. The enhancement is achieved by estimating the sinusoidal model parameters which are the fundamental frequency, amplitudes and phases. The fundamental frequency (nonlinear parameter) is estimated using the recursive prediction error adaptive comb filter; amplitudes and phases are estimated using the recursive least squares (RLS) algorithm. In [Jensen & Hansen 2001], the sinusoidal model, corrupted by additive broadband noise, is used with smoothness constraints imposed on the model parameters. The smoothness condition is induced by the continuous and slow variations with time of the vocal tract transfer function and the pitch. Therefore, this algorithm is restricted only to the voiced speech, while in [Jensen *et al.* 2012], a more general algorithm is proposed, using two filters jointly, one for enhancing voiced speech exploiting its harmonic feature, another for unvoiced speech.

In audio source separation, periodicity has been used exhaustively [Stettiner *et al.* 1993, Chazan *et al.* 1993, Virtanen & Klapuri 2000, Mowlaee *et al.* 2010a, Mowlaee *et al.* 2010b, Christensen & Jakobsson 2010, Jensen *et al.* 2010, Mowlaee *et al.* 2011b]. Specifically, in [Chazan *et al.* 1993, Stettiner *et al.* 1993], the authors consider a multipitch model for voiced speech (referred to also as the long-term model) and introduce a time-warping function which describes pitch variation with time. The separation is achieved by identifying this function and estimating the maximum likelihood (ML) solution of the other usual parameters (amplitudes, phases, etc..). In [Giacobello *et al.* 2009b], the short term prediction (STP) and long term prediction (LTP) aspects of speech are jointly modeled.

In [Bensaid *et al.* 2010b, Schutz & Slock 2010], a joint autoregressive (AR) model (STP + LTP) was introduced for quasiperiodic sources. The LTP part allows to capture the quasiperiodicity (with possible imperfect correlation in time), while the STP part allows to model the spectral envelope (see section 3.2). The modeling of the power spectral density is important to allow power splitting between sources at overlapping harmonics in the source extraction operation. In [Bensaid *et al.* 2010b, Schutz & Slock 2010] Bayesian approaches were adopted for source and parameter estimation, using expectation maximization (EM)-Kalman and variational Bayes (VB) techniques respectively. In [Schutz & Slock 2011], the joint AR model was used for mono-microphone source separation in the frequency domain. Using Gaussian source models, the source extraction is simply the linear MMSE (LMMSE) (Wiener) estimation. In the parametric approach, the joint model parameters need to be estimated also.

In [Schutz & Slock 2011], three criteria are formulated for the estimation of these parameters on the basis of one frame of data, the Itakura-Saito distance (ISD) and optimally weighted spectrum matrix (OWSM) for matching the parametric spectrum and the observations periodogram. The third criterion is the Gaussian maximum likelihood (GML) [de Carvalho & Slock 1999]. The gradients of these three criteria w.r.t. the AR parameters and hence their extrema are shown to be identical. The results in [Schutz & Slock 2011] are based on asymptotic frequency domain expressions that are only valid for extremely long frames. In this chapter, we extend these results by accounting for the finite window length and by introducing advantageously a non-rectangular window. Nontrivial windows were also introduced in [Schutz & Slock 2010], for the different purposes of source extraction and parameter estimation, passing from time to frequency domain. The approach in [Schutz & Slock 2010] was based on Variational Bayes, in which sources and their parameters are estimated jointly in an alternating optimization fashion. Here we estimate the parameters separately from the sources after elimination of the Gaussian sources from the likelihood function as in [Schutz & Slock 2011]. Due to the introduction of the window, which already limits temporal correlation, we propose to replace the LTP coefficient by its maximum value 1. We reconsider the equivalence of the three criteria mentioned beforehand, but this time based on finite data vectors, for which in frequency domain we can no longer neglect the correlations between different frequencies. The goal of the window design will then be to limit these correlations. The equivalence of multivariate ISD and GML is straightforward [Carlson & Clements 1991a] as we shall see. In the multivariate case, the OWSM results in optimally weighted covariance matrix (OWCM)[Ottersten *et al.* 1998]. The OWCM is again shown to be equivalent to ISD and GML in terms of gradients. After proving equivalence, we will focus on the GML criterium and use it to derive

the expressions of the unknown parameters (STP coefficients, innovations powers and observation noise power).

4.2 Windowing for frame-based processing

As mentioned before, audio signals are by nature non-stationary in global, but they can be considered hence piecewise. That is why they are processed in frames (during 30 ms), which corresponds to time-invariant filtering. Many of the signal processing operations such as linear time-invariant filtering and filter computation could be largely simplified by passing to the frequency domain. However, transforming a frame of signal to the frequency domain directly via the discrete Fourier transform (DFT) (FFT) leads to approximations due to the periodic extension of the frame assumption inherent in the DFT. We shall see later how we can improve these approximations.

Just like the original data signal y_n (n here is the time index) will be cut into a series of windowed frames of length N, a bit like in the Welch method, a processed signal (here the extracted source) will be reconstructed by superposing its reconstructed windowed frame segments. Since the window needs to decay towards its edges, consecutive frames need to overlap. Let M be the hop size from one frame to the next, then a perfect reconstruction (PR) window w_n requires [Kabal 2005]

$$\sum_{i=-\infty}^{\infty} w_{n-iM} \;=\; 1\;,\quad \forall n \tag{4.1}$$

see the figures Fig. 4.1 for the cases of relative overlap of $\frac{(N-M)}{N} = 50\%, 75\%$, both the individual windows (Hann windows) and their sum are shown for a finite set of windows. Note that one could consider extensions to non-PR windows, in which the superposition of windowed signal frames could be followed by a zero-forcing re-scaling with $1/(\sum_{i=-\infty}^{\infty} w_{n-iM})$ or (multi-window) minimum mean square error (MMSE) versions thereof. An example of a PR window is a Hann (or raised cosine) window

$$w_n \;=\; \frac{1}{2}\left[1 - cos\left(2\pi\frac{n}{N}\right)\right]\;,\quad n = 0, 1, \ldots, N-1 \tag{4.2}$$

The continuity of the window at its edges can be expected to be reflected in the continuity of the reconstructed signal and helps reduce blocking artifacts (musical noise). The motivations for the window design will be different however in the parameter estimation part as we shall see. In a separate approach for parameter estimation and source extraction, as considered here, different windows could be used in both parts.

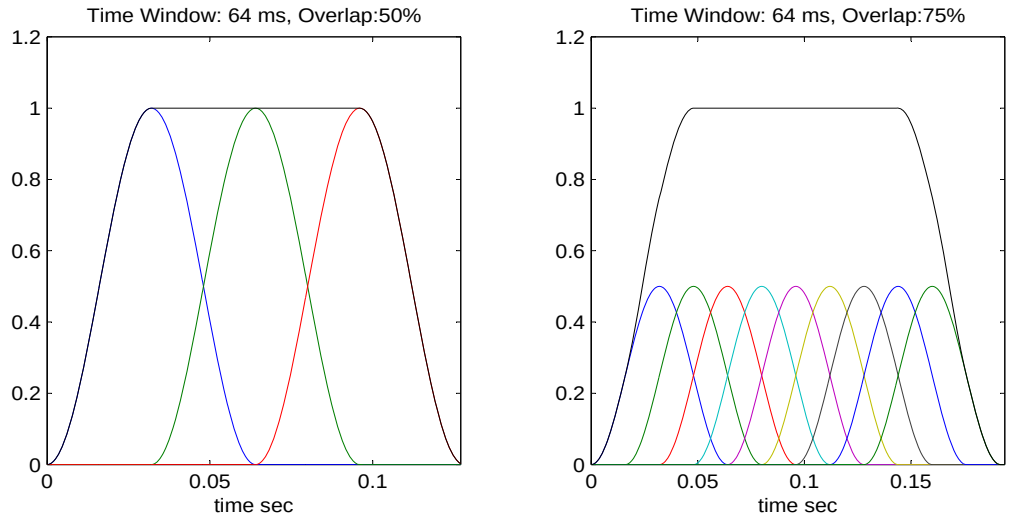

Figure 4.1: *Perfect reconstruction window*

4.3 Equivalence of ISD, GML and OWCM criteria

In what follows we consider a vector of zero mean data $\boldsymbol{y}$ of length N, with covariance matrix $\boldsymbol{R}$, and estimation on the basis of the simple sample covariance $\widehat{\boldsymbol{R}} = \boldsymbol{y}\boldsymbol{y}^H$. We consider the data $\boldsymbol{y}$ to be circular complex Gaussian distributed. Suppose that the covariance matrix $\boldsymbol{R}$ is parameterized by the vector $\boldsymbol{\theta}$: $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\theta})$. The symbols $.^*, .^T, .^H$ denote complex conjugate, transpose and Hermitian transpose respectively.

4.3.1 Itakura-Saito distance

The multivariate ISD is based on the observation that for a nonnegative definite matrix $\boldsymbol{A}$, the tangent hyperplane to $\ln|\boldsymbol{A}|$ at $\boldsymbol{A} = \boldsymbol{I}_N$ is $\mathrm{tr}\{\boldsymbol{A} - \boldsymbol{I}_N\}$, where $\boldsymbol{I}_N$ is the identity matrix of size N. The concavity of $\ln|.|$ then leads to

$$\mathrm{tr}\{\boldsymbol{A} - \boldsymbol{I}_N\} - \ln|\boldsymbol{A}| \geq 0 \tag{4.3}$$

The ISD is a common tool to measure spectral distortion [Carlson & Clements 1991b, Itakura 1975, F.Itakura & S.Saito 1970] and is successfully used as a cost function for non-negative matrix factorization (NMF) based audio processing algorithms [Lefèvre *et al.* 2011, Févotte 2011, Févotte *et al.* 2009]. In our problem, the ISD is used to compare the covariance matrix $\boldsymbol{R}$ and its estimate $\widehat{\boldsymbol{R}}$, and obtained by taking their ratio $\boldsymbol{A} = \widehat{\boldsymbol{R}}\,\boldsymbol{R}^{-1}(\boldsymbol{\theta})$:

$$ISD(\boldsymbol{\theta}) = \mathrm{tr}\{\widehat{\boldsymbol{R}}\,\boldsymbol{R}^{-1}(\boldsymbol{\theta}) - \boldsymbol{I}_N\} - \ln\left|\widehat{\boldsymbol{R}}\,\boldsymbol{R}^{-1}(\boldsymbol{\theta})\right| \tag{4.4}$$

4.3.2 Gaussian maximum likelihood

Assuming a circular complex Gaussian distribution and apart from constants (we mean by constant here independent of $\boldsymbol{\theta}$ like $\ln\left|\widehat{\boldsymbol{R}}\right|$), the negative loglikelihood becomes

$$GML(\boldsymbol{\theta}) = \ln|\boldsymbol{R}(\boldsymbol{\theta})| + \boldsymbol{y}^H\boldsymbol{R}^{-1}(\boldsymbol{\theta})\boldsymbol{y} \tag{4.5}$$

Now note that using a property of the trace operator (see Appendix B)

$$\begin{aligned}\boldsymbol{y}^H\boldsymbol{R}^{-1}(\boldsymbol{\theta})\boldsymbol{y} &= \mathrm{tr}\{\boldsymbol{y}^H\boldsymbol{R}^{-1}(\boldsymbol{\theta})\boldsymbol{y}\}\\ &= \mathrm{tr}\{\boldsymbol{y}\boldsymbol{y}^H\boldsymbol{R}^{-1}(\boldsymbol{\theta})\}\\ &= \mathrm{tr}\{\widehat{\boldsymbol{R}}\,\boldsymbol{R}^{-1}(\boldsymbol{\theta})\}\end{aligned} \tag{4.6}$$

On the other hand, $\ln\left|\widehat{\boldsymbol{R}}\,\boldsymbol{R}^{-1}(\boldsymbol{\theta})\right| = \ln\left|\widehat{\boldsymbol{R}}\right| - \ln|\boldsymbol{R}(\boldsymbol{\theta})|$. Hence, the ISD and GML criteria are identical (in their dependence on $\boldsymbol{\theta}$). Note that the GML criterion only has an estimation motivation, whereas the ISD (and hence GML also) performs jointly approximation and estimation. The approximation part refers to the fact that the true covariance matrix of $\boldsymbol{y}$ may not be of the form $\boldsymbol{R}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$, in which case minimizing the ISD will lead to a $\boldsymbol{\theta}$ that best approximates the data.

4.3.3 Optimally weighted covariance matching

The OWCM is in fact the optimally weighted least-squares applied to a sample covariance. Consider the $vec(.)$ operator which stacks the consecutive columns of a matrix into a vector. Then

$$vec(\widehat{\boldsymbol{R}}) = vec(\boldsymbol{y}\boldsymbol{y}^H) \tag{4.7}$$

$$= \boldsymbol{y}^* \otimes \boldsymbol{y} \tag{4.8}$$

Where $\otimes$ denotes the Kronecker product. The mean of $\boldsymbol{y}^* \otimes \boldsymbol{y}$ is of course $vec(\widehat{\boldsymbol{R}})$. Using expressions for fourth moments of complex Gaussian signal, we get for its covariance matrix $\boldsymbol{R}^* \otimes \boldsymbol{R}(\boldsymbol{\theta})$. After some algebraic computations using rules in Appendix B, the OWCM criterion is then

$$\begin{aligned}OWCM(\boldsymbol{\theta}) &= (\boldsymbol{y}^* \otimes \boldsymbol{y} - vec(\boldsymbol{R}(\boldsymbol{\theta})))^H(\boldsymbol{R}^*(\boldsymbol{\theta}) \otimes \boldsymbol{R}(\boldsymbol{\theta}))^{-1}(\boldsymbol{y}^* \otimes \boldsymbol{y} - vec(\boldsymbol{R}(\boldsymbol{\theta})))\\ &= \mathrm{tr}\{(\widehat{\boldsymbol{R}} - \boldsymbol{R}(\boldsymbol{\theta}))\boldsymbol{R}^{-1}(\boldsymbol{\theta})(\widehat{\boldsymbol{R}} - \boldsymbol{R}(\boldsymbol{\theta}))\boldsymbol{R}^{-1}(\boldsymbol{\theta})\}.\end{aligned} \tag{4.9}$$

Now, it is well-known that the weighting matrices $\boldsymbol{R}^{-1}(\boldsymbol{\theta})$ can be replaced by consistent estimates without modifying the asymptotic covariance matrix of the estimation errors resulting from minimizing the OWCM criterion. Once the $\boldsymbol{R}^{-1}(\boldsymbol{\theta})$ are replaced by a consistent estimate, they are no longer a function of $\boldsymbol{\theta}$. Now, taking the gradient of OWCM w.r.t. a parameter θ_i by only considering the $\boldsymbol{R}(\boldsymbol{\theta})$

appearing in the quadratic "numerator" and using some mathematical rules (see Appendix B) we get

$$\frac{\partial\, OWCM(\boldsymbol{\theta})}{\partial\theta_i} = -2\mathrm{tr}\{\frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{R}^{-1}(\boldsymbol{\theta})(\widehat{\boldsymbol{R}} - \boldsymbol{R}(\boldsymbol{\theta}))\boldsymbol{R}^{-1}(\boldsymbol{\theta})\} \tag{4.10}$$

On the other hand we get for $GML(\boldsymbol{\theta}) = \ln|\boldsymbol{R}(\boldsymbol{\theta})| + \mathrm{tr}\{\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}(\boldsymbol{\theta})\}$ that

$$\begin{aligned}\frac{\partial\, GML(\boldsymbol{\theta})}{\partial\theta_i} &= \mathrm{tr}\{\boldsymbol{R}^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial\theta_i}\} - \mathrm{tr}\{\widehat{\boldsymbol{R}}\,\boldsymbol{R}^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{R}^{-1}(\boldsymbol{\theta})\} \\ &= -\mathrm{tr}\{\frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{R}^{-1}(\boldsymbol{\theta})(\widehat{\boldsymbol{R}} - \boldsymbol{R}(\boldsymbol{\theta}))\boldsymbol{R}^{-1}(\boldsymbol{\theta})\} \end{aligned} \tag{4.11}$$

Comparing (4.10) and (4.11), we see that the extrema of $OWCM(\boldsymbol{\theta})$ and $GML(\boldsymbol{\theta})$ coincide. After establishing the equivalence between the three criteria, we will work with the GML in the next section.

4.4 GML applied to the data DFT

Working in the time domain, we have a full covariance $\boldsymbol{R}$ to work with. By going to the frequency domain, one typically assumes to be able to work with a diagonal $\boldsymbol{R}$ because asymptotically, different frequency components are uncorrelated. We shall analyze more precisely the nonasymptotic regime. Now, let the current frame of N samples be $\boldsymbol{y} = [y_0\, y_1 \cdots y_{N-1}]^T$ and without loss of generality we assume that the first sample starts at time zero. Before applying the DFT, the data get windowed. Let $\boldsymbol{W} = \mathrm{diag}\,\{w_0, w_1, \ldots, w_{N-1}\}$ and $\boldsymbol{F}$ is the $N \times N$ DFT matrix, with inverse DFT $\boldsymbol{F}^{-1} = \frac{1}{N}\boldsymbol{F}^* = \frac{1}{N}\boldsymbol{F}^H$. Then we shall work with the transformed windowed data vector

$$\boldsymbol{Y}^w = \boldsymbol{F}\,\boldsymbol{W}\,\boldsymbol{y} \tag{4.12}$$

The data are assumed to have zero mean so that covariance and correlation matrices are equal. Note now that $\boldsymbol{y}$ is real, but $\boldsymbol{Y}^w$ is complex due to the DFT. $\boldsymbol{Y}^w$ is strictly speaking non-circular as both $\boldsymbol{R}(\boldsymbol{\theta}) = \mathrm{E}\left\{\boldsymbol{Y}^w(\boldsymbol{Y}^w)^H\right\}$ and $\mathrm{E}\left\{\boldsymbol{Y}^w(\boldsymbol{Y}^w)^T\right\}$ are nonzero. However, $\boldsymbol{Y}^w$ is not a genuine complex random vector as only the real vector $\boldsymbol{y}$ is random and the complex aspect is due to a deterministic transformation. As a result we can continue as if $\boldsymbol{Y}^w$ has a circular complex Gaussian distribution which corresponds to a real Gaussian distribution with transposes replaced by Hermitian transposes.

Now, all we need for GML is to compute the expression of $\boldsymbol{R}(\boldsymbol{\theta})$. Note that component Y_k^w of $\boldsymbol{Y}^w = [Y_0^w\, Y_1^w \cdots Y_{N-1}^w]^T$ is in fact the discrete-time Fourier transform (DTFT) of the windowed signal, $Y^w(f)$, evaluated at frequency $f = k/N$. To constitute $\boldsymbol{R}(\boldsymbol{\theta})$, we shall need the correlations between different frequencies

$\mathrm{E}\{Y^w(f_1)\,Y^{w*}(f_2)\}$. For this we consider

$$\begin{aligned} Y^w(f_1) &= \sum_{n=0}^{N-1} w_n\, y_n e^{-j2\pi f_1 n} && (4.13)\\ &= \sum_{n=-\infty}^{\infty} w_n e^{-j2\pi f_1 n}\, y_n && (4.14)\\ &= \sum_{n=-\infty}^{\infty} h_{-n}\, y_n && (4.15)\\ &= h_n * y_n|_{n=0} && (4.16)\\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} H(f)\, Y(f)\, df && (4.17)\\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} W(f_1 - f)\, Y(f)\, df && (4.18) \end{aligned}$$

Where $h_n = w_n e^{j2\pi f_1 n}$. $Y(f)$, $W(f)$ and $H(f)$ are the DTFT of the stationary random process y_n, w_n and h_n respectively. Notice that we zeropadded the finite window to infinity. Now we get

$$\begin{aligned} \mathrm{E}\{Y^w(f_1)\,Y^{w*}(f_2)\} &= \mathrm{E}\left\{\int W(f_1 - f)\, Y(f)\, df \int W^*(f_2 - f_0)\, Y^*(f_0)\, df_0 \right\}\\ &= \int df\, W(f_1 - f) \int df_0 W^*(f_2 - f_0)\, \mathrm{E}\{Y(f)Y^*(f_0)\}\\ &= \int df\, W(f_1 - f) \int df_0 W^*(f_2 - f_0)\, S_{yy}(f, \boldsymbol{\theta})\, \delta_1(f - f_0)\\ &= \int df\, W(f_1 - f)\, W^*(f_2 - f)\, S_{yy}(f, \boldsymbol{\theta}) \qquad (4.19) \end{aligned}$$

where $S_{yy}(f, \boldsymbol{\theta})$ is the spectrum of y_n, and $\delta_1(f) = \sum_{k=-\infty}^{\infty} \delta(f - k)$ is the periodicized delta function. Now let us introduce the vector of DFT frequencies $\underline{f} = [0\,1 \cdots N-1]^T/N$ and the $N \times 1$ vector of ones $\underline{1}$, let $W(\underline{f})$ denote the column vector of $W(.)$ evaluated at the components of $\underline{f}$, then we can write for

$$\boldsymbol{R}(\boldsymbol{\theta}) = \int df\, W(\underline{f} - f\underline{1})\, W^H(\underline{f} - f\underline{1})\, S_{yy}(f, \boldsymbol{\theta}) \qquad (4.20)$$

We get in particular for the diagonal elements $\boldsymbol{R}_{kk} = \int df\, |W(\frac{k-1}{N} - f)|^2\, S_{yy}(f, \boldsymbol{\theta})$ which is the well-known spectrum smearing appearing in the mean of the periodogram. For an example, see figure Fig. 4.2 of a 64 ms length Hann window.

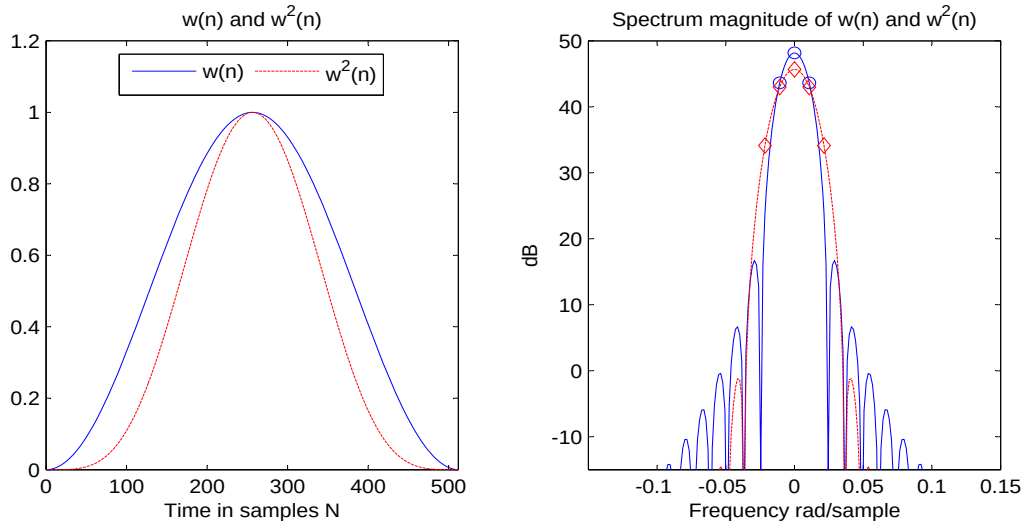

Figure 4.2: *The smearing effect*

Now, to limit complexity in the frequency domain based methods, one should sparsify $\boldsymbol{R}(\boldsymbol{\theta})$ as much as possible. Here is where the window design comes in. For a properly designed window, $W(f)$ can be neglected outside of its main lobe (see figure Fig. 4.2). Note that from this point of view, a rectangular window is (again) not a very good choice since the sidelobes are not much attenuated. If Δf is the doublesided width of the main lobe of $W(f)$, then the correlation in (4.19) can be approximated to zero for $|f_1 - f_2| > \Delta f$. This means that $\boldsymbol{R}(\boldsymbol{\theta})$ can be approximated by a banded matrix with only $\lceil N\,\Delta f \rceil$ non-zero diagonals. The inversion of $\boldsymbol{R}(\boldsymbol{\theta})$ can then be done efficiently using the lower upper (LU) triangular factorization of $\boldsymbol{R}(\boldsymbol{\theta})$ in which the triangular factors will also be banded. Compared to classical frequency-domain asymptotics, the spectrum gets smeared on the diagonal and spills onto the main sub- and super-diagonals, leading to correlations between neighboring frequencies (only). In those classical asymptotics, the smearing effect of $W(f)$ gets neglected, leading to $\boldsymbol{R} = \operatorname{diag}\{S_{yy}(\underline{f}, \boldsymbol{\theta})\}$. If $S_{yy}(f, \boldsymbol{\theta})$ is sufficiently smooth, the integral in (4.20) can be approximated by a sum over frequencies spaced more densely at $\underline{f}'$, containing multiples of $1/N'$, where $N' > N$. This can be obtained by zeropadding the signal from N to N' samples and applying the DFT of size N'. We then get $\boldsymbol{R}'(\boldsymbol{\theta})$ of the form

$$\boldsymbol{R}'(\boldsymbol{\theta}) = \mathcal{C}(W(\underline{f}'))\, \operatorname{diag}\{S_{yy}(\underline{f}', \boldsymbol{\theta})\}\, \mathcal{C}^H(W(\underline{f}')) \tag{4.21}$$

where $\mathcal{C}$ denotes a circulant matrix constructed from the vector argument. The entries of $\boldsymbol{R}'(\boldsymbol{\theta})$ can be downsampled to obtain $\boldsymbol{R}(\boldsymbol{\theta})$ if desired. In the maximization of GML in (4.11), we need to determine the derivatives of $\boldsymbol{R}(\boldsymbol{\theta})$ w.r.t. the components of $\boldsymbol{\theta}$. If we denote by θ_i the i^{th} component of $\boldsymbol{\theta}$ and using the expression of $\boldsymbol{R}(\boldsymbol{\theta})$

given in (4.20), we get for the derivatives

$$\frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial \theta_i} = \int df\, W(\underline{f} - f\underline{1})\, W^H(\underline{f} - f\underline{1})\, \frac{\partial S_{yy}(f,\boldsymbol{\theta})}{\partial \theta_i} \tag{4.22}$$

In the next section, we will introduce the expression of $S_{yy}(f,\boldsymbol{\theta})$ for the joint speech model used in chapter 3.

4.5 Periodic sources with short-term AR spectral envelope

We consider the same speech and observation models as in chapter 3 where the single microphone measurement signal y_n is composed of K quasiperiodic Gaussian sources $s_{k,n}$ plus white zero-mean Gaussian noise v_k of variance σ_v^2. The unknown vector of parameters $\boldsymbol{\theta}$ is defined as in (3.15). Assuming stationarity, the spectrum $S_{yy}(f,\boldsymbol{\theta})$ of y_n can be written as

$$S_{yy}(f,\boldsymbol{\theta}) = S_0(f,\boldsymbol{\theta}) + \sum_{k=1}^{K} S_k(f,\boldsymbol{\theta}_k) \tag{4.23}$$

where $S_0(f) = \sigma_v^2$, and the spectrum $S_k(f,\boldsymbol{\theta}_k)$ of source k is expressed as

$$S_k(f,\boldsymbol{\theta}_k) = \frac{\rho_k}{|A_k(f)|^2 |B_k(f)|^2} \qquad k = 1 \cdots K \tag{4.24}$$

where ρ_k adjusts the source power, $A_k(f)$ and $B_k(f)$ represent the AR spectral envelopes of the STP and LTP filters respectively, expressed as

$$A_k(f) = \sum_{i=0}^{p_k} a_{k,i} e^{-j2\pi f\, i} \qquad k = 1 \cdots K \tag{4.25}$$

$$B_k(f) = 1 + \beta_k e^{-j2\pi f\, T_k} \qquad k = 1 \cdots K \tag{4.26}$$

where f_k and p_k are the pitch and the AR order of source k. For quasiperiodic sources which are observed over a limited time frame where the latter is furthermore windowed with reduced weight towards the edges, we can neglect possible limited long-term correlation and model the source as a Gaussian periodic signal with STP spectral envelope, leading to a spectrum of the form

$$\begin{aligned} S_k(f,\boldsymbol{\theta}_k) &= \frac{\rho_k}{|A_k(f)|^2} \sum_{m=-\lfloor \frac{1}{2f_k} \rfloor}^{\lfloor \frac{1}{2f_k} \rfloor} \delta(f - m\, f_k) \\ &= \rho_k \sum_{m=-\lfloor \frac{1}{2f_k} \rfloor}^{\lfloor \frac{1}{2f_k} \rfloor} \frac{1}{|A_k(m\, f_k)|^2} \delta(f - m\, f_k) \qquad k = 1 \cdots K \end{aligned} \tag{4.27}$$

With the above spectrum model, the parameters $\boldsymbol{\theta}$ are reduced to $\{\sigma_v^2, \rho_k, a_{k,i}, i = 1, \ldots, p_k,\ k = 1, \ldots, K\}$ and we get for $\boldsymbol{R}(\theta)$

$$\boldsymbol{R}(\boldsymbol{\theta}) = \sigma_v^2 \int df\, W(\underline{f} - f\underline{1})\, W^H(\underline{f} - f\underline{1}) + \sum_{k=1}^{K} \rho_k \sum_{m=-\lfloor \frac{1}{2f_k} \rfloor}^{\lfloor \frac{1}{2f_k} \rfloor} \frac{1}{|A_k(m\, f_k)|^2} W(\underline{f} - m\, f_k\underline{1})\, W^H(\underline{f} - m\, f_k\underline{1}) \quad (4.28)$$

In figure Fig. 4.3, we show an example of matrix $\boldsymbol{R}(\boldsymbol{\theta})$ when using two types of windows (Hann window and the rectangular window) in order to illustrate the importance of window's choice in getting the sparse banded from of $\boldsymbol{R}(\boldsymbol{\theta})$. In the next section, we express the GML extrema for the unknown parameters $\boldsymbol{\theta}$ using the covariance matrix in (4.28).

4.6 Parameters estimation

In this section, we will derive the estimate of the unknown parameters $\boldsymbol{\theta}$. The gradient of the spectrum $S_{yy}(f, \boldsymbol{\theta})$ relative to $\boldsymbol{\theta}$ is computed, then injected in (4.11). By equating to zero, the different estimates are deduced. In order to avoid cumbersome notations, we will note the covariance matrix shortly $\boldsymbol{R}$ instead of $\boldsymbol{R}(\boldsymbol{\theta})$.

4.6.1 Estimation of the inputs and observation noise powers

We want to estimate the inputs power $\{\rho_k\}_{k=1:K}$ and the observation noise power σ_v^2 by maximizing the GML criterium. In (4.11), we need to compute the gradient of the covariance matrix w.r.t. these parameters. Let's adopt this notation

$$\boldsymbol{R} = \sum_{k=0}^{K} \rho_k \boldsymbol{G}_k \quad (4.29)$$

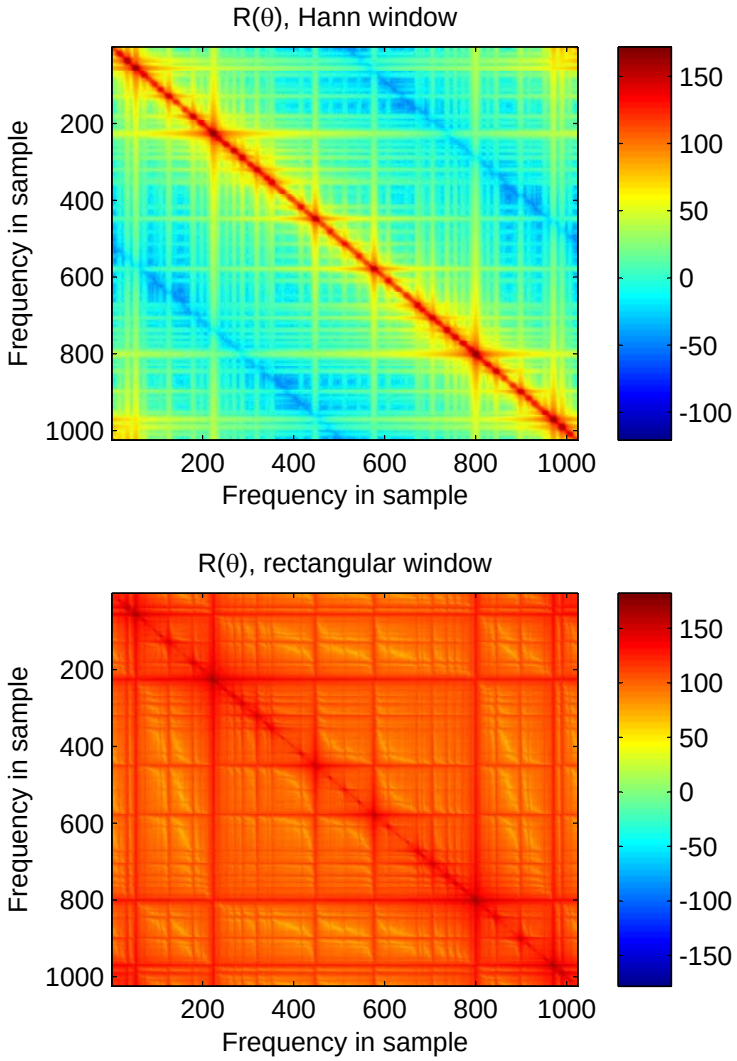

Figure 4.3: *The structure of covariance matrix* $\boldsymbol{R}(\boldsymbol{\theta})$ *when using Hann window and rectangular window*

where $\rho_0 = \sigma_v^2$ and $\boldsymbol{G}_k$ is defined as it follows

$$\boldsymbol{G}_0 = \int df\, W(\underline{f} - f\underline{1})\, W^H(\underline{f} - f\underline{1}) \tag{4.30}$$

$$\boldsymbol{G}_k = \sum_{m=-\lfloor \frac{1}{2f_k} \rfloor}^{\lfloor \frac{1}{2f_k} \rfloor} \frac{1}{|A_k(m\, f_k)|^2} W(\underline{f} - m\, f_k\underline{1})\, W^H(\underline{f} - m\, f_k\underline{1}) \quad k = 1 \cdots K \tag{4.31}$$

Then, the gradient of the covariance matrix relative to ρ_k is simply $\boldsymbol{G}_k$. By injecting

this result in (4.11), we get

$$\begin{aligned}\frac{\partial GML(\theta)}{\partial \rho_k} &= -\text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}(\widehat{\boldsymbol{R}} - \sum_{k'=0}^{K}\rho_{k'}\boldsymbol{G}_{k'})\boldsymbol{R}^{-1}\} \\ &= -\text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\} + \text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}(\sum_{k'=0}^{K}\rho_{k'}\boldsymbol{G}_{k'})\boldsymbol{R}^{-1}\} \\ &= -\text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\} + \text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}(\rho_k\boldsymbol{G}_k + \sum_{k'\neq k}\rho_{k'}\boldsymbol{G}_{k'})\boldsymbol{R}^{-1}\} \\ &= -\text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\} + \rho_k\text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}\boldsymbol{G}_k\boldsymbol{R}^{-1}\} \\ &+ \sum_{k'\neq k}\rho_{k'}\text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}\boldsymbol{G}_{k'}\boldsymbol{R}^{-1}\} \end{aligned} \tag{4.32}$$

Hence, if we denote by $\boldsymbol{\rho} = [\rho_0, \rho_1, ..., \rho_K]^T$ the vector of observation noise and inputs powers, then the estimate of $\boldsymbol{\rho}$ (denoted $\hat{\boldsymbol{\rho}}$) is computed by resolving the system $\boldsymbol{M}\boldsymbol{\rho} = \boldsymbol{d}$, where:

$$M_{kk'} = \text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}\boldsymbol{G}_{k'}\boldsymbol{R}^{-1}\} \qquad k, k' = 0\cdots K \tag{4.33}$$

$$d_k = \text{tr}\{\boldsymbol{G}_k\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\} \qquad k = 0\cdots K \tag{4.34}$$

Since $\boldsymbol{R}$ and $\{\boldsymbol{G}_k\}_{k=1:K}$ depend on $\boldsymbol{\theta}$, what we do in practical is that we update them with the last estimate of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}^i$ to get $\boldsymbol{R}(\widehat{\boldsymbol{\theta}}^i) = \boldsymbol{R}^i$ and $\boldsymbol{G}_k(\widehat{\boldsymbol{\theta}}_k^i) = \boldsymbol{G}_k^i$ (consequently $\boldsymbol{M}^i$ and $\boldsymbol{d}^i$). $\boldsymbol{R}^i$ is then efficiently inverted using LU factorization and forward/backward substitution. The estimate $\hat{\boldsymbol{\rho}}^i$ is then expressed

$$\hat{\boldsymbol{\rho}}^i = \left(\boldsymbol{M}^i\right)^{-1}\boldsymbol{d}^i \tag{4.35}$$

4.6.2 Estimation of the AR coefficients

In this section, we will find out the estimate of the AR coefficients $\{a_{k,i}, i = 1, \ldots, p_k,\, k = 1, \ldots, K\}$. The derivation of $\boldsymbol{R}(\boldsymbol{\theta})$ in (4.28) w.r.t. A_k^* is computed as it follows

$$\frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial A_k^*} = -\sum_{m=-\lfloor\frac{1}{2f_k}\rfloor}^{\lfloor\frac{1}{2f_k}\rfloor} W(\underline{f} - m\,f_k\underline{1})\,W^H(\underline{f} - m\,f_k\underline{1})\frac{S_k(m\,f_k, \boldsymbol{\theta}_k)}{|A_k(m\,f_k)|^2}A_k(m\,f_k) \tag{4.36}$$

After plugging (4.36) in (4.11) and equating to zero, we get

$$\begin{aligned}&\text{tr}\Big\{\sum_{m=-\lfloor\frac{1}{2f_k}\rfloor}^{\lfloor\frac{1}{2f_k}\rfloor} W(\underline{f} - m\,f_k\underline{1})W^H(\underline{f} - m\,f_k\underline{1})\frac{S_k(m\,f_k, \boldsymbol{\theta}_k)}{|A_k(m\,f_k)|^2}A_k(m\,f_k)\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\Big\} \\ &= \text{tr}\Big\{\sum_{m=-\lfloor\frac{1}{2f_k}\rfloor}^{\lfloor\frac{1}{2f_k}\rfloor} W(\underline{f} - m\,f_k\underline{1})W^H(\underline{f} - m\,f_k\underline{1})\frac{S_k(m\,f_k, \boldsymbol{\theta}_k)}{|A_k(m\,f_k)|^2}A_k(m\,f_k)\boldsymbol{R}^{-1}\Big\}\end{aligned} \tag{4.37}$$

If we switch the trace and the finite sum, we get

$$\sum_{m=-\lfloor\frac{1}{2f_k}\rfloor}^{\lfloor\frac{1}{2f_k}\rfloor} \mathrm{tr}\left\{W(\underline{f}-m\,f_k\underline{1})W^H(\underline{f}-m\,f_k\underline{1})\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\right\}\frac{S_k(m\,f_k,\boldsymbol{\theta}_k)}{|A_k(m\,f_k)|^2}A_k(m\,f_k)$$

$$=\sum_{m=-\lfloor\frac{1}{2f_k}\rfloor}^{\lfloor\frac{1}{2f_k}\rfloor} \mathrm{tr}\left\{W(\underline{f}-m\,f_k\underline{1})W^H(\underline{f}-m\,f_k\underline{1})\boldsymbol{R}^{-1}\right\}\frac{S_k(m\,f_k,\boldsymbol{\theta}_k)}{|A_k(m\,f_k)|^2}A_k(m\,f_k) \quad (4.38)$$

Consequently

$$\left[\mathrm{tr}\left\{W(\underline{f}-m\,f_k\underline{1})W^H(\underline{f}-m\,f_k\underline{1})\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\right\}\frac{S_k(m\,f_k,\boldsymbol{\theta}_k)}{|A_k(m\,f_k)|^2}\right]A_k(m\,f_k)$$

$$=\mathrm{tr}\left\{W(\underline{f}-m\,f_k)W^H(\underline{f}-m\,f_k)\boldsymbol{R}^{-1}\right\}\frac{S_k(m\,f_k,\boldsymbol{\theta}_k)}{A_k^*(m\,f_k)} \quad (4.39)$$

As in [Schutz & Slock 2010], This is a Yule-Walker like equation with non zero Right Hand Side (RHS) which is solved iteratively

$$\boldsymbol{T}(\boldsymbol{r}_{k,(0,\ldots,p_k-1)})\;\boldsymbol{a}_k=\boldsymbol{g}_{k,(1,\ldots,p_k)}-\boldsymbol{r}_{k,(1,\ldots,p_k)} \quad (4.40)$$

where $\boldsymbol{T}$ is the Toeplitz matrix constructed from the first p_k elements of $\boldsymbol{r}_k$, $\boldsymbol{a}_k$ is the short term AR coefficients vector of source k as defined in (3.15). The p_k-dimensional vectors $\boldsymbol{r}_k$ and $\boldsymbol{g}_k$ are defined as

$$\boldsymbol{r}_k=\boldsymbol{F}^{-1}\left(\mathrm{tr}\left\{W(\underline{f}-m\,f_k\underline{1})W^H(\underline{f}-m\,f_k\underline{1})\boldsymbol{R}^{-1}\widehat{\boldsymbol{R}}\boldsymbol{R}^{-1}\right\}\frac{S_k(m\,f_k,\boldsymbol{\theta}_k)}{|A_k(m\,f_k)|^2}\right) \quad (4.41)$$

$$\boldsymbol{g}_k=\boldsymbol{F}^{-1}\left(\mathrm{tr}\left\{W(\underline{f}-m\,f_k)W^H(\underline{f}-m\,f_k)\boldsymbol{R}^{-1}\right\}\frac{S_k(m\,f_k,\boldsymbol{\theta}_k)}{A_k^*(m\,f_k)}\right) \quad (4.42)$$

The alternating optimization scheme will consist in iterating (4.40) for source k by updating each time the resulting $\hat{\boldsymbol{a}_k}^i$ in the computation of $\boldsymbol{R}^i$, $\hat{\boldsymbol{\rho}}^i$ and $\boldsymbol{g}_k^i$ until convergence of the latter, then we move to the next source $k+1$ and repeat the same iteration scheme. Global convergence is achieved when for all sources the correlation vectors $\{\boldsymbol{r}_k\}_{k=1:K}$ become stable. We remind that all this algorithm is achieved per processing frame. The parameters of the current processing frame are initialized using the estimates resulting from the previous frame except of the first frame which is initialized randomly. The estimation of pitches is handled as in chapter 3

4.6.3 Source estimation

For a given processing frame l and after estimation of the parameters $\widehat{\boldsymbol{\theta}}^l=\left\{\hat{\boldsymbol{a}}_k^l,\hat{\rho}_k^l,\hat{f}_k^{\,l}\right\}_{k=1..K}$, the sources are estimated using Wiener filtering followed by

a DFT inverse operation

$$\hat{\boldsymbol{s}}_k^l = \boldsymbol{F}^{-1}\, \boldsymbol{R}_{SY}^l\, \boldsymbol{R}^l\, \boldsymbol{Y}^w \tag{4.43}$$

$$\boldsymbol{R}_{SY}^l = \mathrm{E}\{\boldsymbol{S}_k^w (\boldsymbol{Y}^w)^H\} = \mathrm{E}\{\boldsymbol{S}_k^w (\boldsymbol{S}_k^w)^H\} = \hat{\boldsymbol{\rho}}_k^l \boldsymbol{G}_k^l \tag{4.44}$$

where $\boldsymbol{S}_k^w$ is the DFT of the windowed source k and $\hat{\boldsymbol{s}}_k^l$ is the estimate of source k in frame l. The full length estimates of sources are reconstructed from the different piecewise estimates of sources using the overlap-add technique. The overall algorithm is described in algorithm. 3.

Algorithm 3 Spectral envelope estimation algorithm

1: For each processing frame l
2: Initialize the parameters using the estimates of the previous frame $\left\{\hat{\boldsymbol{a}}_k^{l-1}, \hat{\rho}_k^{l-1}, \hat{f}_k{}^l\right\}_{k=1..K}$.
3: Until convergence of $\{\boldsymbol{r}_k^{i_r}\}_{k=1:K}$ and if $i_r \leq i_{r,max}$
4: For all sources k
5: Until convergence of $\boldsymbol{g}_k^{i_g}$ and if $i_g \leq i_{g,max}$, estimate $\boldsymbol{a}_k$ using (4.40), update $\boldsymbol{R}^{i_g}$, $\hat{\boldsymbol{\rho}}^{i_g}$ and $\boldsymbol{g}_k^{i_g}$ using (4.28), (4.35) and (4.42) respectively.
6: Update $\{\boldsymbol{r}_k^{i_r}\}_{k=1:K}$ with the last estimates $\left\{\hat{\boldsymbol{a}}_k^{i_g}\right\}_{k=1..K}$, $\left\{\hat{\rho}_k^{i_g}\right\}_{k=0..K}$ using (4.41). If not converged, back to step 3.
7: Construct sources per frame using (4.43).
8: Construct the full length sources using the overlap-add technique.

4.6.4 Frequency domain Cramér-Rao bound

For a Gaussian process with zero mean, the element (i, j) (pertaining to θ_i and θ_j) of the Fisher information matrix (FIM) is obtained as [Stoica & Moses 2005]

$$FIM_{i,j} = \mathrm{tr}\{\boldsymbol{R}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial \theta_i} \boldsymbol{R}^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial \theta_j}\} \,. \tag{4.45}$$

In the classical asymptotics, the FIM gets then approximated as

$$FIM_{i,j} = \int df\, S_{yy}^{-2}(f, \boldsymbol{\theta}) \frac{\partial S_{yy}(f, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial S_{yy}(f, \boldsymbol{\theta})}{\partial \theta_j} \tag{4.46}$$

$$= \int df\, \frac{\partial \ln S_{yy}(f, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln S_{yy}(f, \boldsymbol{\theta})}{\partial \theta_j} \tag{4.47}$$

4.7 Simulations

Synthetic spectra

In this part, we run the algorithm for synthetic spectra in order to illustrate the parameters estimation part of the algorithm. Like in chapter 3, the source number is

fixed to $K = 2$ and the AR order is fixed to $p_k = 10$ for all sources k. The sampling frequency F_s is set to 8 KHz. We simulate at high source to noise ratio (SNR) of 20 dB. We use a Hann window of length 64 ms. The pitches of the different sources are set to $f_1 = 140.35$ Hz and $f_2 = 250$ Hz. The AR coefficients and the powers vector $\boldsymbol{\rho}$ that are used to generate the synthetic spectra are set as it follows

$$\begin{aligned}
\boldsymbol{a}_1 &= \big[-1.5020, 1.7380, -2.0290, 1.7890, -1.3760, 1.2550, -0.6930, 0.3760, \\
&\quad -0.0800, 0.0330\big]^T \\
\boldsymbol{a}_2 &= \big[0.2153, 0.2153, -0.0176, 0.0806, 0.0127, 0.1569, -0.0218, 0.2146, \\
&\quad -0.7225, -0.3034\big]^T \\
\boldsymbol{\rho} &= \big[7.5473, 1.5, 1\big]^T
\end{aligned}$$

The simulated synthetic spectra of the sources and the noisy mixture are displayed in figure Fig. 4.4. In the figure, **TrueSpec k** and **SpecEnv k** denote the parametric spectrum of source k and its corresponding spectral envelope respectively. In the estimation process, AR coefficients are initialized to zero (flat spectrum) for both sources. The initial vector $\hat{\boldsymbol{\rho}}^0$ is initialized to $0.1 \times \boldsymbol{\rho}$. Here, the pitches are assumed known. In figure Fig. 4.5, we show the estimation results compared to the original values where the algorithm presents a perfect estimation result. In this case, the spectral envelopes of the sources are distinct. In a second simulation, we produce the synthetic spectra of two sources with identical AR envelopes ($\boldsymbol{a}_2$) but different pitches (the same pitches as in the first simulation). We also adjust the input powers in order to keep a low signal to signal ratio (SSR) and set them to $\rho_1 = 3$ and $\rho_2 = 1$. The new sources spectra are displayed in figure Fig. 4.6. The same initialization approach is done here too. In this case, the algorithm manages to extract the parameters perfectly as seen in Fig. 4.7. In fact, the main element that makes the two sources distinguishable is to have different pitches with minimum overlap between their corresponding harmonics from the where the importance to have a good pitch estimation in every processing frame.

4.8 Conclusion

In this chapter, we proposed a frequency method for monomicrophone source separation. Due to the non-stationary nature of speech, a frame-wise processing is proposed raising the importance of choosing carefully the multiplying window in order to assure a PR of the estimated speech. The effect of the window is studied in the finite length case. The algorithm is based on a joint modeling of speech inducing a parametric spectrum model for each sources. The estimation of these spectra is based on optimizing the cost function GML. In previous work [Schutz & Slock 2011], the

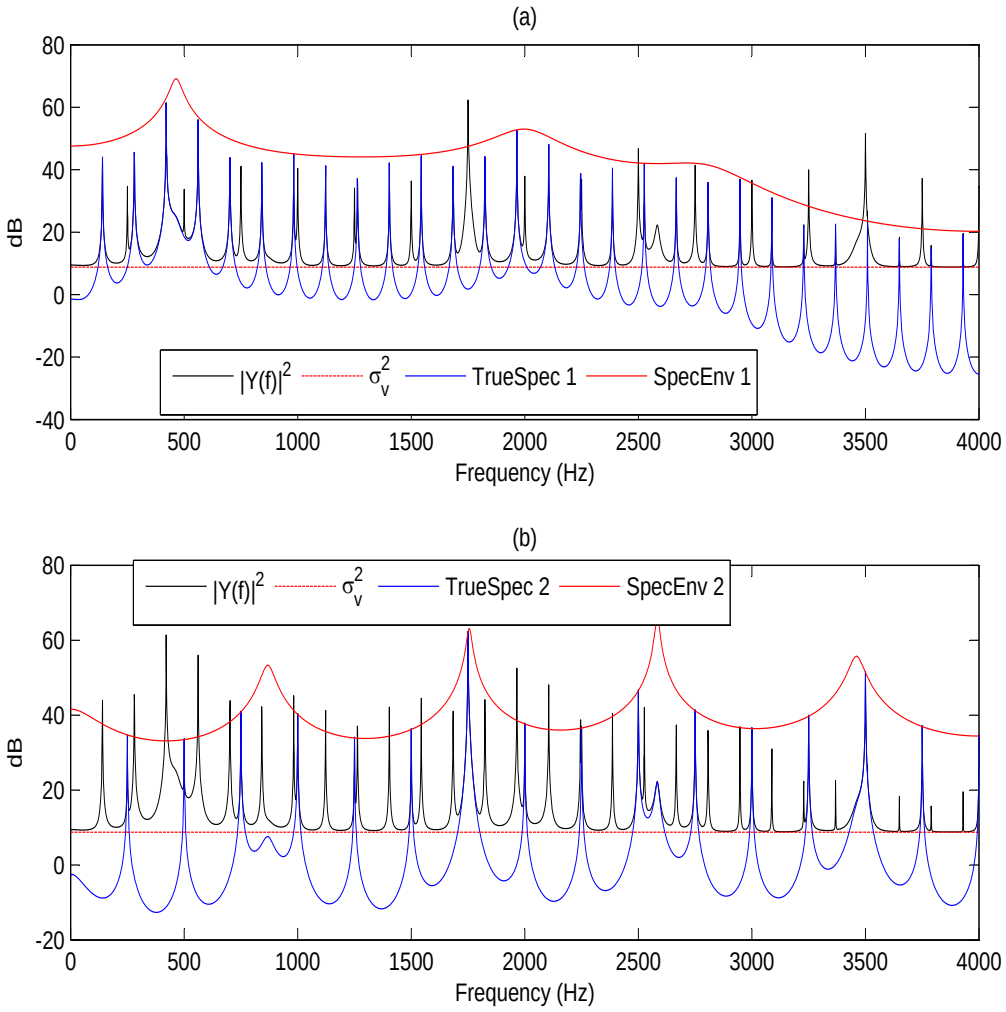

Figure 4.4: *Spectra of sources 1 (a) and 2 (b), and observation data.*

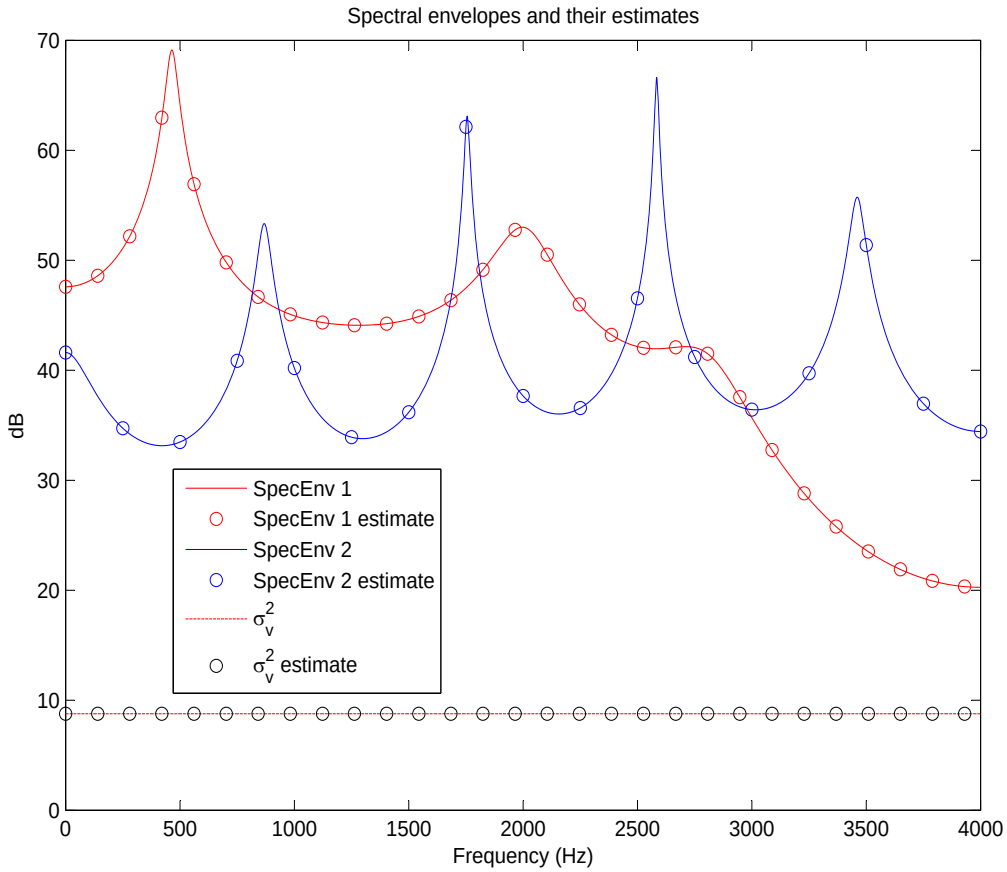

Figure 4.5: *Comparison of original parameters and their estimates (case of different spectral envelopes)*

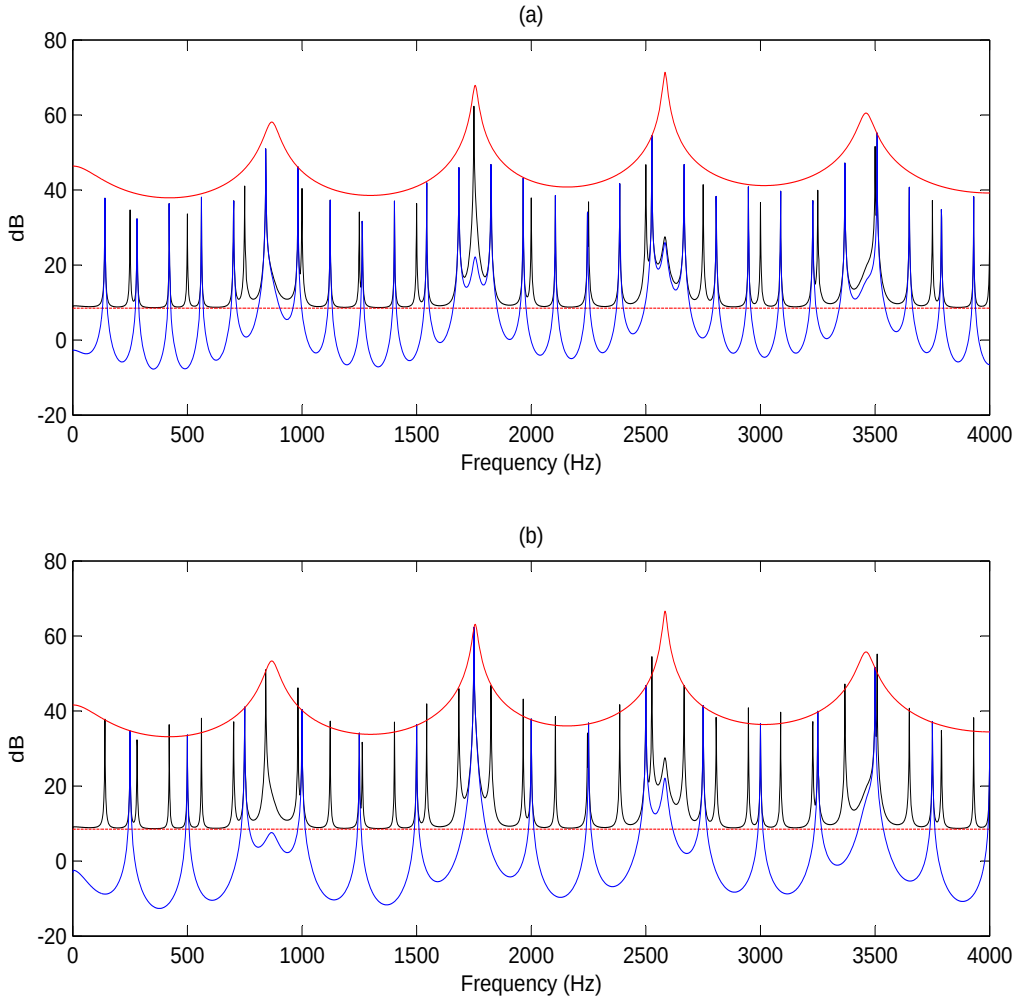

Figure 4.6: *Spectra of sources 1 (a) and 2 (b), and observation data.*

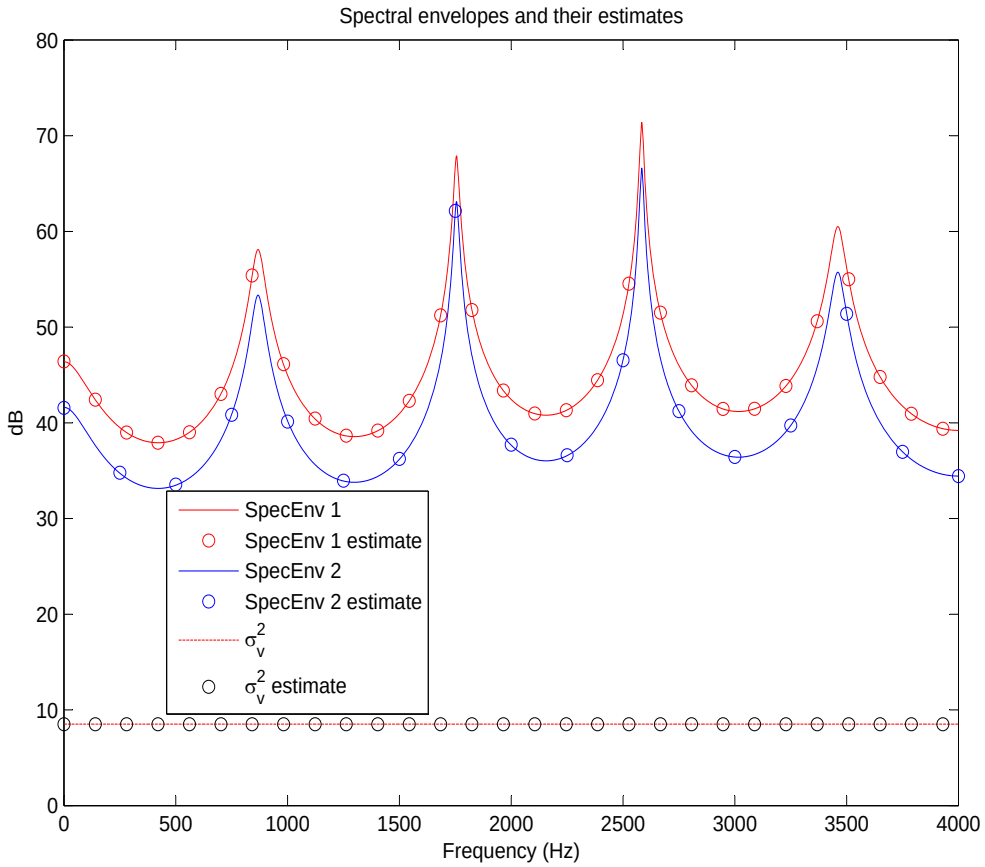

Figure 4.7: *Comparison of original parameters and their estimates (case of similar spectral envelopes)*

GML was proved equivalent to the ISD and OWSM in asymptotic conditions (infinite length frames). In our work, this equivalence is proved for finite length windows for ISD, and forOWCM in terms of gradient and under specific hypothesis. The optimization of the GML finds out the parameters that reduce the difference between the parametric spectra and the observations periodogram, in an iterative fashion. the estimate of the parametric spectrums permits the reconstruction of the final sources using a PR window and the overlap-add technique.

Chapter 5

On the Cramer-Rao like lower bounds for performance evaluation

5.1 Introduction

In chapters 3 and 4, we tackled the problem of estimating the sources, which are random variables, and their short term prediction (STP)+long term prediction (LTP) parameters which are deterministic. In chapter 3, we used a joint (hybrid) estimation strategy, the expectation maximization (EM)-Kalman, whereas in chapter 4, the sources were eliminated (marginalized), the parameters were estimated using only the observation likelihood (Gaussian maximum likelihood (GML) function) then their estimates used to estimate sources. Then, two strategies are present These different estimation strategies raise the question of what is the best strategy? (best in the mean square error (MSE) sense). In terms of estimation performance evaluation, we generally start by studying/comparing methods in the asymptotic regime. An ubiquitous used tool is the Cramér-Rao bound (CRB) which provides a lower bound to the achieved MSE. There are several types of bounds derived in literature. The choice of the bound depends on many factors: the nature of the estimated parameters (random, nonrandom or hybrid case), the presence or not of nuisance parameters, tractability of the likelihood function, etc. In the case of hybrid estimation, two famous bounds are used to evaluate the asymptotic performance of the deterministic parameters estimation, the hybrid Cramér-Rao bound (HCRB) and modified Cramér-Rao bound (MCRB). The latter is more precisely used when the estimation of the random variables is not the main goal. Our question is: can the difference between the HCRB and CRB be informative about the influence of the random variables on the deterministic ones? the same question is raised for the MCRB. In this chapter, we will start by introducing the joint Gaussian framework. Afterward, we will give a tutorial about the existing bounds, their bias condition, tightness and possible comparison between them. For an exhaustive list of papers about lower bounds, the reader may refer to Van Trees book [Van Trees & Bell 2007]. We will then characterize the difference, first, between the inverses of HCRB and CRB, and second, between the inverses of MCRB and CRB. We finally conclude.

5.2 Jointly Gaussian Framework

In this chapter, the conventional notations will be lowercase for scalars, bold lowercase for vectors and bold capital for matrices. Let $\boldsymbol{y}$ denotes the $N \times 1$ measurement vector on the basis of which we want to estimate the $M \times 1$ random process $\boldsymbol{x}$. We assume $\boldsymbol{y}$ and $\boldsymbol{x}$ are both Gaussian of distributions $\mathcal{N}(\boldsymbol{\mu_{yy}}(\boldsymbol{\theta}), \boldsymbol{C_{yy}}(\boldsymbol{\theta}))$ and $\mathcal{N}(\boldsymbol{\mu_{xx}}(\boldsymbol{\theta}), \boldsymbol{C_{xx}}(\boldsymbol{\theta}))$ respectively where $\boldsymbol{\theta}$ is an $L \times 1$ unknown vector that parameterizes the two distributions, and consequently parameterizes the joint likelihood function $f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$. All the considered quantities here are real. In the jointly Gaussian setting, the whole estimation problem is characterized by the joint mean $\boldsymbol{\mu}(\boldsymbol{\theta}) = [\boldsymbol{\mu}_{\boldsymbol{yy}}^T(\boldsymbol{\theta})\, \boldsymbol{\mu}_{\boldsymbol{xx}}^T(\boldsymbol{\theta})]^T$ and the joint covariance matrix $\boldsymbol{C}(\boldsymbol{\theta})$

$$\boldsymbol{C}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{C_{yy}}(\boldsymbol{\theta}) & \boldsymbol{C_{yx}}(\boldsymbol{\theta}) \\ \boldsymbol{C_{xy}}(\boldsymbol{\theta}) & \boldsymbol{C_{xx}}(\boldsymbol{\theta}) \end{bmatrix} \tag{5.1}$$

where $\boldsymbol{C_{xy}}(\boldsymbol{\theta})$ is the cross-covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$. In this context, two scenarios are possible to estimate $\boldsymbol{x}$ and $\boldsymbol{\theta}$. The first scenario consists in integrating out $\boldsymbol{x}$ since it is random and find out the marginalized maximum likelihood (ML) estimate of $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \ln f(\boldsymbol{y}|\boldsymbol{\theta}) \tag{5.2}$$

Where the loglikelihood for the Gaussian model is expressed

$$\ln f(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{2} \ln \det \boldsymbol{C_{yy}}(\boldsymbol{\theta}) - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu_{yy}}(\boldsymbol{\theta}))^T \boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta}) (\boldsymbol{y} - \boldsymbol{\mu_{yy}}(\boldsymbol{\theta})) \tag{5.3}$$

Then the estimate $\widehat{\boldsymbol{\theta}}_{ML}$ is used to estimate $\boldsymbol{x}$. This scenario is also used when $\boldsymbol{x}$ are considered as nuisance parameters and the target estimate is $\boldsymbol{\theta}$. The second scenario is to estimate both jointly using maximum a posteriori (MAP) for $\boldsymbol{x}$ and ML for $\boldsymbol{\theta}$. In the Gaussian case, the MAP solution (estimator that maximizes the posterior distribution) coincides with the minimum mean square error (MMSE) solution. In particular, we get for the posterior distribution

$$f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{x}}(\boldsymbol{\theta}), \boldsymbol{P}(\boldsymbol{\theta})) \tag{5.4}$$

where $\hat{\boldsymbol{x}}(\boldsymbol{\theta})$ is the MMSE estimate and $\boldsymbol{P}(\boldsymbol{\theta})$ is its covariance matrix, which are both defined as

$$\hat{\boldsymbol{x}}(\boldsymbol{\theta}) = \boldsymbol{\mu_{xx}}(\boldsymbol{\theta}) + \boldsymbol{F}(\boldsymbol{\theta})(\boldsymbol{y} - \boldsymbol{\mu_{yy}}(\boldsymbol{\theta})) \tag{5.5}$$

$$\boldsymbol{F}(\boldsymbol{\theta}) = \boldsymbol{C_{xy}}(\boldsymbol{\theta}) \boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta}) \tag{5.6}$$

$$\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{C_{xx}}(\boldsymbol{\theta}) - \boldsymbol{C_{xy}}(\boldsymbol{\theta}) \boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta}) \boldsymbol{C_{yx}}(\boldsymbol{\theta}) \tag{5.7}$$

Notice that the posterior distribution in (5.4) is maximum at $\hat{\boldsymbol{x}}(\boldsymbol{\theta})$ and the value of its maximum is independent of $\boldsymbol{\theta}$

$$\begin{aligned} \arg\max_{\boldsymbol{x}} \ln f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) &= \hat{\boldsymbol{x}}(\boldsymbol{\theta}) \\ \max_{\boldsymbol{x}} \ln f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) &= -\tfrac{1}{2} \ln \det \boldsymbol{P}(\boldsymbol{\theta}) \end{aligned} \tag{5.8}$$

Hence due to this separability, a compressed form of the joint likelihood is obtained (5.9) which remains to be optimized w.r.t. $\boldsymbol{\theta}$

$$\max_{\boldsymbol{x}} \ln\, f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) = \ln\, f(\boldsymbol{y}|\boldsymbol{\theta}) - \frac{1}{2}\ln\det \boldsymbol{P}(\boldsymbol{\theta}) \tag{5.9}$$

This separability was also noted by Yeredor in [Yeredor 2000] and used to study the joint MAP/ML algorithm performance. In the following sections, the lower bounds that govern the performance of these two estimators are derived.

5.3 Lower bounds for deterministic parameters estimation

In this section, we will survey the lower bounds used when the estimated parameters vector $\boldsymbol{\theta}$ is deterministic. We denote by $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ an estimator of $\boldsymbol{\theta}$ which is a function of the observations vector $\boldsymbol{y}$ but we will omit the $\boldsymbol{y}$ later on to simplify notations. An important notion should be reminded which is the bias element $\boldsymbol{b}$. The bias of an estimator $\widehat{\boldsymbol{\theta}}$ is defined [Van Trees 2002, Van Trees 2001]

$$\boldsymbol{b}(\widehat{\boldsymbol{\theta}}) = \mathrm{E}\{\widehat{\boldsymbol{\theta}}\} - \boldsymbol{\theta} \tag{5.10}$$

The probability density function (pdf) according to which the expectation operator in (5.10) is applied depends on the problem context. An estimator $\widehat{\boldsymbol{\theta}}$ is said to be unbiased if and only if $\boldsymbol{b}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{0}$. In general, lower bounds are defined for unbiased estimators.

5.3.1 The Cramér-Rao bound

The CRB was introduced for the first time by Harald Cramér and Calyampudi Radhakrishna Rao in [Cramér 1946, Rao 1945]. For any unbiased estimator $\widehat{\boldsymbol{\theta}}$, the covariance matrix of $\widehat{\boldsymbol{\theta}}$ is lower bounded by the CRB defined as the following

$$\begin{aligned} CRB &\triangleq \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}}\left\{\frac{\partial \ln f\,(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \ln f\,(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right\}^{-1} \\ &= \boldsymbol{J}^{-1}(\boldsymbol{\theta}) \end{aligned} \tag{5.11}$$

where $\boldsymbol{J}(\boldsymbol{\theta})$ is the $L \times L$ Fisher information matrix (FIM). The expectation of the bias expression in (5.10) is taken according to the likelihood $f\,(\boldsymbol{y}|\boldsymbol{\theta})$. The CRB is achieved by the ML estimator [Lehmann & Casella 1998] in asymptotic conditions (high source to noise ratio (SNR) and/or infinite number of samples), in that case the estimator is said asymptotically efficient. A second expression is proved to define the CRB using the Hessian of the likelihood.

$$CRB = \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f\,(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}^{-1} \tag{5.12}$$

The CRB is used as a benchmark in many applications [Stoica & Arye 1989, Stoica & Nehorai 1990, Catovic & Sahinoglu 2004, Larsson 2004, Li *et al.* 2008]. Despite the derivation of other lower bounds, the CRB is the most used one due to its computational advantages.

Gaussian case

Since in most of the time, the involved signals are of Gaussian nature, it is noteworthy to mention the CRB in this case [Stoica & Moses 1997]. Since $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu_{yy}}(\boldsymbol{\theta}), \boldsymbol{C_{yy}}(\boldsymbol{\theta}))$, then the CRB inverse is expressed

$$CRB^{-1} = \frac{1}{2}\frac{\partial \boldsymbol{C}_{\boldsymbol{yy},\boldsymbol{\theta}}^{T}}{\partial \boldsymbol{\theta}}(\boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta})\otimes\boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta}))\frac{\partial \boldsymbol{C}_{\boldsymbol{yy},\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}^{T}} + \frac{\partial \boldsymbol{\mu}_{\boldsymbol{yy}}^{T}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\mu}_{\boldsymbol{yy}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{T}} \quad (5.13)$$

where $\boldsymbol{C}_{\boldsymbol{yy},\boldsymbol{\theta}} = vec\,(\boldsymbol{C}_{\boldsymbol{yy}}(\boldsymbol{\theta}))$.

Biased case

The CRB in (5.11) is only true when the estimator is unbiased. Since it is not always the case, an other derivation of CRB was developed for biased estimators [Van Trees 1968]

$$CRB_{\boldsymbol{b}} = \boldsymbol{b}\boldsymbol{b}^{T} + \left(\boldsymbol{I} + \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\theta}}\right)^{T} CRB\left(\boldsymbol{I} + \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\theta}}\right) \quad (5.14)$$

A biased estimator does not mean necessarily a bad estimator in the MSE sense since this latter depends also on the variance of the estimator. Thus, a good estimator makes a trade-off between the bias and the variance. In many applications [Demoment 1989, Carlson 1988, O'Sullivan 1986], the bias was used as a degree of freedom to improve the estimator performance, yet its choice was done in a *ad-hoc* way and not guaranteed to perform better for all values of $\boldsymbol{\theta}$. In [Eldar 2006], Eldar presents a more rigourous construction of the biased estimator when the bias is linear on the true parameters vector $\boldsymbol{\theta}$. Moreover, she proves that this estimator dominates the ML estimator.

5.3.2 The Bhattacharyya bound

The Bhattacharyya bound was introduced first by P. K. Bhattacharya in [Bhattacharya 1966]. It is considered as a generalization of the CRB. Assuming all the regularization conditions in [Bhattacharya 1966] fulfilled, we consider the problem of estimating the function $g(\theta)$ (notice that θ here is scalar). Let's denote by $\boldsymbol{g}(\theta)$ the k-dimensional vector $[g^{(1)}(\theta), \ldots, g^{(k)}(\theta)]^{T}$ where $g^{(i)}(\theta)$ is the i^{th} derivative of $g(\theta)$ w.r.t θ and by $\hat{g}(\boldsymbol{y})$ an unbiased estimator of this function. Then

the estimation error variance is lower bounded by the k^{th} order Bhattacharya bound

$$\mathrm{E}(\hat{g}(\boldsymbol{y}) - g(\theta))^2 \geq \boldsymbol{g}(\theta)^T \boldsymbol{J}_k^{-1}(\theta)\boldsymbol{g}(\theta) \tag{5.15}$$

Where $\boldsymbol{J}_k$ is the $k \times k$ matrix whose (i, j) element is given by

$$J_k^{ij} = \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}} \left\{ \frac{\partial^i \ln f(\boldsymbol{y}|\theta)}{\partial \theta^i} \frac{\partial^j \ln f(\boldsymbol{y}|\theta)}{\partial \theta^j} \right\} \tag{5.16}$$

In the special case when $g(\theta) = \theta$, the Bhattacharyya bound coincides with the CRB. In [Tanaka & Akahira 2003], the authors prove that the 2^{nd} order Bhattacharya bound is tight for a likelihood function $f(\boldsymbol{y}|\theta)$ belonging to the exponential family.

5.3.3 The Barankin bound

The Barankin bound (BB) was introduced first by Barankin in [Barankin 1949]. It is mainly used to detect the threshold effect where the CRB fails. The threshold effect is the SNR value under which the ML estimator starts to give inaccurate results. If we denote by $g(\boldsymbol{\theta})$ a real-valued scalar function of the vector $\boldsymbol{\theta}$ and $\hat{g}(\boldsymbol{\theta})$ an unbiased estimate of it that depends on the observations vector $\boldsymbol{y}$, the BB is the solution of the following optimization problem

$$\min_{\hat{g}(\boldsymbol{\theta})} \{MSE[\hat{g}(\boldsymbol{\theta})]\} \qquad s.t.\ \mathrm{E}[\hat{g}(\boldsymbol{\theta})] = g(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \tag{5.17}$$

The constraint in (5.17) corresponds to the condition for getting unbiased estimator. In [Barankin 1949], Barankin proposes a form to the solution of (5.17). If we denote by $\{\boldsymbol{\theta}_i\}_{i=1..N_t}$ N_t test points of $\boldsymbol{\theta}$ and by $L(\boldsymbol{y}; \boldsymbol{\theta}_i, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta}_i)/f(\boldsymbol{y}|\boldsymbol{\theta})$ the likelihood ratio of the test points vector $\boldsymbol{\theta}_i$ to the true parameters vector $\boldsymbol{\theta}$, then, the solution of (5.17) can be expressed [Barankin 1949]

$$\boldsymbol{BB} = \lim_{N_t \to \infty} \max_{\boldsymbol{a}, \{\theta_i\}_{i=1..N_t}} \frac{(\sum_{i=1}^{N_t} a_i (g(\boldsymbol{\theta}_i) - g(\boldsymbol{\theta})))^2}{\mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}} \left[(\sum_{i=1}^{N_t} a_i L(\boldsymbol{y}, \boldsymbol{\theta}_i, \boldsymbol{\theta}))^2 \right]} \tag{5.18}$$

Where $\boldsymbol{a} = [a_1, \cdots, a_{N_t}]^T$ is an N_t-dimensional real vector. We can notice form (5.18) that the problem of BB is that it does not have an analytical expression, which makes it impractical to use. Many works later tried to overcome this problem by deriving BB for specific applications such as nonlinear modulation systems in additive Gaussian noise background [McAulay & Seidman 1969], communication and radar problems [McAulay & Hofstetter 1971], source localization in ocean environment [Knockaert 1997], and the estimation of the frequency of a sinusoid in additive white Gaussian noise[Tabrikian & Krolik 1999].

5.4 Lower bounds for deterministic parameters estimation in the presence of nuisance signals

In the previous section, the bounds are introduced in a context where only two elements are involved, the observation data $\boldsymbol{y}$ and the deterministic parameters $\boldsymbol{\theta}$. Nevertheless, there are other scenarios where nuisance random variables $\boldsymbol{x}$ are involved too, and most of the time, the conditional pdf $f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$ is more tractable than the marginal likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$. For this type of scenarios, some bounds are developed. Before reviewing them, we shall introduce the joint vector $\boldsymbol{w} = [\boldsymbol{\theta}^T, \boldsymbol{x}^T]^T$ of the nuisance random vector $\boldsymbol{x}$ and the deterministic parameters $\boldsymbol{\theta}$. We also consider the conditional FIM given by

$$\begin{aligned} \boldsymbol{J}_c &\triangleq \mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}} \left\{ \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{w}} \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{w}^T} \right\} \\ &= \mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}} \left\{ -\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} \right\} \\ &= \begin{bmatrix} \boldsymbol{J}_{\boldsymbol{\theta}} & \boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}} \\ \boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}}^T & \boldsymbol{J}_{\boldsymbol{x}} \end{bmatrix} \end{aligned} \tag{5.19}$$

Where the matrices $\boldsymbol{J}_{\boldsymbol{\theta}}$, $\boldsymbol{J}_{\boldsymbol{x}}$ and $\boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}}$ are defined as

$$\begin{aligned} \boldsymbol{J}_{\boldsymbol{\theta}} &= \mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}} \left\{ -\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}, \quad \boldsymbol{J}_{\boldsymbol{x}} = \mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}} \left\{ -\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} \right\} \\ \boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}} &= \mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}} \left\{ -\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{x}^T} \right\} \end{aligned} \tag{5.20}$$

5.4.1 The modified Cramér-Rao bound

The MCRB was introduced first in [D'Andrea *et al.* 1994] for synchronization problem where the carrier frequency is estimated in the presence of unwanted random parameters $\boldsymbol{x}$ (data symbols, carrier phase and time epoch). An important constraint is that the *a priori* of the nuisance random variables must be independent of the target parameter θ. In [D'Andrea *et al.* 1994], the MCRB was derived when θ is scalar. This derivation was extended to the vectorial case in [Gini *et al.* 1998] where it takes the following form

$$\begin{aligned} MCRB &\triangleq \mathrm{E}_{\boldsymbol{x}} \left\{ \boldsymbol{J}_{\boldsymbol{\theta}} \right\}^{-1} \\ &= \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}} \left\{ \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right\}^{-1} \\ &= \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}} \left\{ -\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}^{-1} \end{aligned} \tag{5.21}$$

The derivation of the MCRB is based on the observation that

$$\begin{aligned} \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^T\right\} &= \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}}\left\{\mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}}\left\{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^T\right\}\right\} \\ &\geq \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}}\left\{\left(\mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}}\left\{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^T\right\}\right)^{-1}\right\} \\ &\geq \left(\mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}}\left\{\mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}}\left\{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^T\right\}\right\}\right)^{-1} \\ &= \left(\mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^T\right\}\right)^{-1} \end{aligned} \tag{5.22}$$

where $\widetilde{\boldsymbol{\theta}}$ denotes the estimation error $(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})$. Notice that here the estimator should be globally unbiased (in average over $\boldsymbol{x}$). The first inequality is an application of the CRB to the estimator $\widehat{\boldsymbol{\theta}}$ for a fixed $\boldsymbol{x}$, and the second inequality uses the Jensen's inequality and the convexity of the function $1/z$ for $z>0$. In [D'Andrea *et al.* 1994, Gini *et al.* 1998], authors prove that the MCRB is looser than the CRB ($MCRB \leq CRB$). We explain it briefly here. The proof makes use of some important relations by expressing the FIM and the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$ in a tricky way as it follows

$$\boldsymbol{J}(\boldsymbol{\theta}) = \int \frac{\partial f(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \frac{1}{f(\boldsymbol{y}|\boldsymbol{\theta})} d\boldsymbol{y} \tag{5.23}$$

$$f(\boldsymbol{y},\boldsymbol{\theta}) = \int f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})\, f(\boldsymbol{x})\, d\boldsymbol{x} \tag{5.24}$$

The derivation of (5.24) relative to $\boldsymbol{\theta}$ yields

$$\begin{aligned} \frac{\partial f(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \int \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \theta} f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\; d\boldsymbol{x} \\ &= \int \left[\frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \sqrt{f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}\right] \sqrt{f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}\, d\boldsymbol{x} \end{aligned} \tag{5.25}$$

By applying the Cauchy-Schwartz inequality to (5.25) and after some algebraic manipulations

$$\boldsymbol{J}(\boldsymbol{\theta}) \leq \int\int \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})\; d\boldsymbol{x}\, d\boldsymbol{y} \tag{5.26}$$

Notice that the independence of the *a priori* of $\boldsymbol{x}$ from $\boldsymbol{\theta}$ is important for the proof to hold. Though it is looser than the CRB, the MCRB is computationally more attractive in some cases where it can be derived in closed form when it is not the case for the CRB [Middleton 1972, Gini *et al.* 1998]. The equality in (5.26) is achieved when $\frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is independent of $\boldsymbol{x}$.

5.4.2 The Miller-Chang bound and its extension

The Miller-Chang bound (MCB) was introduced by Miller et al. in [Miller & Chang 1978] as an extension of the CRB. It is computed by first deriving the CRB for $\boldsymbol{\theta}$ while supposing the nuisance parameters exactly known, and

then averaging over all possible $\boldsymbol{x}$, in contrast to the MCRB where we average over $\boldsymbol{x}$ before inversion.

$$\begin{aligned} MCB &\triangleq \mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta}}\left\{\boldsymbol{J}_{\boldsymbol{\theta}}^{-1}\right\} \\ &= \mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta}}\left\{\frac{1}{\mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}}\left\{\frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right\}}\right\} \\ &= \mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta}}\left\{\frac{1}{\mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}}\right\} \end{aligned} \tag{5.27}$$

Unlike MCRB and CRB, the MCB is applied to locally unbiased estimators, i.e. $\widehat{\boldsymbol{\theta}}$ has to be unbiased for all values of $\boldsymbol{x}$. Moreover, there is no such constraint that the *a priori* of $\boldsymbol{x}$ must be independent of $\boldsymbol{\theta}$. In [Miller & Chang 1978], the MCB was defined for random parameters too ($\boldsymbol{\theta}$ random). It is a parallel definition to the nonrandom case, except that the conditional expectation $\mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}}\{.\}$ is replaced by $\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x}}\{.\}$

$$\begin{aligned} MCB' &\triangleq \mathrm{E}_{\boldsymbol{x}}\left\{\frac{1}{\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x}}\left\{\frac{\partial \ln f(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}}\frac{\partial \ln f(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}^T}\right\}}\right\} \\ &= \mathrm{E}_{\boldsymbol{x}}\left\{\frac{1}{\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x}}\left\{-\frac{\partial^2 \ln f(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{x})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}}\right\} \end{aligned} \tag{5.28}$$

In view of Jensen inequality, MCB is tighter than MCRB ($MCRB \leq MCB$) [Gini & Reggiannini 2000]. Applied to a restricted type of estimators, the MCB is tighter than the true CRB in some applications such as the arrival and separation time estimation of two interfering signals [Miller & Chang 1978]. An extension of MCB is proposed in [Gini & Reggiannini 2000] which consists in computing the CRB for joint estimation of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ considering that $\boldsymbol{x}$ is deterministic and unknown, and then averaging the result over the latter. Hence, the extended Miller-Chang bound (EMCB) is given by

$$\begin{aligned} EMCB &\triangleq \mathrm{E}_{\boldsymbol{x}}\left\{\left[\boldsymbol{J}_c^{-1}\right]_{11}\right\} \\ &= \mathrm{E}_{\boldsymbol{x}}\left\{\frac{1}{\boldsymbol{J}_{\boldsymbol{\theta}} - \boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}}\boldsymbol{J}_{\boldsymbol{x}}^{-1}\boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}}^T}\right\} \end{aligned} \tag{5.29}$$

where $\left[\boldsymbol{J}_c^{-1}\right]_{11}$ is the $L \times L$ upper block of matrix $\boldsymbol{J}_c^{-1}$. Notice that the EMCB is the joint Cramér-Rao bound (JCRB) averaged over $\boldsymbol{x}$ in [Moeneclaey 1998] ($EMCB = E_{\boldsymbol{x}}\{JCRB\}$).

5.4.3 The conditional Cramér-Rao bound

The conditional Cramér-Rao bound (CCRB) is defined for the joint vector $\boldsymbol{w}$ and when we ignore the *a priori* function of the nuisance vector $\boldsymbol{x}$ and consider it as non-random [Noam & Messer 2009]. Then, the covariance of any strict-sense unbiased estimator $\hat{\boldsymbol{w}}$ is lower bounded by

$$\begin{aligned} CCRB &\triangleq \boldsymbol{J}_c^{-1} \\ &= \mathrm{E}_{\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}} \left\{ \frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}\right)}{\partial \boldsymbol{w}} \frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}\right)}{\partial \boldsymbol{w}^T} \right\}^{-1} \end{aligned} \tag{5.30}$$

Hence, the CCRB for the $\boldsymbol{\theta}$ estimation part is deduced

$$CCRB_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left(\boldsymbol{J}_{\boldsymbol{\theta}} - \boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}} \boldsymbol{J}_{\boldsymbol{x}}^{-1} \boldsymbol{J}_{\boldsymbol{\theta},\boldsymbol{x}}^T\right)^{-1} \tag{5.31}$$

Notice that this bound is a function of both $\boldsymbol{x}$ and $\boldsymbol{\theta}$. This bound coincides with the JCRB, consequently its average over $\boldsymbol{x}$ results in the EMCB.

5.4.4 The asymptotic Cramér-Rao bound

The asymptotic Cramér-Rao bound (ACRB) is defined as the true CRB at high SNR. There is no general analytic expression of ACRB because its derivation depends tightly on the context (signal models, Noise model,etc.). Yet, the asymptotic bound derived in [Moeneclaey 1998] seems very interesting, since the observations model used therein is the most general so far and which is described as

$$\boldsymbol{y} = \boldsymbol{s}(\boldsymbol{x},\theta) + \boldsymbol{v} \tag{5.32}$$

Where $\boldsymbol{s}(.,.)$ is a general function of the scalar θ and the nuisance random vector $\boldsymbol{x}$ whose *a priori* function is independent of θ. $\boldsymbol{v}$ is the Gaussian noise vector of i.i.d. samples with variance σ_v^2. To derive the ACRB, Moeneclaey uses a tricky expression of the FIM $J(\theta)$ (for θ scalar) [Moeneclaey 1998] and which consists in

$$\begin{aligned} \mathrm{E}_{\boldsymbol{y}|\theta} \left\{ -\frac{\partial^2 \ln f\left(\boldsymbol{y}|\theta\right)}{\partial \theta^2} \right\} = -1/2\sigma_v^2 \ \mathrm{E}_{\boldsymbol{y}|\theta} \left\{ \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\theta} \left\{ \frac{\partial^2 \left|\boldsymbol{y} - \boldsymbol{s}\left(\boldsymbol{x},\theta\right)\right|^2}{\partial \theta^2} \right\} \right\} \\ - 1/2\sigma_v^2 \ \mathrm{E}_{\boldsymbol{y}|\theta} \left\{ var_{\boldsymbol{x}|\boldsymbol{y},\theta} \left(\frac{\partial \left|\boldsymbol{y} - \boldsymbol{s}\left(\boldsymbol{x},\theta\right)\right|^2}{\partial \theta} \right) \right\} \end{aligned} \tag{5.33}$$

where $var_{\boldsymbol{x}|\boldsymbol{y},\theta}(.)$ denotes the variance using the pdf $f(\boldsymbol{x}|\boldsymbol{y},\theta)$. After mathematical manipulations, Moeneclaey derives approximate expressions to each term of the RHS in (5.33) at high SNR and shows that

$$\begin{aligned} ACRB &\cong \frac{1}{\mathrm{E}_{\boldsymbol{x}} \left\{ \left[\boldsymbol{J}_c^{-1}\right]_{11} \right\}} \\ &= \frac{1}{\mathrm{E}_{\boldsymbol{x}} \left\{ \boldsymbol{J}_{\theta} - \boldsymbol{J}_{\theta,\boldsymbol{x}} \boldsymbol{J}_{\boldsymbol{x}}^{-1} \boldsymbol{J}_{\theta,\boldsymbol{x}}^T \right\}} \end{aligned} \tag{5.34}$$

Recall that all these results are for the case when the *a priori* function of $\boldsymbol{x}$ is independent of θ. If this condition is not fulfilled, the result will change. In [Gini & Reggiannini 2000], important inequalities are derived to compare some of the lower bounds seen before. These inequalities are derived for a specified application (carrier frequency offset estimation) but can be generalized. We summarize them in the following

$$EMCB \geq MCB \geq MCRB \tag{5.35}$$

$$EMCB \geq ACRB \geq MCRB \tag{5.36}$$

5.5 Lower bounds for Bayesian estimators

Bayesian estimation is used when the target parameters vector $\boldsymbol{\theta}$ is random with an *a priori* function denoted by $f(\boldsymbol{\theta})$. The optimal Bayesian estimator of $\boldsymbol{\theta}$ is the MMSE estimator [Kay 1993] (optimal on the MSE sense) defined as the minimizer of the MSE function.

$$\widehat{\boldsymbol{\theta}}_{MMSE} \triangleq \arg\min_{\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}} \left\{ \widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}^T \right\} \tag{5.37}$$

The solution of the optimization problem in (5.37) consists in the *a posteriori* mean $\mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\{\boldsymbol{\theta}\}$. Nevertheless, the *a posteriori* mean can be intractable in some cases. An alternative sub-optimal solution is the MAP estimator [Kay 1993]. It is based on maximizing the *a posteriori* distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$.

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{MAP} &\triangleq \arg\max_{\boldsymbol{\theta}} \left(\ln f(\boldsymbol{\theta}|\boldsymbol{y})\right) \\ &= \arg\max_{\boldsymbol{\theta}} \left(\ln f(\boldsymbol{y}|\boldsymbol{\theta}) + \ln f(\boldsymbol{\theta})\right) \end{aligned} \tag{5.38}$$

In asymptotic conditions (infinite number of samples or high SNR), the contribution of the *a priori* function becomes insignificant, and the MAP solution converges to the ML estimation solution. In a Bayesian context, the classical bounds cited previously are no longer appropriate since they do not take into consideration the *a priori* information, whence the need to invent new benchmarks or Bayesian lower bounds. In the following, we review this type of bounds. It is noteworthy that the bias condition in the Bayesian case is different than the one in the deterministic case in (5.10). In the Bayesian case the bias is defined as

$$\boldsymbol{b} = \mathrm{E}_{\boldsymbol{\theta}} \left\{ \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}} \left\{ \widehat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta} \right\} \right\} \tag{5.39}$$

In the deterministic case, the bias statement is a local condition where the estimator has to be unbiased for each value of $\boldsymbol{\theta}$ (strictly unbiased), whereas in the Bayesian case, the bias statement is a global condition where the estimator has to be unbiased in average (widely unbiased).

5.5.1 The Bayesian Cramér-Rao bound

The Bayesian Cramér-Rao bound (BCRB) was introduced first by Van Trees [Van Trees 1968] as a lower bound to the variance of estimation error when $\boldsymbol{\theta}$ is random. The bias condition of BCRB is that it is applied for any wide sense unbiased estimator $\widehat{\boldsymbol{\theta}}$. The BCRB is formulated as it follows

$$\begin{aligned} BCRB &\triangleq \mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\left\{\frac{\partial \ln f\left(\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\frac{\partial \ln f\left(\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right\}^{-1} \\ &= \mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f\left(\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}^{-1} \\ &= \boldsymbol{J}_B^{-1} \end{aligned} \tag{5.40}$$

where $\boldsymbol{J}_B$ is the Bayesian information matrix (BIM) which can be decomposed into two informative quantities

$$\begin{aligned} \boldsymbol{J}_B &= \mathrm{E}_{\boldsymbol{y}}\left\{\mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}\right\} + \mathrm{E}_{\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f\left(\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\} \\ &= \mathrm{E}_{\boldsymbol{y}}\left\{\boldsymbol{J}(\boldsymbol{\theta})\right\} + \boldsymbol{J}_p \end{aligned} \tag{5.41}$$

where the first term consists in the information brought by the observed data and the second term is the information brought from the *a priori* function of $\boldsymbol{\theta}$. Notice that the BIM is not a function of $\boldsymbol{\theta}$ but constant. In [Van Trees 2001], Van Trees proves that when $f(\boldsymbol{\theta}|\boldsymbol{y})$ satisfies a multivariate Gaussian density, the BCRB can be achieved by $\widehat{\boldsymbol{\theta}}_{MAP}$ that coincides with $\widehat{\boldsymbol{\theta}}_{MMSE}$ in this case.

5.5.2 The Bobrovsky-Zakai bound

The Bobrovsky-Zakai Bound was introduced for the first time in the context of nonlinear diffusion process filtering [Bobrovsky & Zakai 1976]. The bound was meant to assess the suboptimal implementable filters relative to the intractable optimal solution (optimal filter). The Bobrovsky-Zakai Bound is defined as the inverse of the $L \times L$ matrix $\boldsymbol{J}(\delta)$ whose (i,j) element is defined as

$$J_{i,j}(\delta) = \mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\left\{\left(\frac{f\left(\boldsymbol{y},\boldsymbol{\theta}\right) - f\left(\boldsymbol{y},\boldsymbol{\theta}+\delta\,\boldsymbol{e}_i\right)}{\delta\, f\left(\boldsymbol{y},\boldsymbol{\theta}\right)}\right)\left(\frac{f\left(\boldsymbol{y},\boldsymbol{\theta}\right) - f\left(\boldsymbol{y},\boldsymbol{\theta}+\delta\,\boldsymbol{e}_j\right)}{\delta\, f\left(\boldsymbol{y},\boldsymbol{\theta}\right)}\right)\right\} \tag{5.42}$$

where $\boldsymbol{e}_i$ is the i^{th} vector of the standard basis in $\mathbb{R}^L$, and δ is a real number.

5.5.3 The Weiss-Weinstein bound

In [Weiss & Weinstein 1985, Weinstein & Weiss 1988], Weiss et al. prove that for any arbitrary scalar functions $g(\boldsymbol{y})$ and $k(\boldsymbol{\theta})$, the following inequality is true

$$\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\left\{\left(k(\boldsymbol{\theta}) - g(\boldsymbol{y})\right)^2\right\} \geq \frac{(\boldsymbol{a}^T\boldsymbol{u})^2}{\boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a}} \tag{5.43}$$

where $\boldsymbol{a}$ is an arbitrary N_t-dimensional vector, N_t is the number of test points and $\boldsymbol{u}$ is the N_t-dimensional vector whose i^{th} element is given by

$$u_i = \mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\big\{[k(\boldsymbol{\theta}-\boldsymbol{h}_i)-k(\boldsymbol{\theta})]L^{1-s_i}(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{h}_i,\boldsymbol{\theta})\big\} \tag{5.44}$$

$\{\boldsymbol{h}_i\}_{i=1\cdots N_t}$ are arbitrary N_t-dimensional vectors, $\{s_i\}_{i=1\cdots N_t}$ are arbitrary scalars in $[0,1]$ and $L(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{h}_i,\boldsymbol{\theta})$ is given by

$$L(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{h}_i,\boldsymbol{\theta}) \triangleq f(\boldsymbol{y},\boldsymbol{\theta}-\boldsymbol{h}_i)/f(\boldsymbol{y},\boldsymbol{\theta}) \tag{5.45}$$

$\boldsymbol{V}$ is the matrix whose (i,j) element is given by

$$V_{ij} = \mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\Big\{ \big[L^{s_i}(\boldsymbol{y};\boldsymbol{\theta}+\boldsymbol{h}_i,\boldsymbol{\theta})-L^{1-s_i}(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{h}_i,\boldsymbol{\theta})\big] \big[L^{s_j}(\boldsymbol{y};\boldsymbol{\theta}+\boldsymbol{h}_j,\boldsymbol{\theta})-L^{1-s_j}(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{h}_j,\boldsymbol{\theta})\big] \Big\} \tag{5.46}$$

After maximization of the bound relative to the arbitrary vector $\boldsymbol{a}$ and substitution of $k(\boldsymbol{\theta})=\boldsymbol{l}^T\boldsymbol{\theta}$ and $g(\boldsymbol{y})=\boldsymbol{l}^T\widehat{\boldsymbol{\theta}}$ with $\boldsymbol{l}$ arbitrary vector, the final lower bound on the error covariance matrix is

$$\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\big\{(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^T\big\} \geq \boldsymbol{H}\boldsymbol{Q}^{-1}\boldsymbol{H}^T \tag{5.47}$$

where $\boldsymbol{H}$ is the matrix whose columns are the vectors $\{\boldsymbol{h}_i\}_{i=1\cdots N_t}$ and $\boldsymbol{Q}$ is the matrix whose (i,j) element is given by

$$Q_{ij} = V_{ij}/\big(\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\big\{L^{1-s_i}(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{h}_i,\boldsymbol{\theta})\big\}\,\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\big\{L^{1-s_j}(\boldsymbol{y};\boldsymbol{\theta}-\boldsymbol{h}_j,\boldsymbol{\theta})\big\}\big) \tag{5.48}$$

Notice that the BCRB is a limiting case for Weiss-Weinstein bound ($\boldsymbol{H}=h\,\boldsymbol{I}$, $h\to 0$) provided that more regularity conditions are satisfied (integrability of the Hessian of the joint log-likelihood). Hence, the Weiss-Weinstein bound presents a good alternative when the regularity conditions are not fulfilled. Moreover, it is possible to have a bound even tighter than the CRB if the arbitrary parameters are chosen properly. The Bobrovsky-Zakai bound is also a special case of Weiss-Weinstein bound ($s_i=1$) that needs more regularity conditions. In [Weinstein & Weiss 1988], the authors generalize more (5.43) by formulating lower bounds for higher order moments of the parameters estimation error $\mathrm{E}_{\boldsymbol{y},\boldsymbol{\theta}}\big\{(k(\boldsymbol{\theta})-g(\boldsymbol{y}))^k\big\}$ with $k>1$.

5.6 Lower bound for hybrid estimators: the hybrid Cramér-Rao bound

The need for the HCRB was expressed first in the work of Rockah and Schultheiss [Rockah & Schultheiss 1987] where they study the source localization, but restricted to the case where the *a priori* pdf of $\boldsymbol{x}$ is independent of $\boldsymbol{\theta}$. Ten years later, Reuven

and Messer generalize its formulation in [Reuven & Messer 1997] where they extend the BB to the case of a hybrid vector estimation. The HCRB is applied to wide-sense unbiased hybrid estimators. A well known example of hybrid estimator is the MAP/ML estimator defined as

$$\hat{\boldsymbol{w}}_{ML/MAP} = \arg\max_{\boldsymbol{\theta},\boldsymbol{x}} \ln f\left(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}\right) \tag{5.49}$$

It is important to mention that among the regularity conditions in [Reuven & Messer 1997], the statistical independence of the random and nonrandom parameters is required. This constraint is relaxed after in [Bay *et al.* 2008]. Let us define the hybrid information matrix (HIM) as

$$\begin{aligned}\tilde{\boldsymbol{J}} &\triangleq \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{\frac{\partial \ln f\left(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{w}}\frac{\partial \ln f\left(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{w}^T}\right\} \\ &= \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f\left(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{w}\partial \boldsymbol{w}^T}\right\} \\ &= \begin{bmatrix}\tilde{\boldsymbol{J}}_{\boldsymbol{\theta}} & \tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}} \\ \tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}}^T & \tilde{\boldsymbol{J}}_{\boldsymbol{x}}\end{bmatrix}\end{aligned} \tag{5.50}$$

where the matrices $\tilde{\boldsymbol{J}}_{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{J}}_{\boldsymbol{x}}$ and $\tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}}$ are defined as it follows

$$\begin{aligned}\tilde{\boldsymbol{J}}_{\boldsymbol{\theta}} &= \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}, \; \tilde{\boldsymbol{J}}_{\boldsymbol{x}} = \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^T}\right\} \\ \tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}} &= \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{-\frac{\partial^2 \ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{x}^T}\right\}\end{aligned} \tag{5.51}$$

The HCRB for $\boldsymbol{\theta}$ estimation part is defined as the $L \times L$ upper block of matrix $\tilde{\boldsymbol{J}}^{-}1$ [Noam & Messer 2009] computed using the matrix inversion lemma (see Appendix B.2).

$$HCRB = \left(\tilde{\boldsymbol{J}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}}^T\right)^{-1} \tag{5.52}$$

In [Noam & Messer 2009], the authors prove that the CRB is tighter than the HCRB and provide a necessary and sufficient condition for them to be equal without the need to compute the CRB explicitly. They also provide a necessary and sufficient condition for HCRB to be tight and prove that when it is the case, HCRB is achieved by the hybrid estimator MAP/ML. Yet, in general the HCRB is not asymptotically tight unless it coincides with the CRB. In [Gini & Reggiannini 2000], authors prove that HCRB is tighter than MCRB with equality when $\tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}} = \boldsymbol{0}$. The different comparison are summarized in (5.53)

$$CRB \geq HCRB \geq MCRB \tag{5.53}$$

It is noteworthy that the inequalities where MCRB is compared to CRB and HCRB were derived under the constraint of statistical independence between $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

5.7 Characterizing the differences (HIM-FIM) and (FIM-MFIM)

In this section, we investigate the differences (HIM-FIM) and (FIM-MFIM) in an attempt to understand the effect of the random variables $\boldsymbol{x}$ on the estimation of $\boldsymbol{\theta}$.

5.7.1 Difference between HIM and FIM

The relation between the joint pdf $f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})$ and the marginalized pdf $f(\boldsymbol{y}|\boldsymbol{\theta})$ can be simply expressed using Bayes rule in (5.54)

$$\ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) = \ln f(\boldsymbol{y}|\boldsymbol{\theta}) + \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \tag{5.54}$$

we compute the Hessian of both sides in (5.54) w.r.t. $\boldsymbol{w}$ and after applying the joint expectation $\mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\{.\}$, we can express the HIM using the FIM

$$\begin{aligned}\begin{bmatrix} \tilde{\boldsymbol{J}}_{\boldsymbol{\theta}} & \tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}} \\ \tilde{\boldsymbol{J}}^T_{\boldsymbol{\theta},\boldsymbol{x}} & \tilde{\boldsymbol{J}}_{\boldsymbol{x}} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{J}(\boldsymbol{\theta}) & 0 \\ 0 & 0 \end{bmatrix} + \tilde{\boldsymbol{G}} \\ &= \begin{bmatrix} \boldsymbol{J}(\boldsymbol{\theta}) & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \tilde{\boldsymbol{G}}_{\boldsymbol{\theta}} & \tilde{\boldsymbol{G}}_{\boldsymbol{\theta},\boldsymbol{x}} \\ \tilde{\boldsymbol{G}}^T_{\boldsymbol{\theta},\boldsymbol{x}} & \tilde{\boldsymbol{G}}_{\boldsymbol{x}} \end{bmatrix}\end{aligned} \tag{5.55}$$

where $\tilde{\boldsymbol{G}} = -\mathrm{E}_{\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta}}\left\{\frac{\partial^2 \ln f(\boldsymbol{x}|\boldsymbol{y},\theta)}{\partial \boldsymbol{w}\,\partial \boldsymbol{w}^T}\right\}$ is the block matrix partitioned in a similar way as $\tilde{\boldsymbol{J}}$ in (5.50). From (5.55), we deduce the relation between HCRB^{-1} and CRB^{-1} :

$$\begin{aligned} HCRB^{-1} &= \tilde{\boldsymbol{J}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{J}}_{\boldsymbol{\theta},\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}^T_{\boldsymbol{\theta},\boldsymbol{x}} \\ &= \boldsymbol{J}(\boldsymbol{\theta}) + \tilde{\boldsymbol{G}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{G}}_{\boldsymbol{\theta},\boldsymbol{x}}\tilde{\boldsymbol{G}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{G}}^T_{\boldsymbol{\theta},\boldsymbol{x}} \\ &= CRB^{-1} + \tilde{\boldsymbol{G}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{G}}_{\boldsymbol{\theta},\boldsymbol{x}}\tilde{\boldsymbol{G}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{G}}^T_{\boldsymbol{\theta},\boldsymbol{x}} \end{aligned} \tag{5.56}$$

This expression is valid for any distribution and shows that the inverse of HCRB (information in the presence of nuisance parameters) for $\boldsymbol{\theta}$ equals the inverse marginal/separate CRB plus an inverse CRB that would correspond to joint estimation from the posterior density $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$. It is noteworthy that this result can be also derived using the terminology of orthogonal projection used in [Noam & Messer 2009]. The second method of derivation is detailed in Appendix A.1. For the joint Gaussian model presented in 5.2 and using the FIM expression in (5.13), the HCRB inverse is expressed as

$$\begin{aligned}\tilde{\boldsymbol{J}}_{\boldsymbol{\theta}} = &\frac{1}{2}\frac{\partial \boldsymbol{C}^T_{\boldsymbol{yy},\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}(\boldsymbol{C}^{-1}_{\boldsymbol{yy}}(\boldsymbol{\theta}) \otimes \boldsymbol{C}^{-1}_{\boldsymbol{yy}}(\boldsymbol{\theta}))\frac{\partial \boldsymbol{C}_{\boldsymbol{yy},\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}^T} + \frac{\partial \boldsymbol{\mu}^T_{\boldsymbol{yy}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\boldsymbol{C}^{-1}_{\boldsymbol{yy}}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\mu}_{\boldsymbol{yy}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \\ &+ \frac{1}{2}\frac{\partial \boldsymbol{P}^T_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(\boldsymbol{P}^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{P}^{-1}(\boldsymbol{\theta})\right)\frac{\partial \boldsymbol{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} + \frac{\partial \boldsymbol{F}^T_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}\left(\boldsymbol{P}^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{C}_{\boldsymbol{yy}}(\boldsymbol{\theta})\right)\frac{\partial \boldsymbol{F}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}^T}\end{aligned} \tag{5.57}$$

where $\boldsymbol{P}_{\boldsymbol{\theta}}$ and $\boldsymbol{F}_{\boldsymbol{\theta}}$ are the vectorized forms of $\boldsymbol{P}(\boldsymbol{\theta})$ and $\boldsymbol{F}(\boldsymbol{\theta})$ respectively, the last two terms correspond to the difference in inverse CRB, and correspond to the

information for $\boldsymbol{\theta}$ that can be extracted from the covariance and the mean of the Gaussian posterior $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$.

5.7.2 The Difference between FIM and MFIM

In [Moeneclaey 1998], Moeneclaey computes $(MFIM - FIM)$ for the specific case when $f(\boldsymbol{x}|\theta) = f(\boldsymbol{x})$, θ is scalar and the observation noise is white Gaussian with covariance matrix $\boldsymbol{R}_{vv} = \sigma_v^2 \boldsymbol{I}$ independent of $\boldsymbol{\theta}$. The extension of his result to vectorial $\boldsymbol{\theta}$ and general observation noise covariance matrix $\boldsymbol{R}_{vv}$ (but independent of $\boldsymbol{\theta}$) is straightforward and can be written as follows

$$\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{J}_M(\boldsymbol{\theta}) - \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}} \left\{ Cov_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}} \left\{ \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\} \right\} \tag{5.58}$$

where $\boldsymbol{J}_M(\boldsymbol{\theta}) = -\mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}} \left\{ \frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}$ denotes the modified FIM (MFIM) (the inverse of MCRB). Here we shall extend this to the case of $f(\boldsymbol{x}|\boldsymbol{\theta})$. We claim the following result

$$\boldsymbol{J}(\boldsymbol{\theta}) = \boldsymbol{J}_M(\boldsymbol{\theta}) - \tilde{\boldsymbol{G}}_{\boldsymbol{\theta}} + \boldsymbol{J}_{\boldsymbol{x}}(\boldsymbol{\theta}) \tag{5.59}$$

where $\boldsymbol{J}_{\boldsymbol{x}}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{x}|\boldsymbol{\theta}} \left\{ -\frac{\partial^2 \ln f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}$.

Proof. There is a second way to express the joint likelihood in addition to the fashion in (5.54),

$$\ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) = \ln f(\boldsymbol{y}|\boldsymbol{\theta}) + \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \tag{5.60}$$

the Hessian of the two expressions in (5.54) and (5.60) w.r.t. $\boldsymbol{\theta}$ results in

$$\frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \frac{\partial^2 \ln f(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{\partial^2 \ln f(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} + \frac{\partial^2 \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

Applying the $\mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\{.\}$ operator over all random variables and changing the terms to the right sides results in the claimed result. □

When $f(\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{x})$, the last term in the RHS of (5.59) $\boldsymbol{J}_{\boldsymbol{x}}(\boldsymbol{\theta})$ vanishes. The second term in the RHS of (5.59) $\tilde{\boldsymbol{G}}_{\boldsymbol{\theta}}$ can be easily proved equal to the covariance term in (5.58) when we notice that $\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}} \left\{ \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}$ is simply equal to $\frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, and therefore get Moeneclaey's result in (5.58). In terms of interpretation, $\tilde{\boldsymbol{G}}_{\boldsymbol{\theta}}$ may be interpreted as the difference in information between $\boldsymbol{x}$ being deterministic or random. $\boldsymbol{J}_{\boldsymbol{x}}(\boldsymbol{\theta})$ is new and corresponds to the information on $\boldsymbol{\theta}$ in the prior distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$.

5.8 Performance analysis: back to basics

If we summarize the two main estimation scenarios raised in previous sections, given the measurements vector $\boldsymbol{y}$, the latent random (Gaussian) vector $\boldsymbol{x}$ which is considered according to the application to be either nuisance or target parameters, and

the deterministic parameters vector $\boldsymbol{\theta}$. In a *joint* estimation approach, we maximize the joint likelihood $f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta})$ which means alternating the MAP/ML, MAP for $\boldsymbol{x}$ and ML for $\boldsymbol{\theta}$ (the estimator is denoted $\widehat{\boldsymbol{\theta}}^J_{ML}$), with (joint) error covariance matrix $\boldsymbol{C}^J_{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}}$ and performance lower bound HCRB. In a *marginalized* estimation approach, the random variables $\boldsymbol{x}$ can be eliminated (marginalized) from the likelihood, and then the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$ maximized, which means ML for $\boldsymbol{\theta}$ (the estimator is denoted $\widehat{\boldsymbol{\theta}}^M_{ML}$), with (marginalized) ML error covariance matrix $\boldsymbol{C}^M_{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}}$ and associated lower bound CRB. Asymptotically (in infinite data samples sense), we get the performance ordering as follows

$$\boldsymbol{C}^J_{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}} \overset{(i)}{\geq} \boldsymbol{C}^M_{\widetilde{\boldsymbol{\theta}}\widetilde{\boldsymbol{\theta}}} \overset{(ii)}{=} \text{CRB} \overset{(iii)}{\geq} \text{HCRB} \tag{5.61}$$

where (i) is due to the inconsistency of $\hat{\boldsymbol{x}}_{MAP}$ which prevents $\widehat{\boldsymbol{\theta}}^J_{ML}$ from reaching its CRB, (ii) is due to $\widehat{\boldsymbol{\theta}}^M_{ML}$ being consistent, (iii) is proved in [Noam & Messer 2009]. In other words, in terms of actual performance, joint estimation of the state and the parameters leads to worse parameter estimates than when the parameters are estimated in a marginalized fashion, even though the CRBs would indicate otherwise. In [Yeredor 2000], Yeredor proves that the MAP/ML induces a persisting bias even asymptotically.

5.9 Conclusion

In this chapter, we tackled the issue of *joint* vs *marginal* estimation. We reviewed some of the lower bounds used as benchmarks to assess estimators of both types. We then investigated the effect of the presence of random (nuisance) variables on the estimation of deterministic parameters in terms of lower bounds and in comparison to the case when these random variables are eliminated.

Different scenarios of joint estimations are present. One scenario is illustrated by the alternating MAP/ML Kalman filter where the estimate $\widehat{\boldsymbol{\theta}}$ is computed using only $\hat{\boldsymbol{x}}$. Idem for $\boldsymbol{x}$, $\hat{\boldsymbol{x}}$ is computed using only $\widehat{\boldsymbol{\theta}}$. This algorithm converges to the joint MAP/ML solution. Another scenario is illustrated with the EM-Kalman seen in chapter 3, the estimation of $\widehat{\boldsymbol{\theta}}$ is computed differently by using both $\hat{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{x}}$, thus, the estimation error of $\boldsymbol{x}$ is considered and expected to improve the estimation of $\boldsymbol{\theta}$. In fact, it is known that the EM approach converges to the marginalized ML approach, so the EM-Kalman algorithm would be one approach to get this optimal performance. the estimation fashion of $\hat{\boldsymbol{x}}$ is unchanged. The last scenario is illustrated with the variational Bayes (VB)-Kalman filter where $\widehat{\boldsymbol{\theta}}$ is computed the same way as in EM-Kalman algorithm, and the estimation of $\hat{\boldsymbol{x}}$ is improved by using both $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$. This raises the question about a potential improvement of the VB-Kalman estimate of $\boldsymbol{\theta}$ compared to the EM-Kalman estimate.

Chapter 6

On the performance of joint LMMSE filtering and parameter estimation

6.1 Introduction

In chapter 5, we characterized the difference between the hybrid information matrix (HIM) (the deterministic parameter part) and the classical Fisher information matrix (FIM) in order to understand the influence of the random signals on the estimation of deterministic parameter in terms of lower bounds. In this chapter, we focus on the joint and separate estimation in terms of performances. In estimation theory, the choice of an estimator depends closely on the context of the problem. When the unknown parameters are deterministic, the maximum likelihood (ML) estimator is often considered as the best approach. It is typically consistent and asymptotically optimal (attaining the Cramér-Rao bound (CRB))[Van Trees 2001, Wald 1949]. For the random case, the minimum mean square error (MMSE) estimator is used and known (in the Gaussian case) to achieve the Bayesian Cramér-Rao bound (BCRB) introduced by Van Trees [Van Trees & Bell 2007]. When the MMSE estimate is intractable, it is sometimes replaced by the maximum a posteriori (MAP) estimator. An other important estimation problem to be considered is when nuisance (random) parameters are affecting the estimation of the unknown (deterministic) parameters/signals as in synchronization problems [Lindsey 1972] or audio source separation [Bensaid *et al.* 2010b]. Different scenarios have been considered. One scenario is to marginalize out the nuisance parameters which yields the previous problem of ML estimation. In some cases, the marginalization is intractable or very tedious, so we resort to joint estimation (MAP/ML, expectation maximization (EM), variational Bayes (VB)...)[Dempster *et al.* 1977, Beal 2003, Yeredor 2000], which is also relevant when the random signals are of interest [Guarnieri & Tebaldini 2007, Tichavsky & Wong 2004]. One instance of linear MMSE (LMMSE) estimation is the Kalman filter. In the literature, variations of the Kalman filter have been derived to handle the problem of joint filtering and parameter estimation. In this chapter, we start with a review of the different derivatives of Kalman filter in

the joint filtering/estimation framework. We then analyze the performance of the iterative algorithm VB (compared to EM) and introduce second order extended LMMSE (SOELMMSE). Finally, we conclude.

6.2 Adaptive Kalman filtering approaches

Since Rudolf E. Kalman published his famous paper [Kalman 1960], Kalman filter has been the work horse of many processes in different domains (position and speed tracking, optimal parameters estimation...). Yet, the filter may depend on unknown parameters whose knowledge is crucial to perform good estimation. This problem arises in many contexts such as these three examples :

- Bayesian adaptive filtering [Sadiki & Slock 2004] (or wireless channel estimation [Lenardi & Slock 2002, Gao *et al.* 2003]) where the state vector represents the finite impulse response (FIR) filter and the parameters are the power delay profile, autoregressive (AR)(1) dynamics.

- Position tracking (GPS) [Consortium Partners 2010] where the state vector represents acceleration, velocity and position variables, and the parameters are of the acceleration model (e.g. white noise, AR(1))

- Blind audio source separation where the state vector represents source signals and parameters are the short term prediction (STP) and long term prediction (LTP) AR parameters [Bensaid *et al.* 2010a].

In the literature, new varieties of Kalman are derived to overcome this problem like the widely used EM-Kalman algorithm ([Couvreur & Bresler 1995, Gao *et al.* 2003, Feder & Weinstein 1988]) which uses the famous EM technique in order to estimate the unknown parameters and then update them in the equations of Kalman filter to estimate the hidden state in adaptive scheme. Another well-known variety is the extended Kalman filter (EKF) algorithm. In this case, nonlinear models are tackled and the state is extended with the unknown parameters in order to be estimated in parallel. A third derivation is the truncated second order extended Kalman filter (SOEKF) introduced by [Bass *et al.* 1966, Jazwinski 1970] in which nonlinearities are carried to second order, third and higher order statistics are neglected. A corrected derivation of this filter is presented in [Henriksen 1982]. In ([Jazwinski 1970, Athans *et al.* 1968]), the Gaussian SOEKF is derived so that fourth-order terms in Taylor series approximations are retained and approximated by assuming that the underlying probabilities are Gaussian. In [Villares & Vazquez 2004], Villares et al. introduced the quadratic extended Kalman filter (QEKF) where they extend the EKF to a new algorithm using quadratic processing and incorporating fourth order

statistics of the input signal. The problem of uncertainty about the process noise and measurement noise covariance matrices was also addressed in [Mehra 1970] where a test of Kalman filter optimality is used in order to determine the estimation of the unknown noise matrices. The performance of some Kalman filter approaches was studied in literature. In [Dempster *et al.* 1977], the EM-Kalman is proved to converge to the ML performance. The asymptotic behavior of the EKF when depending of unknown parameter is treated in [Ljung 1979] where it is proved that no global convergence is guaranteed. Performance analysis of linear and nonlinear Kalman filters are also treated in terms of CRB computation. In [Tichavsky *et al.* 1998], the BCRB is developed for the discrete nonlinear Kalman filter. Recursive BCRBs were also developed for continuous and discrete nonlinear Kalman filters for many problems. In order to have a global overview, the interested reader may refer to [Trees & Bell 2007].

Before reviewing the different Kalman derivatives, we shall present the general state space model. Let us denote by $\boldsymbol{y}$ the N-dimensional observations vector and by $\boldsymbol{x}$ the M-dimensional AR(1) state vector, then

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{F}_k\, \boldsymbol{x}_k + \boldsymbol{G}_k\, \boldsymbol{w}_k \\ \boldsymbol{y}_k &= \boldsymbol{H}_k\, \boldsymbol{x}_k + \boldsymbol{v}_k \end{aligned} \tag{6.1}$$

for discrete time $k = 1, 2, ...$ and where the initial state $\boldsymbol{x}_0$ is Gaussian distributed with mean $\hat{\boldsymbol{x}}_0$ and covariance $\boldsymbol{P}_0$, the state noise $\boldsymbol{w}_k$ and measurement noise $\boldsymbol{v}_k$ are zero-mean Gaussian processes with covariances $\boldsymbol{Q}_k$ and $\boldsymbol{R}_k$ respectively. All these random quantities are mutually uncorrelated.

6.2.1 Basic Kalman filter

In the following, we introduce the compact notation $\boldsymbol{y}_{1:k} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k\}$ of the k successive samples of $\boldsymbol{y}$. Kalman filter performs Gram-Schmidt orthogonalization of the measurement variables $\boldsymbol{y}_k$. This is done by computing the LMMSE predictor $\hat{\boldsymbol{y}}_{k|k-1}$ of $\boldsymbol{y}_k$ on the basis of $\boldsymbol{y}_{1:k-1}$, leading to the orthogonalized prediction error (or innovation) $\widetilde{\boldsymbol{y}}_k = \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1}$. We introduce the correlation matrix notation $\boldsymbol{C}_{\boldsymbol{x}\boldsymbol{y}} = \mathrm{E}\left\{\boldsymbol{x}\,\boldsymbol{y}^T\right\}$ (correlation matrices will usually also be covariance matrices here since the processes $\boldsymbol{y}_k$ and $\boldsymbol{x}_k$ have zero-mean and also various estimation errors will have (conditional) zero-mean). We denote the covariance matrix $\boldsymbol{C}_{\widetilde{\boldsymbol{y}}_k\widetilde{\boldsymbol{y}}_k} = \boldsymbol{S}_k$. The idea of the innovations approach is that (linear) estimation in terms of $\boldsymbol{y}_{1:k}$ is equivalent to estimation in terms of $\widetilde{\boldsymbol{y}}_{1:k}$ since one set is obtained from the other by an invertible linear transformation. Now, since the $\widetilde{\boldsymbol{y}}_k$ are uncorrelated, estimation

in terms of $\widetilde{\boldsymbol{y}}_{1:k}$ is simplified:

$$\begin{aligned}\hat{\boldsymbol{x}}_{|k} &= \sum_{i=1}^{k} \boldsymbol{C}_{\boldsymbol{x}\widetilde{\boldsymbol{y}}_i}\, \boldsymbol{C}^{-1}_{\widetilde{\boldsymbol{y}}_i\widetilde{\boldsymbol{y}}_i}\, \widetilde{\boldsymbol{y}}_i \\ &= \hat{\boldsymbol{x}}_{|k-1} + \boldsymbol{C}_{\boldsymbol{x}\widetilde{\boldsymbol{y}}_k}\, \boldsymbol{S}_k^{-1}\, \widetilde{\boldsymbol{y}}_k \end{aligned} \tag{6.2}$$

This will be used to obtain *predicted* estimates $\hat{\boldsymbol{x}}_{k|k-1}$ with estimation error $\widetilde{\boldsymbol{x}}_{k|k-1} = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k-1}$ and covariance matrix $\boldsymbol{P}_{k|k-1} = \boldsymbol{C}_{\widetilde{\boldsymbol{x}}_{k|k-1}\widetilde{\boldsymbol{x}}_{k|k-1}}$ and also *filtered* estimates $\hat{\boldsymbol{x}}_{k|k}$ with estimation error $\widetilde{\boldsymbol{x}}_{k|k} = \boldsymbol{x}_k - \hat{\boldsymbol{x}}_{k|k}$ with covariance matrix $\boldsymbol{P}_{k|k} = C_{\widetilde{\boldsymbol{x}}_{k|k}\widetilde{\boldsymbol{x}}_{k|k}}$. Now exploiting the correlation structure in the signal model, this leads to the following two-step recursive procedure to go from $|k{-}1$ to $|k$:

Measurement Update

$$\hat{\boldsymbol{y}}_{k|k-1} = \boldsymbol{H}_k\, \hat{\boldsymbol{x}}_{k|k-1} \tag{6.3}$$

$$\widetilde{\boldsymbol{y}}_k = \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1} \tag{6.4}$$

$$\boldsymbol{S}_k = \boldsymbol{H}_k\, \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^T + \boldsymbol{R}_k \tag{6.5}$$

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\, \boldsymbol{H}_k^T\, \boldsymbol{S}_k^{-1} \tag{6.6}$$

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\, \widetilde{\boldsymbol{y}}_k \tag{6.7}$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\, \boldsymbol{H}_k\, \boldsymbol{P}_{k|k-1} \tag{6.8}$$

Time Update (prediction)

$$\hat{\boldsymbol{x}}_{k+1|k} = \boldsymbol{F}_k\, \hat{\boldsymbol{x}}_{k|k} \tag{6.9}$$

$$\boldsymbol{P}_{k+1|k} = \boldsymbol{F}_k\, \boldsymbol{P}_{k|k}\, \boldsymbol{F}_k^T + \boldsymbol{G}_k\, \boldsymbol{Q}_k\, \boldsymbol{G}_k^T \tag{6.10}$$

$\boldsymbol{K}_k$ is the Kalman filter gain. In the usual case of total absence of prior information on the initial state, one can choose $\hat{\boldsymbol{x}}_0 = 0$, $\boldsymbol{P}_0 = p_0\, I$ with p_0 being a large number.

6.2.2 Extended Kalman filter

For the case of a nonlinear state space model, the idea of the EKF is to apply the Kalman filter to a linearized version of the state space model, via a first-order Taylor series expansion. So we get for the state update and the measurement update equations

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{w}_k) \approx \boldsymbol{F}_k\, \boldsymbol{x}_k + \boldsymbol{G}_k\, \boldsymbol{w}_k \tag{6.11}$$

$$\boldsymbol{y}_k = \boldsymbol{h}(\boldsymbol{x}_k) + \boldsymbol{v}_k \approx \boldsymbol{H}_k\, \boldsymbol{x}_k + \boldsymbol{v}_k \tag{6.12}$$

where

$$\begin{aligned}\boldsymbol{F}_k &= \left.\frac{\partial\, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w})}{\partial\, \boldsymbol{x}^T}\right|_{(\boldsymbol{x},\boldsymbol{w})=(\boldsymbol{x}_k,\boldsymbol{w}_k)} \qquad \boldsymbol{G}_k = \left.\frac{\partial\, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w})}{\partial\, \boldsymbol{w}^T}\right|_{(\boldsymbol{x},\boldsymbol{w})=(\boldsymbol{x}_k,\boldsymbol{w}_k)} \\ \boldsymbol{H}_k &= \left.\frac{\partial\, \boldsymbol{h}(\boldsymbol{x})}{\partial\, \boldsymbol{x}^T}\right|_{\boldsymbol{x}=\boldsymbol{x}_k}\end{aligned} \tag{6.13}$$

So, at this point, the basic Kalman filter can be applied to the obtained approximate linear state space model. The EKF approach can be used to adapt some parameters in an otherwise linear state space model $\boldsymbol{x}'_{k+1} = \boldsymbol{F}' \ \boldsymbol{x}'_k + \boldsymbol{G}' \ \boldsymbol{w}_k$. For instance, consider the case in which one wants to adapt parameters appearing (e.g.) linearly in the matrix $\boldsymbol{F}' = \boldsymbol{F}'(\boldsymbol{\theta})$. One can jointly estimate the unknown constant parameters vector $\boldsymbol{\theta}$ by considering the following state update for them: $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$. Then one can introduce the augmented state and system matrices

$$\boldsymbol{x}_k = \begin{bmatrix} \boldsymbol{x}'_k \\ \boldsymbol{\theta}_k \end{bmatrix}, \ \boldsymbol{F}_k = \begin{bmatrix} \boldsymbol{F}'(\boldsymbol{\theta}_k) & \boldsymbol{C}(\boldsymbol{x}'_k) \\ 0 & I \end{bmatrix}, \ \boldsymbol{G}_k = \begin{bmatrix} \boldsymbol{G}' \\ 0 \end{bmatrix} \tag{6.14}$$

where $\boldsymbol{C}(\boldsymbol{x}'_k) = \dfrac{\partial \boldsymbol{F}'(\boldsymbol{\theta})\boldsymbol{x}'_k}{\partial \boldsymbol{\theta}^T}$. When running the EKF, the state-dependent system matrices have to be filled with the latest state estimates, so in this case

$$\boldsymbol{F}_k = \begin{bmatrix} \boldsymbol{F}'(\widehat{\boldsymbol{\theta}}_{k|k}) & \boldsymbol{C}(\hat{\boldsymbol{x}}'_{k|k}) \\ 0 & I \end{bmatrix} \tag{6.15}$$

The parameters vector $\boldsymbol{\theta}$ is often not really constant and hence need to be tracked adaptively. This can be done either by introducing some process noise in $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k$ (random walk time evolution) or by introducing exponential weighting (at least for the $\boldsymbol{\theta}$ portion) into the Kalman filter updates [Anderson & Moore 1979]. The EKF approach allows fairly and straightforwardly the estimation of parameters in $\boldsymbol{F}_k$, $\boldsymbol{H}_k$, or $\boldsymbol{G}_k$, but much less so in $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$. For adapting (parameters in) $\boldsymbol{Q}$ and $\boldsymbol{R}$, one needs to consider the innovations representation $\hat{\boldsymbol{x}}_{k+1|k} = \boldsymbol{F}_k \, \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{F}_k \boldsymbol{K}_k \, \widetilde{\boldsymbol{y}}_k$ and consider gradients of the Kalman gain $\boldsymbol{K}_k$ w.r.t. these matrices. The EKF has been widely used thanks to its practical usefulness [Pham *et al.* 1998, Dhaouadi *et al.* 1991, Kim *et al.* 1994], nevertheless, its convergence is not always guaranteed [Ljung 1979, Reif *et al.* 1999].

6.2.3 Recursive prediction error method

The recursive prediction error method (RPEM) is an adaptive implementation of ML parameters estimation [Ljung & Söderström 1983, Ljung 2002]. The negative loglikelihood becomes a least-squares criterion in the prediction errors (innovations) and RPEM performs one iteration per new sample. Applied to Kalman filter, the RPEM (referred also as adaptive EKF) can be seen as a more rigorous version of EKF and computes gradients more precisely [Wiklander 2003]. Indeed, in the case of a state transition matrix $\boldsymbol{F}_k = \boldsymbol{F}_k(\boldsymbol{\theta})$, the EKF considers the gradient

$$\frac{\partial \boldsymbol{x}_{k+1}}{\partial \boldsymbol{\theta}^T} = \frac{\partial \boldsymbol{F}_k(\boldsymbol{\theta})\boldsymbol{x}_k}{\partial \boldsymbol{\theta}^T} \tag{6.16}$$

where only the explicit dependence of $\boldsymbol{F}$ on $\boldsymbol{\theta}$ is taken into account, whereas the RPEM considers more correctly

$$\frac{\partial \boldsymbol{x}_{k+1}}{\partial \boldsymbol{\theta}^T} = \frac{\partial \boldsymbol{F}_k(\boldsymbol{\theta})\boldsymbol{x}_k}{\partial \boldsymbol{\theta}^T} + \boldsymbol{F}_k(\boldsymbol{\theta}) \frac{\partial \boldsymbol{x}_k}{\partial \boldsymbol{\theta}^T} \tag{6.17}$$

One characteristic of the RPEM is a higher complexity. More details about RPEM for Kalman filter can be found in [Åström 1980, Wiklander 2003, Ljungquist & Balchen 1993].

6.2.4 Fixed-lag smoothing

In fixed-lag smoothing, the aim is to have an estimate of the state at time $k-L$ given L future measurements where L is the fixed lag. In the case of $L=1$ and using the innovations approach, we have

$$\hat{\boldsymbol{x}}_{k-1|k} = \hat{\boldsymbol{x}}_{k-1|k-1} + C_{\boldsymbol{x}_{k-1}\widetilde{\boldsymbol{y}}_k} \boldsymbol{S}_k^{-1} \widetilde{\boldsymbol{y}}_k \, . \tag{6.18}$$

After a few steps, we get the following lag-1 smoothing equations that need to be added to the basic Kalman filter equations (to be inserted between the measurement update and the time update)

$$\begin{aligned} \boldsymbol{K}_{k;1} &= \boldsymbol{P}_{k-1|k-1} \boldsymbol{F}_{k-1}^T \boldsymbol{H}_k^T & (6.19)\\ \hat{\boldsymbol{x}}_{k-1|k} &= \hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{K}_{k;1} \boldsymbol{S}_k^{-1} \widetilde{\boldsymbol{y}}_k & (6.20)\\ \boldsymbol{P}_{k-1|k} &= \boldsymbol{P}_{k-1|k-1} - \boldsymbol{K}_{k;1} \boldsymbol{S}_k^{-1} \boldsymbol{K}_{k;1}^T & (6.21) \end{aligned}$$

There are more types of optimal smoothing such as fixed-point smoothing where the state estimate at time k is computed using measurements up to and including time j where $j \geq k$, and fixed-interval smoothing where the state is estimated based on fixed set of measurements [Simon 2006].

6.2.5 Gaussian second order extented Kalman filter

In EKF, the linearization of $\boldsymbol{f}(.)$ and $\boldsymbol{h}(.)$ using Taylor series is limited to the first order expansion. In order to handle more accurately the nonlinearity, the SOEKF proposes to extend the Taylor series expansion to the second order [Jazwinski 1970, Athans *et al.* 1968]. If we consider the following state update equation ($\boldsymbol{x}_k$ is the augmented state)

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k) + \boldsymbol{G}_k \boldsymbol{w}_k \tag{6.22}$$

Then, we get for the SOEKF time update equations

$$\hat{\boldsymbol{x}}_{k+1|k} = \boldsymbol{f}(\hat{\boldsymbol{x}}_{k|k}) + \frac{1}{2}\sum_{i=1}^{M} \boldsymbol{\phi}_i \,\mathrm{tr}\left\{\boldsymbol{F}_{i,k}^{xx}\boldsymbol{P}_{k|k}\right\} \tag{6.23}$$

$$\begin{aligned}\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k\boldsymbol{P}_{k|k}\boldsymbol{F}_k^T + \boldsymbol{G}_k\boldsymbol{Q}_k\boldsymbol{G}_k^T \\ &+ \frac{1}{2}\sum_{i,j=1}^{M} \boldsymbol{\phi}_i\boldsymbol{\phi}_j^T \,\mathrm{tr}\left(\boldsymbol{F}_{i,k}^{xx}\boldsymbol{P}_{k|k}\boldsymbol{F}_{j,k}^{xx}\boldsymbol{P}_{k|k}\right)\end{aligned} \tag{6.24}$$

$$\boldsymbol{F}_{i,k}^{xx} = \frac{\partial^2 f_i}{\partial\boldsymbol{x}_k\partial\boldsymbol{x}_k^T}\Big|_{\boldsymbol{x}_k=\hat{\boldsymbol{x}}_{k|k}} \tag{6.25}$$

$$\boldsymbol{F}_k = \frac{\partial\,\boldsymbol{f}}{\partial\boldsymbol{x}_k^T}\Big|_{\boldsymbol{x}_k=\hat{\boldsymbol{x}}_{k|k}} \tag{6.26}$$

where f_i is the i$^{\text{th}}$ component of the vector $\boldsymbol{f}$ and $\boldsymbol{\phi}_i$ is the $M \times 1$ vector with all zeros except for 1 in the i$^{\text{th}}$ element. The same development is made for the measurement update equations

$$\hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\left(\boldsymbol{y}_k - \boldsymbol{h}(\hat{\boldsymbol{x}}_{k|k-1}) - \boldsymbol{\pi}_k\right) \tag{6.27}$$

$$\boldsymbol{\pi}_k = \frac{1}{2}\boldsymbol{K}_k\sum_{i=1}^{M} \boldsymbol{\phi}_i \,\mathrm{tr}\left(\mathbf{H}_{i,k}^{xx}\boldsymbol{P}_{k|k-1}\right) \tag{6.28}$$

$$\mathbf{H}_{i,k}^{xx} = \frac{\partial^2 h_i}{\partial\boldsymbol{x}_k\partial\boldsymbol{x}_k^T}\Big|_{\boldsymbol{x}_k=\hat{\boldsymbol{x}}_{k|k-1}} \tag{6.29}$$

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\mathbf{H}_k^T\left(\mathbf{H}_k\boldsymbol{P}_{k|k-1}\mathbf{H}_k^T + \boldsymbol{R}_k + \boldsymbol{D}_k\right)^{-1} \tag{6.30}$$

$$\boldsymbol{D}_k = \frac{1}{2}\sum_{i,j=1}^{M} \boldsymbol{\phi}_i\boldsymbol{\phi}_j^T \,\mathrm{tr}\left(\mathbf{H}_{i,k}^{xx}\boldsymbol{P}_{k|k-1}\mathbf{H}_{j,k}^{xx}\boldsymbol{P}_{k|k-1}\right) \tag{6.31}$$

$$\mathbf{H}_k = \frac{\partial\,\boldsymbol{h}}{\partial\boldsymbol{x}_k^T}\Big|_{\boldsymbol{x}_k=\hat{\boldsymbol{x}}_{k|k-1}} \tag{6.32}$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\,\boldsymbol{H}_k\,\boldsymbol{P}_{k|k-1} \tag{6.33}$$

where h_i is the i$^{\text{th}}$ component of the vector $\boldsymbol{h}$. The term $\boldsymbol{\pi}_k$ represents a bias correction term (w.r.t. the EKF) that aims to have unbiased estimate of $\boldsymbol{x}_k$. Although the SOEKF often provides improved performance over the EKF, nothing definitive can be said about its performance. In fact, an example of an unstable SOEKF was reported in [Kushner 1967].

6.2.6 Expectation-Maximization Kalman filter

In EM, the parameters are estimated by minimizing expected values of negative loglikelihoods. For the state update, since $\boldsymbol{G}_k$ is typically a tall matrix, $\boldsymbol{G}_k\,\boldsymbol{w}_k$ has a singular covariance matrix. The state update equation can be rewritten as

$$\boldsymbol{G}_k^+\,\boldsymbol{x}_{k+1} = \boldsymbol{G}_k^+\,\boldsymbol{F}_k\,\boldsymbol{x}_k + \boldsymbol{w}_k \tag{6.34}$$

where $\boldsymbol{G}_k^+ = (\boldsymbol{G}_k^T \boldsymbol{G}_k)^{-1} \boldsymbol{G}_k^T$ is the pseudo-inverse of $\boldsymbol{G}_k$. For the parameters involved in the state update equation, the following negative loglikelihood is applicable:

$$\sum_{k=1}^{n} \{\ln\det(\boldsymbol{Q}_k) + (\boldsymbol{x}_{k+1} - \boldsymbol{F}_k\,\boldsymbol{x}_k)^T \boldsymbol{G}_k^{+T} \boldsymbol{Q}_k^{-1} \boldsymbol{G}_k^{+} (\boldsymbol{x}_{k+1} - \boldsymbol{F}_k\,\boldsymbol{x}_k)\} \tag{6.35}$$

where n is the total number of observed samples. For the parameters involved in the measurement equation, the appropriate negative loglikelihood is

$$\sum_{k=1}^{n} \{\ln\det(\boldsymbol{R}_k) + (\boldsymbol{y}_k - \boldsymbol{H}_k\,\boldsymbol{x}_k)^T \boldsymbol{R}_k^{-1} (\boldsymbol{y}_k - \boldsymbol{H}_k\,\boldsymbol{x}_k)\} \tag{6.36}$$

Now the expectation is computed, in principle with the conditional distribution given all data, hence $\mathrm{E}_{|\boldsymbol{y}_{1:n}}\{.\}$. This leads to an iterative algorithm within each iteration a whole fixed-interval smoothing operation is done. An adaptive version [Weinstein *et al.* 1994, Gao *et al.* 2003] can be obtained by replacing fixed-interval smoothing by fixed-lag smoothing and performing one iteration per time sample. Since the state update equation corresponds to a vector AR(1) model, one may expect (as in [Gao *et al.* 2003]) that a lag of 1 should be enough (to guarantee convergence). In [Weinstein *et al.* 1994], Weinstein et al. reduce complexity further by suggesting that filtering might be enough. In that case, the (presumably) slowly varying $\widehat{\boldsymbol{Q}}_{k+1}$ and $\widehat{\boldsymbol{F}}_{k+1}$ (for use in the Kalman filter at time $k+1$) get determined by minimizing $\sum_{i=1}^{k} \lambda^{k-i}\ \mathrm{E}_{|i}\{\text{Terms in ((6.35))}\}$ w.r.t. $\boldsymbol{Q}$ and $\boldsymbol{F}$ ($\boldsymbol{G}$ is known) where they introduce an exponential forgetting factor $\lambda \lessapprox 1$. This is equivalent to

$$\begin{aligned}\left\{\widehat{\boldsymbol{Q}}_{k+1}, \widehat{\boldsymbol{F}}_{k+1}\right\} &= \arg\min_{\boldsymbol{Q},\boldsymbol{F}} \gamma_k^{-1} \ln\det(\boldsymbol{Q}) \\ &\quad + \sum_{i=1}^{k} \lambda^{k-i}\ \mathrm{tr}\{\boldsymbol{G}_i^{+T} \boldsymbol{Q}^{-1} \boldsymbol{G}_i^{+}\ \mathrm{E}_{|\boldsymbol{y}_{1:i}}\{(\boldsymbol{x}_{i+1} - \boldsymbol{F}\,\boldsymbol{x}_i)(\boldsymbol{x}_{i+1} - \boldsymbol{F}\,\boldsymbol{x}_i)^T\}\}\end{aligned} \tag{6.37}$$

where we introduced $\gamma_k^{-1} = \sum_{i=1}^{k} \lambda^{k-i} = \lambda\,\gamma_{k-1}^{-1} + 1$. γ_k^{-1} behaves initially as $1/k$ but saturates eventually at $\gamma_\infty^{-1} = 1/(1-\lambda)$. The expectations that need to be computed in (6.37) are expressed as follows

$$\mathrm{E}_{|\boldsymbol{y}_{1:i}}\left\{\boldsymbol{x}_i \boldsymbol{x}_i^T\right\} = \hat{\boldsymbol{x}}_{i|i} \hat{\boldsymbol{x}}_{i|i}^T + \boldsymbol{P}_{i|i} \tag{6.38}$$

$$\mathrm{E}_{|\boldsymbol{y}_{1:i}}\left\{\boldsymbol{x}_{i+1} \boldsymbol{x}_i^T\right\} = \widehat{\boldsymbol{F}}_i\,\hat{\boldsymbol{x}}_{i|i} \hat{\boldsymbol{x}}_{i|i}^T + \widehat{\boldsymbol{F}}_i\,\boldsymbol{P}_{i|i} \tag{6.39}$$

$$\mathrm{E}_{|\boldsymbol{y}_{1:i}}\left\{\boldsymbol{x}_i \boldsymbol{x}_{i+1}^T\right\} = \hat{\boldsymbol{x}}_{i|i} \hat{\boldsymbol{x}}_{i|i}^T \widehat{\boldsymbol{F}}_i^T + \boldsymbol{P}_{i|i}\,\widehat{\boldsymbol{F}}_i^T \tag{6.40}$$

$$\begin{aligned}\mathrm{E}_{|\boldsymbol{y}_{1:i}}\left\{\boldsymbol{x}_{i+1} \boldsymbol{x}_{i+1}^T\right\} &= \hat{\boldsymbol{x}}_{i+1|i} \hat{\boldsymbol{x}}_{i+1|i}^T + \boldsymbol{P}_{i+1|i} \\ &= \widehat{\boldsymbol{F}}_i\,(\hat{\boldsymbol{x}}_{i|i} \hat{\boldsymbol{x}}_{i|i}^T + \boldsymbol{P}_{i|i})\,\widehat{\boldsymbol{F}}_i^T + \boldsymbol{G}_i \widehat{\boldsymbol{Q}}_i \boldsymbol{G}_i^T\end{aligned} \tag{6.41}$$

In case of time-invariant $\boldsymbol{G}_k \equiv \boldsymbol{G}$, we can rewrite the cost function in (6.37) as

$$\ln\det(\boldsymbol{Q}) + \mathrm{tr}\{\boldsymbol{G}^{+T} \boldsymbol{Q}^{-1} \boldsymbol{G}^{+} (\boldsymbol{M}_k^{11} - \boldsymbol{F}\,\boldsymbol{M}_k^{01} - \boldsymbol{M}_k^{10}\,\boldsymbol{F}^T + \boldsymbol{F}\,\boldsymbol{M}_k^{00}\,\boldsymbol{F}^T)\} \tag{6.42}$$

where

$$\begin{aligned}
\boldsymbol{M}_k^{00} &= (1-\gamma_k)\,\boldsymbol{M}_{k-1}^{00} + \gamma_k\,(\hat{\boldsymbol{x}}_{k|k}\hat{\boldsymbol{x}}_{k|k}^T + \boldsymbol{P}_{k|k}) && (6.43)\\
\boldsymbol{M}_k^{10} &= (1-\gamma_k)\,\boldsymbol{M}_{k-1}^{10} + \gamma_k\,\widehat{\boldsymbol{F}}_k\,(\hat{\boldsymbol{x}}_{k|k}\hat{\boldsymbol{x}}_{k|k}^T + \boldsymbol{P}_{k|k}) && (6.44)\\
\boldsymbol{M}_k^{01} &= (1-\gamma_k)\,\boldsymbol{M}_{k-1}^{01} + \gamma_k\,(\hat{\boldsymbol{x}}_{k|k}\hat{\boldsymbol{x}}_{k|k}^T + \boldsymbol{P}_{k|k})\,\widehat{\boldsymbol{F}}_k^T && (6.45)\\
\boldsymbol{M}_k^{11} &= (1-\gamma_k)\,\boldsymbol{M}_{k-1}^{11} + \gamma_k(\widehat{\boldsymbol{F}}_k\,(\hat{\boldsymbol{x}}_{k|k}\hat{\boldsymbol{x}}_{k|k}^T + \boldsymbol{P}_{k|k})\,\widehat{\boldsymbol{F}}_k^T + \boldsymbol{G}\widehat{\boldsymbol{Q}}_k\boldsymbol{G}^T) && (6.46)
\end{aligned}$$

In case of furthermore time-invariant $\boldsymbol{F}_k \equiv \boldsymbol{F}$ and $\boldsymbol{Q}_k \equiv \boldsymbol{Q}$, then

$$\begin{aligned}
\boldsymbol{M}_k^{10} &= \widehat{\boldsymbol{F}}\,\boldsymbol{M}_k^{00} && (6.47)\\
\boldsymbol{M}_k^{01} &= \boldsymbol{M}_k^{00}\,\widehat{\boldsymbol{F}}^T && (6.48)\\
\boldsymbol{M}_k^{11} &= \widehat{\boldsymbol{F}}\,\boldsymbol{M}_k^{00}\,\widehat{\boldsymbol{F}}^T + \boldsymbol{G}\,\widehat{\boldsymbol{Q}}\,\boldsymbol{G}^T && (6.49)
\end{aligned}$$

As a result, (6.43) can be rewritten as

$$\ln\det(\boldsymbol{Q}) + \mathrm{tr}\{\boldsymbol{Q}^{-1}\widehat{\boldsymbol{Q}}\} + \mathrm{tr}\{\boldsymbol{G}^{+T}\boldsymbol{Q}^{-1}\boldsymbol{G}_i^{+}(\widehat{\boldsymbol{F}}-\boldsymbol{F})\,\boldsymbol{M}_k^{00}\,(\widehat{\boldsymbol{F}}-\boldsymbol{F})^T)\} \qquad (6.50)$$

The optimization of (6.50) now clearly leads to $\widehat{\boldsymbol{F}} = \boldsymbol{F}$, $\widehat{\boldsymbol{Q}} = \boldsymbol{Q}$. So we just get back the quantities that we use in the Kalman filter, without any additional information. Hence, just Kalman filtering in the EM-Kalman is not enough to adapt the state update parameters.

For adapting the parameters in the measurement equation, $\boldsymbol{R}_k$ and $\boldsymbol{H}_k$, Kalman filtering is sufficient. A similar derivation from the expected measurement negative logliklihood in (6.36) leads to

$$\begin{aligned}
\widehat{\boldsymbol{H}}_k &= \widehat{\boldsymbol{C}}_{\boldsymbol{yx},k}\left(\boldsymbol{M}_k^{00}\right)^{-1} && (6.51)\\
\widehat{\boldsymbol{R}}_k &= \widehat{\boldsymbol{C}}_{\boldsymbol{yy},k} - \widehat{\boldsymbol{C}}_{\boldsymbol{yx},k}\left(\boldsymbol{M}_k^{00}\right)^{-1}\widehat{\boldsymbol{C}}_{\boldsymbol{xy},k} && (6.52)
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{C}}_{\boldsymbol{yy},k} &= (1-\gamma_k)\widehat{\boldsymbol{C}}_{\boldsymbol{yy},k-1} + \gamma_k\boldsymbol{y}_k\boldsymbol{y}_k^T && (6.53)\\
\widehat{\boldsymbol{C}}_{\boldsymbol{xy},k} &= (1-\gamma_k)\widehat{\boldsymbol{C}}_{\boldsymbol{xy},k-1} + \gamma_k\hat{\boldsymbol{x}}_{k|k}\boldsymbol{y}_k^T && (6.54)
\end{aligned}$$

6.2.7 Adaptive EM-Kalman filter with fixed-lag smoothing

Consider now the case in which the state space model is essentially time-invariant (or slowly time-varying). In that case the time index k of the system matrices ($\boldsymbol{F}_k$, $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, $\boldsymbol{H}_k$) just reflects at which time they have been adapted. The resulting Kalman filter equations with lag-1 smoothing become

$$\begin{aligned}
\hat{\boldsymbol{y}}_{k|k-1} &= \boldsymbol{H}_{k-1}\,\hat{\boldsymbol{x}}_{k|k-1} && (6.55)\\
\widetilde{\boldsymbol{y}}_k &= \boldsymbol{y}_k - \hat{\boldsymbol{y}}_{k|k-1} && (6.56)\\
\boldsymbol{S}_k &= \boldsymbol{H}_{k-1}\,\boldsymbol{P}_{k|k-1}\,\boldsymbol{H}_{k-1}^T + \boldsymbol{R}_{k-1} && (6.57)\\
\boldsymbol{K}_{k;1} &= \boldsymbol{P}_{k-1|k-1}\,\boldsymbol{F}_{k-1}^T\,\boldsymbol{H}_{k-1}^T && (6.58)\\
\hat{\boldsymbol{x}}_{k-1|k} &= \hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{K}_{k;1}\,\boldsymbol{S}_k^{-1}\,\widetilde{\boldsymbol{y}}_k && (6.59)\\
\boldsymbol{P}_{k-1|k} &= \boldsymbol{P}_{k-1|k-1} - \boldsymbol{K}_{k;1}\,\boldsymbol{S}_k^{-1}\,\boldsymbol{K}_{k;1}^T && (6.60)\\
\boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1}\,\boldsymbol{H}_{k-1}^T\,\boldsymbol{S}_k^{-1} && (6.61)\\
\hat{\boldsymbol{x}}_{k|k} &= \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\,\widetilde{\boldsymbol{y}}_k && (6.62)\\
\boldsymbol{P}_{k|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\,\boldsymbol{H}_{k-1}\,\boldsymbol{P}_{k|k-1} && (6.63)\\
\hat{\boldsymbol{x}}_{k+1|k} &= \boldsymbol{F}_k\,\hat{\boldsymbol{x}}_{k|k} && (6.64)\\
\boldsymbol{P}_{k+1|k} &= \boldsymbol{F}_k\,\boldsymbol{P}_{k|k}\,\boldsymbol{F}_k^T + \boldsymbol{G}\,\boldsymbol{Q}_k\,\boldsymbol{G}^T && (6.65)
\end{aligned}$$

So, the system matrices ($\boldsymbol{F}_k$, $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, $\boldsymbol{H}_k$) should be adapted after the smoothing step and before the filtering and prediction steps. We now adapt the matrices $\boldsymbol{F}$ and $\boldsymbol{Q}$ from the equivalent of (6.37) with $\mathrm{E}_{|\boldsymbol{y}_{1:i}}\{.\}$ replaced by $\mathrm{E}_{|\boldsymbol{y}_{1:i+1}}\{.\}$. This leads to the matrix updates

$$\begin{aligned}
\boldsymbol{M}_k^{00} &= (1-\gamma_k)\,\boldsymbol{M}_{k-1}^{00} + \gamma_k\,(\hat{\boldsymbol{x}}_{k-1|k}\hat{\boldsymbol{x}}_{k-1|k}^T + \boldsymbol{P}_{k-1|k}) && (6.66)\\
\boldsymbol{M}_k^{10} &= (1-\gamma_k)\,\boldsymbol{M}_{k-1}^{10} + \gamma_k\,(\hat{\boldsymbol{x}}_{k|k}\hat{\boldsymbol{x}}_{k-1|k}^T + \boldsymbol{F}_{k-1}\boldsymbol{P}_{k-1|k}\\
&\quad -\boldsymbol{G}\boldsymbol{Q}_{k-1}\boldsymbol{G}^T\boldsymbol{H}_{k-1}^T\boldsymbol{S}_k^{-1}\boldsymbol{K}_{k,1}^T) && (6.67)\\
&= (\boldsymbol{M}_k^{01})^T\\
\boldsymbol{M}_k^{11} &= (1-\gamma_k)\,\boldsymbol{M}_{k-1}^{11} + \gamma_k(\hat{\boldsymbol{x}}_{k|k}\hat{\boldsymbol{x}}_{k|k}^T + \boldsymbol{P}_{k|k}) && (6.68)
\end{aligned}$$

The minimization of the expected negative loglikelihood w.r.t. $\boldsymbol{F}$ and $\boldsymbol{Q}$ leads to the following minimizers

$$\begin{aligned}
\widehat{\boldsymbol{F}}_k &= \boldsymbol{M}_k^{10}\,(\boldsymbol{M}_k^{00})^{-1} && (6.69)\\
\widehat{\boldsymbol{Q}}_k &= \boldsymbol{G}^{+}(\boldsymbol{M}_k^{11} - \boldsymbol{M}_k^{10}\,(\boldsymbol{M}_k^{00})^{-1}\boldsymbol{M}_k^{01})\boldsymbol{G}^{+\,T} && (6.70)
\end{aligned}$$

For adapting $\boldsymbol{H}$ and $\boldsymbol{R}$, the system becomes

$$\begin{aligned}
\widehat{\boldsymbol{H}}_k &= \widehat{\boldsymbol{C}}_{\boldsymbol{yx},k}\,\left(\boldsymbol{M}_k^{11}\right)^{-1} && (6.71)\\
\widehat{\boldsymbol{R}}_k &= \widehat{\boldsymbol{C}}_{\boldsymbol{yy},k} - \widehat{\boldsymbol{C}}_{\boldsymbol{yx},k}\,\left(\boldsymbol{M}_k^{11}\right)^{-1}\,\widehat{\boldsymbol{C}}_{\boldsymbol{xy},k} && (6.72)
\end{aligned}$$

where $\widehat{\boldsymbol{C}}_{\boldsymbol{yx},k}$ and $\widehat{\boldsymbol{C}}_{\boldsymbol{yy},k}$ are the same defined in (6.51). For the initialization, in absence of any side information, one can take $\boldsymbol{M}_0^{00} = 1/p_0\boldsymbol{I}$, $\boldsymbol{M}_0^{10} = \boldsymbol{0}$, $\boldsymbol{M}_0^{11} = \boldsymbol{0}$, $\widehat{\boldsymbol{C}}_{\boldsymbol{xy},0} = 1/p_0\boldsymbol{I}$, $\widehat{\boldsymbol{C}}_{\boldsymbol{yy},0} = \boldsymbol{0}$ where again p_0 is a very large number.

6.2.8 Alternating MAP/ML Kalman filter

The alternating MAP/ML estimate consists in alternating optimization between the MAP estimator for the state sequence $\boldsymbol{x}_k$ and the ML estimator for the parameters $\boldsymbol{\theta}$. The ML estimate of $\boldsymbol{\theta}$ is obtained by performing least-squares estimation given the state sequence replaced by its estimate. The resulting algorithm is similar to the EM-Kalman with only the $\hat{\boldsymbol{x}}$ terms kept in the matrices $\boldsymbol{M}_k^{ij}$.

6.2.9 Variational Bayes Kalman filter

Variational Bayes

The VB is again an application of alternating optimization but this time applied to the Kullback-Leibler divergence (KLD) between the true joint *a posteriori* probability density function (pdf) of hidden data and parameters $f(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$, and an approximate form $q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ called the free distribution [Tzikas *et al.* 2008]. The main simplifying assumption of VB is to constrain $q(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ to a factorised (separable) form $q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{\theta}|\boldsymbol{y})$, in other words, suppose *a posteriori* independence between $\boldsymbol{x}$ and $\boldsymbol{\theta}$. The VB can be interpreted as an extension of EM to the case where the parameters $\boldsymbol{\theta}$ are also considered random. Then, the cost function (3.16) in chapter 3 is modified to

$$\mathcal{F}(q(\boldsymbol{x}|\boldsymbol{y}), q(\boldsymbol{\theta}|\boldsymbol{y})) \triangleq \int q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})}{q(\boldsymbol{x}|\boldsymbol{y})q(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{\theta} \tag{6.73}$$

The optimization of the functional (6.73) w.r.t $q(\boldsymbol{x}|\boldsymbol{y})$ and $q(\boldsymbol{\theta}|\boldsymbol{y})$ results in the variational Bayes for EM (VBEM) algorithm, composed of two main steps VBE and VBM (in parallel to the E-step and M-step of EM respectively).

$$\text{VBM step:} \quad \ln q^{i+1}(\boldsymbol{\theta}|\boldsymbol{y}) \doteq \int q^{i}(\boldsymbol{x}|\boldsymbol{y}) \ln f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \, d\boldsymbol{x} \tag{6.74}$$

$$\text{VBE step:} \quad \ln q^{i+1}(\boldsymbol{x}|\boldsymbol{y}) \doteq \int q^{i+1}(\boldsymbol{\theta}|\boldsymbol{y}) \ln f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \, d\boldsymbol{\theta} \tag{6.75}$$

where $f(\boldsymbol{\theta})$ is the *a priori* function of $\boldsymbol{\theta}$, $\doteq$ denotes equality up to "constant" (constant relative to the variable of the equation, but it may depend on the other variables) and the index i denotes the iteration number. Notice that the difference between the loglikelihood $f(\boldsymbol{y})$ and the cost function in (6.73) corresponds to the KLD cited in the beginning of the section. The privilege of VB is that by using $q(\boldsymbol{\theta}|\boldsymbol{y})$ in the computation of the hidden sate estimate $\hat{\boldsymbol{x}}$, not only the estimate of $\boldsymbol{\theta}$ accounts but also its estimation error.

Conjugate exponential family and variational Bayes

The VBEM is proved successful when applied to the conjugate-exponential (CE) models [Ghahramani & Beal 2001, Beal 2003]. Two main conditions are satisfied by

CE models. The first condition is that the complete-data likelihood $f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})$ has to be in the exponential family

$$f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta}) = g(\boldsymbol{\theta})\ p(\boldsymbol{x},\boldsymbol{y})\ e^{\varphi(\boldsymbol{\theta})^T \boldsymbol{u}(\boldsymbol{x},\boldsymbol{y})} \tag{6.76}$$

where $\varphi(\boldsymbol{\theta})$ is the vector of natural parameters, $p(\boldsymbol{x},\boldsymbol{y})$ and $\boldsymbol{u}(\boldsymbol{x},\boldsymbol{y})$ are the functions that define the exponential family, and $g(\boldsymbol{\theta})$ is a normalisation constant. The second condition is that the parameters prior $f(\boldsymbol{\theta})$ has to be conjugate to the complete-data likelihood which means that the parameters prior and posterior pdfs belong to the same family [Beal 2003]. Thus, the prior function is of the form

$$f(\boldsymbol{\theta}|\eta,\nu) = \zeta(\eta,\nu)\ g(\boldsymbol{\theta})^{\eta}\ e^{\varphi(\boldsymbol{\theta})^T \nu} \tag{6.77}$$

where η and ν are the hyperparameters of $\boldsymbol{\theta}$ and $\zeta(\eta,\nu)$ is a normalisation constant. The two former conditions on the complete-data likelihood and the parameters prior pdfs provide tractable VBEM updates. Given an i.i.d. data set $\boldsymbol{y}_{1:n} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$, the analytical expressions of the approximate posteriors are described as

$$q^{i+1}(\boldsymbol{\theta}|\boldsymbol{y}_{1:n},\eta',\nu'^{\,i}) = \zeta(\eta',\nu'^{\,i})\ g(\boldsymbol{\theta})^{\eta'}\ e^{\varphi(\boldsymbol{\theta})^T \nu'^{\,i}} \tag{6.78}$$

$$q^{i+1}(\boldsymbol{x}_k|\boldsymbol{y}_k) = f(\boldsymbol{x}_k|\boldsymbol{y}_k, \overline{\varphi}^{\,i+1}(\boldsymbol{\theta})) \quad for\ k = 1 \cdots, n \tag{6.79}$$

where the different new quantities are defined as

$$\overline{\varphi}^{\,i+1}(\boldsymbol{\theta}) = \mathrm{E}_{q^{i+1}(\boldsymbol{\theta}|\boldsymbol{y}_{1:n},\eta',\nu'^{\,i})}\{\varphi(\boldsymbol{\theta})\} \tag{6.80}$$

$$\nu'^{\,i} = \nu + \sum_{k=1}^{n} \mathrm{E}_{q^i(\boldsymbol{x}_k|\boldsymbol{y}_k)}\{\boldsymbol{u}(\boldsymbol{x}_k,\boldsymbol{y}_k)\} \tag{6.81}$$

$$\eta' = \eta + n \tag{6.82}$$

Hence, the VBM and VBE steps are simplified to the computation of the expectations of the sufficient statistics $\{\boldsymbol{u}(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1:n}$ under the distributions $\left\{q^i(\boldsymbol{x}_k|\boldsymbol{y}_k)\right\}_{k=1:n}$ and the natural parameters $\varphi(\boldsymbol{\theta})$ under the distributions $q^{i+1}(\boldsymbol{\theta}|\boldsymbol{y}_{1:n},\eta',\nu'^{\,i})$ respectively. For detailed proofs of these results, the reader can refer to [Beal 2003].

The interesting particular case of the joint Gaussian complete-data $(\boldsymbol{x},\boldsymbol{y})$ (presented in section 5.2) satisfies the CE model, where, using the same notations therein

$$g(\boldsymbol{\theta}) = exp(-\frac{1}{2}\boldsymbol{\mu}(\boldsymbol{\theta})^T \boldsymbol{C}(\boldsymbol{\theta})^{-1}\boldsymbol{\mu}(\boldsymbol{\theta}) - \frac{1}{2}\ln 2\pi|\boldsymbol{C}(\boldsymbol{\theta})|) \tag{6.83}$$

$$p(\boldsymbol{x},\boldsymbol{y}) = 1 \tag{6.84}$$

$$\varphi(\boldsymbol{\theta}) = [(\mathrm{vec}\boldsymbol{C}(\boldsymbol{\theta})^{-1})^T \quad \boldsymbol{\mu}(\boldsymbol{\theta})^T \boldsymbol{C}(\boldsymbol{\theta})^{-1}]^T \tag{6.85}$$

$$\boldsymbol{u}(\boldsymbol{x},\boldsymbol{y}) = [-\frac{1}{2}(\boldsymbol{w}\otimes\boldsymbol{w})^T \quad \boldsymbol{w}^T]^T \tag{6.86}$$

where the $(M+N)$-dimensional vector $\boldsymbol{w}$ is the complete-data vector $[\boldsymbol{x}^T, \boldsymbol{y}^T]^T$.

Variational Bayes for Kalman filter

The VB methodology is applied to several statistical models such as hidden Markov model (HMM) [MacKay 1997], mixture of factor analysers [Ghahramani & Beal 2000] and linear dynamical systems where the VB Kalman filter and smoother are derived [Beal 2003]. If we reconsider the state space model in (6.1) but with the matrices of the filter independent of time k, then in the VB Kalman the complete-data is defined by the state and observation vectors $\{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1:n}$, while $\boldsymbol{\theta}$ is defined by the state space model parameters $\{\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}, \boldsymbol{R}\}$. An important preprocessing task consists in defining the prior functions of the different parameters. By respecting the conjugacy constraint, the prior functions of the different parameters are defined as

- For the covariances $\boldsymbol{Q}$ and $\boldsymbol{R}$, Wishart distributions are considered for their inverses $\boldsymbol{Q}^{-1} \sim \mathcal{W}(\boldsymbol{Q}^{-1}|\boldsymbol{S}^Q, \nu^Q)$ and $\boldsymbol{R}^{-1} \sim \mathcal{W}(\boldsymbol{R}^{-1}|\boldsymbol{S}^R, \nu^R)$, where $\boldsymbol{S}^Q$ and $\boldsymbol{S}^R$ are their corresponding hyperparameters, and ν^Q and ν^Q are their degrees of freedom. When the covariance matrix ($\boldsymbol{Q}$ and/or $\boldsymbol{R}$) is reduced to a diagonal matrix, the diagonal components of its inverse are modeled by i.i.d. Gamma distributed processes.

- The matrix $\boldsymbol{F}$ has a zero-mean matrix normal distribution $\boldsymbol{F} \sim \mathcal{MN}(\boldsymbol{F}|\boldsymbol{0}, \operatorname{diag}\boldsymbol{\alpha}^{-1}, \boldsymbol{G}\boldsymbol{Q}\boldsymbol{G}^T)$ where the M-dimensional vector $\boldsymbol{\alpha}$ is its corresponding hyperparameter. Notice that the prior of $\boldsymbol{F}$ depends also on $\boldsymbol{Q}$ due to their interaction in the state evolution equation in (6.1), idem for matrix $\boldsymbol{H} \sim \mathcal{MN}(\boldsymbol{H}|\boldsymbol{0}, \operatorname{diag}\boldsymbol{\beta}^{-1}, \boldsymbol{R})$ where the N-dimensional vector $\boldsymbol{\beta}$ is its corresponding parameter.

- The initial condition $\boldsymbol{x}_0$ has a multivariate Gaussian distribution with mean $\hat{\boldsymbol{x}}_0$ and covariance $\boldsymbol{P}_0$ which are considered as hyperparameters.

Given n observations $\boldsymbol{y}_{1:n}$ and the priors defined above, we deduce the full joint distribution

$$f(\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{x}_{0:n}, \boldsymbol{y}_{1:n}) = f(\boldsymbol{F}|\boldsymbol{Q}, \boldsymbol{\alpha}) f(\boldsymbol{Q}|\boldsymbol{S}^Q, \nu^Q) f(\boldsymbol{H}|\boldsymbol{R}, \boldsymbol{\beta}) f(\boldsymbol{R}|\boldsymbol{S}^R, \nu^R) \\ f(\boldsymbol{x}_0|\hat{\boldsymbol{x}}_0, \boldsymbol{P}_0) \prod_{t=1}^{n} f(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{F}, \boldsymbol{Q}) f(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{H}, \boldsymbol{R}) \quad (6.87)$$

The variational methodology approximates the joint posterior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{x}_{0:n}$ to $q(\boldsymbol{\theta}, \boldsymbol{x}_{0:n})$ where

$$q(\boldsymbol{\theta}, \boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n}) = q(\boldsymbol{\theta}|\boldsymbol{y}_{1:n})\ q(\boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n}) \quad (6.88)$$

$$= q(\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}, \boldsymbol{R}|\boldsymbol{y}_{1:n})\ q(\boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n}) \quad (6.89)$$

$$= q(\boldsymbol{F}|\boldsymbol{Q}, \boldsymbol{y}_{1:n}) q(\boldsymbol{Q}|\boldsymbol{y}_{1:n}) q(\boldsymbol{H}|\boldsymbol{R}, \boldsymbol{y}_{1:n}) q(\boldsymbol{R}|\boldsymbol{y}_{1:n}) q(\boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n}) \quad (6.90)$$

We should distinguish between the first factorization in (6.88) that results from the VB approximation, and the additional factorizations in (6.90) that are motivated by the existing (or nonexisting) interactions implied naturally by (6.1) between the different parameters. Using the previous derivations, the cost function (6.73) is finally formulated for the VB Kalman

$$\begin{aligned}
\mathcal{F} = & \int d\boldsymbol{Q}\; q(\boldsymbol{Q}|\boldsymbol{y}_{1:n}) \ln \frac{f(\boldsymbol{Q}|\boldsymbol{S}^Q,\nu^Q)}{q(\boldsymbol{Q}|\boldsymbol{y}_{1:n})} \\
& + \int d\boldsymbol{Q}\; q(\boldsymbol{Q}|\boldsymbol{y}_{1:n}) \int d\boldsymbol{F}\; q(\boldsymbol{F}|\boldsymbol{Q},\boldsymbol{y}_{1:n}) \ln \frac{f(\boldsymbol{F}|\boldsymbol{Q},\boldsymbol{\alpha})}{q(\boldsymbol{F}|\boldsymbol{Q},\boldsymbol{y}_{1:n})} \\
& + \int d\boldsymbol{R}\; q(\boldsymbol{R}|\boldsymbol{y}_{1:n}) \ln \frac{f(\boldsymbol{R}|\boldsymbol{S}^R,\nu^R)}{q(\boldsymbol{R}|\boldsymbol{y}_{1:n})} \\
& + \int d\boldsymbol{R}\; q(\boldsymbol{R}|\boldsymbol{y}_{1:n}) \int d\boldsymbol{H}\; q(\boldsymbol{H}|\boldsymbol{R},\boldsymbol{y}_{1:n}) \ln \frac{f(\boldsymbol{H}|\boldsymbol{R},\boldsymbol{\beta})}{q(\boldsymbol{H}|\boldsymbol{R},\boldsymbol{y}_{1:n})} \\
& - \int d\boldsymbol{x}_{0:n}\; q(\boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n}) \ln q(\boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n}) \\
& + \int d\boldsymbol{R}\; q(\boldsymbol{R}|\boldsymbol{y}_{1:n}) \int d\boldsymbol{H}\; q(\boldsymbol{H}|\boldsymbol{R},\boldsymbol{y}_{1:n}) \int d\boldsymbol{Q}\; q(\boldsymbol{Q}|\boldsymbol{y}_{1:n}) \int d\boldsymbol{F}\; q(\boldsymbol{F}|\boldsymbol{Q},\boldsymbol{y}_{1:n}) \\
& \int d\boldsymbol{x}_{0:n}\; q(\boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n})\; \ln f(\boldsymbol{x}_{0:n}, y_{1:n}|\boldsymbol{F},\boldsymbol{H},\boldsymbol{Q},\boldsymbol{R}) \\
= & \mathcal{F}(q(\boldsymbol{x}_{0:n}|\boldsymbol{y}_{1:n}), q(\boldsymbol{F}|\boldsymbol{Q},\boldsymbol{y}_{1:n}), q(\boldsymbol{Q}|\boldsymbol{y}_{1:n}), q(\boldsymbol{H}|\boldsymbol{R},\boldsymbol{y}_{1:n}), q(\boldsymbol{R}|\boldsymbol{y}_{1:n})) \qquad (6.91)
\end{aligned}$$

The approximate posteriors of the different parameters are derived from the optimization of (6.91) in the VBM-step. Due to the conjugacy, the posteriors will have the same form of the prior functions. For example, $q(\boldsymbol{F}|\boldsymbol{Q},\boldsymbol{y}_{1:n})$ and $q(\boldsymbol{Q}|\boldsymbol{y}_{1:n})$ will be Gaussian and Wishart distributions respectively as their respective priors. Hence, the task is reduced to estimating the parameters of these posteriors using the parameters of the prior distributions and sufficient statistics of the hidden states computed in the VBE-step. Since the functional $\mathcal{F}$ also depends on the hyperparameters $\left\{\boldsymbol{S}^Q, \nu^Q, \boldsymbol{S}^R, \nu^R, \boldsymbol{\alpha}, \boldsymbol{\beta}, \hat{\boldsymbol{x}}_0, \boldsymbol{P}_0\right\}$, an estimate of the latter can be deduced by maximizing it. In the VBE-step, the variational Kalman filter/smoother derivation always takes into consideration the averaging of the terms depending on the parameters over the latter using their approximate posteriors. Thus, the sufficient statistics of the hidden states will depend on the parameters of the approximate posterior distributions computed in the VBE-step. For more details about the derivation of VB Kalman equations, the reader may refer to [Beal 2003].

6.3 On the performance of iterative ML algorithms

At the end of section 5.8, we mentioned that Arie Yeredor [Yeredor 2000] proves that the joint MAP/ML estimator is biased for θ and that the bias persists asymptotically. The author also provides approximate analytical expressions of the bias

$b^J(\theta)$ and the mean square error (MSE) $e^{2\,J}(\theta)$ in the subasymptotic regime (for θ scalar) where the number of observations n is too small to fit to the asymptotic but still too large to satisfy weaker assumptions. Both $b^J(\theta)$ and $e^{2\,J}(\theta)$ are expressed in terms of the bias and the MSE of the ML estimator in the subasymptotic regime. The asymptotic expressions can be deduced by letting the ML bias and MSE tend to zero and the inverse of the Fisher information $J^{-1}(\theta)$ respectively. The asymptotic bias and MSE are expressed as

$$b^J(\theta) \approx \frac{r'(\theta)}{J(\theta) - r''(\theta)} \tag{6.92}$$

$$e^{2\,J}(\theta) \approx \frac{J(\theta) + r'^2(\theta)}{(J(\theta) - r''(\theta))^2} \tag{6.93}$$

where $r'(\theta)$ and $r''(\theta)$ are respectively the gradient and Hessian of $-\frac{1}{2}\ln\det \boldsymbol{P}(\theta)$ w.r.t. θ ($\boldsymbol{P}(\theta)$ defined in (5.7)). In the following, we will discuss the EM performance.

6.3.1 The EM algorithm performance

The EM algorithm was introduced to iterate towards the ML estimate while having reduced-complexity iterations [Dempster *et al.* 1977]. Let's consider the joint Gaussian model presented in 5.2 and the notations of the MMSE estimation presented therein, but this time in the zero-mean case ($\boldsymbol{\mu} = \boldsymbol{0}$), we mentioned in section 3.3 that the estimate of $\boldsymbol{\theta}$ at iteration $i+1$, $\widehat{\boldsymbol{\theta}}^{i+1}$ is expressed (the M-step)

$$\widehat{\boldsymbol{\theta}}^{i+1} = \arg\max_{\boldsymbol{\theta}} \left\langle \ln f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) \right\rangle_{f(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^i)} = \arg\max_{\boldsymbol{\theta}} \boldsymbol{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^i) \tag{6.94}$$

The expectation term $\left\langle \ln f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) \right\rangle_{f(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^i)}$ is computed in the E-step. Depending on the application, some simplifications may occur. For instance if the parameters of interest only appear in $f(\boldsymbol{x}|\boldsymbol{\theta})$ so that $\ln f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) = \ln f(\boldsymbol{y}|\boldsymbol{x}) + \ln f(\boldsymbol{x}|\boldsymbol{\theta})$, then, apart from additive constants, we get

$$-2\,\boldsymbol{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^i) \doteq \mathrm{tr}\{\widehat{\boldsymbol{C}}^i_{\boldsymbol{xx}} \boldsymbol{C}^{-1}_{\boldsymbol{xx}}(\boldsymbol{\theta}) - \boldsymbol{I}\} - \ln\det(\widehat{\boldsymbol{C}}^i_{\boldsymbol{xx}} \boldsymbol{C}^{-1}_{\boldsymbol{xx}}(\boldsymbol{\theta})) \tag{6.95}$$

where

$$\widehat{\boldsymbol{C}}^i_{\boldsymbol{xx}} = \hat{\boldsymbol{x}}(\widehat{\boldsymbol{\theta}}^i)\hat{\boldsymbol{x}}^T(\widehat{\boldsymbol{\theta}}^i) + \boldsymbol{P}(\widehat{\boldsymbol{\theta}}^i) \tag{6.96}$$

Which is the Itakura-Saito distance (ISD) between $\widehat{\boldsymbol{C}}^i_{\boldsymbol{xx}}$ and $\boldsymbol{C}_{\boldsymbol{xx}}(\boldsymbol{\theta})$. In the general case, (6.94) leads to the ISD minimization between $\widehat{\boldsymbol{C}}^i = \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^i}\{\boldsymbol{w}\boldsymbol{w}^T\}$ with $\boldsymbol{w}^T = [\boldsymbol{x}^T \boldsymbol{y}^T]$ and the joint covariance matrix $\boldsymbol{C}(\boldsymbol{\theta})$. Now, using the block upper diagonal lower (UDL) factorization

$$\boldsymbol{C}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{F}(\boldsymbol{\theta}) \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{P}(\boldsymbol{\theta}) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_{\boldsymbol{yy}}(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{F}^T(\boldsymbol{\theta}) & \boldsymbol{I} \end{bmatrix} \tag{6.97}$$

By considering that both tr{.} and ln det(.) allow cyclic commutation of the factors in their argument, we get

$$\begin{aligned}\boldsymbol{C}^{-1}(\boldsymbol{\theta})\,\widehat{\boldsymbol{C}}^{i} &= \begin{bmatrix}\boldsymbol{P}^{-1}(\boldsymbol{\theta}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta})\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & -\boldsymbol{F}(\boldsymbol{\theta})\\ \boldsymbol{0} & \boldsymbol{I}\end{bmatrix}\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\widehat{\boldsymbol{\theta}}^{i}}\left\{\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}^{T}\right\}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ -\boldsymbol{F}^{T}(\boldsymbol{\theta}) & \boldsymbol{I}\end{bmatrix}\\ &= \begin{bmatrix}\boldsymbol{P}^{-1}(\boldsymbol{\theta}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta})\end{bmatrix}\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\widehat{\boldsymbol{\theta}}^{i}}\left\{\begin{bmatrix}\widetilde{\boldsymbol{x}}\\ \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\widetilde{\boldsymbol{x}}\\ \boldsymbol{y}\end{bmatrix}^{T}\right\}\\ &= \boldsymbol{D}^{-1}(\boldsymbol{\theta})\,\widehat{\boldsymbol{D}}^{i}\end{aligned} \tag{6.98}$$

Where

$$\boldsymbol{D}(\boldsymbol{\theta}) = \begin{bmatrix}\boldsymbol{P}(\boldsymbol{\theta}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_{\boldsymbol{yy}}(\boldsymbol{\theta})\end{bmatrix} \tag{6.99}$$

$$\widehat{\boldsymbol{D}}^{i} = \begin{bmatrix}\boldsymbol{P}(\widehat{\boldsymbol{\theta}}^{i}) & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}\hat{\boldsymbol{x}}(\widehat{\boldsymbol{\theta}}^{i}) - \hat{\boldsymbol{x}}(\boldsymbol{\theta})\\ \boldsymbol{y}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{x}}(\widehat{\boldsymbol{\theta}}^{i}) - \hat{\boldsymbol{x}}(\boldsymbol{\theta})\\ \boldsymbol{y}\end{bmatrix}^{T} \tag{6.100}$$

At convergence we get (with $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{\infty}$)

$$\boldsymbol{D}^{-1}(\boldsymbol{\theta})\,\widehat{\boldsymbol{D}}^{\infty} = \begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{\theta})\,\boldsymbol{y}\boldsymbol{y}^{T}\end{bmatrix} \tag{6.101}$$

Hence the ISD between $\widehat{\boldsymbol{C}}^{i}$ and $\boldsymbol{C}^{-1}(\boldsymbol{\theta})$ converges to the ISD between $\widehat{\boldsymbol{C}}_{\boldsymbol{yy}}^{i} = \boldsymbol{y}\boldsymbol{y}^{T}$ and $\boldsymbol{C}_{\boldsymbol{yy}}(\boldsymbol{\theta})$ and hence to the ML loglikelihood. Actually, $\boldsymbol{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{i})$ does not measure exactly the ISD between $\widehat{\boldsymbol{C}}^{i}$ and $\boldsymbol{C}(\boldsymbol{\theta})$, but

$$-2\,\boldsymbol{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{i}) \doteq \ln\det(\boldsymbol{C}(\boldsymbol{\theta})) + \mathrm{tr}\{\boldsymbol{C}^{-1}(\boldsymbol{\theta})\,\widehat{\boldsymbol{C}}^{i}\} \tag{6.102}$$

which also converges to the ML loglikelihood. The difference between $\boldsymbol{Q}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{i})$ and the ISD is due to the fact that both $\boldsymbol{C}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{C}}^{i}$ depend on $\boldsymbol{\theta}$, leading to the difference term $\ln\det(\widehat{\boldsymbol{C}}^{i})$.

The EM algorithm guarantees to monotonically increase the likelihood function $f(\boldsymbol{y}|\boldsymbol{\theta})$. In the case when $f(\boldsymbol{y}|\boldsymbol{\theta})$ is multimodal, the algorithm converges to one of its stationary points (local maximum or saddle point) depending on the initialization $\widehat{\boldsymbol{\theta}}^{0}$. If $f(\boldsymbol{y}|\boldsymbol{\theta})$ is unimodal, the algorithm converges to the global maximum independently of the initialization value [McLachlan & Krishnan 2008]. Despite its advantages, a common criticism for EM is that its convergence can be very slow. Many solutions were proposed for EM acceleration [Aitkin & Aitkin 1996, Jamshidian & Jennrich 1993, Jamshidian & Jennrich 1997], a famous one is the algorithm space-alternating generalized EM (SAGE) proposed by Fessler et al. in [Fessler & Hero 1994]. The simultaneous maximization in classical EM needs overly informative complete-data spaces, which in turn lead to slow convergence. Updating

the parameters "'sequentially"' in small groups improves the convergence rates. In [Fessler & Hero 1993], it is proved that the less informative the complete-data is, the better the asymptotic convergence rates. Less informative complete-data spaces lead to larger step size and greater likelihood increases in the early iterations. The minimization of the complete-data space information improves convergence rates. However, this can lead to intractable maximization step in classical EM, due to the simultaneous update. For each parameters subset $\boldsymbol{\theta}_S$, hidden data are designed so that it is less informative than the case where there is single complete-data space. This reduction leads to faster convergence. The methods proposed to accelerate EM based on numerical tools (Aitken's acceleration, over-relaxation...) do not guarantee monotone increases in the objective function unless the objective function is explicitly computed, but with SAGE method, monotonicity is guaranteed since it is based on statistical considerations. The choice of hidden data per each subset is tricky. In fact, the hidden data of subset $\boldsymbol{\theta}_S$ play the role of complete data for this latter knowing $\boldsymbol{\theta}_{\bar{S}}$ where $\bar{S}$ is the complementary space of S.

6.3.2 The VB algorithm performance

Even though ln $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ is quadratic and ln $f(\boldsymbol{\theta}|\boldsymbol{y})$ is asymptotically quadratic, the joint ln $f(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ contains products of both quadratic terms and hence is not Gaussian. In VB, apart from approximating the true posterior pdf by a factored form, one can furthermore require the factors to be of a certain parametric form. In the case considered here however, $q(\boldsymbol{x}|\boldsymbol{y})$ is automatically Gaussian, whereas we shall force $q(\boldsymbol{\theta}|\boldsymbol{y})$ to be Gaussian. This is done by taking the mean and covariance of the RHS in (6.75). Note that asymptotically, $q(\boldsymbol{\theta}|\boldsymbol{y})$ becomes Gaussian automatically. Also note that $q(\boldsymbol{x}|\boldsymbol{y})$ and $q(\boldsymbol{\theta}|\boldsymbol{y})$ are not the marginals of $f(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$. They are factors of which the product attempts to approximate the joint pdf as well as possible. The equalities in (6.75) should be interpreted as up to additive "constants" (possibly functions of $\boldsymbol{y}$). Hence $f(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})$ is equivalent to $f(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{y})$ in (6.75). Finally, VB is an approach that normally applies to the fully Bayesian case in which both $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are considered random. However, we shall consider the prior $f(\boldsymbol{\theta})$ to be uniform so that $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})$ becomes equivalent to $f(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{\theta})$.

The EM algorithm can be viewed as a limiting case of the VB approach, in which $\boldsymbol{\theta}$ is treated as deterministic and hence can be viewed as random with prior $f(\boldsymbol{\theta}') = \delta(\boldsymbol{\theta}' - \boldsymbol{\theta})$ (where $\boldsymbol{\theta}$ is the unknown true value). As a result also the posterior becomes of the form $q(\boldsymbol{\theta}|\boldsymbol{y}) = \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$ and hence is characterized solely by the point estimate $\widehat{\boldsymbol{\theta}}$. Under some regularity conditions, the EM estimate is known to converge to the (separate) ML estimate and hence has the same performance. EM can be viewed as a case of VB in which $q^i(\boldsymbol{\theta}|\boldsymbol{y})$ is forced to be of the form $\delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}^i})$ in the M step. The best approximation is obviously obtained for $\widehat{\boldsymbol{\theta}^i} = \arg\max_{\boldsymbol{\theta}} q^i(\boldsymbol{\theta}|\boldsymbol{y})$

where $q^i(\boldsymbol{\theta}|\boldsymbol{y})$ is obtained from the first equation in (6.75). If now furthermore $q^i(\boldsymbol{\theta}|\boldsymbol{y}) = \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^i)$, then we get from the second equation in (6.75)

$$\ln\, q^i(\boldsymbol{x}|\boldsymbol{y}) \doteq \int \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^i)\, \ln\, f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta} = \ln\, f(\boldsymbol{x}, \boldsymbol{y}|\widehat{\boldsymbol{\theta}}^i) \doteq \ln\, f(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^i) \quad (6.103)$$

This finally leads to $\widehat{\boldsymbol{\theta}}^{i+1} = \arg\max_{\boldsymbol{\theta}} \int f(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^i)\, \ln f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\, d\boldsymbol{x}$ for EM.

Now consider the actual VB updates ((6.75)) with Gaussian $q^i(\boldsymbol{\theta}|\boldsymbol{y}) = \mathcal{N}(\widehat{\boldsymbol{\theta}}^i, \boldsymbol{C}^i_{\boldsymbol{\theta}})$. Motivated by asymptotics we shall determine the Gaussian approximation by a second-order Taylor series expansion

$$\ln q^{i+1}(\boldsymbol{\theta}|\boldsymbol{y}) \doteq g^i(\boldsymbol{\theta}) \quad (6.104)$$

$$\doteq -\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^{i+1})^T (\boldsymbol{C}^{i+1}_{\boldsymbol{\theta}})^{-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^{i+1}) \quad (6.105)$$

$$\doteq g^i(\widehat{\boldsymbol{\theta}}^i) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^i)^T \frac{\partial g^i(\widehat{\boldsymbol{\theta}}^i)}{\partial \boldsymbol{\theta}} + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^i)^T \frac{\partial^2 g^i(\widehat{\boldsymbol{\theta}}^i)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^i) \quad (6.106)$$

Equating the last two lines yields

$$\widehat{\boldsymbol{\theta}}^{i+1} = \widehat{\boldsymbol{\theta}}^i + \boldsymbol{C}^{i+1}_{\boldsymbol{\theta}} \frac{\partial \ln g^i(\widehat{\boldsymbol{\theta}}^i)}{\partial \boldsymbol{\theta}} \quad (6.107)$$

$$\boldsymbol{C}^{i+1}_{\boldsymbol{\theta}} = \left(-\frac{\partial^2 \ln g^i(\widehat{\boldsymbol{\theta}}^i)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \right)^{-1} \quad (6.108)$$

This converges to a point $\widehat{\boldsymbol{\theta}}^V$ for which $\frac{\partial \ln g^i(\widehat{\boldsymbol{\theta}}^V)}{\partial \boldsymbol{\theta}} = 0$ and for which $f^V(\boldsymbol{\theta}|\boldsymbol{y}) = \mathcal{N}(\widehat{\boldsymbol{\theta}}^V, \boldsymbol{C}^V_{\boldsymbol{\theta}})$. Now, we have for $g^i(\boldsymbol{\theta})$ in (6.104), and from (6.75)

$$\begin{aligned} g^i(\boldsymbol{\theta}) &= \mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})} \{\ln f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})\} \\ &= \ln f(\boldsymbol{y}|\boldsymbol{\theta}) + \mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})} \{\ln f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})\} \\ &\doteq -\frac{1}{2} \ln \det(\boldsymbol{C}(\boldsymbol{\theta})) - \frac{1}{2}\, \mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})} \left\{ \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}^T \boldsymbol{C}^{-1}(\boldsymbol{\theta}) \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \right\} \\ &= -\frac{1}{2} [\ln \det(\boldsymbol{C}(\boldsymbol{\theta})) + \mathrm{tr}\{\boldsymbol{C}^{-1}(\boldsymbol{\theta})\, \widehat{\boldsymbol{C}}^i\}] \end{aligned} \quad (6.109)$$

where now $\widehat{\boldsymbol{C}}^i = \mathrm{E}_{q^i(\boldsymbol{x}|\boldsymbol{y})} \{\boldsymbol{w}\boldsymbol{w}^T\}$ with $\boldsymbol{w}^T = [\boldsymbol{x}^T \boldsymbol{y}^T]$. Hence, the computation of $g^i(\boldsymbol{\theta})$ in VB is identical to that in EM except that $f(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^i)$ in EM is replaced by $q^i(\boldsymbol{x}|\boldsymbol{y})$ in VB. However, asymptotically, for the second-order expansion in (6.104), $q(\boldsymbol{x}|\boldsymbol{y})$ can equivalently be replaced by $f(\boldsymbol{x}|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}^i)$. Hence, asymptotically there is no information for $\boldsymbol{\theta}$ in $\mathrm{E}_{q(\boldsymbol{x}|\boldsymbol{y})} \{\ln f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})\}$ and

$$f^V(\boldsymbol{\theta}|\boldsymbol{y}) = f^E(\boldsymbol{\theta}|\boldsymbol{y}) = f^M(\boldsymbol{\theta}|\boldsymbol{y}) = \mathcal{N}(\widehat{\boldsymbol{\theta}}^M, \mathrm{CRB}^M) \quad (6.110)$$

For the estimator $\widehat{\boldsymbol{\theta}}$, VB is asymptotically equivalent to ML, nevertheless, this establishes that: (i) asymptotically, one can not do better than ML (and CRB^M!),

(*ii*) the convergence behavior of the VB iterations may be more interesting, and (*iii*) the VB performance may be better than ML non-asymptotically. To summarize the performances of the different estimators of $\boldsymbol{\theta}$:

- In the alternating MAP/ML, the estimate $\widehat{\boldsymbol{\theta}}$ is computed using only $\hat{\boldsymbol{x}}$ (as if $\hat{\boldsymbol{x}} = \boldsymbol{x}$). Reciprocally, $\hat{\boldsymbol{x}}$ is computed from $\widehat{\boldsymbol{\theta}}$ only (as if $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}$). The alternating MAP/ML converges to the joint MAP/ML.
- In EM, the estimate $\widehat{\boldsymbol{\theta}}$ is computed from $\hat{\boldsymbol{x}}$ and its estimation error $\widetilde{\boldsymbol{x}}$, whereas $\hat{\boldsymbol{x}}$ is computed only from $\widehat{\boldsymbol{\theta}}$. EM converges to the marginalized ML approach.
- In VB, $\widehat{\boldsymbol{\theta}}$ is computed from both $\hat{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{x}}$. Reciprocally, $\hat{\boldsymbol{x}}$ is computed from $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$. Asymptotically, VB has the same performance as ML and EM (hence efficient).

Note that all these iterative algorithms require only one iteration to converge if initialized with a consistent $\widehat{\boldsymbol{\theta}}^0$. From (6.75), we get for the VB update of $q(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\hat{\boldsymbol{x}}, \boldsymbol{C}_{\boldsymbol{x}})$

$$\begin{aligned}
\ln q^i(\boldsymbol{x}|\boldsymbol{y}) &\doteq \int q^i(\boldsymbol{\theta}|\boldsymbol{y}) \ln f(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{\theta} \\
&\doteq \int q^i(\boldsymbol{\theta}|\boldsymbol{y}) \ln f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\doteq -\frac{1}{2} \mathrm{E}_{q^i(\boldsymbol{\theta}|\boldsymbol{y})} \left\{ (\boldsymbol{x} - \boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{y})^T \boldsymbol{P}^{-1}(\boldsymbol{\theta}) (\boldsymbol{x} - \boldsymbol{F}(\boldsymbol{\theta})\boldsymbol{y}) \right\} \\
&\doteq \mathrm{E}_{q^i(\boldsymbol{\theta}|\boldsymbol{y})} \{ \boldsymbol{y}^T \boldsymbol{F}(\boldsymbol{\theta})^T \boldsymbol{P}^{-1}(\boldsymbol{\theta}) \boldsymbol{x} - \frac{1}{2} \boldsymbol{x}^T \boldsymbol{P}^{-1}(\boldsymbol{\theta}) \boldsymbol{x} \} \\
&\doteq -\frac{1}{2} (\boldsymbol{x} - \hat{\boldsymbol{x}}^i)^T (\boldsymbol{C}_{\boldsymbol{x}}^i)^{-1} (\boldsymbol{x} - \hat{\boldsymbol{x}}^i) \qquad (6.111)
\end{aligned}$$

where the Gaussian pdf comes out automatically. So we get

$$\hat{\boldsymbol{x}}^i = \boldsymbol{C}_{\boldsymbol{x}}^i \, \mathrm{E}_{q^i(\boldsymbol{\theta}|\boldsymbol{y})} \left\{ \boldsymbol{P}^{-1}(\boldsymbol{\theta}) \boldsymbol{F}(\boldsymbol{\theta}) \right\} \boldsymbol{y} \qquad (6.112)$$

$$\boldsymbol{C}_{\boldsymbol{x}}^i = \mathrm{E}_{q^i(\boldsymbol{\theta}|\boldsymbol{y})} \left\{ \boldsymbol{P}^{-1}(\boldsymbol{\theta}) \right\}^{-1} \qquad (6.113)$$

which can be computed (asymptotically) by second-order expansions in $\boldsymbol{\theta}$ of $\boldsymbol{P}^{-1}(\boldsymbol{\theta})$ and $\boldsymbol{P}^{-1}(\boldsymbol{\theta})\, \boldsymbol{F}(\boldsymbol{\theta})$. Asymptotically, when $q(\boldsymbol{\theta}|\boldsymbol{y})$ becomes $f(\boldsymbol{\theta}|\boldsymbol{y})$, $\hat{\boldsymbol{x}}^V$ attains a CRB corresponding to the following FIM

$$\begin{aligned}
\boldsymbol{J}_{\boldsymbol{x}\boldsymbol{x}}^V &= -\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}} \left\{ \frac{\partial^2 f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} \right\} \\
&= -\mathrm{E}_{\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}} \left\{ \frac{\partial^2 f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} \right\} \\
&= \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}} \left\{ \boldsymbol{P}^{-1}(\boldsymbol{\theta}) \right\} \qquad (6.114)
\end{aligned}$$

So, asymptotically, $\boldsymbol{C}_{\boldsymbol{x}}^V = \left(\boldsymbol{J}_{\boldsymbol{x}\boldsymbol{x}}^V \right)^{-1}$. Note that the VB update of $q(\boldsymbol{x}|\boldsymbol{y})$ is non-iterative in $\boldsymbol{x}$, due to the quadratic nature of $\ln f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$: $q(\boldsymbol{x}|\boldsymbol{y})$ is just a function

of $q(\boldsymbol{\theta}|\boldsymbol{y})$ (whereas $q^{i+1}(\boldsymbol{\theta}|\boldsymbol{y})$ depends on both $q^{i}(\boldsymbol{x}|\boldsymbol{y})$ and $q^{i}(\boldsymbol{\theta}|\boldsymbol{y})$). Hence the VB update for $\boldsymbol{x}$ is an extension of LMMSE estimation, accounting for the model parameter $\boldsymbol{\theta}$ inaccuracies. There has been some work in recent years to account for the channel estimation error in LMMSE receiver design [Piantanida *et al.* 2009].

6.3.3 Second-order extended LMMSE

Inspired by this state of the art presented before, one possible extension of LMMSE to account for parameter estimation performance would be to optimize a linear estimator $\boldsymbol{F}(\boldsymbol{\theta})$ on the basis of an extended MSE criterion

$$\text{MSE} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\big\{\,\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}\big\{||\boldsymbol{x}-\boldsymbol{F}(\boldsymbol{\theta})\,\boldsymbol{y}||^2\big\}\big\} \tag{6.115}$$

leading to

$$\widehat{\boldsymbol{F}} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left\{\boldsymbol{C}_{\boldsymbol{x}\boldsymbol{y}}(\boldsymbol{\theta})\right\}\ \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left\{\boldsymbol{C}_{\boldsymbol{y}\boldsymbol{y}}(\boldsymbol{\theta})\right\}^{-1} \tag{6.116}$$

which would be applicable also in the case of non-Gaussian $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ and $f(\boldsymbol{\theta}|\boldsymbol{y})$.

6.4 Conclusion

In this chapter, we tackled the issue of iterative ML algorithms performance. We surveyed different Kalman filter varieties in order to understand the difference between their behaviours before convergence. Then, we compared between the asymptotic behaviour of EM and VBEM in the Gaussian case. Due to the fact that in the VBEM, we use the estimation error of $\boldsymbol{\theta}$ in the estimation of $\boldsymbol{x}$, we expected that the VBEM performs better than the EM. Yet, asymptotically, we proved that similarly to EM , VBEM converges to the ML solution. Since it is not always possible in many applications to get placed in the asymptotic mode (limited number of samples, low source to noise ratio (SNR)), it would be interesting to compare between EM and VBEM in subasymptotic conditions where VBEM is expected to perform better than EM. In asymptotic performance analysis of ML, joint MAP/ML, etc., the Taylor series expansion of cost functions (marginal loglikelihood and/or joint loglikelihood) takes into consideration terms until $\mathcal{O}(1/N)$ and neglects higher-orders terms. We think that pushing the analysis to higher orders may clarify more the nuance between both.

Chapter 7

Conclusions and perspectives

7.1 Conclusions

In this thesis, we tackled two different topics. The first topic consists in finding new solution to the problem of mono-microphone source separation. Two main algorithms were proposed. The first algorithm, presented in chapter 3, consists in a time domain separation solution. A joint speech model is used where both time correlation and quasi-periodicity are exploited. The short time correlation is described with an autoregressive (AR) model of order 10. The long term correlation is described with a long-term model where only the coefficient of a pitch delay is nonzero. The long term model describes the quasi-periodicity nature of speech signals. The joint speech model is used to formulate two state space models whose parameters, namely, the state transition matrix, the observation and the state noise covariance depend on the speech model parameters, namely, the short term prediction (STP) and the long term prediction (LTP) coefficients, pitches and innovations powers in linear (1^{st} model) and bilinear (2^{nd} model) fashions. The separation problem is reduced to the estimation of the state vector and the parameters in both models linear and bilinear. In such case, an appropriate solution is to use the expectation maximization (EM)-Kalman algorithm. Since pitches are nonlinear parameters in this problem, a pitch estimation algorithm is used to estimate them in parallel of the main separation algorithm. Three varieties of algorithm are derived which are EM-Kalman for linear model, EM-Kalman with and without Rauch-Tung-Striebel (RTS) smoothing for bilinear model. The advantage of these algorithms is that, unlike the model-driven methods, they do not rely on a training step which is a costly pre-processing task. Simulations are achieved for both synthetic and real signals. For synthetic signals and when all the parameters are assumed known (filtering case), the algorithm converges to the wanted solution and shows good separation performance. When the parameters are unknown but appropriately initialized, the algorithm manages to achieve separation with performance close to the filtering case and converges to parameters close to the original ones. For real voiced signals, the proposed methods perform less than in synthetic signals. This is due to the limitations of the proposed methods as well as the used speech model, which are explained hereafter.

In fact, there are several limitation to these methods. First, the performance of

our algorithm is highly-dependent on the pitch estimation accuracy. In our simulations, we always assumed having some knowledge about the pitches of speakers by supposing that we can always estimate them from sequences where only one speaker is active. This also implies that these pitches will not vary in the other voiced sequences. In realistic context where these assumptions are no more true, the performance will decrease drastically. Moreover, the batch-iterative fashion used in our methods is more advantageous (compared to the adaptive processing) when the sources are stationary through the processing interval which is only true for very short voiced sequences. Yet for signals which are long enough to become non stationary, it is no longer the case, especially in terms of pitch variations that make the estimation of LTP coefficients as done here unmanageable. Another limitation of these methods is the initialization issue. Again, if we can not afford non overlapping sequences, and initialize the algorithm randomly, we risk to fall in a local minimum different from the desired solution or take a long time to converge. The second limitation is the number of active sources. In fact, we have to provide the algorithm with this information in order to construct the state space model properly. In a real context where the number of sources changes unpredictably, this model may not be easy to apply.

In addition to the previous limitations which are linked to the algorithm conception, there are limitations induced by the model itself [Chu 2003]. In the case where the involved pitches are fractional, the LTP is interpreted as two coefficients corresponding to the floor and ceil of the real pitch delay which is still an approximation that induces a certain estimation error. Moreover, it is utopian to think that the presented joint model matches perfectly all speech signals. For example, nasal sounds like n and m are better modeled with pole-zero type transfer function as cite in [Lim & Lee 1993].

In the second algorithm presented in chapter 4, we propose a frequency domain based method. We use the same speech model as in the previous chapter. The switch to the frequency domain is beneficial when there are filtering operations in time. Unlike the previous chapter where the parameters and sources are estimated jointly, in this chapter, the Gaussian sources are marginalized (eliminated), the parameters are estimated separately then used to estimate the sources afterward. Here, signals are processed piecewise using non trivial finite length window, which helps to handle the non stationarity better than with the proposed algorithms in chapter 3. The introduction of finite length window induce a smearing effect that we take into account on the formulation of the criterium used to estimate the parameters. We use the Gaussian maximum likelihood (GML) criterium that we show equivalent to Itakura-Saito distance (ISD), and to optimally weighted covariance matrix (OWCM) in terms of gradient (extrema). The smearing effect which is neglected in the asymp-

totic methods (extremely long frames) is considered by transforming the spectrum covariance matrix (used in GML) from diagonal to banded matrix (cross-correlation between only neighboring frequencies are modeled). This banded form is inherited from the non trivial choice of the window which should have highly attenuated sidelobes. Parameters estimation is achieved in an alternating optimization fashion. Like the proposed algorithms in chapter 3, the performance of this algorithm depends highly on the quality of pitch estimation. Moreover, the banded covariance matrix is in reality banded circulant, which means the existence of non zero components in the top right and bottom left corners of the matrix. By zeropadding the involved signals, we increase the total number of frequency samples (dimensions of the covariance matrix) so that the contribution of these components can be neglected and we get a purely banded matrix. Yet, it is definitely not the best way to handle this point, not only because we increase the algorithm complexity by increasing the dimensions of the covariance matrix but also that it is better to find a solution that takes into consideration the contributions of these components.

The second topic deals with the problem of joint optimal filtering and parameters estimation. In chapter 5, we tackled the issue of joint vs marginal estimation. We reviewed some of the lower bounds used as benchmarks to assess estimators of both types. We then investigated the effect of the presence of random (nuisance) variables on the estimation of deterministic parameters in terms of lower bounds in comparison to the case when these random variables are eliminated. A known result is that the Cramér-Rao bound (CRB) is tighter than hybrid Cramér-Rao bound (HCRB). We characterized the difference between both and find out an additive positive term interpreted as an inverse CRB of the joint estimation from the posterior of random variables. In literature, the modified Cramér-Rao bound (MCRB) is always defined in the case when the *a priori* distribution of random variables is independent of the deterministic parameters and this condition is used in the proof where the CRB is shown tighter than MCRB. We extend it to the case when they are dependent (such as the case in our monaural speech separation problem) which induce a change in the relation between the modified FIM (MFIM) and the Fisher information matrix (FIM). In this case, the well-known result of theCRB being tighter than the MCRB is no more true.

In chapter 6, we switch from the lower bounds to the estimation error investigation. Our focus was in comparing to joint estimators EM and variational Bayes (VB) where the former considers deterministic parameters while the latter suppose an *a priori* distribution for them. Our question was about the performance of VB compared to EM that converges asymptotically to the maximum likelihood (ML) solution. Using an ISD interpretation of both criteria in EM and VB and exploiting asymptotic Gaussian approximation for the parameters, we prove that, like EM, the

VB converges asymptotically to the ML solution.

7.2 Perspectives

7.2.1 Monaural speech separation

In this part, we suggest few ideas to improve the separation algorithms proposed in the thesis:

- In our work, we assumed that the number of active sources is known and constant through a time sequence. Since in reality it is not the case, the integration of an estimator of this element is necessary to avoid the problem of overfitting.
- In the case where we can not afford a good initialization of EM-Kalman, we can investigate the idea to use one of the accelerated forms of EM algorithms like space-alternating generalized EM (SAGE) method [Fessler & Hero 1994].
- The EM-Kalman based algorithms in chapter 3 can be extended to VB-Kalman based versions where the AR coefficients and the involved powers are modeled as random Gaussian and Gamma distributed variables respectively.

7.2.2 Joint optimal filtering and parameters estimation

- The proofs in chapter 6 are developed in the case of zero-mean Gaussian observations and latent data (only the covariance matrix depends on the parameters), an extension to the case where means also depend on the parameters should be investigated.
- An investigation of the subasymptotic performance of VB compared to EM and ML is also an interesting issue.

Appendix A

Appendix

A.1 Using orthogonal projection terminology in (HFIM-FIM) characterization

In this section, we show that using the orthogonal projection terminology used by Messer et al. in [Noam & Messer 2009], we can prove the same result in (5.56). We use the same notations used in chapter 5. In [Noam & Messer 2009], the authors prove the following inequality

$$\boldsymbol{J}^{-1} \geq \left(\tilde{\boldsymbol{J}}_{\theta} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^{T}\right)^{-1} \tag{A.1}$$

The RHS term represents the upper block of the HCRB for the estimation part $\boldsymbol{\theta}$. We introduce some important results/formulations presented in [Noam & Messer 2009], which are needed for this proof. In [Noam & Messer 2009], an interesting form of the HCRB inverse is presented

$$\tilde{\boldsymbol{J}}_{\theta} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^{T} = \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{\left(Z_{\theta} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\left(Z_{\theta} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)^{T}\right\} \tag{A.2}$$

where the L-dimensional vector Z_{θ} and the M-dimensional vector $Z_{\boldsymbol{x}}$ are presented respectively

$$Z_{\theta} = \frac{\partial \ln f\left(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}\right)}{\partial \theta} \tag{A.3a}$$

$$Z_{\boldsymbol{x}} = \frac{\partial \ln f\left(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{x}} \tag{A.3b}$$

They also prove the two following results

$$\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}\left\{\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right\} = 0 \tag{A.4a}$$

$$\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}\left\{Z_{\theta}\right\} = \frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \tag{A.4b}$$

and re-express the CRB inverse, $\boldsymbol{J}$, in a different interesting form

$$\boldsymbol{J} = \mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\left\{\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}\left\{\left(Z_{\theta} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\right\}\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}\left\{\left(Z_{\theta} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)^{T}\right\}\right\} \tag{A.5}$$

Using the fact that $\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}\{.\}$ is a projection operator, Messer et al. prove the inequality (A.1). The authors used the inequality between the norm of an element and

the norm of its projection : $\|u\|^2 \geq \|\boldsymbol{P}\, u\|^2$ where $\boldsymbol{P}$ is a given projector. In order to find analytic expression to the difference between the CRB inverse and HCRB inverse, we use the expression $\|u\|^2 = \|\boldsymbol{P}\, u\|^2 + \|(\boldsymbol{I}-\boldsymbol{P})\, u\|^2$. In what follows, the notations $\mathrm{E}_{\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}}\{.\}$ and $\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}\{.\}$ are simplified to $\mathrm{E}\{.\}$ and $\boldsymbol{P}$ respectively. By applying this formula to our problem gives the following expression

$$\tilde{\boldsymbol{J}}_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T = \boldsymbol{J} + \left\|(\boldsymbol{I}-\boldsymbol{P})\left(Z_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\right\|^2 \tag{A.6}$$

using the previous results, we can compute the second term of the RHS in (A.6)

$$\begin{aligned}\left\|(\boldsymbol{I}-\boldsymbol{P})\left(Z_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\right\|^2 = \mathrm{E}\Big\{ &\left((\boldsymbol{I}-\boldsymbol{P})\left(Z_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\right) \\ &\left((\boldsymbol{I}-\boldsymbol{P})\left(Z_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\right)^T\Big\}\end{aligned} \tag{A.7}$$

we have

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{P})\left(Z_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right) &= \frac{\partial \ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \frac{\partial \ln f(\boldsymbol{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}} \\ &= \frac{\partial \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\frac{\partial \ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{\partial\boldsymbol{x}} \\ &= \underbrace{\frac{\partial \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}}_{Y_\theta} - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\underbrace{\frac{\partial \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{x}}}_{Y_{\boldsymbol{x}}}\end{aligned} \tag{A.8}$$

inserting (A.8) in (A.7) results in

$$\begin{aligned}\left\|(\boldsymbol{I}-\boldsymbol{P})\left(Z_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\right\|^2 &= \mathrm{E}\left\{\left(Y_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Y_{\boldsymbol{x}}\right)\left(Y_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Y_{\boldsymbol{x}}\right)^T\right\} \\ &= \mathrm{E}\left\{Y_\theta Y_\theta^T\right\} + \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\mathrm{E}\left\{Y_{\boldsymbol{x}}Y_{\boldsymbol{x}}^T\right\}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T \\ &\quad - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\mathrm{E}\left\{Y_{\boldsymbol{x}}Y_\theta^T\right\} - E\left\{Y_\theta Y_{\boldsymbol{x}}^T\right\}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T\end{aligned} \tag{A.9}$$

It is easy to prove that $\mathrm{E}\left\{Y_{\boldsymbol{x}}Y_{\boldsymbol{x}}^T\right\} = \tilde{\boldsymbol{J}}_{\boldsymbol{x}}$ and $E\left\{Y_\theta Y_\theta^T\right\} = \tilde{\boldsymbol{G}}_\theta$. In fact :

$$\begin{aligned}\mathrm{E}\left\{Y_{\boldsymbol{x}}Y_{\boldsymbol{x}}^T\right\} &= \mathrm{E}\left\{\frac{\partial \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{x}}\frac{\partial \ln f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{x}^T}\right\} \\ &= \mathrm{E}\left\{\frac{\partial \ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{\partial\boldsymbol{x}}\frac{\partial \ln f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta})}{\partial\boldsymbol{x}^T}\right\} \\ &= \tilde{\boldsymbol{J}}_{\boldsymbol{x}}\end{aligned} \tag{A.10}$$

Idem for the term in $\mathrm{E}\left\{Y_\theta Y_\theta^T\right\}$:

$$\begin{aligned}
\mathrm{E}\left\{Y_\theta Y_\theta^T\right\} &= \mathrm{E}\left\{\frac{\partial \ln f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\frac{\partial \ln f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right\} \\
&= \int f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)d\boldsymbol{y}\int \frac{\partial f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\frac{\partial \ln f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}d\boldsymbol{x} \\
&= \int\int \frac{\partial}{\partial \boldsymbol{\theta}}\left(f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)\frac{\partial \ln f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right)d\boldsymbol{x}f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)d\boldsymbol{y} \\
&- \int\int f\left(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}\right)\frac{\partial\partial \ln f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}d\boldsymbol{x}d\boldsymbol{y} \\
&= \mathrm{E}\left\{-\frac{\partial\partial \ln f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\} \\
&= \tilde{\boldsymbol{G}}_\theta
\end{aligned} \tag{A.11}$$

The first term in the third RHS line of (A.11) disappears since when we invert the derivation w.r.t. $\boldsymbol{\theta}$ and the integration w.r.t $\boldsymbol{x}$, we get $\frac{\partial\partial}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\int f\left(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}\right)d\boldsymbol{x}$ which is equal to zero. For the crossed derivation term $\mathrm{E}\left\{Y_{\boldsymbol{x}} Y_\theta^T\right\}$. We notice that

$$\begin{aligned}
\mathrm{E}\left\{Y_{\boldsymbol{x}} Y_\theta^T\right\} &= \mathrm{E}\left\{Z_{\boldsymbol{x}} Z_\theta^T\right\} - \mathrm{E}\left\{Z_{\boldsymbol{x}}\frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right\} \\
&= \tilde{\boldsymbol{J}}_{\boldsymbol{x},\theta} - \mathrm{E}\left\{Z_{\boldsymbol{x}}\frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right\}
\end{aligned} \tag{A.12}$$

By multiplying both sides of (A.12) by $\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}$, we get

$$\begin{aligned}
\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\mathrm{E}\left\{Y_{\boldsymbol{x}} Y_\theta^T\right\} &= \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\boldsymbol{x},\theta} - \mathrm{E}\left\{\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right\} \\
&= \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\boldsymbol{x},\theta} - \mathrm{E}_{\boldsymbol{y}|\boldsymbol{\theta}}\left\{\underbrace{\boldsymbol{P}\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}}_{\mathbf{0}}\frac{\partial \ln f\left(\boldsymbol{y}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^T}\right\}
\end{aligned} \tag{A.13}$$

Replacing (A.10), (A.11) and (A.13) in (A.9) and using the fact that $\tilde{\boldsymbol{G}}_{\theta,\boldsymbol{x}} = \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}$ and $\tilde{\boldsymbol{G}}_{\boldsymbol{x}} = \tilde{\boldsymbol{J}}_{\boldsymbol{x}}$ from (5.55) results in

$$\begin{aligned}
\left\|(\boldsymbol{I}-\boldsymbol{P})\left(Z_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}Z_{\boldsymbol{x}}\right)\right\|^2 &= \tilde{\boldsymbol{G}}_\theta - \tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{J}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{J}}_{\theta,\boldsymbol{x}}^T \\
&= \tilde{\boldsymbol{G}}_\theta - \tilde{\boldsymbol{G}}_{\theta,\boldsymbol{x}}\tilde{\boldsymbol{G}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{G}}_{\theta,\boldsymbol{x}}^T
\end{aligned} \tag{A.14}$$

We end up with the same result in (5.56).

Appendix B

Miscellaneous mathematical utilities

In this appendix we will remind the mathematical tools we used in the thesis to derive the different results.

B.1 Determinant, trace and vec-operator rules

- Suppose $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ three matrices

$$\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{CAB}) = \operatorname{tr}(\boldsymbol{BCA}) \tag{B.1}$$

- If we denote by vec the operator that when applied on a matrix $\boldsymbol{A}$ stacks its columns into a vector, $\boldsymbol{B}$ and $\boldsymbol{X}$ are matrices

$$\begin{aligned} \operatorname{vec}(\boldsymbol{AXB}) &= \left(\boldsymbol{B}^T \otimes \boldsymbol{A}\right) \operatorname{vec}(\boldsymbol{X}) && \text{(B.2)} \\ \operatorname{tr}\left(\boldsymbol{A}^T\boldsymbol{B}\right) &= \operatorname{vec}(\boldsymbol{B})^T \operatorname{vec}(\boldsymbol{B}) && \text{(B.3)} \end{aligned}$$

- if the matrix $\boldsymbol{A}$ is a function of the scalar x

$$\begin{aligned} \frac{\partial \det(\boldsymbol{A})}{\partial x} &= \det(\boldsymbol{A}) \operatorname{tr}\left[\boldsymbol{A}^{-1}\frac{\partial(\boldsymbol{A})}{\partial x}\right] && \text{(B.4)} \\ \frac{\partial \boldsymbol{A}^{-1}}{\partial x} &= -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1} && \text{(B.5)} \end{aligned}$$

B.2 Matrix inversion Lemma

In chapter 5, we dealt with 2×2 block matrices and their inverses. Let's suppose the matrix system

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{G} & \boldsymbol{H} \end{bmatrix} \tag{B.6}$$

Then the block matrices of the RHS of (B.6) are expressed in the following system

$$\begin{aligned} \boldsymbol{E} &= \left(\boldsymbol{A} - \boldsymbol{BD}^{-1}\boldsymbol{C}\right)^{-1} && \text{(B.7)} \\ \boldsymbol{F} &= -\left(\boldsymbol{A} - \boldsymbol{BD}^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{BD}^{-1} && \text{(B.8)} \\ \boldsymbol{G} &= -\boldsymbol{D}^{-1}\boldsymbol{C}\left(\boldsymbol{A} - \boldsymbol{BD}^{-1}\boldsymbol{C}\right)^{-1} && \text{(B.9)} \\ \boldsymbol{H} &= \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\left(\boldsymbol{A} - \boldsymbol{BD}^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{BD}^{-1} && \text{(B.10)} \end{aligned}$$

Which is equivalent to

$$\boldsymbol{E} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \tag{B.11}$$

$$\boldsymbol{F} = -\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \tag{B.12}$$

$$\boldsymbol{G} = -\left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \tag{B.13}$$

$$\boldsymbol{H} = \left(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1} \tag{B.14}$$

Bibliography

[Abdallah & Plumbley 2004] Samer M. Abdallah and Mark D. Plumbley. *Polyphonic transcription by non-negative sparse coding of power spectra.* In ISMIR, 2004. (Cited on page 15.)

[Abeysekera 2004] S.S. Abeysekera. *Multiple pitch estimation of poly-phonic audio signals in a frequency-lag domain using the bispectrum.* In Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on, volume 3, pages III–469–72 Vol.3, 2004. (Cited on page 37.)

[Aitkin & Aitkin 1996] Murray Aitkin and Irit Aitkin. *A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions.* Statistics and Computing, vol. 6, no. 2, pages 127–130, 1996. (Cited on page 100.)

[Anderson & Moore 1979] B.D.O. Anderson and J.B. Moore. Optimal Filtering. Prentice Hall (or more recently: Dover, 2005), 1979. (Cited on page 89.)

[Anguera Miro *et al.* 2012] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals. *Speaker Diarization: A Review of Recent Research.* Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 2, pages 356–370, 2012. (Cited on page 9.)

[Åström 1980] K.J. Åström. *Maximum likelihood and prediction error methods.* Automatica, vol. 16, no. 5, pages 551 – 574, 1980. (Cited on page 90.)

[Atal & Hanauer 1971] B. S. Atal and Suzanne L. Hanauer. *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave.* The Journal of the Acoustical Society of America, vol. 50, no. 2B, pages 637–655, 1971. (Cited on pages 10 and 23.)

[Atal & Remde 1982] B.S. Atal and J. Remde. *A new model of LPC excitation for producing natural-sounding speech at low bit rates.* In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82., volume 7, pages 614–617, 1982. (Cited on page 10.)

[Athans *et al.* 1968] M. Athans, R. Wishner and A. Bertolini. *Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements.* Automatic Control, IEEE Transactions on, vol. 13, no. 5, pages 504 – 514, oct 1968. (Cited on pages 86 and 90.)

[Barankin 1949] E. W. Barankin. *Locally best unbiased estimates.* Ann. Math. Statist., vol. 20, no. 4, pages 477–501, Dec. 1949. (Cited on page 73.)

[Barros & Cichocki 2001] Allan Kardec Barros and Andrzej Cichocki. *Extraction of Specific Signals with Temporal Structure*. Neural Comput., vol. 13, no. 9, pages 1995–2003, 2001. (Cited on page 21.)

[Bass *et al.* 1966] R. D. Bass, V. D. Norum and L. Swartz. *Optimal multichannel nonlinear filtering*. J. Mufh. Anal. Appl., vol. 16, pages 152 – 164, 1966. (Cited on page 86.)

[Bay *et al.* 2008] S. Bay, Benoit Geller, A. Renaux, J. P Barbot and J. Brossier. *On the Hybrid Cramér Rao Bound and Its Application to Dynamical Phase Estimation*. Signal Processing Letters, IEEE, vol. 15, pages 453–456, 2008. (Cited on page 81.)

[Beal 2003] M.J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. Phd, University of Cambridge, UK, 2003. (Cited on pages 2, 85, 95, 96, 97 and 98.)

[Benaroya & Bimbot 2003] Laurent Benaroya and Frédéric Bimbot. *Wiener based source separation with HMM/GMM using a single sensor*. In Proc. 4th Int. Symp. on Independent Component Anal. and Blind Signal Separation, pages 957–961, 2003. (Cited on pages 11 and 14.)

[Benaroya *et al.* 2003] L. Benaroya, L.M. Donagh, F. Bimbot and R. Gribonval. *Non negative sparse representation for Wiener based source separation with a single sensor*. In Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, volume 6, pages VI–613–16 vol.6, 2003. (Cited on page 14.)

[Benaroya *et al.* 2006] L. Benaroya, F. Bimbot and R. Gribonval. *Audio source separation with a single sensor*. Audio, Speech, and Language Process., IEEE Trans. on, vol. 14, no. 1, pages 191–199, 2006. (Cited on pages 11 and 14.)

[Benaroya 2003] E.L. Benaroya. *Séparation de plusieurs sources sonores avec un seul microphone*. PhD thesis, l'Universités de Rennes 1, Rennes, France, Juin 2003. (Cited on page 8.)

[Benetos *et al.* 2006] E. Benetos, M. Kotti and C. Kotropoulos. *Musical instrument classification using non-negative matrix factorization algorithms*. In Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on, pages 1844–1847, 2006. (Cited on page 15.)

[Bensaid & Slock 2013] Siouar Bensaid and Dirk T M Slock. *On the performance of joint linear minimum mean squared error (LMMSE) filtering and parameter*

estimation. In SPAWC 2013, 14th IEEE International Workshop on Signal Processing Advances in Wireless Communications, June 16-19, 2013, Darmstadt, Germany, Darmstadt, ALLEMAGNE, 06 2013. (Cited on page 4.)

[Bensaid *et al.* 2010a] S. Bensaid, A. Schutz and D.T.M. Slock. *Single Microphone Blind Audio Source Separation Using EM-Kalman Filter and Short+Long Term AR Modeling*. In Proc. Int'l Conf. on Latent Variable Analysis and Signal Separation (LVA-ICA), Saint-Malo, France, Sept. 2010. (Cited on page 86.)

[Bensaid *et al.* 2010b] Siouar Bensaid, Antony Schutz and Dirk T M Slock. *Single Microphone Blind Audio Source Separation Using EM-Kalman Filter and Short+Long Term AR Modeling*. In LVA ICA 2010, 9th Int'l Conf. on Latent Variable Analysis and Signal Separation, September 27-30, Saint-Malo, France / Also published in LNCS, 2010, 2010. (Cited on pages 3, 50 and 85.)

[Bertin *et al.* 2010] N. Bertin, R. Badeau and E. Vincent. *Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription*. Audio, Speech, and Lang. Process., IEEE Trans., vol. 18, no. 3, pages 538–549, 2010. (Cited on page 15.)

[Bhattacharya 1966] P. K. Bhattacharya. *Estimating the Mean of a Multivariate Normal Population with General Quadratic Loss Function*. Ann. Math. Statist., vol. 37, no. 6, pages 1819–1824, Dec. 1966. (Cited on page 72.)

[Bobrovsky & Zakai 1976] B. Bobrovsky and M. Zakai. *A lower bound on the estimation error for certain diffusion processes*. Information Theory, IEEE Transactions on, vol. 22, no. 1, pages 45–52, 1976. (Cited on page 79.)

[Bregman 1990] A. S. Bregman. Auditory scene analysis. MIT Press, Cambridge MA, 1990. (Cited on page 12.)

[Carlson & Clements 1991a] B.A. Carlson and M.A. Clements. *A Computationally Compact Divergence Measure for Speech Processing*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1991. (Cited on page 50.)

[Carlson & Clements 1991b] B.A. Carlson and M.A. Clements. *A computationally compact divergence measure for speech processing*. Pattern Analysis and Machine Intelligence, IEEE Trans., vol. 13, no. 12, pages 1255–1260, 1991. (Cited on page 52.)

[Carlson 1988] B.D. Carlson. *Covariance matrix estimation errors and diagonal loading in adaptive arrays*. Aerospace and Electronic Systems, IEEE Transactions on, vol. 24, no. 4, pages 397–401, 1988. (Cited on page 72.)

[Catovic & Sahinoglu 2004] A. Catovic and Z. Sahinoglu. *The Cramer-Rao bounds of hybrid TOA/RSS and TDOA/RSS location estimation schemes.* Communications Letters, IEEE, vol. 8, no. 10, pages 626–628, 2004. (Cited on page 72.)

[Cemgil 2004] A. T. Cemgil. *Bayesian Music Transcription.* PhD thesis, Radboud University of Nijmegen, 2004. (Cited on page 37.)

[Chazan *et al.* 1993] D. Chazan, Y. Stettiner and D. Malah. *Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation.* In Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE Int'l Conf. on, April 1993. (Cited on pages 9, 10, 37 and 49.)

[Chen *et al.* 2006] Jingdong Chen, J. Benesty, Yiteng Huang and S. Doclo. *New insights into the noise reduction Wiener filter.* Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 4, pages 1218–1234, 2006. (Cited on page 13.)

[Cho & Choi 2005] Yong-Choon Cho and Seungjin Choi. *Nonnegative features of spectro-temporal sounds for classification.* Pattern Recognition Letters, vol. 26, no. 9, pages 1327 – 1336, 2005. (Cited on page 15.)

[Cho *et al.* 2003] Yong-Choon Cho, Seungjin Choi and Sung-Yang Bang. *Non-negative component parts of sound for classification.* In Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium on, pages 633–636, 2003. (Cited on page 15.)

[Christensen & Holdt Jensen 2006] Mads G. Christensen and Sø ren Holdt Jensen. *New Results in Rate-Distortion Optimized Parametric Audio Coding.* In Audio Engineering Society Convention 120, May 2006. (Cited on page 37.)

[Christensen & Jakobsson 2010] M.G. Christensen and A. Jakobsson. *Optimal Filter Designs for Separating and Enhancing Periodic Signals.* Signal Processing, IEEE Trans. on, Dec. 2010. (Cited on pages 37 and 49.)

[Christensen *et al.* 2006] M.G. Christensen, A. Jakobsson and S.H. Jensen. *Multi-Pitch Estimation Using Harmonic Music.* In Signals, Systems and Computers, 2006. ACSSC '06. Fortieth Asilomar Conference on, pages 521–524, 2006. (Cited on page 37.)

[Christensen *et al.* 2008] M.G. Christensen, J.H. Jensen, A. Jakobsson and S.H. Jensen. *On Optimal Filter Designs for Fundamental Frequency Estimation.* Signal Processing Letters, IEEE, vol. 15, pages 745–748, 2008. (Cited on page 37.)

[Christensen *et al.* 2009] Mads Christensen, Andreas Jakobsson and Juang B.H. Multi-pitch estimation. Morgan & Claypool, 2009. (Cited on pages 9 and 37.)

[Chu 2003] Wai. C. Chu. Speech coding algorithms: foundation and evolution of standardized coders. John Wiley and Sons, NewYork, 2003. (Cited on pages 10, 23 and 106.)

[Chu 2004] Wai C. Chu. Linear prediction, pages 91–142. John Wiley & Sons, Inc., 2004. (Cited on page 24.)

[Cichocki & Thawonmas 2000] Andrzej Cichocki and Ruck Thawonmas. *On-line Algorithm for Blind Signal Extraction of Arbitrarily Distributed, but Temporally Correlated Sources Using Second Order Statistics.* Neural Process. Lett., vol. 12, no. 1, pages 91–98, 2000. (Cited on page 21.)

[Cichocki *et al.* 2000] A. Cichocki, T. Rutkowski, A.K. Barros and Sang-Hoon Oh. *A blind extraction of temporally correlated but statistically dependent acoustic signals.* In Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop, volume 1, pages 455–464 vol.1, 2000. (Cited on page 10.)

[Comon & Jutten 2010] Pierre Comon and Christian Jutten. Handbook of Blind Source Separation, Independent Component Analysis and Applications. Academic Press (Elsevier), February 2010. (Cited on page 7.)

[Comon 1994] Pierre Comon. *Independent Component Analysis, a New Concept?* Signal Process., vol. 36, no. 3, pages 287–314, April 1994. (Cited on page 14.)

[Consortium Partners 2010] Consortium Partners. *Deliverable D2.3 "Hybrid Localization Techniques".* Technical report, EC FP7 project WHERE, May 2010. URL: http://www.kn-s.dlr.de/where/public_documents_deliverables.php. (Cited on page 86.)

[Cooke *et al.* 2006] Martin Cooke, Jon Barker, Stuart Cunningham and Xu Shao. *An audio-visual corpus for speech perception and automatic speech recognition.* The Journal of the Acoustical Society of America, vol. 120, no. 5, pages 2421–2424, November 2006. (Cited on page 46.)

[Cooke *et al.* 2010] Martin Cooke, John R. Hershey and Steven J. Rennie. *Monaural speech separation and recognition challenge.* Computer Speech & Language, vol. 24, no. 1, pages 1 – 15, 2010. (Cited on page 8.)

[Couvreur & Bresler 1995] C. Couvreur and Y. Bresler. *Decomposition of a mixture of Gaussian AR processes.* Acoustics, Speech, and Signal Processing,

IEEE International Conference on, vol. 3, pages 1605–1608, 1995. (Cited on pages 27 and 86.)

[Cramér 1946] Harald Cramér. Mathematical methods of statistics. Princeton, NJ: Princeton Univ. Press., 1946. (Cited on page 71.)

[D'Andrea *et al.* 1994] A.N. D'Andrea, U. Mengali and R. Reggiannini. *The Modified Cramér-Rao Bound and its Application to Synchronization Problems.* IEEE Trans. Communications, Feb/Mar/Apr 1994. (Cited on pages 74 and 75.)

[Davies & James 2007] M.E. Davies and C.J. James. *Source separation using single channel {ICA}.* Signal Processing, vol. 87, no. 8, pages 1819 – 1832, 2007. (Cited on page 15.)

[de Carvalho & Slock 1999] E. de Carvalho and D.T.M. Slock. *A fast Gaussian maximum-likelihood method for blind multichannel estimation.* In Signal Processing Advances in Wireless Communications, 1999. SPAWC '99. 2nd IEEE Workshop on, 1999. (Cited on page 50.)

[Demoment 1989] G. Demoment. *Image reconstruction and restoration: overview of common estimation structures and problems.* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 37, no. 12, pages 2024–2036, 1989. (Cited on page 72.)

[Dempster *et al.* 1977] A. P. Dempster, N. M. Laird and D. B. Rubin. *Maximum Likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, Series B, vol. 39, no. 1, pages 1–38, 1977. (Cited on pages 27, 28, 31, 85, 87 and 99.)

[Dhaouadi *et al.* 1991] R. Dhaouadi, N. Mohan and L. Norum. *Design and implementation of an extended Kalman filter for the state estimation of a permanent magnet synchronous motor.* Power Electronics, IEEE Transactions on, vol. 6, no. 3, pages 491–497, Jul 1991. (Cited on page 89.)

[Eldar 2006] Y.C. Eldar. *Uniformly Improving the Cramér-Rao Bound and Maximum-Likelihood Estimation.* Signal Processing, IEEE Transactions on, vol. 54, no. 8, pages 2943–2956, 2006. (Cited on page 72.)

[Ellis & Weiss 2006] D.P.W. Ellis and R.J. Weiss. *Model-Based Monaural Source Separation Using a Vector-Quantized Phase-Vocoder Representation.* In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, volume 5, pages 957–960, 2006. (Cited on page 16.)

[Emiya *et al.* 2011] V. Emiya, E. Vincent, N. Harlander and V. Hohmann. *Subjective and Objective Quality Assessment of Audio Source Separation.* Audio, Speech, and Lang. Process., IEEE Trans., vol. 19, no. 7, pages 2046–2057, 2011. (Cited on page 18.)

[Ephraim 1992] Y. Ephraim. *Statistical-model-based speech enhancement systems.* Proceedings of the IEEE, vol. 80, no. 10, pages 1526–1555, 1992. (Cited on page 13.)

[Erkelens 1996] J.S. Erkelens. Autoregressive modelling for speech coding: Estimation, interpolation and quantisation. Delft University Press, 1996. (Cited on page 10.)

[Feder & Weinstein 1988] M. Feder and E. Weinstein. *Parameter estimation of superimposed signals using the EM algorithm .* Acoustics, Speech and Signal Processing, IEEE Trans., vol. 36, no. 4, pages 477–489, Apr 1988. (Cited on pages 27 and 86.)

[Fessler & Hero 1993] J.A. Fessler and A.O. Hero. *Complete-data spaces and generalized EM algorithms.* In Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, volume 4, pages 1–4 vol.4, April 1993. (Cited on page 101.)

[Fessler & Hero 1994] J.A. Fessler and A.O. Hero. *Space-alternating generalized expectation-maximization algorithm.* Signal Process., IEEE Trans., vol. 42, no. 10, pages 2664–2677, 1994. (Cited on pages 100 and 108.)

[Févotte *et al.* 2009] C. Févotte, N. Bertin and J.-L. Durrieu. *Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis.* Neural Computation, vol. 21, no. 3, pages 793–830, Mar. 2009. (Cited on page 52.)

[Févotte 2011] C. Févotte. *Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization.* In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, May 2011. (Cited on page 52.)

[F.Itakura & S.Saito 1970] F.Itakura and S.Saito. *A statistical method for estimation of speech spectral density and formant frequencies.* Electronics & Communications in Japan, pages 53A: 36–43, 1970. (Cited on page 52.)

[Gannot *et al.* 1997] S. Gannot, D. Burshtein and E. Weinstein. *Iterative-Batch and Sequential Algorithms for Single Microphone Speech Enhancement.* Acous-

tics, Speech, and Signal Processing, IEEE International Conference on, vol. 2, page 1215, 1997. (Cited on page 10.)

[Gao *et al.* 2003] Wen Gao, Tsai S. and J.S. Lehnert. *Diversity combining for DS/SS systems with time-varying, correlated fading branches.* Communications, IEEE Transactions on, vol. 51, no. 2, pages 284–295, Feb 2003. (Cited on pages 27, 86 and 92.)

[Ghahramani & Beal 2000] Zoubin Ghahramani and Matthew J. Beal. *Variational Inference for Bayesian Mixtures of Factor Analysers.* In In Advances in Neural Information Processing Systems 12, pages 449–455. MIT Press, 2000. (Cited on page 97.)

[Ghahramani & Beal 2001] Zoubin Ghahramani and Matthew J. Beal. *Propagation Algorithms for Variational Bayesian Learning.* In In Advances in Neural Information Processing Systems 13, pages 507–513. MIT Press, 2001. (Cited on page 95.)

[Giacobello *et al.* 2009a] D. Giacobello, M.G. Christensen, J. Dahl, S.H. Jensen and M. Moonen. *Joint estimation of short-term and long-term predictors in speech coders.* In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, pages 4109–4112, April 2009. (Cited on page 24.)

[Giacobello *et al.* 2009b] D. Giacobello, M.G. Christensen, J. Dahl, S.H. Jensen and M. Moonen. *Joint estimation of short-term and long-term predictors in speech coders.* In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE Int'l Conf. on, April 2009. (Cited on page 49.)

[Gini & Reggiannini 2000] F. Gini and R. Reggiannini. *On the use of Cramér-Rao-like Bounds in the Presence of Random Nuisance Parameters.* IEEE Trans. Communications, Dec. 2000. (Cited on pages 76, 78 and 81.)

[Gini *et al.* 1998] F. Gini, R. Reggiannini and U. Mengali. *The Modified Cramér-Rao Bound in Vector Parameter Estimation.* IEEE Trans. Communications, Jan. 1998. (Cited on pages 74 and 75.)

[Gold & Rabiner 1969] B. Gold and L. Rabiner. *Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain.* The Journal of the Acoustical Society of America, vol. 46, no. 2B, pages 442–448, 1969. (Cited on page 37.)

[Guarnieri & Tebaldini 2007] A.M. Guarnieri and S. Tebaldini. *Hybrid Cramér-Rao Bounds for Crustal Displacement Field Estimators in SAR Interferometry.*

Signal Processing Letters, IEEE, vol. 14, no. 12, pages 1012–1015, 2007. (Cited on page 85.)

[H. Viste & G. Evangelista 2001] H. Viste and G. Evangelista. *Sound Source Separation: Preprocessing for Hearing Aids and Structured Audio Coding.* In Proc. of Digital Audio Effects Conf. (DAFx '01), pages 67–70, Limerick, Ireland, Dec. 2001. (Cited on page 8.)

[Henriksen 1982] R. Henriksen. *The truncated second-order nonlinear filter revisited.* Automatic Control, IEEE Transactions on, vol. 27, no. 1, pages 247 – 251, feb 1982. (Cited on page 86.)

[Hess 1983] W. Hess. Pitch determination of speech signals. Springer-Verlag, Berlin, 1983. (Cited on page 37.)

[Hess 2008] WolfgangJ. Hess. *Pitch and Voicing Determination of Speech with an Extension Toward Music Signals.* In Jacob Benesty, M.Mohan Sondhi and Yiteng(Arden) Huang, editeurs, Springer Handbook of Speech Processing, pages 181–212. Springer Berlin Heidelberg, 2008. (Cited on page 37.)

[Holzapfel & Stylianou 2008] A. Holzapfel and Y. Stylianou. *Musical Genre Classification Using Nonnegative Matrix Factorization-Based Features.* Audio, Speech, and Language Processing, IEEE Transactions on, vol. 16, no. 2, pages 424–434, 2008. (Cited on page 15.)

[Hu & Wang 2004] Guoning Hu and DeLiang Wang. *Monaural speech segregation based on pitch tracking and amplitude modulation.* Neural Networks, IEEE Trans. on, vol. 15, no. 5, pages 1135–1150, 2004. (Cited on pages 12 and 13.)

[Itakura 1975] F. Itakura. *Minimum prediction residual principle applied to speech recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 23, no. 1, pages 67–72, 1975. (Cited on page 52.)

[Jamshidian & Jennrich 1993] Mortaza Jamshidian and Robert I. Jennrich. *Conjugate Gradient Acceleration of the EM Algorithm.* Journal of the American Statistical Association, vol. 88, no. 421, pages 221–228, 1993. (Cited on page 100.)

[Jamshidian & Jennrich 1997] Mortaza Jamshidian and Robert I. Jennrich. *Acceleration of the EM Algorithm by Using Quasi-Newton Methods.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 59, no. 3, pages 569–587, 1997. (Cited on page 100.)

[Jang & Lee 2002] Gil-Jin Jang and Te-Won Lee. *A Probabilistic Approach to Single Channel Blind Signal Separation.* In Suzanna Becker, Sebastian Thrun and Klaus Obermayer, editeurs, NIPS, pages 1173–1180. MIT Press, 2002. (Cited on page 15.)

[Jang *et al.* 2003] Gil-Jin Jang, Te-Won Lee and Yung-Hwan Oh. *Single-channel signal separation using time-domain basis functions.* Signal Processing Letters, IEEE, vol. 10, no. 6, pages 168–171, 2003. (Cited on pages 15 and 16.)

[Jansson & Stoica 1999] Magnus Jansson and Petre Stoica. *Forward-only and forward-backward sample covariances - A comparative study.* Signal Processing, vol. 77, no. 3, pages 235 – 245, 1999. (Cited on page 37.)

[Jazwinski 1970] A. H. Jazwinski. Stochastic processes and filtering theory. Academic Press, 1970. (Cited on pages 86 and 90.)

[Jensen & Hansen 2001] J. Jensen and J.H.L. Hansen. *Speech enhancement using a constrained iterative sinusoidal model.* Speech and Audio Processing, IEEE Transactions on, Oct. 2001. (Cited on page 49.)

[Jensen *et al.* 2010] J.R. Jensen, M.G. Christensen and S.H. Jensen. *An optimal spatio-temporal filter for extraction and enhancement of multi-channel periodic signals.* In Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conf. on, Nov. 2010. (Cited on page 49.)

[Jensen *et al.* 2012] J.R. Jensen, J. Benesty, M.G. Christensen and S.H. Jensen. *Enhancement of Single-Channel Periodic Signals in the Time-Domain.* Audio, Speech, and Language Processing, IEEE Trans. on, Sept. 2012. (Cited on page 49.)

[jin Jang *et al.* 2003] Gil jin Jang, Te won Lee, Te won Lee, Jean françois Cardoso, Erkki Oja and Shun ichi Amari. *A maximum likelihood approach to single-channel source separation.* Journal of Machine Learning Research, vol. 4, pages 1365–1392, 2003. (Cited on pages 15 and 16.)

[Kabal & Ramachandran 1989] P. Kabal and R.P. Ramachandran. *Joint optimization of linear predictors in speech.* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 37, no. 5, pages 642–650, May 1989. (Cited on page 24.)

[Kabal 2005] P. Kabal. *Windows for Transform Processing.* Technical report, electrical & computer engineering, McGill University, Dec. 2005. (Cited on page 51.)

[Kalman 1960] Rudolph Emil Kalman. *A New Approach to Linear Filtering and Prediction Problems*. Transactions of the ASME–Journal of Basic Engineering, vol. 82, no. Series D, pages 35–45, 1960. (Cited on pages 27 and 86.)

[Kay 1993] Steven M. Kay. Fundamentals of statistical signal processing: Estimation theory. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. (Cited on page 78.)

[Kim *et al.* 1994] Y.-R. Kim, Seung-Ki Sul and M.-H. Park. *Speed sensorless vector control of induction motor using extended Kalman filter*. Industry Applications, IEEE Transactions on, vol. 30, no. 5, pages 1225–1233, Sep 1994. (Cited on page 89.)

[Knockaert 1997] L. Knockaert. *The Barankin bound and threshold behavior in frequency estimation*. Signal Processing, IEEE Transactions on, vol. 45, no. 9, pages 2398–2401, 1997. (Cited on page 73.)

[Kostek 2004] B. Kostek. *Musical instrument classification and duet analysis employing music information retrieval techniques*. Proceedings of the IEEE, vol. 92, no. 4, pages 712–729, 2004. (Cited on page 37.)

[Kristjansson *et al.* 2004] Trausti Kristjansson, H. Attias and J. Hershey. *Single microphone source separation using high resolution signal reconstruction*. In Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on, volume 2, pages ii–817–20 vol.2, 2004. (Cited on page 16.)

[Kushner 1967] H.J. Kushner. *Approximations to optimal nonlinear filters*. Automatic Control, IEEE Transactions on, vol. 12, no. 5, pages 546–556, October 1967. (Cited on page 91.)

[Larsson 2004] E.G. Larsson. *Cramer-Rao bound analysis of distributed positioning in sensor networks*. Signal Processing Letters, IEEE, vol. 11, no. 3, pages 334–337, 2004. (Cited on page 72.)

[Lee & Lewicki 2000] T.-W. Lee and M. S. Lewicki. *The generalized Gaussian mixture model using ICA*. International Workshop on Independent Component Analysis (ICA'00), pages 239–244, June 2000. (Cited on page 15.)

[Lee & Seung 1996] Daniel D. Lee and H. Sebastian Seung. *Unsupervised Learning by Convex and Conic Coding*. In NIPS, pages 515–521, 1996. (Cited on page 15.)

[Lee & Seung 2000] Daniel D. Lee and H. Sebastian Seung. *Algorithms for Non-negative Matrix Factorization.* In NIPS, pages 556–562, 2000. (Cited on pages 15 and 16.)

[Lee *et al.* 1999] Te-Won Lee, M.S. Lewicki, M. Girolami and T.J. Sejnowski. *Blind source separation of more sources than mixtures using overcomplete representations.* Signal Processing Letters, IEEE, vol. 6, no. 4, pages 87–90, 1999. (Cited on page 1.)

[Lefèvre *et al.* 2011] A. Lefèvre, F. Bach and C. Févotte. *Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence.* In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Mohonk, NY, Oct. 2011. (Cited on page 52.)

[Lehmann & Casella 1998] E. L. Lehmann and G. Casella. Theory of point estimation. New York, NY:Springer-Verlag, second edition édition, 1998. (Cited on page 71.)

[Lenardi & Slock 2002] M. Lenardi and D.T.M. Slock. *Estimation of Time-Varying Wireless Channels and Application to the UMTS W-CDMA FDD Downlink.* In Proc. European Wireless (EW), Florence, Italy, Feb. 2002. (Cited on page 86.)

[Li *et al.* 2008] Jian Li, Luzhou Xu, Petre Stoica, K.W. Forsythe and D.W. Bliss. *Range Compression and Waveform Optimization for MIMO Radar: A Cramér Rao Bound Based Study.* Signal Processing, IEEE Trans., vol. 56, no. 1, pages 218–232, 2008. (Cited on page 72.)

[Li *et al.* 2010] Peng Li, Yong Guan, Shijin Wang, Bo Xu and Wenju Liu. *Monaural speech separation based on {MAXVQ} and {CASA} for robust speech recognition.* Computer Speech & Language, vol. 24, no. 1, pages 30 – 44, 2010. (Cited on pages 8 and 12.)

[Lim & Lee 1993] Il-Taek Lim and Byeong Gi Lee. *Lossless pole-zero modeling of speech signals.* Speech and Audio Processing, IEEE Transactions on, vol. 1, no. 3, pages 269–276, 1993. (Cited on page 106.)

[Lindblom 2005] J. Lindblom. *A Sinusoidal Voice Over Packet Coder Tailored for the Frame-Erasure Channel.* Speech and Audio Process., IEEE Trans., vol. 13, no. 5, pages 787–798, 2005. (Cited on page 37.)

[Lindsey 1972] W.C. Lindsey. Synchronization systems in communication and control. Pearson Education Ltd, 1972. (Cited on page 85.)

[Ljung & Söderström 1983] L. Ljung and T. Söderström. Theory and Practice of Recursive Identification. MIT Press, Cambridge, MA, 1983. (Cited on page 89.)

[Ljung 1979] L. Ljung. *Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems.* IEEE Transactions on Automatic Control, vol. 24, no. 1, pages 36 – 50, feb 1979. (Cited on pages 87 and 89.)

[Ljung 2002] L. Ljung. System Identification: Theory for the User. Prentice-Hall, Upper Saddle River, NJ, 2002. 2nd edition. (Cited on page 89.)

[Ljungquist & Balchen 1993] Dag Ljungquist and Jens G. Balchen. *Recursive prediction error methods for online estimation in nonlinear state-space models.* In Decision and Control, 1993., Proceedings of the 32nd IEEE Conference on, pages 714–719 vol.1, Dec 1993. (Cited on page 90.)

[MacKay 1997] David J.C. MacKay. *Ensemble Learning for Hidden Markov Models.* Technical report, Cavendish Laboratory, University of Cambridge, 1997. (Cited on page 97.)

[Makino *et al.* 2007] S. Makino, T.W. Lee and H. Sawada. Blind speech separation. Signals and communication technology. Springer, 2007. (Cited on page 14.)

[McAulay & Hofstetter 1971] R. McAulay and E.M. Hofstetter. *Barankin Bounds on Parameter Estimation.* Inf. Theory, IEEE Trans., vol. 17, no. 6, pages 669–676, 1971. (Cited on page 73.)

[McAulay & Seidman 1969] R. McAulay and L.P. Seidman. *A useful form of the Barankin lower bound and its application to PPM threshold analysis.* Information Theory, IEEE Transactions on, vol. 15, no. 2, pages 273–279, 1969. (Cited on page 73.)

[McLachlan & Krishnan 2008] Geoffrey J. McLachlan and Thriyambakam Krishnan. The em algorithm and extensions. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed édition, 2008. (Cited on page 100.)

[Mehra 1970] R. Mehra. *On the identification of variances and adaptive Kalman filtering.* IEEE Transactions on Automatic Control, vol. 15, no. 2, pages 175 – 184, apr 1970. (Cited on page 87.)

[Melia & Rickard 2007] Thomas Melia and Scott Rickard. *Underdetermined Blind Source Separation in Echoic Environments Using DESPRIT.* EURASIP J. Appl. Signal Process., no. 1, pages 90–90, January 2007. (Cited on page 1.)

[Meyer & Simmer 1997] J. Meyer and K.U. Simmer. *Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction.* In Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, volume 2, pages 1167–1170 vol.2, 1997. (Cited on page 13.)

[Middleton 1972] D. Middleton. *Statistical-Physical Models of Urban Radio-Noise Environments - Part I: Foundations.* Electromagnetic Compatibility, IEEE Transactions on, vol. EMC-14, no. 2, pages 38–56, 1972. (Cited on page 75.)

[Miller & Chang 1978] R. Miller and C. Chang. *A Modified Cramér-Rao Bound and its Applications.* IEEE Trans. Information Theory, May 1978. (Cited on pages 75 and 76.)

[Moeneclaey 1998] M. Moeneclaey. *On the True and the Modified Cramér-Rao Bounds for the Estimation of a Scalar Parameter in the Presence of Nuisance Parameters.* IEEE Trans. Communications, Nov. 1998. (Cited on pages 76, 77 and 83.)

[Mowlaee *et al.* 2010a] P. Mowlaee, M.G. Christensen and S.H. Jensen. *Improved single-channel speech separation using sinusoidal modeling.* In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE Int'l Conf. on, March 2010. (Cited on page 49.)

[Mowlaee *et al.* 2010b] P. Mowlaee, R. Saeidi, Z.-H. Tan, M.G. Christensen, P. Fränti and S.H. Jensen. *Joint single-channel speech separation and speaker identification.* In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE Int'l Conf. on, March 2010. (Cited on page 49.)

[Mowlaee *et al.* 2011a] P. Mowlaee, M.G. Christensen and S.H. Jensen. *New Results on Single-Channel Speech Separation Using Sinusoidal Modeling.* Audio, Speech, and Lang. Process., IEEE Trans., vol. 19, no. 5, pages 1265 –1277, july 2011. (Cited on pages 9 and 10.)

[Mowlaee *et al.* 2011b] Pejman Mowlaee, Rahim Saeidi, Zheng-Hua Tan, Mads Græsbøll Christensen, Tomi Kinnunen, Pasi Fränti and Søren Holdt Jensen. *Sinusoidal Approach for the Single-Channel Speech Separation and Recognition Challenge.* In INTERSPEECH, 2011. (Cited on page 49.)

[Mowlaee 2010] P. Mowlaee. New strategies for single-channel speech separation: Ph.d. thesis. Aalborg, Denmark, December 2010. (Cited on page 8.)

[Mukai *et al.* 2001] Ryo Mukai, Shoko Araki and Shoji Makino. *Separation And Dereverberation Performance Of Frequency Domain Blind Source Separation.* In Proc. Eurospeech 2001, pages 2599–2602, 2001. (Cited on page 1.)

[Nehorai & Porat 1986] A. Nehorai and B. Porat. *Adaptive comb filtering for harmonic signal enhancement.* Acoustics, Speech and Signal Processing, IEEE Trans. on, Oct. 1986. (Cited on pages 37 and 49.)

[Noam & Messer 2009] Y. Noam and H. Messer. *Notes on the Tightness of the Hybrid Cramér Rao Lower Bound.* IEEE Trans. Signal Processing, june 2009. (Cited on pages 77, 81, 82, 84 and 109.)

[Noll 1967] A. Michael Noll. *Cepstrum Pitch Determination.* The Journal of the Acoustical Society of America, vol. 41, no. 2, pages 293–309, 1967. (Cited on page 37.)

[O'Sullivan 1986] Finbarr O'Sullivan. *A Statistical Perspective on Ill-posed Inverse Problems.* Statistical Science, vol. 1, no. 4, pages 502–527, 1986. (Cited on page 72.)

[Ottersten *et al.* 1998] B Ottersten, P Stoica and R Roy. *Covariance Matching Estimation Techniques for Array Signal Processing Applications.* Digital Signal Processing, no. 3, 1998. (Cited on page 50.)

[Ozerov *et al.* 2007] A. Ozerov, P. Philippe, F. Bimbot and R. Gribonval. *Adaptation of Bayesian Models for Single-Channel Source Separation and Its Application to Voice/Music Separation in Popular Songs.* Audio, Speech, and Lang.Process., IEEE Trans., vol. 15, no. 5, pages 1564–1578, jul 2007. (Cited on pages 11, 16 and 17.)

[Parry & Essa 2007] R.M. Parry and I. Essa. *Incorporating Phase Information for Source Separation via Spectrogram Factorization.* In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 2, pages II–661–II–664, 2007. (Cited on page 15.)

[Pham *et al.* 1998] Dinh Tuan Pham, Jacques Verron and Marie Christine Roubaud. *A singular evolutive extended Kalman filter for data assimilation in oceanography.* Journal of Marine Systems, vol. 16, no. 3 - 4, pages 323 – 340, 1998. (Cited on page 89.)

[Piantanida *et al.* 2009] P. Piantanida, S. Sadough and P. Duhamel. *On the Outage Capacity of a Practical Decoder Accounting for Channel Estimation Inaccuracies.* IEEE Trans. Communications, May 2009. (Cited on page 104.)

[Quatieri & Danisewicz 1990] T.F. Quatieri and R.G. Danisewicz. *An approach to co-channel talker interference suppression using a sinusoidal model for speech.* Acoustics, Speech and Signal Process., IEEE Trans., vol. 38, no. 1, pages 56–69, 1990. (Cited on pages 9, 12 and 13.)

[Rabiner 1977] L. Rabiner. *On the use of autocorrelation analysis for pitch detection.* Acoustics, Speech and Signal Process., IEEE Trans., vol. 25, no. 1, pages 24–33, 1977. (Cited on page 37.)

[Radfar & Dansereau 2007] M.H. Radfar and R.M. Dansereau. *Single-Channel Speech Separation Using Soft Mask Filtering.* Audio, Speech, and Lang.Process., IEEE Trans., vol. 15, no. 8, pages 2299–2310, 2007. (Cited on pages 11 and 16.)

[Radfar *et al.* 2010] M.H. Radfar, R.M. Dansereau and W.-Y. Chan. *Monaural Speech Separation Based on Gain Adapted Minimum Mean Square Error Estimation.* Journal of Signal Processing Systems, vol. 61, no. 1, pages 21–37, 2010. (Cited on page 16.)

[Raj & Smaragdis 2005] B. Raj and P. Smaragdis. *Latent variable decomposition of spectrograms for single channel speaker separation.* In Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on, pages 17–20, 2005. (Cited on page 15.)

[Ramachandran & Kabal 1987] R.P. Ramachandran and P. Kabal. *Stability and performance analysis of pitch filters in speech coders.* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 35, no. 7, pages 937–946, Jul 1987. (Cited on page 24.)

[Ramachandran & Kabal 1989] R.P. Ramachandran and P. Kabal. *Pitch prediction filters in speech coding.* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 37, no. 4, pages 467–478, April 1989. (Cited on page 24.)

[Rao 1945] Calyampudi Radakrishna Rao. *Information and the accuracy attainable in the estimation of statistical parameters.* Bulletin of the Calcutta Mathematical Society, 1945. (Cited on page 71.)

[Rauch *et al.* 1965] H. E. Rauch, C. T. Striebel and F. Tung. *Maximum Likelihood Estimates of Linear Dynamic Systems.* Journal of the American Institute of Aeronautics and Astronautics, vol. 3, no. 8, pages 1445–1450, August 1965. (Cited on page 31.)

[Reddy & Raj 2004] Aarthi M. Reddy and Bhiksha Raj. *A minimum mean squared error estimator for single channel speaker separation.* In INTERSPEECH. ISCA, 2004. (Cited on page 16.)

[Reddy & Raj 2007] A.M. Reddy and B. Raj. *Soft Mask Methods for Single-Channel Speaker Separation.* Audio, Speech, and Lang. Process., IEEE Trans., vol. 15, no. 6, pages 1766–1776, 2007. (Cited on page 16.)

[Reif *et al.* 1999] K. Reif, S. Gunther, E. Yaz and R. Unbehauen. *Stochastic stability of the discrete-time extended Kalman filter.* Automatic Control, IEEE Transactions on, vol. 44, no. 4, pages 714–728, Apr 1999. (Cited on page 89.)

[Reuven & Messer 1997] I. Reuven and H. Messer. *A Barankin-type Lower Bound on the Estimation Error of a Hybrid Parameter Vector.* IEEE Trans. Information Theory, Mar. 1997. (Cited on page 81.)

[Reyes-gomez *et al.* 2004] Manuel J. Reyes-gomez, Daniel P. W. Ellis and Nebojsa Jojic. *Multiband Audio Modeling for Single-Channel Acoustic Source Separation.* In Proceedings of the International Conference on Audio, Speech and Signal processing, 2004. (Cited on page 16.)

[Rix *et al.* 2001] A.W. Rix, J.G. Beerends, M.P. Hollier and A.P. Hekstra. *Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs.* In Acoustics, Speech, and Signal Processing, 2001. IEEE International Conference on, volume 2, pages 749–752 vol.2, 2001. (Cited on page 18.)

[Rockah & Schultheiss 1987] Y. Rockah and P.M. Schultheiss. *Array shape calibration using sources in unknown locations–Part I: Far-field sources.* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 35, no. 3, pages 286–299, Mar .1987. (Cited on page 80.)

[Rodbro *et al.* 2003] C.A. Rodbro, M.G. Christensen, S.V. Andersen and S.H. Jensen. *Compressed domain packet loss concealment of sinusoidally coded speech.* In Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE Int. Conf. on, volume 1, pages I–104–7 vol.1, 2003. (Cited on page 37.)

[Ross *et al.* 1974] M. Ross, H. Shaffer, A. Cohen, R. Freudberg and H. Manley. *Average magnitude difference function pitch extractor.* Acoustics, Speech and Signal Process., IEEE Trans., vol. 22, no. 5, pages 353–362, 1974. (Cited on page 37.)

[Roweis 2000] Sam T. Roweis. *One Microphone Source Separation.* In NIPS, pages 793–799, 2000. (Cited on page 16.)

[Roweis 2003] Sam T. Roweis. *Factorial models and refiltering for speech separation and denoising.* In INTERSPEECH, 2003. (Cited on page 16.)

[Roy & Kailath 1989] R. Roy and T. Kailath. *ESPRIT-estimation of signal parameters via rotational invariance techniques.* Acoustics, Speech and Signal Process., IEEE Trans., vol. 37, no. 7, pages 984–995, 1989. (Cited on page 37.)

[Sadiki & Slock 2004] T. Sadiki and D.T.M. Slock. *"Bayesian Adaptive Filtering: Principles and Practical Approaches".* In Proc. 12th European Signal Processing Conf. (EUSIPCO), Vienna, Austria, Sept. 2004. (Cited on page 86.)

[Sambur & Jayant 1976] M. Sambur and N. Jayant. *LPC analysis/Synthesis from speech inputs containing quantizing noise or additive white noise.* Acoustics, Speech and Signal Process., IEEE Trans., vol. 24, no. 6, pages 488–494, 1976. (Cited on page 23.)

[sbø ll Christensen *et al.* 2008] Mads Græsbø ll Christensen, Petre Stoica, Andreas Jakobsson and Søren Holdt Jensen. *Multi-pitch estimation.* Signal Processing, vol. 88, no. 4, pages 972 – 983, 2008. (Cited on pages 9 and 37.)

[Schmidt & Olsson 2006a] Mikkel N. Schmidt and Rasmus K. Olsson. *Single-Channel Speech Separation using Sparse Non-Negative Matrix Factorization.* In Proceedings of Interspeech, pages 2614–2617, September 2006. (Cited on pages 15 and 16.)

[Schmidt & Olsson 2006b] Mikkel N. Schmidt and Rasmus K. Olsson. *Single-channel speech separation using sparse non-negative matrix factorization.* In in Proc. INTERSPEECH, 2006. (Cited on page 15.)

[Schmidt & Olsson 2007] M.N. Schmidt and Rasmus K. Olsson. *Linear Regression on Sparse Features for Single-Channel Speech Separation.* In Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on, pages 26–29, 2007. (Cited on page 16.)

[Schmidt *et al.* 2007] M.N. Schmidt, J. Larsen and Fu-Tien Hsiao. *Wind Noise Reduction using Non-Negative Sparse Coding.* In Machine Learning for Signal Process., 2007 IEEE Workshop on, pages 431–436, 2007. (Cited on page 8.)

[Schroeder & Atal 1985] Manfred R. Schroeder and B.S. Atal. *Code-excited linear prediction(CELP): High-quality speech at very low bit rates.* In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85., volume 10, pages 937–940, 1985. (Cited on page 10.)

[Schutz & Slock 2010] Antony Schutz and Dirk T M Slock. *Single-microphone blind audio source separation via Gaussian Short+Long Term AR Models.* In ISCCSP 2010, 4th Int'l Symposium on Communications, Control and Signal Processing, March 3-5, 2010, Limassol, Cyprus, 2010. (Cited on pages 50 and 61.)

[Schutz & Slock 2011] Antony Schutz and Dirk T M Slock. *Blind audio source separation using Short+Long Term AR source models and spectrum matching.* In DSP/SPE 2011, 14th IEEE Digital Signal Processing ; 6th Signal Processing Education Workshop, January 4-7, 2011, Sedona, Arizona, USA, 2011. (Cited on pages 50 and 63.)

[Shao *et al.* 2010] Yang Shao, Soundararajan Srinivasan, Zhaozhang Jin and DeLiang Wang. *A computational auditory scene analysis system for speech segregation and robust speech recognition.* Comput. Speech Lang., vol. 24, no. 1, pages 77–93, January 2010. (Cited on pages 8, 11 and 12.)

[Shumway & Stoffer 1982] R. H. Shumway and D. S. Stoffer. *An approach to time series smoothing and forecasting using the EM algorithm.* Journal of Time Series Analysis, vol. 3, no. 4, pages 253–264, 1982. (Cited on page 31.)

[Simon 2006] Dan Simon. Optimal state estimation: Kalman, h infinity, and nonlinear approaches. Wiley-Interscience, 2006. (Cited on page 90.)

[Smaragdis & Brown 2003] P. Smaragdis and J.C. Brown. *Non-negative Matrix Factorization for Polyphonic Music Transcription.* In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 19-22 2003. (Cited on page 15.)

[Smaragdis 2004] Paris Smaragdis. *Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs.* In CarlosG. Puntonet and Alberto Prieto, editeurs, Independent Component Analysis and Blind Signal Separation, volume 3195 of *Lecture Notes in Computer Science*, pages 494–499. Springer Berlin Heidelberg, 2004. (Cited on page 15.)

[Smith *et al.* 2005] D. Smith, J. Lukasiak and I. Burnett. *Blind speech separation using a joint model of speech production.* Signal Processing Letters, IEEE, vol. 12, no. 11, pages 784–787, Nov. 2005. (Cited on pages 10 and 21.)

[Stettiner *et al.* 1993] Y. Stettiner, D. Malah and D. Chazan. *Estimation of the parameters of a long-term model for accurate representation of voiced speech.* In Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE Int'l Conf. on, April 1993. (Cited on page 49.)

[Stoica & Arye 1989] Petre Stoica and Nehorai Arye. *MUSIC, maximum likelihood, and Cramer-Rao bound.* Acoustics, Speech and Sig. Process., IEEE Trans., vol. 37, no. 5, pages 720–741, 1989. (Cited on page 72.)

[Stoica & Moses 1997] P. Stoica and R.L. Moses. Introduction to Spectral Analysis. Prentice Hall, 1997. (Cited on page 72.)

[Stoica & Moses 2005] P. Stoica and R. Moses. Spectral analysis of signals. Prentice-Hall, New Jersey:, 1 édition, 2005. (Cited on page 62.)

[Stoica & Nehorai 1990] Petre Stoica and Arye Nehorai. *MUSIC, maximum likelihood, and Cramer-Rao bound: further results and comparisons.* Acoustics, Speech and Sig. Process., IEEE Trans., vol. 38, no. 12, pages 2140–2150, 1990. (Cited on page 72.)

[Tabrikian & Krolik 1999] J. Tabrikian and J.L. Krolik. *Barankin bounds for source localization in an uncertain ocean environment.* Signal Processing, IEEE Transactions on, vol. 47, no. 11, pages 2917–2927, 1999. (Cited on page 73.)

[Talkin 1995] D. Talkin. *A Robust Algorithm for Pitch Tracking (RAPT).* In W. B. Kleijn and K. K. Paliwal, editeurs, Speech Coding and Synthesis. Elsevier Science Inc., New York, NY, USA, 1995. (Cited on page 38.)

[Tanaka & Akahira 2003] Hidekazu Tanaka and Masafumi Akahira. *On a family of distributions attaining the Bhattacharyya bound.* Annals of the Institute of Statistical Mathematics, vol. 55, no. 2, pages 309–317, 2003. (Cited on page 73.)

[Tichavsky & Wong 2004] P. Tichavsky and K.T. Wong. *Quasi-fluid-mechanics-based quasi-Bayesian Cramér-Rao bounds for deformed towed-array direction finding.* Signal Processing, IEEE Transactions on, vol. 52, no. 1, pages 36–47, 2004. (Cited on page 85.)

[Tichavsky *et al.* 1998] P. Tichavsky, C.H. Muravchik and A. Nehorai. *Posterior Cramer-Rao bounds for discrete-time nonlinear filtering.* Signal Processing, IEEE Transactions on, vol. 46, no. 5, pages 1386 –1396, may 1998. (Cited on page 87.)

[Tordini & Piazza 2002] F. Tordini and F. Piazza. *A semi-blind approach to the separation of real world speech mixtures.* In Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on, volume 2, pages 1293–1298, 2002. (Cited on page 21.)

[Trees & Bell 2007] Harry L. Van Trees and Kristine L. Bell. Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking. Wiley-IEEE Press, 2007. (Cited on page 87.)

[Tzikas *et al.* 2008] D.G. Tzikas, A.C. Likas and N.P. Galatsanos. *The variational approximation for Bayesian inference.* Signal Processing Magazine, IEEE, vol. 25, no. 6, pages 131–146, November 2008. (Cited on page 95.)

[Van Trees & Bell 2007] H.L. Van Trees and K.L. Bell. Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking. Wiley-IEEE Press, 2007. (Cited on pages 69 and 85.)

[Van Trees 1968] H. L. Van Trees. Detection, Estimation, and Modulation Theory, Part I. John Willey and Sons Inc., 1968. (Cited on pages 72 and 79.)

[Van Trees 2001] H. L. Van Trees. Detection, Estimation, and Modulation Theory, Part I. Wiley-Interscience, 2001. (Cited on pages 71, 79 and 85.)

[Van Trees 2002] H. L. Van Trees. Optimum Array Processing (Detection, Estimation, and Modulation Theory, Part IV). Wiley-Interscience, 2002. (Cited on page 71.)

[Villares & Vazquez 2004] J. Villares and G. Vazquez. *The quadratic extended Kalman filter.* In Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2004, pages 480 – 484, july 2004. (Cited on page 86.)

[Vincent *et al.* 2003] Emmanuel Vincent, Cédric Févotte, Rémi Gribonval, Laurent Benaroya, Xavier Rodet, Axel Röbel, Eric Le Carpentier and Frédéric Bimbot. *Comment évaluer les algorithmes de séparation de sources audio ?* In 19e colloque GRETSI sur le traitement du signal et des images, pages 27–30, Paris, France, 2003. (Cited on page 17.)

[Vincent *et al.* 2006] E. Vincent, R. Gribonval and C. Fevotte. *Performance measurement in blind audio source separation.* Audio, Speech, and Lang. Process., IEEE Trans., vol. 14, no. 4, pages 1462–1469, 2006. (Cited on page 17.)

[Vincent *et al.* 2008] E. Vincent, N. Bertin and R. Badeau. *Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription.* In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pages 109–112, 2008. (Cited on page 15.)

[Virtanen & Klapuri 2000] T. Virtanen and A. Klapuri. *Separation of harmonic sound sources using sinusoidal modeling.* In Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE Int'l Conf. on, volume 2, 2000. (Cited on pages 10, 13 and 49.)

[Virtanen & Klapuri 2001] T. Virtanen and A. Klapuri. *Separation of harmonic sounds using multipitch analysis and iterative parameter estimation.* In Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pages 83–86, 2001. (Cited on page 10.)

[Virtanen 2006a] T. Virtanen. *Sound Source Separation in Monaural Music Signals.* PhD thesis, Tampere University of Technology, Finland, November 2006. (Cited on page 8.)

[Virtanen 2006b] Tuomas Virtanen. *Speech Recognition Using Factorial Hidden Markov Models for Separation in the Feature Space.* In Interspeech, in proc. ISCA, 2006. (Cited on page 8.)

[Virtanen 2007a] T. Virtanen. *Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria.* Audio, Speech, and Language Process., IEEE Trans. on, vol. 15, no. 3, pages 1066–1074, 2007. (Cited on page 15.)

[Virtanen 2007b] T. Virtanen. *Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria.* Audio, Speech, and Lang. Process., IEEE Trans., vol. 15, no. 3, pages 1066–1074, 2007. (Cited on page 16.)

[Wald 1949] Abraham Wald. *Note on the Consistency of the Maximum Likelihood Estimate.* The Annals of Mathematical Statistics, vol. 20, no. 4, pages 595–601, December 1949. (Cited on page 85.)

[Wang 2005] DeLiang Wang. *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis.* In Pierre Divenyi, editeur, Speech Separation by Humans and Machines, pages 181–197. Springer US, 2005. (Cited on page 12.)

[Weinstein & Weiss 1988] E. Weinstein and A.J. Weiss. *A general class of lower bounds in parameter estimation.* Inf. Theory, IEEE Trans., vol. 34, no. 2, pages 338–342, 1988. (Cited on pages 79 and 80.)

[Weinstein *et al.* 1994] E. Weinstein, A.V. Oppenheim, M. Feder and J.R. Buck. *Iterative and sequential algorithms for multisensor signal enhancement.* Signal Process., IEEE Trans., vol. 42, no. 4, pages 846–859, Apr 1994. (Cited on pages 10 and 92.)

[Weiss & Ellis 2010] Ron J. Weiss and Daniel P. W. Ellis. *Speech separation using speaker-adapted eigenvoice speech models.* Comput. Speech Lang., vol. 24, no. 1, pages 16–29, January 2010. (Cited on page 8.)

[Weiss & Weinstein 1985] A.J. Weiss and E. Weinstein. *A lower bound on the mean-square error in random parameter estimation (Corresp.)*. Information Theory, IEEE Transactions on, vol. 31, no. 5, pages 680–682, 1985. (Cited on page 79.)

[Wiklander 2003] J. Wiklander. Performance comparison of the Extended Kalman Filter and the Recursive Prediction Error Method. Master's thesis, Linköping Univ., 2003. (Cited on pages 89 and 90.)

[Yeredor 2000] A. Yeredor. *The Joint MAP-ML Criterion and its Relation to ML and to Extended Least-Squares*. IEEE Trans. Signal Processing, Dec. 2000. (Cited on pages 71, 84, 85 and 98.)