Statistical Signal Processing using a class of Iterative Estimation Algorithms

RLE Technical Report No. 532

September 1987

Meir Feder

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

This work has been supported in part by the Joint Services Electronics Program Contract No. DAAL03-86-K-0002 and in part by the National Science Foundation Grant No. ECS-83-05448.

Statistical Signal Processing using a class of Iterative Estimation Algorithms

Meir Feder

B.Sc., Tel-Aviv University (1980)

M.Sc., Tel-Aviv University (1984)

Submitted in Partial Fulfillment
of the Requirements for the

Degree of

Doctor of Science

at the

Massachusetts Institute of Technology

and the

Woods Hole Oceanographic Institution

September 1987

Signature of Author Meir Feder
Dept. of Electrical Engineering and Computer Science
August 31, 1987

Certified by Alan V. Oppenheim
Prof. Alan V. Oppenheim, Thesis Supervisor

Certified by
Prof. Ehud Weinstein, Thesis Supervisor

Accepted by
Arthur C. Smith
Chairman, EECS Department Committee on Graduate Students, MIT

Accepted by George V. Frisk
Dr. George V. Frisk
Chairman, MIT/WHOI Joint Committee for Oceanographic Engineering

Statistical Signal Processing using a class of Iterative Estimation Algorithms

by

Meir Feder

Submitted in partial fulfillment of the requirements for the degreé of Doctor of Science at the Massachusetts Institute of Technology.
August 31, 1987

Abstract

Many Signal Processing problems may be posed as statistical parameter estimation problems. A desired solution for the statistical problem is obtained by maximizing the Likelihood (ML), the A - Posteriori probability (MAP) or by optimizing other criterion, depending on the a - priori knowledge. However, in many practical situations, the original signal processing problem may generate a complicated optimization problem e.g when the observed signals are noisy and "incomplete".

A framework of iterative procedures for maximizing the likelihood, the EM algorithm, is widely used in statistics. In the EM algorithm, the observations are considered "incomplete" and the algorithm iterates between estimating the sufficient statistics of the "complete data" given the observations and a current estimate of the parameters (the E step) and maximizing the likelihood of the complete data, using the estimated sufficient statistics (the M step). When this algorithm is applied to signal processing problems it yield, in many cases, an intuitively appealing processing scheme.

In the first part of the thesis we investigate and extend the EM framework. By changing the "complete data" in each step of the algorithm we achieve algorithms with better convergence properties. We suggest EM type algorithms to optimize other (non ML) criteria. We also develop sequential and adaptive version of the EM algorithm.

In the second part of the thesis we discuss some applications of this extended framework of algorithms. We consider,

- Parameter estimation of composite signals, i.e signals that can be represented as a decomposition of simpler signals. This problem appear in e.g.
  - Multiple source location (or bearing) estimation
  - Multipath or multi-echo time delay estimation
- Noise canceling in multiple microphone environment, for a speech enhancement problem.

Thesis Supervisor: Alan V. Oppenheim
Title: Professor of Electrical Engineering, MIT.
Thesis Supervisor: Ehud Weinstein
Title: Adjunct Scientist, WHOI.

Acknowledgments

First and foremost I wish to thank Alan Oppenheim and Ehud Weinstein for their essential contributions to my research and to this thesis. Al, with his intuition and uncompromising standards, always pushed me to improve, to learn more and to test my ideas against real problems. He cared about my professional development; for all this, thanks, Al! Udi's close involvement with my research, his knowledge and insights and his collegial approach were invaluable and are greatly appreciated. During the years we have worked together, we have become close friends. Thank you, Udi, for the inspiration, encouragement and friendship.

I am also grateful to Professors Arthur Baggeroer and Bruce Musicus for serving as readers on my thesis committee. Bruce's comments as well as our numerous discussions greatly improved this thesis. I also want to thank Dr. Jules Jaffe for serving as the chairman of my thesis defense.

I was fortunate to be a member of the Digital Signal Processing Group at MIT. I thank all members of the group for many interesting technical discussions and for creating an exciting and enjoyable research environment. In particular, I wish to thank Mike Wengrovitz, who is also a good friend, and Avideh Zakhor, my office-mate. I thank Joe Bondaryk for carefully reading the thesis and also David Izraelevitz, Patrick Van Hove, Michele Covell, Dennis Martinez, Pat Peterson and George Hart for many useful discussions. This thesis could not be produced without Giovanni Alliberti, who kept the network alive and the computers running and was always available for advice.

A special acknowledgment goes to Woods Hole Oceanographic Institution and the MIT/WHOI joint program. I thank WHOI for the financial support, and for providing me excellent research facilities. I am indebted to the joint program for the great, memorable summers in Woods Hole, which proved that doctoral studies can indeed be incorporated with fun!

Finally, I would like to thank my wife, Yael. Her support, encouragement, patience and above all love, greatly contributed to this work and made it worthwhile. To her, and to our children, Liat and Eyal, this thesis is dedicated.

to Yael, Liat and Eyal

Contents

1 Introduction **8**

1.1 Introductory remarks 8

1.2 Preview and organization of the thesis 11

2 The EM method: Review and new developments **14**

2.1 Basic theory of the EM algorithm 15

2.2 Convergence results 22

2.2.1 Convergence of the likelihood sequence 24

2.2.2 Convergence of the parameter estimate sequence 26

2.2.3 Rate of convergence 27

2.3 The Linear Gaussian case 31

2.4 The EM algorithm with varying complete data 34

2.4.1 General theory 35

2.4.2 Motivation and rules for changing the complete data 37

2.5 The EM algorithm for general estimation criteria 39

2.5.1 Formal structure 40

2.5.2 The EM algorithm for the Minimum Information criterion 42

2.6 Possible signal processing applications 45

3 Sequential and Adaptive algorithms **52**

3.1 Sequential EM algorithms based on problem structure 53

3.1.1 Sequential EM with exact EM mapping 54

3.1.2 Sequential EM algorithm based on recursive E and M steps 59

3.2 Sequential EM based on stochastic approximation 62

3.2.1 The EM algorithm: General sequential considerations 62

3.2.2 Stochastic approximation 65

3.2.3 The EM stochastic approximation algorithm 67

3.3 Some properties of the sequential EM algorithms 69

3.3.1 General asymptotic consistency 70

3.3.2 Limit distribution 71

4 Parameter Estimation of Superimposed Signals **75**

4.1 Parameter estimation of superimposed signals: The deterministic case . . . 78

4.1.1 The ML problem 79

4.1.2 Solution using the EM method 80

4.1.3 The extended EM algorithm and the choice of the β_k's 86

4.2 Parameter estimation of superimposed signals: The stochastic Gaussian case 89
4.2.1 The ML problem 89
4.2.2 Solution using the EM method 91
4.3 Application to multipath time-delay estimation 95
4.3.1 The deterministic case 97
4.3.2 Simulation results 100
4.4 Application to multiple source Direction Of Arrival (DOA) estimation 104
4.4.1 The passive multiple source DOA estimation problem 105
4.4.2 The Gaussian case 107
4.4.3 Simulation results 109
4.5 Sequential and adaptive algorithms 112
4.5.1 Sequential algorithms based on exact EM mapping for the Deterministic case 115
4.5.2 Application to sum of sinusoids in noise 118

5 Maximum likelihood noise cancellation 122
5.1 Maximum likelihood formulation of the two-sensor noise cancellation problem 126
5.2 A simplified scenario 131
5.2.1 The ML problem 132
5.2.2 Solution via the EM algorithm 134
5.3 A more general scenario 138
5.3.1 The ML problem 139
5.3.2 Solution via the EM algorithm 140
5.4 Experimental results 148
5.4.1 The simplified scenario 148
5.4.2 The more general scenario 150

6 Information. Relative Entropy and the EM algorithm 153
6.1 Maximum Entropy and Minimum Relative Entropy 155
6.1.1 ME and MRE in comparison to other statistical criteria 155
6.1.2 The rationale of the MRE method 157
6.2 MRE by alternate minimization and the EM algorithm 159
6.2.1 MRE by alternate minimization 160
6.2.2 ML as a special case of the MRE criterion 161
6.2.3 The EM as an alternate minimization algorithm 163
6.2.4 Remark: how not to use the MRE criterion 164
6.3 Minimum Description Length interpretation of the MRE criterion 165
6.3.1 The Minimum Description Length idea 166
6.3.2 Minimum Relative Entropy as Minimum Description Length 168

7 Conclusions and further research 173
7.1 Further research 173
7.1.1 The sequential and adaptive algorithms 174
7.1.2 Other research directions 175
7.2 Signal reconstruction from partial information: Ideas and proposed research 178
7.2.1 General discussion 178
7.2.2 Statistical formulation of signal reconstruction problems 180

7.2.3 Solution using the EM method 183
7.3 Summary and discussion 186
7.3.1 The signal processing applications 186
7.3.2 The theoretical contributions 187

A Convergence theorems of the EM algorithm **190**
A.1 Convergence of the likelihood sequence 190
A.2 Convergence of the parameter estimate sequence 193
A.3 Rate of convergence 194

B Consistency of sequential EM algorithms **197**

Chapter 1

Introduction

1.1 Introductory remarks

Many signal processing problems may be posed as statistical estimation problems. A celebrated example is the work of Wiener, who formulated the fundamental problem of filtering a signal from an additive noise as a statistical problem, whose solution is known now as the "Wiener filter". Other common examples involve parameter estimation; e.g finding the localization and the velocity of targets in radar/sonar environments, or synchronization (i.e timing estimation) problems in communications systems. Many examples of the statistical analysis of signals processing problems may be found in [1], especially in its second and third parts.

In order to formulate the statistical problem, a model assumption is needed. A specific model may generate a simple statistical problem. However, it may not represent the original signal processing problem well. On the other hand, another model that tries to consider too many aspects of the original problem, may generate not only a difficult statistical problem, but also a non-robust, possibly ill-posed problem. The art of good modeling, which captures the important aspects of the real problem without complicating the resulting mathematical or statistical problem, is probably the most important factor in a successfully implemented statistical solution to the underlying real problem.

After the statistical problem is formulated, its desired solution often requires the optimization of some criterion, depending on the a-priori knowledge, and on the (possibly subjective) "risk" criterion. Frequently used criteria are Maximum Likelihood (ML) and Maximum A-Posteriori (MAP). Even with good modeling, these optimization problems may be complicated, e.g when the observed signals are noisy and incomplete. These optimization problems are rarely solved analytically. Instead, standard iterative search methods, e.g. gradient methods, Newton-Raphson method, are often used. The standard methods have some well known numerical problems. Furthermore, these methods may still be complicated since they require the calculation of the gradient and sometimes the Hessian matrix. These standard search methods rarely generate intuitive algorithms for the original real problem.

An interesting alternative to the straightforward gradient or Newton methods has been introduced in [2]. This technique, known as the Estimate-Maximize (EM) algorithm, suggests an iterative algorithm that exploits the properties of the stochastic system under consideration. The EM algorithm is actually a framework of iterative algorithms. To implement an EM algorithm, one has to consider the observations as incomplete with respect to more convenient choice of complete data. The algorithm then iterates between estimating the sufficient statistics of the complete data, given the observations and a current estimate of the parameters (the E step), and maximizing the likelihood of the complete data using the estimated sufficient statistics (the M step).

As will become evident in the course of this thesis, the EM method may yield intuitive processing schemes for the original signal processing problem, by innovatively choosing the complete data. Therefore, it is not surprising that some previously proposed algorithms for solving various signal processing problems can be interpreted in the EM algorithm context. One example is the iterative speech enhancement method suggested by Lim and Oppenheim [3]. We will return to this example later in the thesis. Other algorithms that have been suggested intuitively to solve specific signal processing problems, e.g. the iterative channel estimation algorithm of [4], the iterative reconstruction algorithm of [5], the iterative resolution technique of [6] and more, can also be interpreted as examples of the EM algorithm.

It is particularly important to note, at this point, the work of Musicus [7] and [8]. In this work, a general class of iterative algorithms has been suggested to minimize a special form of the Relative Entropy. In some special cases, the minimum relative entropy criterion reduces to the maximum likelihood criterion, and in those cases, the suggested iterative algorithms reduce to the EM algorithm. This work was an important inspiration for this thesis. We will discuss this approach later in the thesis in conjunction with our work on general information criteria and the EM algorithm.

Our exposure to a variety of estimation problems in oceanography, specifically in underwater acoustics, also had a major impact on this thesis. We have found that many of these problems were approached suboptimally, probably since the standard mathematical models of these problem usually generate statistical problems whose direct solution is complicated. We have suggested the EM iterative algorithm as a better approach to solve these statistical problems. Later in the thesis, we will describe how modeling considerations and EM algorithms have applied to array processing and time delay estimation problems in underwater acoustics, and have generated interesting solution procedures. The important experience with oceanographic signal processing problems established and confirmed our approach for solving statistical signal processing problems in general, and the other problems, presented later in the thesis, in particular.

In summary, this thesis presents a class of iterative and adaptive algorithms, based on the ideas that led to the EM algorithm, to optimize various statistical criteria. In addition, the thesis will address several signal processing problems and show that by using a reasonable model, an appropriate statistical criterion and an EM algorithm, an insightful solution procedure may be achieved and implemented successfully.

1.2 Preview and organization of the thesis

The application of the EM algorithm to a real world problem first requires modeling the problem statistically and then applying the EM algorithm to solve the resulting statistical problem. However, the EM algorithm is not uniquely defined; it depends on the choice of complete data. An unfortunate choice may yield a completely useless algorithm.

In this thesis we will consider the following signal processing problems:

- Parameter estimation of superimposed signals, i.e signals that can be represented as a sum of simpler signals. We will consider specifically the problems of multiple source location (or bearing) estimation, multipath or multi-echo time delay estimation and spectral estimation.

- Noise canceling in a multiple microphone environment. The real world application is a speech enhancement problem.

- Signal reconstruction from partial information. For this problem, we will present ideas and propose further research.

We will suggest statistical models for these signal processing problems and solve the resulting statistical problems by the EM method. In all these problems we will use a natural choice of complete data.

In the process of considering the applic tions mentioned above, we have modified and extended the scope of the EM method and derived explicit forms for some important special cases. Each of these results may be considered as a contribution to the EM algorithm at a theoretical level. We have also developed and analyzed sequential and adaptive algorithms based on the EM algorithm.

As a result of these contributions a general and flexible class of iterative and adaptive estimation algorithms is established. Beyond the theoretical contributions and the specific applications, we believe that this thesis suggests a way of thinking and a philosophy which may be used in a large variety of seemingly complex statist.,al inference and signal processing problems.

This thesis is organized as follows. Chapter 2 and 3 provide the theoretical background and contributions. In Chapter 2, we start with a review of the EM algorithm as developed in [2], and give its basic convergence properties following the considerations in [9]. We then derive the EM algorithm for the linear Gaussian case, whose importance will be evident later in the thesis. We also modify the basic EM algorithm and extend it, so that it may be applied to general estimation criteria.

Any iterative algorithm implies an adaptive or sequential estimation procedure, in which the new iteration takes into account new data points. A derivation of a class of sequential algorithms based on the EM structure is presented in Chapter 3. This class of algorithms may have the important tracking capabilities typical to an adaptive algorithms together with desirable asymptotic convergence results, achieved from the EM theory.

The signal processing applications of this class of iterative and adaptive algorithms are presented in Chapters 4 and 5. In Chapter 4 several problems that arise in radar/sonar signal and array processing are presented. Those problem involve multiple targets and multipath signals. A more general problem is the estimation of parameters of superimposed signals. We will describe an EM solution to the general superimposed signals problem, and apply it to multiple target bearing estimation and to multipath time delay estimation. Sequential algorithms to solve this problem will also be suggested.

The problem of multiple microphone noise cancellation is presented in Chapter 5. Using models of the speech and the noise, a statistical problem is formulated and then solved using the EM algorithm. This solution generates an intuitive processing scheme, that provides a novel solution to this well-investigated problem. An adaptive scheme based on the above algorithm, may be an alternative to Widrow's algorithm [10].

Chapter 6 is entitled "Information, Relative Entropy and the EM algorithm". It presents several interesting results that give an alternate interpretation to the EM algorithm and to information criteria mentioned in the EM algorithm context.

Chapter 7 will conclude and summarize the thesis. We will also suggest in this chapter topics for further research. As one of these topics, we will present specific ideas for solving problems of signal reconstruction from partial information. A statistical framework for these problems is developed and EM algorithms for solving this statistical problem, by optimizing the likelihood, the a-posteriori probability or other appropriate criteria are derived.

Chapter 2

The EM method: Review and new developments

In this Chapter we review the Estimate-Maximize (EM) algorithm for solving maximum likelihood (ML) and maximum-a-posteriori (MAP) estimation problems, and present new developments that extend the scope of the algorithm, and make it more accessible for solving signal processing problems.

The chapter is organized as follows. In section 2.1, the basic EM algorithm is presented, following the considerations in [2]. In section 2.2, we analyze and discuss the convergence properties of the EM algorithm. The results presented here clarify and simplify the convergence analysis presented in [2] and [9].

In section 2.3, the EM algorithm is explicitly derived for the special but important case where the observed (incomplete) data and complete data are jointly Gaussian. related by a linear non-invertible transformation. In perspective, the linear-Gaussian case was an important step towards the application of the EM algorithm to signal processing problems.

In sections 2.4 and 2.5, we present new ideas and results that extend the scope of the EM method. The results in these sections generate a more general, yet more flexible, class of iterative algorithms.

Section 2.6 concludes this chapter, by discussing the possible signal processing applications of the EM framework.

2.1 Basic theory of the EM algorithm

Let $\underline{Y}$ denote a data vector with the associated probability density $f_{\underline{Y}}(\underline{y};\underline{\theta})$, indexed by the parameter vector $\underline{\theta} \in \Theta$, where Θ is a subset of the k-dimensional Euclidean space. Given an observed $\underline{y}$, the maximum likelihood (ML) estimate, $\hat{\underline{\theta}}_{ML}$, is the value of $\underline{\theta}$ that maximizes the log-likelihood, that is,

$$\hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}\in\Theta} \log f_{\underline{Y}}(\underline{y};\underline{\theta}) \tag{2.1}$$

Finding the ML estimator is often desirable since it is, in most cases, asymptotically consistent and efficient. However, in many cases, the maximization problem of (2.1) is complicated.

Suppose that the data vector $\underline{Y}$ can be viewed as being incomplete, and we can specify some data $\underline{X}$ related to $\underline{Y}$ by

$$T(\underline{X}) = \underline{Y} \tag{2.2}$$

where $T(\cdot)$ is a non-invertible (many to one) transformation. If an observation $\underline{x}$ of $\underline{X}$ is given, an observation $\underline{y}$ of $\underline{Y}$ is available too, but not vice versa. $\underline{X}$ will be referred to as the complete data. The probability density of the complete data, denoted $f_{\underline{X}}(\underline{x};\underline{\theta})$, is also indexed by the parameter vector $\underline{\theta}$. Assume that H is specified so that if $\underline{x}$ is available, finding the maximum likelihood estimate of $\underline{\theta}$ is easy, i.e. solving

$$\hat{\underline{\theta}} = \arg\max_{\underline{\theta}\in\Theta} \log f_{\underline{X}}(\underline{x};\underline{\theta}) \tag{2.3}$$

is straightforward. The EM algorithm, presented below, will use the simple procedure for ML estimation in the complete data model, as a part of an iterative algorithm for ML estimation in the observations' model.

Given a sample of the incomplete data $\underline{y}$, the complete data $\underline{x}$ must be a member of the set $\mathcal{X}(\underline{y})$ where,

$$\mathcal{X}(\underline{y}) = \left\{\underline{x} | T(\underline{x}) = \underline{y}\right\} \tag{2.4}$$

Since $\underline{Y}$ is a many to one function of the complete data $\underline{X}$, the probability density functions of the complete and incomplete data satisfy,

$$f_{\underline{Y}}(\underline{y};\underline{\theta}) = \int_{\mathcal{X}(\underline{y})} f_{\underline{X}}(\underline{x};\underline{\theta})d\underline{x} \tag{2.5}$$

The conditional density of $\underline{X}$, given $\underline{Y} = \underline{y}$, is defined over the set $\mathcal{X}(\underline{y})$. This probability density function is given by,

$$f_{\underline{X}/\underline{Y}=\underline{y}}(\underline{x};\underline{\theta}) = \frac{f_{\underline{X}}(\underline{x};\underline{\theta})}{\int_{\mathcal{X}(\underline{y})} f_{\underline{X}}(\underline{x};\underline{\theta})d\underline{x}} = \frac{f_{\underline{X}}(\underline{x};\underline{\theta})}{f_{\underline{Y}}\left(\underline{y};\underline{\theta}\right)}, \quad \forall\underline{x}\in\mathcal{X}(\underline{y}) \tag{2.6}$$

Taking the logarithm on both sides of (2.6) and rearranging, we obtain

$$\log f_{\underline{Y}}\left(\underline{y};\underline{\theta}\right) = \log f_{\underline{X}}(\underline{x};\underline{\theta}) - \log f_{\underline{X}/\underline{Y}=\underline{y}}(\underline{x};\underline{\theta}), \quad \forall\underline{x}\in\mathcal{X}(\underline{y}) \tag{2.7}$$

We can now take the conditional expectation over $\underline{X}$, of both sides of (2.7), given $\underline{Y} = \underline{y}$ and an arbitrary parameter value $\underline{\theta}'$. The left hand side remains unchanged, and we get,

$$\log f_{\underline{Y}}(\underline{y};\underline{\theta}) = E\left\{\log f_{\underline{X}}(\underline{x};\underline{\theta}) \,|\, \underline{Y} = \underline{y};\underline{\theta}'\right\} - E\left\{\log f_{\underline{X}/\underline{Y}}(\underline{x}/\underline{y};\underline{\theta}) \,|\, \underline{Y} = \underline{y};\underline{\theta}'\right\} \tag{2.8}$$

Define, for convenience,

$$L(\underline{\theta}) = \log f_{\underline{Y}}(\underline{y};\underline{\theta}) \tag{2.9}$$

$$Q(\underline{\theta},\underline{\theta}') = E\left\{\log f_{\underline{X}}(\underline{x};\underline{\theta}) \,\middle|\, \underline{Y}=\underline{y};\underline{\theta}'\right\} = \int_{\mathcal{X}(\underline{y})} \log f_{\underline{X}}(\underline{x};\underline{\theta})\cdot f_{\underline{X}/\underline{Y}=\underline{y}}(\underline{x};\underline{\theta}')d\underline{x} \tag{2.10}$$

$$H(\underline{\theta},\underline{\theta}') = E\left\{\log f_{\underline{X}/\underline{Y}=\underline{y}}(\underline{x};\underline{\theta}) \,\middle|\, \underline{Y}=\underline{y};\underline{\theta}'\right\} = \int_{\mathcal{X}(\underline{y})} \log f_{\underline{X}/\underline{Y}=\underline{y}}(\underline{x};\underline{\theta})\cdot f_{\underline{X}/\underline{Y}=\underline{y}}(\underline{x};\underline{\theta}')d\underline{x} \tag{2.11}$$

With these definitions, equation (2.8) reads

$$L(\underline{\theta}) = Q(\underline{\theta},\underline{\theta}') - H(\underline{\theta},\underline{\theta}') \tag{2.12}$$

Equations (2.7) and (2.8) are interesting identities for $L(\underline{\theta})$, the log-likelihood of the observations. Equation (2.7) is true for any $\underline{x} \in \mathcal{X}(\underline{y})$. Equation (2.8), or equivalently equation (2.12) is true for any pair $\underline{\theta},\underline{\theta}' \in \Theta$.

Consider now Jensen's inequality (see e.g. equations 1e.5.6 and 1e.6.6 in [11]) which states that for any two p.d.f.'s f and g defined over the same sample space,

$$E_f\{\log g\} \leq E_f\{\log f\} \tag{2.13}$$

where equality holds if and only if $f = g$ almost everywhere. E_f denotes the expectation, using the p.d.f. f. Let $f = f_{\underline{X}/\underline{Y}}(\underline{x};\underline{\theta}')$ and $g = f_{\underline{X}/\underline{Y}}(\underline{x};\underline{\theta})$, both defined over the sample space $\mathcal{X}(\underline{y})$. Substituting in (2.13), and using the definition of $H(\cdot,\cdot)$, we get,

$$H(\underline{\theta},\underline{\theta}') \leq H(\underline{\theta}',\underline{\theta}') \tag{2.14}$$

Suppose we can find $\underline{\theta}$ such that,

$$Q(\underline{\theta};\underline{\theta}') \geq Q(\underline{\theta}';\underline{\theta}') \tag{2.15}$$

In this case, using (2.14) and (2.12), we conclude that,

$$L(\underline{\theta}) \geq L(\underline{\theta}') \tag{2.16}$$

A procedure for iteratively increasing the likelihood may be suggested based on (2.15) and (2.16) as follows. Given a value of the parameter $\underline{\theta}^{(n)}$, we will find a new value $\underline{\theta}^{(n+1)}$ that satisfies (2.15), and thus increases the likelihood, by maximizing $Q(\underline{\theta};\underline{\theta}^{(n)})$. This procedure is the EM algorithm, which we now formally present.

- Start, $n = 0$: guess $\underline{\theta}^{(0)}$
- Iterate (until some convergence criterion is achieved)
  - **The E step:** calculate

$$Q(\underline{\theta};\underline{\theta}^{(n)}) = E\left\{\log f_{\underline{X}}(\underline{x};\underline{\theta}) \mid \underline{Y} = \underline{y};\underline{\theta}^{(n)}\right\} \tag{2.17}$$

  - **The M step:** solve

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}\in\Theta} Q(\underline{\theta};\underline{\theta}^{(n)}) \tag{2.18}$$

  - $n = n + 1$

The "E" stands for the conditional Expectation, or Estimation performed in the E step, and the "M" stands for the Maximization performed in the M step.

An EM iteration may be summarized by the updating equation,

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}\in\Theta} E\left\{\log f_{\underline{X}}(\underline{x};\underline{\theta}) \mid \underline{y};\underline{\theta}^{(n)}\right\} \tag{2.19}$$

This iteration is justified intuitively as follows. We would like to choose $\underline{\theta}$ that maximizes $\log f_{\underline{X}}(\underline{x};\underline{\theta})$, the log-likelihood of the complete data. However, since $\log f_{\underline{X}}(\underline{x};\underline{\theta})$ is not available to us (because the complete data is not available), we maximize instead its expectation, given the observed data $\underline{y}$, and the current value of the parameters $\underline{\theta}^{(n)}$.

An iterative procedure that increases the likelihood is also achieved, if instead of maximizing $Q(\underline{\theta};\underline{\theta}^{(n)})$, we just increase it. Thus, we may replace the M step by the following step:

$$\underline{\theta}^{(n+1)} = M(\underline{\theta}^{(n)}) \tag{2.20}$$

where $M(\underline{\theta})$ is any mapping that satisfies

$$Q\left(M(\underline{\theta});\underline{\theta}\right) \geq Q(\underline{\theta};\underline{\theta}) \quad \forall\underline{\theta} \in \Theta \tag{2.21}$$

This variation of the algorithm was named the Generalized EM algorithm (GEM) by Dempster et. al [2]. A special case, of course, is the EM algorithm.

The motivation of the GEM algorithm, which also applies to the EM algorithm, is summarized in the following theorem, (theorem 1 of [2]). This theorem carefully states the basic monotonicity property of the GEM algorithm. The proof of this theorem will follow immediately from the considerations above.

Theorem 2.1 *For every GEM algorithm,*

$$L(\underline{\theta}^{(n+1)}) \geq L(\underline{\theta}^{(n)}) \tag{2.22}$$

where equality holds if and only if both

$$Q(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) = Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) \tag{2.23}$$

and

$$f_{\underline{X}/\underline{Y}}(\underline{x};\underline{\theta}^{(n+1)}) = f_{\underline{X}/\underline{Y}}(\underline{x}/\underline{y};\underline{\theta}^{(n)}) \quad \textit{a.e in } \mathcal{X}(\underline{y}) \tag{2.24}$$

Proof: By the definition of the GEM algorithm

$$Q(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) \geq Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)})$$

Thus, since $H(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) \leq H(\underline{\theta}^{(n)};\underline{\theta}^{(n)})$,

$$L(\underline{\theta}^{(n+1)}) \geq L(\underline{\theta}^{(n)})$$

Now by (2.12), equality holds if and only if

$$Q(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) = Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) \text{ and } H(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) = H(\underline{\theta}^{(n)};\underline{\theta}^{(n)})$$

The latter holds if and only if

$$f_{\underline{X}/\underline{Y}}(\underline{x};\underline{\theta}^{(n+1)}) = f_{\underline{X}/\underline{Y}}(\underline{x}/;\underline{\theta}^{(n)}), \quad \text{a.e in } \mathcal{X}(\underline{y})$$

□

This theorem leads to the following corollaries:

Corollary 2.1 *If for some $\underline{\theta}^* \in \Theta$, $L(\underline{\theta}^*) > L(\underline{\theta})$, $\forall \underline{\theta} \in \Theta$, then, for every GEM algorithm*

$$M(\underline{\theta}^*) = \underline{\theta}^*$$

Corollary 2.2 *Suppose for some $\underline{\theta}^* \in \Theta$, $L(\underline{\theta}^*) \geq L(\underline{\theta})$, $\forall \underline{\theta} \in \Theta$. Then, for every GEM algorithm*

$$L(M(\underline{\theta}^*)) = L(\underline{\theta}^*)$$
$$Q(M(\underline{\theta}^*);\underline{\theta}^*) = Q(\underline{\theta}^*;\underline{\theta}^*)$$
$$f_{\underline{X}/\underline{Y}}(\underline{x};M(\underline{\theta}^*)) = f_{\underline{X}/\underline{Y}}(\underline{x}/;\underline{\theta}^*) \quad \textit{a.e in } \mathcal{X}(\underline{y})$$

In other words, if a unique global maximum of the likelihood exists, it is a fixed point of any GEM algorithm. If we have a set of global maxima, the GEM algorithm may move inside this set. However, each new value must satisfy the conditions of corollary 2.2.

We note that the EM algorithm is actually a class of algorithms. There are many complete data specifications $\underline{X}$, that will generate the observed data $\underline{Y}$. The choice of complete data may critically affect the complexity and the convergence properties of the algorithm. An unfortunate choice of complete data will yield a completely useless algorithm. Thus, it takes creativity to apply the EM algorithm to a given problem. This will be demonstrated later in the thesis when we solve specific signal processing problems.

To complete the basic theory of the EM algorithm, we will present in section 2.2 the convergence properties of the algorithm. This presentation, which clarifies and simplifies previous results, may be used as a future reference on these topics.

The EM algorithm for exponential families

Examining the expressions for the EM algorithm, (2.17) and (2.18), we note that, in general, the EM algorithm may be complicated. The calculation of $Q(\underline{\theta}, \underline{\theta}^{(n)})$ in the E step may require multiple integration, and the maximization in the M step is, in general, a non-linear optimization problem. However, in the case of exponential families of distributions, which is now described, the E step has an explicit simple form and the maximization performed in the M step is as complicated as solving a maximum likelihood problem for the complete data, which is assumed to be easy.

Suppose that the p.d.f. of the complete data, $\underline{x}$, belongs to the exponential family of probabilities, i.e.

$$f_{\underline{X}}(\underline{x};\underline{\theta}) = \frac{b(\underline{x})}{a(\underline{\theta})} \exp\{\sum_i \underline{\phi}_i(\underline{\theta})^T \underline{t}_i(\underline{x})\} \tag{2.25}$$

The set of statistics $\{\underline{t}_i(\underline{x})\}$ is the sufficient statistics. This set is denoted $\mathcal{T}(\underline{x})$. Note that the exponential family of distributions includes almost all common p.d.f.'s e.g. Gaussian, binomial, exponential etc.

The log-likelihood of the complete data for exponential families has the form

$$\log f_{\underline{X}}(\underline{x};\underline{\theta}) = -\log a(\underline{\theta}) + \sum_i \underline{\phi}_i(\underline{\theta})^T \underline{t}_i(\underline{x}) + \underbrace{\log b(\underline{x})}_{\text{independent of } \underline{\theta}} \tag{2.26}$$

Due to this special form of the log-likelihood, we need only to calculate, in the E step, the conditional expectation of the sufficient statistics. We then substitute the estimated sufficient statistics in the likelihood of the complete data, and maximize the resulting expression in the M step. The E and M of the EM algorithm steps, for exponential families, reduce thus to

- The E step: Calculate

$$\mathcal{T}^{(n)}(\underline{x}) = E\{\mathcal{T}(\underline{x})/\underline{y};\underline{\theta}^{(n)}\}$$

or, $\forall i$, calculate

$$\underline{t}_i^{(n)} = E\{\underline{t}_i(\underline{x})/\underline{y};\underline{\theta}^{(n)}\} \tag{2.27}$$

- The M step: solve

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} \left\{ -\log a(\underline{\theta}) + \sum_i \underline{\phi}_i(\underline{\theta})^T \underline{t}_i^{(n)}(\underline{x}) \right\} \tag{2.28}$$

The sufficient statistics are usually simple functions of the data, $\underline{x}$, and therefore explicit formulas usually exist for the E step above. The expression to be maximized in the M step has the functional form (w.r.t $\underline{\theta}$) of the log-likelihood of the complete data. Since maximizing the likelihood of the complete data is assumed to be easy, the implementation of the M step above is easy too.

Gaussian distribution belongs to the exponential family of distributions. In section 2.3 we will derive a closed form analytical expression for $Q(\underline{\theta}, \underline{\theta}')$, i.e. the E step, for the case where $\underline{X}$ and $\underline{Y}$ are jointly Gaussian related by linear transformation. The maximization problem in the M step, for this linear-Gaussian case, will be as complicated as solving a maximum likelihood problem in the complete data model.

2.2 Convergence results

The EM (or the GEM) algorithm generates a sequence of parameters, $\{\underline{\theta}^{(n)}\}$, and an associated sequence of log-likelihoods, $\{L^{(n)}\}$, where $L^{(n)} = L(\underline{\theta}^{(n)})$. We have shown that each iteration increases the likelihood, i.e. the likelihood sequence is a monotonic nondecreasing sequence ($L^{(n+1)} \geq L^{(n)}$). However, the EM theory should also answer the following important questions:

- Do the likelihood and the parameter sequences converge?
- To where will they converge?
- How fast will they converge?

These convergence issues will be addressed as follows. The convergence of the likelihood sequence will be considered first. The issue of its convergence to a global maximum, local maximum or a stationary point will be discussed. Then, the convergence of the parameter estimate sequence will be considered, noting that even if the likelihood sequence converges (say to L^*), the associated parameter sequence may not converge, i.e. it may have a set of limit points, each of which corresponds to this likelihood value L^*. For the cases in which the sequences do converge, the rate of convergence in the neighborhood of the convergence point will be calculated.

Our discussion in this section follows the considerations in Wu, [9], and the original paper of Dempster et. al. [2]. Another important reference is [12]. The rate of convergence and the computation of the Fisher information matrix associated with the EM parameter estimate sequence are also discussed in [13], [14] and elsewhere.

The following notation and assumptions are used in this section. Let Θ be the set of possible parameter values, which is assumed to be a subset of the k-dimensional Euclidean space. Θ_0 is the set

$$\Theta_0 = \{\underline{\theta} \in \Theta | L(\underline{\theta}) \geq L(\underline{\theta}^{(0)})\}$$

and it is assumed to be compact for any $L(\underline{\theta}^{(0)}) > -\infty$.

$\mathcal{M}$ will denote the set of local maxima of $L(\cdot)$ while $\mathcal{S}$ will denote the set of stationary points of $L(\cdot)$, in the interior of Θ.

An EM (GEM) iteration may be denoted by

$$\underline{\theta}^{(n)} \longrightarrow \underline{\theta}^{(n+1)} \in M(\underline{\theta}^{(n)})$$

where $M(\cdot)$ is a point to set mapping such that $M(\underline{\theta}^{(n)})$ is the set of maximizers of $Q(\underline{\theta};\underline{\theta}^{(n)})$ over $\underline{\theta} \in \Theta$ for an EM algorithm, and such that

$$Q(\underline{\theta};\underline{\theta}^{(n)}) \geq Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)}), \quad \forall \underline{\theta} \in M(\underline{\theta}^{(n)})$$

for a GEM algorithm.

2.2.1 Convergence of the likelihood sequence

As shown in theorem 2.1, the likelihood sequence, $\{L^{(n)}\}$, is a monotonic nondecreasing sequence. Thus, if this sequence is also bounded, it converges to some value L^*. Only in rare and singular cases can we find a non bounded likelihood sequence. Furthermore, if the likelihood function $L(\cdot)$ is continuous in Θ, the compactness of Θ_0 guarantees that the likelihood sequence, $\{L^{(n)}\}$, is bounded for any starting point $\underline{\theta}^{(0)} \in \Theta$. Thus, the likelihood sequence can be expected to converge in most cases to some L^*.

We want to know whether L^* is a global maximum, a local maximum or at least a stationary value of $L(\underline{\theta})$ over Θ. Unfortunately, as for any general "hill climbing" optimization algorithm, there is no guarantee that the EM algorithm will converge to a global or even a local maximum. It has been reported, in [15] and [16] that, if the log-likelihood, L, has several maxima and stationary points, then convergence of the EM algorithm to either type of point depends on the choice of the starting point. Note that this phenomenon occurs despite the fact that we may perform a global maximization (of Q) in the M step.

In Appendix A we consider the convergence issue more precisely, where, as in many numerical analysis algorithms, the convergence analysis is based on the Global Convergence Theorem which may be found in [17] page 91, and [18] page 187. This theorem provides sufficient conditions that guarantee the convergence of a general iterative procedure

$$\underline{\theta}^{(n+1)} = M(\underline{\theta}^{(n)})$$

to a solution set.

For the EM algorithm, where $M(\underline{\theta}^{(n)})$ is the set of maximizers of $Q(\underline{\theta}; \underline{\theta}^{(n)})$, it is shown in Appendix A, that the simple condition,

$$Q(\underline{\theta}_1; \underline{\theta}_2) \quad \text{is continuous in both } \underline{\theta}_1 \text{ and } \underline{\theta}_2 \tag{2.29}$$

in addition to the compactness of Θ_0, guarantees the convergence to the solution set $\mathcal{S}$, i.e. this condition implies that the likelihood sequence of the EM algorithm converges to a stationary value.

A stronger sufficient condition is needed to guarantee convergence to a local maxima. Again, in Appendix A, it is shown that, if in addition to the continuity condition (2.29) Q satisfies

$$\sup_{\underline{\theta}' \in \Theta} Q(\underline{\theta}'; \underline{\theta}) > Q(\underline{\theta}; \underline{\theta}) \quad \forall \underline{\theta} \in (\mathcal{S} - \mathcal{M}) \tag{2.30}$$

where $(\mathcal{S} - \mathcal{M})$ is the difference set $\{\underline{\theta} \in \mathcal{S} | \underline{\theta} \notin \mathcal{M}\}$, then the likelihood sequence converges to a local maxima, i.e. to the solution set $\mathcal{M}$.

Since condition (2.30) is hard to verify, we may have to be satisfied with a proof of convergence to a stationary point, even when the EM algorithm does converge to a local maximum. Condition (2.30) is not met in general, and the EM algorithm converges to a stationary value, local maximum or global maximum depending on the choice of starting point.

2.2.2 Convergence of the parameter estimate sequence

The convergence of the likelihood sequence does not imply the convergence of the parameter estimate sequence. Suppose that the likelihood sequence converges to L^* and that the conditions, that guarantee the convergence to a stationary point, are satisfied. Define

$$S(L^*) = \{\underline{\theta} \in S | L(\underline{\theta}) = L^*\}$$

The sequence of estimates, $\{\underline{\theta}^{(n)}\}$, may not converge, i.e. it may have a (possibly infinite) set of limit points. We may only say that all limit points of $\{\underline{\theta}^{(n)}\}$ are in $S(L^*)$.

The convergence of the parameter sequence may be guaranteed (trivially), if the solution set, i.e. $S(L^*)$ in the example above, has a single point. An important special case, in which the solution set is a singleton, is when the likelihood function is unimodal in Θ.

The requirement that the solution space has a single point may be relaxed, and it may be shown, see Appendix A, that if the solution set is discrete and

$$\lim_{n \to \infty} ||\underline{\theta}^{(n+1)} - \underline{\theta}^{(n)}|| = 0 \tag{2.31}$$

the parameter estimates sequence will converge.

Condition (2.31) may be easily verified in many applications. For the EM estimate sequence, since $L^{(n)} \to L^*$, and since

$$L^{(n+1)} - L^{(n)} \geq Q(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) - Q(\underline{\theta}^{(n)}; \underline{\theta}^{(n)}) \tag{2.32}$$

a sufficient condition for (2.31) is that there exist a forcing function $\sigma(\cdot)$, such that

$$Q(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) - Q(\underline{\theta}^{(n)}; \underline{\theta}^{(n)}) \geq \sigma(||\underline{\theta}^{(n+1)} - \underline{\theta}^{(n)}||), \quad \forall n \tag{2.33}$$

where a forcing function is a function such that for any sequence, $\{x_n\}$,

$$\lim_{n \to \infty} \sigma(x_n) = 0 \Longrightarrow \lim_{n \to \infty} x_n = 0 \tag{2.34}$$

Taking $\sigma(x) = \lambda x^2$, $\lambda > 0$ as the forcing function, we get the sufficient condition

$$Q(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) - Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) \geq \lambda||\underline{\theta}^{(n+1)} - \underline{\theta}^{(n)}||^2, \quad \forall n \tag{2.35}$$

which may be verified easily in several applications.

One may argue, that the convergence of the parameter sequence is not as important as the convergence of the likelihood sequence to the desired location on the log-likelihood surface. However, one should be aware of the possibility of a non-convergent estimate sequence, e.g. if $L(\underline{\theta})$ has a ridge of stationary points in which $L(\underline{\theta}) = L^*$, then the set $S(L)$ is not discrete and the EM algorithm may move indefinitely on that ridge.

2.2.3 Rate of convergence

When the EM (or GEM) algorithm converges, an interesting and important problem is the determination of its rate of convergence. In this section, after defining the rate of convergence and other terms, that are commonly used in association with it, we will calculate the rate of convergence of the EM algorithm.

Let us denote the differentiation operator D. A differentiation operator with respect to two variable will be denoted D^{ij} as

$$D^{ij} f(a,b) = \left.\frac{\partial^{i+j} f(\theta_1, \theta_2)}{\partial\theta_1^i \partial\theta_2^j}\right|_{\theta_1 = a, \theta_2 = b}$$

The following identities will be needed later when the rate of convergence of the EM algorithm is explicitly developed,

$$DL(\underline{\theta}) = D^{10}Q(\underline{\theta};\underline{\theta}) \tag{2.36}$$

$$D^2L(\underline{\theta}) = D^{20}Q(\underline{\theta};\underline{\theta}) - D^{20}H(\underline{\theta};\underline{\theta}) = D^{20}Q(\underline{\theta};\underline{\theta}) + D^{11}H(\underline{\theta};\underline{\theta}) \tag{2.37}$$

$$D^{11}Q(\underline{\theta};\underline{\theta}) = D^{11}H(\underline{\theta};\underline{\theta}) \tag{2.38}$$

These identities have already been recognized by Fisher [19], and are redeveloped in [2],[13] and in Appendix A.

Definitions and background

Consider a sequence, $\{x_n\}$, converging to a limit, x^*. Each element, x_n, belongs to X, which is a subset of some norm-space (say the k-dimensional Euclidean space). We define the *order of convergence* as follows:

Definition 2.1 *The order of convergence of a sequence, $\{x_n\}$, that converges to x^*, denoted p, is the supremum of the nonnegative numbers, p', for which the following ratio is finite, i.e. for which*

$$\alpha = \overline{\lim_{n \to \infty}} \frac{||x_{n+1} - x^*||}{||x_n - x^*||^{p'}} < \infty$$

Loosely speaking, the order of convergence describes the asymptotic behavior of the error sequence, $\{e_n\}$, where $e_n = x_n - x^*$, i.e. as $n \to \infty$ we have

$$||e_{n+1}|| = \alpha ||e_n||^p$$

So, the larger p is, the faster the sequence, $\{x_n\}$, converges.

Most iterative algorithm generate sequences, whose order of convergence is unity. In this case, the important number is the convergence rate defined by,

Definition 2.2 *The convergence rate of a sequence, $\{x_n\}$, that converges to x^*, denoted α, is*

$$\alpha = \lim_{n \to \infty} \frac{||x_{n+1} - x^*||}{||x_n - x^*||}$$

where $0 \leq \alpha \leq 1$. The sequence is said to converge linearly if $0 < \alpha < 1$, and superlinearly if $\alpha = 0$.

The convergence rate of any sequence, whose order is greater then unity, will be zero. These sequences have superlinear convergence. We note, however, that a sequence with unity order of convergence may also have a superlinear convergence. Linear convergence is sometimes referred to as *geometric* convergence or *exponential* convergence, since in this case the error sequence, $\{e_n\}$, is a geometric sequence.

In many iterative algorithms, the iteration is defined via a mapping, that successively approximates the solution, i.e.

$$x_{n+1} = M(x_n)$$

In this case, we may find the rate of convergence by investigating the Jacobian matrix (or the matrix of derivatives) of this mapping, defined by,

$$[DM(x^*)]_{ij} = \frac{\partial [M(x)]_i}{\partial [x]_j}\bigg|_{x=x^*} = \lim_{n\to\infty} \frac{[M(x_n) - M(x^*)]_i}{[x_n - x^*]_j} \tag{2.39}$$

where $[\cdot]_i$ denotes the i^{th} component of a vector.

Since

$$\alpha = \lim_{n\to\infty} \frac{||x_{n+1} - x^*||}{||x_n - x^*||} = \lim_{n\to\infty} \frac{||M(x_n) - M(x^*)||}{||x_n - x^*||}, \tag{2.40}$$

the largest eigenvalue of the matrix $DM(x^*)$, will provide us with the convergence rate of the iterative algorithm.

Rate of convergence of the EM algorithm

The rate of convergence of the EM algorithm can be calculated by deriving the Jacobian matrix of the mapping, $\underline{\theta}^{(n+1)} = M(\underline{\theta}^{(n)})$, associated with the EM algorithm. We recall that this mapping is defined by

$$\underline{\theta}^{(n+1)} = M(\underline{\theta}^{(n)}) = \arg\max_{\underline{\theta}} Q(\underline{\theta}; \underline{\theta}^{(n)}) \tag{2.41}$$

Using the fact that $D^{10}Q(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) = 0$, this Jacobian matrix can be easily calculated as follows. Since the vector $D^{10}Q(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) = D^{10}Q(M(\underline{\theta}^{(n)}); \underline{\theta}^{(n)}) = 0$, its derivative with respect to the vector $\underline{\theta}^{(n)}$ is the zero matrix, i.e.

$$0 = \frac{\partial}{\partial \underline{\theta}^{(n)}} D^{10}Q(M(\underline{\theta}^{(n)});\underline{\theta}^{(n)}) = DM(\underline{\theta}^{(n)}) \cdot D^{20}Q(M(\underline{\theta}^{(n)});\underline{\theta}^{(n)}) + D^{11}Q(M(\underline{\theta}^{(n)});\underline{\theta}^{(n)}) \tag{2.42}$$

Let $\underline{\theta}^*$ denote the limit of the estimate sequence. Since $\underline{\theta}^{(n+1)} = M(\underline{\theta}^{(n)})$ and $\underline{\theta}^* = M(\underline{\theta}^*)$, in the limit as $\underline{\theta}^{(n)} \to \underline{\theta}^*$, equation (2.42) becomes

$$0 = DM(\underline{\theta}^*)D^{20}Q(\underline{\theta}^*;\underline{\theta}^*) + D^{11}Q(\underline{\theta}^*;\underline{\theta}^*) \tag{2.43}$$

Using (2.38) and then (2.37) will give us the Jacobian matrix,

$$DM(\underline{\theta}^*) = D^{20}H(\underline{\theta}^*;\underline{\theta}^*)\left[D^{20}Q(\underline{\theta}^*;\underline{\theta}^*)\right]^{-1} \tag{2.44}$$

This result appears in theorem 4 of [2], which is repeated in Appendix A.

The rate of convergence of the EM algorithm (2.44) has the following interesting interpretation. The term $D^{20}H(\underline{\theta}^*;\underline{\theta}^*)$ is the Fisher Information matrix $I_{X/Y}$ of X given Y about $\underline{\theta}^*$, i.e. for exponential families it is the variance of the sufficient statistics $t(\underline{x})$ given $\underline{y}$ and $\underline{\theta}^*$. The term $D^{20}Q(\underline{\theta}^*;\underline{\theta}^*)$ is, for regular exponential families, the Fisher Information matrix I_X of X about $\underline{\theta}^*$, i.e. it is the variance of the sufficient statistics $t(\underline{x})$ in the X model without any measurements. From (2.37), the Fisher information I_Y of Y about $\underline{\theta}^*$, for regular exponential families, is given by,

$$I_Y = I_X - I_{X/Y} \tag{2.45}$$

Thus, in the scalar case, the rate of convergence is given by,

$$\alpha = \frac{I_{X/Y}}{I_X} = 1 - \frac{I_Y}{I_X} \tag{2.46}$$

If the complete data is such that it can be predicted well given the observations, i.e. $I_{X/Y}$ is small, then α is small and the EM algorithm converges rapidly On the other hand, if we choose the complete data to be much larger than the observations, then the complete data will carry much more (Fisher) information then the observations: I_Y/I_X will be close to zero, α will be close to unity and the EM algorithm will converge slowly. Indeed, if the complete data is identical to the observations, the EM algorithm converges in one step; however this step is as complicated as solving the original ML problem in the Y model. On the other hand, choosing a complete data that is much larger than the observations, in order to get simple EM steps, will require performing more iterations because the algorithm converges slower.

2.3 The Linear Gaussian case

This section has two objectives. The first objective is to provide an explicit example of the application of the general EM theory developed above. The second and more important goal is to develop results that are referred to later in the thesis in a wide range of applications.

Suppose that the complete data, $\underline{X}$, and the observed (incomplete) data $\underline{Y}$ are related by the linear transformation

$$\underline{Y} = H\underline{X} \tag{2.47}$$

where H is a non invertible matrix. The complete data $\underline{X}$ possesses the following multivariate Gaussian probability density:

$$f_{\underline{X}}(\underline{x};\underline{\theta}) = \left[\det(\frac{2\pi}{\lambda}\Lambda(\underline{\theta}))\right]^{-\lambda/2} \exp\left[-\frac{\lambda}{2}(\underline{x} - \underline{m}(\underline{\theta}))^{\dagger}\Lambda^{-1}(\underline{\theta})(\underline{x} - \underline{m}(\underline{\theta}))\right] \tag{2.48}$$

where $\lambda = 1$ if $\underline{X}$ is real valued, $\lambda = 2$ if $\underline{X}$ is complex valued, and $\dagger$ denotes the conjugate transpose operation. The observations $\underline{Y}$, also possess a Gaussian distribution, and thus the likelihood is given by

$$f_{\underline{Y}}(\underline{y};\underline{\theta}) = \left[\det(\frac{2\pi}{\lambda}\Lambda_y(\underline{\theta}))\right]^{-\lambda/2} \exp\left[-\frac{\lambda}{2}(\underline{y}-\underline{m}_y(\underline{\theta}))^{\dagger}\Lambda_y^{-1}(\underline{\theta})(\underline{y}-\underline{m}_y(\underline{\theta}))\right] \qquad (2.49)$$

where m_y and Λ_y are respectively the mean and the covariance of the observations, given by

$$\underline{m}_y(\underline{\theta}) = H\cdot\underline{m}(\underline{\theta}) \qquad (2.50)$$

$$\Lambda_y(\underline{\theta}) = H\cdot\Lambda(\underline{\theta})\cdot H^{\dagger} \qquad (2.51)$$

We note that our parametric model is such that the parameters define the mean and the covariance of a Gaussian density in a possibly non-linear way. Thus, maximizing the likelihood in this linear Gaussian case may require solving a non-linear optimization problem. Nevertheless, we will be able to invoke results from linear estimation theory and explicitly derive the EM algorithm for this case.

We start developing the EM algorithm for maximizing the likelihood of the observation, $\underline{Y}$, using the complete data, $\underline{X}$, by examining the log-likelihood of $\underline{X}$. By taking the logarithm of (2.48), we get,

$$\begin{aligned}\log f_{\underline{X}}(\underline{x};\underline{\theta}) &= C-\frac{\lambda}{2}\log\det(\frac{2\pi}{\lambda}\Lambda(\underline{\theta}))-\frac{\lambda}{2}(\underline{x}-\underline{m}(\underline{\theta}))^{\dagger}\Lambda^{-1}(\underline{\theta})(\underline{x}-\underline{m}(\underline{\theta}))\\ &= C-\frac{\lambda}{2}\log\det(\frac{2\pi}{\lambda}\Lambda(\underline{\theta}))-\frac{\lambda}{2}\underline{m}^{\dagger}(\underline{\theta})\Lambda^{-1}(\underline{\theta})\underline{m}(\underline{\theta})\\ &-\frac{\lambda}{2}\underline{x}^{\dagger}\Lambda^{-1}(\underline{\theta})\underline{m}(\underline{\theta})+\frac{\lambda}{2}\underline{m}^{\dagger}(\underline{\theta})\Lambda^{-1}(\underline{\theta})\underline{x}-\frac{\lambda}{2}tr(\Lambda^{-1}(\underline{\theta})\underline{x}\underline{x}^{\dagger}) \qquad (2.52)\end{aligned}$$

where C is a constant independent of $\underline{\theta}$ and $tr(\cdot)$ denotes the trace of a matrix. Maximizing this expression with respect to $\underline{\theta}$ is assumed to be easy.

Taking the conditional expectation of (2.52), given $\underline{Y}=\underline{y}$, at a parameter value $\underline{\theta}^{(n)}$, we get,

$$Q(\underline{\theta};\underline{\theta}^{(n)}) = E\left\{\log f_{\underline{X}}(\underline{x};\underline{\theta})/\underline{y};\underline{\theta}^{(n)}\right\}$$

$$\begin{aligned} &= C - \frac{\lambda}{2}\log\det(\frac{2\pi}{\lambda}\Lambda(\underline{\theta})) - \frac{\lambda}{2}\underline{m}^{\dagger}(\underline{\theta})\Lambda^{-1}(\underline{\theta})\underline{m}(\underline{\theta}) + \frac{\lambda}{2}(\underline{x}^{(n)})^{\dagger}\Lambda^{-1}(\underline{\theta})\underline{m}(\underline{\theta}) \\ &\quad + \frac{\lambda}{2}\underline{m}^{\dagger}(\underline{\theta})\Lambda^{-1}(\underline{\theta})\underline{x}^{(n)} - \frac{\lambda}{2}tr(\Lambda^{-1}(\underline{\theta})\Psi^{(n)}) \end{aligned} \tag{2.53}$$

where $\underline{x}^{(n)} = E\left\{\underline{x}/\underline{Y} = \underline{y}; \underline{\theta}^{(n)}\right\}$ and $\Psi^{(n)} = E\left\{\underline{x}\underline{x}^{\dagger}/\underline{Y} = \underline{y}; \underline{\theta}^{(n)}\right\}$.

Maximizing expression (2.53) with respect to $\underline{\theta}$ must be easy, since it has the same functional form, with respect to $\underline{\theta}$, as (2.52).

Since $\underline{X}$ and $\underline{Y}$ are jointly Gaussian, and related by a linear transformation, the conditional expectations required for (2.53) can be computed by straight-forward modifications of known results from linear estimation theory. We obtain,

$$\underline{x}^{(n)} = \underline{m}(\underline{\theta}^{(n)}) + \Gamma(\underline{\theta}^{(n)})\left[\underline{y} - H \cdot \underline{m}(\underline{\theta}^{(n)})\right] \tag{2.54}$$

$$\Psi^{(n)} = \left[I - \Gamma(\underline{\theta}^{(n)}) \cdot H\right]\Lambda(\underline{\theta}^{(n)}) + (\underline{x}^{(n)})(\underline{x}^{(n)})^{\dagger} \tag{2.55}$$

where I is the identity matrix and $\Gamma(\underline{\theta})$ is the "Kalman gain" defined by

$$\Gamma(\underline{\theta}) = \Lambda(\underline{\theta})H^{\dagger}\left[H\Lambda(\underline{\theta})H^{\dagger}\right]^{-1} \tag{2.56}$$

Note that if we set $\underline{\theta}^{(n)} = \underline{\theta}$, equations (2.54) and (2.55) are the well known formulae for the conditional expectation in the Gaussian case, e.g. [20].

The E and M steps of the EM algorithm for the linear Gaussian case may now be stated explicitly as follows. Having a current estimate, $\underline{\theta}^{(n)}$, the algorithm iterates between,

- **The E step**

  Calculate $\underline{x}^{(n)}$ and $\Psi^{(n)}$, by (2.54) and (2.55). Note that the sufficient statistics of Gaussian distribution are composed of linear and quadratic functions of the data.

- **The M step**

  Update $\underline{\theta}$ by maximizing the expression in (2.53). The explicit solution is some functional of the statistics calculated in the E step.

2.4 The EM algorithm with varying complete data

In this section, we present a variation of the EM algorithm , where the complete data may vary from iteration to iteration. This variation is referred to as the Extended EM (EEM) algorithm.

As mentioned above, the choice of complete data is the critical factor in designing an EM algorithm for a given problem. This choice determines the complexity of the algorithm and its convergence rate; it may also affect the convergence point, leading to a different stationary point for a different choice of complete data. An alternative to choosing a fixed complete data, is to let the complete data vary from iteration to iteration. The choice of complete data may vary according to a fixed rule or may depend on the current value of the estimate. By allowing the complete data to vary, we can achieve the following useful properties:

- Additional iterative algorithms are incorporated in the EM framework.
- Simpler algorithms may emerge.
- The algorithm may converge faster.
- Varying the complete data may enable the algorithm to escape from unwanted stationary points.

We start by presenting the algorithm formally, and giving its properties. Then, we will motivate the EEM algorithm and suggest strategies for varying the complete data.

2.4.1 General theory

Suppose we observe $\underline{y} \in Y$, where Y denotes the sample space of the observations, and the probability of $\underline{y}$ is $f_Y(\underline{y};\underline{\theta})$ indexed by $\underline{\theta} \in \Theta$. The observed sequence may be viewed as being incomplete with respect to a *family* of complete data X_β, indexed by $\beta \in \mathcal{B}$, where $\mathcal{B}$ is an arbitrary index set. Each X_β is a sample space with an associated p.d.f., $f_{X_\beta}(\underline{x}_\beta;\underline{\theta})$, also indexed by $\underline{\theta} \in \Theta$. For any β, a sample of the complete data, $\underline{x}_\beta$, is related to the observations by,

$$\underline{y} = T_\beta(\underline{x}_\beta) \tag{2.57}$$

where $\underline{y}$ denote the observations and T_β is a non-invertible transformation.

In complete analogy to (2.8), we may write for all β,

$$\log f_Y(\underline{y};\underline{\theta}) = E\left\{\log f_{X_\beta}(\underline{x}_\beta;\underline{\theta}) \,\middle|\, \underline{y};\underline{\theta}'\right\} - E\left\{\log f_{X_\beta/Y}(\underline{x}_\beta/\underline{y};\underline{\theta}) \,\middle|\, \underline{y};\underline{\theta}'\right\} \tag{2.58}$$

or, using the notation in sectic¬ 2.1,

$$L(\underline{\theta}) = Q_\beta(\underline{\theta},\underline{\theta}') - H_\beta(\underline{\theta},\underline{\theta}') \tag{2.59}$$

Using this relation and invoking Jensen's inequality we may prove the following lemma.

Lemma 2.1 *For a given parameter value $\underline{\theta}_1$, if for any β, another value $\underline{\theta}_2$ satisfies*

$$Q_\beta(\underline{\theta}_2,\underline{\theta}_1) > Q_\beta(\underline{\theta}_1,\underline{\theta}_1)$$

then,

$$L(\underline{\theta}_2) > L(\underline{\theta}_1)$$

Note that this lemma states that we have a procedure for strictly increasing the likelihood, if we can find any complete data, for which the function Q_β may be strictly increased.

The Extended EM algorithm is now presented formally. We note that the choice of complete data in each iteration may depend on the current and previous estimates and on the iteration index.

- Start, $n = 0$: guess $\underline{\theta}^{(0)}$

- Until some convergence criterion is met,

  - Choose a complete data $X_{\beta^{(n)}}$, where,

$$\beta^{(n)} = f(n, \underline{\theta}^{(0)}, \cdots, \underline{\theta}^{(n)}) \tag{2.60}$$

  - **The E step:** calculate

$$Q_{\beta^{(n)}}(\underline{\theta}; \underline{\theta}^{(n)}) = E\left\{ \log f_{X_{\beta^{(n)}}}(\underline{x}_{\beta^{(n)}}; \underline{\theta}) \,\middle|\, \underline{y}; \underline{\theta}^{(n)} \right\} \tag{2.61}$$

  - **The M step:** solve

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta} \in \Theta} Q_{\beta^{(n)}}(\underline{\theta}; \underline{\theta}^{(n)}) \tag{2.62}$$

  - $n = n + 1$

The proposed EEM algorithm preserves the basic monotonicity property of the EM algorithm. Since the convergence properties of the EM algorithm were proved using the Global Convergence theorem, they will follow through to the EEM algorithm, if the conditions developed in section 2.2 holds for every $\beta \in \mathcal{B}$. These properties of the EEM algorithm hold regardless of the rule f we use for changing the complete data. A carefully designed rule may provide an algorithm with better properties, however. Such rules will be suggested in the following section.

2.4.2 Motivation and rules for changing the complete data

The following simple situation may motivate the usage of the EEM algorithm. Suppose that, in a specific problem, two complete data definitions may be considered. Each choice of complete data generates a different algorithm for maximizing the likelihood of the observations with different convergence properties. Following the EEM idea, we may switch between these algorithms. If one specification of complete data generates a simpler but slower algorithm, we will start using the simpler algorithm and then, near the convergence point, switch to the other algorithm to converge to the solution faster. If one algorithm converges to an unwanted stationary point of the likelihood, which is not a fixed point of the other algorithm, we will switch to the other algorithm to avoid it.

In general, the family of complete data specifications is indexed by $\beta \in \mathcal{B}$, where $\mathcal{B}$ is an arbitrary index set, say, a subset of the k-dimensional Euclidean space. The choice of complete data may depend on the current estimate of the parameters. In the general case, the following strategies may be used to obtain algorithms with better convergence properties.

Accelerating the convergence rate

Loosely speaking, the rate of convergence is faster, when the complete data can be better predicted from the observations. Thus, changing the complete data in each iteration, depending on the current parameter model, $\underline{\theta}^{(n)}$, in such a way that it may be better predicted from the observations, will improve the rate of convergence.

More specifically, an EEM iteration is given by $\underline{\theta}^{(n+1)} = M_\beta(\underline{\theta}^{(n)})$. It satisfies

$$D^{10}Q_\beta(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) = 0$$

Since the vector $D^{10}Q_{\beta}(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) = D^{10}Q_{\beta}(M_{\beta}(\underline{\theta}^{(n)});\underline{\theta}^{(n)}) = 0$, its derivative with respect to the vector $\underline{\theta}^{(n)}$ is the zero matrix, i.e.

$$0 = \frac{\partial}{\partial\underline{\theta}^{(n)}} D^{10}Q_{\beta}(M_{\beta}(\underline{\theta}^{(n)});\underline{\theta}^{(n)}) = DM_{\beta}(\underline{\theta}^{(n)})\cdot D^{20}Q_{\beta}(M_{\beta}(\underline{\theta}^{(n)});\underline{\theta}^{(n)}) + D^{11}Q_{\beta}(M_{\beta}(\underline{\theta}^{(n)});\underline{\theta}^{(n)}) \tag{2.63}$$

Now, in the limit as $n \to \infty$, $\underline{\theta}^{(n)} \to \underline{\theta}^{(n+1)} \to \underline{\theta}^{*}$ and we get

$$DM_{\beta}(\underline{\theta}^{(n)}) = -D^{11}Q_{\beta}(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)})\left[D^{20}Q_{\beta}(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)})\right]^{-1} = -D^{11}Q_{\beta}(\underline{\theta}^{(n)};\underline{\theta}^{(n)})\left[D^{20}Q_{\beta}(\underline{\theta}^{(n)};\underline{\theta}^{(n)})\right]^{-1} \tag{2.64}$$

The largest eigenvalue of $DM_{\beta}(\underline{\theta}^{(n)})$ will define the rate of convergence. To accelerate the convergence, we want to choose a complete data (i.e. β), that will minimize this largest eigenvalue.

Depending on the set $\mathcal{B}$, it may be possible to find $\beta \in \mathcal{B}$, in terms of $\underline{\theta}^{(n)}$, that solves the following equation

$$D^{11}Q_{\beta}(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) = f(\beta,\underline{\theta}^{(n)}) = 0 \tag{2.65}$$

In this case, the convergence rate of the EEM algorithm will be *superlinear*.

Avoiding unwanted convergence points

We wish to find a global maximizer of the likelihood function. However, under the conditions of section 2.2, an EM algorithm with a fixed complete data specification is only guaranteed to convergence to a stationary point of the likelihood. Nevertheless, not every stationary point of the likelihood is a fixed point of an EM algorithm. If a family of complete data is given and a specific stationary point is not a fixed point of all the EM iterations that correspond to the members of this family, then following lemma 2.1, we may find a complete data specification, that will take us away from this unwanted stationary point. Once we have avoided a stationary point and we find a parameter value, for which the likelihood is higher, then due to the monotonicity property, we will never return to this stationary point.

When the given set of possible complete data is indexed by β, the following rules for choosing β in each iteration, in order to avoid unwanted convergence points are suggested:

- Choose a random β in its domain.
- Search for β that give the largest increase in the likelihood. If searching the entire domain of β is complicated, search in a sub-domain, which may be picked randomly.

One may argue that these rules are heuristic and ad-hoc. However, the whole area of global optimization of non-convex functions is heuristic, and depends on the specific goal function. Our approach potentially provides an improvement, within the framework of the EM algorithm, in the sense that, even when it fails to find the global maximizer, it finds a better local maximizer.

2.5 The EM algorithm for general estimation criteria

The EM idea may be applied to general inference problems, other than parameter estimation problems, and a variety of estimation methods, other than the ML method. We will start by suggesting a formal structure of the EM algorithm for general estimation methods. Then, using the general Minimum Information criterion, we will show that a wide class of estimation methods reduce to optimizing a criterion composed of the log-likelihood and an additive penalty term. An EM method for optimizing these criteria, analogous to the EM method for ML, will be suggested.

2.5.1 Formal structure

As before, let Y be the sample space of the observations, and X the sample space of the complete data.

Suppose we observe $\underline{x} \in X$. We want to find a model or a structure, π, that will "explain" $\underline{x}$. Since we consider statistical inference methods, a model will define a probability distribution or a p.d.f., $f_X(\underline{x}; \pi)$, over the set X. The model may be as simple as a parameter specification or as complicated as a full, unconstrained description of the underlying probability measure.

The a-priori knowledge, the model complexity and a cost function for measuring goodness-of-fit of the model to the observation, will determine the procedure for estimating this model. There are many ways to incorporate knowledge, complexity and goodness-of-fit measures, which explains the variety of criteria for statistical inference. However, in any inference procedure we may find the following two characteristics:

- *Extraction of sufficient statistics:* Not all the observed data is relevant to the model estimation goal. Extracting only $\mathcal{T}(\underline{x})$ from the data, where $\mathcal{T}(\cdot)$ is many to one function, is sufficient.

- *Optimization:* The possible models are compared and a model estimate $\hat{\pi}$ is generated by a procedure $\mathcal{F}\{\mathcal{T}(\underline{x})\} = \hat{\pi}$, which is usually a result of solving an optimization problem:

$$\hat{\pi} = \arg\min_{\pi} F(\mathcal{T}(\underline{x}); \pi)$$

We assume that given $x \in X$, i.e. given the complete data, we have a satisfactory solution for the model estimation problem. In other words, there exists a way to incorporate the a-priori knowledge, to measure the complexity and goodness-of-fit of a candidate model, to calculate the required statistics and to solve any optimization problem, that is implied from the above. A satisfactory solution consists of a set of formulas, a detailed algorithm or a program. Thus, we may imagine the existence of a "black box", whose input is a measurement $x \in X$ and whose output is a model estimation, $\hat{\pi}$.

Suppose now that we observe the incomplete data, $\underline{y} \in Y$, where $\underline{y} = T(\underline{x})$ and $T(\cdot)$ is non-invertible (many to one) transformation. Any candidate model will define a p.d.f., $f_Y(\underline{y};\pi)$, over the set Y where

$$f_Y(\underline{y};\pi) = \int_{\mathcal{X}(\underline{y})} f_X(\underline{x};\pi)d\underline{x} \tag{2.66}$$

and the set $\mathcal{X}(\underline{y})$ is given by (2.4).

We assume that we *do not* have a satisfactory direct way to determine the model or the structure, given the incomplete observations, either because we cannot specify the procedure for determining the model, or, when the procedure is specified, simply because implementing that procedure (e.g. solving the implied optimization problem) is difficult.

The EM algorithm, which we now formally present, is a possible method for determining the model, given the incomplete observations, by making an essential use of the availability of a satisfactory estimation procedure for complete data observations.

- Start, $n = 0$, initial model $\pi^{(0)}$

- Iterate (until some convergence criterion is met)

  - **The E step:** calculate

$$\mathcal{T}^{(n)} = E\left\{\mathcal{T}(\underline{x}) \,\middle|\, \underline{Y} = \underline{y};\pi^{(n)}\right\} = \int_{\mathcal{X}(\underline{y})} \mathcal{T}(\underline{x}) f_{X/Y}(\underline{x}/\underline{y};\pi^{(n)})d\underline{x} \tag{2.67}$$

- **The M step:** solve

$$\pi^{(n+1)} = \mathcal{F}\left\{\mathcal{T}^{(n)}; \pi\right\} = \arg\min_{\pi} F(\mathcal{T}^{(n)}; \pi) \tag{2.68}$$

2.5.2 The EM algorithm for the Minimum Information criterion

Minimum Information (MI) is a general method for solving inference problems, suggested originally by Solomonoff [21] and recently by Hart [22]. This method generalizes the ML and MAP methods. This method may be applied to situations, where a general structure or model π should be estimated. This method also enables to incorporate a more general a-priori information.

Given data, $\underline{y} \in Y$, the MI method estimates the model π by,

$$\hat{\pi} = \arg\min_{\pi} I(\underline{y}, \pi) \tag{2.69}$$

where $I(\cdot)$ denotes the (self) information. The joint information $I(\underline{y}, \pi)$ may be written as,

$$I(\underline{y}, \pi) = I(\underline{y}/\pi) + I(\pi) \tag{2.70}$$

where $I(\underline{y}/\pi)$ is the conditional (self) information.

The MI criterion implies many estimation procedures, since there are many notions and definitions of information. We will usually use the more quantitative ones:

- *Combinatorial information*, due to Hartley [23].
- *Probabilistic (Shannon) information*, due to Shannon [24] and Wiener [25].
- *Algorithmic (Kolmogorov) information*, due to Solomonoff [21], Kolmogorov [26] and Chaitin [27].

These three notions of information are summarized in [26].

Shannon information is the most adequate for $I(\underline{y}/\pi)$, since a given model provides a probabilistic description of the observations. Thus, $I(\underline{y};\pi) = -\log f_Y(\underline{y};\pi)$, and the MI criterion reduces to the minimization of likelihood-like criterion,

$$G(\pi) = -\log f_Y(\underline{y};\pi) + I(\pi) \tag{2.71}$$

Shannon information may be adequate to describe the information, given by specifying a model π, if we know that all possible models belong to a well defined set Π, and an a-priori p.d.f., $f_\Pi(\pi)$, defined over Π, is given. In this case $I(\pi) = -\log f_\Pi(\pi)$ and the MI criterion reduces to the MAP criterion, i.e. we estimate the model by,

$$\hat{\pi} = \arg\max_{\pi\in\Pi} \left[\log f_Y(\underline{y};\pi) + \log f_\Pi(\pi)\right] \tag{2.72}$$

Other examples will involve the algorithmic notion of information, which measures the information of an observed data by the number of bits needed to describe it. As shown, e.g. [28], the algorithmic (Kolmogorov) information cannot be computed when no constraints on the "language" used to describe the data are specified. However, given a constrained framework, this information may be specified quantitatively.

A special case, that uses the algorithmic information, in a constrained framework, as a criterion to weight a given model, is known as the Minimum Description Length (MDL). This criterion was suggested by Rissanen [29,30,31]. The description length needed to describe the model was given explicitly by Rissanen for the problem of determining the parameters $\theta_1, \cdots, \theta_n$ together with their number n. The conditional information of the data, given the model, is interpreted as the code length needed to describe the observation, given the model. This term is, as above, the log-likelihood, or the Shannon information, of the data, given the model. Thus, in this case the MDL criterion requires solving,

$$\hat{n}, \hat{\underline{\theta}} = \arg\min_{n,\underline{\theta}} [-\log f_Y(\underline{y}; \underline{\theta}) + \frac{1}{2} n \log N] \tag{2.73}$$

where N is the length of the observation sequence.

For all these cases, where the conditional information of the data with respect to the model is the probabilistic (Shannon) information, the MI criterion has the form of (2.71), and the EM algorithm of (2.67),(2.68) becomes,

- Start, $n = 0$, initial model $\pi^{(0)}$
- Iterate (until some convergence criterion is met)
  - **The E step:** calculate

    $$Q(\pi; \pi^{(n)}) = E\left\{\log f_X(\underline{x}; \pi) \,\Big|\, \underline{y}, \pi^{(n)}\right\} \tag{2.74}$$

    Note that this step corresponds to the E step of the regular EM algorithm, (2.17).
  - **The M step:** minimize

    $$\pi^{(n+1)} = \arg\min_{\pi} \left[-Q(\pi; \pi^{(n)}) + I(\pi)\right] \tag{2.75}$$
  - $n = n + 1$

This algorithm was suggested in [2] for the MAP criterion.

It is easy to show that each iteration improves (decreases) the likelihood-like goal function of the observation (2.71). The goal function may be written as

$$G(\pi) = -L(\pi) + I(\pi) = -[Q(\pi; \pi') - H(\pi; \pi')] + I(\pi) \tag{2.76}$$

where $Q(\cdot;\cdot)$ and $H(\cdot;\cdot)$ are defined in (2.11). Thus by a simple extension of theorem 2.1, we conclude that

$$G(\pi^{(n+1)}) \leq G(\pi^{(n)}) \tag{2.77}$$

where equality holds if and only if both

$$-Q(\pi^{(n+1)},\pi^{(n)})+I(\pi^{(n+1)})=-Q(\pi^{(n)},\pi^{(n)})+I(\pi^{(n)}) \tag{2.78}$$

and

$$f_{\underline{X}/\underline{Y}}(\underline{x};\pi^{(n+1)})=f_{\underline{X}/\underline{Y}}(\underline{x}/\underline{y};\pi^{(n)}),\quad \text{a.e in } \mathcal{X}(\underline{y}) \tag{2.79}$$

Due to the monotonicity property, if the goal function is bounded, the sequence, $\{G^{(n)}\}$, where $G^{(n)}=G(\pi^{(n)})$, must converge to a limit G^*. Also, a global optimizer of the goal function must be a fixed point of the algorithm.

For proving other convergence properties, such as convergence to a stationary point, convergence of the model estimate sequence and the rate of convergence, we need that the set of models Π will be a subset of a metric space. Otherwise, the notions of distance, convergence, continuity etc. are undefined. When Π is a subset of a metric space, those convergence properties are readily available. We will use the results developed in section 2.2, where $Q(\underline{\theta};\underline{\theta})$ is is replaced by $-Q(\pi;\pi')+I(\pi)$. Thus, convergence to a stationary point is guaranteed if

$$-Q(\pi,\pi')+I(\pi)\quad\text{is continuous in both } \pi \text{ and } \pi' \tag{2.80}$$

The rate of convergence near the convergence point is the largest eigenvalue of

$$-D^{20}H(\pi^*;\pi^*)\left[-D^{20}Q(\pi^*;\pi^*)+D^2I(\pi^*)\right]^{-1} \tag{2.81}$$

where $D^{20}Q(\underline{\theta}^*;\underline{\theta}^*)$ in equation (2.44) is replaced by $-D^{20}Q(\pi^*;\pi^*)+D^2I(\pi)$.

2.6 Possible signal processing applications

The EM algorithm and its extensions developed in this chapter, will be applied later in this thesis to solve a variety of signal processing problems. We will conclude this chapter by trying to characterize the problems that are naturally solved using the EM method.

In many cases, we will describe the possible applications of the EM algorithm, as having noisy and incomplete observations. As an example, we became interested in the EM algorithm, while considering the problem of power spectrum estimation from a short record of observations. The modern spectral estimation techniques, e.g. Burg's Maximum Entropy technique [32], achieve high resolution by artificially extending the observation period or the autocorrelation support. This solution requires the exact knowledge of the autocorrelation values. However, the sample autocorrelation values are only noisy estimates of the real correlation values. In our opinion, a better approach for high resolution spectrum estimation is to consider the short observations record as noisy and incomplete and to model the spectrum estimation problem as a statistical ML problem. Following these considerations, we have presented in [33], a parametric spectrum estimation method based on the EM algorithm. This method, suggested originally in [33], was later investigated by various authors [34], [35].

In general we can define two classes of possible signal processing applications. The first class contains signal processing problems having partial or distorted observations. The problems of this class are characterized as follows: We may be interested in estimating unknown parameters or even reconstructing a whole waveform. For this task, it is desired to measure some signals. However, we observe only a mapping of the desired signals, such as,

- the magnitude of the signal(s) or
- the sign or the hard limited version of the signal(s) or
- the quantized signal(s) or

- the aliased signal(s)

or any other partial information. Since the observations are distorted and incomplete the statistical problem associated with the signal processing problem is complicated. The EM algorithm provides a natural solution to these problems, where the complete data is defined as the undistorted signals. The algorithm iterates between estimating these undistorted signals and updating the desired parameters.

The second class of applications contains signal processing problems for which the observations are described as a combination of simpler signals. We are interested in estimating signal parameters or reconstructing a signal waveform; however, instead of observing the desired signals, we observe a combination such as,

- sum of signals or
- multiplication of signals or
- convolution of signals

or any other combination. We use a probabilistic modeling of the various signals. With the observations above, the signal processing problem is modeled as doubly (or multiply) stochastic phenomena. The statistical problems generated by doubly stochastic models are usually complicated. The EM algorithm provides a natural solution to these problems, where the complete data is the set consisting of all the separate signal components. In the doubly stochastic case this complete data is equivalent to the set consisting of one of the "hidden" signal components and the combined observations.

Many applications that belong to this class may be considered, since combined signals are common in many practical situations. In chapters 4 and 5, we will consider examples that belong to this class of applications and solve them via the EM algorithm.

We conclude this chapter by briefly presenting two previously suggested applications, which fall naturally within of the EM framework. The first application is the problem of estimating the parameters of a stationary signal in a stationary noise. The EM solution to this problem is referred to as the iterative Wiener filter. The second application is the estimation of the parameters of a Hidden Markov Model (HMM). The EM solution to this problem is widely known as Baum's algorithm. We point out that both these previous applications belong to the class presented above and are thus analogous. We believe that an additional insight to the HMM analysis can be gained by presenting it in terms of an iterative filtering technique.

The iterative Wiener filter

Let $s(t;\underline{\theta})$ be a stationary (discrete) random process, and suppose that this process is observed in additive noise, i.e. we observe

$$y(t) = s(t;\underline{\theta}) + n(t) \tag{2.82}$$

where $n(t)$ is also a stationary process. We are interested in estimating the signal parameters, and, in some cases, in filtering the signal.

This model was suggested in [3] to represent a speech enhancement problem arising from single microphone measurements. In this case, the speech signal is modeled as a stationary autoregressive (AR) process with unknown coefficients, referred to in the speech context as Linear Prediction Coefficients (LPC). Speech signals are frequently modeled as AR processes, since this model captures the important features of the speech signal, at least for a short enough observation window. We may be interested in finding the speech LPC parameters, say for vocoding or for use in a speech recognition system, or in enhancing the speech signal. If the speech signal is observed without the noise, the LPC parameters are easily estimated by solving the appropriate normal linear equations. However, estimating the parameters from noisy observations is complicated.

The iterative Wiener filter suggested in [3] can be interpreted as an EM algorithm as follows. Let the complete data be defined as the speech signal, $s(t;\underline{\theta})$, and the noise signal, $n(t)$, separately. An equivalent definition is to let the complete data be the speech signal, $s(t;\underline{\theta})$ in addition to the observed signal, $y(t)$. From the discussion above, if we observe the signal without the noise, the task of estimating the parameters will be easy. In addition, having the signal parameters, we may estimate the signal, i.e. filter it from the noise, using a Wiener filter. This suggests an EM algorithm, that iterates between Wiener filtering, applied to the observed signal, using the current spectral (or LPC) parameters of the signal (the E step) and updating the spectral parameters using the filtered signal (the M step).

We note that the filtered speech signal, $\hat{s}(t)$, is achieved as a by-product, while implementing the E step of the algorithm.

Hidden Markov Models and Baum's algorithm

Hidden Markov Models (HMM) are interesting and rich statistical models, that have been used frequently to model complex real problems. The iterative Baum's algorithm, suggested in [36], which is now recognized as an instance of the EM algorithm, was suggested for the statistical analysis of HMM. These models are extensively used for modeling the mechanism that generates the speech signal, and are applied in speech recognition systems. A review of HMM may be found in [37], and a review of their application to automatic speech recognition may be found in [38]. We will now briefly present the Hidden Markov Models and Baum's algorithm, from a different perspective.

Suppose we observe the sequence $y_0, y_1, \cdots, y_{N-1}$. A hidden Markov model assumes that the observations are probabilistic functions of a finite alphabet Markov chain. In other words, there exists a hidden Markov sequence, $s_0, s_1, \cdots, s_{N-1}$, such that the observed sequence is a result of combining the Markov sequence with another stochastic contribution, i.e. the model assumes that for each i, there is a conditional probability distribution

$$p(y_i/s_i = \alpha_k) \qquad k = 1, \cdots, M \tag{2.83}$$

where $\{\alpha_1, \cdots, \alpha_M\}$ is the alphabet of the Markov chain.

Using this point of view, the observations, $\{y_i\}$, in a hidden Markov model are the result of a "signal" - the markov chain $\{s_i\}$, combined with "noise".

The unknown "signal" parameters in this case are the transition probabilities represented by the matrix Ψ, and in a non-stationary case, the initial probability vector $\underline{\pi}$. Sometimes, the "noise" parameters, i.e. the parameters that define the conditional probabilities of (2.83), are also unavailable. The ML problem for estimating these parameters is usually too complicated to solve directly.

The complete data will be the hidden "signal" in addition to the observations, which is equivalent to observing the "signal" and the "noise" separately. This complete data is analogous to the complete data used in the iterative Wiener algorithm. Suppose we observe the hidden Markov chain. If the unknown parameters are all entries of the transition matrix, the ML estimate of say $\Psi_{k,\ell}$ is achieved by counting the number of transitions from the symbol α_ℓ to the symbol α_k, divided by the number of occurrences of the symbol α_ℓ. The "noise" parameters may also be estimated easily, given the "noise" realization, which is determined, when the observations and the underlying hidden Markov chain are available.

The specific procedure is defined depending on the specific parameterization.

In the iterative Wiener algorithm where a stationary signal in a stationary noise have been considered, the E step was implemented in the frequency domain using a Wiener Filter. In the HMM case, where the signal is a Markov chain, Bellman's sequential dynamic programming algorithm [39], can be used. Bellman's algorithm is sometimes referred to as the Viterbi algorithm. The E step, which estimates the required statistics of the hidden Markov chain, needed for updating the parameters, is thus implemented by an efficient sequential algorithm.

The detailed Baum's algorithm is presented explicitly in [37], page 11 and elsewhere. The interpretation of this algorithm as an iterative filtering algorithm gives an additional insight, that may help suggest enhancements to this algorithm.

Chapter 3

Sequential and Adaptive algorithms

In this chapter, we will suggest and investigate sequential and adaptive algorithms, that are based on the EM concept. Sequential and adaptive algorithms correspond to the case where the data is processed sequentially and an output is expected, whenever each new block of data is processed. We denote the $n+1^{st}$ data block by $\underline{y}_{n+1}$, and suppose that the desired output is an estimate of a parameter vector, $\underline{\theta}$. The general structure of any sequential (or adaptive) algorithm is,

$$\underline{\theta}^{(n+1)} = G_{n+1}\left(\underline{\theta}^{(n)}; \underline{y}_{n+1}\right) \tag{3.1}$$

The desired output of a sequential algorithm is either identical to or at least asymptotically identical to the result achieved by processing the whole data at once. The advantage of the sequential algorithm over the batch algorithm is not in the final result, but in computational and storage efficiency and in the fact that an output may be provided without having to wait for all data to be processed. Adaptive algorithms correspond to the case where the underlying system features are time varying, and the algorithm is expected to track the varying parameters. In this case, processing all the available data jointly is not desired, even if we can accommodate the computational and storage load of the batch algorithm and can afford to wait to the end of the data, since different data segments correspond to different parameters values.

Sequential and adaptive algorithms may be suggested based on an iterative algorithm in the following way. Given the current estimate, $\underline{\theta}^{(n)}$, the next iteration takes into account the new data block, $\underline{y}_{n+1}$, for generating the updated value, $\underline{\theta}^{(n+1)}$. A well known example is the stochastic gradient algorithm, which is an adaptive version of the iterative gradient algorithm. As another example, the recursive least-squares (RLS) algorithm and the (extended) Kalman algorithm are sequential algorithms based on the iterative Newton-Raphson method. Similarly, the iterative EM algorithm may suggest sequential and adaptive algorithms. These algorithms will be developed in this chapter.

The chapter is organized as follows. In section 3.1 we will develop sequential algorithms based on the EM method, that may be applied only when the underlying estimation problem has a special structure. In section 3.2 we will use approximations and develop sequential and adaptive algorithms, based on the EM method, that may be applied in general. The sequential algorithms, presented in this chapter, will be analyzed in section 3.3.

3.1 Sequential EM algorithms based on problem structure

In this section, we will identify the cases where the underlying estimation problem has a special structure, and suggest sequential EM algorithms that exploit this special structure. The next section will demonstrate that, in general, a sequential algorithm cannot be derived as a direct consequence of the EM method. Only using approximations will we be able to suggest sequential and adaptive algorithms for the general case.

3.1.1 Sequential EM with exact EM mapping

Throughout this chapter, we will consider the observed data as blocks, $\underline{y}_1, \underline{y}_2, \dots, \underline{y}_n, \dots$, to be processed sequentially. The complete data is denoted $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n, \dots$, and is chosen so that each block of observed data $\underline{y}_n$, corresponds to a block of complete data, $\underline{x}_n$, by

$$\underline{y}_n = T_n(\underline{x}_n) \tag{3.2}$$

where $T_n(\cdot)$ is a non-invertible transformation.

In this environment the log-likelihood of the observations, after $n+1$ data blocks have been observed, is given by,

$$L_{n+1}(\underline{\theta}) = \log f_{Y_{n+1} \cdots Y_1}(\underline{y}_{n+1}, \cdots, \underline{y}_1; \underline{\theta}) \tag{3.3}$$

Using the complete data, $\underline{x}_1, \cdots, \underline{x}_{n+1}$, and following the identity (2.12), the log-likelihood of the observations may be written as,

$$L_{n+1}(\underline{\theta}) = Q_{n+1}(\underline{\theta}; \underline{\theta}') - H_{n+1}(\underline{\theta}; \underline{\theta}') \tag{3.4}$$

where

$$Q_{n+1}(\underline{\theta}, \underline{\theta}') = E\left\{\log f_{X_{n+1} \cdots X_1}(\underline{x}_{n+1}, \cdots, \underline{x}_1; \underline{\theta}) \,\Big|\, \underline{y}_{n+1}, \cdots, \underline{y}_1; \underline{\theta}'\right\} \tag{3.5}$$

$$H_{n-1}(\underline{\theta}, \underline{\theta}') = E\left\{\log f_{X_{n+1} \cdots X_1 / \underline{y}_{n+1}, \cdots, \underline{y}_1}(\underline{x}_{n+1}, \cdots, \underline{x}_1; \underline{\theta}) \,\Big|\, \underline{y}_{n+1}, \cdots, \underline{y}_1; \underline{\theta}'\right\} \tag{3.6}$$

An EM algorithm for solving the maximum likelihood problem, given these $n+1$ blocks of data, using the above definition of complete data, is given by the following iteration,

$$\underline{\theta}^{(k+1)} = \arg\max_{\underline{\theta} \in \Theta} Q_{n+1}(\underline{\theta}; \underline{\theta}^{(k)}) = \arg\max_{\underline{\theta} \in \Theta} E\left\{\log f_{X_{n+1} \cdots X_1}(\underline{x}_{n+1}, \cdots, \underline{x}_1; \underline{\theta}) \,\Big|\, \underline{y}_{n+1}, \cdots, \underline{y}_1, \underline{\theta}^{(k)}\right\} \tag{3.7}$$

where k denotes the iteration index and n the data index.

A sequential EM algorithm with exact EM mapping is a method that recalculates in each iteration, as more data is processed, the exact steps of the EM algorithm for maximizing the new likelihood function. For convenience, suppose we perform a single EM iteration for each new observed data block, i.e. the iteration and the data indices are equivalent. This mapping is given by (3.7) where k is replaced by n. This EM mapping is, in general, a function of all given observed data blocks; thus, it may written abstractly as,

$$\underline{\theta}^{(n+1)} = M_{n+1}(\underline{\theta}^{(n)}; \underline{y}_{n+1}, \cdots, \underline{y}_1) \tag{3.8}$$

The exact EM iteration may be implemented recursively, when the effect of the past data blocks, $\underline{y}_n, \cdots, \underline{y}_1$, can be summarized into a small number of simple quantities. We may algebraically manipulate the given expression for the EM iteration and achieve an equivalent expression, that may be written abstractly as the mapping,

$$\underline{\theta}^{(n+1)} = M'_{n+1}\left(\underline{\theta}^{(n)}; \underline{y}_{n+1}, \underline{g}(\underline{y}_n, \cdots, \underline{y}_1)\right) \tag{3.9}$$

where $\underline{g}$ indicates easily stored and updated functions of the past observations.

We will assume that the structure of (3.9) may be achieved for all n. In this case, we suggest the following sequential EM algorithm:

- Start, $n = 0$: Guess $\underline{\theta}^{(0)}$. Initialize $\underline{g}(\cdot, \cdots) = 0$
- For each new data block, $\underline{y}_{n+1}$.
  - **Exact EM mapping:** Update parameters,

$$\underline{\theta}^{(n+1)} = M'_{n+1}\left(\underline{\theta}^{(n)}; \underline{y}_{n+1}, \underline{g}(\underline{y}_n, \cdots, \underline{y}_1)\right) \tag{3.10}$$

  - Update and record $\underline{g}(\underline{y}_{n+1}, \cdots, \underline{y}_1)$ for the next step

– $n = n + 1$

In each step, this algorithm implements the exact EM mapping for maximizing the new likelihood $L_{n+1}(\underline{\theta})$, and thus, $L_{n+1}(\underline{\theta}^{(n+1)}) \geq L_{n+1}(\underline{\theta}^{(n)})$.

This algorithm has been presented abstractly so far. To fix ideas, we now present a simple example, in which a linear least squares problem is solved recursively using this algorithm.

Example: Sequential Least Squares EM algorithm

It is well known that the linear least-squares problem may be posed as a statistical maximum likelihood problem, in the following way. Suppose we observe a vector, $\underline{y} = (y_1, \cdots, y_n)^T$, given by,

$$\underline{y} = A \cdot \underline{\theta} + \underline{n} \tag{3.11}$$

where $\underline{\theta} = (\theta_1, \cdots, \theta_k)^T$ is the unknown parameter vector, $\underline{n} = (n_1, \cdots, n_n)^T$ is the noise vector, where $\{n_i\}$ are i.i.d random variables distributed normally with zero mean and variance σ^2, and A is a given $(n \times k)$ matrix, which may be written by columns as $A = [\underline{a}_1, \cdots, \underline{a}_k]$ or by rows as $A^T = [\underline{\alpha}_1, \cdots, \underline{\alpha}_n]$. In this case maximizing the likelihood of the observation yield a least-squares problem as,

$$\hat{\underline{\theta}}_{ML} = \arg\max_{\underline{\theta}} \log f_Y(\underline{y}; \underline{\theta}) = \arg\min_{\underline{\theta}} \frac{1}{2\sigma^2} \|\underline{y} - A \cdot \underline{\theta}\|^2 \tag{3.12}$$

We start to develop an EM algorithm to this problem by choosing the complete data. Suppose that the vectors $\{\underline{x}_j\}_{j=1}^{k}$ are defined as,

$$\underline{x}_j = \underline{a}_j \cdot \theta_j + \underline{n}_j \tag{3.13}$$

where $\underline{n}_j$ is $(n \times 1)$ noise vector, whose components n_{ji} are zero mean Gaussian i.i.d random variables with variance $\beta_j\sigma^2$. Assuming that $\{\underline{n}_j\}$ are uncorrelated and that $\sum_{j=1}^k \beta_j = 1$, we have

$$\underline{y} = \sum_{j=1}^{k} \underline{x}_j \tag{3.14}$$

The complete data is defined as the set $\{\underline{x}_j\}_{j=1}^k$. Writing the complete data as a long vector, $\underline{x}^T = (\underline{x}_1^T, \cdots, \underline{x}_k^T)$, the relation between complete and incomplete data is given by $\underline{y} = H \cdot \underline{x}$, where H is the non-invertible $(n \times n \cdot k)$ matrix

$$H = \underbrace{[I|I|\cdots|I]}_{k \text{ times}}$$

This is a (simple) Linear-Gaussian case situation; using the results of section 2.3, the E and M steps of an EM algorithm for solving the least-squares problem of (3.12) are given by,

- E step:

$$\underline{x}_j^{(n+1)} = \underline{x}_j^{(n)} - \beta_j \left(\underline{y} - A \cdot \underline{\theta}^{(n)}\right), \quad j = 1, \cdots, k \tag{3.15}$$

- M step

$$\theta_j^{(n+1)} = \arg\min_{\theta_j} ||\underline{x}_j^{(n)} - \underline{a}_j \cdot \theta_j||^2 = \frac{\underline{a}_j^T \underline{x}_j^{(n)}}{||\underline{a}_j||^2}, \quad j = 1, \cdots, k \tag{3.16}$$

Combining these two steps we get the iteration,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} + diag\left(\frac{\beta_1}{||\underline{a}_1||^2}, \cdots, \frac{\beta_k}{||\underline{a}_k||^2}\right) \cdot A^T \cdot \left(\underline{y} - A \cdot \underline{\theta}^{(n)}\right) \tag{3.17}$$

where $diag(\cdot, \cdots, \cdot)$ is a diagonal matrix.

A sequential algorithm, based on the iteration (3.17), according to the exact EM mapping method, may now be easily developed. Define a "correlation matrix", $\mathcal{A}_n$, and a "cross-correlation vector", $\underline{p}_n$, for the least squares problem of order n in the following way,

$$\mathcal{A}_n = \frac{1}{n} A^T A \;=\; \frac{1}{n} \sum_{i=1}^{n} \underline{a}_i \underline{a}_i^T$$

$$\underline{p}_n = \frac{1}{n} A^T \underline{y} \;=\; \frac{1}{n} \sum_{i=1}^{n} \underline{\alpha}_i \cdot y_i \tag{3.18}$$

Given a new measurement. y_{n+1}, we can update $\mathcal{A}_n$ and $\underline{p}_n$ recursively, as,

$$\begin{aligned} \mathcal{A}_{n+1} &= \frac{n}{n+1} \mathcal{A}_n + \frac{1}{n+1} \underline{\alpha}_{n+1} \underline{\alpha}_{n+1}^T \\ \underline{p}_{n+1} &= \frac{n}{n+1} \underline{p}_n + \frac{1}{n+1} \underline{\alpha}_{n+1} \cdot y_{n+1} \end{aligned} \tag{3.19}$$

The exact EM iteration (3.17) may be written as,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} + diag\left(\frac{\beta_1}{\mathcal{A}_{n+1}(1,1)}, \ldots, \frac{\beta_k}{\mathcal{A}_{n+1}(k,k)}\right) \cdot \left(\underline{p}_{n+1} - \mathcal{A}_{n+1} \cdot \underline{\theta}^{(n)}\right) \tag{3.20}$$

which can be calculated recursively, since all required quantities are calculated recursively. The sequential least squares EM algorithm (SLSEM) is completely specified by (3.19) and (3.20).

As a final comment, we may compare this SLSEM algorithm to the well known LMS and RLS algorithms, which also solve this linear least squares problem recursively. The LMS algorithm is specified by the following recursion,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} - \beta \underline{\alpha}_{n+1} \left(y_{n+1} - \underline{\alpha}_{n+1}^T \underline{\theta}^{(n)}\right) \tag{3.21}$$

where $\underline{\theta}_k$ is the i^{th} component of the vector $\underline{\theta}$.

The RLS algorithm, which exactly solves in any step the n^{th} order least-squares problem is specified by,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n+1)} + \mathcal{A}_{n+1}^{-1} \underline{\alpha}_{n+1} \left(y_{n+1} - \underline{\alpha}_{n+1}^T \underline{\theta}^{(n)}\right) \tag{3.22}$$

where $\mathcal{A}_{n+1}^{-1}$ may be calculated recursively as,

$$\mathcal{A}_{n+1}^{-1} = \mathcal{A}_n^{-1} - \frac{\mathcal{A}_n^{-1} \underline{\alpha}_{n+1} \underline{\alpha}_{n+1}^T \mathcal{A}_n^{-1}}{1 + \underline{\alpha}_{n+1}^T \mathcal{A}_n^{-1} \underline{\alpha}_{n+1}} \tag{3.23}$$

We notice immediately that the complexity of the LMS algorithm in each iteration is linear in the number of unknowns k, while the complexity of the SLSEM algorithm and the RLS algorithm is quadratic, of order k^2. The SLSEM algorithm requires less computation, however.

A few experiments with these algorithms indicate that the convergence of the SLSEM is faster than that of the LMS algorithm. The convergence of the SLSEM algorithm is, of course, slower than that of the RLS algorithm since the RLS algorithm exactly solves the least-squares problem in each step. However, the convergence rates of the RLS and SLSEM algorithms to the true value of the parameters, as a function of the data index, are comparable.

3.1.2 Sequential EM algorithm based on recursive E and M steps

The sequential EM algorithm presented now is applicable to the following situation. Suppose that, given the complete data, there is a sequential (or adaptive) algorithm for estimating the parameters. Also, suppose that the required statistics of the complete data may be estimated recursively from the observations. In this case, as each new block is observed, the necessary new complete data statistics are estimated recursively, given the current value of the parameters - the E step. Then, the parameters values are updated sequentially, given the new estimated block of complete data statistics - the M step.

To be more specific, suppose that, if the complete data, $\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_n, \cdots$, is observed, there is a sequential (or adaptive) algorithm for estimating the parameters, i.e.

$$\underline{\theta}^{(n+1)} = G\left(\underline{t}(\underline{x}_{n+1}), \underline{r}(\underline{x}_n, \cdots, \underline{x}_1), \underline{\theta}^{(n)}\right) \tag{3.24}$$

where $\underline{t}(x_{n+1})$ are the statistics to be extracted from the new data x_{n+1}. The statistics

$\underline{r}(\underline{x}_n, \cdots, \underline{x}_1)$ are extracted from the past data. Since (3.24) represents a sequential procedure, these statistics are easily stored and updated.

Unfortunately, the complete data is not available. We only observe the incomplete data, $\underline{y}_1, \underline{y}_2, \cdots, \underline{y}_n, \cdots$, so we cannot estimate the parameters via (3.24). Following the EM idea, we should estimate the complete data statistics, needed for the sequential algorithm of (3.24), by taking their conditional expectation, given $\underline{y}_{n+1}, \cdots, \underline{y}_1$ and tne current value, $\underline{\theta}^{(n)}$, of the parameters. These conditional expectations may be calculated sequentially too. Consider, for example, the case where the complete data is composed of two Gaussian signals, say $s(n)$ and $w(n)$, and the observed data is the sum of these signal, i.e. $y(n) = s(n) + w(n)$. The conditional expectation of the complete data, given the observation, in a non-sequential EM algorithm requires a Wiener filter. However, this conditional expectation may be achieved sequentially, using a causal Wiener filter or a Kalman filter.

A sequential procedure for estimating the required complete data statistics may be described abstractly, e.g.,

$$E\left\{\underline{t}(\underline{x}_{n+1}) \,\middle|\, \underline{y}_{n+1}, \cdots, \underline{y}_1; \underline{\theta}^{(n)}\right\} = \underline{F}(\underline{y}_{n+1}, \underline{g}(\underline{y}_n, \cdots, \underline{y}_1); \underline{\theta}^{(n)}) \tag{3.25}$$

where the functions $\underline{g}$, summarizing the contribution of past observations, are easily stored and calculated.

A sequential (adaptive) EM algorithm, based on recursive E and M steps is presented formally as follows.

- Start, $n = 0$: Guess $\underline{\theta}^{(0)}$. Initial $\underline{g}(\cdot, \cdots) = 0$

- For each new data block $\underline{y}_{n+1}$,

- **Sequential E-step:** calculate

$$\begin{aligned} \hat{\underline{t}} = E\left\{\underline{t}(\underline{x}_{n+1}) \,\Big|\, /\underline{y}_{n+1}, \cdots, \underline{y}_1; \underline{\theta}^{(n)}\right\} &= \underline{F}_t(\underline{y}_{n+1}, \underline{g}_t(\underline{y}_n, \cdots, \underline{y}_1); \underline{\theta}^{(n)}) \\ \hat{\underline{r}} = E\left\{\underline{r}(\underline{x}_n, \cdots, \underline{x}_1) \,\Big|\, \underline{y}_n, \cdots, \underline{y}_1; \underline{\theta}^{(n)}\right\} &= \underline{F}_r(\underline{g}_r(\underline{y}_n, \cdots, \underline{y}_1); \underline{\theta}^{(n)}) \qquad (3.26) \end{aligned}$$

- **Sequential M-step:** Update parameters

$$\underline{\theta}^{(n+1)} = G(\hat{\underline{t}}, \hat{\underline{r}}, \underline{\theta}^{(n)}) \qquad (3.27)$$

- Update and record $\underline{g}_t(\underline{y}_{n+1}, \cdots, \underline{y}_1), \underline{g}_r(\underline{y}_{n+1}, \cdots, \underline{y}_1)$ for the next step

- $n = n + 1$

Suppose that the recursive procedure (3.24), suggested when the complete data is given, increases the likelihood of the complete data $L^c_{n+1}(\underline{\theta})$, i.e. it satisfies,

$$L^c_{n+1}(\underline{\theta}^{(n+1)}) \geq L^c_{n+1}(\underline{\theta}^{(n)}) \qquad (3.28)$$

In this case, it is easy to show that the recursive algorithm suggested by (3.26) and (3.27) increases the current likelihood of the observations, as follows. This is true because the function $Q_{n+1}(\underline{\theta}; \underline{\theta}^{(n)})$ has the same functional form as $L^c_{n+1}(\underline{\theta})$, where the estimated statistics, $\hat{\underline{t}}$ and $\hat{\underline{r}}$, are substituted in place of the statistics, $\underline{t}$ and $\underline{r}$. Thus, if (3.28) is true, then (3.27) implies

$$Q_{n+1}(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) \geq Q_{n+1}(\underline{\theta}^{(n)}; \underline{\theta}^{(n)}) \qquad (3.29)$$

Using (3.4), (3.29) and Jensen's inequality we get,

$$L_{n+1}(\underline{\theta}^{(n+1)}) \geq L_{n+1}(\underline{\theta}^{(n)}) \qquad (3.30)$$

3.2 Sequential EM based on stochastic approximation

The sequential algorithms presented so far were suggested by assuming that the underlying problem had a special structure. In this section, we will address the general situation. Unfortunately, sequential algorithms may not be derived directly from the EM algorithm in the general case. We will therefore suggest algorithms, that approximate the EM iteration, in order to get a recursive implementation. We will be able to show that these algorithms belong to the class of stochastic approximation algorithms, for which a general theory is readily available.

3.2.1 The EM algorithm: General sequential considerations

The log-likelihood of the observations, given $n + 1$ data blocks, is given by (3.3). Define,

$$L_{n+1/n}(\underline{\theta}) = \log f_{Y_{n+1}/Y_n \cdots Y_1}(\underline{y}_{n+1}/\underline{y}_n, \cdots, \underline{y}_1; \underline{\theta}) \tag{3.31}$$

The log-likelihood of the observations may be written recursively as,

$$L_{n+1}(\underline{\theta}) = L_n(\underline{\theta}) + L_{n+1/n}(\underline{\theta}) \tag{3.32}$$

or as,

$$L_{n+1}(\underline{\theta}) = L_1(\underline{\theta}) + \sum_{i=1}^{n} L_{i+1/i}(\underline{\theta}) \tag{3.33}$$

In order to develop a recursive algorithm, we refer to the recursive formula for the log-likelihood (3.32). Analogous to (3.4), the term L_n may be written as,

$$L_n(\underline{\theta}) = Q_n(\underline{\theta}; \underline{\theta}') - H_n(\underline{\theta}; \underline{\theta}') \tag{3.34}$$

where the complete data is defined to be $\underline{x}_1, \cdots, \underline{x}_n$. For the term $L_{n+1/n}$, the complete data is $\underline{x}_{n+1}$ and following the same considerations which lead to (3.34), we may write,

$$L_{n+1/n}(\underline{\theta}) = Q_{n+1/n}(\underline{\theta}; \underline{\theta}') - H_{n+1/n}(\underline{\theta}; \underline{\theta}') \tag{3.35}$$

where

$$Q_{n+1/n}(\underline{\theta},\underline{\theta}') = E\left\{\log f_{X_{n+1}/Y_n \cdots Y_1}(\underline{x}_{n+1}/\underline{y}_n,\cdots,\underline{y}_1;\underline{\theta}) \,\Big|\, \underline{y}_{n+1},\cdots,\underline{y}_1;\underline{\theta}'\right\} \tag{3.36}$$

$$H_{n+1/n}(\underline{\theta},\underline{\theta}') = E\left\{\log f_{X_{n+1}/Y_{n+1} \cdots Y_1}(\underline{x}_{n+1}/\underline{y}_{n+1},\cdots,\underline{y}_1;\underline{\theta}) \,\Big|\, \underline{y}_{n+1},\cdots,\underline{y}_1;\underline{\theta}'\right\} \tag{3.37}$$

Therefore,

$$L_{n+1}(\underline{\theta}) = L_n(\underline{\theta}) + L_{n+1/n}(\underline{\theta}) = Q_n(\underline{\theta};\underline{\theta}') + Q_{n+1/n}(\underline{\theta};\underline{\theta}') - \left[H_n(\underline{\theta};\underline{\theta}') + H_{n+1/n}(\underline{\theta};\underline{\theta}')\right] \tag{3.38}$$

and we have,

$$H_n(\underline{\theta};\underline{\theta}') \leq H_n(\underline{\theta}';\underline{\theta}') \quad \text{and} \quad H_{n+1/n}(\underline{\theta};\underline{\theta}') \leq H_{n+1/n}(\underline{\theta}';\underline{\theta}') \tag{3.39}$$

One could try to achieve a recursive algorithm by maximizing either,

$$Q_n(\underline{\theta};\underline{\theta}^{(n)}) + Q_{n+1/n}(\underline{\theta};\underline{\theta}^{(n)}) \tag{3.40}$$

or,

$$Q_1(\underline{\theta};\underline{\theta}^{(n)}) + Q_{2/1}(\underline{\theta};\underline{\theta}^{(n)}) + \cdots + Q_{n+1/n}(\underline{\theta};\underline{\theta}^{(n)}) \tag{3.41}$$

since maximizing either (3.40) or (3.41) will generate a new value $\underline{\theta}^{(n+1)}$ that increases the likelihood $L_{n+1}(\underline{\theta})$. However, despite their seemingly recursive structure, these maximizations cannot be performed sequentially in general, because:

- Calculating $Q_{n+1/n}$ involves the past data $\underline{y}_n,\cdots,\underline{y}_1$

- For each new parameter value, the conditional expectations needed for the term Q_n, or the terms $Q_1, Q_{2/1},\cdots,Q_{n/n-1}$, should be recalculated. This requires using the past data samples.

An approximate sequential algorithm

From the discussion above we conclude that a general sequential algorithm, that will implement the desired maximizations of (3.40) or (3.41), cannot be specified. However, consider the following sequential algorithm,

```
• Start, $n = 0$ : Initialize $\Psi_0(\underline{\theta}) = 0$. Guess $\underline{\theta}^{(0)}$

• For each new data $\underline{y}_{n+1}$,

    – E-step: calculate

        $Q^a_{n+1/n}(\underline{\theta},\underline{\theta}^{(n)}) = E\left\{\log f_X(\underline{x}_{n+1};\underline{\theta}) \mid \underline{y}_{n+1},\underline{y}_n,\cdots,\underline{y}_{n-m};\underline{\theta}^{(n)}\right\}$        (3.42)

    – M-step: solve

        $\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}\in\Theta} \left[Q^a_{n+1/n}(\underline{\theta},\underline{\theta}^{(n)}) - \beta_n \cdot \Psi_n(\underline{\theta})\right]$        (3.43)

    – Record for next step

        $\Psi_{n+1}(\underline{\theta}) = Q^a_{n+1/n}(\underline{\theta},\underline{\theta}^{(n)}) + \beta_n \cdot \Psi_n(\underline{\theta})$

    – $n = n + 1$
```

This algorithm approximates the desired procedures as follows. First, the term $Q_{n+1/n}(\underline{\theta};\underline{\theta}^{(n)})$ is approximated by $Q^a_{n+1/n}(\underline{\theta};\underline{\theta}^{(n)})$, given by (3.42). We will use in this approximation some past data values, $\underline{y}_n,\cdots,\underline{y}_{n-m}$, as long as $Q^a_{n+1/n}$ is calculated recursively. We note that, if the different observation blocks are independent, $Q_{n+1/n} = Q^a_{n+1/n}$. In general, the weaker the successive observations blocks are correlated, the better this approximation becomes. Second, the previous terms are not recalculated. We calculate each $Q^a_{i+1/i}$, using the corresponding parameter value, $\underline{\theta}^{(i)}$, and we simply accumulate these functions and generate

$\Psi_n(\underline{\theta})$ recursively. Also using this algorithm, the previous terms may be weighted, according to the choice of β_n. By an appropriate choice, we may reduce the contribution of the past data and track varying parameters in the adaptive situation. or we may weight the past data more heavily, to guarantee convergence and consistency, for a sequential algorithm.

Although it seems that this algorithm is based on ad-hoc approximations, we will be able to show that this algorithm belongs to the class of stochastic apprcximation algorithms. Later in the chapter, we will use results developed for stochastic approximation algorithms to calculate the asymptotic distribution of the estimator and prove its consistency. However, before that, we will briefly present the stochastic approximation idea.

3.2.2 Stochastic approximation

In a typical problem for stochastic approximation, a sequence of random variables (vectors), $\{\underline{y}_n\}$, is observed. We assume that the sequence is stationary, in the sense that each $\underline{y}_n$ has the same marginal distribution, and that it is ergodic. At each instance, a function of the observed data and the desired parameters, $L(\underline{\theta}; \underline{y}_n)$, is given. We want either to optimize the unavailable ensemble average of L, i.e. to find

$$\max_{\underline{\theta}} E_{\underline{y}}\left\{L(\underline{\theta}; \underline{y}_n)\right\} = \max_{\underline{\theta}} \bar{L}(\underline{\theta}) \tag{3.44}$$

or to solve an equation, that involves this unavailable expected value, e.g.

$$E_{\underline{y}}\left\{L(\underline{\theta}; \underline{y}_n)\right\} = 0 \qquad \text{or } \bar{L}(\underline{\theta}) = 0 \tag{3.45}$$

The first problem is sometimes referred to as the Kiefer-Wolfowitz (K-W) problem [40]. The second problem is referred to as the Robbins-Monro (R-M) problem [41]. By defining

$$L'(\underline{\theta}; \underline{y}_n) = \frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}; \underline{y}_n) \tag{3.46}$$

a K-W problem for L may be reduced to a R-M problem for L'.

Suppose we have an iterative algorithm for optimizing $\bar{L}(\underline{\theta})$ of (3.44) or for solving the (non-linear) equation (3.45). For example, the gradient iterative algorithm for maximizing (3.44) will be,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} + \beta \cdot \frac{\partial}{\partial \underline{\theta}} \bar{L}(\underline{\theta}^{(n)}) = \underline{\theta}^{(n)} + \beta \cdot E_{\underline{\mathbf{y}}} \left\{ \frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}^{(k)}) \right\} \tag{3.47}$$

We cannot implement this iterative algorithm, since the expected value of L and its derivatives are not available. The stochastic approximation idea is to approximate the expected value by the sample value. Since we have an ergodic sequence of realizations, $\{\underline{y}_n\}$, the next iteration is performed using the next realization. This achieves time-average that approximates the unavailable ensemble average values.

Specifically, the stochastic approximation of the gradient algorithm, referred to as the stochastic gradient algorithm, is given by,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} + \beta_n \frac{\partial}{\partial \underline{\theta}} L(\underline{\theta}^{(n)}; \underline{y}_n) \tag{3.48}$$

In [42] and [43] it is shown that, if the observations sequence is ergodic and if $\{\beta_n\}$ is a sequence of positive numbers such that $\lim_{n \to \infty} \beta_n = 0$ and it satisfies,

$$\sum_{n=0}^{\infty} \beta_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} \beta_n^2 < \infty \tag{3.49}$$

e.g. $\beta_n = \beta/n$, then the stochastic gradient algorithm converges, in probability 1 and in the mean-square sense, to the right solution of (3.44).

We note that, if the observed data is not stationary and we are looking for an adaptive algorithm, then we usually choose constant gain β in some range, instead of $\{\beta_n\}$ as in (3.49). This way, we reduce the weight of past observations and use the new input to track varying parameter values.

3.2.3 The EM stochastic approximation algorithm

The best statistical parameter estimation method, which we can hope to find, is a method that solves the following optimization problem:

$$\hat{\underline{\theta}} = \arg\max_{\underline{\theta}} E_{\underline{y}_{n+1}} \left\{ \log f_{Y_{n+1}}(\underline{y}_{n+1};\underline{\theta}) \right\} = \arg\max_{\underline{\theta}} J(\underline{\theta}) \tag{3.50}$$

This is because the solution to this problem is the *true* parameter value. To prove this claim, we note that using Jensen's inequality, we get

$$J(\underline{\theta}) = \int [\log f_{Y_{n+1}}(\underline{y}_{n+1};\underline{\theta})] f_{Y_{n+1}}(\underline{y}_{n+1};\underline{\theta}_{true}) d\underline{y} \leq J(\underline{\theta}_{true}) \tag{3.51}$$

i.e. $\underline{\theta}_{true}$ maximizes $J(\underline{\theta})$. We note that the equality in (3.50) is achieved, if and only if $f_{Y_{n+1}}(\underline{y}_{n+1};\underline{\theta}) = f_{Y_{n+1}}(\underline{y}_{n+1};\underline{\theta}_{true})$ almost everywhere.

The maximization of $J(\underline{\theta})$ can be accomplished using Newton-Raphson method or any other optimization method. Instead, we will use an iterative algorithm for maximizing $J(\underline{\theta})$, based on the considerations leading to the EM algorithm, as follows.

Using x_{n+1} as the complete data with respect to y_{n+1} and following the method used for deriving (2.12), we may write $J(\underline{\theta})$ as,

$$J(\underline{\theta}) = \bar{Q}(\underline{\theta};\underline{\theta}') - \bar{H}(\underline{\theta};\underline{\theta}') \tag{3.52}$$

where

$$\bar{Q}(\underline{\theta};\underline{\theta}') = E_{\underline{y}_{n+1}} \left\{ E \left\{ \log f_{X_{n+1}}(\underline{x}_{n+1};\underline{\theta}) \,\middle|\, \underline{y}_{n+1};\underline{\theta}' \right\} \right\} \tag{3.53}$$

$$\bar{H}(\underline{\theta};\underline{\theta}') = E_{\underline{y}_{n+1}} \left\{ E \left\{ \log f_{X_{n+1}/Y_{n+1}}(\underline{x}_{n+1}/\underline{y}_{n+1};\underline{\theta}) \,\middle|\, \underline{y}_{n+1};\underline{\theta}' \right\} \right\} \tag{3.54}$$

Considering the function $\bar{H}(\underline{\theta};\underline{\theta}^{(n)})$, it is easy to show, using Jensen's inequality for the expression inside the expectations, that,

$$\bar{H}(\underline{\theta};\underline{\theta}') \leq \bar{H}(\underline{\theta}';\underline{\theta}') \tag{3.55}$$

Analogous to the EM algorithm, an iterative algorithm for maximizing $J(\underline{\theta})$ is given by,

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} \bar{Q}(\underline{\theta}; \underline{\theta}^{(n)}) = \arg\max_{\underline{\theta}} E_{\underline{y}_{n+1}} \left\{ E\left\{ \log f_{X_{n+1}}(\underline{x}_{n+1}; \underline{\theta}) \,\middle|\, \underline{y}_{n+1}; \underline{\theta}^{(n)} \right\} \right\} \quad (3.56)$$

where from (3.54) and (3.55), each iteration of (3.56) increases $J(\underline{\theta})$.

Unfortunately, it is impossible to implement (3.56), since the expected value with respect to $\underline{y}_{n+1}$ is not available. Using the stochastic approximation idea, a stochastic realization of (3.56) is performed as the $(n + 1)^{st}$ data block is observed. Thus, we get the following stochastic approximation algorithm:

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} E\left\{ \log f_{X_{n+1}}(\underline{x}_{n+1}; \underline{\theta}) \,\middle|\, \underline{y}_{n+1}; \underline{\theta}^{(n)} \right\} \quad (3.57)$$

Following the notation used in the beginning of this section we will define

$$Q^{a}_{n-1/n}(\underline{\theta}; \underline{\theta}^{(n)}) = E\left\{ \log f_{X_{n+1}}(\underline{x}_{n-1}; \underline{\theta}) \,\middle|\, \underline{y}_{n+1}; \underline{\theta}^{(n)} \right\} \quad (3.58)$$

In a sequential algorithm, the new data block together with the past data blocks should provide a time-average approximation to the ensemble average of (3.56). This may require weighting the past data more heavily. Thus, we define recursively a function $\Psi_{n+1}(\underline{\theta})$ as,

$$\Psi_{n+1}(\underline{\theta}) = Q^{a}_{n+1/n}(\underline{\theta}, \underline{\theta}^{(n)}) + \beta_n \cdot \Psi_n(\underline{\theta}) \quad (3.59)$$

and the general stochastic EM step will be

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} \Psi_{n+1}(\underline{\theta}) \quad (3.60)$$

which is the algorithm suggested in (3.43).

We note that this algorithm was also suggested in [44], during the investigation of approximations to the stochastic Newton algorithm.

3.3 Some properties of the sequential EM algorithms

Analysis of sequential and adaptive procedures, especially in a statistical or stochastic context, has been the subject of extensive research efforts. The properties of the stochastic approximation method, being the simplest, may be found in various references, e.g. [43,45,46,47] and elsewhere. This topic is probably one of the most difficult subjects in mathematical statistics; investigating convergence of complicated stochastic structures, proving convergence in probability, in probability 1 or in the mean-squares sense and finding the rate of convergence requires using advanced probabilistic tools from Martingale theory and stochastic calculus theory. Thus, a typical assumption made in most of the references above, in order to simplify the analysis, is that the observed data blocks are independent.

The analysis of the sequential EM algorithms for the stationary case, presented below, is far from complete. Nevertheless, the following results were achieved:

- *General asymptotic consistency:* We will show that the estimator, generated by a sequential EM algorithm. is asymptotically consistent, when the ML estimator is consistent and the sequential EM iteration converges to a stationary point.

- *Limit distribution:* The limit distribution of the estimator will be given for some sequential EM algorithms. These results are for independent observations, however.

The properties of the sequential EM algorithms should be investigated further. Detailed analysis may require the use of more advanced mathematical tools. It is an interesting research topic in mathematical statistics. The book by Kushner [47], together with the EM ideas and the preliminary analysis of the sequential EM algorithms, presented in this chapter, should provide the starting point for this research.

3.3.1 General asymptotic consistency

The analysis of sequential algorithms when the observations are not independent may, in general, be quite complicated. However, some sequential EM algorithms have the property that in the limit the convergence point is a stationary point of the corresponding likelihood limit. Using this property we will prove the asymptotic consistency of these algorithms as follows. The sequence of normalized log-likelihood functions at each instance, $\frac{1}{n}L_n(\underline{\theta})$, are shown to converge in probability 1 to a limit, $l(\underline{\theta})$, whose unique maximum is the true parameter value, $\underline{\theta}_{true}$. Under regularity conditions, the derivative of the likelihood also converges. Since the sequence of sequential EM estimates converges to a stationary point of the likelihood, i.e. to zero derivative point, it converges to a zero derivative point of $l(\underline{\theta})$ which is its maximum, i.e. the true parameter value.

Specifically, as discussed in Appendix B, for a class of ergodic sources, which include, for example, all finite Markov sources, the sequence $\frac{1}{n}L_n(\underline{\theta})$ where $L_n(\underline{\theta})$ is given in (3.3), converges uniformly in probability 1 to,

$$l(\underline{\theta}) = \int f_{Y_n,Y_{n-1}\cdots}(y_n, y_{n-1}\cdots;\underline{\theta}_{true}) \log f_{Y_n/Y_{n-1},\cdots}(y_n/y_{n-1},\cdots;\underline{\theta}) dy_n dy_{n-1}\cdots dy_1 \tag{3.61}$$

Intuitively, the sources that belong to this class are ergodic sources, whose memory fades fast enough. This result is also discussed in [48].

The function $l(\underline{\theta})$ achieves its maximum at $\underline{\theta} = \underline{\theta}_{true}$. Under regularity conditions and the convexity of (3.61), $\underline{\theta}_{true}$ is the unique solution to the equation $Dl(\underline{\theta}) = 0$. Now, using this fact and some well known results from analysis, the following theorem may be proved,

Theorem 3.1 *Let the observations $\underline{y}_1, \cdots, \underline{y}_n \cdots$ be generated by an ergodic source for which*

(3.61) holds. Let $\{\underline{\theta}^{(n)}\}$ be an instance of a sequential EM algorithm such that for any realization of the observations,
(i) the sequence of estimates $\{\underline{\theta}^{(n)})\}$ converges to a limit $\underline{\theta}^{*}$
(ii) $\lim_{n\to\infty} D^{10}Q_{n+1}(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) = 0$
Then, in probability 1, as $n \to \infty$, $\underline{\theta}^{(n)} \to \underline{\theta}_{true}$.

The proof of this theorem is also given in appendix B.

The conditions of theorem 3.1 may be verified for many specific sequential EM algorithms. For example the sequential EM algorithm with exact EM mapping satisfy the condition (*ii*) above, for all n; thus, if the observation sequence is ergodic and satisfies (3.61), whenever the algorithm converges, it generates a consistent estimator.

3.3.2 Limit distribution

The asymptotic distribution of several sequential EM algorithms can be calculated using the following technique. The recursion defined by these algorithms will be approximated by a recursion that resembles stochastic approximation algorithms, especially the stochastic Newton method. Having this similarity, we will be able to invoke results developed in the stochastic approximation context and show, in some cases, that in the limit the estimator is distributed Normally around the true parameters value and has $\sqrt{n}$ consistency, i.e. its variance tends to zero as $1/\sqrt{n}$. We note that the possible connection between the sequential EM algorithm and the stochastic Newton method was pointed out in [44].

Consider, for example, the stochastic EM algorithm defined by the recursion (3.60), repeated here,

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} \Psi_{n+1}(\underline{\theta})$$

where Ψ_{n+1} is defined recursively by

$$\Psi_{n+1}(\underline{\theta}) = Q^{a}_{n+1/n}(\underline{\theta}, \underline{\theta}^{(n)}) + \beta_n \cdot \Psi_n(\underline{\theta})$$

and

$$Q^a_{n+1/n}(\underline{\theta},\underline{\theta}^{(n)}) = E\left\{\log f_X(\underline{x}_{n+1};\underline{\theta})/\underline{y}_{n+1});\underline{\theta}^{(n)}\right\} = Q(\underline{\theta},\underline{\theta}^{(n)};\underline{y}_{n+1})$$

Note that by the construction of the stochastic EM algorithm, $\underline{\theta}^{(n)}$ maximizes $\Psi_n(\underline{\theta})$. We will assume regularity conditions, that allow certain operations, e.g. differentiation under the integral sign.

An approximation to the stochastic EM recursion may be achieved if, instead of solving the maximization problem of (3.60), a Newton-Raphson step, starting in $\underline{\theta}^{(n)}$, is performed. The resulting recursion is

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} - [D^2\Psi_{n+1}(\underline{\theta}^{(n)})]^{-1} \cdot D\Psi_{n+1}(\underline{\theta}^{(n)}) \tag{3.62}$$

The gradient vector, $D\Psi_{n+1}$, and the Hessian matrix, $D^2\Psi_{n+1}$, are also given recursively,

$$D\Psi_{n+1}(\underline{\theta}^{(n)}) = D^{10}Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)};\underline{y}_{n+1}) + \beta_n \cdot D\Psi_n(\underline{\theta}^{(n)}) \tag{3.63}$$

$$D^2\Psi_{n+1}(\underline{\theta}^{(n)}) = D^{20}Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)};\underline{y}_{n+1}) + \beta_n \cdot D^2\Psi_n(\underline{\theta}^{(n)}) \tag{3.64}$$

However, $D\Psi_n(\underline{\theta}^{(n)}) = 0$ since $\underline{\theta}^{(n)}$ maximizes Ψ_n. Also, from (2.36)

$$D^{10}Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)};\underline{y}_{n+1}) = D\log f_Y(\underline{y}_{n+1};\underline{\theta}^{(n)}) = S(\underline{y}_{n+1};\underline{\theta}^{(n)})$$

For exponential families, the second derivative of Q is such that $-D^2Q$ is the Fisher information matrix of the complete data, I_X, calculated at $\underline{\theta}^{(n)}$. Thus,

$$D^2\Psi_{n+1}(\underline{\theta}^{(n)}) = -I_X(\underline{\theta}^{(n)})+\beta_n\cdot D^2\Psi_n(\underline{\theta}^{(n)}) = -I_X(\underline{\theta}^{(n)})+\beta_n\left(-I_X(\underline{\theta}^{(n)}) + \beta_{n-1}\cdot\Psi_{n-1}(\underline{\theta}^{(n)})\right) \tag{3.65}$$

and so on. If $\beta_n \equiv 1$ then $D^2\Psi_{n+1}(\underline{\theta}^{(n)}) = -(n-1)I_X(\underline{\theta}^{(n)})$. In this case, from (3.62), the stochastic EM iteration is approximated by,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} + \frac{1}{n+1} I_X^{-1}(\underline{\theta}^{(n)}) \cdot S(\underline{y}_{n+1};\underline{\theta}^{(n)}) \tag{3.66}$$

Recursions like (3.66) are typical in stochastic approximation algorithms. For example, replacing $I_X(\underline{\theta}^{(n)})$ by $I_Y(\underline{\theta}^{(n)}) = D^2L(\underline{\theta}^{(n)})$ in (3.66), yields the stochastic Newton method. It has been shown, e.g. in [49] and [46] that the recursion (3.66) generates an estimator, $\underline{\theta}^{(n)}$, which satisfies, under regularity conditions and provided that $2I_Y(\underline{\theta}_{true}) > I_X(\underline{\theta}_{true})$,

$$\sqrt{n}(\underline{\theta}^{(n)} - \underline{\theta}_{true}) \longrightarrow N\left(0, I_X^{-2}I_Y(2I_YI_X^{-1} - 1)^{-1}\right) \tag{3.67}$$

in distribution, as $n \to \infty$. The Fisher information matrices I_X, I_Y in (3.67) are evaluated at $\underline{\theta}_{true}$.

When $2I_Y(\underline{\theta}_{true}) \leq I_X(\underline{\theta}_{true})$, (3.67) above does not hold, (although the stochastic EM algorithm may still yield a consistent estimator). However, if we choose the coefficients β_n in such a way that the stochastic EM algorithm is approximated by

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} + \frac{1}{n^{(1+\alpha)/2}} I_X^{-1}(\underline{\theta}^{(n)}) \cdot S(\underline{y}_{n+1}; \underline{\theta}^{(n)}) \tag{3.68}$$

and $0 < \alpha < 2I_Y(\underline{\theta}_{true})I_X^{-1}(\underline{\theta}_{true}) < 1$, then according to [46],

$$n^{\alpha/2}(\underline{\theta}^{(n)} - \underline{\theta}_{true}) \longrightarrow N\left(0, I_X^{-2}I_Y(2I_YI_X^{-1} - \alpha)^{-1}\right) \tag{3.69}$$

in distribution, as $n \to \infty$, and the asymptotic Normality and $\sqrt{n}$ consistency hold.

A similar derivation using a Newton-Raphson approximation can be performed for the sequential EM algorithm with exact EM mapping. This algorithm generates estimates according to the mapping (3.7), that is

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} Q_{n+1}(\underline{\theta}, \underline{\theta}^{(n)}; \underline{y}_{n+1}, \cdots, \underline{y}_1)$$

which may be approximated, using a step of Newton-Raphson algorithm by,

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} - [D^2Q_{n+1}(\underline{\theta}^{(n)}, \underline{\theta}^{(n)}; \underline{y}_{n+1}, \cdots)]^{-1} \cdot DQ_{n+1}(\underline{\theta}^{(n)}, \underline{\theta}^{(n)}; \underline{y}_{n+1}, \cdots) \tag{3.70}$$

Using the fact that $DQ_n(\underline{\theta}^{(n)}, \underline{\theta}^{(n)}; \underline{y}_n, \cdots) = 0$ we may write

$$DQ_{n+1}(\underline{\theta}^{(n)}, \underline{\theta}^{(n)}; \underline{y}_{n+1}, \cdots) = DL_{n+1/n}(\underline{\theta}^{(n)}) = S(\underline{y}_{n+1}/\underline{y}_n \cdots ; \underline{\theta}^{(n)}) \tag{3.71}$$

For exponential families, the second derivative of Q will provide the Fisher information, i.e.

$$D^2 Q_{n+1}(\underline{\theta}^{(n)}, \underline{\theta}^{(n)}; \underline{y}_{n+1}, \cdots) = -I_{X_{n+1}X_n \cdots X_1}(\underline{\theta}^{(n)}) \tag{3.72}$$

Thus the approximation (3.70) may be written as

$$\underline{\theta}^{(n+1)} = \underline{\theta}^{(n)} + I^{-1}_{X_{n+1}X_n \cdots X_1}(\underline{\theta}^{(n)}) \cdot S(\underline{y}_{n+1}/\underline{y}_n \cdots ; \underline{\theta}^{(n)}) \tag{3.73}$$

For the case where the observations are coming from a finite markov source, the Fisher information is written recursively as a sum of identical conditional information matrices. We may again use the results of Sacks [49] and Fabian [46] to get Normality and $\sqrt{n}$ consistency of the estimator, as in (3.67), where the conditional information matrices $I_{X_{n+1}/X_n \cdots}$ and $I_{Y_{n+1}/Y_n \cdots}$ replace the matrices I_X and I_Y.

Chapter 4

Parameter Estimation of Superimposed Signals

In this chapter, we will consider the problem of estimating parameters of superimposed signals observed in noise, which occurs in a wide range of signal processing applications. This problem will be approached in this chapter using the iterative EM method, presented in the previous chapters. In the next chapter, we will apply the proposed iterative method to another important signal processing problem, namely, the multiple microphone noise cancellation problem.

A specific example of the applications, that are considered in this chapter, is the multiple source location estimation problem, using an array of sensors. In this problem, we have K sources radiating signals towards an array of M sensors, as illustrated in Figure 4.1. The location of the sensors is known, and we want to use the relative time delay between the observed signals from the different sensors to estimate the location of the sources. The signals, received by the array sensors due to the k^{th} source, may be represented as the

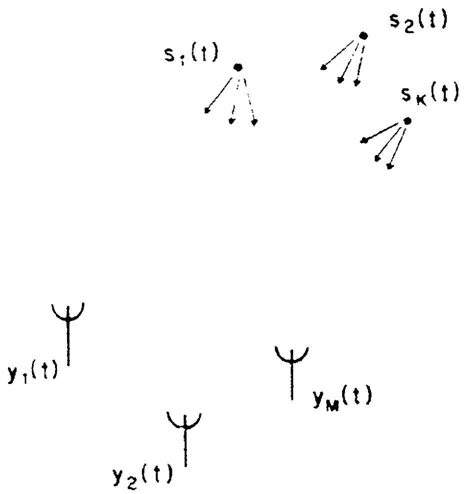


Figure 4.1: Array-Source Geometry

vector signal, $\underline{s}_k(t)$, where the m^{th} component of this vector is the signal received by the m^{th} sensor. The vector signal, $\underline{s}_k(t)$, is dependent upon a vector of unknown parameters, $\underline{\theta}_k$, associated with the k^{th} source and is denoted $\underline{s}_k(t;\underline{\theta}_k)$. In our problem, $\underline{\theta}_k$ is the vector of unknown source location parameters. We will denote by the vector $\underline{y}(t)$ the total signals observed by the array sensors. This observed signal vector is a result of superimposing the various $\underline{s}_k(t;\underline{\theta}_k)$ and an additive noise vector, i.e.

$$\underline{y}(t) = \sum_{k=1}^{K} \underline{s}_k(t;\underline{\theta}_k) + \underline{n}(t) \tag{4.1}$$

Our problem is to estimate the location parameters, given the observations, $\underline{y}(t)$.

The general problem of interest in this chapter is characterized by the model (4.1). The basic structure of (4.1) applies to a wide range of signal signal processing problems, in addition to the multiple source location estimation problem. Consider, for example, the problem of multi-echo time delay estimation. In this case each signal component is the scalar $s_k(t;\underline{\theta}_k)$, representing the k^{th} echo signal, and the parameters $\underline{\theta}_k$ are the time delay and attenuation of the k^{th} echo. Another example is frequency estimation of multiple sinusoids in noise, where the unknown parameters, $\underline{\theta}_k$, are the amplitude and the frequency of the k^{th} sinusoid.

Using the mathematical model of (4.1) together with a stochastic model for the various signals, one can formulate a statistical maximum likelihood problem for the underlying real problem. In many cases, the direct solution of this ML problem is complicated. Using the EM algorithm, we will develop in this chapter computationally efficient schemes for the joint estimation of $\underline{\theta}_1, \underline{\theta}_2, \cdots, \underline{\theta}_K$. The idea is to decompose the observed signal, $\underline{y}(t)$, into its components, and then to estimate the parameters of each signal component separately. Stating this idea in the terminology of the EM algorithm, we choose the complete data to be the contribution of each signal component separately. Thus the algorithm iterates between decomposing the observed data, i.e. estimating the complete data using the current parameter estimates, (the E step), and updating the parameter estimates, having the decomposed signals, (the M step).

So far the superimposed signals problem has been stated in its most general form. In different applications, additional specific modeling assumptions are needed. In a large variety of problems, we may assume that the noise signals, $\underline{n}(t)$, are sample functions from a stationary Gaussian process with a given spectrum. However, the modeling of the signal components in (4.1) varies according to our a-priori knowledge and the nature of the underlying real problem. Generally, these signals may be deterministic or stochastic, and various constraints may be applied on their waveforms or on their power spectra.

The problem of parameter estimation of superimposed signals in noise and its solution via the EM algorithm is presented in this chapter, in a variety of situations, as follows. In section 4.1, we will present the statistical ML problem and its EM solution procedures in the deterministic signals case. In section 4.2, a similar presentation is given for the stochastic (Gaussian) signals case. These procedures are then used, in sections 4.3 and 4.4, to solve the multipath time delay estimation problem and the passive multiple source location estimation problem. We will conclude this chapter, in section 4.5, by presenting sequential and adaptive algorithms, and applying these algorithms to the problem of estimating the frequencies of multiple sinusoids in noise.

4.1 Parameter estimation of superimposed signals: The deterministic case

The signal components, $\underline{s}_k(t;\underline{\theta}_k)$, are naturally modeled as deterministic signals in a variety of applications. Consider for example an active radar or sonar environment, where a known waveform pulse is transmitted. We observe the echoes of this pulse returning from several targets. Assuming perfect propagation conditions, the observed signal is a result of superimposing deterministic signals (the pulse echoes), which are known up to some parameters (e.g. the time delay). A statistical problem for estimating the unknown parameters of the superimposed signals is achieved in this case, when a stochastic model for the noise components of (4.1) is assumed.

In this section, we will present this statistical maximum likelihood problem and show that its direct solution is complicated, even in the simplest case, when we assume that the noise is white. Thus. we will develop methods based on the EM algorithm to solve this problem.

4.1.1 The ML problem

Consider the model of (4.1) under the following assumptions:

- The signal vectors, $\underline{s}_k(t;\underline{\theta}_k)$ $k = 1,\cdots,K$, are conditionally known up to a vector of parameters, $\underline{\theta}_k$.

- The $\underline{n}(t)$ are vector zero-mean white Gaussian processes whose covariance matrix is

$$E\{\underline{n}(t)\underline{n}(\sigma)\} = Q\cdot\delta(t-\sigma)$$

  where Q is a positive definite constant matrix and $\delta(\cdot)$ is the impulse function.

- The signals are observed over a finite duration, say $T_i \le t \le T_f$.

Under these assumptions, the log-likelihood function is given by,

$$\log f_{\underline{Y}}(\underline{y};\underline{\theta}) = C - \frac{\lambda}{2}\int_{T_i}^{T_f}\left[\underline{y}(t) - \sum_{k=1}^{K}\underline{s}_k(t;\underline{\theta}_k)\right]^{\dagger} Q^{-1}\left[\underline{y}(t) - \sum_{k=1}^{K}\underline{s}_k(t;\underline{\theta}_k)\right] dt \tag{4.2}$$

where $\dagger$ denotes the conjugate transpose operator. $\lambda = 1$, if $\underline{n}(t)$ is real valued, $\lambda = 2$, if $\underline{n}(t)$ is complex valued. C is a normalization constant independent of $\underline{\theta}$. This result is just a straightforward multi-channel extension of the known (deterministic) signal in white noise problem ([1], chap. 4). If the observed signal is discrete i.e. we observe $\underline{y}(t_i)$, $i = 1,\cdots,N$ the log-likelihood is still given by (4.2), where the integral over t is replaced by the sum over the t_i's.

Thus, the joint ML estimation of the $\underline{\theta}_k$'s is obtained by solving

$$\min_{\underline{\theta}_1,\cdots,\underline{\theta}_K}\int_{T_i}^{T_f}\left[\underline{y}(t) - \sum_{k=1}^{K}\underline{s}_k(t;\underline{\theta}_k)\right]^{\dagger} Q^{-1}\left[\underline{y}(t) - \sum_{k=1}^{K}\underline{s}_k(t;\underline{\theta}_k)\right] dt \tag{4.3}$$

Or for discrete observations,

$$\min_{\underline{\theta}_1,\cdots,\underline{\theta}_K}\sum_{i=1}^{N}\left[\underline{y}(t_i) - \sum_{k=1}^{K}\underline{s}_k(t_i;\underline{\theta}_k)\right]^{\dagger} Q^{-1}\left[\underline{y}(t_i) - \sum_{k=1}^{K}\underline{s}_k(t_i;\underline{\theta}_k)\right] \tag{4.4}$$

In either case, we have a complicated multi-parameter optimization problem. Of course, standard techniques, such as Gauss-Newton or some other iterative gradient search algorithm, can always be used to solve the problem. However, when applied to the problem at hand, these methods tend to be computationally complex and time consuming.

We note that the more general problem, where the noise vector, $\underline{n}(t)$, has an arbitrarily given power spectrum matrix, $N(\omega)$, may be reduced to the problem presented above, where the noise vector is white, by using an appropriate whitening filter. Let $\underline{\tilde{s}}_k(t;\underline{\theta}_k)$ be the output of the whitening filter to the input $\underline{s}_k(t;\underline{\theta}_k)$. In this case the likelihood of the observations is still given by (4.2), where we use $\underline{\tilde{s}}_k(t;\underline{\theta}_k)$ instead of $\underline{s}_k(t;\underline{\theta}_k)$.

4.1.2 Solution using the EM method

Having in mind the EM algorithm and the class of iterative algorithms, developed in the first part of the thesis, we want to simplify the optimization problem associated with the direct ML approach.

In order to apply an EM algorithm to the problem at hand, we need to specify the complete data. A natural choice of a complete data, $\underline{x}(t)$, is obtained by decomposing $\underline{y}(t)$ into its signal components, that is

$$\underline{x}(t) = \begin{bmatrix} \underline{x}_1(t) \\ \underline{x}_2(t) \\ \cdot \\ \cdot \\ \underline{x}_K(t) \end{bmatrix} \tag{4.5}$$

where

$$\underline{x}_k(t) = \underline{s}_k(t;\underline{\theta}_k) + \underline{n}_k(t) \tag{4.6}$$

and the $\underline{n}_k(t)$ are obtained by arbitrarily decomposing the total noise signal, $\underline{n}(t)$, into K components, so that

$$\sum_{k=1}^{K} \underline{n}_k(t) = \underline{n}(t) \tag{4.7}$$

From (4.1), (4.6) and (4.7), the relation between the complete data $\underline{x}(t)$ and the incomplete data $\underline{y}(t)$ is given by

$$\underline{y}(t) = \sum_{k=1}^{K} \underline{x}_k(t) = H \cdot \underline{x}(t) \tag{4.8}$$

where

$$H = \overbrace{[I|I|\cdots|I]}^{K\ terms}$$

We will find it most convenient to choose the $n_k(t)$ to be statistically independent zero-mean and Gaussian with a covariance matrix

$$E\left\{\underline{n}_k(t)\underline{n}_k(\sigma)\right\} = Q_k \cdot \delta(t-\sigma)$$

where $Q_k = \beta_k Q$ and the β_k's are arbitrary real valued scalars satisfying

$$\sum_{k=1}^{K} \beta_k = 1, \qquad \beta_k \geq 0 \tag{4.9}$$

We will discuss methods for choosing specific β_k's later.

In this case the log-likelihood of the complete data $\underline{x}(t)$ is given by

$$L_c(\underline{\theta}) = \log f_X(\underline{x};\underline{\theta}) = C - \frac{\lambda}{2}\int_{T_i}^{T_f} [\underline{x}(t) - \underline{s}(t;\underline{\theta})]^{\dagger}\Lambda^{-1}[\underline{x}(t) - \underline{s}(t;\underline{\theta})]dt \tag{4.10}$$

where C contains all the terms that are independent of $\underline{\theta}$. The vector $\underline{s}(t;\underline{\theta})$ is given by,

$$\underline{s}(t;\underline{\theta}) = \begin{bmatrix} \underline{s}_1(t;\underline{\theta}_1) \\ \underline{s}_2(t;\underline{\theta}_2) \\ \cdot \\ \cdot \\ \underline{s}_K(t;\underline{\theta}_K) \end{bmatrix} \tag{4.11}$$

and it is the mean of $\underline{x}(t)$. The matrix Λ is the covariance matrix of $\underline{x}(t)$,

$$\Lambda = \begin{bmatrix} Q_1 & & & & \\ & Q_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & Q_K \end{bmatrix} \tag{4.12}$$

The notation in (4.12) indicates that Λ is a block-diagonal matrix. Thus, the log-likelihood of the complete data may be written as

$$L_c(\underline{\theta}) = C - \frac{\lambda}{2} \sum_{k=1}^{K} \int_{T_i}^{T_f} [\underline{x}_k(t) - \underline{s}_k(t; \underline{\theta})]^{\dagger} Q_k^{-1} [\underline{x}_k(t) - \underline{s}_k(t; \underline{\theta})] dt \tag{4.13}$$

From this expression, we notice that, if the complete data was available, then maximizing its likelihood with respect to $\underline{\theta}$ is equivalent to the minimization of each of the terms in the sum above ***separately***, which is simpler than solving a multi-variable optimization problem with respect to all $\underline{\theta}_k$'s at once. We also notice that the sufficient statistics of the complete data contain only linear terms, since the quadratic terms in (4.10) are independent of the unknown parameters.

Of course, we do not observe the complete data. However, we take advantage of the special structure of the likelihood of the complete data by using an EM algorithm with this specification of complete data. This EM algorithm will iterate between estimating $\underline{x}(t)$ and using the estimated value in (4.13) to updating the parameters by a separate optimization with respect to each $\underline{\theta}_k$.

More specifically, from (2.19), an EM iteration is summarized by

$$\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} Q(\underline{\theta}; \underline{\theta}^{(n)}) = \arg\max_{\underline{\theta}} E\left\{ \log f_X(\underline{x}; \underline{\theta}) \,\middle|\, \underline{y}, \underline{\theta}^{(n)} \right\} \tag{4.14}$$

However, in this case

$$\arg\max_{\underline{\theta}} Q(\underline{\theta};\underline{\theta}^{(n)}) = \arg\max_{\underline{\theta}} \left[C - \frac{\lambda}{2}\int_{T_i}^{T_f} [\hat{\underline{x}}^{(n)}(t) - \underline{s}(t;\underline{\theta})]^{\dagger}\Lambda^{-1}[\hat{\underline{x}}^{(n)}(t) - \underline{s}(t;\underline{\theta})]dt\right] \quad (4.15)$$

where

$$\hat{\underline{x}}^{(n)}(t) = E\left\{\underline{x}(t)/\underline{y},\underline{\theta}^{(n)}\right\} = \underline{s}(t;\underline{\theta}^{(n)}) + \Lambda H^{\dagger}\cdot[H\Lambda H^{\dagger}]^{-1}[\underline{y}(t) - H\cdot\underline{s}(t;\underline{\theta}^{(n)})] \quad (4.16)$$

Substituting (4.11) and (4.12) in (4.15) and following straight forward matrix manipulations, we see that the maximization of (4.15) is equivalent to the minimization cf the sum,

$$\min_{\underline{\theta}_1,\cdots,\underline{\theta}_K} \sum_{k=1}^{K}\int_{T_i}^{T_f}\left[\hat{\underline{x}}_k^{(n)}(t) - \underline{s}_k(t;\underline{\theta}_k)\right]^{\dagger} Q_k^{-1}\left[\hat{\underline{x}}_k^{(n)}(t) - \underline{s}_k(t;\underline{\theta}_k)\right]dt \quad (4.17)$$

where $\hat{\underline{x}}_k^{(n)}$ is the k^{th} component of $\hat{\underline{x}}^{(n)}$. This minimization of the sum is equivalent to the minimization of each of its components ***separately***, with respect to $\underline{\theta}_k$.

Also, substituting (4.12) in (4.15), the gain matrix becomes

$$\Lambda H^{\dagger}[H\Lambda H^{\dagger}]^{-1} = diag(\beta_1,\beta_2,\cdots,\beta_K) \quad (4.18)$$

where ***diag***$(\cdots)$ indicates a diagonal matrix.

Summarizing all these relations, we may now write the E and M steps of the EM algorithm for this problem as follows:

- **The E step:** For $k = 1,2,\cdots,K$ compute

$$\underline{x}_k^{(n)}(t) = \underline{s}_k(t;\underline{\theta}_k^{(n)}) + \beta_k\left[\underline{y}(t) - \sum_{\ell=1}^{K}\underline{s}_\ell(t;\underline{\theta}_\ell^{(n)})\right] \quad (4.19)$$

where the β_k's are any real-valued positive scalars satisfying

$$\sum_{k=1}^{K}\beta_k = 1$$

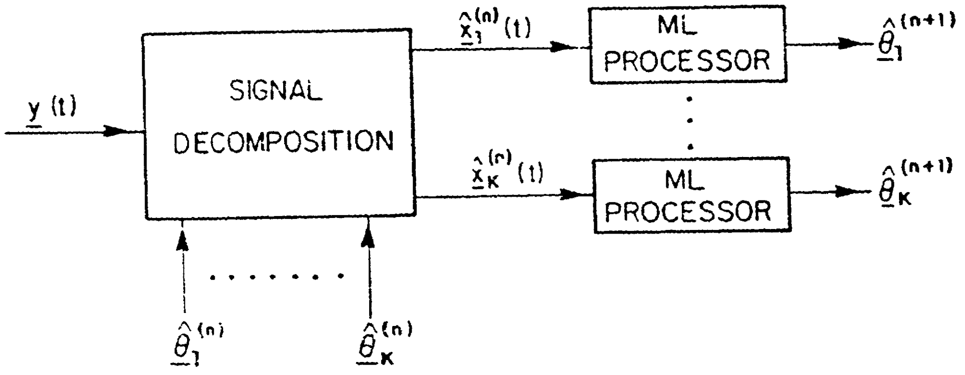

y (t)
SIGNAL DECOMPOSITION
$\hat{\underline{x}}_1^{(n)}(t)$
$\hat{\underline{x}}_K^{(n)}(t)$
ML PROCESSOR
ML PROCESSOR
$\hat{\underline{\theta}}_1^{(n+1)}$
$\hat{\underline{\theta}}_K^{(n+1)}$
$\hat{\underline{\theta}}_1^{(n)}$
$\hat{\underline{\theta}}_K^{(n)}$

Figure 4.2: The EM algorithm for Deterministic (known) signals

- **The M step:** For $k = 1, 2, \cdots, K$

$$\underline{\theta}_k^{(n+1)} = \arg\min_{\underline{\theta}_k} \int_{T_i}^{T_f} \left[\underline{x}_k^{(n)} - \underline{s}_k(t; \underline{\theta}_k)\right]^{\dagger} Q_k^{-1} \left[\underline{x}_k^{(n)} - \underline{s}_k(t; \underline{\theta}_k)\right] dt \qquad (4.20)$$

We observe that $\underline{\theta}_k^{(n+1)}$ is in fact the ML estimate of $\underline{\theta}_k$ based on $\hat{\underline{x}}_k^{(n)}$. This algorithm is illustrated schematically in Figure 4.2. We note that in the case of discrete observations, the integral of (4.20) is replaced by the sum over the points $\{t_i\}$.

The most striking feature of this algorithm is that it decouples the complicated multi-parameter optimization into k separate ML optimizations. Hence, the complexity of the algorithm is essentially unaffected by the assumed number of signal components. As K increases, we have to increase the number of ML processors in parallel: however, each ML processor is maximized separately. Since the algorithm is based on the EM method, each iteration cycle increases the likelihood until convergence is accomplished.

Unknown signal waveforms

The algorithm developed above assumes that the signal waveforms, $\underline{s}_k(t)$, are known a-priori. In practice, one is unlikely to have a detailed prior knowledge of these waveforms, in which case they must be estimated jointly with the parameters, $\underline{\theta}_k$. We will consider the samples of the unknown waveforms as additional parameters; using the same statistical formulation as above, we will get an ML problem for estimating the waveforms. Following the same considerations as above, we can specify the E and M steps of an EM algorithm for estimating the waveforms and the parameters, as,

- **The E step:** For $k = 1, 2, \cdots, K$ compute

$$\underline{x}_k^{(n)}(t) = \underline{s}_k^{(n)}(t;\underline{\theta}_k^{(n)}) + \beta_k \left[\underline{y}(t) - \sum_{\ell=1}^{K} \underline{s}_\ell^{(n)}(t;\underline{\theta}_\ell^{(n)})\right] \tag{4.21}$$

- **The M step:** Minimize

$$\sum_{k=1}^{K} \int_{T_i}^{T_f} \left[\underline{x}_k^{(n)} - \underline{s}_k(t;\underline{\theta}_k)\right]^{\dagger} Q_k^{-1} \left[\underline{x}_k^{(n)} - \underline{s}_k(t;\underline{\theta}_k)\right] dt \tag{4.22}$$

  with respect to $\underline{\theta}_1, \cdots, \underline{\theta}_K$ and $\underline{s}_1(t), \cdots, \underline{s}_K(t)$.

The E step is identical to (4.19), where instead of using the a-priori given waveforms, we use the current estimated waveforms, $\underline{s}_k^{(n)}(t)$. The M step requires a more complicated maximization. However, we will be able to give an explicit example for this M step later in the chapter.

We note that the ML problem for estimating the waveforms in addition to the unknown parameters is ill-posed, since there may be too many unknowns. To make the problem well-posed, we have to incorporate some constraints on the possible signal waveforms. However, we have to make sure that these constraints will correspond to the real, physical situation.

4.1.3 The extended EM algorithm and the choice of the β_k's

The EM algorithm presented above corresponds to a family of complete data definitions. Each choice of the β_k's implies a different choice of complete data, $\underline{x}(t)$; the probability sample space and the corresponding p.d.f. of $\underline{x}(t)$ depend on the choice of the β_k's. The convenient feature of this family of complete data definitions is that each member of the family satisfies the same relation between complete and incomplete data given by (4.8). This feature allowed the presentation of the EM algorithm for the entire family at once.

This family of complete data definitions may be further extended, while keeping the simple structure of the algorithm steps (4.19) and (4.20), in the following way. We could model the complete data, $\underline{x}(t)$, as a Gaussian process, whose mean is $\underline{s}(t;\underline{\theta})$ as in (4.11), but whose variance is time dependent and given by,

$$\Lambda(t) = \begin{bmatrix} Q_1(t) & & & & \\ & Q_2(t) & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & Q_K(t) \end{bmatrix} = \begin{bmatrix} \beta_1(t)Q & & & & \\ & \beta_2(t)Q & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \beta_K(t)Q \end{bmatrix} \tag{4.23}$$

where $\beta_k(t)$ are arbitrary real values, satisfying for all t

$$\sum_{k=1}^{K} \beta_k(t) = 1, \qquad \beta_k(t) \geq 0 \qquad \forall t\ T_i \leq t \leq T_f \tag{4.24}$$

Any member of this extended family of complete data definitions corresponds to decomposing the observation noise, $\underline{n}(t)$, into statistically independent zero-mean Gaussian ***non-stationary*** components, $n_k(t)$, whose covariance matrix is

$$E\left\{\underline{n}_k(t)\underline{n}_k(\sigma)\right\} = Q_k(t)\cdot\delta(t-\sigma) = \beta_k(t)Q\delta(t-\sigma)$$

The E step of the EM algorithm for each member of this extended family is similar to (4.19), except that $\beta_k(t)$'s replace the time invariant β_k's. The M step is similar to (4.20), where again the time varying $Q_k(t)$'s replace the time invariant matrices Q_k's

From the discussion above, we see that the suggested class of EM algorithms has many degrees of freedom. We may look for a choice of $\beta_k(t)$'s that give the simplest and fastest algorithm. Furthermore, following the discussion in section 2.4, instead of fixing this choice, the $\beta_k(t)$'s may vary from iteration to iteration, according to some a-priori rule or depending on the current parameter values, $\underline{\theta}^{(n)}$. The EM algorithm, where the complete data definition varies from iteration to iteration, has been referred to as the extended EM (EEM) algorithm. Examples for applying extended EM algorithms to our problem are now given.

Suppose that in some iteration one of the β_k's, say β_ℓ, is chosen to be unity; the remaining β_k's must be zero. In the next iteration, we will choose $\beta_{\ell+1}$ to be unity and so on (in a cyclic way so that after β_K is unity, β_1 will be unity). Substituting these β_k's in (4.19) and (4.20), we notice that the resulting algorithm is equivalent to a coordinate search or alternate maximization algorithm of (4.2). In each iteration, $\underline{\theta}_k^{(n+1)} = \underline{\theta}_k^{(n)}$ for all k's that correspond to a zero β_k, and $\underline{\theta}_\ell$ is updated by,

$$\underline{\theta}_\ell^{(n-1)} = \arg\min_{\underline{\theta}_\ell} \int_{T_i}^{T_f} \left[\underline{y}(t) - \sum_{k\neq\ell} \underline{s}_k(t;\underline{\theta}_k^{(n)}) - \underline{s}_\ell(t;\underline{\theta}_\ell)\right]^\dagger Q^{-1} \left[\underline{y}(t) - \sum_{k\neq\ell} \underline{s}_k(t;\underline{\theta}_k^{(n)}) - \underline{s}_\ell(t;\underline{\theta}_\ell)\right] \tag{4.25}$$

where ℓ corresponds to the unity β_ℓ.

While in the previous example, we have shown how, by varying the complete data, the EM algorithm has been reduced to a simple (but not necessarily efficient) algorithm, we will now show how an algorithm with a superlinear convergence rate may be achieved. To simplify the exposition, we will discuss a degenerate scalar case, where the unknown parameters are given by the scalar θ.

From (2.65), in order to achieve a superlinear convergence rate, we have to choose, in each iteration, β_k's (or complete data) that are the solution to the equation,

$$D^{11}Q(\theta^{(n)};\theta^{(n)}) = \frac{\partial^2}{\partial\theta_1\partial\theta_2}Q(\theta_1;\theta_2)\Big|_{\theta_1=\theta^{(n)},\theta_2=\theta^{(n)}} = 0 \tag{4.26}$$

Following (4.15),(4.16) and (4.17), the expression for $Q(\theta_1;\theta_2)$ in this case is given by,

$$Q(\theta_1;\theta_2) = C - \sum_{k=1}^{K}\int_{T_i}^{T_f}\Big[\hat{\underline{x}}_k^{(n)}(t;\theta_2) - \underline{s}_k(t;\theta_1)\Big]^{\dagger} Q_k^{-1}(t)\Big[\hat{\underline{x}}_k^{(n)}(t;\theta_2) - \underline{s}_k(t;\theta_1)]dt\Big] \tag{4.27}$$

where $\hat{\underline{x}}_k^{(n)}(t;\theta_2)$ is given by,

$$\underline{x}_k^{(n)}(t;\theta_2) = \underline{s}_k(t;\theta_2) - \beta_k(t)\left[\underline{y}(t) - \sum_{\ell=1}^{K}\underline{s}_\ell(t;\theta_2)\right] \tag{4.28}$$

Thus, a possible solution of (4.26) is to choose

$$\beta_k(t) = \frac{\frac{\partial}{\partial\theta}s_k(t;\theta)|_{\theta^{(n)}}}{\sum_{\ell=1}^{K}\frac{\partial}{\partial\theta}s_\ell(t;\theta)|_{\theta^{(n)}}} \tag{4.29}$$

If this choice of $\beta_k(t)$ is allowed, the convergence rate of the resulting EM algorithm will be superlinear.

Another desired feature of an EM algorithm with varying complete data is that it may avoid convergence to unwanted stationary points. Following the discussion in section 2.4, the simplest procedure is to randomly choose $\beta_k(t)$ in each iteration. These randomly chosen $\beta_k(t)$ have to satisfy the constraints of (4.24), however. A more complicated procedure is to search for the choice of $\beta_k(t)$ in the domain, defined by the constraints of (4.24), that will give the largest increase in the likelihood. Since searching the entire domain of possible $\beta_k(t)$ may be too complicated, we will search only in a sub-domain, which is randomly chosen, in each iteration.

4.2 Parameter estimation of superimposed signals: The stochastic Gaussian case

In the previous section, the signal components, $\underline{s}_k(t;\underline{\theta}_k)$, were deterministic. In this section, we will present the statistical maximum likelihood problem and its solutions using the EM algorithm for the case where the signal components, $\underline{s}_k(t;\underline{\theta}_k)$, of the observed composite signal are modeled as sample functions from a Gaussian stochastic processes. This modeling is natural in a variety of applications. Consider, for example, a passive sonar environment, where the targets generate noise-like acoustic signals. The signals from several targets are superimposed and measured by our array sensors with an additional background noise. We may or may not know the spectral characteristics of the targets' signals. However, we are usually interested in finding the geometrical parameters, i.e. the location or the bearing of the targets.

By assuming that the signal components and the background noise are Gaussian processes, we get a statistical maximum likelihood problem for estimating the unknown parameters (which are the geometrical parameters and maybe some spectral parameters of the signals in the example above). It is difficult to solve this statistical problem directly; indeed, in many applications, suboptimal procedures were suggested. We, however, will present in this section procedures, based on the EM algorithm and its extensions, whose goal is to be optimal, i.e. to solve this maximum likelihood problem.

4.2.1 The ML problem

Consider the model of (4.1) under the following assumptions:

- The signals $\underline{s}_k(t;\underline{\theta}_k)$ $k = 1,\cdots,K$ are mutually uncorrelated, wide sense stationary (WSS), zero-mean, vector Gaussian stochastic processes, whose power spectrum matrices are $S_k(\omega;\underline{\theta}_k)$, $k = 1,2,\cdots,K$.
- The noise, $\underline{n}(t)$, is a WSS, zero-mean, vector Gaussian processes with a given power spectrum matrix, $N(\omega)$.
- The observation interval, $T = T_f - T_i$, is long compared with the correlation time (inverse bandwidth) of the signals and the noise, i.e. $WT/2\pi >> 1$.

Under the above assumptions, the observed signals, $\underline{y}(t)$, are also WSS, zero-mean and Gaussian. WSS processes with a long observation time are conveniently analyzed in the frequency domain. Fourier transforming $\underline{y}(t)$ we obtain

$$\underline{Y}(\omega_\ell) = \frac{1}{\sqrt{T}} \int_{T_i}^{T_f} \underline{y}(t) e^{-j\omega_\ell t} dt, \qquad \omega_\ell = \frac{2\pi}{T} \cdot \ell \tag{4.30}$$

For $WT/2\pi >> 1$, the $\underline{Y}(\omega_\ell)$'s are asymptotically uncorrelated, zero-mean and Gaussian with the covariance matrix $P(\omega_\ell;\underline{\theta})$, where $P(\omega;\underline{\theta})$ is the spectral density matrix of $\underline{y}(t)$ given by,

$$P(\omega;\underline{\theta}) = \sum_{k=1}^{K} S_k(\omega;\underline{\theta}_k) + N(\omega) \tag{4.31}$$

The log-likelihood function observing the $\underline{Y}(\omega_\ell)$'s is therefore given by.

$$L(\underline{\theta}) = -\sum_{\ell} \left[\log \det \pi P(\omega_\ell;\underline{\theta}) + \underline{Y}^{\dagger}(\omega_\ell) \cdot P^{-1}(\omega_\ell;\underline{\theta}) \cdot \underline{Y}(\omega_\ell)\right] \tag{4.32}$$

where the summation in (4.32) is carried over all ω_ℓ in the signal frequency band. In the case of discrete observations, the log-likelihood is still given by (4.32), where the $\underline{Y}(\omega_\ell)$'s are the discrete Fourier transform (DFT) of the observed signals.

In either case, to obtain the ML estimate of the various $\underline{\theta}_k$'s we must solve the following joint optimization problem

$$\min_{\underline{\theta}_1,\underline{\theta}_2,\cdots,\underline{\theta}_K} \sum_{\ell} \left[\log \det P(\omega_\ell;\underline{\theta}) + \underline{Y}^\dagger(\omega_\ell) \cdot P^{-1}(\omega_\ell;\underline{\theta}) \cdot \underline{Y}(\omega_\ell)\right] \tag{4.33}$$

This is usually a complicated joint optimization problem. Standard search techniques, such as gradient or Newton-Raphson methods, tend to be complex, when applied to this problem. Thus, we will propose using the EM method to by-pass this complicated multi-parameter optimization.

4.2.2 Solution using the EM method

Following the same considerations as in the deterministic signal case, a natural choice of complete data, $\underline{x}(t)$, will be obtained by decomposing the observed signal, $\underline{y}(t)$, into its signal components. Thus, repeating equations (4.5) and (4.6), the complete data, $\underline{x}(t)$, is given by,

$$\underline{x}(t) = \begin{bmatrix} \underline{x}_1(t) \\ \underline{x}_2(t) \\ \cdot \\ \cdot \\ \underline{x}_K(t) \end{bmatrix} \tag{4.34}$$

where

$$\underline{x}_k(t) = \underline{s}_k(t;\underline{\theta}_k) + \underline{n}_k(t) \tag{4.35}$$

Again, the $\underline{n}_k(t)$ are chosen to be mutually uncorrelated, zero-mean and Gaussian, whose spectral density matrices are $N_k(\omega) = \beta_k \cdot N(\omega)$, where the β_k's are arbitrary real-valued constants subject to (4.9). Thus the relation between complete and incomplete data is given again by $\underline{y}(t) = H \cdot \underline{x}(t)$ as in (4.8).

The log-likelihood of the complete data $\underline{x}(t)$ is given by

$$\begin{aligned} L_c(\underline{\theta}) = \log f_X(\underline{x};\underline{\theta}) &= -\sum_{\ell}\left[\log\det \pi\Lambda(\omega_\ell;\underline{\theta}) + \underline{X}^{\dagger}(\omega_\ell)\cdot\Lambda^{-1}(\omega_\ell;\underline{\theta})\cdot\underline{X}(\omega_\ell)\right] \qquad (4.36) \\ &= -\sum_{\ell}\left[\log\det \pi\Lambda(\omega_\ell;\underline{\theta}) + tr\left\{\Lambda^{-1}(\omega_\ell;\underline{\theta})\cdot\underline{X}(\omega_\ell)\underline{X}^{\dagger}(\omega_\ell)\right\}\ \right] \end{aligned}$$

where $\underline{X}(\omega_\ell)$ is obtained by Fourier transforming $\underline{x}(t)$, i.e. any of its components, $\underline{X}_k(\omega_\ell)$, is given by,

$$\underline{X}_k(\omega_\ell) = \frac{1}{\sqrt{T}}\int_{T_i}^{T_f} \underline{x}_k(t)e^{-j\omega_\ell t}dt, \qquad \omega_\ell = \frac{2\pi}{T}\cdot\ell \qquad (4.37)$$

The matrix $\Lambda(\omega_\ell;\underline{\theta})$ is the power spectrum density matrix of the complete data. It is a block diagonal matrix given by

$$\Lambda(\omega_\ell;\underline{\theta}) = \begin{bmatrix} \Lambda_1(\omega_\ell;\underline{\theta}_1) & & & & \\ & \Lambda_2(\omega_\ell;\underline{\theta}_2) & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \Lambda_K(\omega_\ell;\underline{\theta}_K) \end{bmatrix} \qquad (4.38)$$

where

$$\Lambda_k(\omega;\underline{\theta}_k) = S_k(\omega;\underline{\theta}) + \beta_k\cdot N(\omega) \qquad (4.39)$$

Exploiting the block diagonal form of $\Lambda(\omega;\underline{\theta})$, the likelihood of the complete data may be written as,

$$L_c(\underline{\theta}) = -\sum_{k=1}^{K}\sum_{\ell}\left[\log\det \pi\Lambda_k(\omega_\ell;\underline{\theta}_k) + tr\left\{\Lambda_k^{-1}(\omega_\ell;\underline{\theta}_k)\cdot\underline{X}_k(\omega_\iota)\underline{X}_k^{\dagger}(\omega_\ell)\right\}\ \right] \qquad (4.40)$$

From this expression, we notice that, if the complete data was available, maximizing its likelihood with respect to $\underline{\theta}$ is equivalent to minimizing each of the terms in the sum above with respect to $\underline{\theta}_k$ ***separately***. This is much simpler than solving a multi-variable optimization problem with respect to all $\underline{\theta}_k$'s at once. The sufficient statistics of the complete data is composed of the quadratic terms $\left\{\underline{X}_k(\omega_\ell)\underline{X}_k^\dagger(\omega_\ell)\right\}$.

In the suggested EM algorithm, we take advantage of this simple form of the likelihood of the complete data. The E step will estimate the required quadratic statistics and the M step will use the estimated statistics in (4.40) and update the parameters via the simple maximization, associated with the likelihood of the complete data.

The required statistics are the diagonal blocks of the matrix $\underline{X}(\omega_\ell)\underline{X}^\dagger(\omega_\ell)$. The relation between complete and incomplete data is linear (i.e. $\underline{Y}(\omega_\ell) = H \cdot \underline{X}(\omega_\ell)$) and the data is Gaussian. Thus, using the results developed for the linear Gaussian case (see (2.55)), the conditional expectation of the matrix $\underline{X}(\omega_\ell)\underline{X}^\dagger(\omega_\ell)$, given the observations, $\underline{Y}(\omega_\ell)$, and an assignment, $\underline{\theta}'$, to the parameters, is given by

$$\begin{aligned} \Psi(\omega_\ell) &= E\left\{\underline{X}(\omega_\ell)\underline{X}^\dagger(\omega_\ell)/\underline{Y}(\omega_\ell;\underline{\theta}'\right\} \\ &= \left[I - \Gamma(\omega;\underline{\theta}') \cdot H\right] \Lambda(\omega_\ell;\underline{\theta}') + \Gamma(\omega;\underline{\theta}')\underline{Y}(\omega_\ell) \cdot \underline{Y}^\dagger(\omega_\ell)\Gamma^\dagger(\omega;\underline{\theta}') \qquad (4.41) \end{aligned}$$

where $\Gamma(\omega;\underline{\theta}')$ is the "Kalman gain"

$$\Gamma(\omega;\underline{\theta}') = \Lambda(\omega_\ell;\underline{\theta}')H^\dagger\left[H\Lambda(\omega_\ell;\underline{\theta}')H^\dagger\right]^{-1}$$

Using straight forward algebraic manipulations the (k,k) block of $\Psi(\omega_\ell)$ is given by

$$\begin{aligned} \Psi_k(\omega_\ell) &= \Lambda_k(\omega_\ell;\underline{\theta}_k') - \Lambda_k(\omega_\ell;\underline{\theta}_k')P^{-1}(\omega_\ell;\underline{\theta}')\Lambda_k(\omega_\ell;\underline{\theta}_k') + \\ &\quad +\Lambda_k(\omega_\ell;\underline{\theta}_k')P^{-1}(\omega_\ell;\underline{\theta}')\underline{Y}(\omega_\ell) \cdot \underline{Y}^\dagger(\omega_\ell)P^{-1}(\omega_\ell;\underline{\theta}')\Lambda_k(\omega_\ell;\underline{\theta}_k') \qquad (4.42) \end{aligned}$$

where $P(\omega_\ell;\underline{\theta})$ is defined by (4.31).

These estimated statistics are used instead of the unavailable statistics of the complete data. The maximization in the M step will be equivalent to K separate minimizations with respect to each $\underline{\theta}_k$. The E and M steps of the EM algorithm for the Gaussian superimposed signal problem may be summarized as follows:

- **The E step:** For $k = 1, 2, \cdots, K$ compute

$$\begin{aligned}\Psi_k^{(n)}(\omega_\ell) &= \Lambda_k(\omega_\ell; \underline{\theta}_k^{(n)}) - \Lambda_k(\omega_\ell; \underline{\theta}_k^{(n)}) P^{-1}(\omega_\ell; \underline{\theta}^{(n)}) \Lambda_k(\omega_\ell; \underline{\theta}_k^{(n)}) - \qquad (4.43)\\ &\quad + \Lambda_k(\omega_\ell; \underline{\theta}_k^{(n)}) P^{-1}(\omega_\ell; \underline{\theta}^{(n)}) \underline{Y}(\omega_\ell) \cdot \underline{Y}^\dagger(\omega_\ell) P^{-1}(\omega_\ell; \underline{\theta}^{(n)}) \Lambda_k(\omega_\ell; \underline{\theta}_k^{(n)})\end{aligned}$$

- **The M step:** For $k = 1, 2, \cdots, K$

$$\underline{\theta}_k^{(n+1)} \arg\min_{\underline{\theta}_k} \sum_\ell \left[\log\det \Lambda_k(\omega_\ell; \underline{\theta}_k) + tr\left\{\Lambda_k^{-1}(\omega_\ell; \underline{\theta}_k) \cdot \Psi_k^{(n)}(\omega_\ell)\right\}\ \right] \qquad (4.44)$$

We observe that $\underline{\theta}_k^{(n+1)}$ is the ML estimate of $\underline{\theta}_k$, where $\underline{X}_k(\omega_\ell)\underline{X}_k^\dagger(\omega_\ell)$ is replaced by its current estimate, $\Psi_k^{(n)}(\omega_\ell)$. The algorithm is illustrated in Figure 4.3. The most attractive feature of the algorithm is that it decouples the full multi-dimensional optimization of equation (4.33) into optimizations in smaller dimensional parameter subspaces. As in the deterministic signal case, the complexity of the algorithm is essentially unaffected by the assumed number of signal components. As K increases, we have to increase the number of parallel ML processors; however, each ML processor operates independently. Since the algorithm is based on the EM method, each iteration cycle increases the likelihood until convergence is accomplished.

As in the deterministic case, this EM algorithm corresponds to a family of complete data definitions. A specific member of the family is associated with a specific choice of the β_k's. This family of complete data definitions can be extended by allowing a different choice of β_k's in each frequency. The EM algorithm for any member of this extended family will keep the structure of the algorithm steps (4.43) and (4.44), where $\beta_k(\omega_\ell)$ is used in the definition of $\Lambda_k(\cdot)$, (4.39).

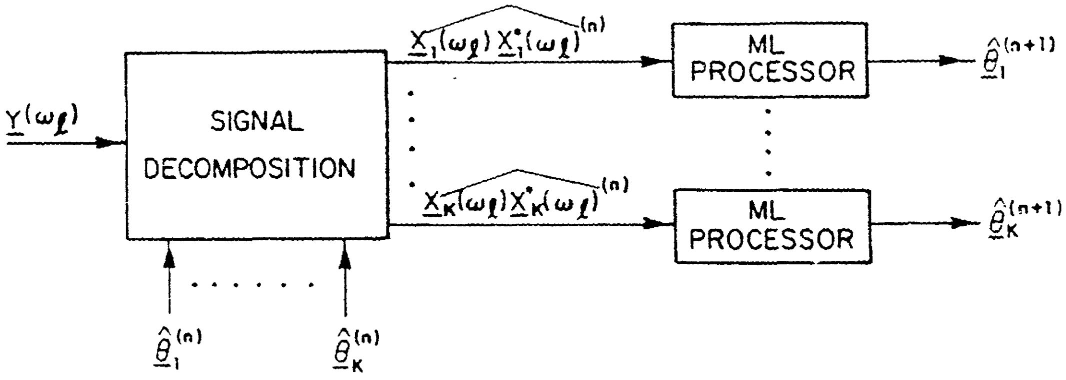

$\underline{Y}(\omega_\ell)$
SIGNAL DECOMPOSITION
$\widehat{\underline{X}_1(\omega_\ell)\underline{X}_1^*(\omega_\ell)}^{(n)}$
$\widehat{\underline{X}_K(\omega_\ell)\underline{X}_K^*(\omega_\ell)}^{(n)}$
ML PROCESSOR
ML PROCESSOR
$\hat{\underline{\theta}}_1^{(n+1)}$
$\hat{\underline{\theta}}_K^{(n+1)}$
$\hat{\underline{\theta}}_1^{(n)}$
$\hat{\underline{\theta}}_K^{(n)}$

Figure 4.3: The EM algorithm for stochastic Gaussian signals

We may choose a fixed complete data definition throughout the algorithm iterations, or we may vary the complete data definition from iteration to iteration. Varying the complete data within this family of complete data definitions will correspond to choosing different $\beta_k(\omega_\ell)$'s in each iteration, but otherwise the algorithm steps remain the same Possible strategies for choosing the complete data and varying it from iteration to iteration were discussed in Chapter 2 and previously in this chapter, for the deterministic signal case. These discussions are relevant in this case too.

4.3 Application to multipath time-delay estimation

Time delay estimation is a common problem in underwater acoustics as well as in radar. Geometrical parameters (such as range and location of targets) and physical parameters (such as velocity and temperature profiles of the ocean) are typically found via time delay analysis. Consider, for example, the ocean tomography experiment, described in [50] and [51]. An acoustic transducer, located at a given point in the ocean, transmits a signal which is received time-delayed by known location sensors. The estimated delay times between the transmitted signal and the received signals are used as an input to an inverse problem that finds the ocean profile in the experiment area.

Multipath may occur due to reflections and propagation modes. The received signal in this case contains several echoes of the transmitted signal having different time-delays and attenuations, i.e. it may be written as

$$y(t) = \sum_{k=1}^{K} \alpha_k s(t - \tau_k) + n(t) \tag{4.45}$$

The existence of more than one path is undesired in some cases; in the ocean tomography experiment, the additional echoes interfere with and corrupt the interesting direct path signal. However, in other cases, additional important information may be obtained from finding the time delay of the other paths. A single sensor may determine the range and the depth of a target, if we can find the delay times of the direct path and of the paths that result from a single bottom or surface reflection.

We will be interested in this section in estimating the delay times of the multipath signal (4.45). In a variety of applications, we may model the components of the multipath signal as deterministic or stochastic. In applications such as ocean tomography, active sonar/radar and many more, a deterministic known waveform signal (pulse) is transmitted. In a passive determination of range and depth of a target by a single sensor, the target may generate a noise-like acoustic signal, naturally modeled as a sample signal from a stochastic Gaussian process. In both cases, we will be able to apply the results of the previous sections to obtain an estimate of the delay times via the EM algorithm.

We will present, in this section, the detailed algorithm and experimental results for the deterministic signal case. In [52], we have presented the EM algorithm for multipath time delay estimation in the Gaussian signal case.

4.3.1 The deterministic case

Suppose that the observed signal is given by (4.45), where $s(t)$ is a known signal waveform, the noise $n(t)$ is Gaussian with a flat spectrum over the receiver frequency band, and the observation time is $T_i \leq t \leq T_f$. The problem is to estimate the pairs (α_k, τ_k) for $k = 1, 2, \cdots, K$.

In this case, the direct ML approach given by (4.3), reduces to,

$$\min_{\substack{\tau_1, \tau_2, \cdots, \tau_K \\ \alpha_1, \alpha_2, \cdots, \alpha_K}} \int_{T_i}^{T_f} \left| y(t) - \sum_{k=1}^{K} \alpha_k s(t - \tau_k) \right|^2 dt \tag{4.46}$$

This optimization problem is addressed in [53], where it is shown that the optimal α_k's may be expressed explicitly in terms of the optimal τ_k's. Thus, the $2K$-dimensional search can be reduced into a K-dimensional search. However, as pointed out in [53], for $K \geq 3$ the required computations become too intensive. Consequently, ad-hoc approaches and sub-optimal solutions have been proposed. The most common solution consists of correlating $y(t)$ with a replica of $s(t)$ and searching for the K highest peaks of the correlation function. If the various paths are resolvable, i.e. the difference between τ_k and τ_ℓ is long compared with the temporal correlation of the signal for *all* combinations of k and ℓ, this approach yields near optimal estimates. However, in situations where the signal paths are unresolvable, this approach is distinctly sub-optimal.

We identify the model in (4.45) as a special case of (4.1). Therefore, in correspondence with equations (4.19) and (4.20), we obtain the following algorithm:

- Start, $n = 0$, initialize $\alpha_k^{(0)}$ and $\tau_k^{(0)}$, $k = 1, \cdots, K$.

- Iterate (until some convergence criterion is met)

  - **The E step:** For $k = 1, 2, \cdots, K$ compute

$$x_k^{(n)}(t) = \alpha_k^{(n)} s(t - \tau_k^{(n)}) + \beta_k \left[y(t) - \sum_{\ell=1}^{K} \alpha_\ell^{(n)} s(t - \tau_\ell^{(n)}) \right] \qquad (4.47)$$

    where the β_k's are any real-valued positive scalars satisfying

$$\sum_{k=1}^{K} \beta_k = 1$$

  - **The M step:** For $k = 1, 2, \cdots, K$

$$\alpha_k^{(n+1)}, \tau_k^{(n+1)} = \arg\min_{\alpha, \tau} \int_{T_i}^{T_f} \left| x_k^{(n)} - \alpha s(t - \tau) \right|^2 dt \qquad (4.48)$$

  - $n = n + 1$

Assuming that the observation interval, T, is long compared to the duration of the signal and to the maximum expected delay, the two parameter maximization required in (4.48) may be simplified, and can be carried out explicitly as follows:

$$\tau_k^{(n+1)} = \arg\max_{\tau} |g_k^{(n)}(\tau)| \qquad (4.49)$$

$$\alpha_k^{(n+1)} = \frac{g_k^{(n)}(\tau_k^{(n+1)})}{E} \qquad (4.50)$$

where $E = \int_{T_i}^{T_f} |s(t)|^2 dt$ is the signal energy, and

$$g_k^{(n)}(\tau) = \int_{T_i}^{T_f} x_k^{(n)}(t) s^*(t - \tau) dt \qquad (4.51)$$

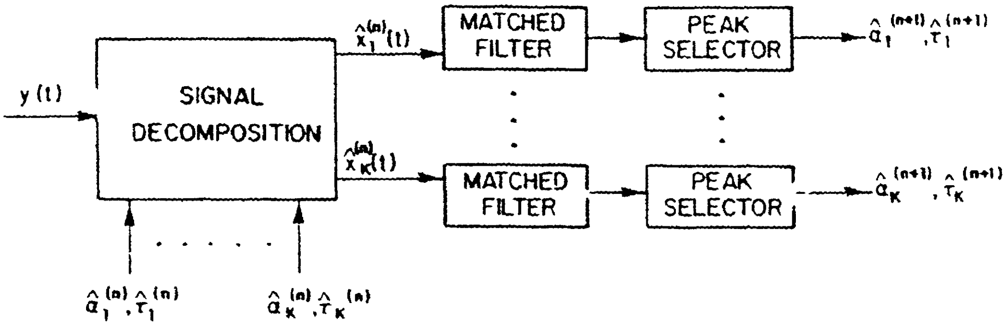

y(t)
SIGNAL DECOMPOSITION
$\hat{x}_1^{(n)}(t)$
$\hat{x}_K^{(n)}(t)$
MATCHED FILTER
MATCHED FILTER
PEAK SELECTOR
PEAK SELECTOR
$\hat{\alpha}_1^{(n+1)}, \hat{\tau}_1^{(n+1)}$
$\hat{\alpha}_K^{(n+1)}, \hat{\tau}_K^{(n+1)}$
$\hat{\alpha}_1^{(n)}, \hat{\tau}_1^{(n)}$
$\hat{\alpha}_K^{(n)}, \hat{\tau}_K^{(n)}$

Figure 4.4: The EM algorithm for multipath time delay estimation

Note that $g_k^{(n)}(\tau)$ can be generated by passing $x_k^{(n)}(t)$ through a filter matched to $s(t)$. The algorithm is illustrated in Figure 4.4. This computationally attractive algorithm iteratively decreases the objective function in (4.46) without ever going through the indicated multi-parameter optimization. The complexity of the algorithm is essentially unaffected by the assumed number of signal paths. As K increases we increase the number of matched filters in parallel; however, each matched filter output is maximized separately.

We note that the algorithm can be extended to the case where the signal waveform, $s(t)$, is unknown. The general EM algorithm steps for the case where the signal waveforms are unknown, are given by equations (4.21) and (4.22). For our problem, the E step is similar to (4.47) where we use the current estimated waveform, $s^{(n)}(t)$, instead of the a-priori given $s(t)$. The M step requires a more complicated maximization with respect to the unknown signal waveform values and the unknown parameters. Using an alternate maximization procedure for the M step, the M step of (4.22) will reduce to,

$$\alpha_k^{(n+1)}, \tau_k^{(n+1)} = \arg\min_{\alpha,\tau} \int_{T_i}^{T_f} \left| x_k^{(n)} - \alpha s^{(n)}(t-\tau) \right|^2 dt, \quad k = 1, \cdots, K \tag{4.52}$$

$$s^{(n+1)}(t) = \frac{\sum_{k=1}^{K} \frac{1}{\beta_k} \alpha_k^{(n+1)} x_k^{(n+1)}(t + \tau_k^{(n+1)})}{\frac{1}{\beta_k}(\alpha_k^{(n+1)})^2} \tag{4.53}$$

We have discussed the unknown signal waveform case in [54], following the considerations above.

For the case in which the number of signals, K, is unknown, several criteria for its determination have been developed in [55] and elsewhere. Usually, these criteria are composed of the likelihood function above and an additional penalty term. Thus, as discussed in section 2.5, these criteria can be incorporated into an EM algorithm, similar to the algorithm above.

4.3.2 Simulation results

To demonstrate the performance of the algorithm, we have considered the following example: The observed signal, $y(t)$, consists of three signal paths in additive noise,

$$y(t) = \sum_{k=1}^{3} \alpha_k s(t - \tau_k) - n(t)$$

where $s(t)$ is a trapezoidal pulse

$$s(t) = \begin{cases} \frac{t}{20} & 0 \leq t < 5 \\ \frac{1}{4} & 5 \leq t < 15 \\ \frac{t-10}{20} & 15 \leq t < 20 \end{cases}$$

The observed data consists of 100 time samples, indexed $-40 \leq t < 60$. The additive noise is spectrally flat with a spectral level of $\sigma^2 = 0.025$, so that the post-integration signal to noise ratio (SNR) is approximately 16 dB.

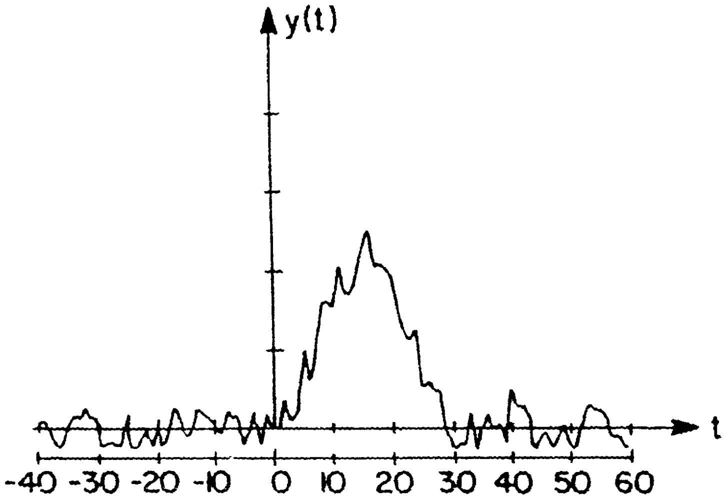


Figure 4.5: The observed data

The actual delays are

$$\tau_1 = 0, \qquad \tau_2 = 5, \qquad \tau_3 = 10$$

and the amplitude scales are

$$\alpha_k = 1, \qquad k = 1,2,3.$$

In Figure 4.5, we have plotted the observed data. In Figure 4.6, we have plotted the matched filter output as a function of the delay. As we can see, the conventional method cannot resolve the various signal paths and estimate their parameters.

First, as a reference, we computed the ML estimates by a direct minimization of the objective function (4.46), using exhaustive search. We obtained,

$$\hat{\tau}_1 = 0.0117 \qquad \hat{\tau}_2 = 5.0031 \qquad \hat{\tau}_3 = 9.9884$$

$$\hat{\alpha}_1 = 1.1511 \qquad \hat{\alpha}_2 = 0.7799 \qquad \hat{\alpha}_3 = 0.9471$$

The value of the objective fnnction at the minimum (corresponding to the value of the

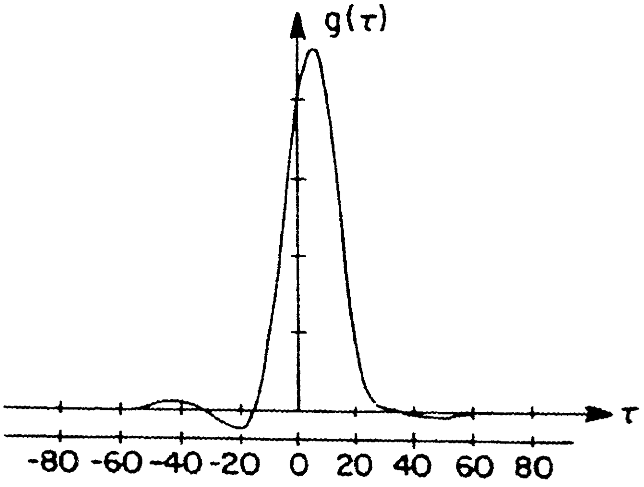


Figure 4.6: The conventional matched filter response

log-likelihood function at the maximum) is,

$$J = 0.45879$$

We also computed lower bounds on the root mean square (r.m.s) error of each parameter, using the Cramer-Rao inequality. We obtained,

$$\sigma(\hat{\tau}_1) = 0.028 \qquad \sigma(\hat{\tau}_2) = 0.030 \qquad \sigma(\hat{\tau}_3) = 0.028$$

($\hat{\tau}$ $\qquad \sigma(\hat{\alpha}_1) = 0.076 \qquad \sigma(\hat{\alpha}_2) = 0.079 \qquad \sigma(\hat{\alpha}_3) = 0.07908$

$_k$) denotes the attainable r.m.s error in the estimate of τ_k, and $\sigma(\hat{\alpha}_k)$ denotes the attainable r.m.s error in the estimate of α_k.

We have applied our algorithm. In Figure 4.7, we have plotted the matched filter response to the various signal paths, as they evolve during the iterations. In addition to this experiment, we have tried this algorithm using several arbitrarily selected starting points; the algorithm has converge, within the Cramer-Rao lower bound, to the ML estimate of all the unknown parameters, after 10 to 15 iterations, regardless of the initial guess.

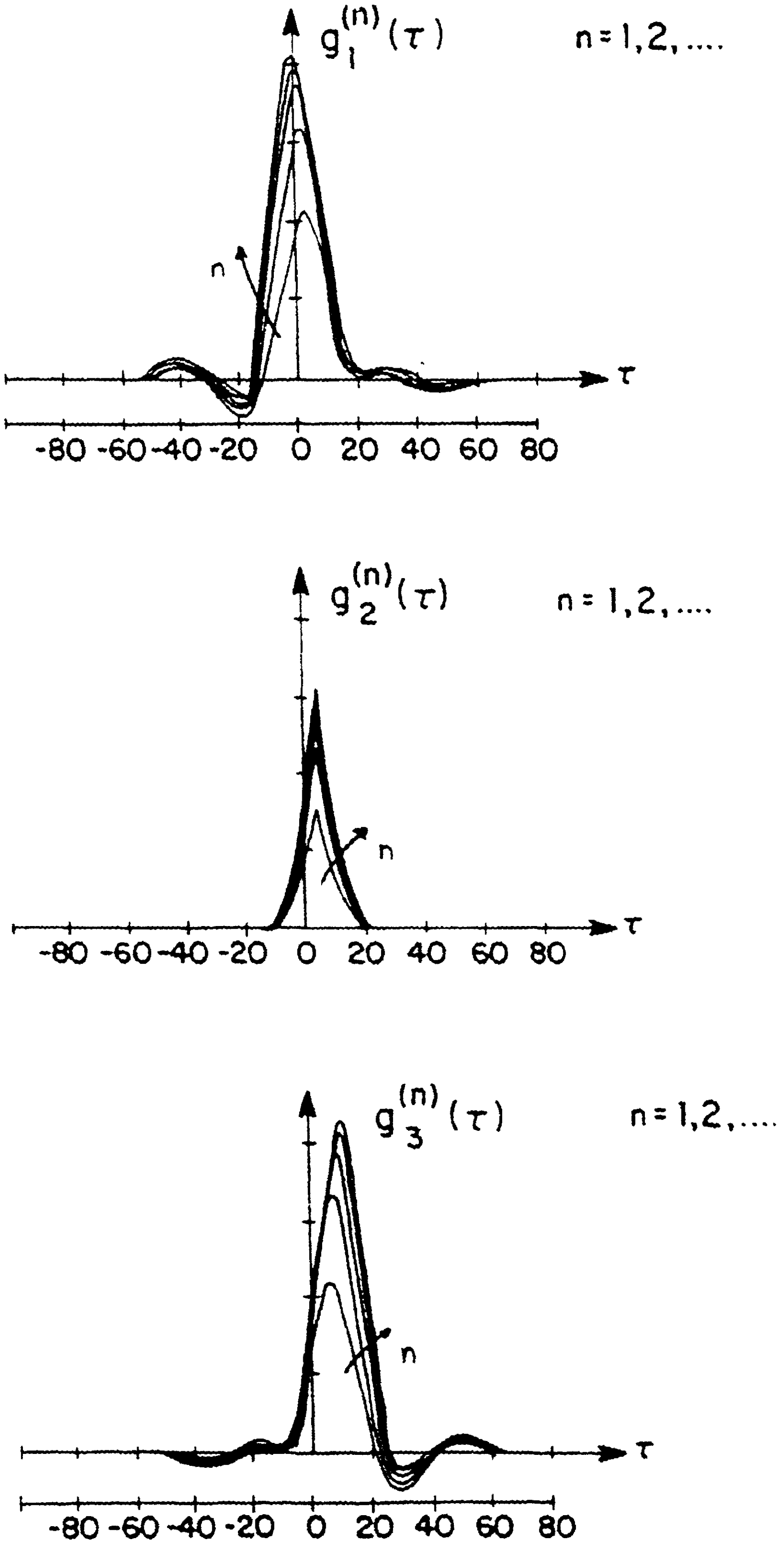


Figure 4.7: The matched filter response to each signal path

Using the asymptotic efficiency and lack of bias of the ML estimates, we can claim with some confidence that the r.m.s error performance of the algorithm is the minimum attainable, characterized by the Cramer-Rao lower bound.

Additional simulation results, which present additional examples of the deterministic multipath time delay estimation, may be found in [52].

4.4 Application to multiple source Direction Of Arrival (DOA) estimation

Passive direction of arrival estimation (DOA) using an array of sensors is a common problem in underwater acoustics, radar and geophysical seismic environments. Using an array of M spatially distributed sensors, the bearing of a source, radiating toward the array, can be determined by estimating the phase differences or the time delays among the signals received in the array sensors.

The standard technique for DOA analysis is known as beamforming. For any given direction, the array signals are delayed and added accordingly, and an output signal is generated. The energy of the output signal is recorded as a function of the direction, and the DOA's estimates correspond to "peaks" of this function. This is an intuitively appealing approach, and indeed, when only a single source exists, the maximum likelihood method, in a variety of modeling assumptions, reduces to maximizing the beamformer output. When several sources exist, this approach is nearly optimal, if the various signal sources are widely separated. However, if the sources are closely spaced this approach is distinctly suboptimal.

The radiating sources generate signals, that may be modeled as deterministic or stochastic. In some radar environments, the targets transmit known waveform pulses which are received by the antenna array of the receiver. Similarly, a seismic pulse received by a sensor array is naturally modeled as a deterministic signal. On the other hand, the acoustic signal, generated by a target and received by a passive sonar array, is a noise-like signal, which is typically modeled as a WSS Gaussian stochastic process.

We will concentrate in this section on the stochastic signal case. We note, however, that methods for multiple source DOA estimation of deterministic signals via the EM algorithm were presented in [56],[57]. We will start by presenting the general mathematical model of the multiple source DOA estimation problem. We will then assume that the signals are Gaussian processes and present the resulting statistical maximum likelihood problem. The solution of this ML problem, using the EM algorithm, will then be presented in detail, and we will describe the simulation results of a specific example.

4.4.1 The passive multiple source DOA estimation problem

We will assume that K spatially distributed sources are radiating signals towards an array of M spatially distributed sensors. Assuming perfect propagation conditions in the medium and ignoring amplitude attenuations of the signal wavefront across the array, the actual waveform observed at the m^{th} sensor output is

$$y_m(t) = \sum_{k=1}^{K} s_k(t - \tau_{km}) + n_m(t) \qquad m = 1, 2, \cdots, M \tag{4.54}$$

where $s_k(t)$ is the k^{th} source signal, $n_m(t)$ is the additive noise at the m^{th} sensor output, and τ_{km} is the travel time of the signal wavefront from the k^{th} source to the m^{th} sensor.

Information concerning the various source location parameters can be extracted by measuring the various τ_{km}. In the passive case, one can only measure the travel time differences. Let the M^{th} sensor be the reference sensor, and set $\tau_{kM} = 0$, then τ_{km} measures the travel time difference of the k^{th} signal wavefront between the (m, M) sensor pair.

We assume that the various signal sources are relatively far-field, so that the observed signal wavefronts are essentially planar across the array. In this case, the unknown source location parameters are their bearings or directions of arrival. To simplify the exposition, we suppose further that the array sensors are co-linear. Then,

$$\tau_{km} = \frac{d_m}{c} \sin\theta_k \tag{4.55}$$

where d_m is the spacing between the sensor m and the reference sensor M, c is the velocity of propagation in the medium, and θ_k is the angle of arrival of the k^{th} signal wavefront relative to the boresight.

Substituting (4.55) into (4.54) and concatenating the various equations, we obtain

$$\underline{y}(t) = \sum_{k=1}^{K} \underline{s}(t;\underline{\theta}_k) + \underline{n}(t) \tag{4.56}$$

where

$$\underline{s}_k(t;\underline{\theta}_k) = \begin{bmatrix} s_k(t - \gamma_1 \sin\theta_k) \\ s_k(t - \gamma_2 \sin\theta_k) \\ \cdot \\ \cdot \\ s_k(t - \gamma_{M-1} \sin\theta_k) \\ s_k(t) \end{bmatrix} \tag{4.57}$$

and $\gamma_m = d_m/c$. We note that this is a special case of the superimposed signal problem of (4.1).

A statistical ML problem for estimating the unknown directions of arrival is achieved by a further statistical modeling of the various signals in (4.56). We will now present the

ML problem and its solution using the EM algorithm, where we model the signals, $\{s_k(t)\}$, and the noise as Gaussian processes.

4.4.2 The Gaussian case

Suppose that the various $s_k(t)$ and the various $n_m(t)$ are mutually independent, WSS, zero-mean Gaussian processes with spectral densities $S_k(\omega)$ and $N_m(\omega)$ respectively. We will also assume that the observation time, $T = T_f - T_i$, is long compared with the correlation time (inverse bandwidth) of the signals and the noises. Under these assumptions we may write the likelihood of the observations, $\underline{y}(t)$, in the frequency domain; the ML estimates of $\theta_1, \theta_2, \cdots, \theta_k$ will be achieved by, (see Eq. (4.33)),

$$\min_{\theta_1,\theta_2,\cdots,\theta_K} \sum_{\ell} \left[\log\det P(\omega_\ell;\underline{\theta}) + \underline{Y}^\dagger(\omega_\ell)\cdot P^{-1}(\omega_\ell;\underline{\theta})\cdot \underline{Y}(\omega_\ell)\right] \tag{4.58}$$

where $\underline{Y}(\omega_\ell)$ are the Fourier transform coefficients (or the DFT coefficients in the discrete case) of $\underline{y}(t)$ and

$$P(\omega;\underline{\theta}) = \sum_{k=1}^{K} S_k(\omega)\underline{V}(\omega;\theta_k)\underline{V}^\dagger(\omega;\theta_k) + N(\omega) \tag{4.59}$$

where

$$\underline{V}(\omega;\theta_k) = \begin{bmatrix} e^{-j\omega\tau_1\sin\theta_k} \\ e^{-j\omega\tau_2\sin\theta_k} \\ \cdot \\ \cdot \\ e^{-j\omega\tau_{M-1}\sin\theta_k} \\ 1 \end{bmatrix} \tag{4.60}$$

and $N(\omega)$ is the diagonal matrix

$$N(\omega;\underline{\theta}) = \begin{bmatrix} N_1(\omega) & & & & \\ & N_2(\omega) & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & N_M(\omega) \end{bmatrix} \tag{4.61}$$

Again as in the previous examples, the resulting ML problem requires the solution of a complicated multi-parameter optimization problem in several unknowns. Consequently, numerous ad-hoc solutions and sub-optimal approaches have been proposed in the recent literature, e.g. [58,59,60,61,62], to by-pass this ML problem. Still, the most common approach consists of beamforming and searching for the K highest peaks. As noted above, if the various sources are widely separated, this approach is nearly optimal. However, when the sources are closely spaced we are likely to obtain poor estimates of the various DOA's.

Identifying the model in (4.56) as a special case of the superimposed signal in noise case, the algorithm specified by (4.43) and (4.44) is directly applicable, where

$$\Lambda_k(\omega;\underline{\theta}_k) = S_k(\omega)\underline{V}(\omega;\theta_k)\underline{V}^{\dagger}(\omega;\theta_k) + \beta_k N(\omega) \tag{4.62}$$

This special form of the matrix Λ_k allows the following simplifications. We may write

$$\det \Lambda_k(\omega;\underline{\theta}_k) = \left[1 + \frac{1}{\beta_k} S_k(\omega)\underline{V}^{\dagger}(\omega;\theta_k)N^{-1}(\omega)\underline{V}(\omega;\theta_k)\right] \cdot \beta_k \det N(\omega) \tag{4.63}$$

and

$$\Lambda_k^{-1}(\omega;\underline{\theta}_k) = \frac{1}{\beta_k}\left[N^{-1}(\omega) - \frac{S_k(\omega)}{\beta_k + S_k(\omega)\underline{V}^{\dagger}(\omega;\theta_k)N^{-1}(\omega)\underline{V}(\omega;\theta_k)} N^{-1}(\omega)\underline{V}(\omega;\theta_k)\underline{V}^{\dagger}(\omega;\theta_k)N^{-1}(\omega)\right] \tag{4.64}$$

Substituting (4.63) and (4.64) into (4.44) and ignoring the terms that are independent of θ_k, the M step of the algorithm will be simplified. The resulting EM algorithm is:

- Start, $n = 0$, initialize $\theta_k^{(0)}$, $k = 1, \cdots, K$.

- Iterate (until some convergence criterion is met)

  - **The E step:** For $k = 1, 2, \cdots, K$ compute

$$\begin{aligned}\Psi_k^{(n)}(\omega_\ell) &= \Lambda_k(\omega_\ell; \theta_k^{(n)}) - \Lambda_k(\omega_\ell; \theta_k^{(n)}) P^{-1}(\omega_\ell; \theta^{(n)}) \Lambda_k(\omega_\ell; \theta_k^{(n)}) + \\ &\quad + \Lambda_k(\omega_\ell; \theta_k^{(n)}) P^{-1}(\omega_\ell; \theta^{(n)}) \underline{Y}(\omega_\ell) \cdot \underline{Y}^\dagger(\omega_\ell) P^{-1}(\omega_\ell; \theta^{(n)}) \Lambda_k(\omega_\ell; \theta_k^{(n)})\end{aligned} \tag{4.65}$$

  - **The M step:** For $k = 1, 2, \cdots, K$

$$\theta_k^{(n+1)} = \arg\max_{\underline{\theta}} \sum_\ell F_k(\omega_\ell) \cdot \underline{V}^\dagger(\omega_\ell; \theta) N^{-1}(\omega) \cdot \Psi_k^{(n)}(\omega_\ell) \cdot N^{-1}(\omega) \underline{V}(\omega_\ell; \theta) \tag{4.66}$$

    where $F_k(\omega)$ is a shaping filter, given by,

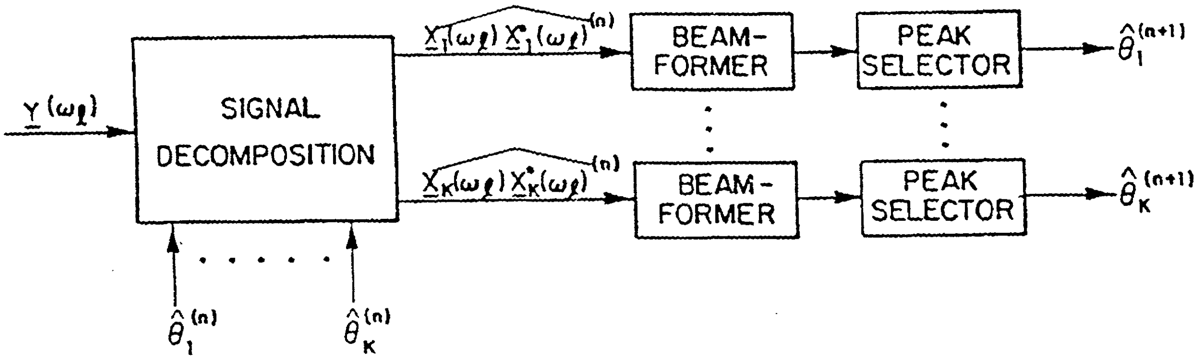
$$F_k(\omega) = \frac{S_k(\omega)}{\beta_k + S_k(\omega) \sum_{m=1}^{M} 1/N_m(\omega)} \tag{4.67}$$

  - $n = n + 1$

We note that the objective function in (4.66) is the array beamformer, where the product $\underline{X}_k(\omega_\ell)\underline{X}_k^\dagger(\omega_\ell)$ is substituted by its current estimate, $\Psi_k^{(n)}(\omega_\ell)$. The algorithm is illustrated in Figure 4.8. This computationally attractive algorithm iteratively decreases the objective function in (4.58) without ever going through the indicated multi-parameter optimization. Again, the complexity of the algorithm is essentially unaffected by the assumed number of signals sources. As K increases, we have to increase the number of beamformers in parallel; however, each beamformer output is maximized separately.

4.4.3 Simulation results

To demonstrate the performance of the algorithm, we have considered the following example. Our array of sensors consists of five, co-linear, evenly-spaced sensors. There are



$\underline{Y}(\omega_\ell)$
SIGNAL DECOMPOSITION
$\widehat{\underline{X}_1(\omega_\ell)\underline{X}_1^*(\omega_\ell)}^{(n)}$
BEAM-FORMER
PEAK SELECTOR
$\hat{\theta}_1^{(n+1)}$
$\widehat{\underline{X}_K(\omega_\ell)\underline{X}_K^*(\omega_\ell)}^{(n)}$
BEAM-FORMER
PEAK SELECTOR
$\hat{\theta}_K^{(n+1)}$
$\hat{\theta}_1^{(n)}$
$\hat{\theta}_K^{(n)}$

Figure 4.8: The EM algorithm for multiple source DOA estimation

two far-field signal sources at the bearings

$$\theta_1 = 0^\circ \qquad \theta_2 = 10^\circ$$

relative to the boresight. The array-source geometry is shown in Figure 4.9. The signals and the noises are spectrally flat with $S_k(\omega) = S$ and $N_m(\omega) = N$, over the frequency band $[-W/2,\ W/2]$. We assume that $S/N = 1$, and that $WT/2\pi = 20$ (so that the post integration SNR per channel is approximately 23dB). The array length is taken to be $L = 6\lambda$ where λ is the wavelength associated with the highest frequency in the signal band.

In Figure 4.10, we have plotted the array beamformer response as a function of the bearing angle. As we may see, the conventional beamformer cannot resolve the signal sources and thus cannot estimate their bearings.

The ML estimates, obtained by direct minimization using exhaustive search of the ob-

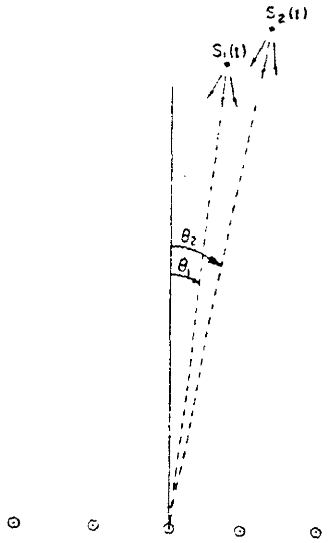


Figure 4.9: Array-Source geometry

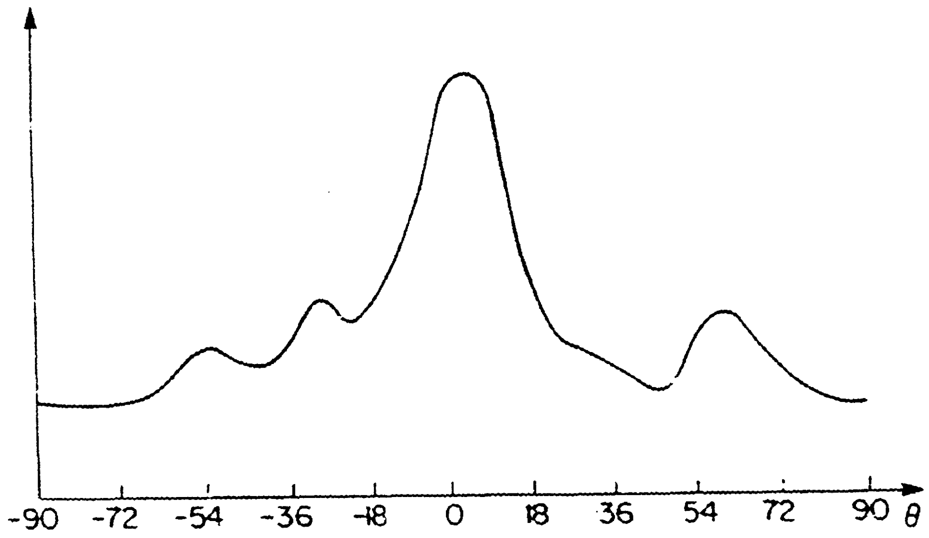


Figure 4.10: The conventional Beamformer

jective function in (4.58), are

$$\hat{\theta}_1 = -0.0563 \qquad \hat{\theta}_2 = 10.4556$$

The value of the objective function at the minimum (corresponding to the value of the log-likelihood function at the maximum) is

$$J = 159.0137$$

We have also computed the Cramer-Rao bound on the r.m.s error of each parameter estimate. We obtain

$$\sigma(\hat{\theta}_1) = 0.2680 \qquad \sigma(\hat{\theta}_2) = 0.2722$$

We now apply our algorithm. In Figure 4.11, we have plotted the beamformer response to the various signal sources as they evolve with iterations. In Figure 4.12, we have tabulated the results using several arbitrarily selected initial guesses. We see that in all cases, after 5 to 10 iterations, the algorithm essentially converges, within the Cramer-Rao lower bound, to the ML estimates of all unknown bearing parameters simultaneously; therefore the various signal sources are correctly resolved.

4.5 Sequential and adaptive algorithms

Sequential and adaptive algorithms for estimating the parameters of superimposed signals in noise, based on the EM algorithm, may be suggested following the consideration of chapter 3. As discussed in chapter 3, in general, any given iterative batch EM algorithm may be transformed into a sequential algorithm using the stochastic approximation idea. The EM algorithms suggested in this chapter for both the deterministic case and the stochastic case have a structure that may support recursive E and M steps. However, we will concentrate

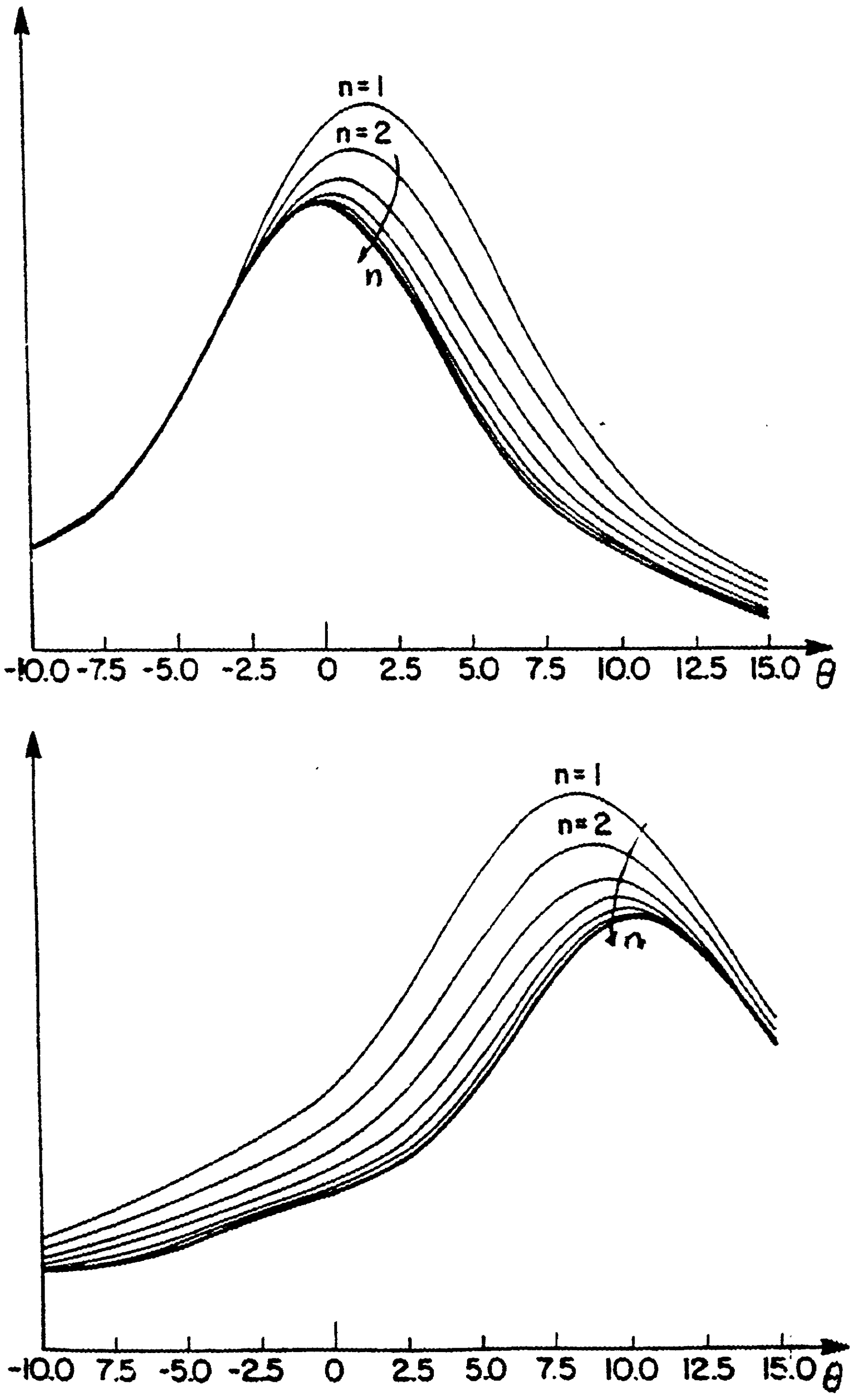


Figure 4.11: The Beamformer response to each signal source

n	DOA 1	DOA 2	J	n	DOA 1	DOA 2	J
0	2.000	8.000	269.66	0	-5.000	13.000	520.29
1	1.577	8.465	241.12	1	-2.469	11.949	294.71
2	1.122	8.974	210.30	2	-1.263	11.145	196.55
3	0.693	9.457	184.13	3	-0.674	10.716	167.77
4	0.372	9.859	168.36	4	-0.379	10.529	161.05
5	0.157	10.127	161.74	5	-0.245	10.429	159.61
6	0.050	10.288	159.72	6	-0.164	10.421	159.21
7	-0.003	10.368	159.19	7	-0.111	10.421	159.07
8	-0.030	10.421	159.04	8	-0.084	10.421	159.04

n	DOA 1	DOA 2	J	n	DOA 1	DOA 2	J
0	4.000	7.000	378.57	0	7.000	13.000	497.05
1	3.641	7.259	351.24	1	6.053	12.699	448.01
2	3.293	7.527	328.51	2	4.954	12.378	385.83
3	2.918	7.822	306.11	3	3.829	12.083	318.25
4	2.489	8.144	281.83	4	2.783	11.842	256.99
5	2.033	8.519	254.93	5	1.953	11.627	214.14
6	1.524	8.948	224.28	6	1.363	11.413	189.18
7	1.015	9.403	194.46	7	0.934	11.225	175.12
8	0.586	9.805	173.51	8	0.639	11.038	167.40
9	0.291	10.100	163.52	9	0.452	10.904	163.66
10	0.131	10.261	160.36	10	0.318	10.797	161.58
11	0.050	10.368	159.36	11	0.211	10.716	160.39
12	-0.003	10.421	159.07	12	0.130	10.637	159.66
13	-0.030	10.449	159.02	13	0.077	10.582	159.33

Figure 4.12: Tables of results for multiple source DOA estimation

in this section on the deterministic superimposed signals problem, for which we were able to develop sequential EM algorithms based on exact EM mapping. These algorithms will be presented explicitly, and their application to solving the problem of multiple sinusoids in noise will be described.

4.5.1 Sequential algorithms based on exact EM mapping for the Deterministic case

Sequential EM algorithms, based on exact EM mapping, are achieved by examining the expression of the EM iteration, which depends, in general, on the entire past observations, and recognizing the terms that depend on the current data and the terms that depend only on past data. Hopefully, the terms that depend on past data may be summarized into a compact form, that will be subsequently updated and recorded. Based on these recorded quantities and the new measurements, the parameters will be updated using an exact EM iteration.

Thus, let us consider the EM iteration for the deterministic signal case, given by equations (4.19) and (4.20). We will assume that the signals are discrete so that the integral in (4.20) is replaced by a sum. Assume that we observe $y(1), \cdots, y(n)$, i.e. the observation index is $t = 1, \cdots, n$. The E and M steps of this EM algorithm are given by,

- **The E step:** For $k = 1, 2, \cdots, K$ and $t = 1, \cdots, n$ compute

$$\underline{x}_k^{(n)}(t) = \underline{s}_k(t; \underline{\theta}_k^{(n)}) + \beta_k \left[\underline{y}(t) - \sum_{\ell=1}^{K} \underline{s}_\ell(t; \underline{\theta}_\ell^{(n)}) \right] \tag{4.68}$$

- **The M step:** For $k = 1, 2, \cdots, K$

$$\underline{\theta}_k^{(n+1)} = \arg \min_{\underline{\theta}_k} \sum_{t=1}^{n} \left[\underline{x}_k^{(n)} - \underline{s}_k(t; \underline{\theta}_k) \right]^{\dagger} Q_k^{-1} \left[\underline{x}_k^{(n)} - \underline{s}_k(t; \underline{\theta}_k) \right] \tag{4.69}$$

We will assume that $Q_k = \gamma_k I$. Combining the E and M steps, and ignoring the terms that are independent of $\underline{\theta}$, we may write the EM iteration as follows. For $k = 1, 2, \cdots, K$

$$\underline{\theta}_k^{(n+1)} = \arg\min_{\underline{\theta}_k} \quad \sum_{t=1}^{n} |\underline{s}_k(t;\underline{\theta}_k)|^2 + \beta_k \sum_{t=1}^{n} \cdot 2Re \sum_{\ell=1}^{K} \underline{s}_\ell(t;\underline{\theta}_\ell^{(n)})^\dagger \underline{s}_k(t;\underline{\theta}_k) - \\ - \sum_{t=1}^{n} \underline{s}_k(t;\underline{\theta}_k^{(n)})^\dagger \underline{s}_k(t;\underline{\theta}_k) - \beta_k \sum_{t=1}^{n} \cdot 2Re[\underline{y}(t)^\dagger \underline{s}_k(t;\underline{\theta}_k)] \tag{4.70}$$

We notice that the term that depends on the observations, $\underline{y}(t)$, is the cross-correlation between $\underline{y}(t)$ and the various signals, $\underline{s}_k(t;\underline{\theta}_k)$. We will denote this term by,

$$p_n(\underline{\theta}_k) = \sum_{t=1}^{n} \underline{y}(t)^\dagger \underline{s}_k(t;\underline{\theta}_k) \tag{4.71}$$

Suppose we record $p_n(\underline{\theta}_k)$. At time $n+1$, when a new measurement, $\underline{y}(n+1)$, arrives, this term may be updated recursively as,

$$p_{n+1}(\underline{\theta}_k) = p_n(\underline{\theta}_k) + \underline{y}(n+1)^\dagger \underline{s}_k(n+1;\underline{\theta}_k) \tag{4.72}$$

The other terms depend only on the a-priori given waveforms, $\{\underline{s}_k(t;\underline{\theta}_k)\}$. In many cases the expressions

$$R_n(\underline{\theta}_\ell, \underline{\theta}_k) = \sum_{t=1}^{n} \underline{s}_\ell(t;\underline{\theta}_\ell)^\dagger \underline{s}_k(t;\underline{\theta}_k) \tag{4.73}$$

may be given for each n by an a-priori analytic formula. However, even if the algorithm needs to calculate these terms explicitly, they may be calculated recursively using the following formula,

$$R_{n+1}(\underline{\theta}_\ell, \underline{\theta}_k) = R_n(\underline{\theta}_\ell, \underline{\theta}_k) + \underline{s}_\ell(n+1;\underline{\theta}_\ell)^\dagger \underline{s}_k(n+1;\underline{\theta}_k) \tag{4.74}$$

The term

$$E_n(\underline{\theta}_k) = \sum_{t=1}^{n} |\underline{s}_k(t;\underline{\theta}_k)|^2 \tag{4.75}$$

which represent the total energy of the signals $\underline{s}_k(t)$ may also be calculated recursively or may be given by an analytic formula. In many cases this energy term takes the form of

$$E_n(\underline{\theta}_k) = n \cdot |a_k|^2 \tag{4.76}$$

i.e. it depends only on the amplitude parameter a_k, and it is independent of the other parameters,

Thus, the sequential EM algorithm for superimposed, deterministic signals is given by:

- Start, $n = 0$: Guess $\underline{\theta}^{(0)}$. Initial $p_0 = R_0 = E_0 = 0$.
- While data is observed
  - Update the p_{n+1}'s by (4.72), the R_{n+1}'s by (4.74) and the E_{n+1}'s by (4.75).
  - Update the parameters: For $k = 1, 2, \cdots, K$,

$$\underline{\theta}_k^{(n+1)} = \arg\min_{\underline{\theta}_k} \left[E_{n+1}(\underline{\theta}_k) + \beta_k \cdot 2Re \sum_{\ell=1}^{K} R_{n+1}(\underline{\theta}_\ell^{(n)}, \underline{\theta}_k) - R_{n+1}(\underline{\theta}_k^{(n)}, \underline{\theta}_k) - \beta_k \cdot p_{n+1}(\underline{\theta}_k) \right] \tag{4.77}$$

  - Store $p_{n+1}, R_{n+1}, E_{n+1}$.
  - $n = n + 1$.

We note that, as in any exact mapping sequential EM algorithm, we can perform few iterations for each observed data point. The advantage is that we have to update the quantities p_n, R_n and E_n only once for each new measurement. Sometimes it will be more efficient to perform few more iterations before moving to the new data. However, in other cases, exhausting the previous data cannot improve the parameters; it is more efficient to proceed and add the new data points.

4.5.2 Application to sum of sinusoids in noise

The sequential EM algorithm above has been applied to the following problem. Let the observed signal be the sum of complex exponential in noise, i.e.

$$y(t) = \sum_{k=1}^{K} a_k e^{j\omega_k t} + n(t) \tag{4.78}$$

where $n(t)$ is a white Gaussian noise with variance σ^2. The unknown parameters are the frequencies of the complex exponentials, $\{\omega_k\}$, and the complex amplitudes, $\{a_k\}$. We note that this problem is essentially the problem of sinusoids in noise, and in this case, we write $a_k = r_k e^{j\phi_k}$, where the $\{\phi_k\}$'s are the unknown phases of the sinusoids. We will assume that the observations are given at time points $t = 0, 1, \cdots, n, \cdots$.

The complete data for this deterministic superimposed signal example is the set of signals, $\{x_k(t)\}_{k=1}^{K}$, where

$$x_k(t) = a_k e^{j\omega t} + n_k(t) \tag{4.79}$$

and $\{n_k(t)\}_{k=1}^{K}$ are independent white noise signals whose sum is $n(t)$. Each $n_k(t)$ has a variance $\beta_k \sigma^2$.

In this application the sequential EM algorithm presented above is further simplified. We first note that the M step of an EM iteration in this case requires solving the following maximization problems, for $k = 1, \cdots, K$

$$\omega_k^{(n+1)} = \arg\max_{\omega} \left|\sum_{t=0}^{n} x_k^{(n)}(t) e^{-j\omega t}\right|^2 = \left|X_k^{(n)}(\omega)\right|^2 \tag{4.80}$$

where $X_k^{(n)}(\omega)$ is the Fourier Transform of the signal, $x_k^{(n)}(t)$, estimated in the E step. The amplitude coefficients may be found either as implied by the EM iteration,

$$a_k^{(n+1)} = \frac{1}{n+1} \sum_{t=0}^{n} x_k^{(n)}(t) e^{-j\omega_k^{(n)} t} \tag{4.81}$$

or by solving a linear least squares problem, noticing that, given $\{\omega_k\}$, determining $\{a_k\}$ is the solution to:

$$a_1^{(n+1)}, \cdots, a_k^{(n+1)} = \arg \min_{a_1, \cdots, a_k} \sum_{t=0}^{n} \left[y(t) - \sum_{k=1}^{K} a_k e^{-\jmath \omega_k^{(n)} t} \right]^2 \tag{4.82}$$

Another simplification comes from the fact that $R_n(\underline{\theta}_1, \underline{\theta}_2)$ may be written analytically as,

$$R_n(a_1, \omega_1, a_2, \omega_2) = \sum_{t=0}^{n} a_1^{*} e^{-\jmath \omega_1 t} a_2 e^{\jmath \omega_2 t} = a_1^{*} a_2 \text{sinc}_{n+1}^{*}(\omega_1 - \omega_2) \tag{4.83}$$

where

$$\text{sinc}_{n+1}^{*}(\omega) = \frac{\sin(\frac{(\omega)(n+1)}{2})}{\sin(\frac{(\omega)}{2})} e^{-\jmath \omega N/2} \tag{4.84}$$

Thus, a step of the sequential EM algorithm for this problem, observing a new measurement, $y(n)$, is given as follows:

- Update the Fourier Transform of the observations, i.e.

$$Y_{n+1}(\omega) = Y_n(\omega) + y(n) e^{-\jmath \omega n} \tag{4.85}$$

- Update the estimates of the frequencies: For $k = 1, \cdots, K$

$$\begin{aligned} \omega_k^{(n+1)} &= \arg \max_{\omega} \left| a_k^{(n)} \text{sinc}_{n+1}^{*}(\omega_k^{(n)} - \omega) - \beta_k \cdot \sum_{\ell=1}^{K} a_\ell^{(n)} \text{sinc}_{n+1}^{*}(\omega_\ell^{(n)} - \omega) + Y_{n+1}(\omega) \right|^2 \\ &= \arg \max_{\omega} \left| X_k^{(n)}(\omega) \right|^2 \end{aligned} \tag{4.86}$$

- Update the amplitudes

  - either by an EM iteration,

$$a_k^{(n+1)} = \frac{1}{n+1} \left| X_k^{(n)}(\omega_k^{(n+1)}) \right|^2 \qquad k = 1, \cdots, K \tag{4.87}$$

- or by a solving the least squares problem of (4.82), i.e.

$$\underline{a}^{(n+1)} = S^{-1}\underline{Y}_{n+1} \tag{4.88}$$

where $\underline{a}$ is the vector of the amplitudes, S is the matrix whose (k,ℓ) element is given by

$$S_{k,\ell} = \text{sinc}^{*}_{n+1}(\omega_k^{(n+1)} - \omega_\ell^{(n+1)}) \tag{4.89}$$

and $\underline{Y}_{n+1}$ is a vector whose κ^{th} component is $Y_{n+1}(\omega_k^{(n+1)})$.

Numerical simulation example

This algorithm has been tested using the following example. The sequentially arriving observed signal, $y(t)$, is complex and discrete; it consists of three complex exponentials in additive white noise, i.e.

$$y(t) = \sum_{k=1}^{3} a_k e^{j\omega_k t} + n(t), \qquad t = 0, 1, \cdots \tag{4.90}$$

The additive noise is spectrally flat with spectral level $\sigma^2 = 0.1$. The normalized frequencies of the complex exponentials were chosen to be,

$$\omega_1 = 0.025, \qquad \omega_2 = 0.03, \qquad \omega_3 = 0.04$$

The magnitude of the complex amplitudes were chosen to be uniformly 1, and their phases chosen as,

$$\phi_1 = 0, \qquad \phi_2 = \pi/6, \qquad \phi_3 = \pi/4$$

We have tested the algorithm given by (4.86) and (4.88), sequentially using 500 data points. A single EM iteration has been performed for each new data point. In Figure 4.13 we have tabulated the estimates of the frequencies as a function of time. We notice that

Time	Source 1	Source 2	Source 3	Time	Source 1	Source 2	Source 3
10	0.016	0.047	0.048	70	0.023	0.043	0.046
20	0.023	0.047	0.048	80	0.024	0.042	0.045
30	0.023	0.046	0.048	90	0.024	0.042	0.045
40	0.022	0.047	0.048	100	0.024	0.042	0.044
50	0.022	0.046	0.048	110	0.024	0.041	0.030
60	0.023	0.043	0.04	120	0.025	0.040	0.030

Figure 4.13: Frequency estimates as a function of time

this efficient sequential algorithm correctly estimates the various frequencies after observing 120 data points. This data record is shorter than the record needed to correctly resolve these sinusoids, using the standard spectral estimation methods.

Chapter 5

Maximum likelihood noise cancellation

The problem of noise cancellation in single and multiple microphone environments has been extensively studied [63]. The performance of the various techniques in the single microphone case seems to be limited. However, enhancement systems with two or more microphones have been more successful due to the availability of reference signals.

In this chapter, noise cancellation, based on a two sensor scenario as indicated in Figure 5.1, is considered. One sensor (the primary microphone) measures a signal that consists of speech with noise. The second sensor (the reference microphone), located away from the speaker, measures a signal that consists mainly of the noise. The signal measured in the reference microphone is used to cancel the noise in the primary microphone. A reasonably general model for this scenario is shown in Figure 5.2.

The most widely used approach to noise cancellation, based on two microphones, was suggested by Widrow et al. [10]. In this approach, it is assumed that the system B is

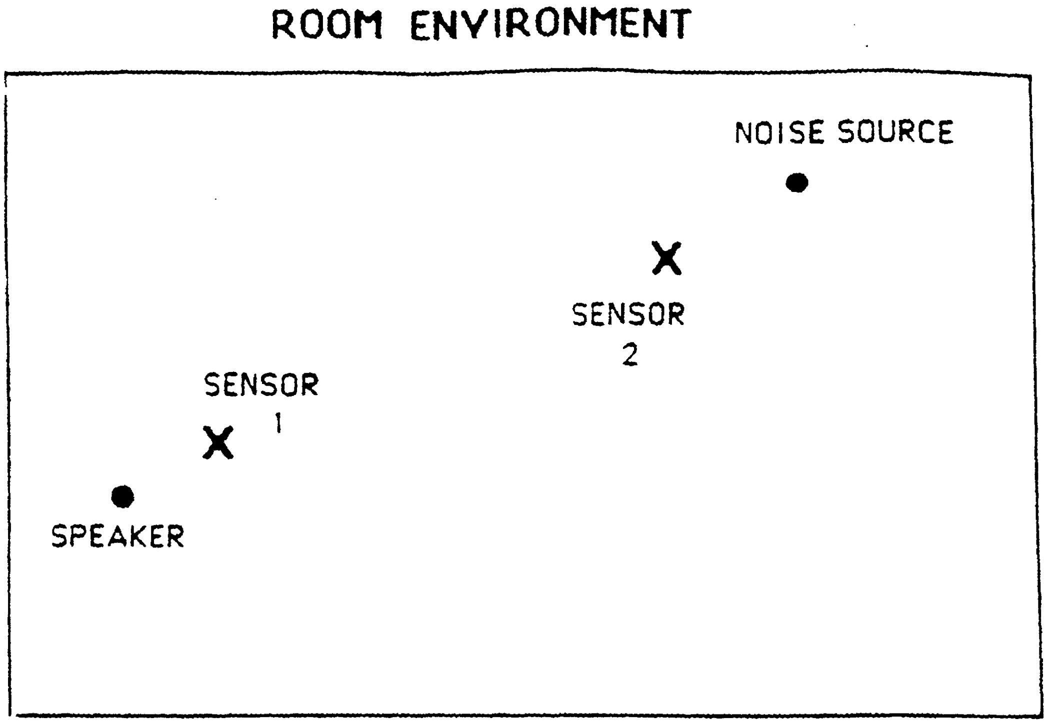


Figure 5.1: The acoustic environment

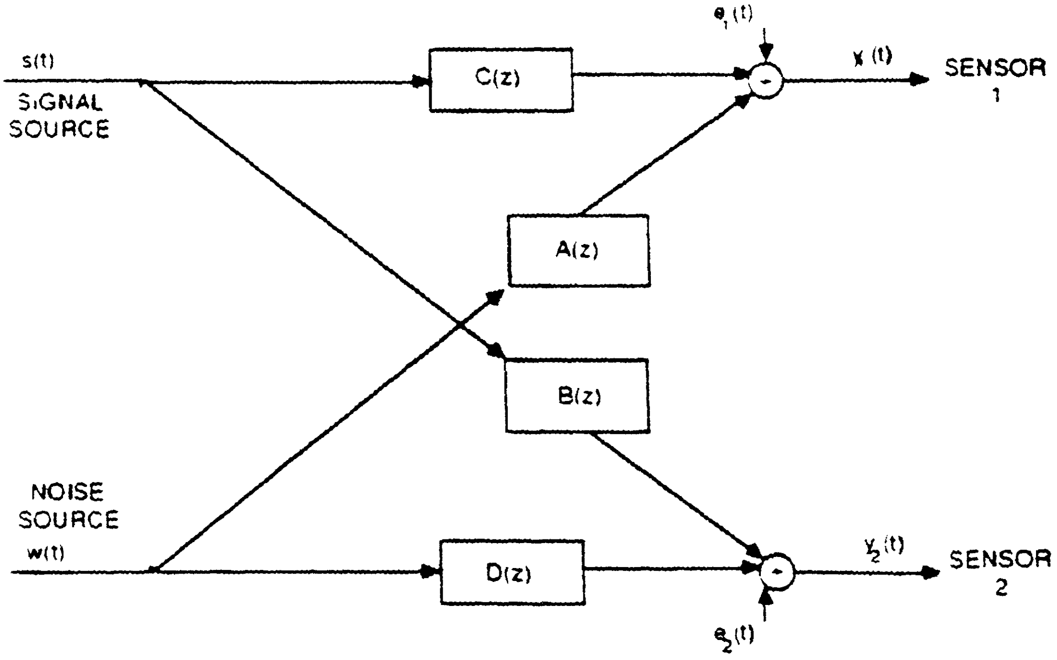


Figure 5.2: The noise cancellation problem

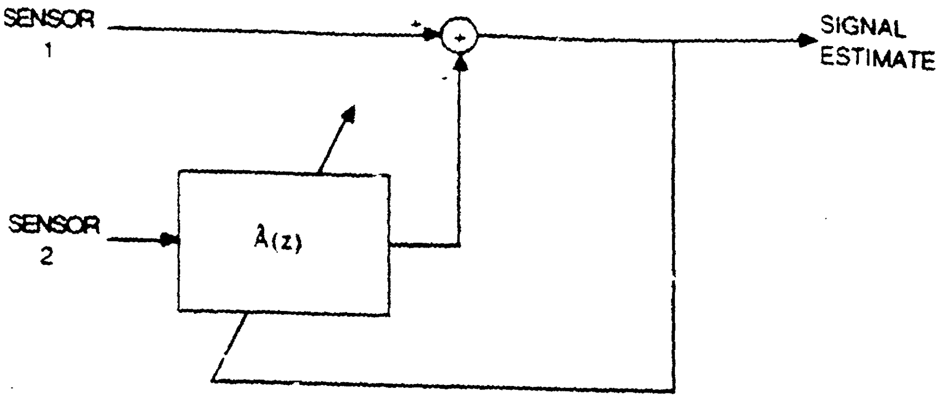


Figure 5.3: "Classical" noise canceling scheme

zero and that C and D are identity, so that the output of the reference microphone is due only to the noise, and that the noise component in the primary microphone is the output of an unknown linear system with transfer function $A(z)$, whose input is the signal measured in the reference microphone. The coefficients of the impulse response of this system are estimated by a least-squares fitting of the reference microphone signal to the primary microphone signal. This method will be referred to later in this chapter as the least-squares method.

Widrow et. al proposed an adaptive solution to this least-squares problem, based on the LMS algorithm. This approach, illustrated in Figure 5.3, has been applied in a speech enhancement context, e.g. [64] and [65]. Adaptive algorithms based on the RLS algorithm also exist, e.g. [66] and [67].

A major limitation of the least-squares method, especially when the reference signal is correlated with the desired (speech) signal, is that a portion of the desired signal may be canceled together with the noise. Since the desired signal may be canceled with some time delay, the resulting effect is to introduce a reverberant distortion in the output.

Our approach consists of formulating the problem as a statistical maximum likelihood problem. This approach will allow us to consider a more general model, that includes the effect of "cross talk", i.e. the coupling of the desired signal into the reference microphone. As in many examples throughout this thesis, solving the resulting ML problem directly is difficult, and so it is solved using the EM method. The proposed algorithm iterates between estimating the speech and the noise source signals (E step) and solving a set of linear equations for the coefficients of the acoustic impulse response (M step).

It is interesting to note that the proposed algorithm is similar to the iterative speech enhancement method for single microphone suggested in [3]. As already noted, the iterative Wiener filter used in [3] is an instance of the EM algorithm. In that respect, the procedures developed in this chapter, may be considered as extensions of the method in [3] to two microphones.

The methods presented in this chapter, can be used as an alternative to the least-squares method of [10] and its derivatives, e.g. [68] and [69]. Simulation results indicate that the proposed schemes tend to eliminate the reverberant distortion encountered in the least-squares method. Adaptive versions of the proposed algorithms are also possible. We finally note that the proposed scheme can easily be extended to the more general, multiple microphone case.

This chapter is organized as follows. In section 5.1, we develop the general maximum likelihood formulation of the noise cancellation problem. In section 5.2, we apply an EM algorithm to solve the ML problem in a simplified scenario, that basically makes the same assumptions as in [10]. We then describe, in section 5.3, the EM algorithm for a more general scenario that includes "cross talk". We conclude this chapter, in section 5.4, by presenting several simulation results including some that use a simulated realistic room impulse response.

5.1 Maximum likelihood formulation of the two-sensor noise cancellation problem

The mathematical ML formulation, encountered in a two-microphone noise cancellation problem, is based on the following scenario. A desired (speech) signal source and a noise source both exist in some acoustic environment, say a living room or an office. We have two microphones used in such a way that one microphone is intended to measure mainly the speech source, while the other is intended to measure mainly the noise source.

The desired signal and the noise are both coupled into each microphone by the acoustic field in this environment. This situation is illustrated in Figure 5.2, and is represented by the equations [1].

$$y_1(t) = C\{s(t)\} + A\{w(t)\} + e_1(t) \tag{5.1}$$

$$y_2(t) = B\{s(t)\} + D\{w(t)\} + e_2(t) \tag{5.2}$$

where $s(t)$ denotes the desired (speech) signal and $w(t)$ denotes the noise source signal. The systems A, B, C and D are assumed to be linear systems, representing the acoustic transfer functions between the sources and the microphones. We will assume that these systems are time invariant in our analysis window. The additional noise sources $e_1(t)$ and $e_2(t)$ are included to represent modeling errors, microphone and measurements noise etc.

[1] The mathematics and the algorithms will be formulated in terms of discrete time signals with the independent variable t representing normalized sample time

Under these assumptions we may write the observed signals in the frequency domain as,

$$Y_1(\omega) = C(\omega)S(\omega) + A(\omega)W(\omega) + E_1(\omega) \tag{5.3}$$

$$Y_2(\omega) = B(\omega)S(\omega) + D(\omega)W(\omega) + E_2(\omega) \tag{5.4}$$

where $Y_1(\omega)$ and $Y_2(\omega)$ are the Fourier transforms of the the observed signals $y_1(t)$ and $y_2(t)$, i.e.

$$Y_i(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} y_i(t) e^{-j\omega t} \tag{5.5}$$

In the more general case of M microphones and K noise sources, the observed signal may be written (in the frequency domain) as,

$$\underline{Y}(\omega) = \underline{A}(\omega)S(\omega) + \underline{B}(\omega)\underline{W}(\omega) + \underline{E}(\omega) \tag{5.6}$$

where $\underline{Y}(\omega)$, $\underline{A}(\omega)$ and $\underline{E}(\omega)$ are $1 \times M$ vectors, $\underline{W}(\omega)$ is $1 \times K$ vector and $\underline{B}(\omega)$ is $K \times M$ matrix.

To formulate a statistical maximum likelihood problem, we make the following assumptions. The noise source signal, $w(t)$, is assumed to be a sample from a Gaussian random process. The desired speech signal, $s(t)$, is modeled in many cases as an AR Gaussian random process, whose parameters (the LPC parameters) are slowly time varying. For our purposes, in a short analysis window, we assume that those parameters are constant, and thus, in the mathematical formulation, the desired signal is also assumed to be a sample from a stationary AR Gaussian process. The error signals $e_1(t)$ and $e_2(t)$, are modeled as white Gaussian noise processes. The signals $s(t), w(t), e_1(t)$ and $e_2(t)$ are assumed to be uncorrelated.

The unknown parameters are the coefficients of the various systems and some spectral parameters of the signals. We denote the power spectra of $s(t)$ and $w(t)$ by $P_s(\omega)$ and

$P_w(\omega)$ respectively. $\sigma_{e_1}^2$ and $\sigma_{e_2}^2$ will denote the error signals variances. $A(\omega), B(\omega), C(\omega)$ and $D(\omega)$ are the frequency responses of the four linear systems in Figure 5.2.

We formulate the problem in terms of short-time processing so that the signals and the system parameters can be slowly time varying; consequently, a sliding window is applied. As already noted, the window length, T, must be short enough so that the parameters are constant over its duration. However, we will also assume that it is long enough so that the short-time DFT coefficients of $s(t), w(t), e_1(t)$ and $e_2(t)$ at different frequencies are uncorrelated. Under this assumption, the likelihood of the observations ($y_1(t)$ and $y_2(t)$) with respect to the parameters above is easily expressed in the frequency domain, and is written as, (see e.g. [1], chapter 4),

$$\log p(y_1(t), y_2(t); \underline{\theta}) = -\sum_{\omega_\ell} \left(\log\det\Lambda(\omega_\ell; \underline{\theta}) + \underline{Y}(\omega_\ell)^\dagger \Lambda^{-1}(\omega_\ell; \underline{\theta}) \underline{Y}(\omega_\ell)\right) \quad \omega_\ell = \frac{2\pi}{T}\cdot\ell \quad (5.7)$$

where $\underline{Y}(\omega)$ is a vector whose components are $Y_1(\omega)$ and $Y_2(\omega)$. The matrix $\Lambda(\omega; \underline{\theta})$ is the power spectrum matrix, i.e.

$$\Lambda(\omega; \underline{\theta}) = E\left\{\underline{Y}(\omega)\underline{Y}(\omega)^\dagger\right\} =$$

$$= \begin{bmatrix} C(\omega)P_s(\omega)C^*(\omega) + A(\omega)P_w(\omega)A^*(\omega) + \sigma_{e_1}^2 & C(\omega)P_s(\omega)B^*(\omega) + A(\omega)P_w(\omega)D^*(\omega) \\ B(\omega)P_s(\omega)C^*(\omega) + D(\omega)P_w(\omega)A^*(\omega) & B(\omega)P_s(\omega)B^*(\omega) + D(\omega)P_w(\omega)D^*(\omega) + \sigma_{e_2}^2 \end{bmatrix} \quad (5.8)$$

For the M microphone case, the likelihood function is again (5.7) where the matrix Λ is now the $M \times M$ power spectrum matrix $E\{\underline{Y}(\omega)\underline{Y}^\dagger(\omega)\}$.

The general maximum likelihood problem, represented by eqs. (5.7) and (5.8), is not only complicated but may also be ill-posed. The likelihood function depends on the parameters only through the matrix $\Lambda(\omega; \underline{\theta})$, and all possible solutions that generate the same $\Lambda(\omega; \underline{\theta})$ have the same likelihood. If indeed all the associated systems and the power spectra are unknown and their structure is allowed to be arbitrary, then we expect a non-unique solution, since any value of $A(\omega; \underline{\theta})$ may correspond to many different values for the parameters. Therefore, some constraints must be imposed on the parameters. For example, we may assume that some of the parameters are known, or that there is a simple structure to the systems. Of course, the more constraints there are, the more this ML problem becomes well-posed mathematically. However, we must limit ourselves to constraints that are consistent with the noise cancellation problem under consideration.

We will constrain the systems that represent the room acoustics to be causal, and to have a finite impulse response. Thus, for example, $A(\omega)$ is a frequency response of an FIR filter, i.e.

$$A(\omega) = \sum_{k=0}^{q} a_k e^{-j\omega k} \tag{5.9}$$

As mentioned earlier, we will usually assume that $s(t)$, the desired signal, is a sample from an AR process of order p, so that its power spectrum, $P_s(\omega)$, is of the form

$$P_s(\omega) = \frac{G}{|1 - \sum_{i=1}^{p} h_i e^{-j\omega i}|^2} \tag{5.10}$$

In the next two sections, more specific situations will be considered, and additional constraints and assumptions, based on the additional knowledge about the underlying scenario, will be made. In both sections, the resulting ML problem is constrained enough so that it is not ill-posed.

We note that even with these assumptions and constraints, the required maximization of the likelihood function (5.5) with respect to the signal and system parameters is still complicated. Therefore the EM algorithm will be proposed for its solution. In the cases considered in the next sections, the unavailable desired signal, $s(t)$, will be a component of the chosen complete data. Thus, as a by product of the use of the EM algorithm, an estimate of the desired signal will become explicitly available while implementing the E step.

Relation to array processing and the previous chapter

The system model, presented above and illustrated in Figure 5.2, may also model the array processing situation and the DOA estimation problem considered in the previous chapter. In the array processing case, the various systems A, B, C and D are simple delays, e.g. $A(\omega) = e^{-j\omega\tau_a}$. The additional noise sources, $e_i(t)$, will represent in this case the background noise. Maximizing the likelihood function (5.7) will be equivalent to maximizing the likelihood function developed for the Gaussian DOA estimation problem, i.e. solving (4.58).

The EM algorithms suggested in this chapter are different and do not reduce to the algorithms presented in the previous chapter. We simply choose a different complete data for solving the same ML problem. The choice of complete data in this chapter is adequate for the noise cancellation problem, since the resulting EM algorithm generates an estimate of the desired signal in the E step and since in the case where an estimate of a complete impulse response and not just time delay is desired, updating the parameters using this choice of complete data is easier. On the other hand, for the DOA estimation problem, where the systems are simple delays, the choice of complete data used in the previous chapter has generated an EM algorithm with attractive properties, such as the simple parallel structure.

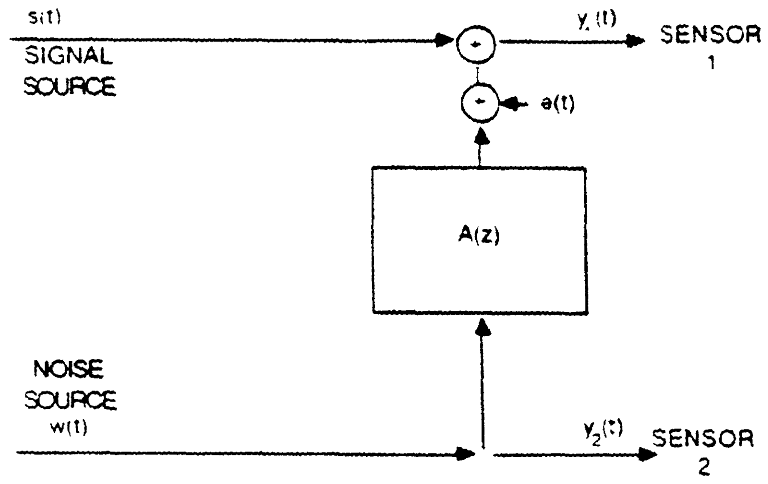


Figure 5.4: The observations; simplified scenario

5.2 A simplified scenario

In this section, a simplified version of the problem is assumed, corresponding to the restriction of $B(z)$ to be zero and both $C(z)$ and $D(z)$ to be unity in Figure 5.2, so that Figure 5.2 reduces to Figure 5.4. This scenario is assumed (at least implicitly) by Widrow et al. in [10]. In this scenario, one microphone measures the desired (speech) signal with additive noise, while the second microphone measures a reference noise signal, which is correlated with the noise component of the signal measured in the first microphone, but has no signal component present.

We will start by presenting the ML problem for this scenario. This ML problem is a special case of (5.7). However, as we shall see, the availability of a reference signal, which contains only noise, makes this likelihood function particularly simple. We will then present an EM algorithm for maximizing this likelihood. The complete data will be composed of the signal part and the noise part of the primary microphone signal *separated* in addition to the reference microphone signal.

5.2.1 The ML problem

As indicated in Figure 5.4, the observed signals are $y_1(t)$ and $y_2(t)$, $A(z)$ is an FIR filter, $e(t)$ is Gaussian white noise, and $s(t)$ is the desired signal.

Specifically, then

$$y_1(t) = s(t) + n(t) \tag{5.11}$$

where the noise component in the primary microphone is

$$n(t) = \sum_{k=0}^{q} a_k y_2(t-k) + e(t) \tag{5.12}$$

Equivalently, equations (5.11) and (5.12) may be written as,

$$y_1(t) = s(t) + \sum_{k=0}^{q} a_k w(t-k) + e(t) \tag{5.13}$$

$$y_2(t) = w(t) \tag{5.14}$$

and the connection to the general scenario is now clear.

As before, we assume that the desired signal, $s(t)$, can be represented as a sample function from a stationary Gaussian process, whose spectrum is known up to some parameters. The unknown parameters, $\underline{\theta}$, are the system coefficients, $\{a_k\}$, the spectral parameters of $s(t)$ (which will be denoted $\underline{\phi}$), and σ^2, the variance of $e(t)$.

The likelihood of the observation is again expressed in the frequency domain. This case is simpler than the general case, discussed in the previous section. The likelihood may be obtained without referring to the general formula of (5.7).

Specifically, under the assumptions made in the previous chapter, the likelihood of the observations may be written in the frequency domain as,

$$L(\underline{\theta}) = \log f_{y_1,y_2}(y_1(t), y_2(t); \underline{\theta}) = \sum_{\omega_\ell} \log f_{Y_1,Y_2}(Y_1(\omega_\ell), Y_2(\omega_\ell); \underline{\theta}) \tag{5.15}$$

Now, at each frequency ω_ℓ

$$\log f_{Y_1,Y_2}(Y_1(\omega_\ell), Y_2(\omega_\ell); \underline{\theta}) = \log f_{Y_1/Y_2}(Y_1(\omega_\ell)/Y_2(\omega_\ell); \underline{\theta}) + \log f_{Y_2}(Y_2(\omega_\ell)) \tag{5.16}$$

where $\log f_{Y_2}(Y_2(\omega_\ell))$ is independent of $\underline{\theta}$. The conditional density of $Y_1(\omega_\ell)$ given $Y_2(\omega_\ell)$ is given by

$$\log f_{Y_1/Y_2}(Y_1(\omega_\ell)/Y_2(\omega_\ell); \underline{\theta}) = -\left[\log \pi \left(P_s(\omega_\ell) + \sigma^2\right) + \frac{|Y_1(\omega_\ell) - A(\omega_\ell) \cdot Y_2(\omega_\ell)|^2}{P_s(\omega_\ell) + \sigma^2}\right] \tag{5.17}$$

Thus maximizing the likelihood (5.15) in this case is equivalent to minimizing,

$$\sum_{\omega_\ell}\left[\log\left(P_s(\omega_\ell) + \sigma^2\right) + \frac{|Y_1(\omega_\ell) - A(\omega_\ell) \cdot Y_2(\omega_\ell)|^2}{P_s(\omega_\ell) + \sigma^2}\right] \tag{5.18}$$

with respect to σ^2 and the coefficients of $P_s(\omega)$ and $A(\omega)$.

We will assume that $A(\omega)$ is the frequency response of an FIR filter of a given order q, i.e. it is of the form of (5.9). Also, we will assume that $s(t)$ is an AR process of order p with coefficients $\{h_i\}_{i=1}^{p}$ and gain G, so that its power spectrum $P_s(\omega)$ is given by (5.10).

In some applications, like LPC vocoding and speech recognition of noisy data, we will be interested mainly in the spectral parameters of the speech signal. In this case, solving this ML problem explicitly provides these desired parameters. In other applications, we will be interested in the speech signal itself. So, using the estimated signal parameters, $\{a_k\}$, we will cancel the noise in the primary microphone and obtain an enhanced speech signal. As mention above, this speech signal estimate will be available as a by product of the EM algorithm suggested below, while implementing the E step.

5.2.2 Solution via the EM algorithm

Direct minimization of (5.18) is complicated; therefore we consider the use of the EM algorithm. In this approach, the complete data is chosen to be the set of signals $\{s(t), n(t), y_2(t)\}$. This choice of complete data is motivated by the simple maximum likelihood solution available if indeed $s(t), n(t)$ and $y_2(t)$ are observed separately.

Loosely speaking, if this complete data is available, the maximum likelihood estimate of $\{a_k\}$ and σ^2 is achieved by least squares fitting of $y_2(t)$ to $n(t)$. The spectral parameters of $s(t)$ are also easily estimated by solving, for example, the normal equation using the sample correlation of $s(t)$, for LPC parameters.

More specifically, the likelihood of the complete data, $L_c(\underline{\theta})$, satisfy

$$\begin{aligned} L_c(\underline{\theta}) &= \log f_{s,n,y_2}(s(t), n(t), y_2(t); \underline{\theta}) \\ &= \log f_{s,n/y_2}(s(t), n(t)/y_2(t); \underline{\theta}) + \log f_{y_2}(y_2(t)) \end{aligned} \tag{5.19}$$

where $\log f_{y_2}(y_2(t))$ is independent of $\underline{\theta}$. Also, given $y_2(t)$, the signals $s(t)$ and $n(t)$ are statistically independent, and thus

$$\log f_{s,n/y_2}(s(t), n(t)/y_2(t); \underline{\theta}) = \log f_{s/y_2}(s(t)/y_2(t); \underline{\theta}) + \log f_{n/y_2}(n(t)/y_2(t); \underline{\theta}) \tag{5.20}$$

Now, $\log f_{n/y_2}(n(t)/y_2(t); \underline{\theta})$ depends only on $\{a_k\}$ and σ^2, and it is defined by the p.d.f of $e(t)$, i.e.

$$\log f_{n/y_2}(n(t)/y_2(t); \underline{\theta}) = -\sum_{t=0}^{N-1} \left[\log \sigma^2 + \frac{1}{\sigma^2} \left(n(t) - \sum_{k=0}^{q} a_k y_2(t-k) \right)^2 \right] \tag{5.21}$$

In general, the signal $y_2(t)$ may be related to $s(t)$. However, this relation is arbitrary and unknown. Therefore, we will assume that the probability distribution of $s(t)$ given $y_2(t)$ is the a-priori distribution of $s(t)$. This probability distribution is the distribution of a stationary random process with power spectrum $P_s(\omega)$ and it depends only on the spectral parameters, $\underline{\phi}$, of $s(t)$, thus,

$$\log f_{s/y_2}(s(t)/y_2(t);\underline{\theta}) = \log f_s(s(t);\underline{\phi}) = \sum_{\omega_\ell}\left[\log P_s(\omega_\ell;\underline{\phi}) + \frac{|S(\omega_\ell)|^2}{P_s(\omega_\ell;\underline{\phi})}\right] \tag{5.22}$$

where $S(\omega)$ is the Fourier transform of $s(t)$, i.e.

$$S(\omega) = \frac{1}{\sqrt{N}}\sum_{t=0}^{N-1} s(t)e^{-j\omega t}$$

Thus, estimating $\underline{\theta}$ by maximizing the likelihood of the complete data is equivalent to estimating σ^2 and $\{a_k\}$ by minimizing

$$\frac{1}{\sigma^2}\sum_{t=0}^{N-1}\left(n(t) - \sum_{k=0}^{q} a_k y_2(t-k)\right)^2 + N\cdot\log\sigma^2 \tag{5.23}$$

and estimating the spectral parameters $\underline{\phi}$ by minimizing

$$\sum_{\omega_\ell}\left[\log P_s(\omega_\ell;\underline{\phi}) + \frac{|S(\omega_\ell)|^2}{P_s(\omega_\ell;\underline{\phi})}\right] \tag{5.24}$$

Note that when $s(t)$ is assumed to be an AR process, minimizing (5.24) is equivalent to solving the Yule-Walker equation, using the sample autocorrelation of $s(t)$.

From eqs. (5.23) and (5.24) we note that the sufficient statistics of the complete data is $n(t)$ and $|S(\omega_\ell)|^2$. The sufficient statistics is linear for the noise part and quadratic for the signal part. Thus, the E step of the algorithm requires the following expectations:

$$\hat{n}(t) = E\left\{n(t)/y_1(t), y_2(t);\underline{\theta}^{(n)}\right\} \tag{5.25}$$

and

$$M_S(\omega_\ell) = E\left\{|S(\omega_\ell)|^2/Y_1(\omega_\ell), Y_2(\omega_\ell);\underline{\theta}^{(n)}\right\} \tag{5.26}$$

where $\underline{\theta}^{(n)}$ denotes the parameters $\{a_k\},\sigma^2$ and $\underline{\phi}$ in the n^{th} iteration. These conditional expectations are immediately available using the results developed for the linear Gaussian case.

The E and the M steps of the EM algorithm for minimizing (5.18) may now be stated explicitly. We will denote by $\underline{\theta}^{(n)}$ (or by $\{a_k^{(n)}\}$, $(\sigma^2)^{(n)}$ and $P_s^{(n)}(\omega)$) the current estimate of the parameters.

- **The E step, the n^{th} iteration:**
  - Generate a signal $x(t)$

$$x(t) = y_1(t) - \sum_{k=0}^{q} a_k^{(n)} y_2(t-k) \tag{5.27}$$

    Note, that if the true coefficients $\{a_k\}$ were known, then $x(t) = s(t) + e(t)$

  - Apply a Wiener filter to $x(t)$ to obtain the conditional expectation or the minimum mean square error estimate of $s(t)$ (or $S(\omega_\ell)$) and $|S(\omega_\ell)|^2$. Specifically, for all ω_ℓ, generate an estimate of $S(\omega_\ell)$, $E(\omega_\ell)$ and $|S(\omega_\ell)|^2$ as

$$\hat{S}(\omega_\ell) = \frac{P_s^{(n)}(\omega_\ell)}{P_s^{(n)}(\omega_\ell) + (\sigma^2)^{(n)}} \cdot X(\omega_\ell) \tag{5.28}$$

$$\hat{E}(\omega_\ell) = X(\omega_\ell) - \hat{S}(\omega_\ell) \tag{5.29}$$

$$M_S(\omega_\ell) = |\hat{S}(\omega_\ell)|^2 + \frac{P_s^{(n)}(\omega_\ell) \cdot (\sigma^2)^{(n)}}{P_s^{(n)}(\omega_\ell) + (\sigma^2)^{(n)}} \tag{5.30}$$

    where $X(\omega)$ is the Fourier transform of $x(t)$ and $\hat{E}(\omega)$ is the Fourier transform of the signal $\hat{e}(t)$.

  - The conditional expectation (estimate) of $n(t)$ is

$$\hat{n}(t) = \sum_{k=0}^{q} a_k^{(n)} y_2(t-k) + \hat{e}(t) \tag{5.31}$$

- **The M step, the n^{th} iteration**

  Substitute the conditional expectations of (5.30) and (5.31) into equations (5.23) and (5.24). Specifically,

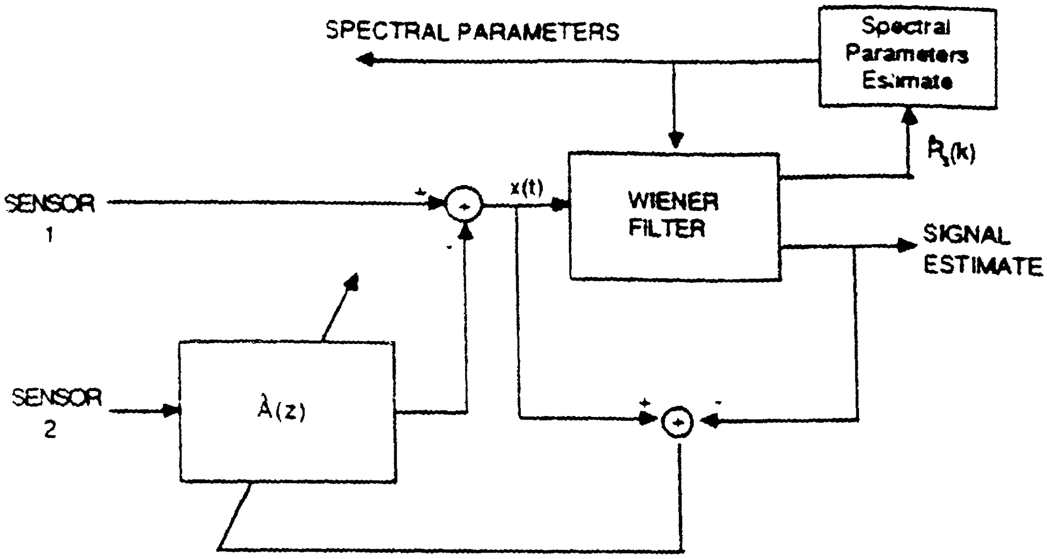


Figure 5.5: The EM algorithm; simplified scenario

- Update $\{a_k\}$ by solving the least-squares problem of (5.23) with (5.31) substituted for $n(t)$, i.e.

$$\{a_k^{(n+1)}\} = \arg\min_{\{a_k\}} \sum_{n=0}^{N-1} \left(\sum_{k=0}^{q} (a_k^{(n)} - a_k) y_2(t-k) + \hat{e}(t) \right)^2 \qquad (5.32)$$

- Update the spectral parameters by solving (5.24) with $M_S(\omega_\ell)$ substituted for $|S(\omega_\ell)|^2$. For LPC parameters, solve the Yule-Walker equation using the estimated correlation values, obtained by inverse Fourier transforming $M_S(\omega)$.

The EM algorithm above iterates, until some convergence criterion is met. This algorithm is summarized in Figure 5.5.

5.3 A more general scenario

The modeling of the two-microphone noise cancellation situation in the previous section ignores the possible coupling of the desired signal, $s(t)$, into the reference microphone. In the classical least-squares approach, this coupling results in a reverberant quality to the output, because the desired signal is partially canceled together with the noise. Since the ML problem of the previous section also ignores this coupling, the resulting EM noise canceling algorithm has a similar problem.

In the ML approach considered in this section, this coupling is taken into account. Specifically, we now include the presence of the system B in Figure 5.2, but still assume that $C \equiv 1$ and $D \equiv 1$, corresponding to the assumption that the first sensor is close to the signal source and that the second sensor is close to the noise source. The resulting model is shown in Figure 5.6. We also assume that $A(z)$ and $B(z)$ are both FIR systems. These assumptions are important, because without them the problem is ill-posed. For example, if A, B, C and D are arbitrary, intuitively one can see that there is a symmetry to the problem, that precludes the algorithm distinguishing between the signal and the noise components in each sensor. With the stated assumptions this symmetry is removed.

We will start by explicitly presenting the ML problem for this scenario. We will then present an EM algorithm for maximizing this likelihood, where the complete data will be composed of the desired speech signal, $s(t)$, and the noise source signal, $w(t)$, in addition to the observed signals $y_1(t)$ and $y_2(t)$.

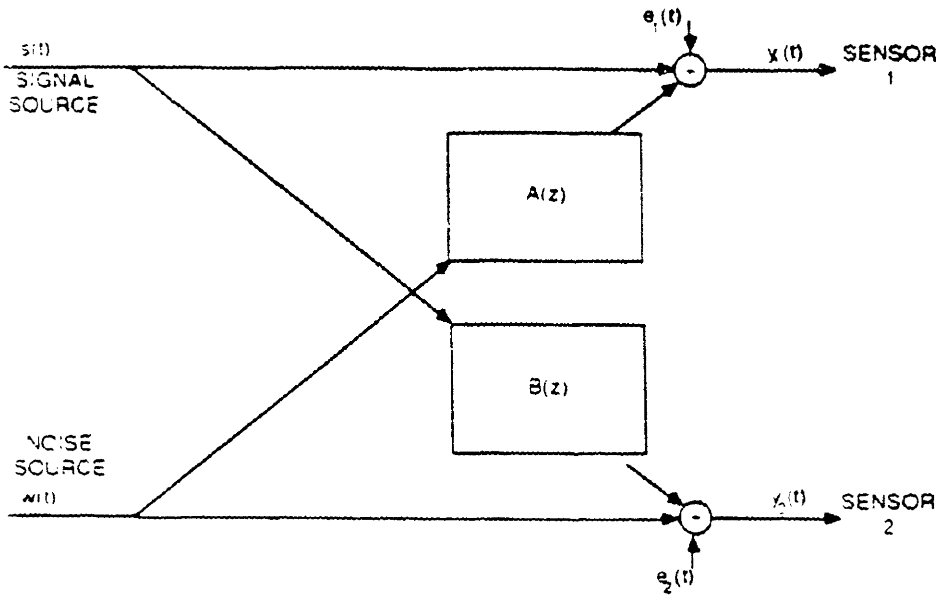


Figure 5.6: The observations; more general scenario

5.3.1 The ML problem

The situation assumed in this section is indicated in Figure 5.6. The mathematical model that corresponds to this situation is given by,

$$y_1(t) = s(t) + A\{w(t)\} + e_1(t) \tag{5.33}$$

$$y_2(t) = B\{s(t)\} + w(t) + e_2(t) \tag{5.34}$$

where, as before, $s(t)$ is the desired signal, $w(t)$ is the noise source signal, and $e_1(t)$ and $e_2(t)$ are the measurement and modeling error signals in the two microphones. As in the general problem, $s(t)$ and $w(t)$ are assumed to be sample signals from Gaussian random processes. The error signals $e_1(t)$ and $e_2(t)$ are white Gaussian noise processes. The unknown parameters, $\underline{\theta}$, are the impulse response coefficients $\{a_k\}$ and $\{b_k\}$ of the systems A and B, the spectral parameters of the signals $s(t)$ and $w(t)$ denoted $\underline{\phi}_s$ and $\underline{\phi}_w$ respectively, and the variances σ_{e_1} and σ_{e_2} of the noises $e_1(t)$ and $e_2(t)$.

With these assumptions, the likelihood of the observations is given again by (5.7). However, with $C(\omega) \equiv D(\omega) \equiv 1$, the power spectrum matrix, $\Lambda(\omega)$, is simplified to

$$\Lambda(\omega;\underline{\theta}) = E\left\{\underline{Y}(\omega)\underline{Y}(\omega)^{\dagger}\right\} =$$

$$= \begin{bmatrix} P_s(\omega) + A(\omega)P_w(\omega)A^*(\omega) + \sigma_{\epsilon_1}^2 & P_s(\omega)B^*(\omega) + A(\omega)P_w(\omega) \\ B(\omega)P_s(\omega) + P_w(\omega)A^*(\omega) & B(\omega)P_s(\omega)B^*(\omega) + P_w(\omega) + \sigma_{\epsilon_2}^2 \end{bmatrix} \tag{5.35}$$

We will assume again that $A(\omega)$ and $B(\omega)$ are frequency responses of FIR filters, i.e. their structure is given by (5.9). The orders of those FIR filters are assumed to be known, and are denoted q_a and q_b respectively. The desired signal is assumed to be a sample from an AR process of a given order p, so that $P_s(\omega)$ will have the structure of (5.10). We further assume that $w(t)$ is a white noise signal, i.e. $P_w(\omega)$ is constant. Even with these assumptions, the underlying ML problem is complicated, so again we will use the EM algorithm for its solution.

For applications, such as LPC vocoding, where only the spectral parameters of the speech signals are required, solving this ML problem will explicitly provide these desired parameters. For applications where the speech signal is required, the MMSE estimate of the speech signal using the ML estimate of the parameters will be suggested. This MMSE estimate will be available for each current parameter value, as a by product, while implementing the E step of the suggested EM algorithm.

5.3.2 Solution via the EM algorithm

The complete data suggested for defining the EM algorithm in the current context is the set of signals $\{s(t), w(t), y_1(t), y_2(t)\}$. The complete data is chosen this way, since, if indeed the signals $s(t)$ and $w(t)$, the input to the two channel system of Figure 5.6, are observed, in addition to the signals $y_1(t)$ and $y_2(t)$, the output of this system, then there will be a simple procedure for ML estimation of the parameters of this two channel system.

Specifically, suppose that this complete data is available. To estimate the parameters we will maximize its likelihood given by

$$\begin{aligned} L_c(\underline{\theta}) &= \log f_{y_1,y_2,s,w}(y_1(t), y_2(t), s(t), w(t); \underline{\theta}) \\ &= \log f_{y_1,y_2/s,w}(y_1(t), y_2(t)/s(t), w(t); \underline{\theta}) + \log f_{s,w}(s(t), w(t); \underline{\theta}) \end{aligned} \tag{5.36}$$

The signals $y_1(t)$ and $y_2(t)$ are uncorrelated, given $s(t)$ and $w(t)$. The signals $s(t)$ and $w(t)$ are uncorrelated by assumption, thus,

$$\begin{aligned} L_c(\underline{\theta}) &= \underbrace{\log f_{y_1/s,w}(y_1(t)/s(t), w(t); \underline{\theta})}_{I} + \underbrace{\log f_{y_2/s,w}(y_2(t)/s(t), w(t); \underline{\theta})}_{II} \\ & \underbrace{\log f_s(s(t); \underline{\theta})}_{III} + \underbrace{\log f_w(w(t); \underline{\theta})}_{IV} \end{aligned} \tag{5.37}$$

Term I depends only on $\{a_k\}$ and $\sigma^2_{e_1}$ and is the log probability of the sequence $e_1(t)$. Similarly, term II depends only on $\{b_k\}$ and $\sigma^2_{e_2}$ and is the log probability of the sequence $e_2(t)$. Term III is the log probability of the stationary signal $s(t)$ and depends only on its spectral parameters $\underline{\phi}_s$. Similarly term IV is the log probability of the stationary signal $w(t)$ and depends only on its spectral parameters $\underline{\phi}_w$. Maximizing the likelihood of the complete data with respect to $\underline{\theta}$ is equivalent to maximizing each of the terms $I - IV$ separately with respect to the parameters they depend on.

Thus, given the complete data, $\underline{\phi}_s$ are estimated by,

$$\hat{\underline{\phi}}_s = \arg\max_{\underline{\phi}_s} \log f_s(s(t); \underline{\phi}_s) = \arg\min_{\underline{\phi}_s} \sum_{\omega_\ell} \left[\log P_s(\omega_\ell; \underline{\phi}_s) + \frac{|S(\omega_\ell)|^2}{P_s(\omega_\ell; \underline{\phi}_s)} \right] \tag{5.38}$$

and $\underline{\phi}_w$ are estimated by,

$$\hat{\underline{\phi}}_w = \arg\max_{\underline{\phi}_w} \log f_w(w(t); \underline{\phi}_w) = \arg\min_{\underline{\phi}_w} \sum_{\omega_\ell} \left[\log P_w(\omega_\ell; \underline{\phi}_w) + \frac{|W(\omega_\ell)|^2}{P_w(\omega_\ell; \underline{\phi}_w)} \right] \tag{5.39}$$

where $S(\omega)$ and $W(\omega)$ are the Fourier transforms of $s(t)$ and $w(t)$ respectively, i.e.

$$S(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} s(t)e^{-j\omega t}$$

$$W(\omega) = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} w(t)e^{-j\omega t}$$

The maximization in (5.38) is sometimes simpler, e.g. when $s(t)$ is assumed to be an AR process, in which case maximizing (5.38) is equivalent to solving the Yule-Walker equation, using the sample autocorrelation of $s(t)$. Similarly, solving (5.39) is sometimes simpler. If $w(t)$, the noise source signal, is a white noise signal, then solving (5.39) is equivalent to finding the (constant) spectrum level, P_w, by,

$$\hat{P_w} = \frac{1}{N} \sum_{t=0}^{N-1} w^2(t) = \sum_{\omega_f} |W(\omega)|^2 \tag{5.40}$$

Estimating the impulse response coefficients, $\{a_k\}$, and the variance, $\sigma^2_{e_1}$, given the complete data, requires solving a least-squares problem, since

$$\begin{aligned} \hat{\sigma^2_{e_1}}, \{\hat{a_k}\} &= \arg \max_{\sigma^2_{e_1}, \{a_k\}} \log f_{y_1/s,w}(y_1(t)/s(t), w(t); a_0, \cdots, a_{q_a}, \sigma^2_{e_1}) \qquad (5.41) \\ &= \arg \min_{\sigma^2_{e_1}, \{a_k\}} \frac{1}{\sigma^2_{e_1}} \sum_{t=0}^{N-1} \left(y_1(t) - s(t) - \sum_{k=0}^{q_a} a_k w(t-k) \right)^2 + N \cdot \log \sigma^2_{e_1} \end{aligned}$$

Similarly, estimating $\{b_k\}$ and $\sigma^2_{e_2}$ given the complete data requires solving the following least squares problem,

$$\begin{aligned} \hat{\sigma^2_{e_2}}, \{\hat{b_k}\} &= \arg \max_{\sigma^2_{e_2}, \{b_k\}} \log f_{y_2/s,w}(y_2(t)/s(t), w(t); b_0, \cdots, b_{q_b}, \sigma^2_{e_2}) \qquad (5.42) \\ &= \arg \min_{\sigma^2_{e_2}, \{b_k\}} \frac{1}{\sigma^2_{e_2}} \sum_{t=0}^{N-1} \left(y_2(t) - w(t) - \sum_{k=0}^{q_b} b_k s(t-k) \right)^2 + N \cdot \log \sigma^2_{e_2} \end{aligned}$$

The explicit solution of the least-squares problems, implied by equations (5.41) and (5.42), is achieved by solving the following "normal" linear equations:

$$\underline{R}_w \cdot \underline{a} = \underline{r}_{wy_1} - \underline{r}_{ws} \tag{5.43}$$

$$\mathcal{R}_s \cdot \underline{b} = \underline{r}_{sy_2} - \underline{r}_{sw} \tag{5.44}$$

where $\mathcal{R}_w$ is the correlation matrix of $w(t)$ of order q_a i.e.

$$\mathcal{R}_w = \left[\begin{array}{cccccc} r_w(0) & r_w(1) & r_w(2) & \cdot & \cdot & r_w(q_a) \\ r_w(1) & r_w(0) & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & & \cdot & \cdot & \cdot & \\ \cdot & & & \cdot & \cdot & r_w(1) \\ r_w(q_a) & \cdot & \cdot & r_w(2) & r_w(1) & r_w(0) \end{array}\right] \quad \text{where } r_w(k) = \frac{1}{N}\sum_{t=0}^{N-1} w(t)w(t-k) \tag{5.45}$$

$\mathcal{R}_s$ is the order q_b correlation matrix of $s(t)$

$$\mathcal{R}_s = \left[\begin{array}{cccccc} r_s(0) & r_s(1) & r_s(2) & \cdot & \cdot & r_s(q_b) \\ r_s(1) & r_s(0) & \cdot & & \cdot & \\ \cdot & \cdot & \cdot & \cdot & & \\ \cdot & & \cdot\cdot & \cdot & \cdot & \\ \cdot & & & \cdot & \cdot & r_s(1) \\ r_s(q_b) & \cdot & \cdot & r_s(2) & r_s(1) & r_s(0) \end{array}\right] \quad \text{where } r_s(k) = \frac{1}{N}\sum_{t=0}^{N-1} s(t)s(t-k) \tag{5.46}$$

The vectors $\underline{r}_{wy_1}, \underline{r}_{ws}, \underline{r}_{sy_2}$ and $\underline{r}_{sw}$ represent the appropriate cross correlation between the signals, e.g.

$$\underline{r}_{ws} = \left[\begin{array}{c} r_{ws}(0) \\ \cdot \\ \cdot \\ \cdot \\ r_{ws}(q_a) \end{array}\right] \quad \text{where } r_{ws}(k) = \frac{1}{N}\sum_{t=0}^{N-1} s(t)w(t-k) \tag{5.47}$$

and the vectors $\underline{a}$ and $\underline{b}$ are the unknown impulse response coefficients of the systems A and B.

By observing the required procedures for maximizing the likelihood of the complete data, i.e. equations (5.38),(5.39) and (5.43),(5.44), we see that the sufficient statistics of the complete data contains quadratic terms, which are the sample autocorrelation (or the sample spectrum) and the sample cross correlation (or cross spectrum) of the various signals, in addition to the linear terms (i.e. the signals themselves). Thus, the E step of the algorithm (with the above choice of complete data) requires the expectations:

$$\hat{s}(t) = E\left\{s(t)/y_1(t), y_2(t); \underline{\theta}^{(n)}\right\} \tag{5.48}$$

$$\hat{w}(t) = E\left\{w(t)/y_1(t), y_2(t); \underline{\theta}^{(n)}\right\} \tag{5.49}$$

and the quadratic terms:

$$\hat{r}_s(k) = E\left\{r_s(k)/y_1(t), y_2(t); \underline{\theta}^{(n)}\right\} \tag{5.50}$$

$$\hat{r}_w(k) = E\left\{r_w(k)/y_1(t), y_2(t); \underline{\theta}^{(n)}\right\} \tag{5.51}$$

$$\hat{r}_{sw}(k) = E\left\{r_{sw}(k)/y_1(t), y_2(t); \underline{\theta}^{(n)}\right\} = \hat{r}_{ws}(-k) \tag{5.52}$$

We will implement the E step in the frequency domain, since for stationary processes with large observation time, the DFT coefficients at each frequency are statistically independent and can be processed separately. In each frequency ω_ℓ the observation may be written as

$$\begin{bmatrix} Y_1(\omega_\ell) \\ Y_2(\omega_\ell) \end{bmatrix} = \begin{bmatrix} 1 & A(\omega_\ell) \\ B(\omega_\ell) & 1 \end{bmatrix} \cdot \begin{bmatrix} \dot{S}(\omega_\ell) \\ W(\omega_\ell) \end{bmatrix} \tag{5.53}$$

The E step requires the conditional expectation of $S(\omega_\ell)$, $W(\omega_\ell)$, $|S(\omega_\ell)|^2$, $|W(\omega_\ell)|^2$ and $S(\omega_\ell)W^*(\omega_\ell)$.

At each step of the algorithm, the current values of the parameters are used. We will denote by $A^{(n)}(\omega)$ and $B^{(n)}(\omega)$ the current estimate of the frequency responses of the unknown systems A and B, and by $P_s^{(n)}(\omega)$ and $P_w^{(n)}(\omega)$ the current estimate of the power spectra of $s(t)$ and $w(t)$. Let $H(\omega_\ell)$ denote the matrix

$$H(\omega_\ell) = \begin{bmatrix} 1 & A^{(n)}(\omega_\ell) \\ B^{(n)}(\omega_\ell) & 1 \end{bmatrix} \tag{5.54}$$

and let $\Phi(\omega_\ell)$ and Σ denote the power spectra matrices

$$\Phi(\omega_\ell) = \begin{bmatrix} P_s^{(n)}(\omega_\ell) & 0 \\ 0 & P_w^{(n)}(\omega_\ell) \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_{e_1}^2 & 0 \\ 0 & \sigma_{e_2}^2 \end{bmatrix} \tag{5.55}$$

The required conditional expectations are readily available, since again this is a linear Gaussian case. These conditional estimates may be interpreted as performing a two-channel Wiener filter (see [20]) and calculating its error covariance matrix. Thus, the estimate of the linear terms is given by

$$\begin{bmatrix} \hat{S}(\omega_\ell) \\ \hat{W}(\omega_\ell) \end{bmatrix} = K(\omega_\ell) \cdot \begin{bmatrix} Y_1(\omega_\ell) \\ Y_2(\omega_\ell) \end{bmatrix} \tag{5.56}$$

where $K(\omega_\ell)$ is the matrix

$$K(\omega_\ell) = \Phi(\omega_\ell) \cdot H(\omega_\ell)^\dagger \left(H(\omega_\ell) \cdot \Phi(\omega_\ell) \cdot H(\omega_\ell)^\dagger + \Sigma \right)^{-1} \tag{5.57}$$

For the quadratic terms, we have to calculate the error covariance matrix of this Wiener filter, i.e.

$$\begin{aligned} \tilde{P}(\omega_\ell) &= \left(\Phi^{-1}(\omega_\ell) + H(\omega_\ell) \cdot \Sigma^{-1} \cdot H(\omega_\ell)^\dagger \right)^{-1} \\ &= \Phi(\omega_\ell) - \Phi(\omega_\ell) H(\omega_\ell)^\dagger \left(H(\omega_\ell) \cdot \Phi(\omega_\ell) \cdot H(\omega_\ell)^\dagger + \Sigma \right)^{-1} H(\omega_\ell) \Phi(\omega_\ell) \end{aligned} \tag{5.58}$$

and the quadratic terms are obtained by,

$$M_S(\omega_\ell) = E\left\{|S(\omega_\ell)|^2/Y_1(\omega_\ell), Y_2(\omega_\ell)\right\} = |\hat{S}(\omega_\ell)|^2 + \hat{P}_{11}(\omega_\ell) \tag{5.59}$$

$$M_W(\omega_\ell) = E\left\{|W(\omega_\ell)|^2/Y_1(\omega_\ell), Y_2(\omega_\ell)\right\} = |\hat{W}(\omega_\ell)|^2 + \hat{P}_{22}(\omega_\ell) \tag{5.60}$$

$$M_{SW}(\omega_\ell) = E\left\{S(\omega_\ell)W^*(\omega_\ell)/Y_1(\omega_\ell), Y_2(\omega_\ell)\right\} = \hat{S}(\omega_\ell)\hat{W}^*(\omega_\ell) + \hat{P}_{12}(\omega_\ell) \tag{5.61}$$

The E and M steps of the EM algorithm for maximizing the likelihood of the observations (given by (5.7) and (5.35)) for this more general case may now be stated explicitly.

- **The E step, the n^{th} iteration:**
  - Calculate the conditional expectations $\hat{S}(\omega_\ell)$ and $\hat{W}(\omega_\ell)$ by (5.56).
  - Calculate $M_S(\omega_\ell), M_W(\omega_\ell)$ and $M_{SW}(\omega_\ell)$ by (5.59)-(5.61).
  - The signal estimates $\hat{s}(t)$ and $\hat{w}(t)$, and the correlation estimates $\hat{r}_s(k), \hat{r}_w(k)$ and $\hat{r}_{sw}(k)$ are achieved by inverse Fourier transforming $\hat{S}(\omega), \hat{W}(\omega), M_S(\omega), M_W(\omega)$ and $M_{SW}(\omega)$ respectively.
- **The M step, the n^{th} iteration:**
  - Solve the linear equations of (5.43) and (5.44) for $\underline{a}$ and $\underline{b}$, using the estimates $\hat{r}_s(k), \hat{r}_w(k)$ and $\hat{r}_{sw}(k)$ from the E step, and with

$$\hat{r}_{wy_1}(k) = \frac{1}{N}\sum_{t=0}^{N-1} \hat{w}(t)y_1(t-k)$$

$$\hat{r}_{sy_2}(k) = \frac{1}{N}\sum_{t=0}^{N-1} \hat{s}(t)y_2(t-k) \tag{5.62}$$

    The result is the updated coefficients $\underline{a}^{(n+1)}$ and $\underline{b}^{(n+1)}$ of the systems A and B.
  - Update the spectral parameter estimate, by solving (5.38) and (5.39), using $M_S(\omega_\ell)$ and $M_W(\omega_\ell)$ instead of $|S(\omega_\ell)|^2$ and $|W(\omega_\ell)|^2$. For LPC parameters of the speech signal $s(t)$, solve the Yule-Walker equations, using $\hat{r}_s(k)$.

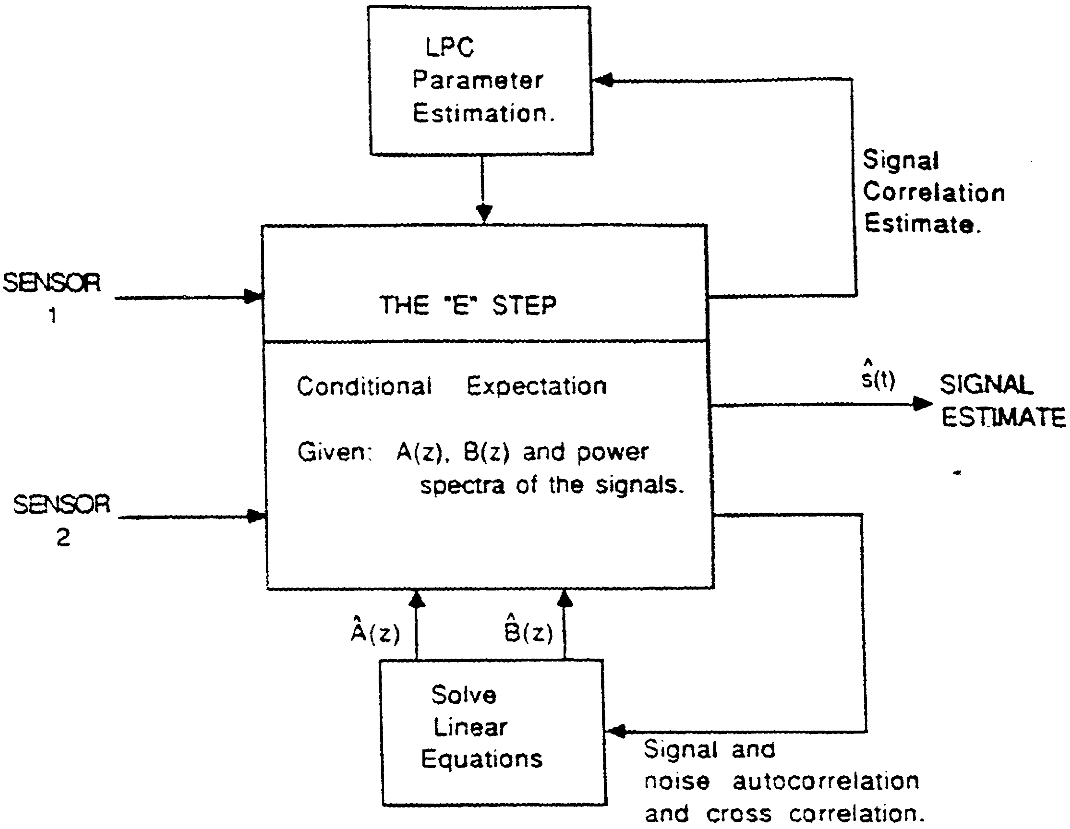


Figure 5.7: The EM algorithm; more general scenario

The EM algorithm above iterates, until some convergence criterion is met. This algorithm is summarized in Figure 5.7.

The procedures, suggested in this section and also in the previous section, are implemented in each iteration on the entire data. Adaptive and sequential procedures, based on the discussion in chapter 3, are also possible. These algorithms may process new data in each new iteration. The parameters will be updated according to one of the suggested strategies in chapter 3, and a new segment of enhanced signal will be produced.

Examining the suggested procedure illustrated in Figure 5.7, a sequential EM algorithm based on recursive E and M steps, comes in mind. The Wiener filter of the E step will be replaced by the sequential Kalman filter, and the linear least-squares problems of the M step will be solved via a sequential RLS type algorithm. The details, the analysis and experiments with this adaptive version are now under investigation and are the subject of further research.

5.4 Experimental results

The algorithms developed in this chapter for both the simplified scenario and the more general scenario have been applied and tested in a simulated environment. A realistic acoustic environment has been created by generating the impulse response coefficients of the systems, representing the room acoustics, using a well tested acoustic simulation program [70]. In this section we will discuss the results of our simulation experiments.

5.4.1 The simplified scenario

The simplified scenario of Figure 5.4 has been implemented with $s(t)$, a speech signal, and $y_2(t)$, a band limited noise signal with a flat spectrum from zero to 3 KHz. The FIR filter, $A(z)$, was of order 10. $y_1(t)$ was generated according to Figure 5.4. The SNR in $y_1(t)$ was approximately -20 db. The results were compared with a "batch" version of the least-squares algorithm, corresponding to estimating the $\{a_k\}$'s via the least-square problem

$$\min_{\{a_k\}} \sum_t \left(y_1(t) - \sum_{k=1}^{q} a_k y_2(t-k) \right)^2$$

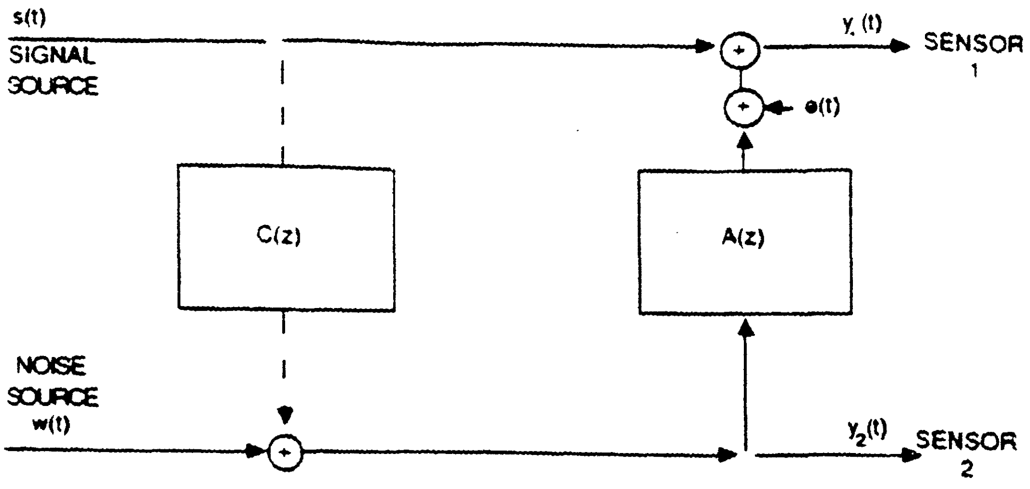


Figure 5.8: Correlation between the reference and d^esired signals. $C(z) = 0.1z^{-5}$

and then canceling the noise and estimating the signal as

$$s(t) = y_1(t) - \sum_{k=1}^{q} a_k y_2(t-k)$$

Both algorithms produced good enhancement of the speech signal, and although there were perceptible differences, the overall quality of both was similar.

The direct least-square approach assumes that $y_2(t)$ and $s(t)$ are uncorrelated. This assumption is critical. Our algorithm, on the other hand, does not require this assumption. In a second experiment, $y_2(t)$ included a delayed version of the speech signal, as illustrated in Figure 5.8. (Note, that this scheme is different then the scheme considered in the more general scenario. since we have a direct measurement of the input to the system $A(z)$).

In this experiment, the SNR in $y_1(t)$ was again -20 db. The direct least squares approach canceled part of the signal together with the noise, resulting in poor quality. In comparison, the performance of our algorithm was still good.

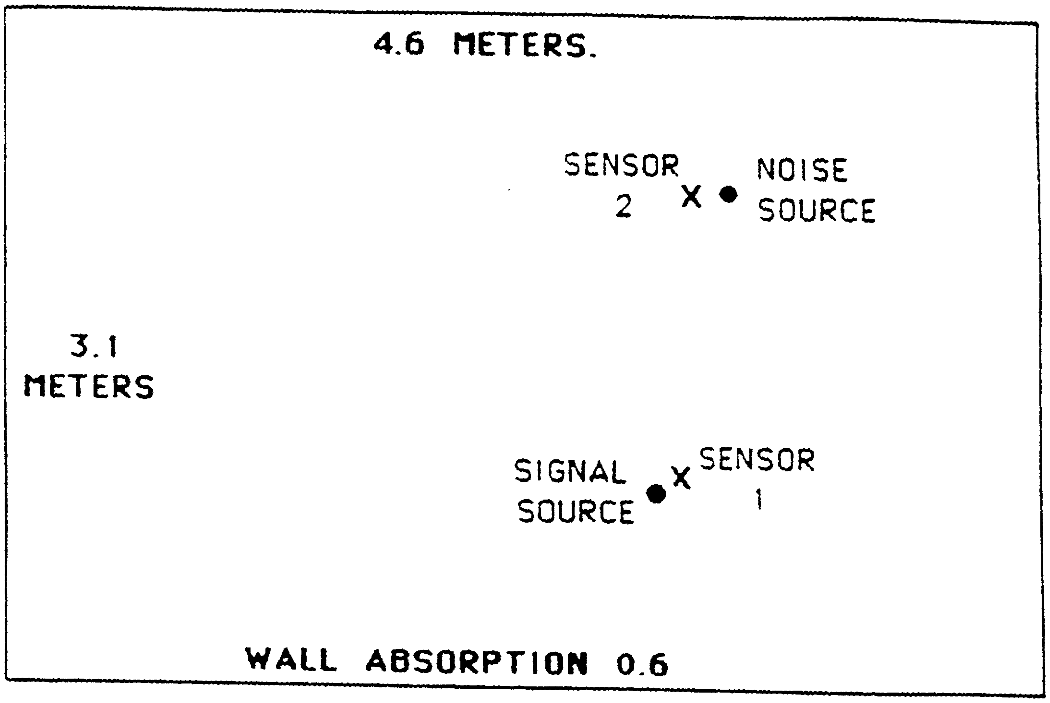


Figure 5.9: The living room acoustic environment

5.4.2 The more general scenario

The more general scenario assumed in Figure 5.6 was simulated, where again $s(t)$ was a speech signal and $w(t)$ was a white noise signal. In order to simulate a realistic scenario, we assumed a living room environment with the signal and noise sources located as illustrated in Figure 5.9. We used a simulation program developed by Peterson ([70]), and we generated FIR impulse responses having 2000 coefficients for each of the systems A and B. The first 500 coefficients of these impulse responses are plotted in Figure 5.10. By monitoring the level of the noise source, we have generated examples in which the SNR of $y_1(t)$ was +20dB, 0dB and -20dB.

We implemented the EM algorithm described in Figure 5.7 and compared the results to the least-squares method, by informal listening. Both algorithms estimated up to 500

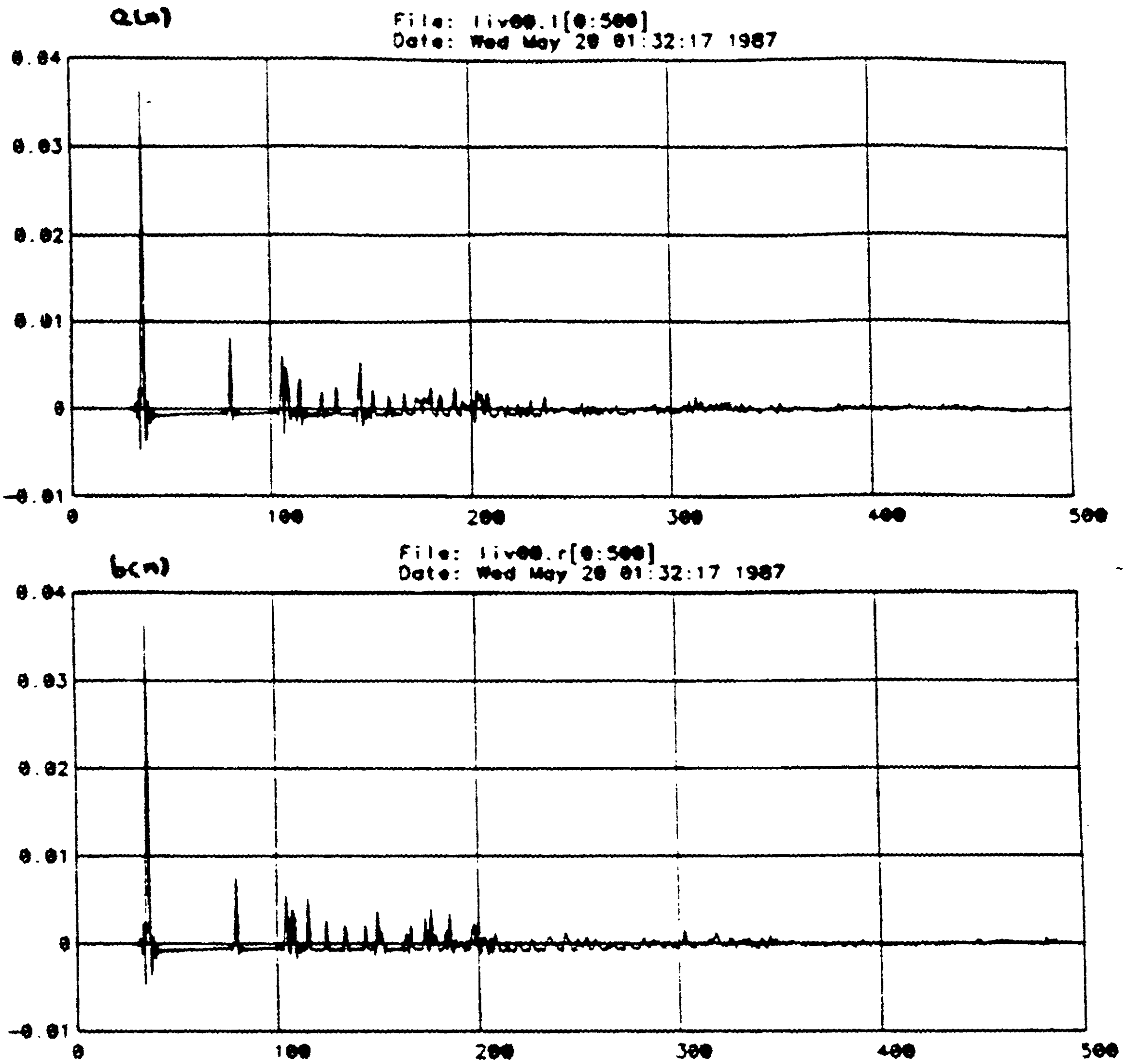


Figure 5.10: Simulated "room acoustics" impulse responses.

coefficients of the impulse response. In all SNR levels our algorithm performed better, and its output, unlike the least square output, was reverberation free.

At high SNR (+20dB), the output of the least squares method output sounded worse than the unprocessed measurement signal, due to the signal canceling effect. The output of our method sounded better than the original measurement signal.

At 0dB, the least squares output sounded better than the measurement signal. However, it sounded much worse than the output of our algorithm, which at this SNR level generated an almost clean signal.

At -20dB SNR, the output of the ML method sounded better then the least-squares method. However, the distinction between the two was not as significant as in the case of 0dB SNR. This is perhaps a result of the fact that, in order to generate a low SNR, we increased the level of the noise source. This resulted in a high Noise to Signal Ratio in the reference microphone, which in turn resulted in lower signal cancellation, since the situation became closer to that assumed by the least squares method.

Chapter 6

Information, Relative Entropy and the EM algorithm

At this point in the thesis, we have already presented the main results and contributions. The class of iterative and sequential algorithms has been presented and motivated. We have also demonstrated several applications of these algorithm to real world signal processing problems. In the course of the thesis, several subjects, related to information and the philosophical essence of the inference process, have been mentioned briefly. We want to use the opportunity of this chapter to discuss these issues further. We will consider in this chapter some interesting topics, in the context of information theory, that are related to the EM method and to the notion of complete and incomplete data relations.

Information measures and statistical inference criteria are closely related. The books by Kullback [71] and Pinsker [72] are devoted to information theory and statistics. The Minimum Description Length (MDL) and the Minimum Information (MI) criteria, mentioned in Chapter 2, provide examples of the application of information measures to inference problems. Another criterion, based on information theory, the Maximum Entropy (ME) or its generalization the Minimum Relative Entropy (MRE), is the main tool used in [71] to relate information theory to statistics. It will be interesting to show that the ME/MRE criterion is a special case of the MDL criterion in a special complete-incomplete data context.

Another topic, discussed in this chapter, is the alternative derivation of the EM algorithm using the MRE criterion, which has been suggested by Musicus in [7] and [8], and by Csiszar et. al. [73]. This derivation is presented with our interpretation and comments. In the context of this chapter, where general information criteria are analyzed, this derivation may be viewed in the right perspective.

This chapter is not organized as coherent theory, rather as "variations on the themes" above. The common "motive" is the relation to the EM method. We start our presentation with the ME or the MRE criterion, point out that its philosophy is distinct from the philosophy of the ML and other statistical criteria and give its common justifications. We then show that the MRE criterion sometimes reduces to the ML criterion, and, in these cases, its minimization using the alternate minimization algorithm, reduces to the EM algorithm. However, we will raise some doubt concerning the rationale behind using the Minimum Relative Entropy in some contexts, including the context that led to the alternative derivation of the EM algorithm. The chapter ends with an important result; we prove that the ME or MRE criterion can be viewed as an interesting implementation of the MI/MDL ideas in a special complete/incomplete data situation.

6.1 Maximum Entropy and Minimum Relative Entropy

So far in the thesis we have presented several methods and criteria for statistical inference. Maximum Entropy and Minimum Relative Entropy methods flow from a different philosophy, however. This philosophy will be presented and compared to the philosophy of other statistical inference criteria.

6.1.1 ME and MRE in comparison to other statistical criteria

The "output" of every statistical inference method is a choice of probability function, which we believe, according to the specific criterion we use, represents best the behavior of the phenomena that we observe. However, ME or MRE methods consider the observations in a unique manner and accept only certain forms of data.

To be more specific, let us describe the common situation that leads to the application of ME or MRE methods. As mentioned above, the goal of any inference process is to find a p.d.f., $p(x)$, defined over the sample space, X, of the observed phenomena. Suppose we know that $p(\cdot)$ belong to a set P, where this set is usually defined by the knowledge of some averages,

$$P = \{p(x) \mid E_p[g(x)] = \bar{g}\} \tag{6.1}$$

The given averages are the only information observed from the underlying phenomena in the ME framework. The choice of probability function is then made by,

$$\hat{p} = \arg\max_{p \in P} \mathcal{E}(p) = \arg\max_{p \in P} \left[- \int_x p(x) \log p(x) dx \right] \tag{6.2}$$

In the MRE framework, an a-priori probability, $q(\cdot)$, for the observed phenomena is also given, and thus the choice of probability function is made by,

$$\bar{p} = \arg\min_{p\in P} \mathcal{H}(p;q) = \arg\min_{p\in P} \left[\int_x p(x)\log\frac{p(x)}{q(x)}dx\right] \tag{6.3}$$

We immediately note that the MRE criterion reduces to the ME criterion if $q(\cdot)$ is the uniform prior. Therefore, in the rest of the chapter, we will discuss only the MRE method; all the results will also apply to the ME method.

The basic limitation of the MRE method is that one cannot incorporate the information provided by a specific observation sequence. The only information that can be incorporated is in the form of given averages or other constraints on the probability function. Nevertheless, the MRE method is sometimes used when a specific observation sequence is given. In this case, one usually calculates some sample averages and uses them as constraints on the relative entropy minimization. This approach is certainly not an optimal one, since not all available information about the phenomena is incorporated and since errors are introduced in the inference process, because the sample averages differ from the statistical averages.

The other statistical inference methods consider the observed data directly. For example, in the Maximum Likelihood framework, we have an observation $x \in X$ and we assume that the probability distribution that describes x is characterized by some unknown parameter vector, $\underline{\theta} \in \Theta$. The ML criterion will choose $\bar{p}$ by choosing $\hat{\underline{\theta}} \in \Theta$ according to,

$$\bar{\underline{\theta}} = \arg\max_{\underline{\theta}\in\Theta} \log p(x;\underline{\theta}) \tag{6.4}$$

The basic limitation of the ML is the need for modeling assumptions. Without those restrictions the method will break down; for example, if we allow the any probability function, maximizing the likelihood will lead to the trivial (but unacceptable) result $p(\alpha) = \delta(\alpha - x)$, where $\delta(\cdot)$ is the Dirac delta function. On the other hand, in the ME/MRE framework, we do not assume any model; the method will work even if the set P in (6.2) and (6.3) contains all possible probability functions. The constraints on the probability functions are derived from the data.

To summarize, the weakness of the MRE method is the unique and restricted manner in which it can accept input. The MRE method can consider the data only in the form of averages. Its strength, on the other hand, lies in the facts that no modeling assumptions are needed and that a full probability function is estimated. There are situations in which giving the input in form of averages is natural; for example, in statistical mechanics, the observations at the macroscopic level are indeed some averages of a stochastic phenomena, that occurs at the microscopic level. In these cases, following the rationale of the MRE method given below, the usage of the MRE method is justified.

6.1.2 The rationale of the MRE method

The commonly used rationale for justifying the Maximum Entropy method is advocated mainly by Jaynes [74] and [75]. Here, we briefly repeat this rationale.

Suppose that the sample space of the underlying phenomena is discrete and finite, i.e. the random variable, X, whose instances, x, we may observe, takes its values over the finite set $\{1, \cdots, m\}$. This random variable has an a-priori probability assignment, $\{q_1, \cdots, q_m\}$. Suppose we observe an infinite i.i.d. sequence, $\{x_1 \cdots x_n \cdots\}$, of realizations of X. A question we might ask is what will the sample frequencies (or histogram) of this sequence typically be ? Naturally, by the strong law of large numbers, the answer is the a-priori probability, $\{q_i\}_{i=1}^m$. However, suppose we have additional information in the form of constraints on the possible histograms, maybe a knowledge of some averages, that rules out the a-priori assignment. In this case, we will compare the probabilities of all possible histograms; the typical histogram will be the one with the highest probability.

Specifically, consider first the case of a finite sequence, $\underline{x} = \{x_1 \cdots x_N\}$, of i.i.d. realization of X. With the given probability assignment, $\{q_i\}_{i=1}^m$, on X, the probability of getting a specific sequence, whose sample frequencies are $\{p_i\}_{i=1}^m$, where $p_i = k_i/N$ and k_i is the number of times the outcome i appeared in $\underline{x}$, is

$$p(\underline{x}) = \prod_{i=1}^{m} q_i^{k_i} = \prod_{i=1}^{m} q_i^{N \cdot p_i} \tag{6.5}$$

There are, however, $\frac{N!}{k_1! \cdots k_m!}$ sequences with these sample frequencies. Thus, the probability of the event "the sample frequencies of $\underline{x}$ are $\{p_i\}_{i=1}^m$", denoted $Prob\{\underline{p}/\underline{q}\}$, is given by,

$$Prob\{\underline{p}/\underline{q}\} = \frac{N!}{k_1! \cdots k_m!} \cdot \prod_{i=1}^{m} q_i^{k_i} = \frac{N!}{k_1! \cdots k_m!} \cdot \prod_{i=1}^{m} q_i^{N \cdot p_i} \tag{6.6}$$

It is easy to prove, using Stirling's formula for factorial, that

$$S = \frac{N!}{k_1! \cdots k_m!} = e^{N[\mathcal{E}(\underline{p}) + o(\frac{\log N}{N})]} \tag{6.7}$$

where $\mathcal{E}(\underline{p}) = -\sum_{i=1}^{m} p_i \log p_i$ is the entropy associated with the frequencies $p_i = k_i/N$. Thus, as $N \to \infty$ equation (6.6) becomes

$$Prob\{\underline{p}/\underline{q}\} = e^{N[\mathcal{E}(\underline{p})]} \cdot \prod_{i=1}^{m} q_i^{N \cdot p_i} = e^{N \cdot \sum_{i=1}^{m} p_i \log \frac{q_i}{p_i}} \tag{6.8}$$

or equivalently,

$$\log Prob\{\underline{p}/\underline{q}\} = -N \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} = -N \cdot \mathcal{H}(\underline{p}; \underline{q}) \tag{6.9}$$

where $\mathcal{H}(\underline{p}; \underline{q})$ is the relative entropy between $\{p_i\}_{i=1}^m$ and $\{q_i\}_{i=1}^m$.

From equation (6.9) we see that the relative entropy is directly proportional to $-\log Prob\{\underline{p}/\underline{q}\}$. Thus the histogram with the highest probability is the one that minimizes the relative entropy. This histogram is also the typical histogram in the following sense. Consider the histogram $\underline{p}$ that minimizes the relative entropy and any other histogram $\underline{p}'$ with higher relative entropy. We claim that the probability of $\underline{p}$ is overwhelmingly larger than the probability of $\underline{p}'$, since, as $N \rightarrow \infty$, the ratio between these probabilities goes exponentially fast to infinity,

$$\frac{Prob\{\underline{p}/\underline{q}\}}{Prob\{\underline{p}'/\underline{q}\}} = e^{-N\left[H(\underline{p};\underline{q})-H(\underline{p}';\underline{q})\right]} = e^{N\,\alpha},\ \alpha > 0 \tag{6.10}$$

This concludes the justification of the minimum relative entropy criterion.

(The argument above, the probability calculations and the term "typical sequence" are frequently used in Shannon's development of information theory [24], especially for his source coding theorem.)

6.2 MRE by alternate minimization and the EM algorithm

In this section, an interesting interpretation of the EM algorithm is provided, using the MRE criterion. This interpretation is based on the fact that the MRE criterion, used in a special way, reduces to the ML or the MAP criterion. Minimizing the MRE criterion is usually difficult; thus, the iterative alternate minimization (or coordinate search) algorithm may be suggested. In the special case where the MRE criterion reduces to the ML criterion, this alternate minimization algorithm reduces to the EM algorithm, where minimizing with respect to one density is equivalent to the E step, and minimizing with respect to the other density is equivalent to the M step.

As already mentioned, the minimization of the relative entropy by alternate minimization and its relation to the maximum likelihood criterion were originally suggested in [7] and [8]. The alternative minimization method and its properties were also developed in [73], where an explicit relation to the EM algorithm was established.

6.2.1 MRE by alternate minimization

As discussed above, the goal of any statistical inference process is to find a probability distribution, that explains the observed phenomenon under some a-priori knowledge and constraints. Suppose that the observations and the a-priori knowledge may be summarized into constraints on the desired probability density functions, i.e. that $p(x) \in \mathcal{P}$ and $q(x) \in \mathcal{Q}$ where $\mathcal{P}$, $\mathcal{Q}$ are sets of p.d.f. A version of the MRE criterion will solve our inference problem via the following functional minimization:

$$\hat{p}(x), \hat{q}(x) = \arg \min_{p(x)\in\mathcal{P},\, q(x)\in\mathcal{Q}} \mathcal{H}\left(p(x); q(x)\right) = \arg \min_{p(x)\in\mathcal{P},\, q(x)\in\mathcal{Q}} \int_X p(x) \log \frac{p(x)}{q(x)} dx \quad (6.11)$$

The minimization of the relative entropy in (6.11) is complicated in general. It is also a functional optimization problem; thus, numerical methods cannot be easily applied either. To solve this problem, an alternate minimization method (or coordinate search algorithm) is suggested. In this method, we generate a sequence of solutions $\{p^{(n)}(x), q^{(n)}(x)\}$ as follows:

- Start: guess $p^{(0)}(x), q^{(0)}(x)$
- Iterate, until some convergence criterion is achieved,

$$p^{(n+1)}(x) = \arg \min_{p\in\mathcal{P}} \mathcal{H}\left(p(x); q^{(n)}(x)\right) \quad (6.12)$$

$$q^{(n+1)}(x) = \arg \min_{q\in\mathcal{Q}} \mathcal{H}\left(p^{(n+1)}(x); q(x)\right) \quad (6.13)$$

Any alternate minimization algorithm has the desired monotonicity property. Thus, each iteration improves (in our case decreases) the goal function. Specifically,

$$\mathcal{H}\left(p^{(n+1)}(x); q^{(n+1)}(x)\right) \leq \mathcal{H}\left(p^{(n+1)}(x); q^{(n)}(x)\right) \leq \mathcal{H}\left(p^{(n)}(x); q^{(n)}(x)\right) \quad (6.14)$$

The monotonicity property implies that, if the goal function is bounded, then it also converges to some value H^*. A continuous function, like the minimum relative entropy, will be bounded when it is defined over a compact set. For our iterative algorithm, in order to show that $\mathcal{H}(p^{(n)}; q^{(n)}) \to \mathcal{H}^*$, it is sufficient to show that the set $\mathcal{P}_0 \times \mathcal{Q}_0$, where

$$\mathcal{P}_0 \times \mathcal{Q}_0 = \left\{(p \in \mathcal{P}, q \in \mathcal{Q} \mid \mathcal{H}(p; q) \leq \mathcal{H}(p^{(0)}; q^{(o)})\right\} \tag{6.15}$$

is compact.

Differentiability of the goal function will imply that H^* is a stationary value. If the sets $\mathcal{P}$ and $\mathcal{Q}$ are convex, then the convergence point H^* is a global minimizer. Unfortunately, in the special case where the desired density functions are defined parameterically, these sets are rarely convex, even when the possible set in the parameter space is convex.

A comprehensive discussion on the properties of this algorithm, especially the convergence issues, may be found in [8] pp. 107-133 and in [73].

6.2.2 ML as a special case of the MRE criterion

Let X be a sample space, referred to as the complete data sample space, and suppose that a parametric family of probability distributions is defined over it. This parametric family is indexed by the vector $\underline{\theta} \in \Theta$, where Θ is a subset of the k-dimensional Euclidean space. Thus, the set of possible complete data densities is

$$\mathcal{Q}_\Theta = \{q(\underline{x}; \underline{\theta}) \mid \underline{\theta} \in \Theta\} \tag{6.16}$$

Suppose that we do not observe $\underline{x} \in X$, but instead we observe an incomplete data $\underline{y} = T(\underline{x})$, where $T(\cdot)$ is a many-to-one mapping. The new sample space, given the incomplete data observations, is $\mathcal{X}(\underline{y})$ as in (2.4). The possible p.d.f., given $\underline{y}$, are therefore constrained to the set

$$\mathcal{P}_{\underline{y}} = \left\{p(\cdot) \mid p(\underline{x}) = 0 \; \forall \underline{x} \notin \mathcal{X}(\underline{y}), \; \int_{\mathcal{X}(\underline{y})} p(\underline{x}) d\underline{x} = 1\right\} \tag{6.17}$$

Note that a sample space, Y, for the observation also exists, and each $\underline{\theta}$ defines a p.d.f. over this space, given by

$$f_Y(\underline{y};\underline{\theta}) = \int_{\mathcal{X}(\underline{y})} q(\underline{x};\underline{\theta})d\underline{x} \tag{6.18}$$

The probability densities, $q(\cdot) \in \mathcal{Q}_\Theta$ and $p(\cdot) \in \mathcal{P}_{\underline{y}}$, are estimated via the minimum relative entropy criterion. Thus, we have to solve

$$\hat{p}(\underline{x}), \hat{q}(\underline{x}) = \arg \min_{p(\underline{x}) \in \mathcal{P}_{\underline{y}},\ q(\underline{x}) \in \mathcal{Q}_\Theta} \int_X p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x})} d\underline{x} \tag{6.19}$$

Note that estimating $q(\cdot) \in \mathcal{Q}_\Theta$ is equivalent to estimating $\underline{\theta} \in \Theta$, i.e. we have to solve

$$\hat{p}(\underline{x}), \hat{\underline{\theta}}_{MRE} = \arg \min_{p(\underline{x}) \in \mathcal{P}_{\underline{y}},\ \underline{\theta} \in \Theta} \int_X p(\underline{x}) \log \frac{p(\underline{x})}{q(\underline{x};\underline{\theta})} d\underline{x} \tag{6.20}$$

where $\hat{\underline{\theta}}_{MRE}$ denotes the minimum relative entropy estimator of the parameters.

The ML criterion determines the estimate $\underline{\theta} \in \Theta$ by maximizing the likelihood of the observation, i.e.

$$\hat{\underline{\theta}}_{ML} = \arg \max_{\underline{\theta} \in \Theta} f_Y(\underline{y};\underline{\theta}) = \arg \max_{\underline{\theta} \in \Theta} \int_{\mathcal{X}(\underline{y})} q(\underline{x};\underline{\theta})d\underline{x} \tag{6.21}$$

We will now show that estimating $\underline{\theta}$ by the minimum relative entropy (i.e. by 6.20) is equivalent to estimating $\underline{\theta}$ by maximum likelihood (i.e. by 6.21).

For any fixed $\underline{\theta}$ we will minimize (6.20) with respect to $p(\underline{x})$. This minimization problem may be solved explicitly using the following lemma, which is a direct result of the convexity of the relative entropy function.

Lemma 6.1 *For any measurable set* Λ

$$\int_\Lambda p(\underline{x}) \frac{p(\underline{x})}{q(\underline{x})} d\underline{x} \geq P(\Lambda) \log \frac{P(\Lambda)}{Q(\Lambda)} \tag{6.22}$$

where $P(\Lambda) = \int_\Lambda p(\underline{x})d\underline{x}$ *and* $Q(\Lambda) = \int_\Lambda q(\underline{x})d\underline{x}$. *Equality holds if and only if*

$$\frac{p(\underline{x})}{P(\Lambda)} = \frac{q(\underline{x})}{Q(\Lambda)}, \quad a.e \text{ in } \Lambda \tag{6.23}$$

The proof of this lemma may be found in [8] page 373.

Let A be the set $\mathcal{X}(\underline{y})$. In this case, $P(\mathcal{X}(\underline{y})) = 1$ and $Q(\mathcal{X}(\underline{y})) = f_Y(\underline{y};\underline{\theta})$. By the lemma above, a probability density function that, will satisfy (6.23), will minimize the relative entropy. Thus the $p(\underline{x})$ that minimizes (6.20) for any fixed $\underline{\theta}$ is

$$\hat{p}(\underline{x}) = P(\mathcal{X}(\underline{y}))\cdot\frac{q(\underline{x};\underline{\theta})}{Q(\mathcal{X}(\underline{y}))} = \frac{q(\underline{x};\underline{\theta})}{f_Y(\underline{y};\underline{\theta})} = f_{X/Y}(\underline{x}/\underline{y};\underline{\theta}) \tag{6.24}$$

which is the conditional probability of $\underline{x}$ given $\underline{y}$. The value of the relative entropy at this point is given again by the lemma above (6.22), i.e.

$$P(\mathcal{X}(\underline{y}))\log\frac{P(\mathcal{X}(\underline{y}))}{Q(\mathcal{X}(\underline{y}))} = -\log f_Y(\underline{y};\underline{\theta}) \tag{6.25}$$

The relations above and the equivalence of the minimum relative entropy estimator, $\hat{\underline{\theta}}_{MRE}$, and the maximum likelihood estimator, $\hat{\underline{\theta}}_{ML}$, are summarized in the following equation:

$$\hat{\underline{\theta}}_{MRE} = \arg\min_{\underline{\theta}\in\Theta}\left[\min_{p(\underline{x})\in P_{\underline{y}}}\int_{\mathcal{X}(\underline{y})} p(\underline{x})\log\frac{p(\underline{x})}{q(\underline{x};\underline{\theta})}d\underline{x}\right] = \arg\min_{\underline{\theta}\in\Theta}\left[-\log f_Y(\underline{y};\underline{\theta})\right] = \hat{\underline{\theta}}_{ML} \tag{6.26}$$

6.2.3 The EM as an alternate minimization algorithm

Since the MRE criterion reduces to the ML criterion in the special case summarized above, applying the alternate minimization algorithm of (6.12) and (6.13), will provide an iterative algorithm for maximum likelihood. This iterative algorithm is the EM algorithm, as shown below.

In each step of the alternate minimization algorithm, we have the current estimates, $p^{(n)}(\underline{x})$ and $q^{(n)}(\underline{x}) = q(\underline{x};\underline{\theta}^{(n)})$, of the p.d.f.'s. Applying first (6.12), we get

$$p^{(n+1)} = \arg\min_{p(\underline{x})\in P_{(\underline{y})}} \mathcal{H}\left(p(\underline{x}); q(\underline{x};\underline{\theta}^{(n)})\right) = f_{X/Y}(\underline{x}/\underline{y};\underline{\theta}^{(n)}) \tag{6.27}$$

where this explicit solution is based, again, on lemma 6.1. Now applying the second step, i.e. (6.13), we get

$$\underline{\theta}^{(n+1)} = \arg\min_{\underline{\theta}\in\Theta} \mathcal{H}\left(p^{(n+1)}(\underline{x}); q(\underline{x};\underline{\theta})\right) = \arg\min_{\underline{\theta}\in\Theta} \int_{\mathcal{X}(\underline{y})} f_{X/Y}(\underline{x}/\underline{y};\underline{\theta}^{(n)}) \log \frac{f_{X/Y}(\underline{x}/\underline{y};\underline{\theta}^{(n)})}{q(\underline{x};\underline{\theta})} \tag{6.28}$$

Using the notations of (2.11), we may write

$$\underline{\theta}^{(n+1)} = \arg\min_{\underline{\theta}\in\Theta} H(\underline{\theta}^{(n)},\underline{\theta}^{(n)}) - Q(\underline{\theta};\underline{\theta}^{(n)}) = \arg\max_{\underline{\theta}\in\Theta} Q(\underline{\theta};\underline{\theta}^{(n)}) \tag{6.29}$$

Equation (6.29) is exactly an iteration of the EM algorithm !

6.2.4 Remark: how not to use the MRE criterion

The relative entropy is sometimes interpreted as a distance measure between two probability distributions: the "Kullback-Leiber" measure. However, it lacks one of the desired features of a distance measure, namely, it is not symmetric,

$$\mathcal{H}(\underline{p};\underline{q}) \neq \mathcal{H}(\underline{q};\underline{p}) \tag{6.30}$$

Furthermore, following the common rationale of the MRE method, the relative entropy has meaning only for comparing possible probability measures, $\underline{p}$, given an a-priori assignment, $\underline{q}$. Thus, we prefer to interpret the relative entropy as representing the conditional likelihood of an assignment $\underline{p}$, given the assignment $\underline{q}$, as summarized in equation (6.9), i.e.

$$\mathcal{H}(\underline{p};\underline{q}) = -\frac{1}{N}\log Prob\{\underline{p}/\underline{q}\} \tag{6.31}$$

Having this interpretation, it makes sense to minimize the relative entropy with respect to the first argument, $\underline{p}$, only. Unfortunately, the alternative derivation of the EM algorithm is based on minimizing the relative entropy with respect to both $\underline{p}$ and $\underline{q}$. Thus, with respect to this derivation, we agree that for any given probability assignment, $f_X(\underline{x})$, for the complete data X, the best assignment over the set $\mathcal{X}(\underline{y})$ is the conditional density, $f_{X/Y}(\underline{x}/\underline{y})$. However, minimizing then the relative entropy with respect to $f_X(\underline{x})$, in this case, is not justified.

This remark does not detract from the mathematical elegance and the additional insight that may be gained by the alternative derivation of the EM algorithm, but it does suggest that justifying the maximum likelihood criterion via this relation is a poor use of the MRE criterion.

6.3 Minimum Description Length interpretation of the MRE criterion

Despite the criticism, the MRE criterion is used and justified in several statistical problems. It can estimate an entire probability distribution function. It is also the basis of the alternative derivation of the EM algorithm. Thus an additional interpretation of the MRE method is desirable.

After a brief review of the philosophical idea of the Minimum Description Length (MDL) and the Minimum Information (MI) criteria, we will prove the main result of this chapter, namely, that the MRE criterion is a special case of the MDL criterion, in a certain context. On one hand, this result clarifies the appropriate context in which the MRE should be used. On the other hand it motivates and supports the MDL criterion by showing that, in the appropriate special case, it reduces to another proven criterion.

The idea of complete and incomplete data specifications that is so important in the development of the EM algorithm, also plays an important role in the definition and the proof of this relationship between the MDL and MRE criteria. For showing the relation between the MDL and the MRE criteria, the MDL criterion is used in a mode where it considers a *set* of possible observation. An immediate example for such situation is the set $\mathcal{X}(\underline{y})$ of the possible complete data, given an observation $\underline{y}$, which is used in the EM algorithm context.

6.3.1 The Minimum Description Length idea

We have already mentioned, in Chapter 2, the Minimum Description Length criterion, suggested by Rissanen [29,30,31] and the more general Minimum Information criterion, suggested originally by Solomonoff [21] and recently by Hart [22]. The philosophical foundation of the MDL/MI methods is the claim that the most compact description of the observation provides the best explanation of the phenomena we observe. In other words, if we have the best method to encode or compress the observation, we have actually estimated the probability distribution that "explains" the data best. This philosophy is intuitively reasonable and is consistent with universal philosophical principles, such as the *Ockham Razor* principle. We strongly believe in these "principles of parsimony", and, since these principles can be made precise by the quantitative measures of information and complexity, we strongly advocate their use.

Using these criteria, we may overcome some of the limitations of the ML method and generate methods that can accept less restrictive modeling assumptions. For example, the ML method fails, when the number of parameters is unknown, since the more parameters we choose, the larger the likelihood can be. In this case, a specific application of the MDL criterion tries to estimate the parameters together with their number by, (see also (2.73) and [31])

$$\hat{n}, \hat{\underline{\theta}} = \arg\min_{n,\underline{\theta}} [-\log p(x; \underline{\theta}) + \frac{1}{2} n \log N] \tag{6.32}$$

where N is the length of the observation sequence.

In the MDL method, the abstract principle of shortest description is translated into a mathematical criterion in the following way. The description (or code) length above is influenced by two factors. If we knew the probability distribution, the "ideal" code length [76], that is required to represent the specific observation, is the (self) information of the observation, i.e. $-\log p(x)$. The second factor, $\frac{1}{2} n \log N$, is the code length needed to represent the model or the parameters, considering the precision required for encoding continuous parameters.

To show that the ME and MRE methods are special cases of the MDL criterion, we have to extend the MDL method somewhat. Suppose that the information given about the underlying phenomena is not a single observation sequence, but rather a set of such sequences. This type of information is available either by having several independent observation sequences or by having constraints, that define a possible set of observation sequences. The MDL criterion for this type of information will suggest that we choose the probability distribution that minimizes the weighted combination of all code lengths, by some a-priori weight $q(x)$, where all members of this set of possible observation sequences are encoded using the proposed distribution.

We can now adapt the MDL criterion to the MRE framework, in which the given information about the underlying phenomena is in terms of constraints on the probability distribution. We claim that if we try to represent all possible observation sequences, whose "histogram" or sample frequencies satisfy those constraints, the minimum weighted code length is achieved by the MRE probability distribution.

6.3.2 Minimum Relative Entropy as Minimum Description Length

In the MRE framework, the given information is the knowledge of some averages. Now recall that the strong law of large numbers implies,

$$E[g(x)] = \bar{g} \iff \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} g(x_i) = \bar{g} \;\; a.s., \tag{6.33}$$

where x_i are i.i.d. observations, distributed as x. So, this MRE-type information is equivalent to the information that the observations lie in the set of all infinitely long sequences whose sample averages equals the given averages.

Considering the above argument, we will show that minimizing the relative entropy, subject to some constraints on the probability distribution, is asymptotically equivalent to minimizing the weighted combined code length needed to represent all the sequences, whose "histogram" or sample frequencies satisfy the given constraints.

To clarify our argument, let us start with a simple example. Suppose we want to estimate the probability of "1" (success) in a simple binary (Bernoulli) trial. We denote $p(\text{"1"}) = \theta$ and $p(\text{"0"}) = 1 - \theta$. Suppose we do not have any observations, so we only know that for any N trials, that we will perform, we may observe any of the 2^N possible sequences of "1"s and "0"s.

Equipped with the Minimum Description Length philosophy, and applying a uniform weight to all code lengths, we will choose θ so that all 2^N sequences can be represented by the shortest possible code. Now, for each sequence $\underline{x}$ we need about

$$-\log p(x;\theta) = -\log[\theta^k(1-\theta)^{N-k}] \tag{6.34}$$

bits, where k in the number of "1"s in x. We will denote by $\mathcal{X}$ the set of all such sequences and by $L(\mathcal{X})$ the combined code length required to represent all members of the set. Now for any k we have $\begin{pmatrix} N \\ k \end{pmatrix}$ sequences with k "1"s, so that

$$L(\mathcal{X}) = \sum_{k=0}^{N} \begin{pmatrix} N \\ k \end{pmatrix} [-\log \theta^{k}(1-\theta)^{N-\kappa}] =$$

$$= -\left[\sum_{k=o}^{N} \begin{pmatrix} N \\ k \end{pmatrix} k\right] \log\theta \; - \left[\sum_{k=0}^{N} \begin{pmatrix} N \\ k \end{pmatrix} (N-k)\right] \log(1-\theta) \qquad (6.35)$$

Noting that

$$\sum_{k=0}^{N} \begin{pmatrix} N \\ k \end{pmatrix} k = \sum_{k=0}^{N} \begin{pmatrix} N \\ k \end{pmatrix} (N-k) = N2^{N-1} \equiv \alpha$$

We see that

$$L(\mathcal{X}) = -\alpha\log\theta - \alpha\log(1-\theta) \qquad (6.36)$$

which is minimized, as expected, by $\theta = 1/2$. Observe that this probability function is the same as that given by the Maximum Entropy principle (or the MRE principle with uniform prior) with no constraints, on the binary random variable.

Notice that here we have ignored the term $1/2\log N$, required to represent the code length for describing the single parameter θ, because it has no effect on the minimizing value.

We are ready now to prove the general claim, stated as the following theorem.

Theorem 6.1 *Let X be a random variable that takes its values over the finite set $\{1, \cdots, m\}$. Let $\underline{x} = x_1x_2\cdots x_N$ be a sample of N independent trials of X. Let $f_i(\underline{x}) = k_i(\underline{x})/N$ be the frequencies of each outcome in this sample. The vector $\underline{f}(\underline{x}) = [f_1(\underline{x}), \cdots, f_m(\underline{x})]^T$ will be called the histogram of the sample. Let $\mathcal{F}$ be any fixed set of histograms. Let $\mathcal{X}_N$ be the set*

$$\mathcal{X}_N = \{\underline{x} = x_1\cdots x_N \mid \underline{f}(\underline{x}) \in \mathcal{F}\}.$$

Let the weighted code length, with the weights $\underline{q} = [q_1, \cdots, q_m]$, that results when the whole set $\mathcal{X}_N$ is encoded by a code book designed according to $\underline{p} = [p_1, \cdots, p_m]$, be denoted $L(\mathcal{X}_N, \underline{p}, \underline{q})$ where

$$L(\mathcal{X}_N, \underline{p}, \underline{q}) = \sum_{\underline{x} \in \mathcal{X}_N} q(\underline{x}) \left[-\log p(\underline{x})\right]$$

Then the probability assignment $\underline{p} = [p_1, \cdots, p_m]$ that minimizes $L(\mathcal{X}_N, \underline{p}, \underline{q})$ is,

$$\hat{\underline{p}}_N = \frac{\sum_{\underline{f} \in \mathcal{F}} \frac{q_1^{k_1} \cdots q_m^{k_m} \cdot N!}{k_1! \cdots k_m!} \underline{f}}{\sum_{\underline{f} \in \mathcal{F}} \frac{q_1^{k_1} \cdots q_m^{k_m} \cdot N!}{k_1! \cdots k_m!}} \tag{6.37}$$

Furthermore, as $N \to \infty$

$$\hat{\underline{p}} = \lim_{N \to \infty} \hat{\underline{p}}_N = \arg\min_{\underline{p} \in \mathcal{F}} H(\underline{p}; \underline{q}) = \arg\min_{\underline{p} \in \mathcal{F}} \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} \tag{6.38}$$

Proof: The code length required to represent a sample, $\underline{x} = x_1 \cdots x_N$, in a code book designed by $\underline{p}$ (within the term $m/2 \log N$ required to encode the probabilities p_i) is,

$$L(\underline{x}) = -\log p(\underline{x}) = -\log \left(\prod_{i=1}^{m} p_i^{k_i} \right) = -\sum_{i=1}^{m} k_i \log p_i \tag{6.39}$$

where k_i is the number of occurrences of the i^{th} outcome in $\underline{x}$. The weighted code length is given by

$$L(\mathcal{X}_N, \underline{p}, \underline{q}) = \sum_{\underline{x} \in \mathcal{X}_N} q_1^{k_1} \cdots q_m^{k_m} \left(-\sum_{i=1}^{m} k_i \log p_i \right) \tag{6.40}$$

Now, there are $\frac{N!}{k_1! \cdots k_m!}$ possible sequences having the same frequencies or the same number of occurrences, $\underline{k} = [k_1, \cdots, k_m]^T$. Since the constraints are only on the frequencies (or on $\underline{k}$), we can write the weighted code length as,

$$L(\mathcal{X}_N, \underline{p}, \underline{q}) = \sum_{\text{admissible } \underline{k}} \frac{q_1^{k_1} \cdots q_m^{k_m} \cdot N!}{k_1! \cdots k_m!} \left(-\sum_{i=1}^{m} k_i \log p_i \right) = -\sum_{i=1}^{m} \left(\sum_{\underline{f} \in \mathcal{F}} \frac{q_1^{k_1} \cdots q_m^{k_m} \cdot N!}{k_1! \cdots k_m!} k_i \right) \log p_i \tag{6.41}$$

We will denote

$$\beta_i = \sum_{\underline{f} \in \mathcal{F}} \frac{q_1^{k_1} \cdots q_m^{k_m} \cdot N!}{k_1! \cdots k_m!} k_i \tag{6.42}$$

The weighted code length is thus

$$L(\mathcal{X}_N, \underline{p}, \underline{q}) = -\sum_{i=1}^{m} \beta_i \log p_i \tag{6.43}$$

which is minimized (using Jensen's inequality) by

$$\hat{p}_{N,i} = \frac{\beta_i}{\sum_{l=1}^{m} \beta_l} \tag{6.44}$$

Substituting (6.42) in (6.44) and recalling that $k_i / \sum_{l=1}^{m} k_l = f_i$, yields (6.37).

In general, we will get the MRE distribution only in the limit as $N \rightarrow \infty$ as follows. Following the derivation of (6.7) and (6.8), we get

$$\frac{q_1^{k_1} \cdots q_m^{k_m} \cdot N!}{k_1! \cdots k_m!} = e^{-N[\mathcal{H}(\underline{f};\underline{q}) + o(\frac{\log N}{N})]} \tag{6.45}$$

where $\mathcal{H}(\underline{f};\underline{q}) = \sum_{i=1}^{m} f_i \log \frac{f_i}{q_i}$ is the relative entropy between the frequencies $f_i = k_i/N$ and q_i. Substituting (6.45) in (6.37) and taking the limit as $N \rightarrow \infty$ yields

$$\underline{\hat{p}} = \lim_{N\rightarrow\infty} \frac{\sum_{\underline{f}\in\mathcal{F}} \underline{f} e^{-N\cdot\mathcal{H}(\underline{f};\underline{q})}}{\sum_{\underline{f}\in\mathcal{F}} e^{-N\mathcal{H}(\underline{f};\underline{q})}} \tag{6.46}$$

Let us assume that the function $\mathcal{H}(\underline{f};\underline{q})$ has in $\mathcal{F}$ a single global minima, at $\underline{f}_{min}$. Now as $N \rightarrow \infty$

$$\lim_{N\rightarrow\infty} e^{-N[\mathcal{H}(\underline{f};\underline{q}) - \mathcal{H}(\underline{f}_{min};\underline{q})]} = \begin{cases} 0 & if \;\; \underline{f} \neq \underline{f}_{min} \\ 1 & if \;\; \underline{f} = \underline{f}_{min} \end{cases} \tag{6.47}$$

so, we can write (6.46)

$$\lim_{N\rightarrow\infty} \frac{\sum_{\underline{f}\in\mathcal{F}} \underline{f} e^{-N\cdot\mathcal{H}(\underline{f};\underline{q})}}{\sum_{\underline{f}\in\mathcal{F}} e^{-N\mathcal{H}(\underline{f};\underline{q})}} = \lim_{N\rightarrow\infty} \frac{\sum_{\underline{f}\in\mathcal{F}} \underline{f} e^{-N[\mathcal{H}(\underline{f};\underline{q}) - \mathcal{H}(\underline{f}_{min};\underline{q})]}}{\sum_{\underline{f}\in\mathcal{F}} e^{-N[\mathcal{H}(\underline{f};\underline{q}) - \mathcal{H}(\underline{f}_{min};\underline{q})]}} = \underline{f}_{min} \tag{6.48}$$

i.e. $\underline{\hat{p}}$ is the MRE distribution. □

Note that in general, if $\mathcal{H}(\cdot; \underline{q})$ has several global minima, $\underline{f}_1, \cdots, \underline{f}_n$, then the result in (6.46) will be

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \underline{f}_i$$

However, since $H(\cdot; \underline{q})$ is convex, whenever the constraint set is convex, there is only a single minimum.

We claim that the above theorem can be extended, following the same lines of proof, to the case where the random variable takes its values over an infinite set. The relation between the MDL and the MRE criteria is thus fully established.

Chapter 7

Conclusions and further research

In this final chapter, we will summarize and discuss the results and the contributions of the thesis and try to convey a general philosophy for solving signal processing problems that may be established from this thesis. Before that, we will suggest further research directions. A particularly interesting suggestion for further research, for which we have specific ideas, is the application of the EM method to the the problem of signal reconstruction from partial information.

7.1 Further research

Many research directions may be suggested to complete and extend the work presented in this thesis. Interesting theoretical problems as well as interesting signal processing applications may be explored further. In this section, we will indicate a few of these research projects. We will start by discussing an important subject, that has not been explored enough in the thesis, namely, the analysis and the applications of the sequential and adaptive EM algorithms.

We have developed specific ideas and mathematical formulations for solving problems of signal reconstruction from partial information, in the EM context. These ideas should be investigated further and should be applied to real reconstruction problems. We will present these ideas and the further proposed research in the next section.

7.1.1 The sequential and adaptive algorithms

Considering the theory of the sequential EM algorithm, we suggest the following research directions:

- **Non asymptotic statistical analysis**

  We have considered only limit distribution and consistency. However, interesting questions arise in the non-asymptotic case. A complete analysis of, say, the variance of the estimator as the iteration proceeds is desirable.

- **Rate of convergence and tracking**

  A topic that is close, but not identical, to the previous topic, is the convergence rate of the sequential and adaptive algorithms. For the adaptive algorithms we are interested in the tracking capabilities, which improve if the algorithm converge faster.

- **Limit distribution for non i.i.d. case**

  This research topic will complete the results presented in Chapter 3.

- **Other approximations**

  We have suggested, in Chapter 3, sequential EM algorithms, based on a specific approximation of the batch EM algorithm. However, there may also be other possible approximations that should be investigated.

The sequential and adaptive algorithms developed in the thesis have not been applied yet, in a serious way, to signal processing problems. As a first step, we suggest that these algorithms be applied extensively to the problems that have been solved in Chapters 4 and 5 of the thesis, via the batch EM algorithm. However, other signal processing problems also call for adaptive solutions.

We may want to start with a simpler example. Consider the problem of a stationary signal in a stationary noise, where the suggested batch EM algorithm is the iterative Wiener filter algorithm. Assume that, given the signal without the noise, the parameters may be estimated sequentially, say, by the RLS algorithm. It will be interesting to try a recursive algorithm that uses a Kalman filter (instead of a Wiener filter) to get the signal, and then use this signal recursively, to estimate the parameters.

Other examples, that come to mind, are parameter estimation of dynamic systems, the problem of tracking the trajectories of multiple targets using tracking radar, analysis of sequences of images and so on.

7.1.2 Other research directions

Here, we will briefly present other research directions that come to mind, both at the theoretical level and the signal processing application level. We will start with the theoretical research directions:

- **Global Optimization:**

  The EM algorithm can guarantee convergence only to a stationary point of the likelihood. One might investigate the combination of the EM algorithm with standard methods, summarized in the book [77]. Recently, a new technique, the simulated annealing [78], has become increasingly popular. The EM algorithm ideas may be combined with this technique to achieve an algorithm for global optimization.

- **The EEM algorithm:**

  The idea of changing the complete data, presented in Chapter 2, should be investigated further. The rules suggested for varying the complete data could be stated more concisely and their properties should be analyzed. This research direction may be combined with the previous one, since one of the motivations for varying the complete data is to escape from unwanted stationary points.

- **EM algorithms for general estimation criteria:**

  We have presented in the thesis the EM method for a class of estimation criteria. Further research may extend this class. The properties of this method for general estimation criteria, should be further explored.

- **Other iterative algorithms for ML:**

  Recently, other iterative algorithms for ML estimation, given incomplete data, have been suggested by statisticians, based on some ideas from the EM theory, e.g. [14]. These algorithms should be explored further; their sequential versions could be developed and applied to signal processing problems.

Applications

Many further signal processing applications of the EM method can be considered. We will briefly present here some examples.

- **Separating a narrow band signal from a wide band signal**

This problem occurs in many practical situations, such as the problem of enhancing the periodic acoustical signal of an helicopter in a wide band noise of a jet plane. This problem is analogous to the problem of filtering a stationary signal from a stationary noise, solved using the iterative Wiener filter. However, in this problem, we cannot model the narrow band signal as a stationary Gaussian signal. Modeling this problem correctly, maybe with the EM algorithm in mind, and solving the resulting statistical problem, is an interesting topic for further research.

- **Separating two speakers: The "Cocktail Party" problem:**

  This problem is analogous to the previous problem. However, we do not expect to gain much by modeling the speech signal as a Gaussian stationary signal, since the spectral distribution of both speakers is identical. We suggest that by modeling the speech signals using their periodic nature, we may be able to distinguish the speakers based on their different periodicities and phases. The resulting statistical problem may be complicated. However, it might be solved using the EM method.

- **Joint estimation of pitch and spectral parameters of speech:**

  Usually the pitch and the LPC parameters of the speech signal are estimated independently. Furthermore the LPC parameters are estimated by modeling the speech signal as a stationary AR process, a model that is clearly not adequate for voiced speech. We suggest using the pulse excitation model. We will choose the complete data to be the pulse train, modeled as a stochastic point process, in addition to the observed speech signal. An EM algorithm for estimating the pitch and LPC parameters might be suggested, using this complete data.

Following the guidelines presented in section 2.6 and these examples, many more signal processing application could be suggested.

7.2 Signal reconstruction from partial information: Ideas and proposed research

In this section, we will present specific ideas and the mathematical formulation for solving problems of signal reconstruction from partial information, using statistical modeling and the EM algorithm.

7.2.1 General discussion

The problem of signal reconstruction from partial information has been investigated by many researchers, e.g. [79,80,81,82] to mention a few. Traditionally, a deterministic formulation of this problem has been adopted where the given partial information provided (non-linear) deterministic equations and constraints for the unknown signal samples. A major research effort was allocated to answer the questions of existence and uniqueness of a solution, and, as a result, statements such as "Phase retrieval is impossible for one dimensional signal however it is possible for a two dimensional signal", were declared. Other research efforts led to algorithms which perform the reconstruction task by finding solutions that satisfy the constraints and the equations, either directly, e.g. [83], or via iterative procedures, e.g. [5],[84], [85].

This deterministic approach assumes, at least implicitly, noiseless measurements. The effects of the noise, which may result from measurement and computation errors, were considered only by investigating the robustness of the algorithms designed for the noiseless case. It has been observed that some reconstruction problems, for which a solution exist, have shown poor noise immunity, and thus are probably ill-posed and practically unsolvable in this deterministic framework. For example, the reconstruction of a two dimensional signal from its Fourier transform magnitude is an open problem, despite the fact that a unique solution to this problem exists. In general, one can calculate condition numbers for deterministic reconstruction algorithms and predict his chances of solving a specific reconstruction problem in a real situation.

We will suggest a statistical formulation of the reconstruction problem. In this formulation, we model the noise, measurement or computation noise, naturally. The signal reconstruction problem becomes a statistical estimation problem, for which well known performance bounds, like the Cramer-Rao bound, exist. These performance bounds play the role of the condition numbers in the deterministic formulation. However, the statistical performance bounds provide more information and insight.

The performance of a reconstruction problem can be improved by incorporating a-priori information about the signal. An important advantage of the statistical formulation is that a wide class of a-priori information may be easily incorporated. We note that in the deterministic formulation, regularization methods were suggested in an attempt to improve the performance. Some of the methods to regularize ill-posed deterministic problems are equivalent to assigning simple a-priori probabilities to the signal in the stochastic formulation.

The statistical problems, in the proposed formulation, require maximizing the likelihood, the a-posteriori probability or other statistical criterion depending on the a-priori information. These problems are naturally solved using the EM method. Since the observations are partial and distorted, the complete data is the undistorted signals. For some ML problems, where no special a-priori information is incorporated, the resulting EM algorithm is equivalent to standard iterative algorithms such as Gerchberg-Saxton algorithm [5], and thus, unfortunately, has a similar poor performance in ill-posed reconstruction problems. However, we believe that open reconstruction problems may become well behaved by using statistical models that incorporate enough realistic information. The EM method will suggest a practical solution to the resulting statistical problems.

7.2.2 Statistical formulation of signal reconstruction problems

The first element in the suggested statistical formulation, is a definition of the quantities that should be estimated and inferred. These quantities may be the signal samples, the samples of some underlying "hidden" process or some unknown parameters. We will denote these quantities by the vector $\underline{s}$.

The next element in the suggested framework is the definition of a stochastic process, denoted $\underline{x}$, that depends on the desired quantities in a simple way, i.e.

$$\underline{x} = \mathcal{F}(\underline{s}) \tag{7.1}$$

where $\mathcal{F}$ is a stochastic function, which defines a (simple) conditional probability, $f_{\underline{X}/\underline{S}}(\underline{x}/\underline{s})$. If $\underline{x}$ is observed, $\underline{s}$ may be estimated easily. We admit that the statistical formulation is defined having the EM idea in mind; this stochastic process will be used as the complete data or as part of the complete data in the suggested EM algorithm.

Another element is the measurement procedure. In defining this element, we will model the partial information aspect and the measurement noise aspect of the real reconstruction problem. Denote the observation by $\underline{y}$. We may write,

$$\underline{y} = H(\underline{x}) + \underline{v} \tag{7.2}$$

where $H(\cdot)$ is a non-invertible transformation representing the fact that only partial information is available. The vector, $\underline{v}$, represents the measurements errors. With a probabilistic description of $\underline{v}$, we achieve a stochastic description of the observations in terms of the desired quantities, via the conditional probability, $f_{\underline{Y}/\underline{S}}(\underline{y}/\underline{s})$.

Using the stochastic description of the observations and an appropriate statistical criterion, determining the desired quantities reduces to solving a mathematical optimization problem. If this problem is ill-posed, then additional information or assumptions and maybe a different statistical criterion should be considered. In the statistical framework, the a-priori knowledge about the desired quantities can be easily incorporated, in a quantitative manner. The a-priori knowledge will also define the statistical criterion which will be used. This element of the statistical formulation is important, since, by incorporating additional information, it is possible to solve real reconstruction problems, which are ill-posed otherwise.

We will now present three examples of different statistical models, that follow the descriptions above. All three models may be used for solving reconstruction problems. In the first example, the desired quantities, $\underline{s}$, are the samples of the signal or the pixels of the image to be reconstructed. Assuming a simple stochastic model, $\underline{x}$ is distributed normally with mean $F \cdot \underline{s}$ and variance $\sigma^2 I$, where F is an invertible linear transformation, say the Fourier transform. Thus, we may write

$$\underline{x} = F \cdot \underline{s} + \underline{n} \tag{7.3}$$

where $\underline{n}$ is a discrete white noise signal. Suppose, for example, that we want to reconstruct the signal from the magnitude of its Fourier transform. The measurements in this case will be modeled as,

$$[\underline{y}]_i = |[\underline{x}]_i| \tag{7.4}$$

where $[\cdot]_i$ denotes the i^{th} component of a vector.

The second modeling example comes from Radioactive and Positron Emission Tomography medical imaging problems. At each point i in (the discrete) space, there is a radiating Poisson source, with parameter λ_i. The desired quantities in this model, denoted $\underline{\lambda}$, will be these parameters. With a perfect imaging system, we may observe the vector $\underline{x}$, where its i^{th} component is the number of photons or particles emitted by the source in the i^{th} point, i.e. we may write,

$$f(\underline{x}/\underline{\lambda}) = \prod_i e^{-\lambda_i} \frac{\lambda_i^{x_i}}{x_i!} \tag{7.5}$$

Our imaging system is not perfect, however. In tomography, for example, we measure noisy projections of $\underline{x}$, which may be modeled as,

$$\underline{y} = H \cdot \underline{x} + \underline{v} \tag{7.6}$$

where H, the projection operator, is a non-invertible linear transformation. We note that, in this case, the relation between complete and incomplete data is linear. However, we cannot use the result of the Linear Gaussian case here, since $\underline{x}$ has Poisson distribution. Statistical models similar to the model above were suggested in medical tomography context in [86], [35] and elsewhere.

The third example is as follows. Modeling images using Markov Random Fields (MRF) has recently become increasingly popular. Using interesting simulation algorithms (e.g. [87],[88],[89]),[90]), realizations of MRF with various parameters were generated. These samples resembled realistic images surprisingly well. We strongly suggest reading [88] to see how realistic images with different characteristics can emerge by the innovative choice of the parameters of the MRF. Consider now the following statistical framework of the signal reconstruction problem. Let the desired quantities be the parameters of the MRF. denoted $\underline{\theta}$. The process $\underline{x}$ will be the image, i.e. it depends stochastically on $\underline{\theta}$, via the Gibbs distribution, see e.g. [89],

$$f_{\underline{X}}(\underline{x};\underline{\theta}) = \frac{1}{Z(\underline{\theta})} e^{-U(\underline{x};\underline{\theta})/T} \tag{7.7}$$

where Z is the normalization factor. and U is called the "energy function", where the neighborhood structure and the characteristic of the image is defined. The observations, $\underline{y}$, will be a noisy and incomplete function of $\underline{x}$, as in (7.2). The EM algorithm suggested in this formulation is directed at finding $\underline{\theta}$. However, as a possible by-product in the E step, an estimate of $\underline{x}$, the image itself, will be available.

7.2.3 Solution using the EM method

Solving the mathematical problems, generated by the statistical formulation, directly is generally complicated. However, since the statistical framework was suggested with the EM ideas in mind, EM iterative algorithms can be naturally applied. The complete data will be the set of signals $\{\underline{x}, \underline{y}\}$, i.e. it will include the undistorted signal $\underline{x}$ in addition to the observations $\underline{y}$. For the cases where the measurement noise $\underline{v}$ does not exist, the complete data will be the undistorted signal $\underline{x}$.

In the E step of the suggested algorithms, the conditional expectation of the sufficient statistics of $\underline{x}$ is calculated, given $\underline{y}$ and the current value of $\underline{s}$. For example, in the first statistical formulation above, if $\underline{x}$ is Gaussian with mean $\underline{s}$, the sufficient statistics is linear; thus, this conditional expectation may be easily derived. For problems such as reconstruction from Short Time Fourier Transforms, band limited extrapolation, reconstruction from projection and others, the observations, $\underline{y}$, are related to $\underline{x}$ by a non-invertible linear transformation, so we may use the results developed for the Linear Gaussian case.

The M step will be simple, since the process $\underline{x}$ depends on the unknown quantities in a direct way. For example, in the second statistical formulation of (7.5) and (7.6), we may estimate the desired quantities λ_i, having the complete data $\{x_i\}$, as

$$\hat{\lambda}_i = \arg\max_{\lambda_i} \{-\lambda_i + x_i \log \lambda_i\} = x_i \tag{7.8}$$

To fix ideas, we will present explicitly the EM iterative algorithm for maximum likelihood signal reconstruction from the magnitude of its Fourier transform, in the first statistical framework of (7.3) and (7.4). We note that a similar algorithm may be found in [8], pp 344-346.

We assume that the signal is real and has a given finite support. This a-priori knowledge is incorporated in the form of deterministic constraints. We will denote the signal to be reconstructed by $s(n)$ and its Fourier transform by $S(\omega)$. The complete data is given in the frequency domain by,

$$X(\omega) = S(\omega) + N(\omega) \tag{7.9}$$

where $N(\omega)$ is complex Gaussian random variable with variance σ^2. The observations are $Y(\omega) = |X(\omega)|$.

The E and M steps of the EM algorithm in the k^{th} iteration are given by,

- **The E step:** Given $s^{(k)}(n)$ or $S^{(k)}(\omega) = |S^{(k)}(\omega)| e^{j\theta_s^{(k)}(\omega)}$ find,

$$\begin{aligned} X^{(k)}(\omega) &= E\left\{X(\omega)/|X(\omega)| = Y(\omega), S^{(k)}(\omega)\right\} \\ &= Y(\omega)e^{j\theta_s^{(k)}(\omega)} \frac{\int_{-\pi}^{\pi} e^{j(Y(\omega)|S^{(k)}(\omega)|/\sigma^2)\cos\theta} \cos\theta d\theta}{\int_{-\pi}^{\pi} e^{j(Y(\omega)|S^{(k)}(\omega)|/\sigma^2)\cos\theta} d\theta} \\ &= Y(\omega)e^{j\theta_s^{(k)}(\omega)} \frac{I_1(Y(\omega)|S^{(k)}(\omega)|/\sigma^2)}{I_0(Y(\omega)|S^{(k)}(\omega)|/\sigma^2)} \end{aligned} \tag{7.10}$$

where $I_0(\cdot)$ is the zero order modified Bessel function and $I_1(\cdot)$ is the first order modified Bessel function.

- **The M step:** Let $x^{(k)}(n)$ be the inverse Fourier transform of $X^{(k)}(\omega)$, the updated estimates $s^{(k+1)}(n)$ are,

$$s^{(k+1)}(n) = \begin{cases} Re[x^{(k)}(n)] & \text{if } n \text{ is in the signal support} \\ 0 & \text{otherwise} \end{cases} \tag{7.11}$$

We note that since

$$\lim_{x \to \infty} \frac{I_1(x)}{I_0(x)} = 1$$

as the variance σ^2 tends to zero, the E step becomes

$$X^{(k)}(\omega) = Y(\omega)e^{j\theta_s^{(k)}(\omega)} \tag{7.12}$$

i.e. the complete data is estimated by combining the given magnitude with the current estimate of the phase. This algorithm was suggested in [79] and [82].

From this discussion, we gain a new interpretation of previously suggested algorithms. However, we also conclude that reconstructing the signal from the magnitude of its Fourier transform *cannot* be solved by maximizing the likelihood in this framework, since this leads us back to previous algorithms, which perform poorly.

As we repeatedly mentioned, for ill-posed problems more information should be incorporated. In the statistical framework, the information can be easily incorporated. Depending on this information, the adequate statistical criterion will be used. The resulting statistical problem for, say the reconstruction from magnitude problem, will be similar to (7.10) and (7.11), where the E step will be the same; however, in the M step, a different criterion will be invoked.

7.3 Summary and discussion

The thesis may be summarized as follows. We have solved signal processing problems using a class of iterative estimation algorithms. This class of algorithm is based on the EM method suggested in [2]. We, however, have extended this class and developed several general theoretical results. We will discuss the contributions made in this thesis in these two levels, the signal processing applications level and the theoretical contributions level.

7.3.1 The signal processing applications

When we discuss the application of the EM algorithm to a real world problem, we first have to model the problem statistically and then apply the EM algorithm to solve the resulting statistical problem. However, the EM algorithm is not uniquely defined; it depends on the choice of complete data, and, as we have seen, an unfortunate choice yields a completely useless algorithm. The choice of complete data or equivalently the choice of a specific EM algorithm requires creativity, in order to get a practically useful algorithm.

As a general philosophy, we will have the EM algorithm in mind, while suggesting a statistical model to the real signal processing problem. Using this philosophy, we will identify what the desired measurements are, model them statistically, and find their relation to the given observations. The statistical problem, generated this way, can then be solved using the EM algorithm; the desired measurements will be chosen, naturally, to be the complete data.

The main contribution of this thesis is the explicit solution of the important signal processing problems presented in Chapters 4 and 5. As we recall, these real problems are:

- Parameter estimation of superimposed signals. Applications of this model, that have been addressed, are,

  - Multiple source location (or bearing) estimation
  - Multipath or multi-echo time delay estimation

- Noise cancellation in a multiple microphone environment. T application considered is a speech enhancement problem.

In both cases, we have suggested solutions that improve upon the existing state of the art. In the superimposed signals application, the ML approach has been formulated before. However, since its solution is complicated, others have avoided it and suggested suboptimal or ad-hoc solutions. We have tackled this ML problem and succeeded in presenting a practical solution to it. In the noise canceling problem, our contributions include the formulation of the statistical ML problem to model different physical situations. Using the EM method, we were able to suggest practical solutions to the underlying real problem.

We may consider Chapters 4 and 5 as a demonstration of our suggested philosophy for solving signal processing problems. In these chapters, we have demonstrated this philosophy through all stages of the solution, from modeling, through the suggestion of an algorithm, to the numerical solution. Thus, these chapters will serve as a reference for further applications.

7.3.2 The theoretical contributions

The basic EM method has been suggested in [2]. However, in the process of considering the applications mentioned above, we have extended and modified the original EM algorithm. We have also derived explicit forms for some special cases. These extensions and derivations made the method more suitable for signal processing applications. We will now present and discuss these contributions.

- **The Linear Gaussian case**

  We derived closed form analytical expressions for the EM algorithm for the case where the complete and incomplete data are jointly Gaussian, related by a linear transformation. We note that, in general, a closed form analytical expression cannot be obtained and that the EM algorithm may require complex operations like multiple integration. In retrospect, this derivation appears to be a significant contribution, since it covers a wide range of applications

- **EM algorithms for general estimations criteria**

  Originally, the EM algorithm was developed and suggested as a technique for maximizing the likelihood. However, other criteria are more appropriate for some problems. We have developed EM algorithms for optimizing other criteria, specifically the Minimum Information criterion. We note that, in Chapters 2 and 6 of the thesis, a general discussion on the Minimum Information criterion, its properties and its relations to other statistical methods, is presented.

- **Extended EM: varying the complete data in each iteration**

  As mentioned above, the choice of the complete data may critically affect the complexity and the rate of convergence of the algorithm. It may also affect the convergence point, leading to a different stationary point for different choices of complete data. We have suggested, in the thesis, an interesting alternative to a fixed choice of complete data: we suggest varying the definition of the complete data in each step of the algorithm. This way, we may get simpler schemes, we may get algorithms that converge faster or algorithms that may escape from unwanted stationary points.

- **Sequential and Adaptive versions**

  Sequential and adaptive versions of the EM algorithm have been developed in Chapter 3 and some of their properties have been derived. We have identified sequential algorithms, that are based on problem structures and we have used the stochastic approximation idea to derive sequential EM algorithms in the general case. We have applied these sequential algorithms in few exampies. However, important topics for further research are the applications of these sequential algorithms to a variety of signal processing problems and a further theoretical analysis of these algorithms.

As a result of these contributions, a general and flexible class of iterative estimation algorithms has been established. Beyond the theoretical contributions and the specific applications, we believe that this thesis suggests a way of thinking and a philosophy, which may be used in a large variety of seemingly complex statistical inference and signal processing problems.

Appendix A

Convergence theorems of the EM algorithm

The convergence theorem of the EM algorithm are given in this Appendix. The theorems will be presented in parallel to the discussion of Chapter 2, i.e. we will start with the convergence properties of the likelihood sequence, then the convergence properties of the parameter estimates sequence and we will end by discussing the rate of convergence.

A.1 Convergence of the likelihood sequence

We start by quoting the Global Convergence Theorem from [17] and [18]. This theorem is frequently used to prove convergence of iterative algorithms in numerical analysis. Recall that a point to set map $M(x)$, where $x \in X$, is called closed at x if

$$x_k \rightarrow x, x_k \in X \text{ and } y_k \rightarrow y, y_k \in M(x_k) \implies y \in M(x)$$

For a point to point map, continuity implies closedness.

Theorem A.1 (**Global Convergence Theorem**) *Let the sequence $\{x_k\}$ be generated by $x_k \rightarrow x_{k+1} \in M(x_k)$, where M is a point to set map. Let a solution set Γ be given, and suppose that:*

(*i*) *all points x_k are contained in a compact set C*

(*ii*) *M is closed over the complement of Γ*

(*iii*) *there is a continuous function L such that*

(*a*) *if $x \notin \Gamma$, $L(y) > L(x)$ $\forall y \in M(x)$*

(*b*) *if $x \in \Gamma$, $L(y) \geq L(x)$ $\forall y \in M(x)$*

Then, all the limit points of $\{x_k\}$ are in the solution set Γ and $L(x_k)$ converges monotonically to $L(x)$ for some $x \in \Gamma$

The proof may be found in [17] page 91 and [18] page 187.

We are interested in applying the theorem above to the case where $L(\cdot)$ is the log-likelihood function defined over Θ, the solution set is either the set of local maxima $\mathcal{M}$ or the set of stationary points S, and $M(\cdot)$ is the point to set map implied by the EM (GEM) algorithm. In this case, condition (*i*) is met by the assumption that Θ_0 is compact, condition (*iii*)(*b*) is true by theorem 2.1, see eq. (2.22); thus, we have the following corollary of theorem A.1:

Corollary A.1 *Let $\{\underline{\theta}^{(n)}\}$ be a GEM sequence generated by*

$$\underline{\theta}^{(n)} \rightarrow \underline{\theta}^{(n+1)} \in M(\underline{\theta}^{(n)})$$

and suppose that

(*i*) *M is a closed point to set map over the complement of S ($\mathcal{M}$)*

(*ii*) *$L(\cdot)$ is continuous and $L(\underline{\theta}^{(n+1)}) > L(\underline{\theta}^{(n)})$ for all $\underline{\theta}^{(n)} \notin S$ ($\mathcal{M}$)*

Then all limit points of $\{\underline{\theta}^{(n)}\}$ are stationary points (local maxima) of L, and $L^{(n)}$ converges monotonically to $L^* = L(\underline{\theta}^*)$ for some $\underline{\theta}^* \in S$ ($\mathcal{M}$).

For the EM algorithm, where $M(\underline{\theta}^{(n)})$ is the set of maximizers of $Q(\underline{\theta}; \underline{\theta}^{(n)})$, the following continuity condition

$$Q(\underline{\theta}_1; \underline{\theta}_2) \quad \text{is continuous in both } \underline{\theta}_1 \text{ and } \underline{\theta}_2 \tag{A.1}$$

implies the closedness of M, i.e. it implies condition (i) in the corollary above. Now, if we are interested only in convergence to a stationary point, where the solution set is S, then the continuity condition also implies condition (ii) above; thus we have the following theorem:

Theorem A.2 *Suppose Q satisfy the continuity condition (A.1). Then all the limit points of any instance $\{\underline{\theta}^{(n)}\}$ of an EM algorithm are stationary points of L and $L(\underline{\theta}^{(n)})$ converges monotonically to $L^* = L(\underline{\theta}^*)$ for some stationary point $\underline{\theta}^*$.*

Proof: Suppose that for some $\underline{\theta}^{(n)} \notin S$ condition (ii) above is not met, i.e.

$$L(\underline{\theta}^{(n+1)}) = L(\underline{\theta}^{(n)}) \tag{A.2}$$

where $\underline{\theta}^{(n+1)} \in M(\underline{\theta}^{(n)})$, i.e. it is a global maximizer of $Q(\cdot;\underline{\theta}^{(n)})$. Since $\underline{\theta}^{(n)}$ is the global maximizer of $H(\cdot;\underline{\theta}^{(n)})$ (by Jensen's inequality, eq. (2.14)), the equality in (A.2) implies

$$Q(\underline{\theta}^{n+1};\underline{\theta}^{(n)}) = Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) \text{ and } H(\underline{\theta}^{n+1};\underline{\theta}^{(n)}) = H(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) \tag{A.3}$$

or in particular, that $\underline{\theta}^{(n)}$ is also a global maximizer of $Q(\cdot;\underline{\theta}^{(n)})$. Now

$$L(\underline{\theta}^{(n)}) = Q(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) - H(\underline{\theta}^{(n)};\underline{\theta}^{(n)}) = q(\underline{\theta}^{(n)}) - h(\underline{\theta}^{(n)}) \tag{A.4}$$

and $\underline{\theta}^{(n)}$ is a global maximizer of $q(\cdot)$ and $h(\cdot)$, thus, $\underline{\theta}^{(n)}$ must be a stationary point of $L(\cdot)$ □

The convergence to the set of local maxima, $\mathcal{M}$, is not guaranteed by conditions above, since we may find $\underline{\theta}^{(n)}$ outside $\mathcal{M}$, but inside S, for which indeed condition (ii) of corollary A.1 is not met. The following theorem imposes an additional condition, and thus provides sufficient conditions for convergence to the set of local maxima $\mathcal{M}$.

Theorem A.3 *Suppose that in addition to the continuity condition (A.1), Q satisfies*

$$\sup_{\underline{\theta}' \in \Theta} Q(\underline{\theta}'; \underline{\theta}) > Q(\underline{\theta}; \underline{\theta}) \quad \forall \underline{\theta} \in (\mathcal{S} - \mathcal{M}) \tag{A.5}$$

where $(\mathcal{S} - \mathcal{M})$ *is the difference set* $\{\underline{\theta} \in \mathcal{S} : \underline{\theta} \notin \mathcal{M}\}$

Then all the limit points of any instance $\{\underline{\theta}^{(n)}\}$ *of an EM algorithm are local maxima of* L *and* $L(\underline{\theta}^{(n)})$ *converges monotonically to* $L^* = L(\underline{\theta}^*)$ *for some local maximum* $\underline{\theta}^*$.

Proof: Condition (A.5) excludes the possibility that condition (*ii*) of corollary A.1 is not met by some $\underline{\theta}^{(n)} \in \mathcal{S} - \mathcal{M}$. Theorem A.2 proved that this condition is met for all $\underline{\theta}^{(n)} \notin \mathcal{S}$, and thus it is met for all $\underline{\theta}^{(n)} \notin \mathcal{M}$. Thus, using corollary A.1, this theorem follows immediately. □

A.2 Convergence of the parameter estimate sequence

The convergence of the likelihood sequence does not imply the convergence of the parameter estimate sequence. However, if the likelihood sequence converges to a solution set that contains a single point, the convergence of the parameter sequence is guaranteed (trivially), as stated in the following theorem:

Theorem A.4 *Let* $\{\underline{\theta}^{(n)}\}$ *be an instance of a GEM algorithm, with a corresponding likelihood sequence* $\{L^{(n)}\}$ *that converges to some* L^* *and satisfy conditions* $(i), (ii)$ *of corollary A.1. Let the solution set* ($\mathcal{S}(L^*)$ *or* $\mathcal{M}(L^*)$) *be the singleton* $\{\underline{\theta}^*\}$. *Then,* $\underline{\theta}^{(n)} \longrightarrow \underline{\theta}^*$.

An important special case of this theorem is when the likelihood function is unimodal in Θ. This case is stated in the following corollary of the theorem above:

Corollary A.2 *Suppose that* $L(\underline{\theta})$ *is unimodal in* Θ *with* $\underline{\theta}^*$ *being the only stationary point and that* $Q(\underline{\theta}'; \underline{\theta})$ *is continuous in both* $\underline{\theta}'$ *and* $\underline{\theta}$. *Then any EM sequence* $\{\underline{\theta}^{(n)}\}$ *converges to the unique maximizer* $\underline{\theta}^*$ *of* $L(\underline{\theta})$.

The requirement that the solution space is singleton may be relaxed, if the sequence of estimates is such that $\|\underline{\theta}^{(n-1)} - \underline{\theta}^{(n)}\| \to 0$ as $n \to \infty$. In this case, $\{\underline{\theta}^{(n)}\}$ will converge, if the solution set is discrete, as shown in the following theorem. We note that a discrete set is a set whose only connected components are singletons.

Theorem A.5 *Let $\{\underline{\theta}^{(n)}\}$ be an instance of a GEM algorithm, with a corresponding likelihood sequence $\{L^{(n)}\}$ that converges to some L^* and satisfies conditions (i), (ii) of corollary A.1. If $\|\underline{\theta}^{(n+1)} - \underline{\theta}^{(n)}\| \to 0$ as $n \to \infty$, then all the limit points of $\{\underline{\theta}^{(n)}\}$ are in a connected and compact subset of $\mathcal{S}(L^*)$ (or $\mathcal{M}(L^*)$). In particular, if $\mathcal{S}(L^*)$ (or $\mathcal{M}(L^*)$) is discrete, then $\underline{\theta}^{(n)}$ converges to some $\underline{\theta}^*$ in $\mathcal{S}(L^*)$ (or $\mathcal{M}(L^*)$).*

Proof: The sequence $\{\underline{\theta}^{(n)}\}$ is bounded (by our assumptions). The set of limit points of a bounded sequence with $\|\underline{\theta}^{(n+1)} - \underline{\theta}^{(n)}\| \to 0$ as $n \to \infty$ is connected and compact (see, e.g. theorem 28.1 of [91]). Since all the limit points of $\{\underline{\theta}^{(n)}\}$ are in $\mathcal{S}(L^*)$ (or $\mathcal{M}(L^*)$), the theorem follows. □

A.3 Rate of convergence

We start by presenting identities for the derivatives of the log-likelihood function, i.e. $DL(\underline{\theta})$ and $D^2L(\underline{\theta})$, which are needed for calculating the expression for the rate of convergence of the EM algorithm. To prove those identities, the following well known results concerning the *score function* are needed.

Let Ω be a sample space and $f(\omega;\phi)$ be a p.d.f. defined over this space, parameterized by ϕ. Let the score function be defined as,

$$S(\omega;\phi) = \frac{\partial \log f(\omega;\phi)}{\partial \phi}$$

Then,

$$E_\phi\{S(\omega;\phi)\} = \int_\Omega \frac{\partial \log f(\omega;\phi)}{\partial \phi} f(\omega;\phi)d\omega = 0 \tag{A.6}$$

and

$$Var_{\phi}\{S(\omega;\phi)\} = \int_{\Omega}\left[\frac{\partial \log f(\omega;\phi)}{\partial \phi}\right]^2 f(\omega;\phi)d\omega = -E_{\phi}\left\{\frac{\partial^2 \log f(\omega;\phi)}{\partial \phi^2}\right\} \tag{A.7}$$

Suppose now that the sample space is $\mathcal{X}(\underline{y})$ and that $f_{X/Y}(\underline{x}/\underline{y};\underline{\theta})$ is defined over it. Equations (A.6) and (A.7) become

$$E\left\{\frac{\partial}{\partial \underline{\theta}}\log f_{X/Y}(\underline{x}\,|\,\underline{y};\underline{\theta}) \;/\; \underline{y},\underline{\theta}\right\} = D^{10}H(\underline{\theta};\underline{\theta}) = 0 \tag{A.8}$$

$$Var\left\{\frac{\partial}{\partial \underline{\theta}}\log f_{X/Y}(\underline{x}/\underline{y};\underline{\theta})\,|\underline{y},\underline{\theta}\right\} = D^{11}H(\underline{\theta};\underline{\theta}) = -D^{20}H(\underline{\theta};\underline{\theta}) \tag{A.9}$$

Differentiating both sides of (2.12) and using (A.8) and (A.9) above, we get the following identities

$$DL(\underline{\theta}) = D^{10}Q(\underline{\theta};\underline{\theta}) = S(\underline{y};\underline{\theta}) \tag{A.10}$$

$$D^2L(\underline{\theta}) = D^{20}Q(\underline{\theta};\underline{\theta}) - D^{20}H(\underline{\theta};\underline{\theta}) = D^{20}Q(\underline{\theta};\underline{\theta}) + D^{11}H(\underline{\theta};\underline{\theta}) \tag{A.11}$$

$$D^{11}Q(\underline{\theta};\underline{\theta}) = D^{11}H(\underline{\theta};\underline{\theta}) \tag{A.12}$$

The rate of convergence of a class of GEM algorithms is now given in the following theorem.

Theorem A.6 *Let $\{\underline{\theta}^{(n)}\}$ be a sequence of a GEM algorithm such that*

(i) $\underline{\theta}^{(n)} \to \underline{\theta}^*$

(ii) $D^{10}Q(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) = 0$

(iii) $D^{20}Q(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)})$ *is negative definite with eigenvalues bounded away from zero*

i.e. $\underline{\theta}^{(n+1)}$ *is a local maximizer of* $Q(\underline{\theta};\underline{\theta}^{(n)})$*. Then,* $DL(\underline{\theta}^*) = 0$, $D^{20}Q(\underline{\theta}^*;\underline{\theta}^*)$ *is negative definite, and*

$$DM(\underline{\theta}^*) = D^{20}H(\underline{\theta}^*;\underline{\theta}^*)\left[D^{20}Q(\underline{\theta}^*;\underline{\theta}^*)\right]^{-1} \tag{A.13}$$

Proof: Differentiating (2.12) we get

$$DL(\underline{\theta}^{(n+1)}) = D^{10}Q(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) - D^{10}H(\underline{\theta}^{(n+1)};\underline{\theta}^{(n)}) \tag{A.14}$$

where the first term of (A.14) is zero by the assumptions, the second term is zero in the limit as $n \to \infty$ by (A.8) and $DL(\underline{\theta}^*) = 0$. Similarly, $D^{20}Q(\underline{\theta}^*; \underline{\theta}^*)$ is negative definite, since it is the limit of $D^{20}Q(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)})$ whose eigenvalues are bounded away from zero.

The last part of the theorem, i.e. showing (A.13), was proved in Chapter 2 using the identities above, see (2.42)-(2.44). □

Appendix B

Consistency of sequential EM algorithms

We will start by presenting the following theorem, which is also used for showing consistency of the ML estimator:

Theorem B.1 *Let $\underline{y}_1, \cdots, \underline{y}_n, \cdots$ be the output of a stationary Markov source with a finite memory p, i.e.*

$$f_{Y_n/Y_{n-1},\cdots,Y_1}(\underline{y}_n/\underline{y}_{n-1},\cdots,\underline{y}_1) = f_{Y_n/Y_{n-1},\cdots,Y_{n-p}}(\underline{y}_n/\underline{y}_{n-1},\cdots\underline{y}_{n-p})$$

Let $L_n(\underline{\theta})$ be the log-likelihood function given $\underline{y}_1, \cdots, \underline{y}_n$, as in (3.3). The sequence of functions $l_n(\underline{\theta}) = \frac{1}{n}L_n(\underline{\theta})$ converges uniformly in probability 1 to a limit $l(\underline{\theta})$ where under regularity conditions, the global maximum of $l(\underline{\theta})$ and the unique solution to the equation $Dl(\underline{\theta}) = 0$ is the true parameter value $\underline{\theta}_{true}$.

Proof: The likelihood $L_n(\underline{\theta})$ may be written as,

$$\begin{aligned} L_n(\underline{\theta}) &= \log f_{Y_n,\cdots,Y_1}(\underline{y}_n,\cdots,\underline{y}_1;\underline{\theta}) \qquad (B.1) \\ &= \log f_{Y_n/Y_{n-1},\cdots,Y_1}(\underline{y}_n/\underline{y}_{n-1},\cdots,\underline{y}_1;\underline{\theta}) + \log f_{Y_{n-1}/Y_{n-2},\cdots,Y_1}(\underline{y}_{n-1}/\underline{y}_{n-2},\cdots,\underline{y}_1;\underline{\theta}) + \cdots \end{aligned}$$

For $n >> p$,

$$l_n(\underline{\theta}) = \frac{1}{n}L_n(\underline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\log f_{Y_i/Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_i/\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta}) \tag{B.2}$$

Using the strong law of large numbers, we get

$$\lim_{n\to\infty} l_n(\underline{\theta}) = E\left\{\log f_{Y_i/Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_i/\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta})\right\} = l(\underline{\theta}) \tag{B.3}$$

in probability 1.

The function $l(\underline{\theta})$ may be written explicitly as,

$$l(\underline{\theta}) = \int \log f_{Y_i/Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_i/\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta}) f_{Y_i,\cdots,Y_{i-p}}(\underline{y}_i,\cdots,\underline{y}_{i-p};\underline{\theta}_{true})d\underline{y}_i\cdots d\underline{y}_{i-p} \tag{B.4}$$

or,

$$\begin{aligned} l(\underline{\theta}) \quad = \quad & \int d\underline{y}_{i-1}\cdots d\underline{y}_{i-p}\Big[f_{Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta}_{true})\cdot \\ & \cdot\int \log f_{Y_i/Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_i/\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta}) f_{Y_i/Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_i/\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta}_{true})d\underline{y}_i\Big] \end{aligned} \tag{B.5}$$

Invoking Jensen's inequality on the inner integral, we conclude that

$$l(\underline{\theta}) \leq l(\underline{\theta}_{true}) \tag{B.6}$$

where equality is achieved if and only if,

$$f_{Y_i/Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_i/\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta}) = f_{Y_i/Y_{i-1},\cdots,Y_{i-p}}(\underline{y}_i/\underline{y}_{i-1},\cdots,\underline{y}_{i-p};\underline{\theta}_{true}) \quad \text{a.e} \tag{B.7}$$

Under the identifiability condition, the equality in (B.7) is achieved only if $\underline{\theta} = \underline{\theta}_{true}$, i.e. $\underline{\theta}_{true}$ is the unique global maximum of $l(\underline{\theta})$. Using the differentiability condition, and the convexity of $l(\underline{\theta})$ we also conclude that $\underline{\theta}_{true}$ is the unique solution to the equation $Dl(\underline{\theta}) = 0$. □

This theorem may be extended to more general ergodic sources, whose memory is fading fast enough. The more general conditions may be found in [48], appendix A.

Using this theorem, we may now state the main consistency result:

Theorem B.2 *Let the observations $\underline{y}_1, \cdots, \underline{y}_n \cdots$ be generated by an ergodic source for which theorem B.1 holds. Let $\{\underline{\theta}^{(n)}\}$ be an instance of a sequential EM algorithm such that for any realization of the observations,*

(i) the sequence of estimates $\{\underline{\theta}^{(n)}\}$ converges to a limit $\underline{\theta}^*$

(ii) $\lim_{n\to\infty} D^{10}Q_{n+1}(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) = 0$

Then, in probability 1, as $n \to \infty$, $\underline{\theta}^{(n)} \to \underline{\theta}_{true}$.

Proof: From the assumption (*i*), and using the identity (A.8), we may write,

$$\lim_{n\to\infty} D^{10}H_{n+1}(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) = \lim_{n\to\infty} D^{10}H_{n+1}(\underline{\theta}^*; \underline{\theta}^*) = 0 \tag{B.8}$$

From theorem (B.1) the sequence $l_n(\underline{\theta}) = \frac{1}{n}L_n(\underline{\theta})$ converges uniformly in probability 1 to some $l(\underline{\theta})$. The sequence of derivatives $Dl_{n+1}(\underline{\theta}^{(n+1)})$ may be written as,

$$Dl_{n+1}(\underline{\theta}^{(n+1)}) = \frac{1}{n}\left[D^{10}Q_{n+1}(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) - D^{10}H_{n+1}(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)})\right] \tag{B.9}$$

Thus, from the assumption (*ii*), and from (B.8) above.

$$\lim_{n\to\infty} Dl_{n+1}(\underline{\theta}^{(n+1)}) = 0 \tag{B.10}$$

Since $l_{n+1}(\underline{\theta})$ converges uniformly to $l(\underline{\theta})$, and using (B.10) we conclude that

$$\lim_{n\to\infty} Dl(\underline{\theta}^{(n+1)}) = 0 \tag{B.11}$$

From theorem B.1 and from (B.11), our desired result follows, i.e.

$$\lim_{n\to\infty} \underline{\theta}^{(n)} = \underline{\theta}_{true}$$

in probability 1. □

Bibliography

[1] H. L. Van Trees. *Detection Estimation and Modulation theory*. Wiley, New York, 1968.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal cf the Royal Statistical Society*, Ser. 39:1–38, 1977.

[3] J. S. Lim and A. V. Oppenheim. All pole modeling of degraded speech. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-26:197–210, 1978.

[4] J. R. Grindon. *Estimation theory for a class of nonlinear random vector channels*. PhD thesis, Washington University, St. Louis, MO., 1970.

[5] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[6] M. Segal. *Resolution and estimation techniques with applications to vectorial miss-distance indicator*. Master's thesis, Technion, Israel Institute of Technology, 1981.

[7] B. R. Musicus. *An iterative technique for maximum likelihood estimation with noisy data*. Master's thesis, Massachusetts Institute of Technology, Feb. 1979.

[8] B. R. Musicus. *Iterative algorithms for optimal signal reconstruction and parameter identification given noisy and incomplete data*. PhD thesis, Massachusetts Institute of Technology, September 1982.

[9] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11:95–103, 1983.

[10] B. Widrow et al. Adaptive noise canceling: principles and applications. *Proc. IEEE*, 63:1692–1716, 1975.

[11] C. R. Rao. *Linear statistical inference and its applications*. Wiley, New York, 1965.

[12] R. A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Stat. Soc.*, B 44, 1983.

[13] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Stat. Soc.*, B 44:226–233, 1982.

[14] I. Meilijson. A fast improvement to the EM algorithm in its own terms. *Journal of the American Stat. Assoc.*, submitted.

[15] N. Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statis. Assoc.*, 73:805–811, 1978.

[16] D. B. Rubin and D. T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47:69–76, 1982.

[17] W. I. Zangwill. *Nonlinear programming: A unified approach*. Prentice Hall, Englewood Cliffs, NJ, 1969.

[18] D. G. Luenberger. *Linear and Nonlinear programming, second edition*. Addison-Wesley, Reading, MA, 1984.

[19] R. A. Fisher. Theory of statistical estimation. *Proc. Cambridge Phil. Soc.*, 22:700–725, 1925.

[20] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice Hall, Englewood Cliffs, NJ, 1979.

[21] R. J. Solomonoff. A formal theory of inductive inference, part 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[22] G. W. Hart. *Minimum Information estimation of structures*. PhD thesis, Massachusetts Institute of Technology, 1987.

[23] R. V. L. Hartley. Transmission of information. *Bell Sys. Tech. Jour.*, 7:535–563, 1928.

[24] C. Shannon. *Mathematical theory of communication*. U. Illinois Press, Urbana Ill., 1948.

[25] N. Wiener. *Cybernetics: Control and communication in the animal and the machine*. MIT Press, Cambridge, MA., 1948.

[26] A. N. Kolmogorov. Three approaches to the quantitive definition of information. *Problems in infor. trans.*, 1:4–7, 1965.

[27] G. J. Chaitin. On the length of programs for computing finite binary sequences. *Jour. of ACM*, 13:547, 1966.

[28] M. D. Davis and E. J. Weyuker. *Computability, Complexity and Languages*. Academic Press, New York, N.Y., 1983.

[29] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[30] J. Rissanen. A universal prior for integers and estimation by Minimum Description Length. *The Annals of Statistics*, 11:416–431, 1983.

[31] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Trans. Information Theory*, IT-30:629–636, 1984.

[32] J. P. Burg. *Maximum Entropy Spectral Estimation*. PhD thesis, Stanford University, 1975.

[33] M. Feder and E. Weinstein. Parametric spectral estimation via the EM algorithm. In *Proceedings of the DSP workshop, Chatam MA.*, October 1984.

[34] A. Dembo. 1987.

[35] M. I. Miller and D. L. Snyder. The role of likelihood and entropy in incomplete data problems: applications to estimating point-process intensities and toeplitz constrained covariances. *Proceedings of the IEEE*, submitted.

[36] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171, 1970.

[37] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP magazine*, 3:4–16, 1986.

[38] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *Bell Sys. Tech. Jour.*, 62:1035–1074, 1983.

[39] R. E. Bellman and S. E. Dreyfus. *Applied Dynamic Programming*. Princeton Univ. Press, Princeton N.J., 1962.

[40] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23:462–466, 1952.

[41] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

[42] J. R. Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25:382–386, 1954.

[43] A. Dvoretzky. On stochastic approximation. In *Proceedings of 3rd Berkley Symposium on math. stat. and prob.*, pages 35–56, 1956.

[44] D. M. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society*, 46 (B):257–267, 1984.

[45] V. Dupac. On the Kiefer Wolfowitz approximation method. *Casopis Pest. Mat.*, 82:47–75, 1957.

[46] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39:1327–1332, 1968.

[47] H. J. Kushner. *Stochastic approximation methods for constrained and unconstrained systems*. Springer-Verlag, New York, 1978.

[48] Ellis. *Entropy, large variations and statistical mechanics*. Springer-Verlag, New York, 1983.

[49] J. Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29:373–405, 1958.

[50] Jr. K. Metzger. *Signal Processing equipment and techniques for use in measuring ocean acoustic Tmultipath structures.* PhD thesis, University of Michigan, Ann Arbor, MI, 1983.

[51] J. L. Spiesberger, R. C. Spindall, and K. Metzger. Stability and identification of ocean acoustic multipath. *Journal of the Acoustical Society of America*, 67(6):2011–2017, 1980.

[52] M. Feder and E. Weinstein. *Multipath time delay estimation using the EM algorithm.* Technical Report WHOI-87-3, Woods Hole Oceanographic Institution, 1987.

[53] J. E. Ehrenberg, T. E. Ewart, and R. D. Morris. Signal processing techniques for resolving individual pulses in multipath signal. *Journal of the Acoustical Society of America*, 63:1861–1865, 1978.

[54] M. Feder and E. Weinstein. *Optimal Multiple Source Location Estimation via the EM algorithm.* Technical Report WHOI-85-25, Woods Hole Oceanographic Institution, July, 1985.

[55] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 32:387–392, 1985.

[56] M. Feder and E. Weinstein. *Multipath and multi-source array processing via the EM algorithm.* Technical Report WHOI-86-25, Woods Hole Oceanographic Institution, 1986.

[57] M. Feder and E. Weinstein. Multipath and multiple source array processing via the EM algorithm. In *Proceedings of the 1986 International Conference on Acoustics Speech and Signal Processing*, 1986.

[58] S. S. Reddy. Multiple source location - a digital approach. *IEEE Trans. Aerospace and Electronics systems*, AES-15:95–105, 1979.

[59] R. O. Schmidt. *A Signal Subspace approach to multiple emitter location and spectral estimation.* PhD thesis, Stanford University, CA., 1981.

[60] G. Su and M. Morf. The signal subspace approach for multiple wide band emitter location. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-31:1503–1522, 1983.

[61] B. Porat and B. Friedlander. Estimation of spacial and spectral parameters of multiple sources. *IEEE Trans. Information Theory*, IT-29:412–425, 1983.

[62] A. Nehorai, G. Su, and M. Morf. Estimation of time differences of arrival for multiple ARMA sources by pole decomposition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-31:1478–1491, 1983.

[63] J. S. Lim (Editor). *Speech Enhancement.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[64] S. F. Boll and D. C. Pulsipher. Suppression af acoustic noise in speech using two microphone adaptive noise cancellation. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-28:752–753, 1980.

[65] W. A. Harrison, J. S. Lim, and E. Singer. A new application of adaptive noise cancellation. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-34:21–27, 1986.

[66] D. Falconer and L. Ljung. Application of fast kalman estimation to adaptive equalization. *IEEE Trans. Communications*, COMM-26:1439–1446, 1978.

[67] R. A. Monzingo and T. W. Miller. *Introduction to adaptive arrays*. Wiley, New York, NY, 1980.

[68] O. L. Frost III. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60:926–935, 1972.

[69] L. J. Griffiths and C. W. Jim. Linearly constrained adaptive beamforming. *IEEE Trans. Antennas and Propagation*, AP-30:27–34, 1982.

[70] P. M. Peterson. Simulating the response of multiple microphone to a single acoustic source in a reverberant room. *Journal of the Acoustical Society of America*, 76, Nov. 1986.

[71] S. Kullback. *Information theory and statistics*. Dover, New York, 1959.

[72] Pinsker. *Information and information stability of random variables*. Holden Day, San Francis. CA., 1964.

[73] I. Csiszar and G. Tsunady. Information geometry and alternating minimization procedures. *Statistics and decisions*, Supplement issue No. 1:205–237, 1984.

[74] E. T. Jaynes. Prior probabilities. *IEEE Trans. on Systems Sci. and Cybernet.*, SSC-4:227–241, 1968.

[75] E. T. Jaynes. On the rationale of maximum entropy methods. *Proceedings of the IEEE*, 70:939–952, 1982.

[76] J. Rissanen and G. G. Langdon. Universal modeling and coding. *IEEE Trans. Information Theory*, IT-27:12–23, 1984.

[77] L. C. W. Dixon and G. P. Szego (editors). *Towards glogal optimization*. North Holland Publishing, 1978.

[78] S. Kirkpatrick, C. D. Gellat, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:932–943, 1983.

[79] M. H. Hayes. *Signal Reconstruction from Phase or Magnitude*. PhD thesis, Massachusetts Institute of Technology, 1981.

[80] S R. Curtis, J. S. Lim, and A. V. Oppenheim. Signal reconstruction from one bit of Fourier transform phase. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-33:643–657, 1985.

[81] P. L. Van Hove, J. S. Lim, and A. V. Oppenheim. Signal reconstruction from signed Fourier transform magnitude. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-31:1286–1293, 1983.

[82] J. R. Fienup. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*, 3:27–29, 1978.

[83] David Izraelevitz. *Reconstruction of two dimensional signals from the Fourier transform magnitude*. PhD thesis, Massachusetts Institute of Technology, 1986.

[84] A. Levi and H. Stark. Image restoration by the method of generalized projections with application to restoration from magnitude. *Journal of the Optical Society of America - A*, 1:932–943, 1984.

[85] R. W. Schafer, R. M. Mersereau, and M. A. Richards. Constrained iterative restoration algorithms. *Proceedings of the IEEE*, 69:432–450, 1981.

[86] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction in positron emission tomography. *IEEE Trans. Med. Imag.*, MI-1:113–122, 1982.

[87] N. Metropolis et. al. Equation of state calculation by fast computing machines. *Jour. of Chemical Physics*, 1087–1092, 1953.

[88] D. Anastassiou and D. J. Sakrison. A probability model for simple close random curves. *IEEE Trans. Information Theory*, IT-27:376–381, 1981.

[89] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, 1984.

[90] J. Marroquin. *Probabilistic solution of inverse problems*. PhD thesis, Massachusetts Institute of Technology, 1985.

[91] A. M. Ostrowski. *Solution of equations and systems of equations, Second edition*. Academic Press, New York, N.Y., 1966.