Parametric estimation of sparse channels: theory and applications

THÈSE N^{O} 5976 (2014)

PRÉSENTÉE LE 31 JANVIER 2014

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

LABORATOIRE DE COMMUNICATIONS AUDIOVISUELLES

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Yann BARBOTIN

acceptée sur proposition du jury:

Prof. A. Madry, président du jury
Prof. M. Vetterli, directeur de thèse
Prof. T. Blu, rapporteur
Dr O. Lévêque, rapporteur
Prof. B. Ottersten, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

Totus Tuus.

Abstract

The use of parametric methods for the estimation of communication channels requires a diverse suite of components : *models*, *algorithms* (for estimation and detection) and *analytical tools* (error bounds, time-frequency uncertainty limits,...). The coherent study of each one of them is the goal of this thesis.

In the first part, we propose a parametric model, the *sparse common support* model, and study under which conditions it can be used to describe communication channels. We then extend classical subspace estimation methods to work on this model, and propose a fast estimation algorithm having a superlinear complexity and a linear memory footprint with respect to the number of measurements. For comparison, a direct implementation has a cubic complexity and a quadratic memory footprint.

Parametric estimation can only be relevant if it is wisely used, *i.e.* if the right model is first selected. This task is called *detection*, and we outline two procedures to complete it. The first one uses the statistical properties of the noise and is therefore powerful but sensitive to model mismatches. A second detection method is based on the low dimensionality of the model, and uses a convexification of the rank of a matrix, called the *effective rank*, to determine the intrinsic dimension of the model.

The first part is concluded with comparative tests on measured channel impulse responses (with added noise), which shows the proposed model and algorithms improve channel estimation at low signal to noise ratio (SNR).

In the second part, we study the localization on periodic domains, which is necessary to analyze the performance of periodic parameters estimation.

First, we will construct periodic waveforms which have a minimal time-frequency product. These periodic waveforms — obtained from *Mathieu functions* — play a role similar to Gaussian functions with respect to the Heisenberg uncertainty principle. A fundamental difference is that in the periodic case, the Heisenberg limit is only achieved by (infinitely) narrow periodic waveforms, and thus, the maximally compact waveforms we construct provide the achievable lowerbound for any given width.

Then, we show that lowerbounds on the variance of estimators such as the Cramér-Rao bound (CRB), and its siblings (the CRB family), can be expressed in a form equivalent to the Heisenberg uncertainty principle. From this similarity, we build periodic lowerbounds (the periodic CRB, the periodic Hammersley-Chapman-Robbins bound,...) with minimal efforts.

Finally, we obtain an explicit analytical formula for the Barankin bound on a single periodic parameter and observe that this bound is not necessarily tight despite the fact that it is the strongest that can be obtained in the CRB family.

Keywords: *sparse channels, parametric estimation, Heisenberg principle, circular lowerbounds, Barankin bound.*

Résumé

L'utilisation de méthodes paramétriques pour l'estimation de canaux de communication demande l'utilisation de plusieurs composants : des modèles, des algorithmes (pour l'estimation et la détection) et des outils d'analyse (bornes d'erreur, limites sur l'étalement temps-fréquence, ...). L'étude concertée de ces composants est le but de cette thèse.

Dans une première partie, nous proposerons un modèle paramétrique, le modèle à *support clairsemé et commun*, et étudierons sous quelles conditions il peut être appliqué à l'estimation des canaux de communication. Nous adapterons ensuite des méthodes de sous-espaces bien connues à ce modèle, et proposerons un algorithme d'estimation avec complexité superlinéaire et une occupation mémoire linéaire en terme du nombre de mesures effectuées. Comme comparaison, une implémentation directe de cette méthode aurait une complexité cubique et une occupation mémoire quadratique.

Une estimation paramétrique doit-être utilisée à propos pour être efficace, c'est-à-dire que le bon modèle doit d'abord être sélectionné. Cette sélection est appelée *détection*, et nous décrirons deux procédures la réalisant. La première, utilise les propriétés statistiques du bruit sur les mesures, et est donc performante mais sensible aux erreurs de modélisation. La seconde méthode exploite la faible dimension du modèle, et utilise pour cela une convexification du rang matriciel — appelé le *rang effectif* — afin de déterminer la dimension intrinsèque du modèle. Pour conclure cette première partie, des tests comparatifs seront réalisés sur des canaux de communication mesurés sur le terrain et auxquels du bruit est ajouté. Ces tests indiquent que le modèle et les algorithmes proposés améliorent l'estimation des canaux lorsque le rapport de puissance entre le signal et le bruit est faible.

Dans une seconde partie, nous étudierons la localisation sur des domaines périodiques, ce qui est nécessaire pour l'analyse des performances lors de l'estimation de paramètres périodiques.

Premièrement, nous construirons des fonctions périodiques possédant un étalement temps-fréquence minimal. Ces fonctions périodiques — obtenues à partir des *fonctions de Mathieu* — ont des propriétés similaires aux fonctions Gaussiennes vis-à-vis du principe d'Heisenberg dans le domaine périodique. Une distinction fondamentale est que dans le cas périodique, la limite du principe d'Heisenberg ne peut-être atteinte que par des fonctions infiniment concentrées, et donc les fonctions à compacité maximum construites indiquent la limite atteignable, quelque soit la concentration.

Ensuite, nous montrerons que des bornes inférieurs sur la variance d'estimations, comme la borne de Cramér-Rao (BCR) et affiliées (famille BCR), peuvent-être exprimées sous une forme semblable au principe d'Heisenberg. Par cette similitude, nous construirons des bornes périodiques (BCR périodique, borne de Hammersley-Chapman-Robbins périodique) avec un effort minimum.

Finalement, nous obtiendrons une formule analytique explicite de la borne de Barankin pour un seul paramètre périodique et observerons qu'elle ne peut pas toujours être atteinte malgré le fait qu'elle soit la borne la plus performante de la famille BCR.

Mots-clés: *canaux clairsemés, estimation paramétrique, principe d'Heisenberg, bornes circulaires, borne de Barankin.*

Acknowledgments

It is a difficult exercise to express *recognition* through static words since it is — as the prefix "*re-*" suggests it — an act of memory from the heart which should be perpetuated. This page can only be a part of it.

I would like to first thank my supervisor Martin for the support and intellectual freedom he gave me. His broad views on research provided great guidance in the arcane corners of a thesis subject. I will remember above all the priority and preference given to the people before the ideas, which says much knowing his passion for research.

My gratitude also goes towards my dear colleagues at LCAV. Jacqueline, for coping with the organizational mess that are scientists and still being friendly, Reza, Juri, Ivan and all the other persons I had the pleasure to collaborate with[1]. Outside the lab, I would like to thank Thierry for his motivating and insightful comments.

I would also like to thank all the persons who gave me their friendship; friends from EPFL, friends from Lausanne, friends from elsewhere and community brothers and sisters. Listing names would be a sure way to forget someone, and my hope is to let all these friendships grow – now through a radical career shift.

Lastly, I want to thank my parents Philippe and Geneviève, my brother and my sister Pierre-Yves and Solène, for their unconditional love and support. A thesis is not always a leisure trip, and this presence makes the journey possible.

Chabestan – Thursday, August 22nd 2013.

[1] Coffee-breaks also qualify as scientific collaboration.

Contents

Abstract v

Résumé vii

Acknowledgments ix

Introduction 1

I Estimation & Detection of Sparse Channels 9

1 Parametric models for communications 11
1.1 Parametric or not parametric? — that is the question. 11
1.1.1 What are parametric models? 11
1.1.2 Advantages of parametric models 12
1.1.3 What are the limitations? 12
1.2 A parametric model for communications 13
1.2.1 The multipath model 13
1.2.2 The conditions for sparsity 15
1.2.3 The common support assumption 17
1.2.4 An example of sparse common support channels 20
1.3 Conclusion 22

2 Parametric estimation algorithms 25
2.1 Measurements model for OFDM communications 25
2.1.1 OFDM in a nutshell 26
2.1.2 Structure of a frame 26
2.1.3 Pilot layouts and the delay-spread 27
2.1.4 Sparse common support (SCS) channels for OFDM communications 27
2.2 Basic algorithms 28
2.2.1 The annihilating filter 31
2.2.2 Rotation invariance 34
2.2.3 Putting it all together 36
2.3 Fast, in place estimation 38
2.3.1 Krylov subspace projection and Lánczos' algorithm 38
2.3.2 Fast Lánczos iterations for SCS estimation 41

2.3.3 How large must the Krylov subspace be? . . . 42
2.3.4 Numerical tests . . . 44
2.4 Conclusion . . . 44

3 Model detection for sparse channels 47
3.1 A review of signal detection for subspace methods . . . 48
3.1.1 Applicability of the reviewed detection scheme to the SCS channel model . . . 51
3.2 Hypothesis testing for structured data matrices . . . 52
3.2.1 The spectral norm of E . . . 54
3.2.2 Multiple snapshots ($P > 1$) . . . 54
3.2.3 Numerical results . . . 55
3.3 Avoiding overfitting . . . 57
3.3.1 The validation of paths . . . 57
3.3.2 Extremal statistics of the noise projection . . . 60
3.3.3 An algorithm to prevent overfitting . . . 63
3.3.4 On the selection of a false positive error rate . . . 63
3.4 The Partial Effective Rank (PER) criterion . . . 64
3.4.1 The effective rank . . . 65
3.4.2 The partial effective rank (PER) . . . 66
3.4.3 The PER in action . . . 68
3.5 Test-case : the Weikendorf measurements . . . 69
3.6 Conclusion . . . 74

4 Tracking sparse channels 77
4.1 Annihilation as a linear constraint . . . 79
4.2 Two parametrizations . . . 81
4.2.1 Time-domain interpretation for $r_1 = \cdots = r_K = 1$ . . . 82
4.2.2 Dimensionality : a curse or a blessing? . . . 82
4.2.3 Solving the minimization problem . . . 85
4.3 Detection for tracking : Update, Validate and Add . . . 86
4.4 Numerical results . . . 88
4.5 Conclusion . . . 88

II Fundamental Limits on Periodic Localization 93

5 Time-Frequency localization in periodic domains 95
5.1 "Where?" — from linear to periodic . . . 95
5.1.1 Uncertainty principles for periodic waveforms and sequences . . . 98
5.1.2 Chapter outline . . . 100
5.2 Localization and its effect on time-frequency uncertainty . . . 101
5.2.1 Uncertainty Principle for Linear Operators . . . 101
5.2.2 The journey from continuous to discrete . . . 102
5.3 Maximally compact waveforms and sequences . . . 106
5.3.1 Properties of Maximally Compact Sequences . . . 106
5.3.2 The computation of maximally compact waveforms & sequences . . . 107
5.4 Simulation Results . . . 110

5.5 Conclusion . . . 111

6 From uncertainty to estimation error 113
6.1 Some history . . . 114
6.1.1 The Cramér-Rao bounds . . . 114
6.1.2 Improvement : Barankin's bound . . . 116
6.1.3 Other parameter spaces: manifolds, periodicity, ... 117
6.1.4 Problem summary . . . 118
6.2 An uncertainty-like inequality for estimators . . . 118
6.2.1 Preliminary example : the aperiodic and periodic CRB . . . 121
6.3 Replacing momentum to maximize the lowerbound . . . 127
6.3.1 Finding the optimal filter ("momentum") with collinearity . . . 129
6.3.2 An analytical solution for shift-invariant signals . . . 131
6.3.3 Discussion about the "gap" between the Barankin bound and the ML estimator . . . 134
6.4 Conclusion . . . 136

Conclusion 137

A Spatial correlation formula for fading channels 139
A.1 Azimuthal scatterers density distribution . . . 139
A.2 Derivation of the correlation matrix formula . . . 140

B Estimation algorithms 143
B.1 Further numerical tests . . . 143
B.1.1 Results on Exp. A . . . 144
B.1.2 Results on Exp. B . . . 146
B.1.3 Results on Exp. C . . . 146
B.2 Proof of Theorem 2.2 . . . 147

C Detection for sparse common support channels 149
C.1 Proof of Proposition 3.2 . . . 149
C.2 Proof of Proposition 3.3 . . . 151

D Tracking of SCS channels 153
D.1 Tracking of a single path . . . 153
D.2 Addition of a path . . . 155

E Time-Frequency localization in periodic domains 157
E.1 Proof of Lemmas 5.1 and 5.2 . . . 157
E.1.1 Proof of Lemma 5.1 . . . 157
E.1.2 Proof of Lemma 5.2 . . . 158
E.2 Proof of Theorem 5.2 . . . 159
E.3 Proof of Lemma 5.3 . . . 160
E.4 Proof of Theorem 5.3 . . . 161

F From uncertainty to error **163**
F.1 Proof of Lemma 6.1 . . . 163
F.2 Proof of Lemma 6.2 . . . 163
F.3 Proof of Theorem 6.2 . . . 164
F.4 Proof of Lemma . . . 165
F.5 Lowerbounds with modulo-2π localization . . . 165
F.5.1 Problem setup . . . 165
F.5.2 Computation of the MSE lower bound . . . 167
F.5.3 Application . . . 172
F.5.4 Conclusions . . . 173

Bibliography **177**

Curiculum Vitæ **191**

Introduction

Motivation

Sometimes, the landscape of signal processing seems to split in half. On one side, a data oriented labour, where the collection and description of signals is the bulk of the work, and datasets are the deliverables. On the other side, an area where abstract models are the foundations on which theories and algorithms are built.

While it is perfectly natural to fit into one of the category, as one cannot specialize in "everything", keeping an eye on the other side is essential.

The philosopher of science Edmund Husserl[2], declared in his inaugural lecture of May 1917 [69; 87] :

> "Natural objects, for example, must be experienced before any theorising about them can occur.",

which may appear as an endorsement of empiricism over speculative sciences, and in a more extreme view, the reduction of scientific knowledge to the experience and experiments. However in the same lecture, a few paragraphs later, this erroneous interpretation is unambiguously refuted

> "Experience by itself is not science. [...] We would be in a nasty position indeed if empirical science were the only kind of science possible."

This short philosophical digression was not in vain. If we follow the good Edmund, writing "Theory and Applications" in the title of this thesis was a terrible misnomer, since we should have used, "Applications and Theory" instead[3]. Nevertheless, the most important is saved, as both terms are distinct and are essential one to the other.

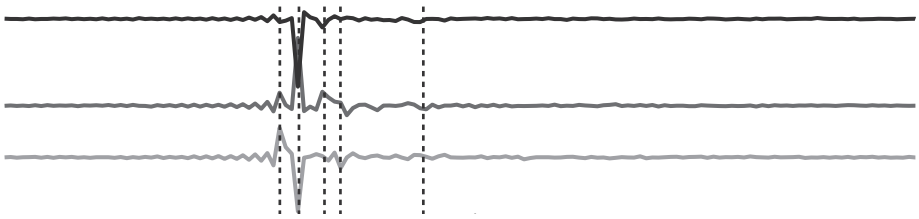
Following this intuition, the observation of a phenomenon should precede its modelization. So, before telling the what, why and how of parametric estimation, we will simply observe.

An observation from mobile communications

An electromagnetic impulse is transmitted over the air, a listening device records the magnetic field with three different antennas:

[2]Edmund Husserl (1859–1938), was a german philosopher and mathematician, founder of *phenomenology*.

[3]This may be the quickest self-rebuttal in academical history.



These signals — called *channels* — can be well described by giving the position and amplitudes of a few spikes. Furthermore, the positions of these spikes seems to be the same from one channel to another. Therefore, we could describe the above image by giving first the positions shared by all the channels (indicated by dashed lines), and than the amplitudes of these spikes for each channel. What we obtain is a *parametric* representation of the signal, where each of the numbers we used for the description (positions, amplitudes) is called a *parameter*. The initial structure we imposed on the signal — namely that it is a succession of 5 spikes with shared locations between channels — is called a *model*.

Parametric estimation is the task of determining the parameter values from a set of observations called *measurements*. The difficulty may lie in the incompleteness of the measurements, their corruption by noise or the computational complexity of the estimation itself.

If we look again at our communication example some time later, the transmitter — which has moved in the meantime — sends another impulse, and this time the listening device records

What is first seen, is that these signals also have a structure, albeit quite different. This difference occurred despite the fact that they both come from the same experimental setup and were recorded within a short period of time.

The central question is now to find how to best represent these signals, which is the task of *modelization*. The unique model we used to describe the channels in the first figure does not describe the second figure adequately.

Ideally, a model should describe a signal with as few numbers as possible (the parameters), so that a minimal amount of information is required to identify a particular instance.

The obvious trade-off is that a short signal description can only represent a limited category of signals, which may pose a problem given the structural changes we have seen in the previous example.

Therefore, a "one size fits all" model is a chimera, and modelization should be concerned not only with the description of a model but also with the identification of its practical range of application. To do so, a model must be linked to the physical

properties of the reality it describes. We will focus on this task in Chapter 1 as a preliminary.

Detection and estimation

Once a collection of models describes adequately what we may observe, one of them must be chosen given an observation (measurements). This task is called *detection*. The goal is to select a model to meet a particular objective, such as minimizing the energy of the difference between an original signal not corrupted by noise and a corrupted observation, or estimating the time at which the first spike arrived, and so on. More fundamentally, a model should capture what is intelligible in an observation, *i.e.* make visible the causes by which an observation is what it is.

Once a model has been chosen, the *estimation* of its parameter can take place, as we mentioned previously.

In Chapters 2–4 We will study the *estimation* and *detection* of the mobile communications channels studied in Chapter 1. We will be concerned with accuracy as well as computational complexity, since mobile communications have strict time and power and complexity constraints.

The tools of the trade

To solve engineering problems, one should have in its toolbox instruments to analyze the performances of a proposed solution. In estimation theory, delimiting a range for the precision with which the numerical value of a parameter can be estimated is a cornerstone in the analysis of performance.

In communications, many parameters are *periodic*. A concrete example of a periodic parameter — not from communications — which can be easily grasped, is the wind direction. For example, its range can be taken from $0°$ to $360°$, and its periodic nature lies in the fact that both ends of this interval are neighbors, e.g. moving $0.1°$ backward from $0°$ leads to $359.9°$ which is at the other end of the range.

From this simple example, we see that the periodic nature of a parameter must be taken into account when judging how far an estimate is from the true value. In Chapter 5 we will get more familiar with the notion of periodic localization through the study of *Heisenberg uncertainty principle* on a periodic domain. Then we will derive lowerbounds on the variance of periodic estimators.

Historical notes

Parametric estimation of communication channels is a well-studied and mature topic. An early model, called the *multipath channel model* can be traced back to the PhD thesis of Turin in 1956 [130]. While the communication medium remains the same, the devices exploiting it have tremendously changed. First, the allocated bandwidth has evolved and since around 1990, multiple antennas systems started to be more widely deployed, especially due to the capacity breakthrough achieved with space-time coding [7; 124]. The evolution of these two factors lead to sensibly different channel models, which require different processing techniques. Also, the computational power

of communication devices has greatly improved thanks to Moore's law, which allows for more advanced processing.

The methods we will consider come from the antenna arrays literature [137; 108; 140] for the most part and use the classical tools of spectral estimation [119]. This field was heavily researched (and funded) for its military applications, e.g. radar detection. In these systems, the number of antennas tends to be large as they are not supposed to fit in a pocket, and it resulted in multiple parallel channels to be estimated. In today's communications, the number of antennas is not as large, but what has been lost in the spatial domain was gained in the temporal domain. By trading off one for the other, we will lose some properties and retain other, and most of the analysis will have to adapt to these modifications.

On the theoretical side, uncertainty principles date back to the beginning of the 20^{th} century [66; 111] and played an important role in the early years of quantum physics, and lead to the characterization of a particle's state by a probability distribution called the *wave function*.

More prosaically, uncertainty principles stem from a lack of commutativity between two operators such as the localization and the momentum or the localization in time and the localization in frequency. The later interpretation made them an important result in time-frequency analysis — with numerous applications in filter design, image processing, multiresolution analysis, etc — the pioneering work is due to D. Gabor in 1946 [58].

Surprisingly, some elementary facts about uncertainty principles for continuous and infinitely supported functions are not available for periodic waveforms or infinite sequences (equivalence using Fourier series), even though a periodic uncertainty principle exists [35] and has been analyzed [99; 100; 126; 52; 72]. What is missing in particular, is the knowledge of periodic waveforms having the same periodic time-frequency characteristics as Gaussians functions have in non-periodic spaces.[4]

Other theoretical tools are lowerbounds on the variance of an estimator, such as the *Cramér-Rao* lowerbound [48]. An equivalence between a bayesian version of the Cramér-Rao bound [132] and the Heisenberg uncertainty principle was shown by A. Dembo in 1990 [50] using the work of Stam on the entropy power inequality (EPI). The equivalence between what is usually called the Cramér-Rao bound – *i.e.* a non-bayesian lowerbound – and the Heisenberg principle could not be found in the literature.

Outline and contributions

The present thesis will be composed of two parts which could be seen as "Applications" and "Theory" — thus following the views of Husserl. Contributions to modelization and algorithms will be found in Part I, while more fundamental and analytical results will be found in part II.

[4] The prolate spheroidal functions of Slepian [118] do not exactly answer this question.

Part I : Estimation & detection of sparse channels

In Chapter 1, the *Sparse common support* (SCS) channel model is proposed and analyzed. It is a joint multipath model, where the path locations are shared among channels. We derive the elementary properties the communication channels should have in order to be fitted by this model. Specifically, the requirements for sparsity are found in (1.2) and involve such quantities as the channel bandwidth, the numbers scatterers and their dimension, the delay-spread and the propagation speed. For the common support property, Proposition 1.1 links the dimension of the antenna array to the bandwidth of the channel. To complete the characterization, a formula for the path amplitudes correlation is given in Proposition 1.2. A concrete example of SCS channels is discussed.

In Chapter 2, we review subspace estimation methods such as the annihilating filter method and ESPRIT from the point of view of the *data matrix* and one of its decomposition called the *Vandermonde decomposition*. We show that these algorithms can be used to *jointly* estimate SCS channels.

Then, we show that the complexity and memory requirements for the joint estimation can be lowered from $\mathcal{O}(PM^3)$ and $\mathcal{O}(PM^2)$ to $\mathcal{O}(PKM\log(M))$ and $\mathcal{O}(PKM)$ respectively, where M is proportional to the number of measurements per channel, P is the number of channels and K is the number of paths per channel. To obtain this result, a convergence theorem is adapted to fit the measurement setup (Theorem 2.3).

For a small and constant K, the estimation can be considered to be *superfast* and *in-place*, to use the classical algorithmic terminology.

Both the original and the accelerated methods are implemented using state of the art libraries and significant speed improvements are shown for a number of measurements per channel larger than 100 and the same estimation error as in the original method.

In Chapter 3, we address the detection problem, specifically the determination of the number of paths. For detection, methods based on information criterion cannot be computed as effectively as it is the case for large antenna arrays. Having less samples in space, it is necessary to replicate measurements to build a data matrix to work on. This replication leads to a non-flat spectral distribution for the noise matrix.

Two criteria based on the properties of the noise are proposed. The first one is fairly coarse and can be evaluated at no extra cost within the superfast estimation algorithm. The second one can be used to target a chosen *false detection rate* and is applied after the estimation as a validation step, to potentially adjust the result.

A third, purely geometric criterion is proposed called the *Partial effective rank*. It monitors the evolution of the intrinsic dimension of the signal subspace in the larger measurement space. This intrinsic dimension is based on a smooth surrogate of the rank, and can therefore cope with the presence of noise – unlike the matrix rank.

To conclude Chapter 3, a test is run on channel impulse responses recorded in a suburban environment to which synthetic noise was added. We compare the proposed superfast estimation algorithm combined with the partial effective rank detection to a method which exploits the joint sample sparsity (few non zero samples) of the channels — and we show that the sparse common support model suits the most at low SNR, and that an appropriate detection method overcomes sudden and radical changes of

the channels (e.g. going through a tunnel).

In Chapter 4, we study how the principles exposed earlier can be used in an iterative fashion. Iterative estimation algorithms are especially useful where an initial estimate is known a priori, yielding a *tracking* algorithm. A connection is made with the family of *rake receiver* algorithms.

Part II : Fundamental Limits on Periodic Localization

The channels studied in the first part were excited by periodic signals, so that the recorded impulse responses were parametrized by non periodic parameters (the paths amplitudes) and periodic parameters (the paths localization, known as *time of arrival*). Periodic parameters are also commonly encountered in bearing estimation problems.

Firstly, to gain some insights on localization in periodic domains, we will study the *time-frequency product* of periodic waveforms, also known as *time-frequency uncertainty* using the terminology of quantum physics. The Fourier dual of a periodic waveform is its Fourier series, and so the time-frequency product of waveforms is by the unitary nature of the Fourier transform equivalent to the time-frequency product of (infinite length) sequences, which is easier to deal with from a numerical point of view.

One of the main difference between the non-periodic and the periodic case is that the lowerbound on uncertainty given by the *Heisenberg/Breitenberger uncertainty principle* cannot be achieved by periodic waveforms, unless they are infinitely narrow over the period (their variance tends to zero) [99]. This limitation contrasts sharply with the non-periodic case, in which gaussian functions of any spread are known to meet exactly the lowerbound. Since the lower limit is not achievable by periodic waveforms with an arbitrary spread, it is of interest to know what is the minimal achievable time-frequency product for a given spread, and which waveforms achieve it.

In Chapter 5, we formulate an optimization program which for a given periodic waveform spread, generates its Fourier series such as to minimize its time-frequency product. We call such a sequence/waveform a *maximally compact sequence/waveform*.

The optimization is a primal semi-definite program (SDP) (Theorem 5.2) for which strong duality holds (Lemma 5.3). By analyzing the boundary of the feasible region of the dual formulation of the SDP, we obtain an analytical formula for maximally compact waveforms (Theorem 5.3). This formula is the *harmonic Mathieu equation* which solutions are *harmonic Mathieu functions* [80]. Among this set of functions, we show that *Mathieu's harmonic cosine of order 0* generates all maximally-compact waveforms up to a shift and/or a modulation.

An interesting parallel shall be made with Slepian's *prolate spheroidal wave functions* (PSWF) [118]. These functions are solutions of a differential equation similar to Mathieu's equation (both equations are *Sturm-Liouville* equations [10]). The main difference between the two solutions is particularly clear from the definition of the problem they respectively solve. PSWF minimize the spread of a periodic waveform under a strict bandlimiting constraint on its Fourier series. In our case, instead of a bandlimiting constraint on the Fourier series, we minimized its spread. It shall be

seen as a "soft" penalization scheme instead of a hard constraint[5].

In Chapter 6, we show that the *Cramér-Rao bound* (CRB) — a lowerbound on the variance of unbiased estimators — can be formulated as an uncertainty principle (Lemma 6.1 and Theorem 6.1). This general formulation makes the derivation of a CRB for periodic parameters straightforward, and we obtain a periodic CRB[6] as a simple corollary (Corollary 6.1). It highlights that the definition of localization in a periodic domain plays a central role. With the definition from Chapter 5, the periodic CRB has the same form as in the non-periodic case (it is the inverse of the *Fisher information*), which allows to use all the existing literature on the subject — e.g. [119] Appendix B.6, [143] — and to obtain lowerbounds for the joint estimation of periodic and non periodic parameters (Theorem 6.2 and Corollary 6.2).

A truly periodic definition of the CRB is by itself of limited interest. However, having a rigorously valid formulation makes it much easier to look for potentially stronger lowerbounds for periodic parameters. Stronger lowerbounds are obtained by replacing the derivative in the formulation of the CRB by a different linear operator. Finding the linear shift invariant filter maximizing the lowerbound is known as the *Barankin bound approximation* problem. We solve this problem analytically for a single periodic parameter (Theorem 6.3 and Corollary 6.3), and observe a gap between our solution and what is achievable in practice (MMSE estimator) — see Examples 6.f–6.g. It indicates that the Barankin bound is not necessarily tight (more thorough discussion in Section 6.3.3).

[5]In practice, our numerical formulation assumes a finite length for the Fourier series, which effect is negligible if it is long enough. For design purposes, this length can be reduced, in which case the solution is no more the Fourier series of a Mathieu function.

[6]The formulation has a range of applications broader than the CRB itself, for example, a periodic version of the Hammersley-Chapman-Robbins bound (HCRB) is shown in Example 6.e

Part I

Estimation & Detection of Sparse Channels

“ Ἐπίστασθαι δὲ οἰόμεθ᾽ ἕκαστον ἁπλῶς, ἀλλὰ μὴ τὸν σοφιστικὸν τρόπον τὸν κατὰ συμβεβηκός, ὅταν τήν τ᾽ αἰτίαν οἰώμεθα γινώσκειν δι᾽ ἣν τὸ πρᾶγμά ἐστιν, ὅτι ἐκείνου αἰτία ἐστί, καὶ μὴ ἐνδέχεσθαι τοῦτ᾽ ἄλλως ἔχειν. ”

Αριστοτέλης ο Σταγειρίτης — Αναλυτικων Υστερων, κεφαλαιον Β΄.

“ Nous pensons savoir les choses d’une manière absolue et non point d’une manière sophistique, purement accidentelle, quand nous pensons savoir que la cause par laquelle la chose existe, est bien la cause de cette chose, et que par suite nous pensons que la chose ne saurait être autrement que nous la savons. ”

Aristote le Stagirite — Seconds Analytiques, ch.2.

“ We suppose ourselves to possess unqualified scientific knowledge of a thing, as opposed to knowing it in the accidental way in which the sophist knows, when we think that we know the cause on which the fact depends, as the cause of that fact and of no other, and, further, that the fact could not be other than it is. ”

Aristotle the Stagirite — Posterior Analytics, ch.2.

Chapter 1

Parametric models for communications

1.1 Parametric or not parametric? — that is the question.

1.1.1 What are parametric models?

A parametric signal model is a mapping from a countable set of scalar numbers — called *parameters* — to a signal space, which can take several forms (vector space, union of subspaces, ...). Therefore, many signals model — bandlimited waveforms, piecewise constant signals, ... — falls into the parametric category.

More specifically, the denomination "parametric" is usually reserved to models where parameters represent the degrees of freedom of the signal. For example, consider a signal $s(t)$ made of a known bandlimited waveform $w(t)$ of period 2π arbitrarily shifted and scaled

$$s(t) = c_0 \cdot w(t - t_0).$$

Since the signal is periodic, it is characterized by its continuous-time Fourier series (CTFS), and because it is bandlimited, it has only a finite number of non-zero CTFS coefficients $[S[-M], \ldots, S[M]]^T \in \mathbb{C}^{2M+1}$.

On the one hand, since the waveform $w(t)$ is known, $s(t)$ is unequivocally characterized by $\begin{bmatrix} c_0 \\ t_0 \end{bmatrix}$.

If both representations admit a finite number of parameters, only the latter one is directly linked to the effective unknowns in the signal — its degrees of freedom.

The estimation of this reduced number of coefficients from a set of samples of $s(t)$ is usually called *parametric estimation*. In the present example, it is a non-linear estimation problem, and the word *parametric* stresses the difference with the estimation of its CTFS coefficients which is a simple linear estimation problem — a projection onto a linear subspace in this particular case.

1.1.2 Advantages of parametric models

From an idealized point of view, parametric models, *i.e.* models having the signal degrees of freedom for parameters, possess mostly benefits.

First, with an adequate sampling scheme, one may sample at a rate close to the *rate of innovation* of the signal (the number of degrees of freedom per unit of time) and still be able to reconstruct it. The theory of *Finite Rate of Innovation* (FRI) sampling, studies the theory and the algorithms for sampling at the rate of innovation. Signal models such as sum of periodic waveforms, piecewise polynomials, exponential splines have been studied in relation to FRI sampling.

The usage of a model with as few parameters as possible makes estimation more robust against noise and measurement errors. Indeed, by reducing the number of parameters, the pre-image space of the signal — *i.e.* the parameter space — shrinks in dimension. In the previous example, the parameters belong to $\mathbb{C} \times [-\pi,\ \pi[$ which is much smaller than the space of its CTFS coefficients $\mathbb{C}^M$ ($M \gg 2$).

Therefore, the inverse mapping from the measurements to the parameters is in general better conditioned.

Also, a reduced number of parameters implies a description of the signal can be encoded on fewer bits. For example, this is important in MIMO communications where the transmitter can form "beams" to increase the communication bit-rate. To do so, it needs an estimate of the communication channels; these estimates are known at the receiver. If the receiver can efficiently encode the channels impulse response, a lesser portion of the channel capacity is used for this exchange of information.

In a nutshell, the key feature of parametric models is to be extremely *rigid*. A model with few parameters can only generate signals with particular attributes, and it makes their estimation more robust.

1.1.3 What are the limitations?

If rigidity is the main advantage of parametric models, it is also their Achille's heel.

To quote the statistician George E.P. Box, "All models are wrong, some are useful". And the more rigid a model, the higher the chances to have a mismatch between this model and the physical phenomenon it describes.

Again, the model *rigidity* plays the central role, for the worse this time.

Let's go back to our toy example and consider the signal is not made of a single waveform, but a cluster of them

$$\tilde{s}(t) = \sum_{k=1}^{K} c_k \cdot w(t - t_0 - \Delta t_k).$$

The number of parameters is now $2K$. If we collect samples corrupted by noise

$$\tilde{s}_n = s(t - nT) + \sigma \cdot E_n,$$

where E_n are identically and independently distributed (iid). standard normal random variable, then we face a dilemma. For Δt_k small enough, there is a noise power σ^2 above which the shifts Δt_k cannot be reliably estimated. In these circumstances it may be preferable to use the "wrong" signal model $s(t)$. Also, when σ is small enough and K is large, the bandlimited signal model becomes the most parsimonious model.

We see from this little experiment that different models may coexist to describe a single physical reality. The more rigid models provide a useful regularization when the signal to noise ratio (SNR) decreases.

1.2 A parametric model for communications

In this section, we motivate a simple and rigid model for multiple output wireless communications. Wireless communications are carried over a finite band of the electromagnetic (EM) spectrum, and so the bandlimited model with Shannon-Nyquist sampling and reconstruction provides a safe and reliable model. The goal is to study which additional properties of the EM channel we can establish to make channel estimation — and thus communication — more accurate with low power (low SNR).

1.2.1 The multipath model

The first thorough study of parametric channel models for wireless communications is the work of G. Turin [130].

A receiving device operates over a channel $\mathcal{H}$

$$x(t) \;=\; \mathcal{H}\{s\}(t),$$

where $s(t)$ is a bandlimited and periodic signal sent by the transmitting device.

The physical properties of the channel are the following

- It is non-dispersive and linear
- It is locally time-invariant
- The propagation medium contains reflecting and scattering objects

The linearity and time-invariance of the channel, imply that the effect of $\mathcal{H}$ on the transmitted waveform can be written as a convolution with an impulse response $h(t)$ — called the *channel impulse response* (CIR),

$$x(t) \;=\; (h * s)(t).$$

Scattering objects are sources of many independent reflections, and so they can be modeled as clusters of reflections

$$h(t) = \sum_{k=1}^{K} \sum_{(A_\ell, \Delta t_\ell) \in \mathcal{C}_k} A_\ell \cdot \varphi(t - t_k - \Delta t_\ell),$$

where $\mathcal{C}_1, \ldots, \mathcal{C}_K$ are clusters centered at time $t_1, \ldots, t_K$, $(A_\ell, \Delta t_\ell) \in \mathbb{C} \times [-\pi\ \pi[$ are the random amplitude and delays of individual reflections, and φ is the channel mask, *i.e.* a waveform which occupies a finite portion of the spectrum. In the Fourier domain, the CIR becomes

$$H[n] = \sum_{k=1}^{K} e^{-j2\pi n t_k} \Phi[n] \cdot \sum_{(A_\ell, \Delta t_\ell) \in \mathcal{C}_k} A_\ell e^{-j2\pi n \Delta t_\ell}\ , \qquad |n| \leq M.$$

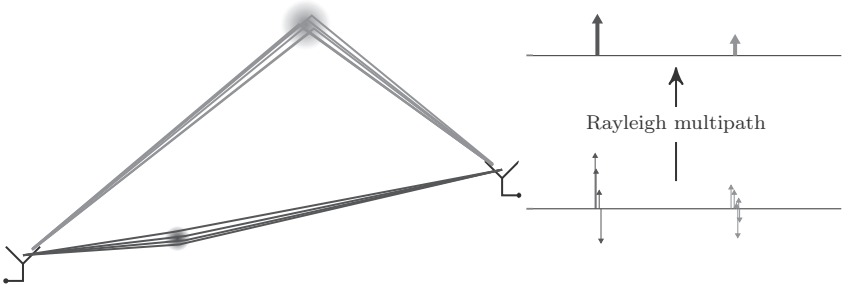


Figure 1.1: *An accurate multipath model comprises clusters of reflections called* scatterers. *The* Rayleigh-fading multipath model *is an approximation of this model in which the reflections from each scatterer are aggregated into a single reflection called a* path. *Because of the large number of reflections and their independent polarization, the central limit theorem implies that the paths amplitudes are complex-valued independently distributed normal random variables. Therefore, the magnitude is Rayleigh distributed, which gave its name to the model.*

where Φ is the CTFS of φ. We can now formulate the two assumptions central to the *Rayleigh fading multipath channel model*

1. If the intra-cluster delays Δt_ℓ are substantially smaller than the inverse bandwidth ($\Delta t_\ell \ll 1/(2M+1)$), then one can make the 0^{th} order approximation

$$e^{-j2\pi n \Delta t_\ell} \approx 1 \ , \qquad n \leq M.$$

2. If the random amplitudes A_ℓ are 0-mean independently distributed random variables (with finite variance), then by the central-limit theorem

$$C_k \ = \sum_{(A_\ell, \Delta t_\ell) \in \mathcal{C}_k} A_\ell$$

are independent normally distributed random variables, $C_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, c_k^2 \mathbb{I})$.

Definition 1.1. *A* Rayleigh fading multipath channel *with K paths has the impulse response*

$$h(t) \ = \ \sum_{k=1}^{K} C_k \varphi(t - t_k) \ , \tag{1.1}$$

where

$$C_k \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, c_k^2 \mathbb{I})$$

are independent random variables. The coefficients C_k are called the path amplitudes *and t_k are called the* times of arrival *(ToA).*

The multipath channel model is also present at the foundations of spread-spectrum communications such as CDMA. The *rake-receiver* uses an estimate of the times of arrival and the path amplitudes to combine the paths coherently with weights chosen to maximize the channel equalization gain. For more on the subject, see [89].

1.2.2 The conditions for sparsity

The multipath model tells us that the channel impulse response is composed of a small number of paths. However, this number should be considered relative to the delay-spread and the bandwidth to be relevant.

Definition 1.2. *Let h be a multipath channel with K paths. Assume uniformly distributed time of arrival $t_1 \leq \cdots \leq t_K$ over an interval defined as the* delay-spread *of the channel, i.e. $\tau \geq t_K - t_1$.*

Then the channel h is sparse *if*

$$\frac{\pi}{\tau} K \ll 2M + 1 \ ,$$

i.e. if the rate of innovation is substantially smaller than the Nyquist rate.

From this definition, we can already see two competing trends for EM channels to be sparse

1. To be sparse, the multipath model must hold, which requires a *low enough bandwidth* (so that clusters form paths).
2. To be sparse, the path density must not be too high compared to the bandwidth according to Definition 1.2. Therefore, *the bandwidth must be high enough*.

This trade-off shall be kept in mind when trying to apply sparse methods to estimate EM channels; Figure 1.2 shows how, for a fixed cluster width, the delay-spread and bandwidth influence the relevance of a sparse model.

The delay-spread, the cluster width and the inverse bandwidth can be expressed in terms of physical distances — with c the speed of light

1. The delay-spread is the maximum propagation time difference between paths. If $d_{\max}$ is the maximum difference between the distances of propagation between paths, the delay-spread is expressed as
$$\tau = \frac{d_{\max}}{c}.$$
2. The maximal ToA difference within a cluster $\Delta t_{\max}$ is linked to the physical dimension of scatterers. If scatterers have a maximum radius $r_{\max}$ then
$$\Delta t_{\max} \leq \frac{2r_{\max}}{c}.$$
3. The bandwidth is $B = \frac{2M+1}{T}$Hz, the associated physical distance is the *minimum wavelength* $\lambda_{\min} = c \cdot B/2$.

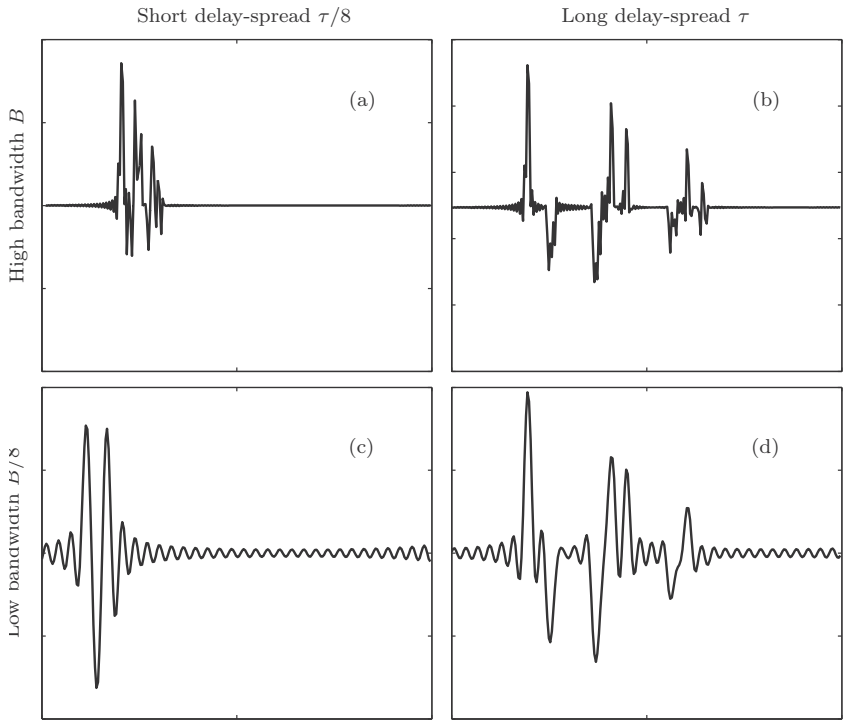


Figure 1.2: *This figure shows how sparsity relates to the channel bandwidth and its delay-spread. All four panels (a)–(d) have the same number of signal components — 80 of them grouped in 8 clusters with exponentially fast energy decay. Signals (a) and (c) cannot be considered sparse as the rate of innovation is close to/greater than the Nyquist rate on the time-lapse corresponding to the delay-spread, and 0 outside it. Signal (b) is weakly sparse, the rate of innovation is for this reason also high. In this setup the sample sparsity approach may be suitable. The signal (d) can be considered sparse as only the 8 clusters will be resolvable in the presence of noise. The rate of innovation of this approximation is much lower than the Nyquist rate. Even though (b) and (d) have the same rate of innovation in a strict sense, (d) can be approximated with a signal having $1/10^{\text{th}}$ the rate of innovation of (b) thanks to its low bandwidth. This approximation motivates the use of a model with a low rate of innovation in the low-SNR regime where the model approximation error has less power than the noise.*

The requisites of having a cluster density substantially lower than the Nyquist-rate and also unresolvable paths within each clusters, is expressed as the inequation

$$\Delta t\text{max} \ll \frac{1}{B} \ll \frac{\tau}{K}.$$

And so, the sparse approximation requires that in the spatial domain

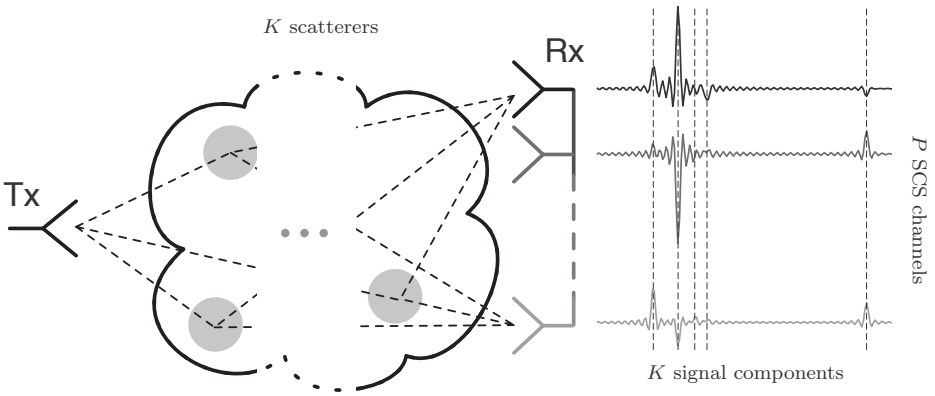


Figure 1.3: *The ideal SCS channel model is a set of P channels of bandwidth B each having K components aligned in time. Assuming complex valued signal components, the total number of unknowns is $(2P+1)K$ instead of $3PK$ for a sparse model with independent time of arrivals (ToA), or $2P$ times the Nyquist Rate for a bandlimited model.*

which implies that the size of the scatterers should be modest compared to the difference between distances of propagation.

$$2r_{\max} \ll \frac{c}{B} \ll \frac{d_{\max}}{K}. \tag{1.2}$$

In Figure 1.2, only (d) verifies (1.2). For the signal Figure 1.2.(b), a model with a small number of non-zero samples in the time-domain (sample sparsity) could be indicated. For Figure 1.2.(a)(c), a bandlimited model limited to a window corresponding to the delay-spread would be the right choice.

It is immediately visible from this simple example that as a model becomes more rigid its range of application narrows, and deciding correctly when to use it becomes critical.

1.2.3 The common support assumption

We have seen how sparsity occurs in point-to-point communications. Modern communications systems go beyond point-to-point communications by having multiple inputs and multiple outputs (multiple transmitting and receiving antennas). We turn our attention to the models describing the channels between a unique transmitter and several receivers[1] (SIMO).

As shown in Figure 1.3, if the assumptions of (1.1) hold, having P receiving antennas creates P multipath channels with *channel impulse responses* given by

[1]The multiple input case (MIMO), is especially relevant for space-time coding [7], we focus here on the estimation of the channels themselves

$$h_p(t) \;=\; \sum_{k=1}^{K} C_{k,p}\varphi(t - t_{k,p}) \;,$$

where $C_{k,p} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, c_k^2\mathbb{I})$.

If the amplitude of reflections from different scatterers are independent, the correlation between $C_{k,p}$ and $C_{k,q}$, $p \neq q$, amplitudes of reflections from a same scatterer in different channels, requires some thoughts and we will propose a model in the next subsection.

For the *times of arrival* $t_{k,p}$, $k = 1, \ldots, K$, $p = 1, \ldots, P$ a *common support property* may immediately hold :

Proposition 1.1. *If the maximum distance between antennas is bounded from above by* $\Delta_{\max}$ *and*

$$\Delta_{\max} \;\ll\; \frac{c}{B} \;,$$

then, one may make the common support *approximation for all* $k = 1, \ldots, K$

$$t_k \;\stackrel{\text{def}}{=}\; t_{k,1} \approx t_{k,2} \approx \cdots \approx t_{k,P}.$$

Spatial correlation

The last important piece to finish the modelization is to determine the correlation between the path amplitudes $C_{k,p}$ and $C_{k,q}$. To this end a modelization of the scatterers is required. If we assume propagation on a 2-dimensional plane[2], each scatterer can be modeled as a bundle of reflections which localizations are independently drawn from a bivariate probability density. With the isotropic multivariate normal distribution and following the methodology in [110], the path amplitudes correlation is given by

Proposition 1.2. *In the physical layout of Figure 1.4, the correlation of path amplitudes is accurately modeled as*

$$\begin{aligned}\frac{\mathbb{E}\left[C_{k,p}C_{k,q}^*\right]}{\sqrt{\mathbb{E}\left[|C_{k,p}|\right]\mathbb{E}\left[|C_{k,q}|\right]}} \;&= \delta_{k-\ell}\left[J_0\left(2\pi\frac{d_{p,q}}{\lambda_c}\right)\right. \\ &\left. +2\sum_{l=1}^{\infty} j^l \frac{I_l(\kappa_k)}{I_0(\kappa_k)} J_l\left(2\pi\frac{d_{p,q}}{\lambda_c}\right)\cdot\cos\left[l\left(\theta_{k,p,q} - \frac{\pi}{2}\right)\right]\right],\end{aligned} \tag{1.3}$$

[2] A similar analysis holds in 3-dimensions.

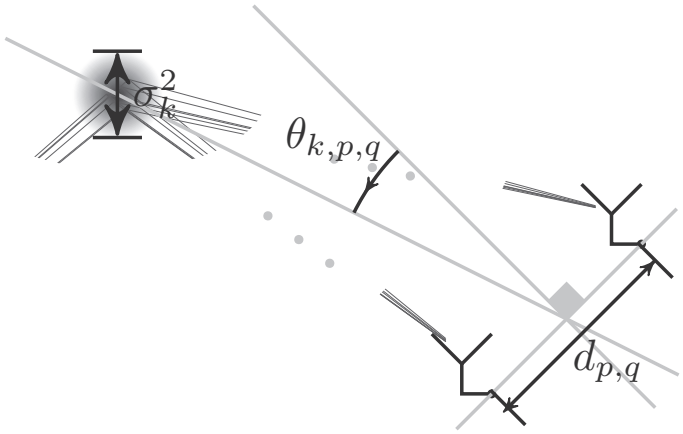


Figure 1.4: *The model makes a "far-field" assumption in which the azimuth of the scatterer is the same for every antenna. With this assumption, the angle between the normal to the antenna pair (p, q) in the array and the azimuth of cluster k relative to antennas p or q are the same. This angle is noted $\theta_{k,p,q}$. The cluster width is noted σ_{k^2} and the distance between the antenna array and the cluster is Δ_k (the far-field assumption makes the distance between the cluster and an antenna the same for each antenna of the array). The euclidean distance between antennas p and q is noted $d_{p,q}$.*

where $\Delta_k^2/\sigma_k^2 \approx (1 - e^{-3\kappa_k/4})\kappa_k$, J_n is the n^{th} Bessel function of the first kind and I_n is the n^{th} modified Bessel function of the first kind.

Proof.

See Appendix A. □

Corollary 1.1. *For $\kappa_k \to 0$ (narrow scatterer):*

$$\frac{\mathbb{E}\left[C_{k,p}C_{\ell,q}^*\right]}{\sqrt{\mathbb{E}\left[|C_{k,p}|\right]\mathbb{E}\left[|C_{\ell,q}|\right]}} = \delta_{k-\ell}J_0\left(2\pi\frac{d_{p,q}}{\lambda_c}\right), \tag{1.4}$$

where λ_c is the wavelength of the carrier frequency.

With (1.4), the often used antenna distances $d_{p,q} \geq \lambda_c/2$ yield a correlation of magnitude of at most 0.4, as can be seen in Figure 1.5.

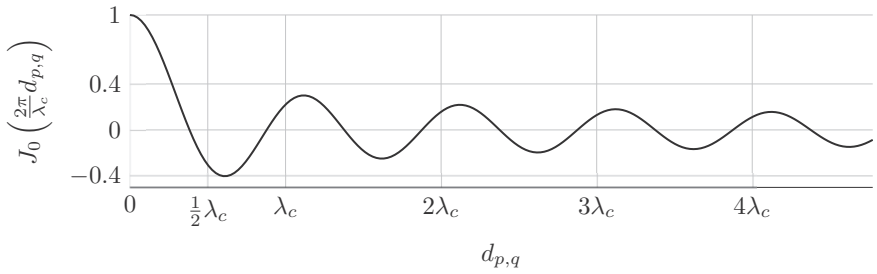


Figure 1.5: *Correlation between path amplitudes at antennas p and q separated by a distance $d_{p,q}$ according to the model (1.4) for narrow scatterers.*

Definition 1.3. *The P channels of a* Sparse Common Support *model with K Rayleigh fading paths have for CIR*

$$h_p(t) \;=\; \sum_{k=1}^{K} C_{k,p}\varphi(t-t_k)\;, \tag{1.5}$$

with path amplitudes $C_{k,p} \sim \mathcal{N}(\mathbf{0}, c_k^2\mathbb{I})$ and cross-correlation defined by (1.3).

Therefore the channels of the sparse common support model share a set parameters $\{t_k\}_{k=1,\dots,K}$ and have their own path amplitudes.

Measurements of sparse common support channels will therefore share a common set of parameters and have also parameters of their own. This calls for *joint estimation techniques* which we will develop in the next chapter.

1.2.4 An example of sparse common support channels

We now validate the sparse common support model on a set of measurements collected by the FTW laboratory in Vienna [67]. We will subsequently call these data the *Weikendorf dataset*, from the place it was collected in.

The properties of the Weikendorf dataset are listed in Table 1.1, and Figure 1.6.(a) shows the CIR over time.

The properties necessary for the *sparse common support* hold if :

1. $\frac{c}{B} \;=\; \frac{3\cdot 10^8 \mathrm{m\cdot s^{-1}}}{120\mathrm{MHz}} \;=\; 2.5\mathrm{m}$. Approximately, scatterers should have a size up to one meter, and the distance between them should be more than 10m.

2. Receiving antennas form a linear array with 8 elements separated by a distance $\lambda_c/2 = 7.5\mathrm{cm}$. Therefore the maximum distance between antennas is

$$\Delta_{\max} \;=\; 60\mathrm{cm} \;\ll\; 2.5\mathrm{m} \;=\; \frac{c}{B}.$$

The common support assumption is relevant.

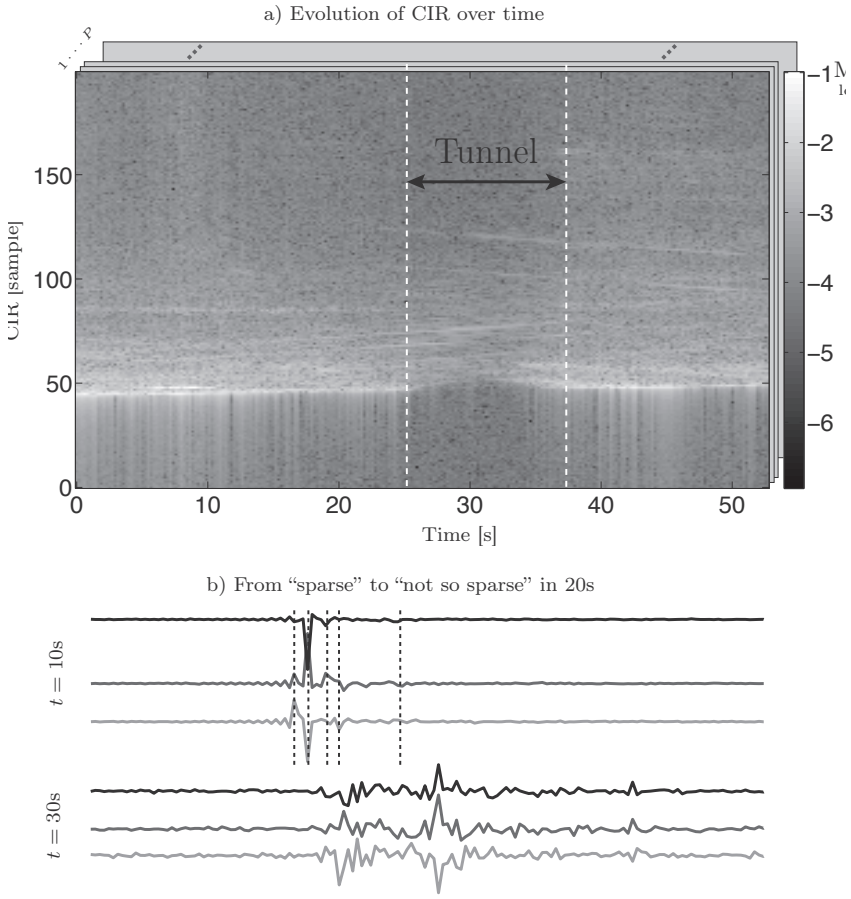


Figure 1.6: *Field measurements from the Weikendorf dataset [67]. The receiver is a base-station with $P = 8$ antennas, and the transmitter is mobile. The image a) shows the magnitude of the first antenna's CIR. The channel is qualitatively sparse except when the transmitter goes through a tunnel. The real part of the CIR for three different antennas is shown in b) confirming the common support property and the transient nature of sparsity.*

3. If the properties necessary for sparsity hold, and one can use (1.4) to show that the path amplitude cross-correlation does not exceed 0.4 as shown in Figure 1.5, and therefore path amplitudes are not shared between the channels.

Sparsity as a transient property

The bandwidth, propagation speed and antenna topology are fundamental properties of a communication system which is not subject to change over time. On the other

Table 1.1: *Properties of the Weikendorf dataset [67]*

Property	Value
Center frequency f_c	2 GHz
Center wavelength λ_c	15 cm
Bandwidth	120 MHz
Mobile Tx	15 monopole antennas uniformly arranged on a 30 cm diameter circle at 1.5 m from the ground
Static Rx	8 patch antennas separated by 7.5 cm ($\lambda_c/2$) forming a linear array at 20 m from the ground
Time interval between snapshots	21 ms
Tx speed	3 to 6 km/h.
Recording	About 1 minute. The Tx travelled a distance of about 50–80 m) and went through a tunnel

hand, the landscape of scatterers may evolve as the transmitter or the receiver move (mobile communications) or as reflecting objects move (mobile environment).

Therefore, the *sparsity pattern* of a channel may change over time. This is specially true in the Weikendorf dataset — see Figure 1.6.(b) — in which the transmitter goes through a tunnel : the many reflections inside the tunnel invalidate the sparse multipath assumption.

The dynamic nature of sparsity leads to two separate algorithmic issues

Tracking Over time, the parameters of a sparse multipath model may vary smoothly, which hints at an estimation which *tracks* the signal parameters.

Model selection An estimator of sparse models should be able to evaluate if the model assumptions are met. E.g., failure to do so would result in poor performances in the tunnel.

1.3 Conclusion

This study of the physics at work behind sparsity in wireless communications allowed us to delimitate the range of application of a rigid channel model like the Rayleigh multipath model. Our conclusions corroborate those of Berger [30] on sparse estimation of underwater acoustic channels. In the acoustic setup, the low propagation speed of sound waves ($\approx 1.5 \times 10^3 \mathrm{m} \cdots \mathrm{s}^{-1}$ in water) compared to EM waves invalidates most of the properties required by the multipath model, and in this case the less rigid sparse model used in the compressed sensing framework provides a good trade-off between the multipath model and the bandlimited model.

The sparse common support model is shown to be relevant for outdoor communications with a medium bandwidth — up to 200MHz approximately — as found in the Weikendorf dataset. Nevertheless, it does not rule out its application to ultrawideband communications as the physical properties of the channel are entirely different for short range communications[42; 73].

The challenges laid out in this chapter — which we will study next — can be summarized as

- Developing joint estimation algorithms for the *sparse common support* model.
- Developing tracking algorithms to take advantage of smooth variations of the parameter values over time.
- Developing efficient and accurate tools for *model order estimation*. This task is an instance of a *detection* problem.
- Detecting transient properties such as sparsity — which is a corollary of model order estimation (detection).

After presenting estimation and detection methods in Chapters 2 and 3 respectively, the fitness of the sparse common support model will be tested on the Weikendorf dataset presented above at the end of Chapter 3.

Chapter 2

Parametric estimation algorithms

The parametric model for communications developed in Chapter 1 sets up the fundamental — because dictated by the physics — properties of wireless communications.

If we picture communications as a succession of layers — the physical properties forming the first one — the sampling scheme completes the picture. This second layer is not dictated directly by physics, nevertheless standardization leaves few options (for good reasons). In this chapter, we will focus on OFDM based communications which are the most common in modern standards — e.g. WLAN, digital radio and TV, 4G communications[1] — and we quickly outline in Section 2.1 the features relevant to the channel estimation problem.

With this base, we then develop channel estimation algorithms in two steps. Section 2.2 reviews line spectra estimation techniques, and extends them to sparse common support (SCS) channels (see Chapter 1) which require a joint estimation of the support. After identification of the computational bottlenecks, we propose a less demanding algorithm in Section 2.3 with guarantees on the accuracy. The reduction of the computational complexity is crucial since channel estimation is a core block in the communication stack of mobile devices.

2.1 Measurements model for OFDM communications

The processing chain from the electromagnetic radiation measured by an antenna to a sequence of samples can be schematized as

- A demodulator, which converts a real-valued bandpass signal into its complex-valued baseband equivalent
- A lowpass filter

[1] We may cite a few examples: IEEE 802.11a/g/n for WLAN; DAB, DVB-T, DVB-H, and MediaFLO for digital radio and TV; 3GPP-LTE, IEEE 802.16e/802.20 for 4G communications.

- An analog to digital converter sampling uniformly and quantizing the input. We will assume an ideal uniform sampler and neglect the distortion introduced by quantization.

It is assumed that white gaussian noise is present in the measurements of the channels (no interference). Since demodulation is a linear operation, the baseband equivalent channel after demodulation is

$$y(t) \;=\; (x * h)(t) + \varepsilon(t), \tag{2.1}$$

where $x(t)$ is the transmitted signal, $h(t)$ is the channel impulse response (CIR) and $\varepsilon(t)$ is a 0-mean white gaussian process with variance σ^2.

The signal $x(t)$ is filtered with an ideal lowpass filter[2] of cutoff pulsation $\omega_0 = \frac{2\pi}{T_s}$ and critically sampled at $1/T_s$ Hz, resulting in a sequence with DTFT

$$Y(e^{j\omega}) \;=\; X(e^{j\omega}) \cdot H(e^{j\omega}) + E(e^{j\omega}) \;, \tag{2.2}$$

where $E(e^{j\omega})$ is a white 0-mean gaussian process over $[-\pi,\ \pi[$ with variance σ^2.

2.1.1 OFDM in a nutshell

Nothing was said about the properties of $x(t)$ so far as it depends on the communication protocol itself. We choose *Orthogonal Frequency Division Multiplexing*, which has the following properties

- The signal $x(t)$ is a succession of *frames*. Each frame has a duration T_f.
- Symbols are coded on the coefficients of an N-points DFT. This DFT coefficients are called *subcarriers*, and their inverse DFT is interpolated with a lowpass filter to yield a periodic waveform $x_d(t)$ of period T_d. A subset of subcarriers is reserved for signalling data called *pilots*, the other subcarriers are for the transmitted data.

2.1.2 Structure of a frame

A frame of duration T_f is composed in time of a *cyclic* prefix and the *data block* x_d as shown in Figure 2.1.

The cyclic prefix is a periodic padding of the data block. Using the notation of Figure 2.1

$$x(t) \;=\; x(t + T_d) \;, \quad t \in [t_0 - \tau_{\max},\ t_0[.$$

It serves two purposes

- Helps synchronization at the receiver side by detecting the correlation between the prefix and the matching portion in the data block.
- Makes the convolution between the CIR and the OFDM frame appear "circular" if the CIR is supported on $[0,\ \tau_{\max}[$.

[2] In practice the sampling prefilter has an impulse response close to an ideal filter.

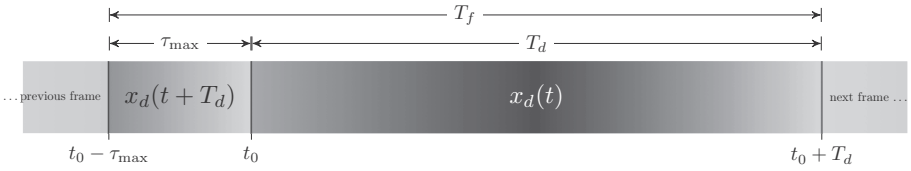


Figure 2.1: *OFDM frame format : $x_d(t)$ is the data block of the frame and a cyclic prefix of length $\tau_{\max}$ is prepended to form the frame.*

Indeed, with the latter property for $t \in [t_0 - \tau_{\max},\ t_0 - \tau_{\max} + T_d[$

$$\begin{aligned} y(t) &= (x * h)(t) + \varepsilon(t)\ , \\ &= (\widetilde{x}_d * h)(t) + \varepsilon(t)\ . \end{aligned}$$

where $\widetilde{x}_d$ is the periodization of x_d.

After sampling uniformly y over $[t_0 - \tau_{\max},\ t_0 - \tau_{\max} + T_d[$ at a rate $\frac{N}{T_d}$, the N-points DFT of the sampled signal is

$$Y[n] \ = \ X_d[n] \cdot H[n] + E[n]\ , n = 0, \ldots, N-1. \tag{2.3}$$

2.1.3 Pilot layouts and the delay-spread

To estimate the impulse response spectrum H, the receiver possess the DFT coefficients of the measurements Y.

In the noiseless case ($E[n] = 0$), if H is only supposed to be bandlimited and critically sampled, a simple dimensional argument shows that X_d must be perfectly known and everywhere non-null in order for the mapping from H to Y to be invertible; *i.e.* for the effect of the channel to be reversed.

Nevertheless, the use of a cyclic prefix was motivated by the fact that $h(t)$ is supported on $[0,\ \tau_{\max}[$, where $\tau_{\max} < T_d$ is an upperbound on the delay-spread of the CIR, *i.e.* H is also time limited.

With this assumption on the delay-spread, the classical Shannon-Nyqvist sampling theory [114] used dually — time-limited replaces bandlimited — states that H is unambiguously defined by its DFT coefficients decimated by $\lfloor \frac{T_d}{\tau_{\max}} \rfloor$ [134].

Therefore, only a decimated set of the coefficients can be reserved for values known by both the transmitting parties, called *pilots*. Without loss of generality we index these pilots on a the set $\{-M,\ \ldots,\ M\}$, yielding an (odd) number of $2M+1$ pilots per frame.

The resulting *pilot layouts* are shown in Figure 2.2. The most popular layout is by far the *scattered layout*, with the rationals that it provides the densest time-frequency sampling grid for a fixed average number of pilots.

2.1.4 Sparse common support (SCS) channels for OFDM communications

The problem of parametric channel estimation in OFDM communications with multiple antennas is know well defined. The *sparse common support model* from Defini-

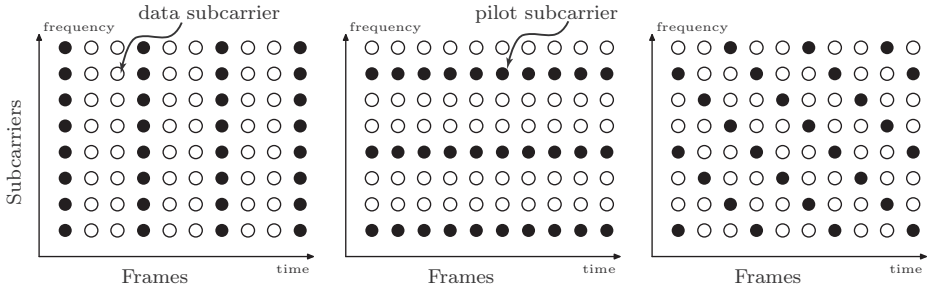


Figure 2.2: *In pilot assisted OFDM communications, subcarriers (time-frequency slots) are reserved to transmit data (○) or to probe the channel with predefined values called "*pilots*" (●). The pilot layout forms a sampling scheme for the CIR in time and frequency. The pilots form a regular lattice in time and frequency. The gap between pilots in frequency (vertical gap) is of particular interest. The separation between pilots in frequency D is called the decimation factor. In b) and c), D = 3.*

tion 1.3 is compatible with the OFDM signal specifications since it was assumed that a bandlimited and periodic signal is transmitted.

Assuming the pilots[3] take the value 1, the demodulated measurements for P SCS channels are

$$Y_p[m] \;=\; \sum_{k=1}^{K} c_{k,p} e^{-j2\pi(Dm+m_0)t_k/T_d} + E[m]\;, \quad p=1,\ldots,P,$$

and $m \in \{-\lfloor (N-1)/2D \rfloor, \ldots, \lfloor N/2D \rfloor\}$. The integer numbers D and m_0 are the gap and offset of pilots in the DFT domain[4]. A schematic view of the receiver frontend is shown in Figure 2.3.

To avoid a cluttered notation we assume without loss of generality $m_0 = 0$ and $m \in \{-M, \ldots, M\}$, and define $\omega_k = 2\pi t_k/T_d \bmod \pi$, so that the channel measurements are in the DFT domain

$$Y_p[m] \;=\; \sum_{k=1}^{K} c_{k,p} e^{-jD\omega_k m} + E[m]\;, \qquad |m| \le M. \tag{2.4}$$

2.2 Basic algorithms

In this section, we study the estimation of the time of arrival (ToA) and amplitude of paths according to the measurement model (2.4). These algorithms are generalizations of line spectra estimation techniques to multiple inputs.

[3]Pilots are usually complex numbers of unit-modulus — e.g. see [55; 54; 89] — setting their value to 1 does not restrict the range of application as a 'demodulation" of the pilots would not change the noise statistics.

[4]Varying the values of these parameters generates the different layouts shown in Figure 2.2.

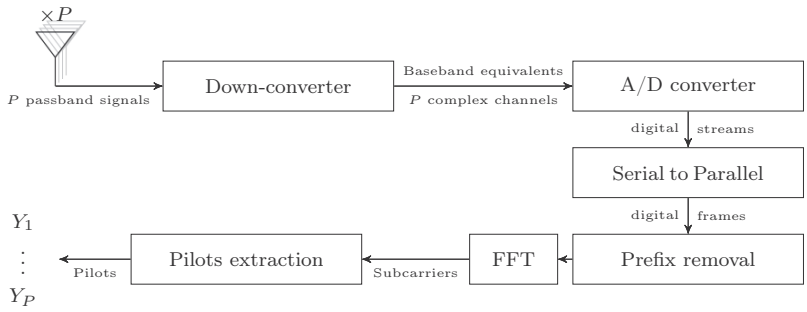


Figure 2.3: *Schematic view of the receiver from the antenna to the pilots.*

The *data matrix* is a central object that will be used in the different algorithms we will study in this chapter and the upcoming ones.

Definition 2.1. ***(Data Matrix)***

Let $\{Y_p\}_{p=1,\dots,P}$ *be measurements defined in (2.4). The* data matrix $\boldsymbol{T}$ *of dimension* L *is the block Toeplitz matrix of size* $P(2(M+1)-L) \times L$

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_1 \\ \vdots \\ \boldsymbol{T}_P \end{bmatrix},$$

such that

$$\boldsymbol{T}_p = \overset{\times L}{\begin{bmatrix} Y_p[-M+L-1] & Y_p[-M+L-2] & \dots & Y_p[-M] \\ Y_p[-M+L] & Y_p[-M+L-1] & \dots & Y_p[-M+1] \\ \vdots & \vdots & \ddots & \vdots \\ Y_p[M] & Y_p[M-1] & \dots & Y_p[M-L+1] \end{bmatrix}} \left.\vphantom{\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}}\right\} \times 2(M+1)-L$$

Definition 2.2. ***(Vandermonde decomposition)***

In the noiseless case ($\Pr[E[n]=0]=1$*) , the Toeplitz blocks* $\boldsymbol{T}_p$ *in Definition 2.1 have a* Vandermonde decomposition

$$\boldsymbol{T}_p \;=\; \mathcal{V}_{2(M+1)-L} \boldsymbol{D}_p \mathcal{V}_L^*,$$

where $\boldsymbol{D}_p$ *is a diagonal matrix of size* K*,*

$$\boldsymbol{D}_p \;=\; \mathrm{diag}(c_{1,p}, \dots, c_{K,p})$$

and $\mathcal{V}_n$ is a Vandermonde matrix of dimensions $n \times K$

$$\mathcal{V}_n = \begin{bmatrix} 1 & 1 & \dots & 1 \\ e^{-jD\omega_1} & e^{-jD\omega_2} & \dots & e^{-jD\omega_K} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-jD\omega_1 n} & e^{-jD\omega_2 n} & \dots & e^{-jD\omega_K n} \end{bmatrix}$$

From the Vandermonde decomposition, a useful lemma immediately follows

Lemma 2.1. *In the noiseless case*

$$\operatorname{rank}(\boldsymbol{T}) \leq K.$$

Proof.

Using Definition 2.2,

$$\boldsymbol{T} = \begin{bmatrix} \mathcal{V}_{(2(M+1)-L)}\boldsymbol{D}_1 \\ \vdots \\ \mathcal{V}_{(2(M+1)-L)}\boldsymbol{D}_P \end{bmatrix} \mathcal{V}_L^*.$$

Therefore, $\boldsymbol{T}$ is the sum of K rank-1 matrices and has a rank of K at most. □

If the Vandermonde decomposition is useful to prove a rank property on the data matrix, the value of the factors in this decomposition is unknown, and finding them amounts to solving the channel estimation problem.[5]

Example 2.a — Joint multipath estimation : the data matrix

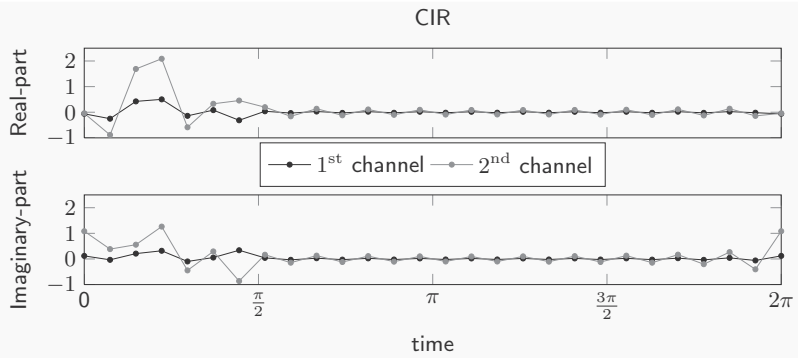
In this chapter, a small toy example will be used throughout the sections, to show the formulas "in action". This example has $P = 2$ channels with $K = 3$ paths, and 9 DFT pilots are available for each channel ($M = 4$). The gap between each pilot is $D = 3$.

The times of arrival (ToA) are respectively 0.1, 0.6 and 1.4. The power of each scatterer is $[1,\ 1.5,\ 0.5]$. The amplitudes $c_{k,p}$ for each path are realizations of uncorrelated complex normal random variables (the channels are independent Rayleigh fading). A realization of the amplitudes is for example

$$\boldsymbol{C} = \begin{bmatrix} -0.205 + 0.0929\text{j} & 0.718 + 0.423\text{j} & -0.26 + 0.385\text{j} \\ -0.556 + 1.25\text{j} & 2.95 + 1.51\text{j} & 0.697 - 0.648\text{j} \end{bmatrix},$$

yielding the CIR

[5] If t_k/T_d are integers, then the Vandermonde decomposition and the SVD coincide up to a rotation of the diagonal entries in $\boldsymbol{D}_p$ to make them real and non-negative. Furthermore the columns of $\mathcal{V}_n$ are the $n+1$ points DFT basis vectors.



The DFT coefficients of the CIRs are

$$\begin{bmatrix} \cdots & -0.33+0.67\mathrm{j} & -0.24+0.42\mathrm{j} & -0.28+0.64\mathrm{j} & \boxed{0.25+0.9\mathrm{j}} & 0.97+0.377\mathrm{j} & 0.85-0.66\mathrm{j} & \cdots \\ \cdots & -3.95+3.27\mathrm{j} & -1.57+5.25\mathrm{j} & 1.66+4.67\mathrm{j} & \boxed{3.09+2.11\mathrm{j}} & 2.34+0.08\mathrm{j} & 1.31-0.49\mathrm{j} & \cdots \end{bmatrix}.$$

The DC coefficient of the DFTs is the framed one, and one out of three coefficient is a pilot known at the receiver (shaded boxes).

From this data the receiver can build the block-Toeplitz data matrix

$$\boldsymbol{T} = \begin{bmatrix} \boxed{0.25+0.9\mathrm{j}} & -0.33+0.67\mathrm{j} & -0.87-1.16\mathrm{j} & 0.31-0.01\mathrm{j} & 0.4+0.72\mathrm{j} \\ -0.13-1.06\mathrm{j} & \boxed{0.25+0.9\mathrm{j}} & -0.33+0.67\mathrm{j} & -0.87-1.16\mathrm{j} & 0.31-0.01\mathrm{j} \\ -0.48+0.15\mathrm{j} & -0.13-1.06\mathrm{j} & \boxed{0.25+0.9\mathrm{j}} & -0.33+0.67\mathrm{j} & -0.87-1.16\mathrm{j} \\ -0.17+1.43\mathrm{j} & -0.48+0.15\mathrm{j} & -0.13-1.06\mathrm{j} & \boxed{0.25+0.9\mathrm{j}} & -0.33+0.67\mathrm{j} \\ 0.56-0.5\mathrm{j} & -0.17+1.43\mathrm{j} & -0.48+0.15\mathrm{j} & -0.13-1.06\mathrm{j} & \boxed{0.25+0.9\mathrm{j}} \\ \boxed{3.09+2.11\mathrm{j}} & -3.95+3.27\mathrm{j} & -2.95-1.01\mathrm{j} & 2.44-1.6\mathrm{j} & -1.66+2.87\mathrm{j} \\ 0.86-0.93\mathrm{j} & \boxed{3.09+2.11\mathrm{j}} & -3.95+3.27\mathrm{j} & -2.95-1.01\mathrm{j} & 2.44-1.6\mathrm{j} \\ -3.98+1.03\mathrm{j} & 0.86-0.93\mathrm{j} & \boxed{3.09+2.11\mathrm{j}} & -3.95+3.27\mathrm{j} & -2.95-1.01\mathrm{j} \\ 2.01+3.78\mathrm{j} & -3.98+1.03\mathrm{j} & 0.86-0.93\mathrm{j} & \boxed{3.09+2.11\mathrm{j}} & -3.95+3.27\mathrm{j} \\ 4.21+0.46\mathrm{j} & 2.01+3.78\mathrm{j} & -3.98+1.03\mathrm{j} & 0.86-0.93\mathrm{j} & \boxed{3.09+2.11\mathrm{j}} \end{bmatrix}, \tag{2.5}$$

which is not hermitian as the original time-domain measurements are complex-valued.

As expected from the Vandermonde decomposition, $\boldsymbol{T}$ has rank 3, and one can verify that its singular values are $[17.19,\ 6.78,\ 5.25,\ 0,\ 0]$.

2.2.1 The annihilating filter

The first estimation technique for the times of arrivals we review, is called the *annihilating filter* method, or *Prony's method*. It dates back to [101] and is used in spectral estimation [129] and FRI sampling [133; 31]. We can show from Lemma 2.1

Proposition 2.1. *Let $\boldsymbol{T}$ be the noiseless data matrix with $K+1$ columns, then there exists $\boldsymbol{a} \in \mathbb{C}^{K+1} \backslash \{\boldsymbol{0}\}$ such that*

$$\boldsymbol{T}\boldsymbol{a} = \boldsymbol{0}. \tag{2.6}$$

The vector $\boldsymbol{a}$ *is called the* annihilating filter *of* $\boldsymbol{T}$.

Proof.

The proof is a corollary of Lemma 2.1, by which $\boldsymbol{T}$ of column dimension $K+1$ is singular. Therefore its nullspace contains $\boldsymbol{a} \neq \boldsymbol{0}$. □

Moreover, the annihilating filter $\boldsymbol{a}$ is unique up to scaling if and only if the noiseless data matrix $\boldsymbol{T}$ is of rank K — necessary and sufficient conditions are found in [15].

Equation(2.6) can also be seen as a K terms *linear recursion* on the measurements. Indeed, if one scales the annihilating filter so that $a_{K+1} = -1$, then

$$[\boldsymbol{T}]_{:,1:K}[\boldsymbol{a}]_{1:K} \;=\; [\boldsymbol{T}]_{:,K+1}, \tag{2.7}$$

which shows that any DFT coefficient[6] of the signal can be written as a linear combination of the previous K ones. This interpretation is popular in coding theory, e.g. see the *Berlekamp-Massey algorithm* for linear feedback shift registers (LFSR) on finite fields [85].

The time of arrival are found as the roots of the annihilating polynomial — the polynomial of degree K which coefficients in the canonical form are the entries of $\boldsymbol{a}$.

Lemma 2.2. ***[15]***

Given $Y_p[m] = \sum_{k=1}^{K} c_{k,p} W^{m t_k}$ *for* $m = -M+K, \ldots, M$ *and* $t_i \neq t_j,\ \forall i \neq j$, *there exists a unique set of coefficients* $\{a_k\}_{k=1,\ldots,K}$ *such that:*

$$Y_p[m] = a_1 Y_p[m-1] + a_2 Y_p[m-2] + \cdots + a_K Y_p[m-K]$$

where

$$w^K - a_1 w^{K-1} - \cdots - a_{K-1} w - a_K$$

is the polynomial with roots $\{W^{t_k}\}_{k=1,\ldots,K}$.

Proof.

A linear recursion of degree K can be written as:

$$w_n = a_1 w_{n-1} + \cdots + a_K w_{n-K},\ a_K \neq 0. \tag{2.8}$$

Its characteristic equation is:

$$w^K - a_1 w^{K-1} - \cdots - a_{K-1} w - a_K = 0. \tag{2.9}$$

If λ_w is a solution of (2.9) then multiplying both sides of the equation by λ_w^{n-K} ($\neq 0$ since $a_K \neq 0$) shows that λ_w^n is a solution of (2.8). Moreover by linearity, any linear combination of solutions of (2.8) is still a solution, and

[6] Except for the first K coefficients which do not have K predecessors

if (2.9) has K distinct solutions, $\{a_k\}_{k=1,\dots,K}$ is uniquely defined by a set of K independent linear equations.

Hence, for $\sum_{k=1}^{K} c_{k,p} W^{m t_k}$ "solution" of (2.8), $t_k \not\equiv t_l \bmod \boldsymbol{T}_d$ for all $k \neq l$, there exists a unique set $\{a_k\}_{k=1,\dots,K}$ such that $\{W^{t_k}\}_{k=1,\dots,K}$ are the K distinct roots of $w^K - a_1 w^{K-1} - \cdots - a_{K-1} w - a_K$. □

Example 2.b — Joint multipath estimation : the annihilating filter property

Continuing the previous example, Lemma 2.2 implies that the DFT coefficients in the pilot sequence can be written as a linear combination of the 3 previous coefficients. This linear prediction property (2.7) is written as (the DC coefficient is boxed for reference)

$$\left[\begin{array}{ccc}
-0.87-1.16\mathrm{j} & 0.31-0.01\mathrm{j} & 0.40+0.72\mathrm{j} \\
-0.33+0.67\mathrm{j} & -0.87-1.16\mathrm{j} & 0.31-0.01\mathrm{j} \\
\boxed{0.25+0.9\mathrm{j}} & -0.33+0.67\mathrm{j} & -0.87-1.16\mathrm{j} \\
-0.13-1.06\mathrm{j} & \boxed{0.25+0.9\mathrm{j}} & -0.33+0.67\mathrm{j} \\
-0.48+0.15\mathrm{j} & -0.13-1.06\mathrm{j} & \boxed{0.25+0.9\mathrm{j}} \\
-0.17+1.43\mathrm{j} & -0.48+0.15\mathrm{j} & -0.13-1.06\mathrm{j} \\
\hline
-2.95-1.01\mathrm{j} & 2.44-1.6\mathrm{j} & -1.66+2.87\mathrm{j} \\
-3.95+3.27\mathrm{j} & -2.95-1.01\mathrm{j} & 2.44-1.6\mathrm{j} \\
\boxed{3.09+2.11\mathrm{j}} & -3.95+3.27\mathrm{j} & -2.95-1.01\mathrm{j} \\
0.86-0.93\mathrm{j} & \boxed{3.09+2.11\mathrm{j}} & -3.95+3.27\mathrm{j} \\
-3.98+1.03\mathrm{j} & 0.86-0.93\mathrm{j} & \boxed{3.09+2.11\mathrm{j}} \\
2.01+3.78\mathrm{j} & -3.98+1.03\mathrm{j} & 0.86-0.93\mathrm{j}
\end{array}\right]
\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}
=
\left[\begin{array}{c}
-0.33+0.67\mathrm{j} \\
\boxed{0.25+0.9\mathrm{j}} \\
-0.13-1.06\mathrm{j} \\
-0.48+0.15\mathrm{j} \\
-0.17+1.43\mathrm{j} \\
0.56-0.5\mathrm{j} \\
\hline
-3.95+3.27\mathrm{j} \\
\boxed{3.09+2.11\mathrm{j}} \\
0.86-0.93\mathrm{j} \\
-3.98+1.03\mathrm{j} \\
2.01+3.78\mathrm{j} \\
4.21+0.46\mathrm{j}
\end{array}\right].$$

Solving this system yields the annihilating filter coefficients

$$[a_0,\ \dots,\ a_3] = [1,\ -0.24+0.4\mathrm{j},\ 0.24+0.39\mathrm{j},\ -1+0.02\mathrm{j}],$$

The roots of the corresponding polynomial are $[0.96-0.3\mathrm{j},\ -0.23-0.97\mathrm{j},\ -0.49+0.87\mathrm{j}]$. The phases of the roots are $[0.3,\ 1.8,\ 4.2]$, which are the original ToAs multiplied by 3 — the decimation factor.

If AWGN is added to the measurements, neither the *least-square* nor the *total least-square* approach are optimal to solve the linear prediction equation (2.7). To mitigate the effect of the noise, a denoising step is performed on the data first — see Figure 2.5.

In the presence of AWGN, (2.7) is a linear system with both coefficients and constant terms corrupted by white gaussian noise. If the noise realizations were independent from one coefficient to another, a consistant estimate of $\boldsymbol{a}$ would be obtained by solving the system in the *Total Least Square* (TLS) sense. However, the noise realizations in $[\boldsymbol{T}]_{:,1:K}$ are constant along the diagonal.

To remedy this shortcoming a denoising step, called *Cadzow denoising* [40], is used in [31] for a single channel. This denoising method uses a *a lift and project* approach by enforcing the low rank constraint on the Toeplitz data-matrix (lift) followed by restoration of the Toeplitz structure. This two step procedure is iterated until convergence.

The estimation performances are found to be best for a data-matrix of roughly equal dimensions [31], *i.e.* for $L = M + 1$.

Since $M \geq K$, the computational cost of the annihilating filter followed by Cadzow denoising is dominated by the SVD used in the *lift* operation in the denoising steps

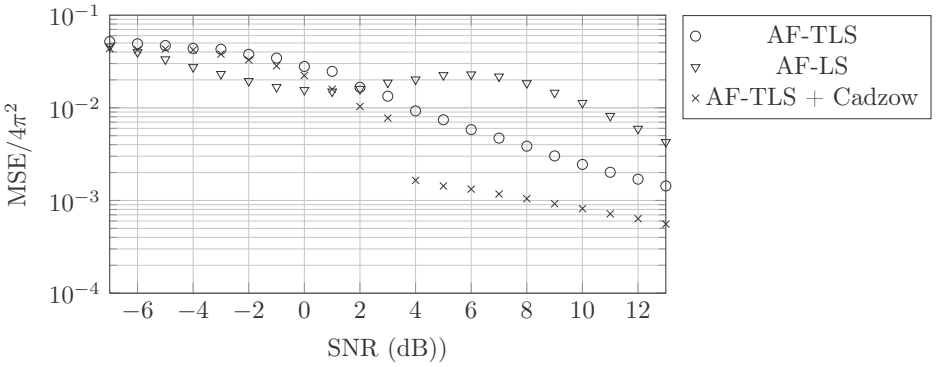


Figure 2.4: *Simulations with $P = 4$ non-fading channels with $K = 5$ paths and 63 DFT pilots ($M = 31$). Denoising the measurements with Cadzow algorithm greatly improves the performances of the annihilating filter (AF) method.*

which requires $\mathcal{O}(PM^3)$ operations[7] as seen in Figure 2.5.

2.2.2 Rotation invariance

The annihilating filter method used the low-rank property of the data matrix to compute the estimate of the time of flights. The Vandermonde decomposition only entered the picture indirectly when linking the roots of the annihilating polynomial to the ToA, *i.e.* after the estimation has taken place. The denoising iterations remedy this shortcoming by taking into account the Toeplitz structure of the data matrix during the estimation.

We now shift our attention to the ESPRIT algorithm [108] which takes into account the peculiar structural properties of $\boldsymbol{T}$ to estimate the time of arrivals.

The ESPRIT algorithm relies on the *rotation invariance* property of the data matrix column space

Proposition 2.2. *Let $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^*$ be a noiseless data-matrix — see Definition 2.1 — of rank K written in the form of its singular value decomposition. The submatrices $\boldsymbol{V}^{\uparrow} = [\boldsymbol{V}]_{1:\mathrm{end}-1,1:K}$ and $\boldsymbol{V}^{\downarrow} = [\boldsymbol{V}]_{2:\mathrm{end},1:K}$ verify*

$$\boldsymbol{V}^{\uparrow} = \boldsymbol{V}^{\downarrow}\boldsymbol{R}\ ,$$

where the eigenvalues of $\boldsymbol{R}$ are $\lambda_k(\boldsymbol{R}) = e^{jD\omega_k}$.

[7] This complexity assumes a "constant" and thus negligible number of iterations for denoising. In comparison solving the annihilating filter in the TLS sense requires $\mathcal{O}(PMK^2)$ operations.

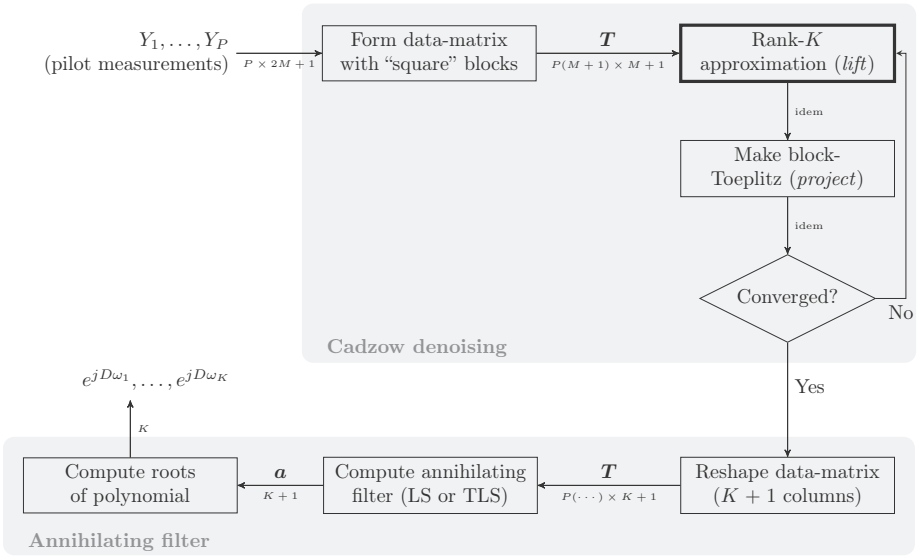


Figure 2.5: *Estimation of the ToAs using the annihilating filter property and Cadzow denoising. The denoising step is an iterative* lift and project *algorithm [40]. The* lift *operation (thick frame) requires $\mathcal{O}(PM^3)$ flops and its computational cost is the most significant.*

Proof.

The rotation invariance property obviously holds on the Vandermonde decomposition of $\boldsymbol{T}$ (see Definition 2.2)

$$\mathcal{V}^{\uparrow} = \mathcal{V}^{\downarrow} \underbrace{\begin{bmatrix} e^{jDw_1} & & & \\ & e^{jD\omega_2} & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & e^{jD\omega_K} \end{bmatrix}}_{\stackrel{\text{def}}{=} \boldsymbol{D}_R}.$$

The matrices $\mathcal{V}$ and $[\boldsymbol{V}]_{:,1:K}$ span the same subspace, the column space of $\boldsymbol{T}$. Therefore there exists an invertible $K \times K$ matrix $\boldsymbol{A}$ such that $\boldsymbol{V} = \mathcal{V}\boldsymbol{A}$. Then

$$\boldsymbol{V}^{\uparrow} = \mathcal{V}^{\uparrow}\boldsymbol{A} = \mathcal{V}^{\downarrow}\boldsymbol{D}_R\boldsymbol{A} = \underbrace{\mathcal{V}^{\downarrow}\boldsymbol{A}}_{\boldsymbol{V}^{\downarrow}} \underbrace{\boldsymbol{A}^{-1}\boldsymbol{D}_R\boldsymbol{A}}_{\stackrel{\text{def}}{=} \boldsymbol{R}}.$$

□

This establishes that rotation invariance can be used for joint ToA estimation on common support channels, and the procedure is summarized in Figure 2.6.

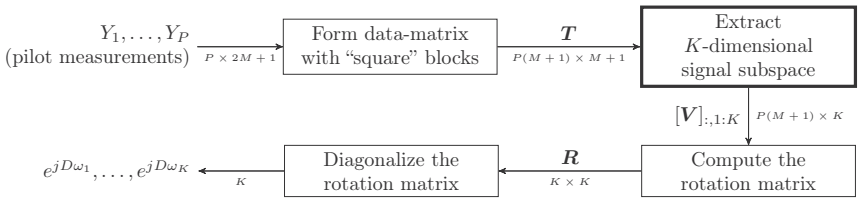


Figure 2.6: *Estimation of the ToAs using the rotation invariance property of the block-Toeplitz data-matrix. The computational complexity is dominated by the estimation of the signal subspace (thick frame) requiring $\mathcal{O}(PM^3)$ flops.*

Example 2.c — Joint multipath estimation : the rotation invariance property

Continuing the previous examples, we obtain an orthonormal basis for the 3-dimensional signal space of $\boldsymbol{T}$ by taking the 3 principal singular-vectors of its column-space :

$$\boldsymbol{V} = \begin{bmatrix} 0.4 & -0.39 & 0.43 \\ -0.06-0.48\mathrm{j} & -0.42-0.024\mathrm{j} & -0.23+0.33\mathrm{j} \\ -0.4+0.14\mathrm{j} & -0.46+0.27\mathrm{j} & -0.31-0.43\mathrm{j} \\ 0.24+0.33\mathrm{j} & -0.3+0.43\mathrm{j} & 0.37+0.02\mathrm{j} \\ 0.27-0.37\mathrm{j} & -0.10+0.31\mathrm{j} & -0.22+0.42\mathrm{j} \end{bmatrix}$$

Because the data-matrix in (2.5) is exactly rank 3 (no noise), the columns of $\boldsymbol{T}$ belong to this subspace; in the presence of noise, this step orthogonally projects the columns of the data-matrix on a subspace of dimension 3, fitting the data to the low-rank model.

Then, the two minors $\boldsymbol{V}^{\uparrow}$ (solid frame) and $\boldsymbol{V}^{\downarrow}$ (dashed frame) verify the rotation invariance property stated in Proposition 2.2. The matrix $\boldsymbol{R}$ solution of $\boldsymbol{V}^{\uparrow} = \boldsymbol{V}^{\downarrow}\boldsymbol{R}$ is

$$\boldsymbol{R} = \begin{bmatrix} -0.20-0.98\mathrm{j} & -0.22-0.2\mathrm{j} & 0.13-0.07\mathrm{j} \\ -0.10+0.059\mathrm{j} & 0.91-0.29\mathrm{j} & 0.22+0.03\mathrm{j} \\ -0.023-0.026\mathrm{j} & 0.094-0.11\mathrm{j} & -0.47+0.87\mathrm{j} \end{bmatrix}.$$

This matrix is not diagonal because the column vectors of $\boldsymbol{V}$ are not the 3 original phasors but a linear combination of them. The diagonal rotation matrix is obtained by undoing the similarity which transformed it into $\boldsymbol{R}$, *i.e.* by computing the eigenvalue decomposition of $\boldsymbol{R}$. The diagonal rotation elements are the eigenvalues of $\boldsymbol{R}$

$$\boldsymbol{z} = \begin{bmatrix} -0.227-0.974\mathrm{j} \\ 0.955-0.296\mathrm{j} \\ -0.49+0.872\mathrm{j} \end{bmatrix}, \text{ so } \quad \angle\boldsymbol{z} = \begin{bmatrix} 1.8 \\ 0.3 \\ 4.2 \end{bmatrix} = 3 \cdot \begin{bmatrix} 0.6 \\ 0.1 \\ 1.4 \end{bmatrix}.$$

2.2.3 Putting it all together

The paths amplitudes are estimated independently for each antenna, by solving a $(2M+1) \times K$ inhomogeneous linear system of equations, e.g. see [31; 15].

Numerical comparisons

Before any further investigations, we can take a step back and compare on synthetic data how the proposed algorithms fare compared one to another, and so retain only

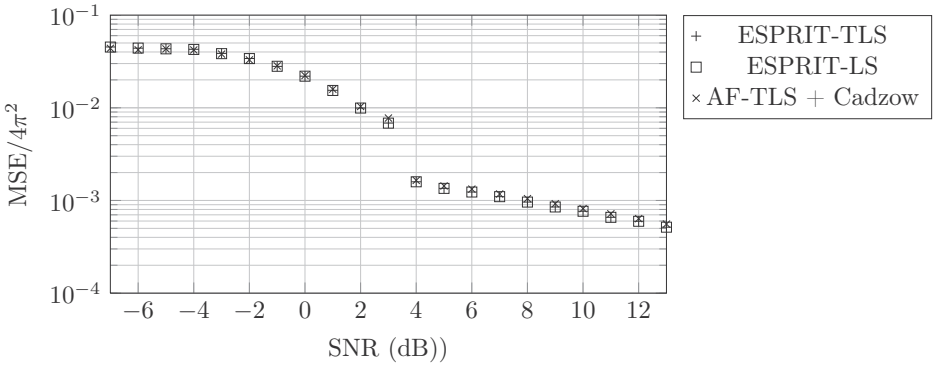


Figure 2.7: *Simulations with $P = 4$ non-fading channels with $K = 5$ paths and 63 DFT pilots ($M = 31$). The performances of the ESPRIT algorithm match those of the annihilating filter method combined with Cadzow denoising.*

the promising ones. Tests on physically measured channel impulse responses are deferred to the end of Chapter 3.

Non-fading channels We start with a crude SCS model having deterministic path amplitudes (non-fading) — the parameters are given in Figure 2.7.

As a first observation, the two serious contenders are ESPRIT based methods and the annihilating filter method with Cadzow denoising. As seen in Figure 2.4, the annihilating filter method alone is insufficient, and this observation made on joint estimation corroborates the results made on single channels [31].

Further simulations on fading channels More comprehensive simulations on Rayleigh fading channels are presented in Appendix B, together with lowerbounds on the parameters estimation to assess the performances.

Computational cost

Since $M \geq K$, the computational cost of the annihilating filter followed by Cadzow denoising is dominated by the SVD used in the *lift* operation in the denoising steps which requires $\mathcal{O}(PM^3)$. operations[8]. Like Cadzow denoising algorithm, the computational cost of ESPRIT algorithm is also dominated on a SVD of size $P(M+1) \times (M+1)$ used for the signal subspace identification, resulting in a cost of $\mathcal{O}(PM^3)$. However a single SVD of this size is required by the ESPRIT algorithm. Costs are summarized in Table 2.1.

The ESPRIT based estimation and the annihilating filter with denoising have the same computational cost order, however the non-iterative nature of ESPRIT makes it more attractive.

[8]This complexity assumes a "constant" and negligible number of iterations for denoising. In comparison solving the annihilating filter in the TLS sense requires $\mathcal{O}(PMK^2)$ operations.

2.3 Fast, in place estimation

The amount of computations and memory consumed by the proposed parametric estimation algorithms is problematic for M large — especially for an embedded use in communication devices. As an example, the 3GPP-LTE standard [54] uses between 72 and 1200 resource elements[9] depending on the bandwidth mode. For higher bandwidth modes, it renders the proposed algorithms fairly incompatible with the real-time and embedded requirements of mobile communications.

For a signal sampled at a rate far above the rate of innovation [31] — *i.e.* $K \ll M$ — the extraction of the K dimensional signal subspace from the $M+1$ dimensional column space of $\boldsymbol{T}$ is particularly inefficient for two reasons :

1. A total of $M+1$ singular vectors are computed, and only the K principal are used.
2. The data matrix is well structured and it can be represented in memory by $P \cdot (2M+1)$ complex numbers, and in some cases, affords faster algebraic manipulations. Algorithms used to compute the SVD may not exploit this structure and may also destroy it right from the beginning (e.g. Householder reflections, and Givens rotations).

In this section, we outline a numerical procedure able to identify the signal subspace in $\mathcal{O}(PKM\log(M))$ flops and operating in $\mathcal{O}(PKM)$ memory, greatly reducing the computational complexity of ESPRIT based algorithm. It is based on projection onto a Krylov subspace using Lánczos' algorithm. In general, identification of a signal subspace with Lánczos' algorithm is not new and can be found in [141] for example where it is used on low-rank covariance matrices.

In our setup, the additional structure of the data matrix allows to reduce the complexity not only to $\mathcal{O}(PKM^2)$ but further down to $\mathcal{O}(PKM\log(M))$. A necessary result to achieve this complexity order is to extend the convergence result of [142] such as to apply when the spectral measure of a noise-only and properly scaled data-matrix is unbounded as M grows. This result will be presented in Theorem 2.3.

2.3.1 Krylov subspace projection and Lánczos' algorithm

A pedestrian introduction to Krylov subspaces The Krylov subspace of dimension L of an hermitian symmetric matrix $\boldsymbol{H}$ of dimension $M+1$ is the subspace spanned by vectors obtained by applying $L-1$ times the *power method* to an initial vector $\boldsymbol{f}_0$. The power method computes at iteration ℓ

$$\boldsymbol{f}_\ell \;=\; \boldsymbol{H}^\ell \boldsymbol{f}_0 \;,$$

For $L-1$ iterations the associated *Krylov subspace* of dimension L is

$$\mathcal{K}_L(\boldsymbol{H};\boldsymbol{f}_0) \;=\; \mathrm{span}\{\boldsymbol{f}_\ell\}_{\ell=0,\ldots,L-1}.$$

It is intuitive that with the adequate normalization, $\boldsymbol{f}_\ell$ converges to $\boldsymbol{\xi}_1$ the principal eigenvector of $\boldsymbol{H}$ provided that $\langle \boldsymbol{\xi}_1 \,,\, \boldsymbol{f}_0\rangle \neq 0$, to show it

[9] roughly the number of pilots

$$\boldsymbol{f}_\ell \;=\; \boldsymbol{\Xi}\boldsymbol{\Lambda}^\ell\boldsymbol{\Xi}^*\boldsymbol{f}_0 \;=\; \sum_m \lambda_m^\ell \,\langle \boldsymbol{\xi}_m \,,\, \boldsymbol{f}_0\rangle\, \boldsymbol{\xi}_m,$$

where the first term in the summation will outweigh the other terms as $\ell \to \infty$.

Definition 2.3. *The eigenvalues (resp. eigenvectors) of the orthogonal projection of an hermitian matrix $\boldsymbol{H}$ onto the Krylov subspace $\mathcal{K}_L(\boldsymbol{H}, \boldsymbol{f}_0)$ are called the* Ritz values *(resp.* Ritz vectors*) of $\boldsymbol{H}$ in $\mathcal{K}_L(\boldsymbol{H}, \boldsymbol{f}_0)$.*

Witnessing the convergence of the power method to the principal eigenvector, one may follow the intuition that a Krylov subspace tends to "align" with the subspace spanned by the principal eigenvectors of the original matrix $\boldsymbol{H}$, implying the Ritz vectors and Ritz values would provide an approximation of the principal eigenpairs of $\boldsymbol{H}$.

The formalization of this intuition is called *Rayleigh-Ritz* theory, a good introduction is found in [97].

The quantification of the error made by approximating the principal eigenvectors and eigenvalues of the hermitian matrix $\boldsymbol{T}^*\boldsymbol{T}$ with Ritz vectors/values will be treated in Section 2.3.3.

Orthogonal projection onto a Krylov subspace: Lánczos algorithm The most straightforward way to obtain the eigenbasis of $\mathcal{K}_L(\boldsymbol{H}; \boldsymbol{f}_0)$ numerically, is to orthogonalize the set of vectors $\{\boldsymbol{f}_\ell\}_{\ell=0,\dots,L-1}$. Done naively (Gram-Schmidt process, Householder reflections, ...), the cost of this operation is $\mathcal{O}(L^2M)$.

A more efficient orthogonalization scheme is obtained by using the fact that vectors in a Krylov subspace are in bijection with polynomials in $\boldsymbol{H}$. A vector $\boldsymbol{a} \in \mathcal{K}_L(\boldsymbol{H}; \boldsymbol{f}_0)$ is expanded as

$$\boldsymbol{a} \;=\; \underbrace{\left[\sum_{\ell=0}^{L-1} \alpha_\ell \cdot \boldsymbol{H}^\ell\right]}_{\stackrel{\text{def}}{=} p_{\boldsymbol{a}}(\boldsymbol{H})} \boldsymbol{f}_0 \;,$$

and so, orthogonality of $\boldsymbol{a}$ and $\boldsymbol{a}'$ in a Krylov subspace for any initial vector $\boldsymbol{f}_0$ amounts to orthogonality of polynomials

$$\langle \boldsymbol{a} \,,\, \boldsymbol{a}'\rangle = \boldsymbol{0} \quad \Leftarrow \quad \langle p_{\boldsymbol{a}}(\boldsymbol{H}) \,,\, p_{\boldsymbol{a}'}(\boldsymbol{H})\rangle \;=\; \boldsymbol{0}.$$

A sequence of orthogonal polynomials of increasing degree has the peculiar property to verify a three terms recursion

Theorem 2.1. ***(Szegö [123])***
Let $p_0(t),\ p_1(t),\ \dots$ be a sequence of orthogonal polynomials of increasing order, i.e. p_i is of degree i and $\langle p_i \,,\, p_j\rangle \neq 0$ iff $i = j$. Then for all $i > 0$

$$p_{i+1}(t) \;=\; (a_i t + b_i)p_i(t) + c_i p_{i-1}(t)\ ,$$

where $b_i = \frac{\alpha_{i+1}}{\alpha_i}$, $a_i = b_i\left(\frac{\beta_{i+1}}{\alpha_{i+1}} - \frac{\beta_i}{\alpha_i}\right)$, $c_i = \frac{\alpha_{i+1}\alpha_{i-1}\|p_i\|^2}{\alpha_i^2\|p_{i-1}^2\|}$
and α_i, β_i *are the leading coefficients of* p_i, *i.e.* $p_i(t) = \alpha_i t_i + \beta_i t_{i-1} + \cdots$.

Proof.

See [123]. □

An important implication of Theorem 2.1 is that an orthonormal basis $\{\boldsymbol{q}_0,\ \ldots,\ \boldsymbol{q}_{L-1}\}$ spanning $\mathcal{K}_L(\boldsymbol{H};\boldsymbol{f}_0)$ can be obtained by computing recursively

$$\begin{aligned}\boldsymbol{q}_{i+1} &\stackrel{\text{def}}{=} p_{i+1}(\boldsymbol{H})\boldsymbol{f}_0,\\ &= (a_i\boldsymbol{H} + b_i)p_i(\boldsymbol{H})\boldsymbol{f}_0 + c_i p_{i-1}(\boldsymbol{H})\boldsymbol{f}_0,\\ &= (a_i\boldsymbol{H} + b_i)\boldsymbol{q}_i + c_i\boldsymbol{q}_{i-1}\ ,\end{aligned}$$

starting with $\boldsymbol{q}_0 = \boldsymbol{f}_0/\left\|\boldsymbol{f}_0\right\|$ and $\boldsymbol{q}_1 = \frac{\boldsymbol{H}\boldsymbol{q}_0 - \langle \boldsymbol{H}\boldsymbol{q}_0\ ,\ \boldsymbol{q}_0\rangle \boldsymbol{q}_0}{\|\cdots\|}$.

This recursive procedure, which orthogonalizes each basis vector against the two previous ones is *Lánczos algorithm* [74]. Another implication of this three terms recursion is that the orthogonal projection of $\boldsymbol{H}$ onto $\mathcal{K}_L(\boldsymbol{H};\boldsymbol{f}_0)$ can be written as a tridiagonal similarity [97]

$$\mathrm{proj}_{\mathcal{K}_L(\boldsymbol{H};\boldsymbol{f}_0)} \;=\; \boldsymbol{Q}^*\begin{bmatrix}\alpha_0 & \beta_1 & & \mathbf{0} & \\ \beta_1 & \alpha_1 & \ddots & & \\ & \ddots & \ddots & \ddots & \beta_{L-1}\\ & \mathbf{0} & & \beta_{L-1} & \alpha_{L-1}\end{bmatrix}\boldsymbol{Q},$$

where $\boldsymbol{Q} = [\boldsymbol{q}_0\ \cdots\ \boldsymbol{q}_{L-1}]$.

Therefore, computing the Ritz pairs amounts to evaluating the eigenvalue decomposition of a symmetric tridiagonal matrix of dimension L. This eigenvalue problem has an $\mathcal{O}(L\log^2 L)$ or $\mathcal{O}(L^2)$ solution depending on whether the eigenvectors are to be computed or not [62]; so that for $L \ll M$, the cost of the Lanczos iteration itself dominates the cost of the whole procedure.

In practice, Lánczos' algorithm suffers from unavoidable numerical instabilities, which mitigation has been thoroughly studied [97; 32; 117]. A key observation is that there exists stable algorithms [76; 12] computing an orthogonal basis for $\mathcal{K}_L(\boldsymbol{H};\boldsymbol{f}_0)$ in $\mathcal{O}(L\times\mathsf{mvm}(M))$ where $\mathsf{mvm}(M)$ is the cost of a matrix vector multiplication of size M.

2.3.2 Fast Lánczos iterations for SCS estimation

The data matrix $\boldsymbol{T}$ is not hermitian[10], nevertheless the ESPRIT algorithm only requires its right singular vectors $\boldsymbol{V}$. The mapping

$$\boldsymbol{T} \mapsto \boldsymbol{T}^*\boldsymbol{T} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^*,$$

transforms the data matrix into an hermitian symmetric matrix, which we call the *data autocorrelation matrix*. Its eigenvectors are $\boldsymbol{V}$ and eigenvalues are $\lambda_k = \sigma_k^2$, where σ_k are the singular values of $\boldsymbol{T}$. A downside of this mapping is to square the condition number, however since our interest is limited to the upper-end of the spectrum, this is not a critical issue.

The data autocorrelation matrix is simply the sum of the block's autocorrelation matrices

$$\boldsymbol{T}^*\boldsymbol{T} \;=\; \sum_{p=1}^{P} \boldsymbol{T}_p^*\boldsymbol{T}_p \;,$$

therefore a matrix-vector multiplication can be done in parallel for each block.

For square Toeplitz matrices of dimension $M+1$, matrix vector multiplications can be done in $\mathcal{O}(M \log M)$ using the FFT

Proposition 2.3. *Let $\boldsymbol{T}_p$ be a square Toeplitz matrix of dimension $M+1$ with a first-row $\boldsymbol{t}_r^T$ and first column $\boldsymbol{t}_c$.*

The matrix-vector product $\boldsymbol{T}_p\boldsymbol{f}$ is computed in $\mathcal{O}(M \log M)$ with $\mathcal{O}(M)$ memory as

$$\boldsymbol{T}_p\boldsymbol{f} \;=\; \left[\mathrm{IDFT}\left\{\boldsymbol{g}_p \cdot \mathrm{DFT}\left\{[\boldsymbol{f}^T,\; 0,\; \cdots]\right\}\right\}\right]_{1:M+1},$$

where

$$\boldsymbol{g}_p \;=\; \mathrm{DFT}\left\{\begin{bmatrix}\boldsymbol{t}_c^T & 0 & t_{r,M+1} & t_{r,M} & \cdots & t_{r,2}\end{bmatrix}^T\right\}.$$

Proof.

This is a well-known result — see [14] — which relies on embedding $\boldsymbol{T}_p$ into a circulant matrix of dimension $2(M+1)$. □

[10] An algorithm similar to Lánczos algorithm exists for non hermitian symmetric matrices — *Arnoldi's method*, though, its stability is in general poor.

Table 2.1: *"$\mathcal{O}$" complexity [134] for subspace identification.*

Algorithm	Main computation	Storage	Latency	Processing units
Krylov	$KPM \log M$	KM	KM [144]	$P \times$ FFT engines $(2(M+1)$ points)
Full SVD (serial)	PM^3	PM^2	PM^3	1 multipurpose processor
Full SVD (systolic array) [36; 43]	PM^3	PM^2	$M(\log M + P)$	$M^2\times$ 2-by-2 SVD pu.

The full SVD is done with Jacobi rotations and can be massively parallelized using the systolic array method of Brent *et al.* [36]. Parallelism greatly reduces the latency of the system, but since it does not reduce the number of computations it comes at the cost of using multiple processing units.

Corollary 2.1. *Projection of $\boldsymbol{T}^*\boldsymbol{T}$ onto an L dimensional Krylov subspace requires L matrix vector multiplications which can be computed with $P(4L+1)+1$ FFT of length $2(M+1)$.*

Proof.

This complexity is obtained by precomputing the P FFTs for the generators of the data matrix and of the FFT for the initial vector. Then each of the L Lanczos iterations requires one matrix-vector multiplication (mvm) with the data autocorrelation matrix, *i.e.* two toeplitz mvm, which amounts to four FFTs. □

Assuming the K principal eigenpairs of $\boldsymbol{T}$ match the K principal Ritz pairs of $\mathcal{K}_L(\boldsymbol{H}; \boldsymbol{f}_0)$ for $L \sim \mathcal{O}(K)$ the signal subspace can be estimated in $\mathcal{O}(KPM \log(M))$ using only $\mathcal{O}(KM)$ memory to process the input measurements.

Table 2.1 summarizes the computational and memory consumption of the ESPRIT based estimation using the proposed acceleration compared to a full SVD (done serially or in parallel with additional hardware).

2.3.3 How large must the Krylov subspace be?

The key assumption motivating the Krylov subspace approach is that the signal subspace can be accurately estimated from a subspace of dimension $L \ll M+1$, rather than from a full SVD of size $M+1$. This argument relies implicitly on the convergence of the K principal Ritz pairs to the K principal eigenpairs of $\boldsymbol{T}^*\boldsymbol{T}$.

Several bounds on the convergence rate of the Ritz pairs to the eigenpairs exist in the literature [71; 109; 140]. They bound the distance between the Ritz values and the eigenvalues and the angle between the Ritz vectors and the eigenvectors, which is the quantity of interest in our case.

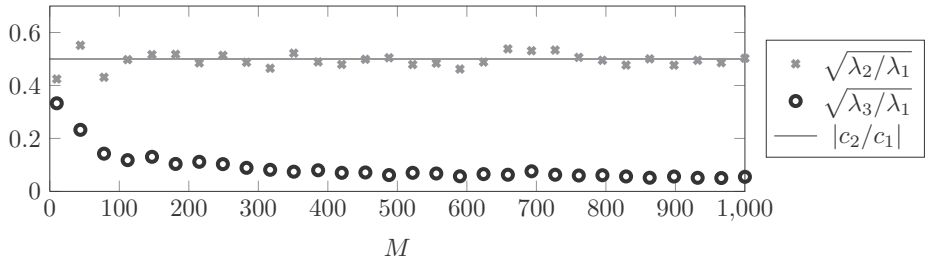


Figure 2.8: *Normalized magnitudes of the K^{th} and $(K+1)^{th}$ singular values of a data-matrix of size $M+1$ (they are the square roots of the eigenvalues of the hermitian product of the data-matrix with itself). The signal used to form the data-matrix contains two paths ($K=2$) with amplitudes $c_1=1$ and $c_2=0.5$. The noise is iid normally distributed with a variance of 2. It illustrates the gap forming between λ_K and λ_{K+1} as M grows.*

Nevertheless, it is noted in [140] that the general bounds found in [71; 109] are not tight and hard to use in practice. The bound of Xu [140] remedy this limitations using two additional assumptions :

$$\exists \varepsilon \ : \quad \lambda_k - \frac{\lambda_{K+1}+\lambda_{M+1}}{2} > \varepsilon \quad \text{and} \quad \lambda_{K+1} - \lambda_{M+1} < \varepsilon. \tag{2.10}$$

To see if this two conditions are met, a *separation theorem* is needed.

Theorem 2.2. *Consider channel measurements as in (2.4)*

$$Y_p[m] \ = \sum_{k=1}^{K} c_{k,p} e^{-jD\omega_k m} + E[m] \ , \qquad |m| \leq M.$$

where K is the number of paths, and $E[m]$ are iid normally distributed with a fixed and finite variance.

Define $\boldsymbol{T}$ the block Toeplitz data-matrix with square blocks of dimension $M+1$ as in Definition 2.1.

Then, for M large enough, the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{M+1} \geq 0$ of $\boldsymbol{T}^*\boldsymbol{T}$ verify

$$\frac{\lambda_m}{\lambda_1} \sim \begin{cases} \mathcal{O}(1) & , \ m \leq K \ , \\ \mathcal{O}\left(\frac{\log M}{\sqrt{M}}\right) & , \ else. \end{cases}$$

Proof.

See Appendix B.2. □

With this separation theorem, the error estimate in [140] yields

Theorem 2.3. *In the setup of Theorem 2.2, consider the Ritz values and Ritz vectors obtained from a projection onto a Krylov subspace of dimension $L > K$.*

The approximation of the K principal eigenvalues/vectors of $\boldsymbol{T}^*\boldsymbol{T}$ by the principal Ritz values/vectors has an error of order

$$\mathcal{O}\left(\left(\frac{\log M}{\sqrt{M}}\right)^{2(L-K)}\right).$$

The definition of the error for the eigenvalues is a normalized difference with the corresponding Ritz values

$$\frac{\lambda_k - \widehat{\lambda}_k}{\lambda_k - \frac{\lambda_K - \lambda_{\min}}{2}},$$

where $\widehat{\lambda}_k$ is the k^{th} Ritz-value and $\lambda_{\min}$ is the least eigenvalue.

For the eigenvectors the sine squared of the principal angle with the corresponding Ritz vectors

$$\sin^2 \angle(\boldsymbol{\xi}_k, \widehat{\boldsymbol{\xi}}_k).$$

Proof.

This is a direct application of Theorem 2.2 to Theorem 3.2 in [140]. □

Theorem 2.3 indicates an acceptable approximation error is obtained in $\mathcal{O}(K)$ for $M \gg K$, and numerical simulation further support this assumption.

2.3.4 Numerical tests

We apply Lanczos algorithm to estimate the signal subspace in the ESPRIT algorithm. Figure 2.9 shows that the accelerated implementation is competitive for $M > 60$. The input has only one channel $P = 1$, and for $P > 1$ an implementation having P FFT engines running in parallel should have essentially the same runtime.

The accuracy of the accelerated implementation is the same as the plain one (the Ritz approximation error is negligible).

2.4 Conclusion

In this chapter, we have proposed an algorithm for the joint estimation of multipath channels based on a classical line spectra estimation technique (ESPRIT).

The computational and memory requirements of a straightforward implementation of this algorithm showed two issues. First, the computational load is independent of

[12] The exact value of M matters for the efficiency of the FFT computation; we report in this plot sizes for which FFTW3 found a good optimization scheme.

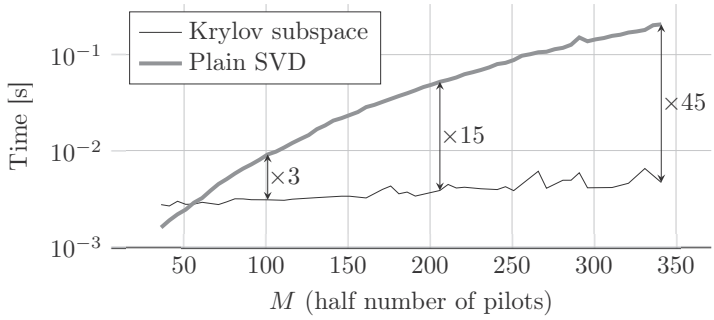


Figure 2.9: *Median runtime of ESPRIT-TLS ($K = 5$, $P = 1$) on a single channel for a given number of pilots $2M + 1$. The test is coded in Python and uses the LAPACK library to compute SVDs and the ARPACK library for the Lanczos iterations. The fast matrix-vector multiplication uses the FFTW3 library*[12].

the sparsity level $K/(2M+1)$ and grows cubically with the number of measurements. Second, the memory footprint was quadratic in the number of measurements.

To adress these issues, we proposed a method based on projection onto Krylov subspaces, which yields a practical solution requiring an amount of memory proportional to the number of measurements and relies on the FFT for the most demanding operations. This improvement is made possible by the fact that the data-matrix, on which the algorithm relies, is never explicitely constructed thanks to its Toeplitz structure. The theoretical results are confirmed by numerical tests.

To meaningfully measure the estimation accuracy gained over a classical non-parametric method, tests on synthetically generated data are insufficient. At the end of Chapter 3, we will undertake such tests. To do so, we will first need to include *signal detection* in the picture (Chapter 3) — i.e. add the estimation of K to the problem.

Chapter 3

Model detection for sparse channels

In Chapter 2 we treated *estimation* problems, *i.e.* the task consisting in fitting a model with a fixed set of unknowns to a set of measurements. We assumed that the sensed channels were realizations of multipaths channels with a known and static number of paths and a common support. Starting from this a priori known state, we proposed solutions to estimate the parameters of the model. Another issue is to obtain this initial state — the identification of a precise model — from what can be reasonably inferred from the general characteristics of the problem.

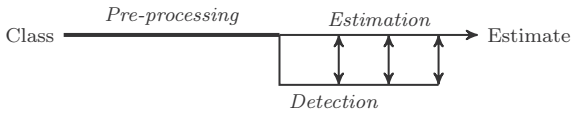
Usage of a single fixed model is too narrow for practical applications. E.g., in multipath channel models, the number of paths is not known a priori and may vary over time. Also we have seen in Chapter 1 that the multipath model may not always be relevant as the classical bandlimited model is preferable in cluttered environments.

Therefore, the starting point of an estimation problem is not a single parametric model but a collection of them, a *class of models*. The selection of the suitable one is called *detection* [132]. Detection differs from estimation in the sense that it alters the number or the nature of the estimated parameters. The simplest way to look at the problem is sequential

However, estimation cannot be decoupled with detection as the fitness of the model is revealed by the estimation. A more elaborate approach is to successively refine the model selection using a *feedback loop*[1].

The issue with an iterative approach is its obvious computational cost and the delay it may introduce. Ideally, detection and estimation should go side-by-side as much as it is possible, e.g.

[1] E.g., "optimum feedback systems" [132].



This is the kind of processing flow we target; and so, the continuity with Chapter 2 is necessary as detection and estimation are now intertwined.

In this chapter, we study how the algorithms outlined in Chapter 2 can be modified to include detection at a minimal cost. Rather than a sequential approach, we will try to achieve a joint detection/estimation.

After a review of the main techniques, we propose two criterion. The first one is based on an hypothesis which can be tested during the fast signal subspace identification developed in Chapter 2. If need be, the validity of the detection can be verified after estimation has taken place, in order to increase the model order if necessary.

The second criterion is a purely geometric one, based on the convexification of the rank of a matrix called the *effective rank* [107].

3.1 A review of signal detection for subspace methods

An exhaustive listing of model order selection techniques for subspace identification algorithms such as ESPRIT would be an overwhelming task based on their sheer proliferation[2] [4; 105; 141; 142; 137; 145; 148].

We will only review methods based on the *likelihood* of the model. There exists other methods such as covariance matching estimation techniques (COMET) which we will not address, see [93] for an in depth review and applications. The selection criteria based on the likelihood can be subdivided with respect to their mode of selection. Namely, they use either *thresholds* or *penalty* functions.

The likelihood of a model Given a set of parameters $\Theta^{(K)}$ where K is the model order, let $\widehat{\Theta}^{(k)}$ be an estimation based on an estimated model order k and a set of observations $\boldsymbol{Y} \in \mathbb{C}^{N\times P}$ distributed according to a probability density f. Each of the P columns of $\boldsymbol{Y}$ is usually called a *snapshot* in the literature [137], and the number of rows N is the number of samples per snapshot[3].

The *likelihood function* – or simply *likelihood* – is the conditional probability law $f(\boldsymbol{Y}|\widehat{\Theta}^{(k)})$, *i.e.* the probability of the observation given the estimated parameter values. For distributions of the exponential family, it is often easier to work with the natural logarithm of the likelihood function

$$\mathcal{L}(\widehat{\Theta}^{(k)}) \;=\; \log f(\boldsymbol{Y}|\widehat{\Theta}^{(k)}),$$

[2]Proposing yet another detection criterion feels in itself shameful; I can only invoke the specificity of the problem to make it somehow acceptable.

[3]In the sparse common support model of Chapter 2, P would be the number of antennas, N the number of pilots per frame and K the number of paths — note that K is not necessarily the number of degrees of freedom in the model, but it provides an index for it.

which has the same extrema as f by concavity of the logarithm.

The *maximum likelihood estimate* of Θ is defined as

$$\widehat{\Theta}_{\mathrm{ML}}^{(k)} \;=\; \arg\max_{\widehat{\Theta}^{(k)}} \mathcal{L}(\widehat{\Theta}^{(k)}),$$

Finding the maximum-likelihood estimate or its approximation is the task of estimation and choosing k is the task of detection.

For detection, two options are to be considered

1. **A priori detection** : Estimation of $\mathcal{L}(\widehat{\Theta}_{\mathrm{ML}}^{(k)})$ without explicit computation of $\widehat{\Theta}_{\mathrm{ML}}^{(k)}$, which can be seen as a decoupling of estimation and detection. The feasibility of this approach depends on the problem.

2. **A posteriori detection** : Estimation of $\mathcal{L}(\widehat{\Theta}_{\mathrm{ML}}^{(k)})$ after explicit computation of $\widehat{\Theta}_{\mathrm{ML}}^{(k)}$ or a close approximation of it. This approach necessitates the estimation of the parameters for different model orders, which can be costly.

Not much needs to be said about *a posteriori detection*. Once the maximum-likelihood estimate is known or approximated, one simply needs to plug these values into the probability density of the signal model

A priori detection For a priori detection, we review the pioneering work of Wax and Kailath [137]. Consider the measurement model

$$Y_p[n] \;=\; \sum_{\ell=1}^{K} c_{\ell,p} X_{\theta_\ell}[n] + E_p[n],$$

where $E_p[n]$ are iid zero mean complex-valued gaussian random variables with variance σ^2.

Using matrix notation,

$$\boldsymbol{Y} \;=\; \boldsymbol{X}\boldsymbol{C} + \boldsymbol{E},$$

where $\boldsymbol{Y}$ and $\boldsymbol{E}$ are $N \times P$ matrices and $\boldsymbol{X}$, $\boldsymbol{C}$ have size $N \times K$ and $K \times P$ respectively

$$[\boldsymbol{Y}]_{n,p} = Y_p[n]\ , \quad [\boldsymbol{X}]_{n,\ell} = X_{\theta_\ell}[n]\ , \quad [\boldsymbol{C}]_{\ell,p} = c_{\ell,p}\ , \quad [\boldsymbol{E}]_{n,p} = E_p[n].$$

Since the noise covariance matrix is a multiple of the identity, the covariance matrix of the measurements is given by [137]

$$\boldsymbol{R} \;=\; \mathbb{E}\,[\boldsymbol{Y}\boldsymbol{Y}^*] \;=\; \underbrace{\boldsymbol{X}\mathbb{E}\,[\boldsymbol{C}\boldsymbol{C}^*]\,\boldsymbol{X}^*}_{\stackrel{\mathrm{def}}{=}\,\boldsymbol{\Phi}} + \sigma^2\mathbb{I}.$$

The noiseless covariance matrix $\boldsymbol{\Phi}$ is symmetric and has rank-K, we call $\{\boldsymbol{\varphi}_\ell\}_{\ell=1,\ldots,K}$ its K principal eigenvectors.

Under the hypothesis that the signal has only k ($\leq K$) components[4], the estimated covariance matrix $\boldsymbol{R}$ is truncated

[4]Evidently, the hypothesis $k = K$ yields the measurements covariance matrix.

$$\boldsymbol{R}^{(k)} \;=\; \sum_{\ell=1}^{k}(\lambda_\ell - \sigma^2)\boldsymbol{\varphi}_\ell\boldsymbol{\varphi}_\ell^* + \sigma^2\mathbb{I}, \tag{3.1}$$

by identification, $(\lambda_\ell, \boldsymbol{\varphi}_\ell)_{\ell\leq k}$ are the K principal eigenpairs of $\boldsymbol{R}^{(k)}$. Because this parametrization is in bijection with the original one, $\Theta^{(k)}$ identifies with $\boldsymbol{R}^{(k)}$.

The noise $\boldsymbol{E}$ is additive white and gaussian, so the computation of the log-likelihood function[5] is an easy task [132]

$$\mathcal{L}(\widehat{\Theta}^{(k)}) \;=\; -P\cdot\ln\det\widehat{\boldsymbol{R}}^{(k)} - \mathrm{Tr}\left\{[\widehat{\boldsymbol{R}}^{(k)}]^{-1}\widehat{\boldsymbol{R}}\right\},$$

where $\widehat{\boldsymbol{R}} \;=\; \frac{1}{P}\boldsymbol{Y}\boldsymbol{Y}^*$ is the *sample covariance matrix* and $\widehat{\boldsymbol{R}}^{(k)}$ is the maximum likelihood estimate of $\boldsymbol{R}^{(k)}$.

This estimate can be built from (3.1) taking $(\widehat{\lambda}_\ell, \widehat{\boldsymbol{\varphi}}_\ell)$ as the eigenpairs of the sample covariance matrix $\widehat{\boldsymbol{R}}$; and

$$\widehat{\sigma}^2 \;=\; \frac{1}{\min(N,P)-k}\sum_{\ell=k+1}^{\min(N,P)}\widehat{\lambda}_\ell,$$

as the estimation of the noise power.

This yields the log-likelihood formula found in [137]

$$\mathcal{L}(\widehat{\Theta}^{(k)}) \;=\; \log\left(\frac{\prod_{\ell=k+1}^{P}\widehat{\lambda}_\ell^{1/(P-k)}}{\frac{1}{P-k}\sum_{\ell=K+1}^{P}\widehat{\lambda}_\ell}\right)^{(P-k)N},$$

which can be evaluated directly from the spectrum of the sample covariance matrix.

Penalization schemes In a seminal paper [4], Akaike showed that maximization of the log-likelihood function leads to a model order estimate with a positive bias.

To overcome this bias, he proposed the addition of a penalty term, such as to minimize the expected Kullback-Liebler divergence between the selected model and the true model. Under certain regularity conditions – not discussed here, see [6] – it is shown that the penalty is equal to the number of degrees of freedom in the model

$$\text{Akaike Information Criterion}: \quad \mathrm{AIC}(k) \;=\; -\mathcal{L}(\widehat{\Theta}^{(k)}) + \underbrace{\text{Degrees of freedom}}_{\text{penalty}}.$$

The model order estimate is then

$$\widehat{K} \;=\; \arg\min_k \mathrm{AIC}(k).$$

Many other penalty functions have been proposed, the most popular ones are Rissanen's MDL criterion [105] and the modified AIC [5].

[5] Terms independent of k are dropped.

Thresholding schemes Bartlett and Lawley proposed a statistical test [28; 75] on the likelihood of the following hypotheses

$$\mathrm{H}_k \ : \quad \lambda_{k+1} = \lambda_{k+2} = \lambda_{k+3} = \cdots ,$$

i.e. the hypotheses that the "left-over" eigenvalues of the covariance matrix are contributed by the noise only — under the assumption of a flat noise spectrum. Starting from $k = 0$, these hypotheses are sequentially tested according to an approximate χ^2-test. The index of the first accepted hypothesis is the model order estimate.

A central question is to find a relevant threshold value on which to base the decision of the tests. This difficulty made penalty based methods much more popular than threshold based methods.

3.1.1 Applicability of the reviewed detection scheme to the SCS channel model

The mentioned methods rely heavily on the whiteness and the gaussian nature of the noise. Indeed, the noise being jointly gaussian, the likelihood function can be expressed in term of the measurements covariance matrix $\boldsymbol{R}$. Also, if the noise covariance matrix is a multiple of the identity, the measurements covariance matrix is simply the sum of the signal covariance matrix and the noise covariance matrix. This property is known in linear algebra as *deflation/inflation*, as it shifts the eigenvalues.

In the sparse common support (SCS) model, the number of snapshots with independent noise realizations, P, is the number of receiving antennas.

Therefore, to obtain a reasonably accurate estimate of the covariance matrix, it is necessary that $P \gg K$ (see Example 3.a) — which is to say, many more antennas than signal "paths" are required. This assumption is in general not satisfied in mobile communications.

Another angle of attack is to consider the columns of the data matrix $\boldsymbol{T}$ introduced in Definition 2.1 as snapshots. There, the number of snapshots is roughly half the number of pilots which is in general much larger than the number of paths.

However, independence between the noise samples had to be sacrificed, since they are replicated along the diagonals of the Toeplitz blocks. We have shown in Chapter 2 that the spectrum of a noise-only data matrix is far from being flat [88]. Its distribution was recently defined by Bryc et al. in [39] via its moments.

To summarize the difficulties :

- The noise and signal spectrums interact in a non-trivial way, *i.e.* it is more complex than a simple inflation by σ^2.
- We conjecture that the likelihood function cannot be simply evaluated from the spectrum of the data-matrix (or the covariance matrix $\propto \boldsymbol{T}^*\boldsymbol{T}$) as no closed-form formula is known for the noise spectral distribution.
- For practical applications, the measurements are not only corrupted by AWGN but also by potential model mismatches, which makes it harder to establish an exact spectral profile for the noise.

Example 3.a — Convergence of the sample covariance matrix

The convergence of $\widehat{\boldsymbol{R}}$ to $\boldsymbol{R}$ is relatively slow with respect to P/N the ratio between the number of snapshots and the dimension of the covariance matrix. Ideally, the sample covariance matrix of the noise should be a multiple of the identity. The ratio between its largest and smallest eigenvalue provides a measure of how far it is from the assumed flat spectrum.

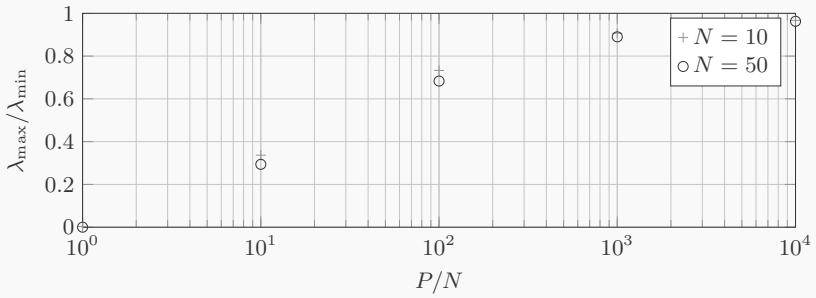


Figure 3.1: *Convergence of the spectrum of the sample covariance matrix to a flat spectrum. The covariance matrix of dimension N is computed from P iid random vectors containing N iid standard gaussian random variables*

Figure 3.1 shows that the number of independent snapshots needs to be orders of magnitude larger than the dimension of the covariance matrix in order to obtain a flat spectrum. In the SCS channel model, where independent snapshots are provided by the antennas of the receiver, this assumption is unpractical.

Therefore a viable alternative seems to use a criterion with an *a posteriori* approximation of the ML estimator likelihood. This implies that the available estimation algorithm has performances close to the ML estimator and it also necessitates the estimation of the signal parameters for all plausible model orders.

This "brute-force" solution, which consists in testing many models and keeping only the best one, may not be suitable for channel estimation on mobile devices in terms of energy efficiency and computational power.

To be more efficient, this approach would require to first obtain a rough estimate of the model to narrow down the number of hypotheses.

3.2 Hypothesis testing for structured data matrices

Let $\boldsymbol{E}$ be a random square Toeplitz matrix of size $n \times n$

$$\boldsymbol{E} = \begin{bmatrix} \varepsilon_0 & \varepsilon_{-1} & \varepsilon_{-2} & \cdots & \cdots & \varepsilon_{-n+1} \\ \varepsilon_1 & \varepsilon_0 & \varepsilon_{-1} & \ddots & & \vdots \\ \varepsilon_2 & \varepsilon_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \varepsilon_{-1} & \varepsilon_{-2} \\ \vdots & & \ddots & \varepsilon_1 & \varepsilon_0 & \varepsilon_{-1} \\ \varepsilon_{n-1} & \cdots & \cdots & \varepsilon_2 & \varepsilon_1 & \varepsilon_0 \end{bmatrix}$$

with iid diagonal entries ε_k with independent real and imaginary parts distributed as standard gaussian random variables.

Remember the block-Toeplitz data-matrix defined in Definition 2.1

$$\boldsymbol{T} \;=\; \boldsymbol{T}_{\text{sig}} + \sigma^2 \boldsymbol{T}_{\text{noise}},$$

where

$$\boldsymbol{T}_{\text{noise}} \;=\; \begin{bmatrix} \boldsymbol{E}_1 \\ \vdots \\ \boldsymbol{E}_P \end{bmatrix}, \quad \boldsymbol{E}_1 \sim \cdots \sim \boldsymbol{E}_p \sim \boldsymbol{E}, \text{ iid.}$$

Definition 3.1. *Define two hypotheses based on the observations $\boldsymbol{T}$, where $\sigma_1(\boldsymbol{T}) \geq \sigma_2(\boldsymbol{T}) \geq \cdots$ are the singular values of $\boldsymbol{T}$*

- $H_1(k)$ *: The signal model order is* $\geq k$.
- $H_2(k)$ *:* $\sigma_k(\boldsymbol{T}) > \|\boldsymbol{T}_{\text{noise}}\|$.

Then

Proposition 3.1.

$$H_2(k) \;\Rightarrow\; H_1(k). \tag{3.2}$$

Proof.

By Weyl's theorem [138]

$$|\sigma_k(\boldsymbol{T}) - \sigma_k(\boldsymbol{T}_{\text{sig}})| \leq \|\boldsymbol{T}_{\text{noise}}\|.$$

Therefore $\neg H_1(k) \Rightarrow \sigma_k(\boldsymbol{T}_{\text{sig}}) = 0 \Rightarrow \neg H_2(k)$, where $\neg$ is the negation. This proves the proposition by contraposition. □

Proposition 3.1 indicates that as long as $H_2(k)$ is true with a certain confidence, the signal subspace dimension can be increased with the same or higher confidence.

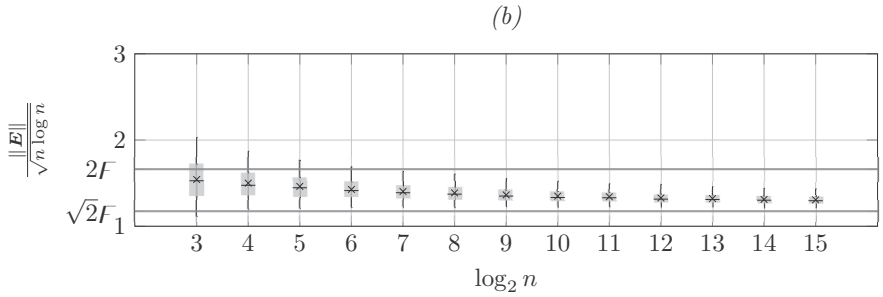


Figure 3.2: *Statistics on the spectral norm of* $\boldsymbol{E}$ *over* 1000 *realizations. The crosses represent the empirical mean and the box plots show — from the bottom whisker to the top whisker— the quantiles* 0.05*,* 0.25*,* 0.5*,* 0.75*, and* 0.95*. The spectral norm measure concentrates as* $n \to \infty$ *to a value in the interval* $[\sqrt{2}\mathcal{F},\ 2\mathcal{F}]$*.*

Failure of the test does not indicate that the model order is overestimated — *i.e.* the relation in (3.2) is not an equivalence. Nevertheless, it may provide a valuable underestimate[6] of the model order, from which to start.

We have already established in Theorem 2.2 that $\|\boldsymbol{T}_{\text{noise}}\| \sim \mathcal{O}(\sqrt{n \log n})$, for the hypothesis testing, a more precise characterization is needed.

3.2.1 The spectral norm of E

We begin our study of $\|\boldsymbol{T}_{\text{noise}}\|$ with one of its blocks, a simpler non-symmetric square Toeplitz matrix $\boldsymbol{E}$.

Proposition 3.2.

$$\sqrt{2} \cdot \mathcal{F} \leq \lim_{n\to\infty} \frac{\|\boldsymbol{E}\|}{\sqrt{n \log n}} \leq 2 \cdot \mathcal{F}\ ,$$

where $\mathcal{F} \approx 0.8288\cdots$.

Proof.

See Appendix C.1 □

In Figure 3.2, the interval found in Proposition 3.2 is confirmed by simulations.

3.2.2 Multiple snapshots ($P > 1$)

The analysis done for a single Toeplitz block does not trivially extend to multiple stacked blocks. The lowerbound stated in Proposition 3.2 for $P = 1$ is also a (trivial)

[6] It is an underestimate if the confidence level used to test the hypothesis is high enough.

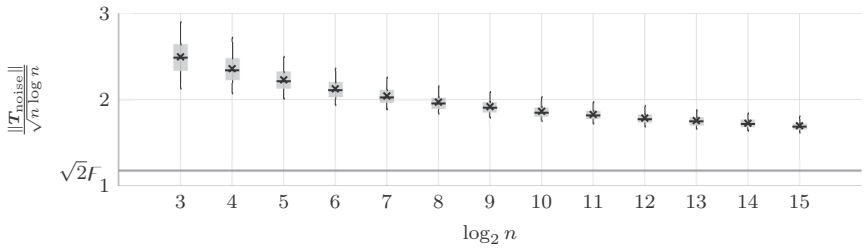


Figure 3.3: *Same setup as in Figure 3.2, now with $P = 4$. The asymptotic lowerbound on the spectral norm for $P = 1$, is also (trivially) a lowerbound for $P > 1$.*

lowerbound for $P > 1$, as shown in Figure 3.3.

For an accurate characterization in function of n and P, numerical simulation can be carried out offline to estimate precisely the spectral distribution of $\boldsymbol{T}_{\text{noise}}$.

Based on this estimation one can compute a threshold to implement the testing of hypothesis $H_2(k)$ at a given level of confidence. This procedure is given as Algorithm 3.1.

3.2.3 Numerical results

To assess the relevance of Algorithm 3.1, we simulate a $P = 4$ SCS channels with $K = 7$ paths each. Each channel impulse response is uniformly sampled at a critical rate of $N = 255$ (large sample regime) and $N = 31$ (small sample regime) samples per period.

In wireless communications, the ambient noise level σ^2 is usually known, as it depends mostly on the gain levels of the amplifiers in the radio-frequency (RF) frontend or can be estimated during idle time. Therefore, we assume σ^2 is known and fixed, and the relative amplitude of the received signal is varied to simulate different SNR regimes.

In Figures 3.4 and 3.5, the model order estimate is given by largest index for which the hypothesis H_2 is accepted

$$\widehat{K} = \max_{H_2(k) \text{ is true}} k,$$

and the box plots schematize its empirical distribution. The model order estimate can be compared to the real model order K, but this comparison becomes meaningless below a certain SNR, where some of the paths cannot be reliably estimated.

A better comparison is to use an oracle which outputs the model order maximizing the quality of the estimation. In the context of communications, maximization of the SNR between the true CIR and the estimated CIR is the goal pursued by channel estimation as it translates into a higher bit-rate if proper coding is applied [47]. Therefore, the standard for comparison is the oracle which outputs the model order for which the estimation algorithm achieves the highest SNR gain.

In view of Proposition 3.1, acceptation of the hypothesis $H_2(k)$ is the affirmation that the true model order is equal or greater than k, therefore the estimate $\widehat{K}$ is bound

Algorithm 3.1 Composite detection method : The hypothesis H_2 and an information criterion

Input: Sparse common support channel measurements $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_P$ (DFT domain), and τ a threshold for the hypothesis H_2

Output: $\widehat{K}$ a model order estimate and $(\widehat{t}_1, \ldots, \widehat{t}_{\widehat{K}})$, $\{(\widehat{c}_{1,p}, \ldots, \widehat{c}_{\widehat{K},P})\}_{p=1:P}$ the SCS model parameters

$i \leftarrow 0$; $k \leftarrow 1$; $\mathcal{Q} \leftarrow \emptyset$; iterate $\leftarrow$ `true`; $\boldsymbol{f} \leftarrow$ `random()`
while iterate **and** $k \leq M+1$ **do**
  $i \leftarrow i+1$
  $\boldsymbol{f} \leftarrow$ `blockToeplitzMult`$(\boldsymbol{y}_1, \ldots, \ldots \boldsymbol{y}_P; \boldsymbol{f})$
  Add $(q_i, \alpha_i, \beta_{i-1}) \leftarrow$ `LanczosIteration`$(\mathcal{Q}; \boldsymbol{f})$ to $\mathcal{Q}$
  $\mathcal{R}_i \stackrel{\text{def}}{=} \{(\widehat{\lambda}_j, \widehat{\boldsymbol{v}}_j)\}_{j=1:i} \leftarrow$ `RitzPairs`$(\mathcal{Q})$
  $\ell \leftarrow$ `ConvergedRitzPairs`$(\mathcal{R}_i, \mathcal{R}_{i-1})$
  while iterate **and** $k \leq \ell$ **do**
    if $\widehat{\lambda}_k > \tau$ **then**
      $k \leftarrow k+1$
    else
      Estimate $(\widehat{t}_1, \ldots, \widehat{t}_k)$ using the Ritz pairs $(\widehat{\lambda}_j, \widehat{\boldsymbol{v}}_j)$
      Estimate $\{(\widehat{c}_{1,p}, \ldots, \widehat{c}_{k,p})\}_{p=1:P}$ based on $(\widehat{t}_1, \ldots, \widehat{t}_k)$ and the measurements
      Compute L_k the log-likelihood of the estimated values
      $\mathrm{ITC}_k \leftarrow -L_k +$ `penalty`(k)
      if $\mathrm{ITC}_k < \mathrm{ITC}_{k-1}$ **then**
        $k \leftarrow k+1$
      else
        iterate $\leftarrow$ `false`
      end if
    end if
  end while
end while

to underestimate the true model order when H_2 is tested with a high confidence, such as 95%.

This observation seems to hold also in the conducted tests. As the number of samples N increases, the measure of the noise spectral norm concentrates and the SVD of $\boldsymbol{T}_{\text{sig}}$ tends to its Vandermonde decomposition — which is to say that each of its singular value corresponds to a path gain $\sigma_k(\boldsymbol{T}_{\text{sig}}) \to \sum_p |c_{k,p}|^2$. This explains why the predicted model order is closer to the oracle in Figure 3.4 than in Figure 3.5.

3.3 Avoiding overfitting

In the previous section, we developed a hypothesis, which when it is accepted with high confidence provides a reliable underestimate of the model order. The advantage of this hypothesis is to be entirely based on the higher end of the data-matrix spectrum which allows for an online evaluation during the identification of the signal subspace.

If the hypothesis is accepted with a low confidence, a positive bias is introduced and the model order estimate is likely to be overestimated — e.g. setting the confidence at 50% leads to often accept spurious dimensions.

We saw in Chapter 2 that the data-matrix spectrum only reveals a low-rank property and not necessarily its inherent Vandermonde decomposition structure.

Therefore, there ought to be a more robust way to estimate the model-order *a posteriori*, *i.e.* after estimation is fully completed. This step can be seen as an extra validation used to reject paths which could be explained solely by the noise.

3.3.1 The validation of paths

We assume that a channel estimation gives us K path estimates for P SCS channels

$$(\widehat{\omega}_k,\ \widehat{c}_{k,1},\ \ldots,\ \widehat{c}_{k,P}), \quad k \in 1,\ \ldots,\ K.$$

Moreover, the path indices have been classified in two sets :

- $\mathcal{K}_+$ contains indices of paths which have been accepted
- $\mathcal{K}_-$ contains indices of paths which status is uncertain.

This classification is feasible in practice since $\mathcal{K}_+ = \emptyset$ is a valid choice.

We first compute the total residual energy for each path in $\mathcal{K}_-$

$$r_k \overset{\text{def}}{=} \sum_{p=1}^{P} |\widehat{c}_{k,p}|^2 \cdot \left(1 - \frac{1}{2M+1} \left\| \text{proj}_{\mathcal{S}_+}[\cdots,\ e^{-j\omega_k m},\ \cdots]^T \right\| \right)^2 , k \in \mathcal{K}_-, \quad (3.3)$$

where

$$\mathcal{S}_+ \overset{\text{def}}{=} \text{span}\{[e^{j\omega_\ell M},\ \cdots,\ e^{-j\omega_\ell m},\ \cdots]^T\}_{\ell \in \mathcal{K}_+}.$$

For each of these residual energy r_k, we should compute the probability to observe it under the assumption that $c_{k,1} = \cdots = c_{k,P} = 0$, *i.e.* roughly[7] the probability of a spurious detection. In the next section we compute a threshold for r_k to target a given false detection rate.

[7] To be absolutely exact, one should also account for interactions between paths within $\mathcal{K}_-$.

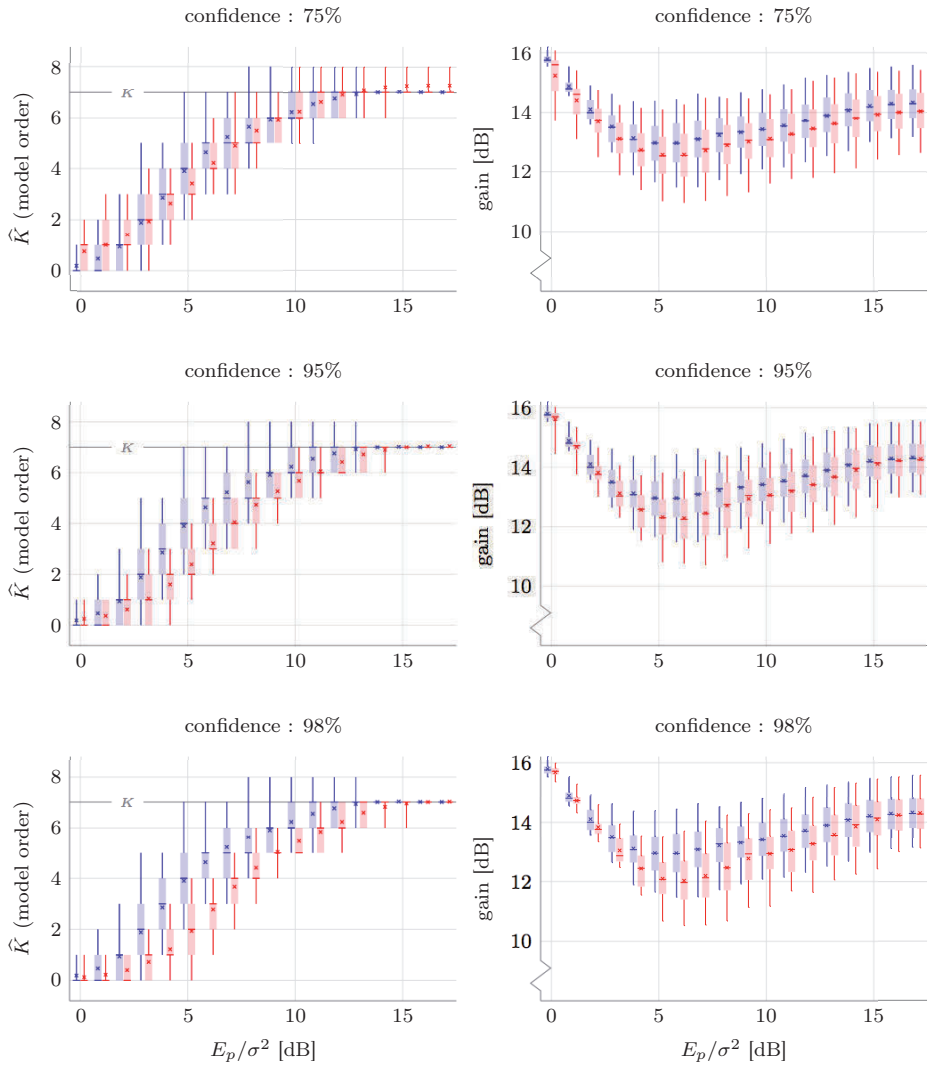


Figure 3.4: *These figures show the model order estimation and equalization gain obtained with an oracle (in blue) — the oracle chooses the model order maximizing the equalization gain — and with the simple hypothesis test of $H_2(k)$ (in red) at different levels of confidence for each row. As expected, using as a model order the maximal index k for which $H_2(k)$ is believed true, leads to a slight underestimation. However, the estimation gain achievable with this model order estimate is close to the one obtained with the oracle.*

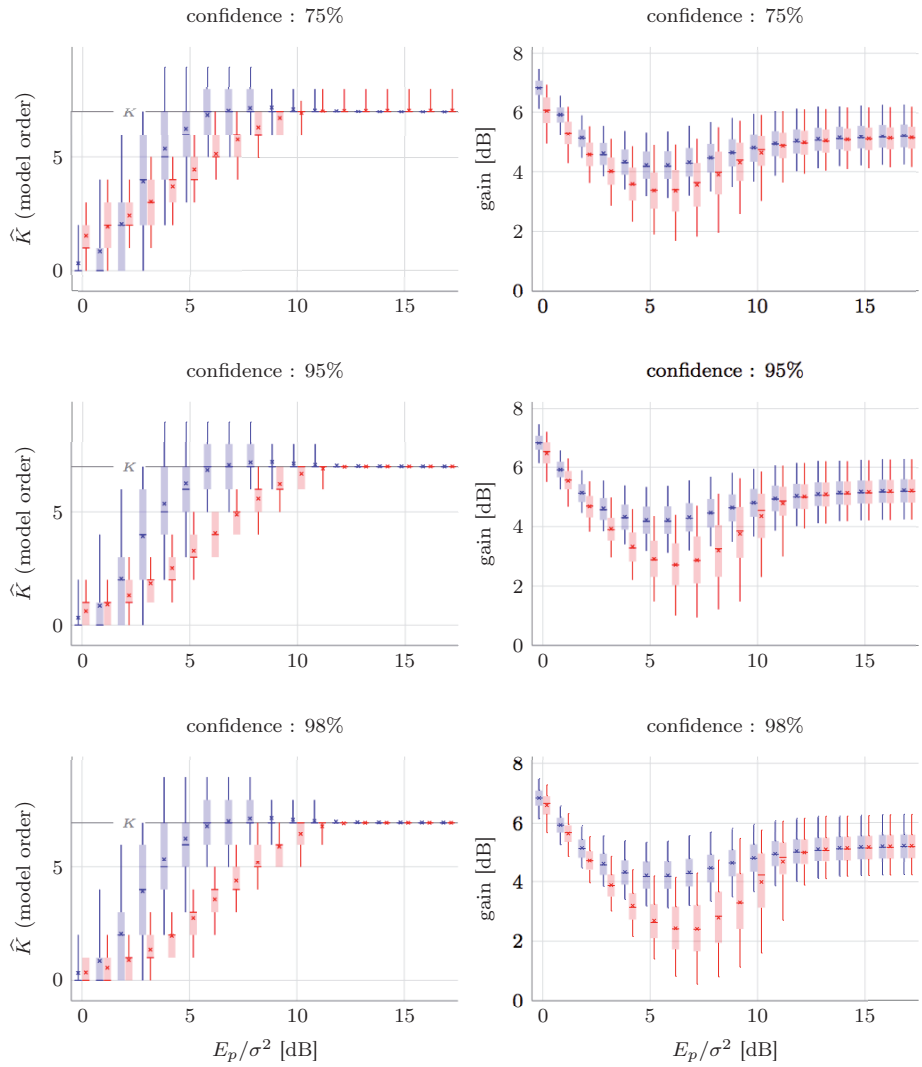


Figure 3.5: *A smaller number of samples leads to a less concentrated measure of the noise matrix spectral norm, which can explain a larger gap. Nevertheless, this hypothesis testing still provides a reliable lowerbound on the model order which can be evaluated online with the fast estimation algorithms developed in Chapter 2.*

3.3.2 Extremal statistics of the noise projection

Define the random variable

$$\rho_{\max}(P, 2M+1) \stackrel{\text{def}}{=} \frac{2}{\sigma^2} \max_{\omega} \sum_{p=1}^{P} \left| \sum_{m=-M}^{M} \frac{e^{j\omega m}}{\sqrt{2M+1}} E_p[m] \right|^2 , \tag{3.4}$$

which is the maximal energy correlation between the noise and the phasor $\frac{e^{j\omega m}}{\sqrt{2M+1}}$ scaled by a factor $\frac{2}{\sigma^2}$.

For $\omega \in [-\pi\ \pi[$, the distribution of $\rho_{\max}$ is not easily evaluated as it is the maximum in a continuum of correlated χ^2 random variables. If we restrict ω to the set $\{\frac{2\pi\ell}{2M+1}\}_{\ell=-M,\ldots,M}$, the phasors form an orthogonal set and the random variable

$$\frac{2}{\sigma^2} \max_{\ell=-M,\ldots,M} \sum_{p=1}^{P} \left| \sum_{m=-M}^{M} \frac{e^{j\frac{\pi\ell m}{2M+1}}}{\sqrt{2M+1}} E_p[m] \right|^2 .$$

is the maximum in a set of iid χ^2 distributed random variables, for which we can easily evaluate the distribution. Because of the bandlimitedness, the difference between the true value of $\rho_{\max}(P, 2M+1)$ and its approximation obtained by restricting ω to the discrete set cannot be arbitrarily large[8], therefore

Proposition 3.3. *A fixed target* false detection rate α *— the expected probability to have at least one detection caused by noise (the probability of overfitting) — can be achieved approximately by the selection criterion*

$$\textbf{\textit{Validate the } } k^{\text{th}} \textbf{\textit{ path if :}} \qquad r_k \ \geq\ \frac{\sigma^2}{c_0 \cdot (2M+1)} t_\alpha \ , \tag{3.5}$$

where r_k is defined as in (3.3), t_α is the solution of

$$\gamma(P, t_\alpha) \ =\ (P-1)! \cdot (1-\alpha)^{\frac{1}{2M+1}} \ ,$$

c_0 is a scalar to be chosen within $]0.77,\ 1]$.

The function $\gamma(P,t) \stackrel{\text{def}}{=} \int_0^t x^{P-1} e^{-x}\, dx$, *is the* lower incomplete gamma function *[2].*

Proof.

See Appendix C.2. □

[8] See Appendix C.2 for an exact quantification.

Corollary 3.1. *For a single channel ($P = 1$), the selection criterion is*

$$\textbf{\textit{Validate the } } k^{\text{th}} \textbf{\textit{ path if :}} \quad r_k \geq -\frac{\sigma^2}{c_0 \cdot (2M+1)} \log\left(1 - (1-\alpha)^{\frac{1}{2M+1}}\right) ,$$
$$= \frac{\sigma^2}{c_0 \cdot (2M+1)} \left[\log \frac{2M+1}{\alpha} + \mathcal{O}(\alpha)\right] .$$

Proof.

Use the identity $\gamma(1,t) = 1 - e^{-t}$, [2] to obtain the first equality. The second equality follows from the Taylor expansion with respect to α in the neighborhood of 0. □

The selection criterion stated in Proposition 3.3 is only an approximation, mostly due to the fact that we assumed the signal did not interfere with the noise when a spurious path is detected, and also that the estimation algorithm maximizes the correlation between the pilot measurements and a phasor.

A second approximation was made in linking the distributions of $\rho'_{\max}$ and $\rho_{\max}$, where we introduced a correction factor of c_0 which reduces the bias of $\rho'_{\max}$ as an estimate of $\rho_{\max}$, but their respective distributions differ slightly[9]. Nevertheless, the dependence on M, P and σ is correctly captured.

Note that at the limit — when $M \to \infty$ — $\rho'_{\max}$ follows a Gumbel distribution [63], which parameters can be found in [51](p. 156).

The quantity E_p/N_0 is the preferred way to measure the SNR in the communication community. At first sight in Proposition 3.3 — if one ignores t_α — it appears that the selection criterion is inversely proportional to E_p/N_0, which would be a natural rule of thumb to follow. In Corollary 3.1, the Taylor expansion makes it more obvious that the selection criterion does not only depend on E_p/N_0 but also on the number of pilots if this number is not negligible as compared to $1/\alpha$, the inverse false positive rate.

Example 3.b — Extremal statistics of the noise

We can first verify how accurately the observed false positive rate matches the target when the selection criterion of Proposition 3.3 is used in an *idealized setup*. In the idealized setup, we consider measurements containing exclusively noise (no path). Each plot display the probability (vertical axis) to have a correlation between the noise measurements and a pulse shape — *i.e.* a matched-filter output — less than a value t (horizontal axis). In each plot, we report three curves (upper and lower envelopes and a fit) and simulated data points. The upper envelope is obtained if the maximum of the matched filter output were to be

[9] For all tested values of M and P, $c_0 \approx 0.9$ gave accurate results. See Appendix C.2 for more details, especially Figure C.1.

taken among $2M+1$ uniform shifts, and it corresponds to the choice of $c_0 = 1$. It is an upper envelope because the output of the matched filter must be greater than the output of the matched filter taken on a small set of shifts, therefore the value of the cumulative distribution is over estimated. The lower envelope is obtained from an upperbound on the maximum of the matched filter output and the plotted curve is thus below the expected correlation, and corresponds to $c_0 = 0.77$.

The true expected maximum of the matched-filter output is obtained by simulation and the probability it exceeds a threshold t is shown as a data point. These simulation data are accurately approximated by choosing the value $c_0 = 0.9$.

The number of channels increases from one line of plots to the next, and the number of pilots increases from one column to the next.

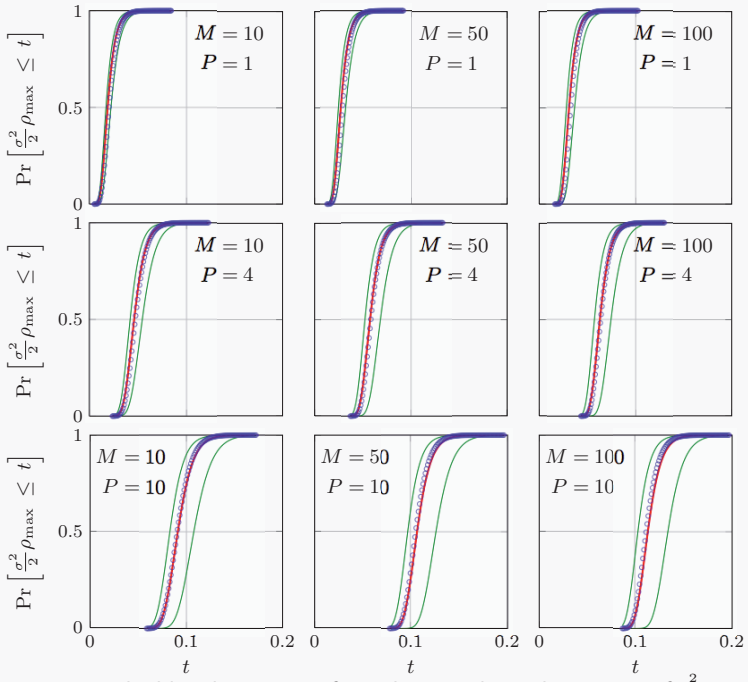


Figure 3.6: *The blue data-points form the cumulative histogram of $\frac{\sigma^2}{2}\rho_{\max}$ — the maximum of the correlation between complex-valued AWGN and a phasor for arbitrary shifts in $[-\pi,\ \pi[$. The green curves are the CDF obtained with Proposition 3.3 for the extreme values of c_0 (0.77 and 1). The red curve is obtained with $c_0 = 0.9$ and provides a good fit of the empirical data over all the tested values of M and P.*

Therefore one can accurately estimate the probability that a noise only signal would have generated a detection for a given magnitude.

It does not address other discrepancies such as the difference between the maximum of the matched-filter and what is obtained with a subspace method such as the joint ESPRIT algorithm we proposed in Chapter 2; or the interferences between the signal and the noise.

3.3.3 An algorithm to prevent overfitting

We may combine the precedent results into an algorithm used after estimation as an extra validation step.

The main idea is to start with an empty set $\mathcal{K}_+$ of validated paths, and to add one-by-one the paths based on the magnitude of their residual r_k. Then, every time a path is added to $\mathcal{K}_+$, the projections into the span of $\mathcal{K}_+$ can be approximated by a "Gram-Schmidt"-like orthogonalization. Because the correlation between two phasors has a closed-form evaluation, the total cost of the procedure is $\mathcal{O}(K^2)$, making it negligible compared to the estimation.

3.3.4 On the selection of a false positive error rate

The key parameter α is left undetermined. It is not intrinsically bad for a model order selection method to have a "knob" to turn — it would be if the corresponding parameter has no operational meaning or interpretation. Here α is linked to a particular type of error which can be critical for some application.

For example, in secure ranging the estimation of the ToF should not include false positive, *i.e.* a far-away eavesdropper should not be able to appear in the vicinity[10]. A false positive could result in unwarranted access, while a false negative result in the denial of a legitimate service. The second issue is annoying while the first one is catastrophic. For this application, the threshold in Proposition 3.3 would be computed for a low value[11] of α.

On the other hand, if the task is to maximize the SNR between the estimated and the true channel impulse response, α should be chosen close to $1/2$ so that the expected SNR gain when rejecting or validating a path with a gain close to the threshold value is roughly 0 dB[12].

Example 3.c — Using validation to prevent overfitting

We generate $P = 4$ SCS channels with $K = 5$ paths and have $2M + 1 = 101$ pilots for each of them. We start from an overestimated a priori model order of 7.

To get rid of the overfitting, we reject estimated paths according to Proposition 3.3 for a targeted false detection rate α. For different values of α and different SNR ($E_p/(\sigma^2(2M + 1))$) we measure α' the fraction of the time the validated model order exceeds 5 :

[10] One of the main application of secure ranging is to use proximity as a trust-metric. With this one can pair devices based on their distances.

[11] To get an idea of how low α can be, the paranoid gold standard of "5σ" to validate a discovery in particle physics amounts to a false acceptance rate of the order 10^{-7}, while the false acceptance rate for an "evidence" is of the order 10^{-3}, and for "coverage by press agencies" of the order $1/2$. But we are not physicists, aren't we?

[12] The conclusion that communication systems work like news agencies is correct from a non-Bayesian perspective. However the prior of faster than light neutrinos is not quite comparable with the hypothesis of having 6 paths instead of 5.

SNR	$\alpha =$	0.5	0.1	0.05	0.01
-10 dB		0.056	0.004	0	0.0005
-5 dB		0.196	0.06	0.038	0.0125
0 dB		0.266	0.088	0.056	0.021
5 dB		0.27	0.092	0.062	0.0215
10 dB		0.274	0.088	0.06	0.021

The observed overfitting rate α' only matches α loosely — the numerous simplifications made on the interactions between the signal and the noise are a probable cause. Below 0 dB , α' does not match α because of how we defined it : validating 5 paths or less is not equivalent to not having spurious detection when all the 5 paths cannot be reliably estimated.

Not overfitting is not a goal in itself, we must also verify that the validation does not reject legitimate paths. For $\alpha = 0.1$, the statistical frequencies of the validated model order are

SNR	$\widehat{K} =$	1	2	3	4	$5 = K$	6	7
-10 dB		0	0.02	0.288	0.566	0.122	0.004	0
-5 dB		0	0	0.002	0.302	0.636	0.06	0
0 dB		0	0	0	0.002	0.91	0.088	0
5 dB		0	0	0	0	0.908	0.092	0
10 dB		0	0	0	0	0.912	0.088	0

The behavior of the validation procedure is intuitive. At high SNR, the result concentrates around the true value, and below a certain SNR, some paths are missed, and the distribution of the validated model order spreads since the weakest are not always reliably estimated. Moreover the paths used for the simulation do not have equal amplitudes ($[1,\ 1,\ -1,\ 0.7,\ 0.4]$), and so the validated model order does not drop suddenly with the SNR.

3.4 The Partial Effective Rank (PER) criterion

In the previous sections, we used the well defined AWGN noise model. We have seen how one can then formulate and test hypotheses on the spectrum of the data matrix $\boldsymbol{T}$ and on the estimated path amplitudes.

A possible caveat of this approach, is its strict dependance on the noise model. If thermal and (some) background noises are accurately modeled by a white gaussian process, other model mismatches (interferences, backscatter, ...) are harder to characterize precisely, and a less powerful but more robust criterion not depending directly on the noise statistics may be desirable.

Therefore one is lead to think about the detection problem in its most general terms at an early stage: we observe a block Toeplitz data matrix which ideally shall be low-rank.

The main difficulty in exploiting the low-rank structure is the inherent "roughness" of the rank. Indeed, if we start with a data matrix of rank K much smaller than the full rank, the addition of the least random perturbation of the original data immediately

makes it fullrank; and the original property is completely lost.

In this section, we propose to use a "smooth" functional to replace the notion of rank. This well behaved functional is called the *effective rank* introduced by Roy *et al.* [107] based on [41].

We will use the *effective rank* to devise a heuristic model order selection criterion[13] called the *partial effective rank* (PER).

The core idea is to monitor the increase of the effective rank with the increase of the model order. The criterion is then to detect a transition between two regimes which indicate a probable transition between signal and noise dominance.

3.4.1 The effective rank

The effective rank, is a matrix functional introduced by Roy [107] which may be seen as a "convexification" of the rank.

Definition 3.2. *Let* $\boldsymbol{A}$ *be a non-null matrix with singular values* $\boldsymbol{\sigma} = [\sigma_1, \ldots, \sigma_M]^T$ *in decreasing order, and singular values distribution (SVD) equal to*

$$p_m = \sigma_m / \|\boldsymbol{\sigma}\|_1 , \quad m = 1, \ldots, M.$$

The Effective Rank *of* $\boldsymbol{A}$ *is*

$$\operatorname{erank}(\boldsymbol{A}) = e^{\mathcal{H}(p_1,\ldots,p_M)},$$

where $\mathcal{H}$ *is the entropy of the singular values distribution*

$$\mathcal{H}(p_1, \ldots, p_M) = -\sum_{m=1}^{M} p_m \log_e p_m.$$

By convention $\operatorname{erank}(\mathbf{0}) = 0$.

This definition shares some similitudes with the *entropy power* [47], in which the effective rank could be seen as the square-root of the entropy power of a discrete random variable having the spectrum of $\boldsymbol{A}$ for distribution.

Unfortunately, the entropy power of a discrete random variable does not inherit the properties enjoyed in the continuous case — notably, the entropy power inequality (EPI) is known not to hold in general [113; 116].

The main properties of the effective ranks are

Proposition 3.4. ***(Roy et al. [107])***

[13] Note that contrary to the hypothesis testing, the PER criterion is — for this application — less grounded in theory.

For any matrix $\boldsymbol{A} \neq \boldsymbol{0}$ of dimension $M \times N$ and singular values $\sigma_1 \geq \sigma_2 \geq \ldots$

- $1 \leq \operatorname{erank}(\boldsymbol{A}) \leq \operatorname{rank}(\boldsymbol{A}) \leq \min(M, N)$.
- $\operatorname{erank}(\boldsymbol{A}) = 1$ *iff* $\sigma_2 = \sigma_3 = \cdots = 0$.
- $\operatorname{erank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A})$ *iff its non-0 singular values are all identical.*
- *The effective rank is invariant with respect to scaling, unitary transformation, transposition (real or hermitian).*

For Hermitian positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

- $\operatorname{erank}(\boldsymbol{A} + \boldsymbol{B}) \leq \operatorname{erank}(\boldsymbol{A}) + \operatorname{erank}(\boldsymbol{B})$.

Proof.

See [107]. □

3.4.2 The partial effective rank (PER)

Consider a tall matrix[14] $\boldsymbol{A}$ of dimension $M \times N$ of rank K having the singular value decomposition

$$\boldsymbol{USV}, \quad \boldsymbol{U} \in \mathbb{C}^{M \times M}, \ \boldsymbol{V} \in \mathbb{C}^{M \times N},$$

and $\boldsymbol{S}$ a real-valued diagonal matrix with diagonal coefficients

$$\sigma_1 \geq \cdots \geq \sigma_K \geq \sigma_{K+1} = \cdots = \sigma_M = 0.$$

Denote $\boldsymbol{A}^{(k)}$ the best rank k approximation of $\boldsymbol{A}$ with respect to the Frobenius norm, which is obtained by taking a partial SVD expansion including only the k principal SVD basis vectors.

The rank of the matrices in the sequence $\left(A^{(k)}\right)_k$ increase one by one with k until $k > K$, at which point it stalls to the value K. In general one may say

$$\operatorname{rank}\left(\boldsymbol{A}^{(k+1)}\right) - \operatorname{rank}\left(\boldsymbol{A}^{(k)}\right) = \begin{cases} 1 & , \ \sigma_{k+1} > 0 \\ 0 & , \ \sigma_k = 0 \end{cases}.$$

The progression of the rank as the number of considered singular values increases follows an "all or nothing" rule, which is highly sensitive to small perturbations of the spectrum.

The effective rank can used in place of the rank to mitigate this sensitivity. We define the partial effective rank of degree k as $\operatorname{erank}\left(\boldsymbol{A}^{(k)}\right)$ and obtain a measure indicating the relative importance of a dimension in the singular basis of a matrix

[14] This requirement is used to simplify the notation, and incurs no loss of generality since $\operatorname{erank}(\boldsymbol{A}) = \operatorname{erank}(\boldsymbol{A}^*)$

Definition 3.3. *The* relative weight *of the* k^{th} *singular dimension of a matrix* $\boldsymbol{A}$ *is*

$$\eta_k(\boldsymbol{A}) \stackrel{\text{def}}{=} \operatorname{erank}\left(\boldsymbol{A}^{(k)}\right) - \operatorname{erank}\left(\boldsymbol{A}^{(k-1)}\right) \ , k \geq 1.$$

Theorem 3.1. *For all* $k \geq 1$ *and any matrix* $\boldsymbol{A}$ *with absolutely summable singular values* $\sigma_1 \geq \sigma_2 \geq \cdots$,

$$0 \leq \eta_k(\boldsymbol{A}) \leq 1.$$

The lowerbound is met iff $\sigma_k = 0$ *and the upperbound is met iff* $\sigma_k = \cdots = \sigma_1$.

Proof.

Let

$$\begin{aligned} p_\ell &= \frac{\sigma_\ell}{\sum_{i=1}^{k} \sigma_i}, \ \ell \leq k, \\ q_\ell &= \frac{\sigma_\ell}{\sum_{i=1}^{k+1} \sigma_i}, \ \ell \leq k+1, \end{aligned}$$

be respectively the normalized singular values of $\boldsymbol{A}^{(k)}$ and $\boldsymbol{A}^{(k+1)}$.

Then for $\lambda = \frac{\sigma_{k+1}}{\sum_{i=1}^{k} \sigma_k}$

$$\begin{aligned} \mathcal{H}(\boldsymbol{q}) &= -\sum_{\ell=1}^{k} \frac{p_k}{1+\lambda} \log\left(\frac{p_k}{1+\lambda}\right) - \frac{\lambda}{1+\lambda} \log\left(\frac{\lambda}{1+\lambda}\right) \ , \\ &= \frac{\mathcal{H}(\boldsymbol{p})}{1+\lambda} + \frac{\log(1+\lambda) - \lambda \log \frac{\lambda}{1+\lambda}}{1+\lambda} \ , \\ &= \frac{\mathcal{H}(\boldsymbol{p}) - \lambda \log(\lambda)}{1+\lambda} + \log(1+\lambda) \ \geq \ \frac{\mathcal{H}(\boldsymbol{p}) - \lambda \log(\lambda)}{1+\lambda}. \end{aligned}$$

Since the singular values are non increasing, $\lambda \leq 1/k$. Also, the effective rank is majorized by the rank, thus

$$e^{\mathcal{H}(\boldsymbol{p})} \leq k \leq \frac{1}{\lambda}$$

which implies $\mathcal{H}(\boldsymbol{p}) \leq -\log \lambda$.

We conclude that $\mathcal{H}(\boldsymbol{q}) \geq \mathcal{H}(\boldsymbol{p})$, proving that $\eta_k(\boldsymbol{A}) \geq 0$.

Also, the term $\log(1+\lambda) \geq 0$ we dropped vanishes iff $\lambda = 0$. Therefore $\sigma_{k+1} = 0$ is a necessary condition to have $\eta_k(\boldsymbol{A}) = 0$, and one can verify equality is indeed met in this case.

The proof of the upper-bound is obtained using the subadditivity of the effective rank for hermitian positive semidefinite matrices. The subadditivity property extends to arbitrary matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ if span $\boldsymbol{A} \perp$ span $\boldsymbol{B}$:

$$\begin{aligned}
\operatorname{erank}(\boldsymbol{A}+\boldsymbol{B}) &= \operatorname{erank}\left(\sqrt{(\boldsymbol{A}+\boldsymbol{B})^*(\boldsymbol{A}+\boldsymbol{B})}\right) \\
&= \operatorname{erank}\left(\sqrt{\boldsymbol{A}^*\boldsymbol{A}}+\sqrt{\boldsymbol{B}^*\boldsymbol{B}}\right) \\
&\leq \operatorname{erank}\left(\sqrt{\boldsymbol{A}^*\boldsymbol{A}}\right)+\operatorname{erank}\left(\sqrt{\boldsymbol{B}^*\boldsymbol{B}}\right) \\
&= \operatorname{erank}(\boldsymbol{A})+\operatorname{erank}(\boldsymbol{B}).
\end{aligned}$$

Therefore

$$\begin{aligned}
\operatorname{erank}\left(\boldsymbol{A}^{(k+1)}\right) &\leq \operatorname{erank}\left(\boldsymbol{A}^{(k)}\right)+\operatorname{erank}\left(\sigma_{k+1}\cdot\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^*\right) \\
&= \operatorname{erank}\left(\boldsymbol{A}^{(k)}\right)+1.
\end{aligned}$$

□

Theorem 3.1 essentially indicates that the evolution of the effective rank smoothes the evolution of the rank, which is the desired behavior. The non-increasing ordering of the singular values is essential, and the theorem would not be true otherwise[15].

3.4.3 The PER in action

We have verified the evolution of the partial effective rank obeys rules which makes it a smooth approximation of the rank evolution, but it is not clear yet how this new quantity shall be used in order to estimate the intrinsic dimensionality of a matrix.

To gain insight, we test it on a Toeplitz data-matrix $\boldsymbol{T}$ having an intrinsic dimension 7. The generator of the matrix is a multipath signal corrupted by AWGN. The evolution of the effective rank is shown in Figure 3.7 at various SNR.

A knee in the curve is clearly visible when crossing the intrinsic dimension. Note that as the SNR decreases the knee shifts to the left, indicating that the dimensions of the signal space having the lowest power are missed.

Based on this insight, we propose to use a very simple knee detector on the sequence $(\eta_k)_{k=0,\ldots}$, based on the local maximum of the second derivative. This method is purely heuristical and incurs a small overhead of computed singular values.

We can now use this criterion to estimate the model order[16], we report the results in Figure 3.4.3.

[15] As a little thought experiment, assume that σ_{k+1} strongly overpowers all the preceding singular values which are all equal. The PER would drop from k to 1_+.

[16] One may combine the PER criterion with the test of the hypothesis H_2, since the latter is essentially "free" to test. It is not done in this experiment.

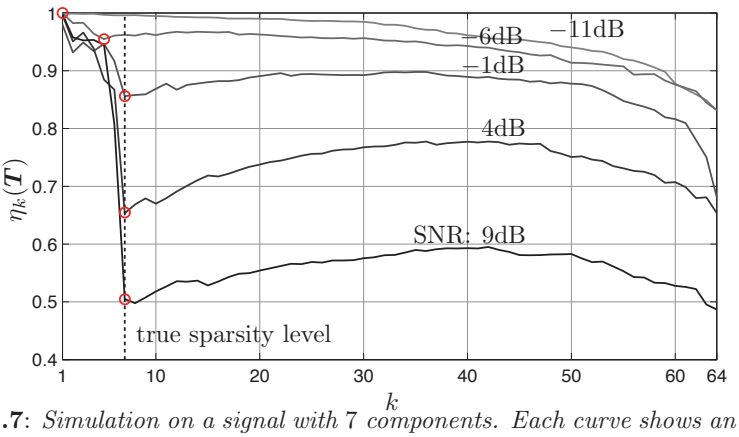


Figure 3.7: *Simulation on a signal with* 7 *components. Each curve shows an average over* 10 *different noise realizations at a particular SNR. The variations in the curve show a clear inflection at* $K = 7$ *(indicated by a dashed line). As the SNR diminishes, the inflection occurs at lower values of* K *and completely disappears at SNRs* $< -10dB$*. The circled markers* ○ *indicate a knee in the curve, or the origin if no knee is present in the curve. The curves are not monotonous, as the evolution of the PER reflects how significant are each dimension of the matrix* ***compared to the previous ones****. The first seven dimensions are all significant (signal space) but the first* 19 *are not, and so the PER increases more for the* 20^{th} *dimension than for the* 8^{th}*.*

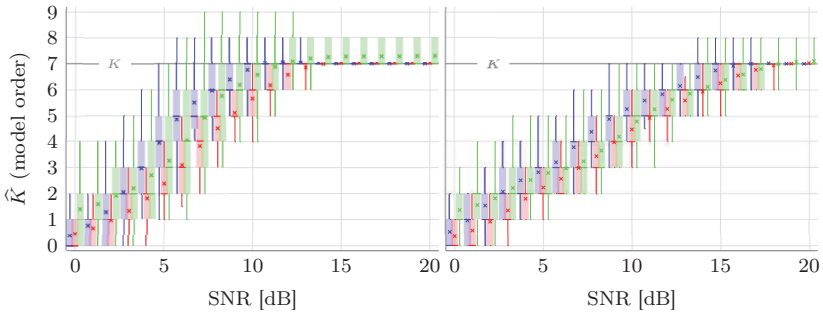


Figure 3.8: *Model order estimation with the PER (green) compared with the* H_2 *hypothesis (red, confidence set at* 95%*) and an oracle which maximizes the equalization SNR gain (blue). The left panel has* $N = 31$ *pilots and the right panel* $N = 63$ *pilots. The number of channels is* $P = 4$*.*

3.5 Test-case : the Weikendorf measurements

Now that both estimation (Chapter 2) and detection (current chapter) have been discussed for the SCS channel model (Chapter 1), it is time for the "reality-test" [14].

Example 3.d — Suburban propagation scenario (Weikendorf)

The Weikendorf measurements [67] are epitomical of a suburban environment exhibiting a strong line of sight path and a couple reflection paths.

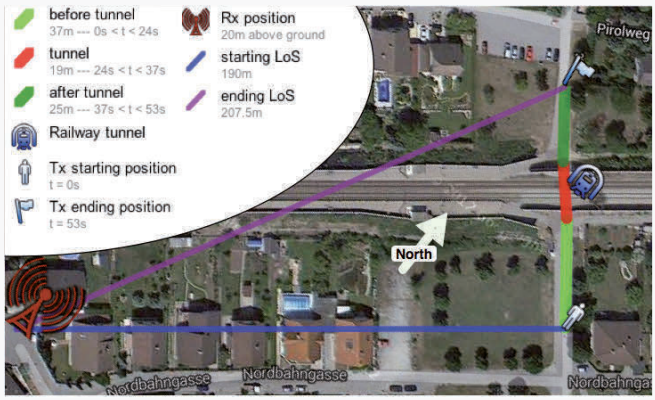


Figure 3.9: *The Weikendorf measurements scenario.*

As it was shown in Figure 1.6, the transmitter goes through a tunnel. Within the tunnel, the channel impulse response is not well described by the SCS model. This particularity makes us anticipate interesting results for an algorithm detecting the presence of sparsity.

As a reminder (see Table 1.1), the parameters of the experiment are

Property	Value
Center frequency f_c	2 GHz
Center wavelength λ_c	15 cm
Bandwidth	120 MHz
Mobile Tx	15 monopole antennas uniformly arranged on a 30 cm diameter circle at 1.5 m from the ground
Static Rx	8 patch antennas separated by 7.5 cm ($\lambda_c/2$) forming a linear array at 20 m from the ground
Time interval between snapshots	21 ms
Tx speed	3 to 6 km/h.
Recording	About 1 minute. The Tx travelled a distance of about 50–80 m) and went through a tunnel

The Weikendorf measurements have a high SNR, and we will simulate a transmission with less power by adding synthetic AWGN to the measurements. The DFT pilots are uniformly laid-out every $D = 3$ DFT bin. We will use the estimated CIR to demodulate 4-PSK coded data symbols occupying the DFT bins allotted to the data, and the obtained *Symbol Error Rate* is the quality metric we will use to benchmark the estimation (the lower the better).

For the task of estimation, we use the ESPRIT based algorithm with the Krylov subspace method for the identification of the signal subspace.

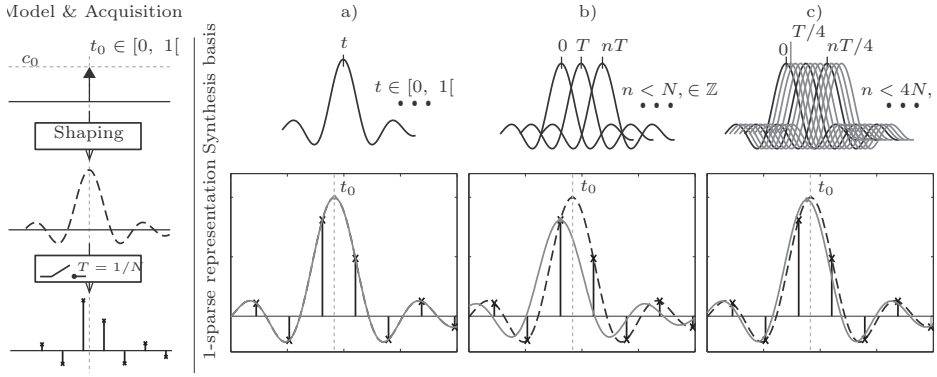


Figure 3.10: *Consider a single pulse (– – –, dashed curve) critically sampled as shown on the left column. From the samples (×, stems), 1-sparse representations of the original signal are computed. In a), the problem is treated as a parametric estimation of* t_0 *and* c_0 *as in the FRI framework. Conceptually, the signal component is chosen from the infinite and uncountable set of the pulse shape and its shifts in* $[0,\ 1[$*. The original signal has a perfect 1-sparse representation in this setup. In b), the signal component is chosen in a finite set of functions forming a basis of the signal space. The original signal can be represented by these functions, but it does not have a 1-sparse representation in general. In c), three times more synthesis functions are added to the set, to form a frame. The signal has a much closer 1-sparse representation in this frame thanks to the shift invariance introduced by the redundancy between the synthesis functions, but the estimation becomes combinatorially more complex. The estimation frameworks b) and c) are referred as* discrete sparsity *which is used in* Compressed Sensing *(CS), and the estimation is subject to a trade-off between accuracy and complexity.*

For comparison, we will also apply two other algorithms

- The "classical" lowpass interpolation method. The lowpass interpolation of the channel spectrum from the pilot measurements is more easily understood in the time-domain where it consists in truncating the measured CIR to a fraction of the frame interval.
- A parametric method using "sample sparsity" called RA-ORMP [49] (*Rank aware orthogonal recursive matching pursuit*). This method comes from the field of compressed sensing and efficiently exploits jointly sparse patterns, where sparsity is to be understood in term of non-zero samples in the time domain. The difference between sample sparsity and the notion we use is explained in Figure 3.10. When the pilots are contiguous, a trivial speed-up is possible using the FFT, we will call this algorithm fast RA-ORMP. We fix the sparsity level appropriately by hand.

We use the PER criterion to estimate the model order. One of the reason is to have a fairer comparison : since we add synthetic AWGN, estimating the model order based

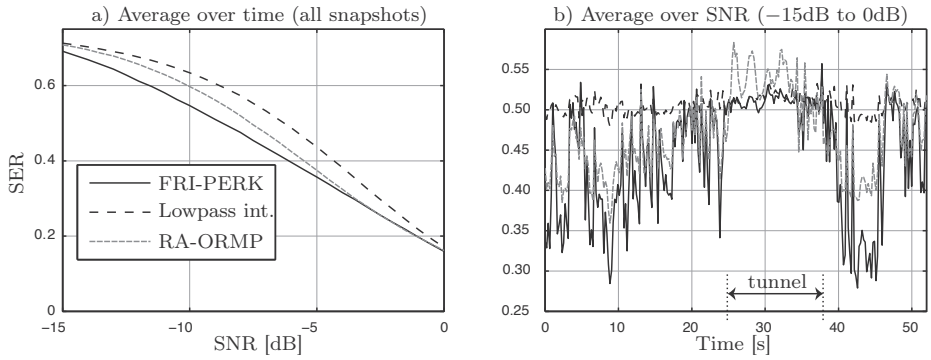


Figure 3.11: *From the left panel a), we conclude that sparse recovery lowers the SER compared to non-sparse recovery below* 0*dB of SNR. If we look at the SER over time — right panel b) — we see that FRI-PERK is robust in the sense that if the input signal is not sparse, its performs approximately as well as a non-sparse recovery.*

on the noise statistics could give an unfair advantage. Whenever the PER criterion reaches the predefined maximum model order (we set it to 10), the algorithm falls back to lowpass interpolation of the spectrum, *i.e.* it gives up on the sparse property of the channel — a similar mechanism was not included in the RA-ORMP algorithm, so we will be able to see if it improves the performances in the tunnel.

To please the acronym deity, we call the resulting algorithm FRI-PERK (FRI with PER detection and Krylov method) — things could have been worse, we could have gone recursive[17].

Interpretation of the results

From Figure 3.11 we may conclude that

- The channels do not exactly fit the SCS model, therefore the modelization error becomes larger than the noise at high SNR
- The SCS property helps in lowering the symbol error rate at medium to low SNR (below 0 dB)
- The "sparsity" model assumed by FRI (few reflections) matches the field measurements better than the one assumed by CS (few non-0 coefficients) as seen in Figure 3.12.
- Any algorithm exploiting sparsity must be "*introspective*", i.e. it must detect when sparsity does not occur, and fall-back to a non-sparse method whenever it happens. It is exemplified by the stroll through the tunnel.

[17]Trivia : *What does the letter "B" stands for in "Benoit B. Mandelbrot" ?*
Answer : „Benoit B. Mandelbrot"

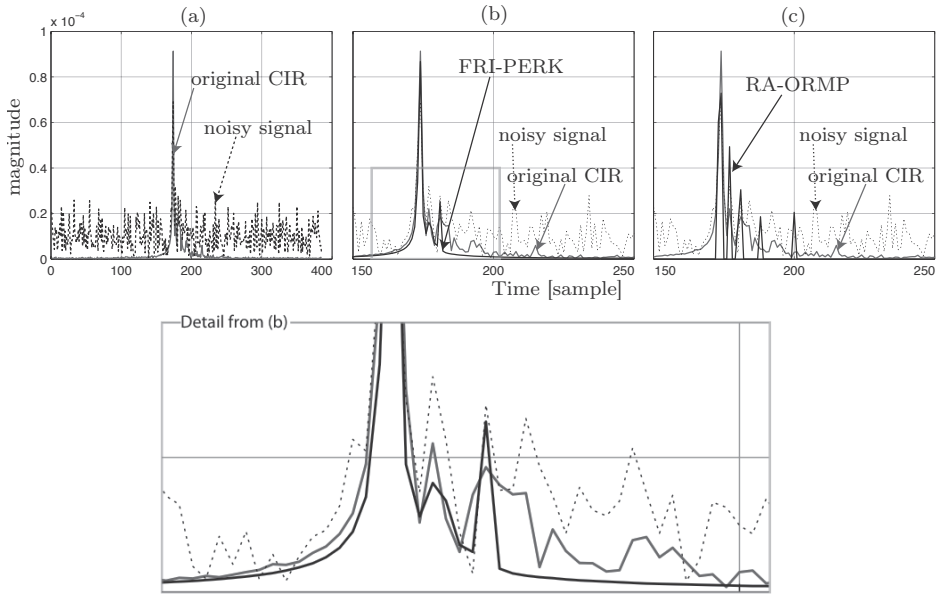


Figure 3.12: *This figure compares the estimation result of FRI-PERK and RA-ORMP. The input signal is the first frame received at the first antenna corrupted with AWGN to obtain $-5dB$ of SNR. Panel (a) shows the original and noisy CIR. Panel (b) shows a portion of interest of the CIR estimated with FRI-PERK. The PER criterion estimates $K = 3$, and by visual inspection on the "detail" panel, the three signal components found match the largest ones of the original signal. Also, it is visible that the envelope of the LoS path is accurately reproduced. Panel (c) shows the result obtained with RA-ORMP. The discrete sparsity model causes the estimation to be more sensitive to uncorrelated noise, as spurious spikes contributed by noise are estimated as signal components. It is important to remember that the noisy CIR is not completely observed, but only a subset of its DFT coefficients, which explains the reconstruction may be worse than the "noisy signal" curve itself.*

For the comparison to be complete between the sparse methods, Figure 3.13 shows a benchmark on synthetic fading SCS data[18]. The comparison of the execution time shows that the algorithmic cost of the fast implementations — panel (a) — have the same complexity with respect to M and P (FRI-PERK is slower by a constant factor of 4). However a fast implementation of RA-ORMP could not be found for non-contiguous pilots which are commonly used when the delay-spread is only a fraction of the frame duration. In that case — panel (c) — RA-ORMP has the same algorithmic complexity as the non accelerated FRI-PER algorithm (RA-ORMP is slower by a constant factor of 5).

For additional details on the experiments, see [14].

[18] All the algorithms were implemented in MATLAB, timing are indicative, not absolute.

3.6 Conclusion

Hopefully, testing the proposed methods on "real-life" channel impulse responses has made clear that estimation is only half of the picture, the other half being detection /model selection.

The addition of a robust detection within the estimation algorithms from Chapter 2 is a necessity to transform them from an academic project into a practically usable algorithm, for the simple reason that it provides the model flexibility required by mobile communications.

The goal of this chapter was to study how early in the algorithmic chain a model can be selected in order to avoid unnecessary computations. To this end, we followed the estimation process, and saw what could be said about the model as it progresses. The first meaningful clues we had are the partial spectrum of the data matrix, being progressively uncovered. We analyzed the properties of this spectrum from a random matrix theory point-of-view, and obtained a biased estimate based on a hypothesis test. Depending on the confidence used for this test the estimate had either a positive or negative bias. From this point of view, we saw how paths could be validated at the end of the estimation procedure.

In a second approach, we deliberately ignored the noise statistics and focused solely on the geometry of the data matrix. We proposed a criterion based on a convexification of the rank. Because this criterion ignores most of the specificities of the problem, it is surely not optimal, but as a side-effect it should also cope fairly well with undefined model mismatches and may have applications in fields other than communications (collaborative filtering, recommendation systems, ...). A possible application of this criterion in conjunction with the estimation algorithm developed in Chapter 2 was shown on a measured CIR (with the addition of AWGN). More tests would be required, particularly on CIRs with many strong reflections (urban environment for example).

This chapter concludes our study on sparse channels estimation at a particular time instant. In the next chapter, we will study the evolution of sparse channels over time and see if any property can be exploited.

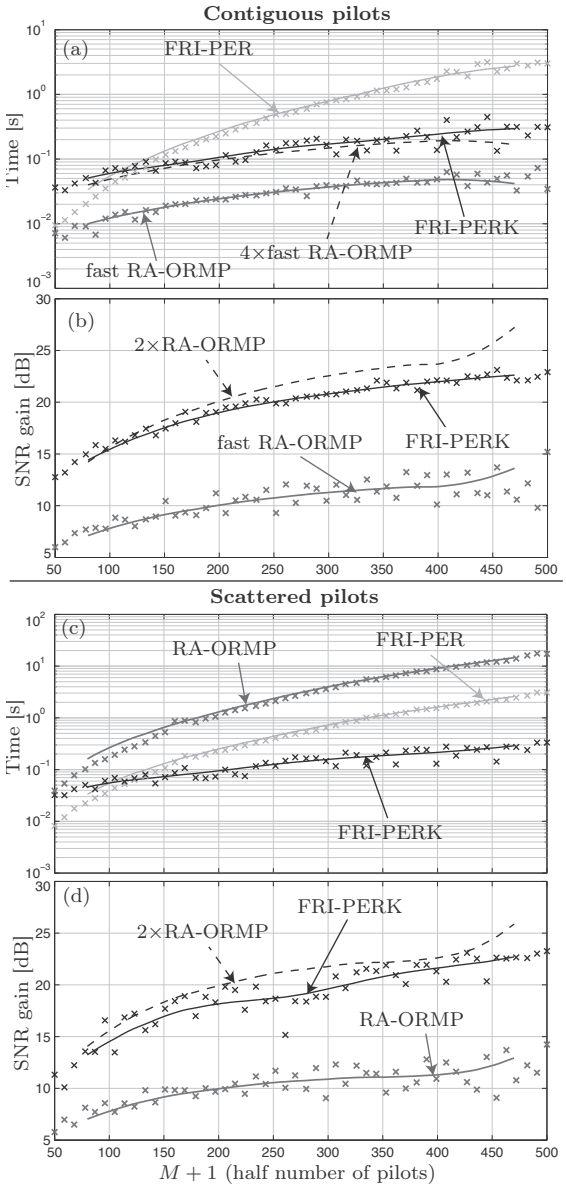


Figure 3.13: *A benchmark is run for pilot sequences of length* $2M + 1 = 101$ *to* $2M + 1 = 1001$. *The top row compares FRI-PERK with FRI-PER — same algorithm using a full SVD instead of Krylov subspace projection — and a fast implementation of RA-ORMP using FFTs, which is possible since the pilots are contiguous in frequency* ($D = 1$). *The bottom row follows the same procedure, but with scattered pilots* ($D = 3$) *and a delay-spread smaller than the frame length. There is no straightforward fast implementation of RA-ORMP in this case. The curves labeled "*$2\times$*" and "*$4\times$*" are simply copies of the curves in the bottom doubled or quadrupled (their use is for visual comparison and they do not correspond to an algorithm).*

Chapter 4

Tracking sparse channels

In Chapter 2, we considered the problem of estimating sparse common support (SCS) channels without prior information about the channels parameters. In practice, measurements from the past may give some information about the current state of the channels because of the temporal correlation of parameters. Exploiting this correlation across time is called *tracking*.

When is tracking relevant? A particular case where temporal correlation can be exploited is when the times of arrival (ToA) vary slowly[1], *i.e.* if between measurements collected T seconds apart, the ToA vary by a small amount — say less than Δt. This is the case if[2]

- The receiver moves at a low speed
- The scatterer moves at a low speed
- The transmitter moves at a low speed

The variation is then upper bounded by

$$\Delta t \leqq \frac{2T}{c}(\text{receiver speed} + \text{scatterer speed} + \text{transmitter speed}).$$

The magnitude of this time variation is not meaningful in itself, and it is to be compared with the inverse bandwidth of the channel. As a quick example, pedestrian speeds together with bandwidths on the order of the MHz yield variations of less than a sample per second. This slow and smooth evolution is visible on the field measurements in Example 4.a, shown hereafter.

From this first rudimentary but practical example, it appears that tracking shall be considered as a useful complement for *estimation*; additional *detection* questions — such as appearance/disappearance of a path — are left for further work.

[1] The exploitation of correlation of paths amplitudes over time [70] is also an interesting problem, not treated in this chapter.

[2] Neglecting clock-speed discrepancies and taking into account first order reflections only.

Example 4.a — Suburban propagation scenario (Weikendorf)

In the Weikendorf measurements [67] used in Chapter 1, the transmitter (Tx) is moving at an approximately constant speed of 5.5 km/h, which is roughly walking speed.The physical layout is given in Figure 4.1.

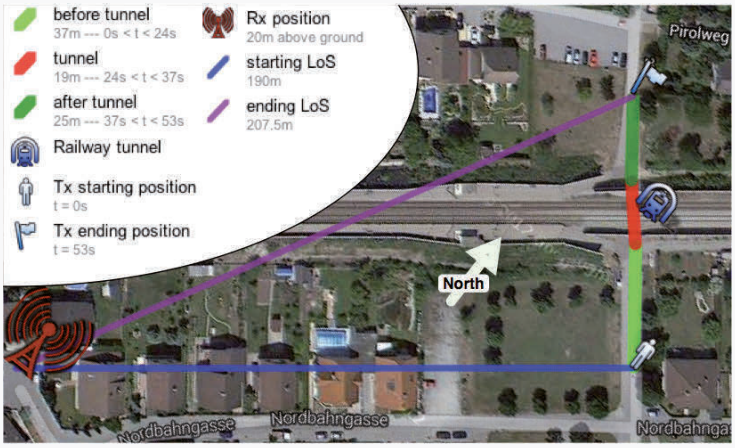


Figure 4.1: *The Weikendorf measurements scenario.*

The other properties of the Weikendorf measurements were previously listed in Table 1.1. Given the 17.5 m difference in the line of sight (LoS) distance between the beginning and the end of the measuring period and the sampling rate of 120 MHz, the time of arrival of the LoS path should be delayed by 7 samples.

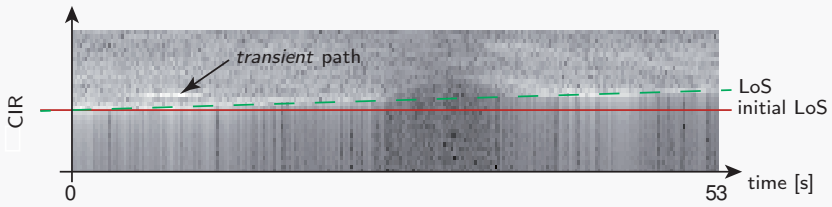


Figure 4.2: *Portion of the CIR from the Weikendorf dataset. The transmitter moves at a pedestrian speed, and each pixel represents a sample. One can see the direct path, always present (except when the transmitter goes through a tunnel around time 30s). Between time 5s and 10s a second path appears suddenly, likely caused by a highly reflective object.*

The constant speed of 5.5 m/s and the Tx movements are consistent with the CIR shown in Figure 4.2. Also in these measurements, a transient path is visible. This can be caused by several factors — e.g. masking by a building.

The benefits of tracking Two aspects must be taken into account. First, tracking may be used to reduce the computational complexity using prior knowledge to initialize parameter values to a likely estimate and then search locally for a more accurate estimation. Second, it may also provide robustness to noise by focusing the estimation on portions of the signal with a higher prior probability of containing paths — these methods are commonly referred to as "warm start optimization".

These considerations point to iterative algorithms with a natural time-domain interpretation, such as the family of *Rake receivers* [89] with a Gauss-Newton optimization for example [121].

In this chapter, we start from the *annihilating filter* property and derive an iterative procedure similar to the *structured TLS*[3] of Lemmerling et al. [77].

This approach leads to two different parametrizations. One of them has a time-domain interpretation which connects it with Rake receivers algorithms. The second parametrization is "*complexified*", thus having twice the number of unknowns. Interestingly, the extra dimensions prove useful to overcome local minima. Comparison of these two interpretations shows the potential benefits of the complexified, annihilating filter based tracking over traditional rake receiver methods.

Tracking does not imply an iterative estimation procedure shall be used, and *vice-versa*; but they are naturally suited for this task when the solution is refined from one iteration to the other.

4.1 Annihilation as a linear constraint

In Chapter 2, it was shown in (2.7) that the *annihilating filter* $\boldsymbol{a} \in \mathbb{C}^{K+1}$ can be computed as the solution of a linear equation

$$[\boldsymbol{T}]_{:,1:K}[\boldsymbol{a}]_{1:K} \;=\; [\boldsymbol{T}]_{:,K+1},$$

with $a_{K+1} = -1$, and $[\boldsymbol{T}]_{:,1:K}$ indexes the first K columns of $\boldsymbol{T}$ (we use the same notation as in Golub & Van Loan [60]).

Initial experiments compared the Least Squares (LS) and the Total Least Squares (TLS) method to solve this system. Unsurprisingly, TLS fared better since both the system coefficients and the objective of the equation are corrupted by noise.

This general rule, has its roots in the formal property

Lemma 4.1. *[82]*
The TLS solution $\boldsymbol{X}_{TLS}$ of the linear system $\boldsymbol{AX} = \boldsymbol{B}$ verifies

$$\{\boldsymbol{X}_{TLS},\ \boldsymbol{\Delta A}_{TLS}, \boldsymbol{\Delta B}_{TLS}\} \;=\; \underset{\boldsymbol{X},\boldsymbol{\Delta A},\boldsymbol{\Delta B}}{\arg\min} \left\|[\boldsymbol{\Delta A}\ \boldsymbol{\Delta B}]\right\|_F \qquad (4.1)$$
$$\text{s.t.}\ \ (\boldsymbol{A}+\boldsymbol{\Delta A})\boldsymbol{X} \;=\; (\boldsymbol{B}+\boldsymbol{\Delta B}).$$

[3] Thanks to F. Quick from Qualcomm Inc., for pointing out the equivalence.

Applied to the computation of the annihilating filter, the TLS method fails to take into account the block-Toeplitz structure of the data-matrix $\boldsymbol{T}$ (the corrections $\boldsymbol{\Delta A}$ and $\boldsymbol{\Delta B}$ should themselves be block-Toeplitz). Usage of Cadzow denoising proved to solve this issue.

Since it is not the annihilation property itself which is at fault for the poor performances, but rather the method used to estimate the annihilating filter; it may be used as a constraint in an optimization problem where we control the objective function :

$$\begin{aligned} \boldsymbol{a}_{\mathsf{opt}} \;=\; \underset{\boldsymbol{a},\boldsymbol{X}}{\arg\min} \quad & \|\boldsymbol{Y}-\boldsymbol{X}\|_F \\ \text{s.t.} \quad & \boldsymbol{a} * [\boldsymbol{Y}]_p \;=\; \boldsymbol{0}, \quad \forall p = 1,\ \ldots,\ P. \end{aligned} \tag{4.2}$$

where $\boldsymbol{Y}$ is a $(2M+1)\times P$ matrix containing the P measurement vectors, "$*$" is the linear convolution without padding, and $\|\boldsymbol{Z}\|_F = \mathrm{Tr}\{\boldsymbol{Z}^*\boldsymbol{Z}\}$ is the *Frobenius norm*.

The solution of 4.2 is the set of P signals having a common annihilating filter of length $K+1$ which is the closest to the original measurements vectors with respect to the Frobenius norm[4].

The optimization (4.2) decomposes in two subproblems

$$\begin{aligned} (4.2) \Leftrightarrow \qquad \boldsymbol{a}_{\mathsf{opt}} \;=\; \underset{\boldsymbol{a}}{\arg\min}\ \min_{\boldsymbol{X}} & \|\boldsymbol{Y}-\boldsymbol{X}\|_F \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{X} \;=\; \boldsymbol{0}, \quad \forall p = 1,\ \ldots,\ P, \end{aligned} \tag{4.3}$$

where $\boldsymbol{A} = \mathrm{Toeplitz}(\boldsymbol{a})$.

The minimization with respect to $\boldsymbol{X}$ is a linear optimization problem, which can be solved in closed form using *Lagrange's multipliers* [34], and thus

Proposition 4.1. *The optimization problem (4.2) is equivalent to the non-linear Least-Squares problem*

$$a_{\mathit{opt}} \;=\; \underset{\boldsymbol{a}}{\arg\min} \|f(\boldsymbol{a})\|_2^2 . \tag{4.4}$$

where

$$f(\boldsymbol{a}) \;=\; \mathsf{vec}\left(\boldsymbol{A}^\dagger \boldsymbol{A}\boldsymbol{Y}\right), \quad \boldsymbol{A} = \mathrm{Toeplitz}(\boldsymbol{a}),$$

and vec *is the linear operation transforming a matrix in a column-vector by stacking-up its columns.*

Non-linear least-squares programs do not have a closed form solution in general and are tackled with iterative methods [56]. Another approach would be to use Bresler's IQML method [37]. It is not consistent in itself, but it can be made so,

[4]The total error is the sum of the ℓ_2 norm of the error made on each channel measurements. A weighted 2-norm can also be used with minimal modifications.

using a prior denoising step on $\boldsymbol{Y}$ to make it low-rank[120]. The resulting algorithm is called MODE[5] (*Method Of Direction Estimation*).

Iterative methods require an initial estimate, and are therefore particularly suited for tracking.

4.2 Two parametrizations

In the previous section, we eluded an important question

> "Over which domain is $\boldsymbol{a}$ defined in (4.4)?"

First, notice that the optimization (4.4) is invariant with the scale of the annihilating filter. Hence, as in Chapter 2, we set a_{K+1} a priori, which leaves us with K complex-valued unknowns. We face the following dilemma in the definition of the remaining K coefficients. They could be

- Coefficients of degree K polynomials :

  $$[\boldsymbol{a}]_{1:K} \in \mathbb{C}^K.$$

  The search space has dimension $2K$ since every a_k has a real and an imaginary part.

- Coefficients of degree K polynomials with unit modulus roots[6] :

  $$[\boldsymbol{a}]_{1:K} \in \mathbb{C}^K \cap \left\{ \boldsymbol{a} \,\middle|\, |z_k| = 1,\ \boldsymbol{z} = \mathsf{roots}(\boldsymbol{a}) \right\}.$$

  The search space has dimension K, the unknowns are the phases of the roots.

These two options are made explicit by re-parametrizing the optimization (4.4) in terms of the modulus and phase of the annihilating filter's roots

$$\underset{\substack{\omega_1,\ldots,\omega_K \in [-\pi,\pi[\\ r_1,\ldots,r_K \ \in \ \mathbb{R}^+ \textbf{or } \{1\}}}{\arg\min} \left\| f([\mathsf{poly}(r_1 e^{j\omega_1},\ \ldots,\ r_K e^{j\omega_K}),\ 1]) \right\|_2^2, \tag{4.5}$$

where the function poly computes a polynomial's coefficients from its roots. The definition of the coefficients $r_1,\ \ldots,\ r_K$ over $\mathbb{R}^+$ or trivially over the singleton $\{1\}$ yields the two options mentioned before.

[5] We would like to thank Pr. Ottersten for mentioning both IQML and MODE to us.

[6] Polynomial with roots on the unit-circle have hermitian symmetric coefficients, which may be exploited before the rooting operation. We did not investigated in this direction. Thanks to Pr. Ottersten for the suggestion to use this property.

4.2.1 Time-domain interpretation for $r_1 = \cdots = r_K = 1$

The non-linear function f is thus simply the orthogonal projection of the measurements $\boldsymbol{Y}$ into range $(\boldsymbol{A}^*)$ — the *row-space* of $\boldsymbol{A}$ — put in vector form.

The row-space of $\boldsymbol{A}$ contains signals which are linear combinations of eigenfunctions of the filtering by $\boldsymbol{a}$ operation

$$\boldsymbol{x} \in \text{range}\,(\boldsymbol{A}^*) \qquad \Leftrightarrow \qquad \boldsymbol{x} \,=\, \sum_{\ell} \alpha_\ell \boldsymbol{x}_\ell \;,$$

$$\text{s.t. } \boldsymbol{x}_\ell * \boldsymbol{a} \,=\, \lambda_\ell \cdot \boldsymbol{x}_\ell \;, \quad \lambda_\ell \neq 0 \;, \quad \text{and } \boldsymbol{x}_1,\, \boldsymbol{x}_2,\, \ldots \text{ are linearly independent.}$$

Basic algebra tells us that orthogonal projection into the kernel of a linear operator is the orthogonal complement of the projection into its row-space

$$\text{proj}_{\ker(\boldsymbol{A})} \,=\, \mathbb{I} - \text{proj}_{\text{range}(\boldsymbol{A}^*)}.$$

We know the kernel of $\boldsymbol{A}$ from the definition of the annihilating filter itself : it is spanned by complex exponentials which radices are the roots of the annihilating filter.

The orthogonal projection into the kernel of $\boldsymbol{A}$ is thus simply the orthogonal projection into the subspace spanned by these complex exponentials. In the time-domain, it is the orthogonal projection into the subspace spanned by Dirichlet kernels (of corresponding bandwidth) shifted by the phase of the roots of $\boldsymbol{a}$.

Therefore, the function f can be interpreted in the time-domain as the residual obtained after correlation with the path estimates. This interpretation is similar to the one found in the family of *Rake receiver* algorithms, in which the energy of the residual is minimized by updating the paths estimates — "moving the *fingers* of a *rake*".

4.2.2 Dimensionality : a curse or a blessing?

The optimization on the K dimensional space has a satisfying time-domain interpretation, why would one want to perform it in a space twice as large? — usually the number of local minima in a non-linear optimization problem increases with the dimension of the space to be searched. As a simple counter-example, for $K = 1$, the number of *critical points* does not decrease when r — the modulus of the annihilating filter root — is added as an optimization variable, but the nature of these critical points changes in such a way that the number of local-minima seems to reduce down to one, providing sure convergence for second-order methods [92].

From minima to saddles The unusual potential benefit of additional dimensions is illustrated in Figure 4.3. This figure shows for $K = 1$ the value of the objective function as a function of the root $re^{j\omega}$.

By inspection on Figure 4.3, the projection of this curve on the unit-circle has $N-1$ local minima. Using the time-domain interpretation, these local minima simply correspond to the shifts for which the side-lobes of a Dirichlet kernel align with the path ToA. Local minima pose a serious threat to any descent algorithms, as they may become the point of convergence.

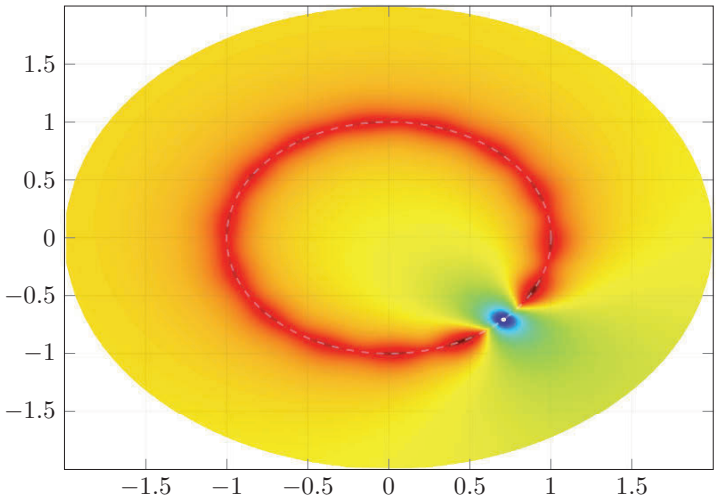


Figure 4.3: *LS residual after annihilation on* 15 *DFT samples of a single phasor of radix* $e^{-j\pi/4}$.

Example 4.b — Avoiding local minima with overparametrization

To verify the intuition developed in Figure 4.3, we perform a simple tracking test on a signal containing only one path which ToA (vertical axis) follows a sampled Brownian motion :

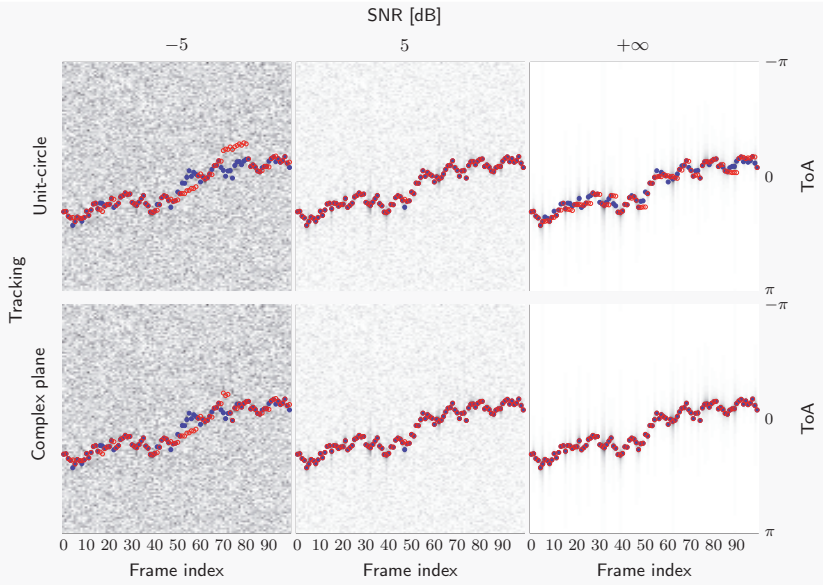


Figure 4.4: *Tracking of a single path in AWGN. The correct ToA is indicated by • and its tracking estimation by ◦.*

The top-row shows tracking of a single-pulse using Levenberg-Marquardt algorithm (details in Section 4.2.3) over a unidimensional space — *i.e.* considering a unique annihilating filter root of unit-modulus. In the second-row, the modulus of the root is considered as an unknown augmenting the dimensionality of the problem by one. This additional dimension yields a more robust estimation at very high SNR ($+\infty$) and, surprisingly, at low SNR (-5dB). The visible improvement is that the algorithm recovers from a mistracking quickly in the 2-dimensional space, which could be explained by the ability to "move around hills" in the objective function thanks to the fact that its level-sets in $\mathbb{R}^2$ are *path-connected* [90] while they are not in $\mathbb{R}$.

On the other hand, the same curve in Figure 4.3 considered over the disk of radius 2 shows a single local minimum which corresponds to the true parameter values. What appeared as local minima on the unit-circle are now *saddle points*. Second order descent methods such as Newton-Raphson method or Gauss-Newton method converge surely to a minimum even in the presence of saddle points. We study the critical points of the objective function for $K = 1$ in Appendix D.1. By a simple inspection of Figure 4.3, the objective function seems to have a single minimum.

The existence of local minima when the parameter space is restricted to the unit-circle prevents tracking from working properly in the noiseless case — as seen in Figure 4.4. When the parameter space is enlarged to the whole complex-plane, convergence to the true time of arrival is achieved.

Regarding local minima, the perturbation introduced by the noise proves to be beneficial, and this can be seen in Figure 4.4 where the performances are better for an SNR of 5dB than for an infinite SNR. This leads to the following constatation

- The noise permits to overcome the shallow local minima when optimization is

done over the unit-circle,

- Slow variations of the ToAs ensures starting from the same attraction basin from one frame to the next, making local minima less threatening to convergence,

which, for a specific application, leaves open the question of whether or not the optimization in the larger space is to be preferred or not.

Time-domain interpretation of the overparametrization To the best of our knowledge, this overparametrization does not correspond to any known algorithm in the time-domain. Indeed, varying the magnitude of the roots modulus varies the decay rate of the pulse shape in the time-domain [11] which does not lend itself to an interpretation similar to the rake receiver or other correlation based methods.

4.2.3 Solving the minimization problem

Non-linear Least-Square problems are non-convex. It is quite obvious that the objective in the channel estimation problem possesses several local minima in general, *e.g.* shifts corresponding to the side-lobes of the pulse autocorrelation function are local minima.

Numerical methods for non-linear least-squares optimization fall into two categories for the most part :

- **Differential methods**, such as the Gauss-Newton algorithm, Levenberg-Marquardt algorithm, Gradient descent, Conjugate gradients descent, ...
- **Simulation based methods** Monte-Carlo simulations, simulated annealing, particle swarm methods, Tabu search, ...

Differential methods evaluate the topography of the objective function around a current estimate, and move it towards a promising direction. This operation is repeated until a local minimum (or a saddle point for some methods) is reached.

Simulations methods follow a stochastic process to search for an optimal solution. The probability measure of the process evolves towards a point-mass over time. The initial randomness of simulation methods make them less sensitive to local minima than differential methods.

Ideally, signal tracking assumes a good estimate of the global optimum is known a priori. If so, *differential methods* may converge to the global optimum — if not, the methods developed in Chapter 2 are indicated. For this reason, we do not include *simulation algorithms* in this study.

Levenberg-Marquardt algorithm The *Levenberg-Marquardt algorithm* [78; 83] can be seen as a Gauss-Newton method regularized by a gradient descent term. A tuning parameter allows to adjust the importance of the Gauss-Newton term relative to the gradient-descent term. The goal is to provide a short overview of the algorithm in order to understand the whereabouts of solving (4.5) numerically; and it is by no mean a comprehensive study of the subject, which can be found in optimization textbooks [56; 92].

As a short reminder, the *Gauss-Newton* method tries to minimize a quadratic norm $\|\boldsymbol{g}(\boldsymbol{\theta})\|^2$, where $\boldsymbol{g}$ is a real-valued multidimensional function of a real-valued vector of parameters $\boldsymbol{\theta}$. For example, we would cast the optimization (4.5) in this framework by taking $\boldsymbol{g}$ to be the concatenation of the real and imaginary parts of $\boldsymbol{f}$ and $\boldsymbol{\theta} = [\omega_1, \ldots, \omega_K]^T$ or $\boldsymbol{\theta} = [\omega_1, \ldots, \omega_K, r_1, \ldots, r_K]^T$ — whether the roots of the annihilating polynomial are restricted to have unit-modulus or not.

Using the multidimensional Taylor expansion of $\boldsymbol{g}$, we write

$$\boldsymbol{g}(\boldsymbol{\theta} + \boldsymbol{\Delta}) = g(\boldsymbol{\theta}) - \boldsymbol{J}\boldsymbol{\Delta} + \mathcal{O}(\|\boldsymbol{\Delta}\|^2) \cdot \mathbf{1} ,$$

where $\boldsymbol{J}$ is the Jacobian matrix of $\boldsymbol{g}$, *i.e.* $[\boldsymbol{J}]_{m,n} = \frac{\partial g_m(\boldsymbol{\theta})}{\partial \theta_n}$.

Then, for $\boldsymbol{\Delta}$ small, one can make the first order approximation

$$\|\boldsymbol{g}(\boldsymbol{\theta} + \boldsymbol{\Delta})\|^2 = \|\boldsymbol{g}(\boldsymbol{\theta})\|^2 - 2\boldsymbol{g}(\boldsymbol{\theta})^T\boldsymbol{J}\boldsymbol{\Delta} + \boldsymbol{\Delta}^T\boldsymbol{J}^T\boldsymbol{J}\boldsymbol{\Delta} . \tag{4.6}$$

The derivative of the error with respect to $\boldsymbol{\Delta}$ should vanish when a (bounded) optimum is reached. Thus, setting the derivative of (4.6) to $\mathbf{0}$ yields the Gauss-Newton *normal equation*

$$\boldsymbol{J}^T\boldsymbol{J}\boldsymbol{\Delta} = \boldsymbol{J}^T\boldsymbol{g}(\boldsymbol{\theta}) ,$$

which can be efficiently solved using Choleski or QR type factorizations.

A regularization term is added to this normal equation in order to obtain

$$\left(\boldsymbol{J}^T\boldsymbol{J} + \lambda \cdot \mathrm{diag}(\boldsymbol{D}^T\boldsymbol{D})\right)\boldsymbol{\Delta} = \boldsymbol{J}^T\boldsymbol{g}(\boldsymbol{\theta}) , \quad \lambda \geq 0, \tag{4.7}$$

which is the update equation of the *Levenberg-Marquardt* algorithm.

Scaling the regularization term by $\mathrm{diag}(\boldsymbol{D}^T\boldsymbol{D})$ in (4.7) ensures it is scaled proportionally with the gradient of $\boldsymbol{g}$ for each of its dimensions.

For a small value of λ, the algorithm behaves like the Gauss-Newton algorithm, which is preferable if the error decayed rapidly in the previous iterations. For a large value of λ, the algorithm behaves like a weighted gradient descent, which is to be preferred if the error decayed slowly in the previous iterations. The tuning of λ dynamically during the optimization process is a well-studied topic, and we point to [92] for a review.

4.3 Detection for tracking : Update, Validate and Add

One of the strength of tracking is the ability to recover paths which have faded enough not to be accurately recovered by a memoryless estimation procedure, but not so much so that their recovery with a locally focused algorithm is still beneficial for equalization.

The issues introduced by the exploitation of a temporal correlation are two folds

1. On which grounds should a tracked path be discarded from the model?
2. On which grounds should a new path be added to the model?

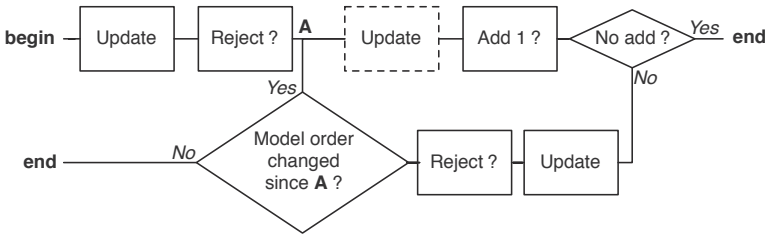


Figure 4.5: *A combination of Update/Reject/Add steps for a dynamical tracking of a SCS signal. A dashed box indicates the operation is optional. Different configurations are realizable.*

These two questions are *detection related* and we propose to adjust the criteria developed in Chapter 3 to cope with them. A possible solution is to split the detection in two stages which consist in first rejecting spurious paths followed by the addition of a new path if necessary. Each of these stages can be preceded and/or followed by an update of the estimate.

Updating an estimate From an initial estimate $\widehat{\boldsymbol{a}}_0$ of the annihilating filter coefficients, apply *Levenberg-Marquardt* algorithm using the evolution equation (4.7) until convergence.

Validation of a path To reject a potentially spurious detection a simple test is to estimate the amplitudes for each path in each channel and reject if observing greater or equal magnitudes can be attributed solely to the noise with a probability exceeding a predefined false positive rate α_{reject}. In the presence of AWGN, this procedure was shown in Proposition 3.3. To have robustness to a temporary fading of a path, one could use the weighted sum of the estimated path energy over time — e.g. average the path energy of the current and previous frame. This provides a trade-off between being robust to temporary fading and adapting quickly to a changing CIR.

Addition of a path After an estimate $\widehat{\boldsymbol{a}}_{\mathsf{opt}}$ of the solution of (4.4) is obtained, one is left with the residual error $f(\widehat{\boldsymbol{a}}_{\mathsf{opt}})$ (f is defined in Proposition 4.1).

The inspection of this residual may provide valuable information about the adequacy of the estimated model order. In Appendix D.2 a statistical test is outlined.

A more straightforward approach is to increase the model order by a predefined number and process the result with the validation algorithm.

The addition of a single path can be done by finding the maximum correlation between the residual and the pulse shape for different shifts (extremum of the matched filter output).

Summary of the tracking process The complete tracking procedure is a succession of "Update", "Reject" and "Add" steps. Figure 4.5 shows a potential flow graph.

4.4 Numerical results

Preliminary study Previously in this chapter, we used the Levenberg-Marquardt algorithm[7] to solve (4.5) on the unit-circle and on the complex-plane.

The plots in Figure 4.4 showed the tracking of a single path which trajectory is a sampled Brownian motion.

The existence of local minima when the parameter space is restricted to the unit-circle prevents tracking from working properly in the noiseless case. When the parameter space is enlarged to the whole complex-plane, convergence to the true time of arrival is achieved. This simple observation confirms the conclusions made in Section 4.2.2.

Regarding local minima, the perturbation introduced by the noise proves to be beneficial, and this can be seen in Figure 4.4 where the performances are better for an SNR of 5dB than for an infinite SNR. This leads to the following constatation

- The noise permits to overcome the shallow local minima when optimization is done over the unit-circle,
- Slow variations of the ToAs ensures starting from the same attraction basin from one frame to the next, making local minima less threatening to convergence,

which, for a specific application, leaves open the question of whether or not the optimization in the larger space is to be preferred or not.

Tracking test We simulate a Rayleigh fading[8] set of $P = 4$ SCS channels. The tracking algorithm follows the simple rule

$$\text{Update} \;\rightarrow\; \text{Validate} \;\rightarrow\; \text{Add } 1 \;\rightarrow\; \text{Validate}.$$

Validation is based on a threshold which targets a false detection rate of magnitude close to $1/2$ — see Proposition 3.3.

Each channel has $K = 5$ paths of amplitude [1, 0.5, 1, 0.7, 0.4]. The paths ToA either drift by a constant amount from one frame to the other or vary according to a sinusoidal trajectory as shown in Figure 4.6.

The results are shown and commented in Figures 4.7–4.8.

4.5 Conclusion

We studied jointly iterative estimation and tracking, even so one does not reduce to the other. Iterative estimation was derived from the annihilating equation. It was shown that the modification of the error measure in the annihilating equation leads to a more robust yet non-linear estimation. A connection with correlation methods such as the Rake receiver was made, and it suggests that an overparametrization may help to overcome some local minima encountered by these methods. Numerical simulations

[7] Namely, the implementation provided by the MATLAB (R2012a) function `lsqnonlin()`.

[8] The correlation between antennas was set to $J_0(\pi) \approx -0.3$ (narrow scatterers, see (1.4) in Chapter 1).

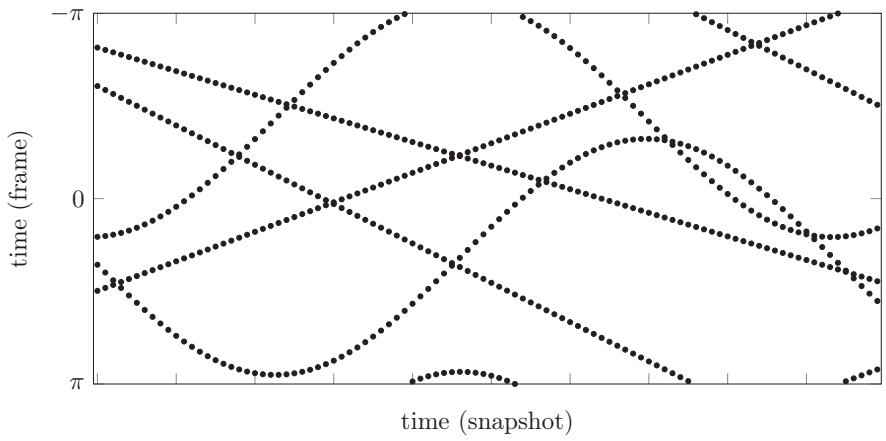


Figure 4.6: *The evolution of five paths over time. Each path follows a linear or a sinusoidal trajectory. The horizontal axis represents snapshots (successive OFDM frames), and the vertical axis represents the period over which is recorded each frame. The data points represent the time of arrival of each path.*

showed that at low SNR, the local view of a tracking algorithm made the detection and estimation more robust, as they happened in streaks, despite the independence of noise and fading between frames. The incorporation of a global detection/estimation step and of a validation step into the algorithm provided the responsiveness required by transient nature of mobile communications channels.

The subject of channel tracking is vast, and this chapter only considered a few of its aspects, principally to show how the methods developed in Chapter 2 and Chapter 3 naturally apply to this subject.

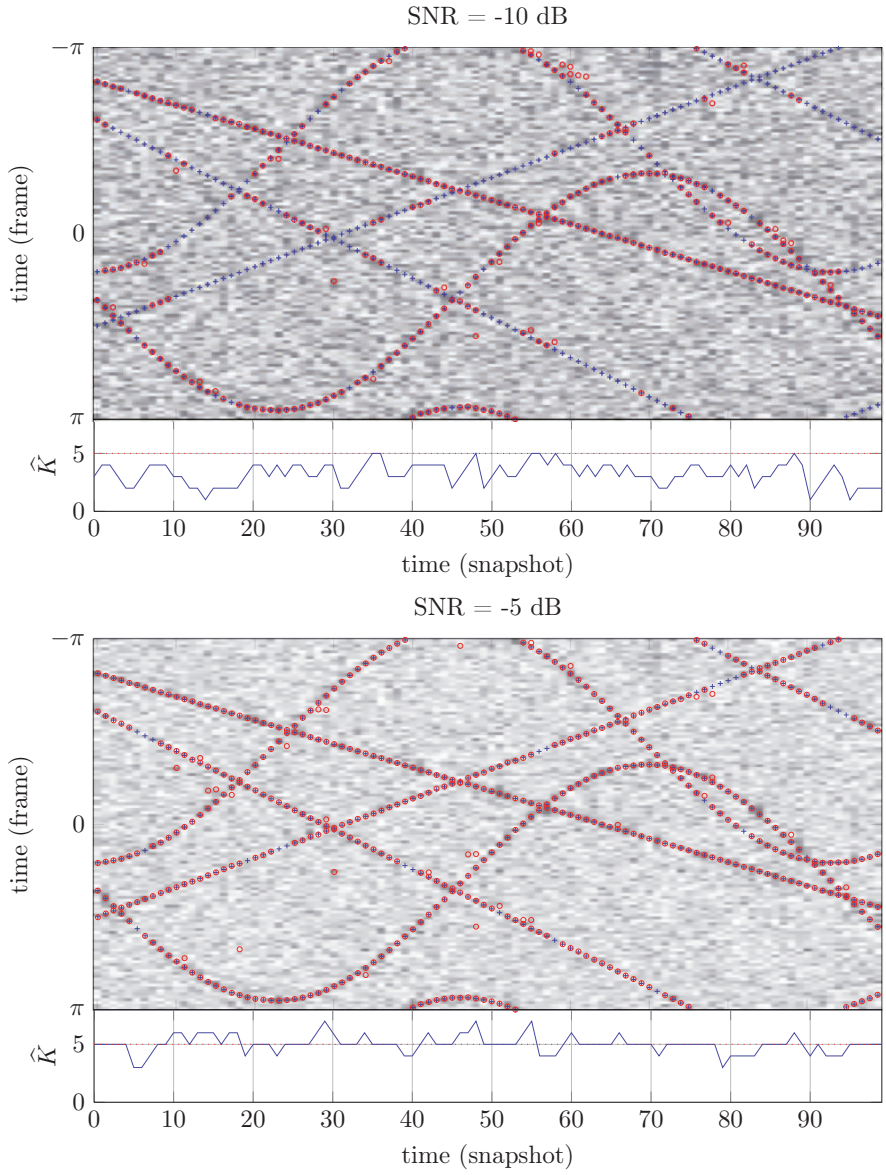


Figure 4.7: *Tracking at a lower SNR. During validation, a path which energy is below the threshold is discarded only if its energy average with the previous frame is below the threshold. Paths are paired from one frame to the next if they are closer than four times the inverse bandwidth of the pulse shape — i.e. pairing is done over a distance which can be interpreted as "tracking". This simple "Markov-chain" type of dependency already introduces robustness against fading as correct estimations happen in streaks (visible on the top panel) despite the fact that the paths energy are independently distributed over time. The fact that only one path can be added from one frame to another is not optimal and could be iterated as proposed in 4.5.*

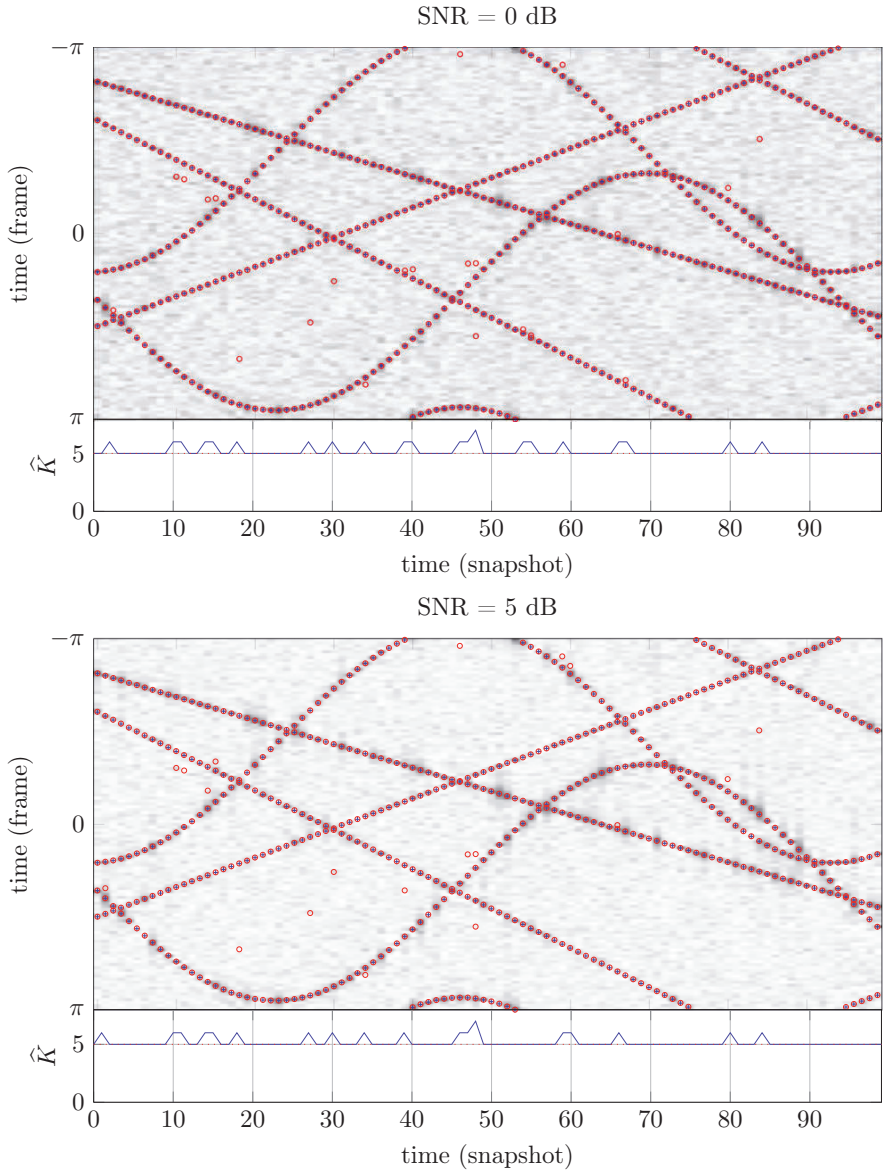
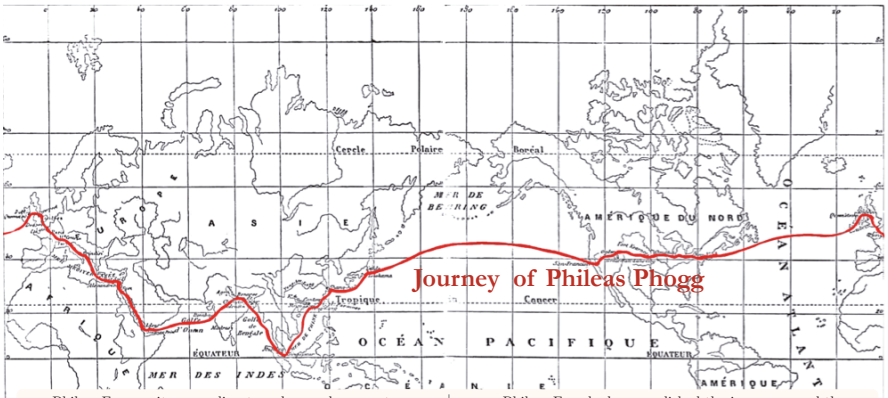


Figure 4.8: *Tracking at higher SNR. The noise is low enough so that the* 5 *paths are always detected. The proportion of spurious detections does not change between* 0 *dB and* 5 *dB. This is to be expected since we chose the threshold such as to yield a constant false detection rate. This could be mitigated by introducing a notion of dynamic range, which necessitates not only the knowledge of* σ^2 *but also some information about the distribution of the paths energy. We did not introduced this notion. The observed false detection rate is approximately* 0.25.

Part II

Fundamental Limits on Periodic Localization



“Phileas Fogg avait accompli ce tour du monde en quatre-vingts jours !...

Phileas Fogg avait gagné son pari de vingt mille livres !

Et maintenant, comment un homme si exact, si méticuleux, avait-il pu commettre cette erreur de jour ? Comment se croyait-il au samedi soir, 21 décembre, quand il débarqua à Londres, alors qu'il n'était qu'au vendredi, 20 décembre, soixante dix neuf jours seulement après son départ ?

Voici la raison de cette erreur. Elle est fort simple.

Phileas Fogg avait, « sans s'en douter », gagné un jour sur son itinéraire, – et cela uniquement parce qu'il avait fait le tour du monde en allant vers l'est, et il eût, au contraire, perdu ce jour en allant en sens inverse, soit vers l'ouest.

En effet, en marchant vers l'est, Phileas Fogg allait au-devant du soleil, et, par conséquent les jours diminuaient pour lui d'autant de fois quatre minutes qu'il franchissait de degrés dans cette direction. Or, on compte trois cent soixante degrés sur la circonférence terrestre, et ces trois cent soixante degrés, multipliés par quatre minutes, donnent précisément vingt-quatre heures, – c'est-à-dire ce jour inconsciemment gagné. En d'autres termes, pendant que Phileas Fogg, marchant vers l'est, voyait le soleil passer quatre-vingts fois au méridien, ses collègues restés à Londres ne le voyaient passer que soixante-dix-neuf fois. C'est pourquoi, ce jour-là même, qui était le samedi et non le dimanche, comme le croyait Mr. Fogg, ceux-ci l'attendaient dans le salon du Reform-Club.”

Jules Verne — Le tour du monde en 80 jours.

“Phileas Fogg had accomplished the journey round the world in eighty days!

Phileas Fogg had won his wager of twenty thousand pounds!

How was it that a man so exact and fastidious could have made this error of a day? How came he to think that he had arrived in London on Saturday, the twenty-first day of December, when it was really Friday, the twentieth, the seventy-ninth day only from his departure?

The cause of the error is very simple.

Phileas Fogg had, without suspecting it, gained one day on his journey, and this merely because he had travelled constantly eastward; he would, on the contrary, have lost a day had he gone in the opposite direction, that is, westward.

In journeying eastward he had gone towards the sun, and the days therefore diminished for him as many times four minutes as he crossed degrees in this direction. There are three hundred and sixty degrees on the circumference of the earth; and these three hundred and sixty degrees, multiplied by four minutes, gives precisely twenty-four hours—that is, the day unconsciously gained. In other words, while Phileas Fogg, going eastward, saw the sun pass the meridian eighty times, his friends in London only saw it pass the meridian seventy-nine times. This is why they awaited him at the Reform Club on Saturday, and not Sunday, as Mr. Fogg thought.”

Jules Verne — Around the world in 80 days.

Chapter 5

Time-Frequency localization in periodic domains

5.1 "Where?" — from linear to periodic

The estimation of periodic parameters leads to estimators with periodic distributions. Localization of periodic phenomena is not straightforward. Interpreting them in a linear way by "unwrapping" them is bound to create confusion. To get a taste of these issues, we can simply look at world maps. As we can see in Figure 5.1, everyone wants to be the center of the world and one's view of the world will be quite different if one is japanese, french or american — or even australian for a more disturbing 'upside-down" view[1].

What is lost there is the perfect symmetry of the sphere, one has to choose a reference point from which to elaborate the desired projection. One of the issue, is the appearance of *boundaries* — e.g. in Figure 5.1, the Pacific ocean or the Atlantic ocean are cut in half and the fact that different countries use different cuts shows that this boundary is rather artificial.

A proper periodic localization should not presuppose any reference point, and is similar— by analogy — to the computerized mapping tools where the cursor is the reference rather than a fixed point.

The matter of this chapter is joint-work with R. Parhizkar, with elements from [94].

[1] Fortunately, the rotation axis of the earth made it so that no one proposed a "slanted" map (with a vertical axis pointing in the south-east direction for example)... yet.

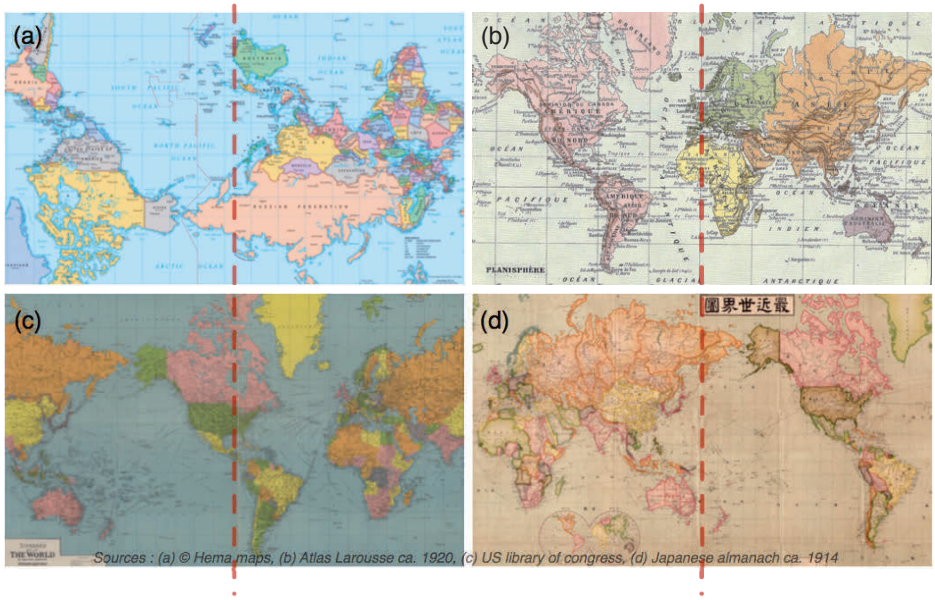
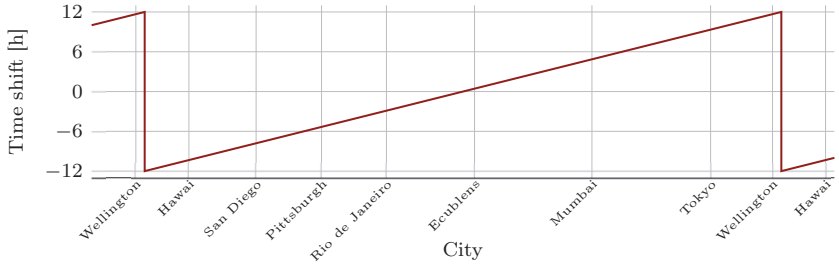


Figure 5.1: *The world as it is seen from (a) Australia, (b) France, (c) the USA or (d) Japan. The dashed line is the vertical centered fold of the map.*

The confusion may turns into an analytical nightmare. For example, if one takes the Greenwich meridian as a reference and a linear time shift (the phase of the principal angle, it is an approximation of the time-zone system), a rather troublesome discontinuity appears somewhere in the pacific ocean[2] :



If one can perfectly accomodate with that for travelling purposes, it causes some serious problems to differential calculus, which we will see later.

Keeping these considerations in mind, we can start the study of a fundamental property on the localization of periodic functions (waveforms) : the relationship between their spread — the variance of the localization — and the total energy of their variations — the variance of the momentum. For non-periodic functions, the most

[2]This discontinuity is not caused by taking a fixed reference point but rather by the way a relative position is computed

classical result is the *Heisenberg uncertainty principle* [66], which lowerbounds the product of these two variances, and there exists an extension to periodic waveforms [35], which we will study carefully.

The question can also be posed only in term of "spread" in the time and frequency domains — the time-domain momentum being the localization in the frequency domain and the mapping from one to the other being unitary[3]. This time-frequency duality also allows us to look at the problem from the perspective of sequences, since periodic functions can be seen as the discrete time Fourier transform (DTFT) of sequences. So, by convention, we choose the frequency domain to be periodic and the time domain to be discrete.

We adopt this point of view since its discrete nature of sequences lends itself better to numerical analysis.

In opposition to sequences, the notions of time and frequency spreads are well defined and established for continuous-time signals [66; 134] and their properties are studied thoroughly in the literature. For a continuous-time signal, we can define the time and frequency characteristics of a signal as in Table 5.1. Note the connection of these definitions with the mean and variance of a probability distribution function $|x(t)|^2 / \|x\|^2$. The value of Δ_t^2 is considered as the spread of the signal in the time domain while $\Delta_{\omega_c}^2$ represents its spread in the frequency domain. We say that a signal is compact in time (or frequency) if it has a small time (or frequency) spread.

The Heisenberg uncertainty principle [66] states that continuous-time signals cannot be arbitrarily compact in both domains. Specifically, for any $x(t) \in L^2(\mathbb{R})$,

$$\eta_c = \Delta_t^2 \, \Delta_{\omega_c}^2 \geq \frac{1}{4}, \tag{5.1}$$

where the lower bound is achieved for Gaussian signals of the form $x(t) = \gamma e^{-\alpha t^2}$, $\alpha > 0$ [134]. The subscript c stands for continuous-time definitions. We call η_c the *time-frequency spread* of x.

Although the continuous/non-periodic Heisenberg uncertainty principle is widely used in theory, in practice we often work with discrete-time signals (e.g. filters and wavelets), or periodic signals. Thus, equivalent definitions for discrete-time sequences and their periodic spectra are needed in signal processing. In the next section we study two common definitions of center and spread available in the literature.

Note that we consider periodic, *analog* spectra (continuous frequency domain), *i.e.* spectra of infinite sequences. For discrete spectra Donoho & Stark studied the uncertainty linked to the ℓ_0-norm, [84] studied it with respect to the ℓ_2-norm and [102] with and information measure (entropy).

For infinite sequences and their periodic analog spectra, results can be found in [35] and [136]. The most comprehensive work on the uncertainty relations for discrete sequences is found in [100]. The authors show that 1/4 is a lowerbound on the time-frequency spread, which can only be achieved asymptotically as the sequence spreads in time. With our approach, we will obtain *constructively* the achievable lowerbound on the time-frequency spread for any given frequency spread.

[3] One should normalize the Fourier transform properly.

domain	center	spread
time	$\mu_t = \frac{1}{\|x\|^2}\int_{t\in\mathbb{R}} t\|x(t)\|^2 dt$	$\Delta_t^2 = \frac{1}{\|x\|^2}\int_{t\in\mathbb{R}}(t-\mu_t)^2\|x(t)\|^2 dt$
frequency	$\mu_{\omega_c} = \frac{1}{2\pi\|x\|^2}\int_{\omega\in\mathbb{R}}\omega\|X(\omega)\|^2 d\omega$	$\Delta_{\omega_c}^2 = \frac{1}{2\pi\|x\|^2}\int_{\omega\in\mathbb{R}}(\omega-\mu_\omega)^2\|X(\omega)\|^2 d\omega$

Table 5.1: *Time and frequency centers and spreads for a continuous time signal $x(t)$.*

domain	center	spread
time	$\mu_n = \frac{1}{\|x\|^2}\sum_{n\in\mathbb{Z}} n\|x_n\|^2$	$\Delta_n^2 = \frac{1}{\|x\|^2}\sum_{n\in\mathbb{Z}}(n-\mu_n)^2\|x_n\|^2$
frequency	$\mu_{\omega_\ell} = \frac{1}{2\pi\|x\|^2}\int_{-\pi}^{\pi}\omega\|X(e^{j\omega})\|^2 d\omega$	$\Delta_{\omega_\ell}^2 = \frac{1}{2\pi\|x\|^2}\int_{-\pi}^{\pi}(\omega-\mu_\omega)^2\|X(e^{j\omega})\|^2 d\omega$

Table 5.2: *Time and frequency centers and spreads for a discrete time signal x_n as extensions of Table 5.1 [134].*

5.1.1 Uncertainty principles for periodic waveforms and sequences

An obvious and intuitive extension of the definitions in Table 5.1 for discrete-time signals is presented in Table 5.2, where

$$X(e^{j\omega}) = \sum_{n\in\mathbb{Z}} x_n e^{-j\omega n} \quad \omega\in\mathbb{R}\,, \tag{5.2}$$

is the discrete-time Fourier transform (DTFT) of x_n.

Using the definitions [134] in Table 5.2, we can also state the Heisenberg uncertainty principle for discrete-time signals. Under the condition $X(e^{j\pi}) = 0$, the following holds

$$\eta_\ell = \Delta_n^2\,\Delta_{\omega_\ell}^2 > \frac{1}{4}\,, \quad x_n\in\ell^2(\mathbb{Z})\,, \tag{5.3}$$

where the subscript ℓ stands for *linear* in reference to the definition of the frequency spread. Note the extra assumption on the Fourier transform of the signal. This assumption is necessary for the result to hold.

Example 5.a — Beating the $\frac{1}{4}$ barrier

Take $x_n = \delta_n + 7\delta_{n-1} + 2\delta_{n-2}$. The Fourier transform of x_n is shown in Figure 5.2. Observe that $\left|X(e^{j\pi})\right| = 0.22 \neq 0$, which violates the condition $X(e^{j\pi}) = 0$. The linear time-frequency spread of this signal according to Table 5.2 is $\eta_\ell = 0.159 < 1/4$.

The dependency of the uncertainty limit on the value of the waveform at $e^{j\pi}$ is a consequence of the discontinuity of the "saw-tooth" function $\widetilde{\omega}$ which is used for localization in the periodic frequency domain. Because of this restriction on the

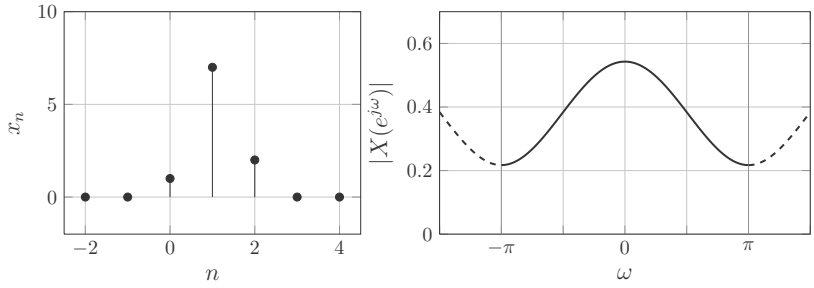


Figure 5.2: *A signal that violates $X(e^{j\pi}) = 0$ and does not satisfy the linear Heisenberg uncertainty principle ($\eta_\ell = 0.159 < \frac{1}{4}$).*

domain	**center**	**spread**
time	$\mu_n = \frac{1}{\|x\|^2} \sum_{n\in\mathbb{Z}} n\|x_n\|^2$	$\Delta_n^2 = \frac{1}{\|x\|^2} \sum_{n\in\mathbb{Z}} (n-\mu_n)^2\|x_n\|^2$
frequency	$\mu_{\omega_p} = 1 - \tau(x)$	$\Delta_{\omega_p}^2 = \frac{1-\|\tau(x)\|^2}{\|\tau(x)\|^2} = \left\|\frac{\|x\|^2}{\sum_{n\in\mathbb{Z}} x_n\, x_{n+1}^*}\right\|^2 - 1$

Table 5.3: *Time and frequency centers and spreads for a discrete time signal x_n using circular moments, where $\tau(x)$ is defined in* (5.4).

applicability of the Heisenberg uncertainty principle, the definitions in Table 5.2 do not fully capture the periodic nature of $X(e^{j\omega})$ for the frequency center and spread. In the search for more natural properties, we can adopt definitions for circular moments widely used in quantum mechanics [35] and directional statistics [81]; and we will study and motivate these definitions in the next section.

For a sequence x_n, $n \in \mathbb{Z}$, with a 2π periodic DTFT, $X(e^{j\omega})$ as in (5.2), the *first trigonometric moment* is defined as [99; 100]

$$\tau(x) = \frac{1}{2\pi \|x\|^2} \int_{-\pi}^{\pi} e^{j\omega} |X(e^{j\omega})|^2 d\omega \,. \tag{5.4}$$

The first trigonometric moment was originally defined for probability distributions on a circle. With proper normalization, this definition applies also to periodic functions.

Using (5.4), the *periodic frequency spread* is defined as [35]:

$$\Delta_{\omega_p}^2 = \frac{1 - |\tau(x)|^2}{|\tau(x)|^2} = \left| \frac{\|x\|^2}{\sum_{n\in\mathbb{Z}} x_n\, x_{n+1}^*} \right|^2 - 1 \,. \tag{5.5}$$

The definition of Δ_n^2 remains unchanged as in Table 5.2. These definitions are summarized in Table 5.3.

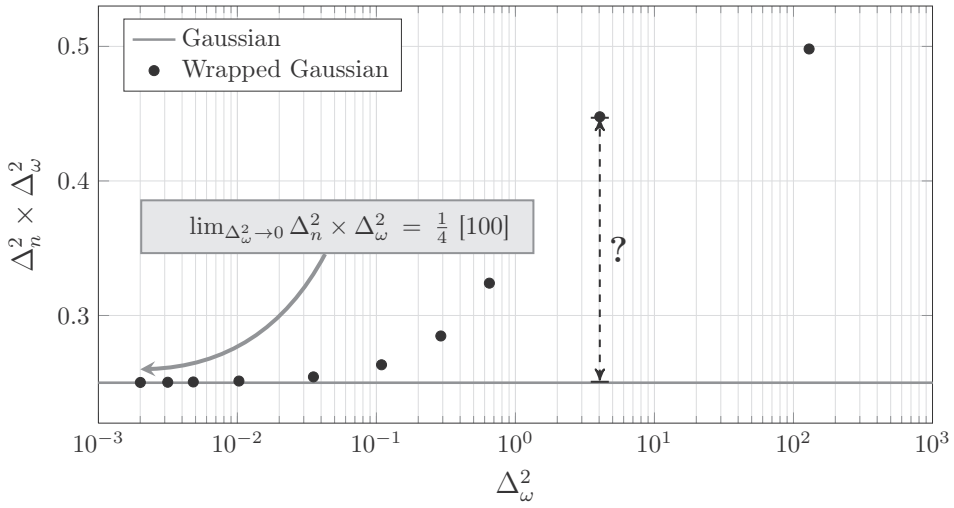


Figure 5.3: *The time-frequency spread of the wrapped gaussian (its Fourier series are samples of a gaussian) only tends to the Heisenberg limit (0.25) as* $\Delta_\omega^2 \to 0$.

5.1.2 Chapter outline

We address the fundamental yet unanswered question : If someone asks us to design a sequence with a certain frequency spread ($\Delta_{\omega_p}^2$ fixed), can we return the sequence with minimal time spread Δ_n^2?

Answering this question, can be formulated as the equation

$$\begin{aligned} \Delta_{n,\text{opt}}^2 \;&= \underset{x_n}{\text{minimize}} \quad \Delta_n^2 \\ &\text{subject to} \quad \Delta_{\omega_p}^2 \;=\; \text{fixed}\,. \end{aligned} \tag{5.6}$$

The solution of (5.6) is called a *maximally compact* sequence and its spectrum is in turn a maximally compact periodic waveform.

For non-periodic functions, gaussians are the well-known solutions of this problem, and they reach the Heisenberg limit of 1/4 regardless of their spread. To make the transition from non-periodic functions to waveforms, one may try as a first guess to periodize the gaussian function by wrapping it — which corresponds to sampling its Fourier transform uniformly. We show the result in Figure 5.3. For narrow waveforms, the wrapped-gaussian tends to the Heisenberg limit — agreeing with [100]. However, as the variance of the wrapped gaussian is increased, its time-frequency spread increases and saturate at the value of 1/2.

Framing the design of maximally compact sequences as an optimization problem, we show that contrary to the continuous case, it is not possible to reach a constant time-frequency lower bound for arbitrary time or frequency spreads. We further develop a simple optimization framework to find maximally compact sequences in the time domain for a given frequency spread. It means that we can design exactly periodic waveforms of a given spread having a minimal time-frequency spread. As a

corollary, it provides a *sharp uncertainty principle for sequences and periodic waveforms* since the optimal can be computed — see Figure 5.6 for an illustration.

We also show that maximally compact waveforms can all be formed from a template which is *Mathieu's harmonic cosine of order* 0.

5.2 Localization and its effect on time-frequency uncertainty

The choice of $\Delta^2_{\omega_p}$ as an *angular variance* must be motivated. The most thorough study on the subject in [35] provides many heuristic reasons. If these reasons all have their merits, we would also like to show that the definition of a frequency spread is unambiguously linked to a pseudo-differential operator, providing a simple and unique selection criterion. We will show that $\Delta^2_{\omega_p}$ corresponds to the *finite difference* operator which is the simplest among discrete differentials.

The *Heisenberg uncertainty principle* is rooted in particle physics, and therefore thought in terms of *position-momentum* uncertainty rather than *time-frequency* spread as in the signal processing community. These two interpretations are equivalent, but the intuitions behind them are useful to make the transition from continuous to discrete time.

5.2.1 Uncertainty Principle for Linear Operators

Consider a Hilbert space $\mathcal{H}$ with the inner-product $\langle \cdot\,,\,\cdot \rangle$ and the induced norm $\|x\| \stackrel{\text{def}}{=} \sqrt{\langle x\,,\,x\rangle}$. Define linear operators $L,\ M\ :\ \mathcal{H} \mapsto \mathcal{H}$, and the mean-value

$$\mu_L(x) \stackrel{\text{def}}{=} \langle Lx\,,\,x\rangle \,/\, \|x\|^2 .$$

If L is self-adjoint, $\mu_L(x) \in \mathbb{R}$.

For any pair of linear operators, the *commutator*

$$[L\,,\,M] \stackrel{\text{def}}{=} LM - ML ,$$

vanishes if and only if its arguments commute [38].

With these definitions, the *Heisenberg uncertainty principle* is

Theorem 5.1. *(Schrödinger 1930 [9; 111])*
For L and M self-adjoint,

$$\|(L - \mu_L(x))x\|^2 \, \|(M - \mu_M(x))x\|^2 \ \geq\ \frac{1}{4} \left|\langle [L\,,\,M]\, x\,,\,x\rangle\right|^2 \ . \qquad (5.7)$$

Proof.

Using Cauchy-Schwarz inequality and the self-adjointness of L and M,

$$\begin{aligned}\|(L-\mu_L(x))x\|^2\,\|(M-\mu_M(x))x\|^2 &\geq |\langle (M-\mu_M(x))x\,,\,(L-\mu_L(x))x\rangle|^2\ ,\\ &= \left|\langle LMx\,,\,x\rangle - \mu_L(x)\,\mu_M(x)\,\|x\|^2\right|^2\ ,\\ &= \Big|\frac{1}{2}\,\langle (LM-ML)x\,,\,x\rangle \\ &\qquad + \frac{1}{2}\,\langle (LM+ML)x\,,\,x\rangle \\ &\qquad -\mu_L(x)\,\mu_M(x)\,\|x\|^2\Big|^2\ .\end{aligned}$$

The two halves within the modulus are respectively imaginary and real, therefore with the *anticommutator* $\{L\,,\,M\} \stackrel{\text{def}}{=} LM+ML$, one obtains the *Schrödinger uncertainty principle*

$$\begin{aligned}\|(L-\mu_L(x))x\|^2\,\|(M-\mu_M(x))x\|^2 \geq\ & \frac{1}{4}\,|\langle [L\,,\,M]\,x\,,\,x\rangle|^2 \\ & + \frac{1}{4}\,|\langle \{L\,,\,M\}\,x\,,\,x\rangle \\ & \qquad -2\mu_L(x)\,\mu_M(x)\,\|x\|^2\Big|^2\ ,\end{aligned}$$

from which the weaker *Heisenberg uncertainty principle* follows. □

Remark : The self-adjointness of the operators can be relaxed a little. For example, a multiplication by a number of modulus 1 could be applied to localization or momentum to make them self-adjoint. Also, in particular cases — as in [35] — the same inequality may hold even though one of the operator is not self-adjoint.

5.2.2 The journey from continuous to discrete

We apply the general uncertainty relation (5.7) to continuous and discrete-time signals to make some connections explicit. In particular, we study two schemes to make the transition from continuous-time to discrete-time.

Continuous time-frequency uncertainty

Let $x(t) \in L_2(\mathbb{R})$, the localization L and momentum M operators are

$$\begin{aligned}Lx(t) &\stackrel{\text{def}}{=} t\cdot x(t)\ ,\\ Mx(t) &\stackrel{\text{def}}{=} \frac{dx}{dt}(t)\ .\end{aligned}$$

Moreover, because of the similarity between the continuous-time domain and the CTFT domain, localization and momentum are Fourier dual of each other :

$$\begin{aligned} Lx(t) &\overset{\text{CTFT}}{\longleftrightarrow} j\frac{dX}{d\omega}(\omega) \overset{\text{def}}{=} M_{\mathcal{F}}X(\omega) \ , \\ Mx(t) &\overset{\text{CTFT}}{\longleftrightarrow} j\omega X(\omega) \overset{\text{def}}{=} L_{\mathcal{F}}X(\omega) \ . \end{aligned}$$

This duality shows that the *localization-momentum* and *time-frequency localization* interpretations are the sides of the same coin.

With the evaluation of

$$\langle [t, d/dt]x \ , \ x \rangle \ = \ \|x\|^4 \ ,$$

where the square brackets indicate the commutator and the use of Parseval's equality

$$\left\|\overline{L_{\mathcal{F}}}X\right\|^2 / \|X\|^2 = \left\|\overline{M}x\right\|^2 / \|x\|^2 \ ,$$

the uncertainty relation (5.7) yields the continuous-time *Heisenberg uncertainty principle* (5.1)

$$\Delta t^2 \Delta_{\omega_c}^2 \geq \frac{1}{4} \ ,$$

where $\Delta t^2 = \left\|\bar{L}x\right\|^2 \Big/ \|x\|^2$ and $\Delta_{\omega_c}^2 = \left\|\overline{L_{\mathcal{F}}}X\right\|^2 \Big/ \|X\|^2$ as found in Table 5.1.

First attempt at discrete-time uncertainty: the DSP point of view

In the same way that we defined the localization and momentum in time and frequency for continuous-time signals, we define them for sequences. The process can be seen either from the time-domain as a discretization or from the frequency domain as a periodization.

If we follow the time-frequency interpretation, it is natural to discretize L and periodize $L_{\mathcal{F}}$

$$\begin{array}{rcl} \text{Designed Localization:} & & \text{Implied Momentum:} \\ Lx_n & \overset{\text{def}}{=} & n \cdot x_n \\ L_{\mathcal{F}}X(e^{j\omega}) & \overset{\text{def}}{=} & \widetilde{\omega} \cdot X(e^{j\omega}) \end{array} \quad \overset{\text{DTFT}}{\longleftrightarrow} \quad \begin{array}{rcl} j\frac{dX}{d\omega}(e^{j\omega}) & \overset{\text{def}}{=} & M_{\mathcal{F}}X(e^{j\omega}) \\ \frac{(-1)^n}{jn} * x_n & \overset{\text{def}}{=} & Mx_n \end{array} \tag{5.8}$$

where $\widetilde{\omega}$ represents the sawtooth wave with period 2π. The implied frequency-domain momentum $M_{\mathcal{F}}$ has the properties one would expect for momentum as it measures the variation of the spectrum locally by differentiation; on the other side, the implied time-domain momentum M does not lead to an easy interpretation.

This choice of operators not only causes interpretation problems. The uncertainty relation is

$$\Delta_n^2 \Delta_{\omega_\ell}^2 \geq \frac{1}{4} \left| 1 - \frac{\left|X(e^{j\pi})\right|^2 - 2\text{Re}\left\{X'(e^{j\pi})X^*(e^{j\pi})\right\}}{\|X\|^2} \right|^2 \ , \tag{5.9}$$

where $\Delta_n^2 \overset{\text{def}}{=} \|Lx\|^2 / \|x\|^2$ and $\Delta_{\omega_\ell}^2 \overset{\text{def}}{=} \|L_{\mathcal{F}}X\|^2 / \|X\|^2$, as found in Table 5.2.

Figure 5.4: ***Two ways to obtain an uncertainty principle for sequences.*** *On the left side* a) *—by periodization—one obtains the definitions of Table 5.2, on the right side* b) *—by discretization—one obtains the definitions of Table 5.3. Note that periodization and discretization yield the same result for the top plane, but not for the bottom plane.*

Under the condition $X(e^{j\pi}) = 0$, the uncertainty principle (5.3) is obtained

$$\Delta_n^2 \Delta_{\omega_\ell}^2 \geq \frac{1}{4} \ , \quad X(e^{j\pi}) = 0 \ .$$

However the necessity of a root at π severely reduces the utility of this result, excluding sequences such as the discretized gaussian kernel, binomial filters, etc. Also natural periodic features such as shift invariance in frequency (modulation) or scaling invariance are lost.

Second attempt at discrete-time uncertainty: the physicist point of view

Failure of the first attempt can be pinned down on the definition of the frequency-domain localization. Instead of designing both localization operators, we may design both time-domain operators. The task is thus to discretize localization (already done) and momentum. For the latter, the *finite difference* filter is a natural candidate

$$\begin{array}{lll} \text{Designed in the time-domain} & & \text{Implied in the frequency-domain} \\ Lx_n \stackrel{\text{def}}{=} n \cdot x_n & \overset{\text{DTFT}}{\longleftrightarrow} & j\frac{dX}{d\omega}(e^{j\omega}) \stackrel{\text{def}}{=} M_{\mathcal{F}} X(e^{j\omega}) \\ Mx_n \stackrel{\text{def}}{=} x_n - x_{n-1} & & (1 - e^{-j\omega}) \cdot X(e^{j\omega}) \stackrel{\text{def}}{=} L_{\mathcal{F}} X(e^{j\omega}) \end{array} \tag{5.10}$$

The implied operator $L_{\mathcal{F}}$ has intuitive properties. For example,

$$\|L_{\mathcal{F}} X\|^2 = 2 \int_{-\pi}^{\pi} (1 - \cos(\omega)) \left| X(e^{j\omega}) \right|^2 d\omega \ ,$$

shows that the spread is measured with respect to the smooth kernel $2(1 - \cos(\omega))$ instead of $\widetilde{\omega}^2$ in (5.8).

Note that $2(1 - \cos(\omega)) = \omega^2 + \mathcal{O}(\omega^4)$, so the two definitions of spread coincide for $\omega \to 0$, and should yield asymptotically equal results for sequences with a narrow spectrum.

The operators $M_{\mathcal{F}}$ and $L_{\mathcal{F}}$ are the ones most often used in directional statistics [81].

In [35], the same uncertainty principle as (5.7) is shown to hold for this particular choice of operators[4], and the right-hand side of the inequality evaluates to

$$\left| \left\langle \left[1 - e^{-j\omega} \ , \ jd/d\omega \right] X \ , \ X \right\rangle \right|^2 \ = \ \|X\|^4 \, |\tau(x)|^2 \ . \tag{5.11}$$

For $\tau(x) \neq 0$ the corresponding uncertainty relation [35] is therefore

$$\Delta_n^2 \Delta_{\omega_p}^2 \geq \frac{1}{4} \ , \quad \tau(x) \neq 0, \tag{5.12}$$

where $\Delta_{\omega_p}^2 \stackrel{\text{def}}{=} \frac{\left\| \overline{L_{\mathcal{F}}} X \right\|^2}{|\tau(x)|^2 \|X\|^2}$ and Δ_n^2 are as found in Table 5.3.

The study of continuous and discrete uncertainty principle shows that the definitions of time and frequency spreads found in Table 5.3 are not arbitrary and follow from the most natural discretization of the continuous-time definitions found in Table 5.1.

[4] It is not a corollary of (5.7) since one operator lacks self-adjointness.

5.3 Maximally compact waveforms and sequences

The main objective is to design maximally compact waveforms/sequences as solutions of (5.6). Thus we are interested in solving

$$\begin{aligned} \Delta^2_{n,\text{opt}} &= \underset{x_n}{\text{minimize}} \quad \Delta^2_n \\ &\qquad \text{subject to} \quad \Delta^2_{\omega_p} = \sigma^2 \,. \end{aligned} \tag{5.13}$$

where σ^2 is the fixed, given frequency spread of the sequence. We saw in (5.12) that the time-frequency spread of such sequences is naturally bounded from below by the Heisenberg uncertainty bound. Prestin et al. in [99] show that the lower bound in (5.12) is achievable only asymptotically when the frequency spread of the sequence tends to zero. Thus, the question is "what is the minimal achievable uncertainty for sequences with a given frequency spread?". The answer to this question lies in the solution of (5.13).

We start with some properties of maximally compact sequences. These properties will greatly facilitate the task of solving (5.13).

5.3.1 Properties of Maximally Compact Sequences

In the definitions of time and frequency spreads in Table 5.3 we considered complex sequences and their DTFTs. In the following, we establish two lemmas that make the search for maximally compact sequences easier. In the following we assume that $\|x\| = 1$.

Lemma 5.1. *For any fixed Δ^2_n or $\Delta^2_{\omega_p}$,*

$$x_n \text{ maximally compact} \quad \Longrightarrow \quad |x_n| \text{ maximally compact.}$$

Proof.

See Appendix E.1.1. □

Consider also the shift operator

$$x_{n+\nu} \quad \overset{\text{DTFT}}{\longleftrightarrow} \quad e^{j\omega\nu} X(e^{j\omega}) \,, \quad \nu \in \mathbb{R}, \tag{5.14}$$

whose principal effect is to shift the time center of a sequence

$$\mu_n(x_{n+\nu}) \;=\; \mu_n(x) - \nu. \tag{5.15}$$

Notice that ν might be non-integer, in which case $x_{n-\nu}$ is a shorthand for sinc resampling on a grid shifted by ν in the time-domain.

Lemma 5.2. *If x is a maximally compact sequence, then $x_{n-\mu_n(x)}$ is also maximally compact.*

Proof.
See Appendix E.1.2. □

Lemmas 5.1 and 5.2 greatly reduce the complexity of the problem, and from here on we only consider — without loss of generality — real, positive sequences x, with $\mu_n(x) = 0$ and $\|x\|^2 = 1$.

5.3.2 The computation of maximally compact waveforms & sequences

Theorem 5.2. *For finding the solution of* (5.13)*, it is sufficient to solve the following semi-definite program (SDP)*

$$\begin{aligned} \underset{\boldsymbol{X}}{\text{minimize}} \quad & tr(\boldsymbol{A}\boldsymbol{X}) \\ \text{subject to} \quad & tr(\boldsymbol{B}\boldsymbol{X}) = \alpha \\ & tr(\boldsymbol{X}) = 1, \quad \boldsymbol{X} \succeq 0\,, \end{aligned} \tag{5.16}$$

where $\alpha = \frac{1}{\sqrt{1+\sigma^2}}$. Further, $\boldsymbol{X}^{opt}$, the solution to (5.16) *has rank one and $\boldsymbol{X}^{opt} = \boldsymbol{x}^{opt}\,\boldsymbol{x}^{opt^T}$, with $\boldsymbol{x}^{opt}$ the solution of* (5.13)*. Matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined as*

$$\boldsymbol{A} = \begin{bmatrix} \ddots & & & & & \\ & 2^2 & & & \mathbf{0} & \\ & & 1^2 & & & \\ & & & 0 & & \\ & & & & 1^2 & \\ & \mathbf{0} & & & & 2^2 \\ & & & & & & \ddots \end{bmatrix}, \quad \boldsymbol{B} = \begin{bmatrix} \ddots & & & & \mathbf{0} \\ & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & \frac{1}{2} & 0 & \frac{1}{2} \\ \mathbf{0} & & & & \ddots \end{bmatrix}.$$

Proof.
See Appendix E.2. □

The SDP in (5.16) can be solved to an arbitrary precision by using existing optimization toolboxes; for example using the cvx software package [61]. This gives a constructive way to design sequences that are maximally compact in the time domain with a given frequency spread.

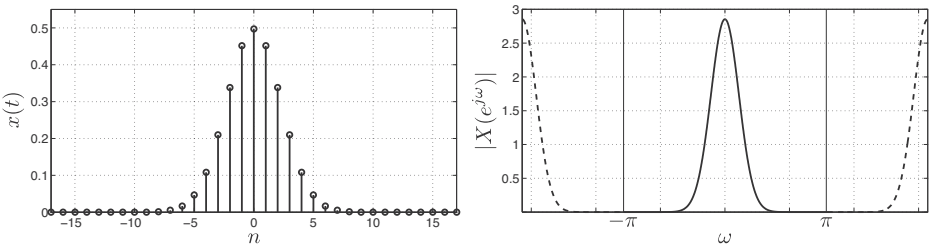


Figure 5.5: ***An example solution of*** (5.16)**.** *The output of the SDP in* (5.16) *with* $\sigma^2 = 0.1$ *using* cvx *in Example 5.b. The optimal value for* Δ_n^2 *is found to be* 2.62 *which results in a time-frequency spread of* $\eta_p = 0.262$.

Example 5.b — Computing a maximally compact sequence with cvx

Take $\sigma^2 = 0.1$ to be the fixed and given frequency spread of the sequence. We can use cvx to solve the semi-definite program (5.16) and find the optimal value of $\Delta_n^2 = 2.62$. This results in the time-frequency spread of $\eta_p = 0.262$. The code in MATLAB is:

```
cvx_begin
variable X(n,n);
minimize(trace(A*X))
subject to
    trace(B*X) == alpha
    trace(X) == 1
    X == semi-definite(n)
cvx_end;
```

Note that contrary to continuous-time signals, we cannot reach the 0.25 lower bound for sequences. The resulting sequence and its DTFT are shown in Figure 5.5.

The dual of SDR (5.16) is [34]:

$$\begin{aligned} \underset{\lambda_1,\lambda_2}{\text{maximize}} \quad & \alpha\,\lambda_1 + \lambda_2 \\ \text{subject to} \quad & \boldsymbol{A} - \lambda_1\,\boldsymbol{B} - \lambda_2\,\boldsymbol{I} \succeq 0 \end{aligned} \tag{5.17}$$

Lemma 5.3. *For the primal problem* (5.16) *and the dual* (5.17)*, strong duality holds.*

Proof.

We refer the reader to Appendix E.3 for the proof. □

Thus, for finding the time-frequency spread of maximally compact sequences, solving the dual problem suffices.

Note that although Theorem 5.2 provides a constructive way to find maximally compact sequences, it does not specify the closed form for these sequences. One would be interested to see if—in analogy to continuous-time—sampled gaussians are maximally compact? The answer is negative as the spectrum of maximally compact sequences is related to *Mathieu functions* as shown by this theorem:

Theorem 5.3. *The DTFT spectrum, $X(e^{j\omega})$ of maximally compact sequences are Mathieu functions. More specifically,*

$$X(e^{j\omega}) \;=\; \gamma_0 \cdot \mathrm{ce}_0(-2\lambda_1\,;\,(\omega-\omega_0)/2)e^{j\mu\omega}\;, \tag{5.18}$$

where $|\gamma_0| = \|\mathrm{ce}_0(-2\lambda_1\,;\,(\omega-\omega_0)/2)\|^{-1}$, ω_0, μ are shifts in frequency or time and λ_1 is the optimal solution of the dual problem (5.17)*. $\mathrm{ce}_0(.\,;.)$ is Mathieu's harmonic cosine function of order zero.*

For the proof of the theorem and further insights about Mathieu functions, we refer the reader to Appendix E.4.

Using the constructive method presented in Theorem 5.2, we can find the achievable (and tight) uncertainty principle bound for discrete sequences. This is shown and discussed more in Section 5.4 and Figure 5.6. However, a numerically computed boundary may not always be practical, and even though the numerical solution exactly solves the problem, its accuracy may be challenged. Therefore, we characterize the asymptotic behavior of the time-frequency bound:

Theorem 5.4. *If x_n is maximally compact for a given $\Delta^2_{\omega_p} = \sigma^2$, then*

$$\eta_p = \Delta_n^2\,\Delta^2_{\omega_p} \geq \sigma^2\left(1-\sqrt{\frac{\sigma^2}{1+\sigma^2}}\right). \tag{5.19}$$

Proof.

See Appendix E in [94]. □

This fundamental result states that for a given frequency spread, we cannot design sequences which achieve the classic Heisenberg uncertainty bound. We will see how this curve compares to the classic Heisenberg bound in Section 5.4.

The lower bound in (5.19) converges to 1/2 as the value of σ^2 grows, and "pushes up" the time-frequency spread of maximally compact sequences towards 1/2 which is also an asymptotic upper bound on the time-frequency spread as $\Delta^2_{\omega_p} \to \infty$. Indeed, one may construct the unit-norm sequence $x_n^{(\varepsilon)} = \varepsilon\delta_{n+1} + \sqrt{1-2\varepsilon^2}\delta_n + \varepsilon\delta_{n-1}$, which verifies $\lim_{\varepsilon\to 0}\eta_p(x^{(\varepsilon)}) = 1/2$.

Theorem 5.5. *For small values of σ^2, maximally compact sequences satisfy*

$$\eta_p = \Delta_n^2 \, \Delta_{\omega_p}^2 \leq \frac{\sigma^2}{8} \left(\frac{\sqrt{1+\sigma^2}}{\sqrt{1+\sigma^2}-1} - \frac{1}{2} \right) . \tag{5.20}$$

Proof.

See Appendix F in [94]. □

For small values of σ^2, the upper bound in (5.20) converges from above to $1/4$, thus "pushing down" the time-frequency spread of maximally compact sequences towards the Heisenberg uncertainty bound $1/4$ from above.

Finite-Length Sequences

The theory that we have provided so far holds for infinite sequences. For computational purposes, we have to assume finite length for the sequences in the time domain, which is not an issue if the sequence length is chosen to be long enough such as to truncate samples below machine precision. As a side benefit, a length constraint on the sequence may be put at will to meet design requisites.

5.4 Simulation Results

In order to show the behaviour of the results obtained in Theorems 5.2, 5.4 and 5.5, we ran some simulations. To this end, we assumed that the designed filter is finite length with 201 taps in the time domain (the length is long enough not to pose restrictions on the solution). For different values of $\Delta_{\omega_p}^2 = \sigma^2$, we solved the semi-definite program (5.16) using the cvx toolbox in MATLAB.

The resulting values of Δ_n^2 are then multiplied with the corresponding $\Delta_{\omega_p}^2$ to produce the time-frequency spread of maximally compact sequences. The time-frequency spread of maximally compact sequences versus their frequency spread is shown with the solid curve in Figure 5.6. This means—numerically—that any time-frequency spread under this curve is not achievable. The dotted line in this figure shows the classic Heisenberg uncertainty bound. Comparing the two curves shows the gap between the classic Heisenberg principle and what is achievable in practice. The dashed lines represent analytical lower and upper bounds for the time-frequency spread of maximally compact sequences.

Further, to give an insight on how the time-frequency spread of some common filters compare to that of maximally compact sequences, we plot their time-frequency spread together with the new uncertainty bound in Figure 5.7. By changing the length of each filter in time, we can find its time and frequency spreads which results in a point on the figure. We observe that as shown by Prestin et al. in [100], asymptotically when the frequency spread of sequences are very small, sampled Gaussians converge to the lower bound for maximally compact sequences.

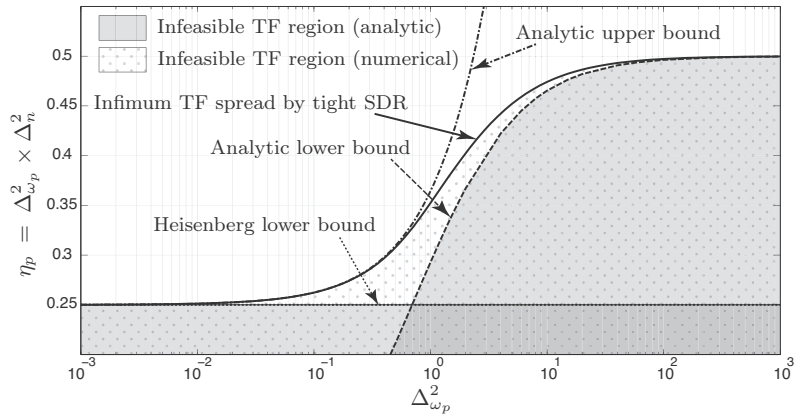


Figure 5.6: ***New uncertainty bounds***. *The solid line shows the results of solving the SDP in* (5.16). *The dotted line shows the classic Heisenberg lower bound. The dashed lines show the lower and upper bounds found in Theorems 5.4 and 5.5, respectively.*

5.5 Conclusion

In this chapter, we have seen that localization in a periodic domain is not as straightforward as one would initially think. It was shown that the adoption of a localization with a discontinuity made the definition of a meaningful time-frequency localization lowerbound difficult.

The localization operator used in circular statistics and quantum mechanics overcomes this difficulty, and we motivated its choice by showing its Fourier series corresponds to the simplest discretization of a derivative.

By working on the Fourier series of periodic waveforms, we derived a numerical procedure to compute sequences which have a minimal time-frequency spread. In an unusual way, the analysis of the numerical formulation yielded an analytical formula.

We conclude that although wrapped gaussians are not maximally compact periodic waveforms, they are extremely close contenders, and could be considered so for applications. This settles the question which was posed in Figure 5.3 about the nature of the gap between the wrapped gaussian and Heisenberg principle. With a better understanding of periodic phenomena, we will apply the same line of reasoning to obtain lowerbounds on the variance of periodic estimates in the next chapter.

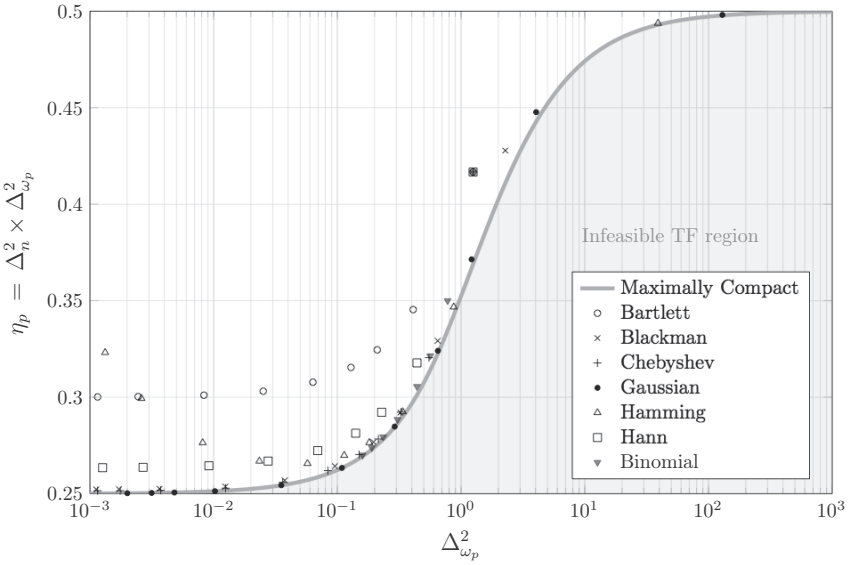


Figure 5.7: ***Time-frequency spread of common FIR filters.*** *By changing the length of the filters in time, we compute the time and frequency spreads for each type of the filters. For small values of frequency spread, Gaussian filters are good approximations of Mathieu functions (as shown also in [100]).*

Chapter 6

From uncertainty to estimation error

Determination of the minimal error achievable by an estimator based on a given measurement model is not only of interest to benchmark the performances of an algorithm, it can also be useful for operational purposes to assess whether or not the result of an estimation can be trusted.

In this chapter, we use the insight gained on the notion of localization in Chapter 5 to derive estimation lowerbounds of the *Cramér-Rao* family. The first contribution is to take into account periodic parameters (supported on $[-\pi, \pi[$) and not just aperiodic parameters (supported on $\mathbb{R}$). Path locations in OFDM transmissions (Chapters 1–4) or bearings are examples of periodic parameters. Simply thinking about the estimation of a wind direction should motivate the formulation of periodic lowerbounds on the variance of an estimator. Indeed, if the difference between a true direction of $0°$ and an estimated direction of $359°$ is taken to measure the estimation error, one obtains a distance of $359°$ which defies common-sense. If one tries to get around this issue by shifting the parameter interval so that $359°$ becomes $-1°$, the same problem is still there for parameter values close to the interval boundaries.

Taking into account the periodicity of the parameter space from the beginning does not make the derivation of estimation lowerbounds harder, rather, it simplifies many aspects. For example the existence of unbiased estimators is compromised by the existence of boundaries at the end of the interval, and this issue is related in the introduction of [149].

To obtain a periodic formulation of the Cramér-Rao bound (CRB), we first show that the CRB, in general, can be formulated as Heisenberg's uncertainty principle. The immediate consequences are not only on the level of the interpretation, but also on the level of the application. We saw in Chapter 5 two ways to make the transition from an aperiodic uncertainty principle to a periodic one. We experimented that the choice

of the periodic localization played an important role. Having a similar formulation for the CRB, we can make the transition from the aperiodic to the periodic in the same way with very little efforts.

The periodic Cramér-Rao bound (CRB) we will obtain is very similar in its form to the CRB for aperiodic parameters (the "classical" CRB), so estimation of periodic parameters with aperiodic and periodic nuisance parameters is readily available. The practical implication is that this kind of lowerbounds directly applies to the joint channel estimation problem we studied in Chapters 1–4, where the times of arrival are periodic and the paths amplitudes are not.

We then shift our attention to a slightly different problem. To exploit the similarities between Heisenberg principle and the CRB, we can think in terms of localization and momentum. The choice of the localization operator induces a definition for the bias and the variance of an estimator, and we already chose a periodic localization operator when deriving the periodic CRB. The function to which we applied this operator is dictated by probability distribution of the measurements, and so we are in a setup quite different from Chapter 5 where we chose the function which made the uncertainty inequality as tight as possible. What is left for us to use is the momentum operator. The CRB adopts the classical definition of momentum which is the first order derivative, but we are not limited to this particular choice. The path to follow is now clear; which linear operator should we use in place of the derivative to make the lowerbound as tight as possible?

Replacing the derivative by a linear shift invariant filtering operation was proposed by Barankin [13], who proved — non constructively — that an infinite stream of properly chosen delta functions "made the most" out of the inequality[1]. We design analytically a filter which achieves Barankin result. Rather than relying on an infinite stream of delta functions we adopt the approach developed by Swerling [122] for aperiodic parameters.

All of the key concepts discussed in this chapter are present in an univariate setup (one parameter) — with the exception of the multivariate CRB formula for mixed periodic and aperiodic parameters. Therefore, we will in general only consider the estimation of a single parameter.

6.1 Some history

We will first review the history of *Cramér-Rao bounds* (CRB) and *Barankin bounds* (BB), and see their range of application and their limitations.

6.1.1 The Cramér-Rao bounds

In 1946 H. Cramér introduced a lowerbound for unbiased estimators [48].

Let $\widehat{\theta}_X$ be an estimator of $\theta \in \mathbb{R}$ based on the random measurements X. The measurements follows the probability distribution p_θ. The bias of $\widehat{\theta}_X$ is defined as

[1] We voluntarily avoid to say it makes the application of the Cauchy-Schwarz inequality tight, as it could be interpreted that the obtained lowerbound is tight.

$$\mathsf{bias}_{\widehat{\theta}}(\theta) \stackrel{\text{def}}{=} \int_{\mathcal{X}} (\widehat{\theta}_{\boldsymbol{x}} - \theta) p_\theta(\boldsymbol{x}) \boldsymbol{dx} \;=\; \mathbb{E}\left[\widehat{\theta}_X - \theta\right].$$

To derive the CRB, we make the assumption that $\widehat{\theta}_X$ is unbiased

$$\mathsf{bias}_{\widehat{\theta}}(\theta) \;=\; 0, \quad \forall \theta.$$

We then proceed to compute

$$\frac{d}{d\theta}\mathsf{bias}_{\widehat{\theta}}(\theta) \;=\; \int_{\mathcal{X}} (\widehat{\theta}_{\boldsymbol{x}} - \theta) \frac{d}{d\theta} p_\theta(\boldsymbol{x}) \boldsymbol{dx} - \underbrace{\int_{\mathcal{X}} p_\theta(\boldsymbol{x}) \boldsymbol{dx}}_{=1}. \tag{6.1}$$

With the unbiased assumption, the left hand side of (6.1) is null, and rearranging the remaining terms of the equation yields

$$\mathbb{E}_{X|\theta}\left[\underbrace{(\widehat{\theta}_X - \theta)}_{u} \cdot \underbrace{\frac{\frac{d}{d\theta} p_\theta}{p_\theta}}_{v} \right] \;=\; 1.$$

Cauchy-Schwarz inequality states

$$\mathbb{E}\left[|u|^2\right] \mathbb{E}\left[|v|^2\right] \;\geq\; \left|\mathbb{E}\left[uv\right]\right|^2,$$

therefore (we have shown above that $\left|\mathbb{E}\left[uv\right]\right|^2 = |1|^2$),

$$\mathbb{E}\left[\left|\widehat{\theta}_X - \theta\right|^2\right] \geq \mathbb{E}\left[\left(\frac{\frac{d}{d\theta} p_\theta}{p_\theta}\right)^2\right]^{-1} \stackrel{\text{def}}{=}\; J_\theta^{-1}. \tag{6.2}$$

The denominator in the right-hand side of (6.2) is called the *Fisher information* and is noted J_θ. Because we assumed that the estimator is unbiased, it is also a lowerbound on the variance.

Example 6.a — Single pulse estimation

Consider the measurements

$$x_n \;=\; s_\theta[n] + \sigma^2 \varepsilon_n\,, \quad -M \leq n \leq M,\; \theta \in [-\pi,\, \pi[, \tag{6.3}$$

where

$$s_\theta[n] \;=\; D_M\left(\frac{2\pi}{2M+1} n - \theta\right),$$

and D_M is the Dirichlet kernel of bandwidth N, and $\{\varepsilon_n\}$ are realizations of iid standard normal random variables.

The value of θ is estimated from $\boldsymbol{x}$ using the maximum-likelihood estimator $\widehat{\theta}_{\text{ML}}$.

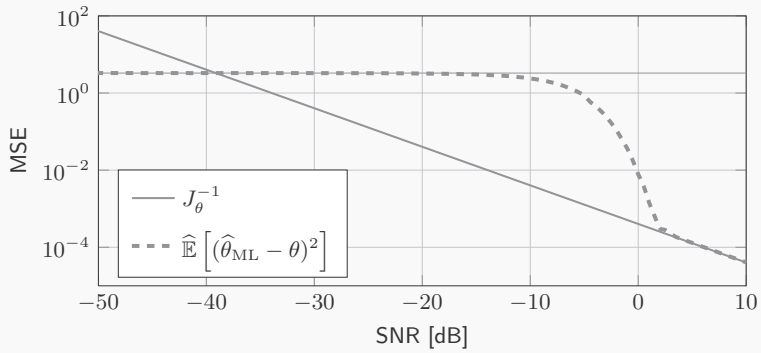


Figure 6.1: *MSE of the maximum-likelihood (ML) estimator – or variance since the ML estimator is unbiased – and the inverse Fisher information.*

The ML estimator minimizes the variance asymptotically as the noise vanishes [132]. As the SNR tends to $-\infty$ dB, the lowerbound diverges to $+\infty$, which causes interpretation issues for an estimation where the unknown parameter is known to be in a finite interval.

In Figure 6.1, the MSE of the maximum-likelihood (ML) estimator is optimal above 2 dB of SNR. It is rather troubling that the CRB diverges to infinity as the SNR goes toward $-\infty$ dB. The distance $\widehat{\theta}_X - \theta$ is upper-bounded by 2π — or π whether we consider the distance modulo π or not. Fatally, the CRB is not a valid lowerbound below a certain SNR as it does not take into account the boundedness or the periodicity of the parameter space into account. More radically, it could cast a doubt on the validity of the bound regardless of SNR. However, as the SNR tends to $+\infty$ the distribution of reasonably good estimators concentrates around θ and the approximation of the parameter space by the infinite line is adequate.

As a conclusion, since physical quantities are bounded or periodic, the CRB can only be used in the high SNR regime where the estimation error is local.

6.1.2 Improvement : Barankin's bound

To derive the CRB, one starts with differentiation of the bias as in (6.1) and then applies the Cauchy-Schwarz inequality to the result. But, instead of the derivative $\frac{d}{d\theta}$, a linear shift invariant filter can be applied to the bias :

$$(g * \mathsf{bias}_{\widehat{\theta}})(\theta) = \int_{\mathcal{X}} (\widehat{\theta}_{\boldsymbol{x}} - \theta)(g(t) * p_t(\boldsymbol{x}))(\theta) d\boldsymbol{x} - \int_{\Theta} g(t) t \int_{\mathcal{X}} p_t(\boldsymbol{x}) d\boldsymbol{x} dt, \tag{6.4}$$

$$= \mathbb{E}\left[(\widehat{\theta}_X - \theta) \cdot \frac{(g(t) * p_t)(\theta)}{p_\theta}\right] + (g(t) * t)(\theta). \tag{6.5}$$

The corresponding MSE lowerbound is

$$\mathbb{E}\left[\left|\widehat{\theta}_X - \theta\right|^2\right] \geq |(g(t) * t)(\theta)|^2 \cdot \mathbb{E}\left[\left|\frac{(g(t) * p_t)(\theta)}{p_\theta}\right|^2\right]^{-1}. \tag{6.6}$$

This lowerbound is not the Barankin lowerbound *per se*, but our own construction which makes it easier to introduce Barankin's idea. The freedom we have gained by using a linear filter is the ability to choose its impulse response g such as to maximize the right-hand side of (6.6).

To this end, some intuition can be gained from euclidean geometry. For any two non-zero vectors u and v in a euclidean space, the Cauchy-Schwarz inequality applied to this pair of vectors becomes an equality iff u and v are *collinear*

$$\|u\|_2^2 \cdot \|v\|_2^2 \;=\; |\langle u\;,\;v\rangle|^2 \quad , \quad \text{iff } u \text{ and } v \text{ are collinear.}$$

Using this notion of collinearity, Barankin showed that the sequence of filters

$$g_n(t) \;=\; \sum_{k=0}^{n} \alpha_k \delta(t - t_k)$$

will maximize[2] the right hand side (RHS) of (6.6) for appropriately sequentially chosen parameters (α_k, t_k) as $n \to \infty$. I.e., the difference between the maximal RHS obtained with g and the RHS obtained with g_k tends to 0 as k grows.

The merit of Barankin's result is to pose the design of g as a greedy procedure making the search complexity manageable. In practice, this procedure is hard to apply for numerical stability reasons [1; 45; 147], and finding a solution to this problem — called *Barankin bound approximation* — has been an open topic for many years [8; 1; 86; 147]. Another open question is that there is no quantification of the rate of convergence with k, the number of delta functions. So, stopping the procedure early because of numerical instability or using different filters opens the question of how close is the lowerbound from the optimal.

6.1.3 Other parameter spaces: manifolds, periodicity, ...

So far, we assumed $\theta \in \mathbb{R}$. Many estimation problems involve parameters supported on finite intervals (e.g. speed estimation — $< 3 \cdot 10^8 \mathrm{m/s}^{-1}$), semi-finite intervals (e.g. sampled variance estimation — ≥ 0), periodic domains (e.g. paths localization in OFDM communications, or in CDMA where the pseudo-random sequences are long but periodic), euclidean manifolds, ...

Most of the CRBs developed for these problems only concern cases for which the parameter space can be assimilated to the infinite real-line. For example, a loose definition of N-dimensional euclidean manifolds is that they look "locally" like the classical euclidean space $\mathbb{R}^N$, allowing for a *linear approximation* as long as the lowerbound is low enough. In this case the error is local and similar to what would happen in the classical euclidean space.

In the periodic case, using the aperiodic approximation is not good enough to yield a reliable Barankin bound approximation. Indeed, the CRB evaluated the derivative around the true parameter value θ ensuring the result depends only on the local neighborhood of θ. For the Barankin bound, this is not the case as the impulse response of the filters g_k may spread over the whole parameter space.

[2] Majorization of the lowerbound does not imply tightness !

To the best of our knowledge, one recent contribution [106] proposed a periodic CRB which is of interest in the field of communications. One limitation of this result is that it requires the knowledge of the distribution of an estimator minimizing the MSE to compute the periodic CRB. One would argue that knowing this distribution limits the application of a MSE lowerbound, as a the minimal MSE is directly available from it. In Appendix F.5, we show how it can be avoided (paying the price of a non closed-form formula).

6.1.4 Problem summary

After this brief review, it appears the challenges are two-folds to obtain good lowerbounds for periodic parameters estimation

1. Include the periodic nature of the parameter into the lowerbound formulation.
2. Quantify the strength of Barankin's bound approximation for periodic parameters or give an explicit exact formula.

It turns out these two problems are linked, as a proper periodic formulation will yield a closed form formula for Barankin's lowerbound.

A similar formula was derived by Swerling [122] in 1959 for an aperiodic parameter, but it found surprisingly little echo in the literature.

6.2 An uncertainty-like inequality for estimators

The connection between the CRB and the Heisenberg principle is known to some extent. In [50] it is shown that the CRB and Heisenberg's uncertainty principle are equivalent for random measurement vectors which depend linearly on the parameters ([50], p.16-17). In the more general case where the dependency is non-linear, equivalence is shown between Heisenberg's uncertainty principle and a Bayesian CRB[3] ([50], p.17-18).

The parallel between the classical CRB — *i.e.* a lowerbound on the variance of an estimator with no bias (or a constant bias) for a deterministic value of θ — and the Heisenberg principle is not immediate. In fact we could not derive the CRB from the Heisenberg principle.

Nevertheless, a similar a formula with the same structure as the Heisenberg uncertainty principle exists under mild conditions on the bias, but it will be proven directly from the Cauchy-Schwarz inequality (without Stam's inequality as in [50] for the bayesian CRB).

Preliminary definitions We define the localization and momentum operators

Definition 6.1. ***Localization*** *multiplies the probability density with a function* $\mathsf{loc}(\widehat{\theta}_X, \theta)$*, measuring a difference between* $\widehat{\theta}_X$ *and* θ

[3] The Bayesian CRB was derived originally by Van Trees [132].

$$Lp_\theta \stackrel{\text{def}}{=} \mathsf{loc}(\widehat{\theta}_X, \theta) \cdot p_\theta.$$

*The **momentum** operator M is a linear operator depending only on θ and independent of the measurements.*

This definition of localization is broad enough to be applied to aperiodic and periodic parameters alike. The definition of momentum includes, of course, derivatives of various order (necessary for the CRB and the Bhattacharyya bound [57]), and linear shift invariant filtering (necessary for the Hammersley-Chapman-Robbins bound [44] or the Barankin bound [13] among others).

Then we define an inner-product

Definition 6.2. *For a bounded, strictly positive probability distribution p_θ and linear operators L and M as in Definition 6.1,*

$$\langle Lp_\theta \, , \, Mp_\theta \rangle \stackrel{\text{def}}{=} \mathbb{E}\left[\frac{Lp_\theta (Mp_\theta)^*}{p_\theta^2}\right] = \int_{\mathcal{X}} Lp_\theta(\boldsymbol{x})(Mp_\theta(\boldsymbol{x}))^* p_\theta^{-1}(\boldsymbol{x}) d\boldsymbol{x}. \quad (6.7)$$

Define also the norms

$$\|Lp_\theta\|^2 = \langle Lp_\theta \, , \, Lp_\theta \rangle \, , \quad \textit{and} \quad \|Mp_\theta\|^2 = \langle Mp_\theta \, , \, Mp_\theta \rangle \, .$$

Note that this inner-product may not be well-defined for any pair of arguments, and its application must be handled with care[4]. The inverse of the probability density function p_θ^{-1} is the positive definite[5] kernel of the inner-product.

As in Chapter 5, we also define the mean of L and M

$$\mu_L(\theta) \stackrel{\text{def}}{=} \langle Lp_\theta \, , \, p_\theta \rangle = \int_{\mathcal{X}} \mathsf{loc}(\widehat{\theta}_{\boldsymbol{x}}, \theta) \cdot p_\theta(\boldsymbol{x}) d\boldsymbol{x} \, , \quad (6.8)$$

which is a generalization of the *bias* of $\widehat{\theta}_X$. The use of loc will allow us to adopt different definition of localization for periodic parameters and choose the most suitable one.

For the momentum, remembering that M is only function of θ, we obtain

$$\mu_M(\theta) \stackrel{\text{def}}{=} \langle Mp_\theta \, , \, p_\theta \rangle = \int_{\mathcal{X}} Mp_\theta(\boldsymbol{x}) d\boldsymbol{x} = M\, 1 \, , \quad (6.9)$$

which is the application of M to the constant function 1. For shorthand, we will often write μ_L and μ_M, omitting θ. From (6.8) and (6.9), we draw natural interpretations for $\mu_L = 0$ and $\mu_M = 0$

$$\begin{aligned} \mu_L &= 0 \quad \Leftrightarrow \quad \text{Unbiased estimator} \, , \\ \mu_M &= 0 \quad \Leftrightarrow \quad M \text{ kills constants.} \end{aligned}$$

[4] In doubt, read [33] with 1g of aspirin.

[5] We assumed a probability density greater than 0 everywhere.

A prototype lowerbound Cauchy-Schwarz inequality implies

$$\|(L-\mu_L)p_\theta\|^2 \;\geq\; \frac{|\langle (L-\mu_L)p_\theta\,,\,(M-\mu_M)p_\theta\rangle|^2}{\|(M-\mu_M)p_\theta\|^2}\;, \tag{6.10}$$

and is indeed an inequality on the variance of $\widehat{\theta}_X$ as

$$\|(L-\mu_L)p_\theta\|^2 \;=\; \mathbb{E}\left[\left|\mathsf{loc}(\widehat{\theta}_X,\theta)\right|^2\right] - \left|\mathbb{E}\left[\mathsf{loc}(\widehat{\theta}_X,\theta)\right]\right|^2 \;\stackrel{\text{def}}{=}\; \mathsf{var}_{\widehat{\theta}}(\theta). \tag{6.11}$$

This inequality does not yet bear any resemblance with an uncertainty principle. It can be transformed into one if some conditions are put on μ_L and μ_M.

Lemma 6.1. *For L and M as in Definition 6.1, μ_L a constant function of θ and $\mu_M = 0$*

$$\mathsf{var}_{\widehat{\theta}}(\theta) \geq \frac{|\langle [L\,,\,M]\,p_\theta\,,\,p_\theta\rangle|^2}{\|Mp_\theta\|^2}, \tag{6.12}$$

Proof.

See Appendix F.1. □

The requirement of a constant μ_L is a constraint on the estimator weaker than being unbiased ($\mu_L = 0$). Moreover we will see that unbiasedness may not always be a sensible requirement while a constant bias is.

Lemma 6.1 will serve as a base from which we will derive several bounds on the variance (or the MSE) of estimators, which structure is broadly speaking of the form:

$$\text{Estimator Variance} \;\geq\; \frac{\text{Lack of commutation between } L \text{ and } M}{\text{“Fisher-like” information measure}}$$

The main advantage of using the uncertainty-like form (6.12) instead of (6.10) is to use results from Chapter 5 as intuitions, and to realize that the lack of commutativity between L and M is crucial to obtain strong lower bounds. The main drawback is the condition put on μ_L and μ_M; so whenever we will feel limited by them, we will go back to (6.10).

Example 6.b — A trivial lowerbound

It is well-known [45] that choosing M to be the identity — $Mp_\theta = p_\theta$ — yields the trivial lowerbound $\mathsf{var}_{\widehat{\theta}}(\theta) \geq 0$. The interpretation of this result in the light of Lemma 6.1 is that M and L commute, so that the numerator of the RHS in (6.12) vanishes.

We can now outline a plan to obtain meaningful and practically computable bounds from (6.12) using insight from Chapter 5 :

1. Design the localization operator L such that
   - The implied estimator variance and MSE definitions are meaningful,
   - The RHS of (6.12) can be evaluated in closed-form. No dependency on the estimator itself, except its first moment (unbiasedness, ...).
2. Find the operator M which maximizes the RHS of (6.12).

6.2.1 Preliminary example : the aperiodic and periodic CRB

Before diving into the two steps plan, we grind our teeth on the CRB, *i.e.* on the case where $Mp_\theta \overset{\text{def}}{=} \frac{d}{d\theta}p_\theta$.

First, one obtains a generic formulation of the CRB for a localization operator as in Definition 6.1.

Theorem 6.1. ***(generic CRB)***
An estimator $\widehat{\theta}_X$ with constant bias

$$\mu_L \;=\; \int_{\mathcal{X}} \mathsf{loc}(\widehat{\theta}_{\boldsymbol{x}},\theta)\cdot p_\theta(\boldsymbol{x})\boldsymbol{dx} \;=\; \text{constant}\;,$$

has its variance lowerbounded by

$$\mathsf{var}_{\widehat{\theta}}(\theta) \;\geq\; \frac{\left|\int_{\mathcal{X}}\left(\frac{d}{d\theta}\mathsf{loc}(\widehat{\theta}_{\boldsymbol{x}},\theta)\right)p_\theta(\boldsymbol{x})\boldsymbol{dx}\right|^2}{J_\theta}, \tag{6.13}$$

which is the CRB.

Proof.

First notice that $\mu_M = 0$, therefore we can apply Lemma 6.1.

The denominator in the RHS of (6.12) is *Fisher's information*

$$\|Mp_\theta\|^2 \;=\; \mathbb{E}\left[\left(\frac{d}{d\theta}p_\theta\right)^2 \Big/\, p_\theta^2\right] \overset{\text{def}}{=} J_\theta \tag{6.14}$$

For the numerator, one obtains

$$\left|\langle [L\;,\;M]\,p_\theta\;,\;p_\theta\rangle\right|^2 \;=\; \int_{\mathcal{X}}\left(\frac{d}{d\theta}\mathsf{loc}(\widehat{\theta}_{\boldsymbol{x}},\theta)\right)p_\theta(\boldsymbol{x})\boldsymbol{dx} \tag{6.15}$$

□

Theorem 6.1 applies immediately to various type of parameters :

Aperiodic parameter The usual localization operator is

$$\mathsf{loc}(\widehat{\theta}_X,\theta) \;=\; \widehat{\theta}_X - \theta,$$

and so Theorem 6.1 implies

$$\mathsf{var}_{\widehat{\theta}}(\theta) \;\geq\; J_\theta^{-1},$$

which is the classical CRB on the variance of an estimator.

Periodic parameter We assume $\theta \in [-\pi,\, \pi[$. Several choices can be made for localization. Because of the similarity with the Heisenberg principle, one can infer that

$$\mathsf{loc}(\widehat{\theta}_X, \theta) \;=\; (\widehat{\theta}_X - \theta)_{\mathrm{mod}\pi},$$

would lead to a dependency on the value of the distribution $\widehat{\theta}_X$ at $\theta = \pi$ in the commutator, and so the lowerbound would depend on the estimator distribution at $\theta = \pi$, which is problematic. This foresighted problem was observed in [106], and yielded a lowerbound dependent on the value of the estimator's CDF.

To avoid this issue, we will choose the localization definition (5.10) in Chapter 5 since it was shown to have many desirable properties

$$\mathsf{loc}(\widehat{\theta}_X, \theta) \;=\; 1 - e^{j(\widehat{\theta}_X - \theta)}. \tag{6.16}$$

We define the *centered first angular moment* of p_θ as

$$\tau(\theta) \;\stackrel{\text{def}}{=}\; \int_{\mathcal{X}} e^{j(\widehat{\theta}_{\boldsymbol{x}} - \theta)} p_\theta(\boldsymbol{x}) \boldsymbol{dx}, \tag{6.17}$$

and so

$$\mu_L(\theta) \;=\; \mathbb{E}\left[\mathsf{loc}(\widehat{\theta}_X, \theta)\right] \;=\; 1 - \tau(\theta).$$

It is important to note that assuming $1 - \tau(\theta) = 0$ is nonsensical. This leads to think of what motivates the unbiasedness assumption in the classical derivation of the CRB. If we assume the estimator distribution is centered on θ regardless of its value — which was the motivation behind the unbiased requirement in the aperiodic parameter case— then $\tau(\theta)$ is real-valued and constant. If the estimator deviates from θ by a constant amount, then $\tau(\theta)$ is constant and complex valued.

Hence, the substitute notion for "unbiasedness" with the localization (6.16) is to have a constant, real-valued, centered first angular moment. The substitute for a "constant bias" is to have a constant centered first angular moment, *i.e.* to have $\frac{d}{d\theta}\tau(\theta) = 0$.

We now have a truly periodic definition of the CRB

Corollary 6.1. ***(The periodic CRB)***
Let $\widehat{\theta}_X$ be an estimator of θ based on measurements X with a constant centered first angular moment $\tau(\theta) = \tau$ defined in (6.17). Then

$$\mathsf{var}_{\widehat{\theta}} \;=\; 1 - |\tau|^2 \;\geq\; \frac{1}{1 + J_\theta}\,, \tag{6.18}$$

where J_θ is the Fisher information *of the measurements X about θ :*

$$J_\theta \stackrel{\text{def}}{=} \mathbb{E}\left[\left(\frac{d}{d\theta}p_\theta\right)^2 \Big/ p_\theta^2\right].$$

Proof.

$$\begin{aligned}\mathsf{var}_{\widehat{\theta}}(\theta) &= \mathbb{E}\left[\left|1-e^{j(\widehat{\theta}_X-\theta)}\right|^2\right] - \left|\mathbb{E}\left[1-e^{j(\widehat{\theta}_X-\theta)}\right]\right|^2 \\ &= 1-\tau-\tau^*_{\widehat{\theta}}+1-(1-\tau-\tau^*_{\widehat{\theta}}+|\tau|^2) \\ &= 1-|\tau|^2,\end{aligned}$$

Then applying (6.13) from Theorem 6.1

$$1-|\tau|^2 \geq \frac{\left|\mathbb{E}\left[\frac{d}{d\theta}\left(1-e^{j(\widehat{\theta}_X-\theta)}\right)\right]\right|^2}{J_\theta} = \frac{|\tau|^2}{J_\theta}.$$

This inequality upperbounds $|\tau|^2$ which yields a lowerbound on the periodic variance $1-|\tau|^2$:

$$|\tau|^2 \leq \frac{J_\theta}{1+J_\theta} \quad\Rightarrow\quad 1-|\tau|^2 \geq \frac{1}{1+J_\theta}.$$

As a side observation, we got extremely lucky since the numerator in the RHS of (6.13) — which includes the commutator of L and M — had a dependence on $\widehat{\theta}$ which is function of its variance only. Therefore, we were able to "resorb" it in the variance itself. In general, this is not the case for any definition of L, as we have seen previously for the "modulo" definition of localization. □

The interest of Theorem 6.1 is its similarity with the classical CRB. The periodic variance ranges from 0 for a point-mass distribution to 1 for the uniform distribution among others.

The lowerbound is the inverse of the Fisher information regularized by the addition of 1 to avoid singularity, therefore the lowerbound ranges from 0 for an infinite amount of information ($J_\theta \to \infty$) to 1 if no information is present at all ($J_\theta = 0$).

Since $1/(1+t) \overset{\infty}{\to} 1/t$, the periodic CRB tends to the classical one as the amount of information increases. This is to be expected as good estimators' distribution will concentrate around the true parameter value, making the real-line parameter space approximation relevant.

This is also consistant with the definition of the periodic variance. Indeed, we saw in Chapter 5 that

$$\left|1-e^{jt}\right|^2 = t^2+\mathcal{O}(t^4),$$

which implies that the aperiodic and the periodic variance are similar for narrow distributions.

Example 6.c — Single pulse estimation (ctd.)

In the continuation of single pulse estimation example (Example 6.a), we estimate the centered first angular moment of the ML estimator

$$\widehat{\tau}_{\mathrm{ML}} \;=\; \mathrm{mean}\left\{ e^{j(\widehat{\theta}_{\mathrm{ML}}-\theta)} \right\} ,$$

where mean$\{\,\bullet\,\}$ is the *sample mean* of a random variable.

The Fisher information is then [31]

$$J_\theta \;=\; \frac{\left\| \frac{d}{d\theta}\boldsymbol{s}_\theta \right\|_2^2}{\sigma^2}.$$

Corollary 6.1 provides two inequalities to assess the dispersion of the ML estimator

$$1-|\tau_{\mathrm{ML}}|^2 \;\geq\; \frac{1}{1+J_\theta}\;, \quad \frac{1-|\tau_{\mathrm{ML}}|^2}{|\tau_{\mathrm{ML}}|^2} \;\geq\; J_\theta^{-1}\;.$$

The first one concerns what we defined as the periodic variance. It ranges from 0 to 1, which makes it intuitive to work with. The second inequality concerns the unbounded periodic variance. Having a definition of variance spanning $\mathbb{R}^+$ for a parameter ranging from $[-\pi,\ \pi[$ is not the most practical, but it is lower-bounded by the inverse Fisher information, just as in the aperiodic case.

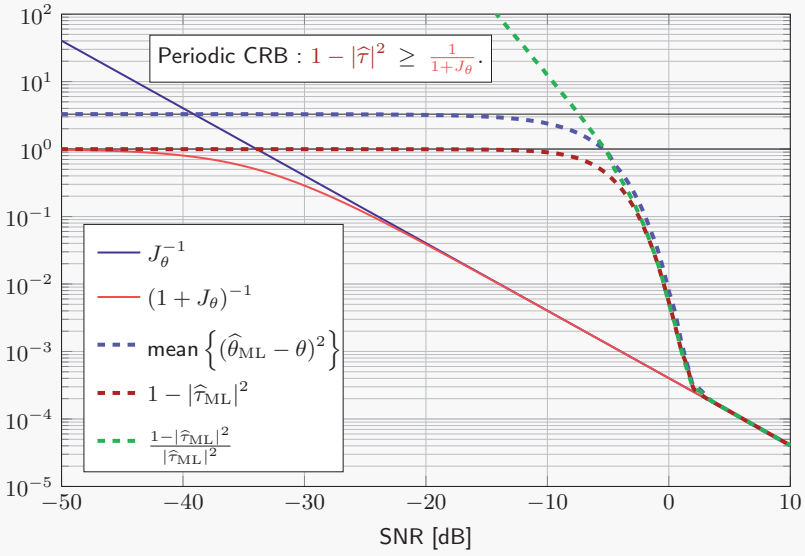


Figure 6.2: *The periodic variance of an estimator $1-|\widehat{\tau}|^2$ is a measure of spread well-suited for both a rigorous analysis and an intuitive interpretation of a periodic estimator's performances.*

Mixed multivariate estimation : Periodic & Aperiodic parameters The classical multivariate CRB establishes a lowerbound on the covariance matrix of a multivariate estimator by the inverse of Fisher's information matrix.

We derived a CRB for the periodic variance $1-|\tau|^2$ and also for a normalized definition of the variance

$$\frac{1-|\tau|^2}{|\tau|^2} \geq J_\theta^{-1}.$$

Since it has the exact same structure as the CRB for aperiodic parameters, a CRB on a heterogeneous collection of parameters ought to exist in a similar form

Theorem 6.2. ***(Mixed aperiodic/periodic multivariate CRB)***

Let $\widehat{\boldsymbol{\theta}}_X$ be an estimator of $\boldsymbol{\theta}$ a vector of K parameters. Each parameter θ_k is either periodic ($\theta_k \in [-\pi, \pi[$) or aperiodic ($\theta_k \in \mathbb{R}$), and for each parameter, we define its localization as

$$\mathsf{loc}(\widehat{\theta}_{X,k}, \theta_k) \overset{\text{def}}{=} \begin{cases} \frac{1-e^{j(\widehat{\theta}_{X,k}-\theta_k)}}{\mathbb{E}\left[e^{j(\widehat{\theta}_{X,k}-\theta_k)}\right]} & \theta_k \text{ is periodic} \\ \widehat{\theta}_{X,k}-\theta_k & \theta_k \text{ is aperiodic} \end{cases}$$

If $\mu_k(\boldsymbol{\theta}) \overset{\text{def}}{=} \mathbb{E}\left[\mathsf{loc}(\widehat{\theta}_{X,k}, \theta_k)\right]$ is constant with respect to $\boldsymbol{\theta}$ for all k, then[6]

$$\mathsf{cov}\left\{\mathsf{loc}(\widehat{\boldsymbol{\theta}}_X, \boldsymbol{\theta})\right\} \succeq \boldsymbol{J}_{\boldsymbol{\theta}}^{-1}, \tag{6.19}$$

where $\mathsf{cov}\left\{\mathsf{loc}(\widehat{\boldsymbol{\theta}}_X, \boldsymbol{\theta})\right\}$ is the covariance matrix with entries

$$\left[\mathsf{cov}\left\{\mathsf{loc}(\widehat{\boldsymbol{\theta}}_X, \boldsymbol{\theta})\right\}\right]_{k,\ell} \overset{\text{def}}{=} \mathbb{E}\left[\left(\mathsf{loc}(\widehat{\theta}_{X,k}, \boldsymbol{\theta}) - \mu_k(\boldsymbol{\theta})\right) \cdot \left(\mathsf{loc}(\widehat{\theta}_{X,\ell}, \theta_k) - \mu_\ell(\theta_k)\right)^*\right].$$

Proof.

See Appendix F.3. □

Evaluation of each of the diagonal coefficients of $\boldsymbol{J}_{\boldsymbol{\theta}}^{-1}$ yields a lowerbound on the estimator variance for the corresponding parameter. For aperiodic parameters, this is the usual definition of the variance, but for periodic parameters, this is the periodic variance normalized by $\left|\tau_{\widehat{\theta}_k}\right|^{-2}$. Since there is a one-to-one map between the periodic variance and its normalized counterpart, it follows that

[6] The relation $\boldsymbol{A} \succeq \boldsymbol{B}$ means $(\boldsymbol{A}-\boldsymbol{B})$ is *positive semidefinite*.

Corollary 6.2. *For a periodic parameter θ_k and $\widehat{\theta}_k$ an estimator with constant centered first angular moment $\tau_{\widehat{\theta}_k}$*

$$\mathit{var}_{\widehat{\theta}}(\theta) \stackrel{\text{def}}{=} 1 - \left|\tau_{\widehat{\theta}_k}\right|^2 \geq \frac{1}{1 + [\boldsymbol{J}_{\boldsymbol{\theta}}^{-1}]_{k,k}^{-1}} \tag{6.20}$$

This corollary makes it possible to use the periodic CRB in the presence of multiple *nuisance parameters* which have either a periodic or aperiodic support. The estimation of the multipath model developed in Chapter 1 is an example of a model combining periodic parameters (the time of arrivals) and aperiodic parameters (the path amplitudes).

Example 6.d — Multipath channel estimation

We consider a 2-paths with AWGN measurement model — *i.e.* a model with four unknown parameters, two aperiodic and two periodic — defined as

$$x_n = s_{\boldsymbol{\theta}}[n] + \sigma^2 \varepsilon_n \,, \quad 0 \leq n < 2M+1,$$

where

$$s_{\boldsymbol{\theta}}[n] = \theta_3 D_M\left(\frac{2\pi}{2M+1}n - \theta_1\right) + \theta_4 D_M\left(\frac{2\pi}{2M+1}n - \theta_2\right),$$

and D_M is the Dirichlet kernel of bandwidth $2M+1 = 31$, and ε_n are realizations of iid standard normal random variables.

The Fisher information matrix is [31]

$$\boldsymbol{J}_{\boldsymbol{\theta}} = \frac{1}{\sigma^2}\boldsymbol{\Phi}^*\boldsymbol{\Phi}\,, \quad \boldsymbol{\Phi} \stackrel{\text{def}}{=} \left[\frac{\partial \boldsymbol{s}_{\boldsymbol{\theta}}}{\partial \theta_1} \quad \cdots \quad \frac{\partial \boldsymbol{s}_{\boldsymbol{\theta}}}{\partial \theta_4}\right].$$

We first set $\boldsymbol{\theta} = [0.2,\ 1,\ 1,\ 1]^T$ such that

$$\boldsymbol{J}_{\boldsymbol{\theta}}^{-1} = \sigma^2 \cdot \begin{bmatrix} 0.0128 & -3.85\text{E-}5 & 1.75\text{E-}4 & 0.0164 \\ -3.85\text{E-}5 & 0.0128 & -0.0164 & -1.75\text{E-}4 \\ 1.75\text{E-}4 & -0.0164 & 1.0212 & 0.0142 \\ 0.0164 & -1.75\text{E-}4 & 0.0142 & 1.0212 \end{bmatrix},$$

then, with Theorem 6.2 and Corollary 6.2

$$\begin{aligned} \mathsf{var}_{\theta_1} = \mathsf{var}_{\theta_2} &\gtrapprox \frac{1}{1 + 78.35/\sigma^2}\,, \\ \mathsf{var}_{\theta_3} = \mathsf{var}_{\theta_4} &\gtrapprox 1.0212 \cdot \sigma^2\,. \end{aligned}$$

In this example, the gap between each pulse is large enough so that the joint estimation error is almost the same as in the univariate scenario

$$[\boldsymbol{J}_{\boldsymbol{\theta}}^{-1}]_{k,k} \approx 1/[\boldsymbol{J}_{\boldsymbol{\theta}}]_{k,k}.$$

This is obviously not the case if this gap is made much smaller as shown in Figure 6.3. In this case the multivariate bound in Theorem 6.2 is absolutely necessary. It is now clear that the fact that both the periodic and the aperiodic CRB are proportional to the inverse Fisher information is not just a nice to have property since it made possible to treat jointly periodic and aperiodic parameters.

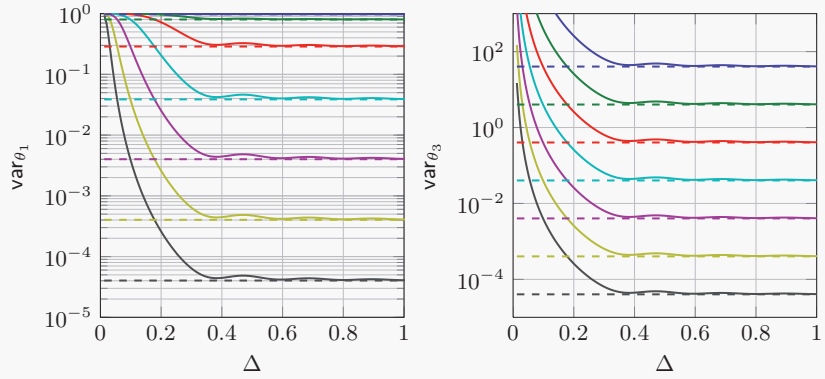


Figure 6.3: *The signal model is the one developed above with* $\boldsymbol{\theta} = [0.2,\ 0.2+\Delta,\ 1,\ 1]$. *The inverse bandwidth of the pulse shape is* $2\pi/31 \approx 0.2$, *and a rule of thumb in communications is to consider twice the inverse bandwidth — i.e.* 0.4 *— to be the distance at which interferences between paths are negligible. The dashed line is the CRB for independent estimation, i.e. if the values of the other parameters are known exactly.*

6.3 Replacing momentum to maximize the lowerbound

In Theorem 6.1, we obtained the CRB for a periodic parameter. It is of interest to see if this lowerbound can be made stronger by using a different "momentum" operator, as done in Barankin's work and its followings. Our starting point will be

Lemma 6.2. *A periodic estimator with constant centered first angular moment* τ *verifies*

$$\frac{1-|\tau|^2}{|\tau|^2} \geq \frac{\left|\int_\Theta (1-e^{jt})g(t)dt\right|^2}{\mathbb{E}\left[\left|(g(t)*p_t)(\theta)\right|^2 \Big/ p_\theta^2\right]}, \tag{6.21}$$

for any square integrable impulse response g.

Proof.

See Appendix F.2. □

Example 6.e — The periodic Hammersley-Chapmann-Robbins bound (HCRB)

The HCRB [44] was originally proposed as an alternative to the CRB for measurements distributions p_θ lacking derivability with respect to θ. The idea is to use a pseudo-differential filter in Lemma 6.2

$$g_\Delta(t) \overset{\text{def}}{=} \frac{\delta_t - \delta_{t-\Delta}}{\Delta},$$

implying

$$(g_\Delta * f)(\theta) = \frac{f_\theta - f_{\theta-\Delta}}{\Delta}.$$

Therefore as $\Delta \to 0$, one obtains the CRB if it exists. By choosing Δ such as to maximize the RHS[7] in (6.21), a lowerbound at least as tight as the CRB is obtained

$$\begin{aligned}\frac{1-|\tau|^2}{|\tau|^2} &\geq \max_{\Delta\in\Theta} \frac{\left|\int_\Theta (1-e^{jt}) g_\Delta(t) dt\right|^2}{\mathbb{E}\left[|(g(t) * p_t)(\theta)|^2 / p_\theta^2\right]}, \\ &= \max_{\Delta\in\Theta} \frac{\left|1-e^{j\Delta}\right|^2}{\mathbb{E}\left[\left(\frac{p_{X|\theta-\Delta}}{p_\theta}\right)^2\right] - 1}. \end{aligned} \tag{6.22}$$

The expectation in the denominator can be evaluated by numerical integration in general. For the AWGN measurement model (6.3) used in the previous examples, it has a simple form[8] which yields

$$\frac{1-|\tau|^2}{|\tau|^2} \geq \max_{\Delta\in\Theta} \frac{\left|1-e^{j\Delta}\right|^2}{e^{\left\|\boldsymbol{s}_\theta - \boldsymbol{s}_{\theta-\Delta}\right\|_2^2/\sigma^2} - 1}. \tag{6.23}$$

This periodic lowerbound is to be compared with the aperiodic HCRB [44]

$$\mathbb{E}\left[(\widehat{\theta}_X - \theta)^2\right] - \mathbb{E}\left[\widehat{\theta}_X - \theta\right]^2 \geq \max_{\Delta\in\Theta} \frac{\Delta^2}{e^{\left\|\boldsymbol{s}_\theta - \boldsymbol{s}_{\theta-\Delta}\right\|_2^2/\sigma^2} - 1}. \tag{6.24}$$

As expected from Chapter 5, the localization kernel Δ^2 is replaced with its periodic counterpart $\left|1-e^{j\Delta}\right|^2 = \Delta^2 + \mathcal{O}(\Delta^4)$. Therefore for Δ small enough, the RHS of both equations coincide, and they both tend to the inverse Fisher information.

As in (6.3) the signal $\boldsymbol{s}_\theta$ is the normalized sampled Dirichlet kernel of bandwidth $N = 2M+1$ such that

$$\left\|\boldsymbol{s}_\theta - \boldsymbol{s}_{\theta-\Delta}\right\|_2^2 = 2 \cdot (1 - D_M(\Delta)) \ , \quad D_M(\Delta) \overset{\text{def}}{=} \frac{\sin((2M+1)\Delta/2)}{(2M+1)\sin(\Delta/2)}. \tag{6.25}$$

So, for the signal model (6.3) specifically, the HCRB on the periodic variance is

$$1 - |\tau|^2 \geq \left(1 + \min_{\Delta \in \Theta} \left[\frac{e^{2(1-D_M(\Delta))/\sigma^2} - 1}{2(1-\cos\Delta)}\right]\right)^{-1}. \tag{6.26}$$

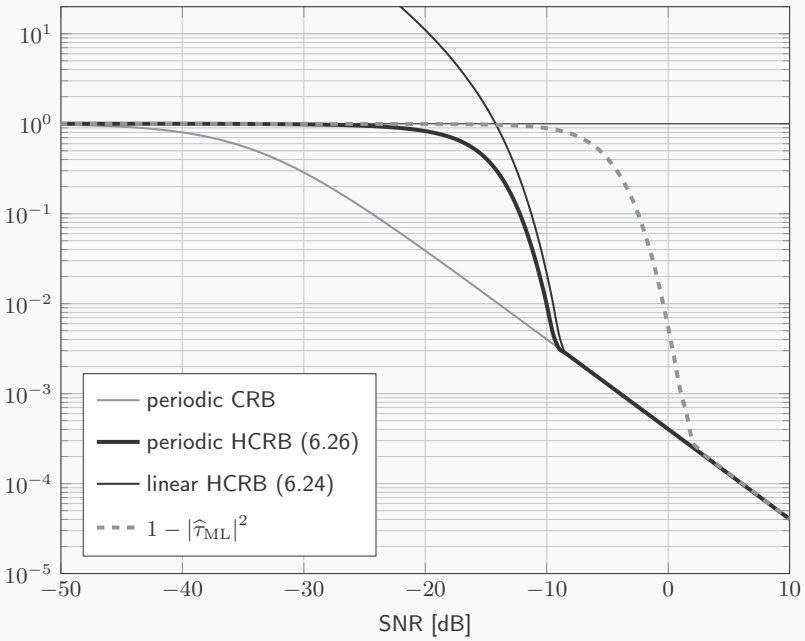


Figure 6.4: *The HCRB is by construction tighter than the CRB, and a threshold SNR is visible. This threshold SNR is significantly lower than the threshold SNR of the ML estimator.*

Note that a multivariate HCRB for a set of K aperiodic and periodic parameters can be easily obtained following the same line from Lemma 6.2, this is left as an exercise. The only difference is that the optimization of Δ is now done over a K-dimensional space which can become difficult for large values of K.

6.3.1 Finding the optimal filter ("momentum") with collinearity

The task is to find the impulse response g which maximizes the RHS of (6.21). It may not be feasible analytically for any distribution p_θ as the RHS of (6.21) depends on an expectation taken with respect to the measurements. Numerical integration or Monte-Carlo methods can solve this problem in general. More specifically, for AWGN measurements an analytical solution can be found.

Lemma 6.3. *Let θ be a periodic parameter and $\boldsymbol{x}$ be measurements corrupted by additive gaussian noise with distribution*

$$p_\theta(\boldsymbol{x}) \overset{\text{def}}{=} \mathcal{N}_{\mathbb{C}}(\boldsymbol{s}_\theta, \boldsymbol{\Sigma}),$$

where $\boldsymbol{s}_\theta$ is a vector depending deterministically on θ and $\boldsymbol{\Sigma}$ is the noise covariance matrix. Then

$$\mathbb{E}\left[|(g(t) * p_t)(\theta)|^2 \Big/ p_\theta^2 \right] = \int_\Theta g(t)\kappa\left(\theta - t, \theta\right)^{-1} \cdot \int_\Theta g^*(t')\kappa\left(\theta, \theta - t'\right)^{-1}\kappa\left(\theta - t, \theta - t'\right)dt'dt, \tag{6.27}$$

where

$$\kappa\left(t, t'\right) \overset{\text{def}}{=} e^{-\|\boldsymbol{s}_t - \boldsymbol{s}_{t'}\|^2_{\boldsymbol{\Sigma}^{-1}}/2}.$$

Proof.

See Appendix F.4. □

The quantity in (6.27) can be seen as a generalization of Fisher information.

It shows that for measurements corrupted by AWGN, the generalization of the Fisher information can be computed via a closed-form formula. We can combine this expression with Lemma 6.2 and use the property of the Cauchy-Schwarz inequality for collinear arguments to make the lowerbound as tight as possible. This results in solving a simple integral equation

Theorem 6.3. *The impulse response*[9] *g solution of*

$$(1 - e^{-jt}) \cdot \kappa\left(\theta - t, \theta\right) = \int_\Theta g(t')\kappa\left(\theta, \theta - t'\right)^{-1}\kappa\left(\theta - t, \theta - t'\right)dt', \tag{6.28}$$

an Harmonic Fredholm integral equation of the first kind*, maximizes the RHS of (6.21). This equation can be solved numerically using* Galerkin's method *or* finite elements*. Then*

$$\mathsf{var}_{\widehat{\theta}} = 1 - |\tau|^2 \geq \frac{1}{1 + \frac{1}{2\pi}(G_0 - G_{-1})^{-1}}, \tag{6.29}$$

where $G_n \overset{\text{def}}{=} \text{DTFS}\{g\}[n] = \frac{1}{2\pi}\int_\Theta g(t)e^{-jtn}dt$.

[9] We use the same symbol g for the optimal filter and an arbitrary one to avoid a cluttered notation, beware!

Proof.

With the expression (6.27), the RHS of (6.21) is

$$\frac{\left|\int_\Theta g(t)(1-e^{jt})dt\right|^2}{\int_\Theta g(t)\kappa\left(\theta-t,\theta\right)^{-1}\int_\Theta g^*(t')\kappa\left(\theta,\theta-t'\right)^{-1}\kappa\left(\theta-t,\theta-t'\right)dt'dt}$$

This fraction is maximized when the numerator is the squared magnitude of the denominator[10]which implies

$$1-e^{jt}\propto\kappa\left(\theta-t,\theta\right)^{-1}\int_\Theta g^*(t')\kappa\left(\theta,\theta-t'\right)\kappa\left(\theta-t,\theta-t'\right)dt'.$$

Scaling the magnitude of g does not change the value of the lowerbound, so without loss of generality, equality can be sought in the previous equation instead of proportionality, yielding (6.28) by taking the conjugate.

With g, the RHS of (6.21) is equal to $\int_\Theta g(t)(1-e^{jt})dt$, which is real-valued and positive. □

Theorem 6.3 provides a constructive way to obtain the Barankin bound for a periodic parameter. Its structure is remarkably similar to the CRB, here $\frac{1}{2\pi}(G_0-G_{-1})^{-1}$ replaces Fisher's information.

6.3.2 An analytical solution for shift-invariant signals

Solving a Fredholm equation — as (6.28) — is numerically an ill-conditioned[11] but well-understood problem [65; 64], and the literature about Ritz-Galerkin methods is abundant[12].

Despite these features, an analytical solution is of interest to gain insight on the inner-workings of the maximization problem.

If the kernel κ is a convolution kernel — *i.e.* if $\kappa\left(t,t'\right)=\kappa\left(t'-t\right)$ —[13] the eigenfunctions of the Fredholm equation (6.28) are the DTFT basis functions (DTFS' dual basis functions), and so the equation is analytically solvable in the DTFS domain.

[10]This principle is a consequence of the fact that Cauchy-Schwarz inequality is an equality for collinear inputs

[11]A small perturbation in the left hand side of (6.28) may drastically change the solution.

[12]See [59] for a remarkable historical perspective.

[13]We abused the notation by using the same symbol κ twice.

Corollary 6.3. *In the setup of Theorem 6.3, if κ is a periodic convolution kernel, $\kappa(t,t') = \kappa(t'-t)$, then*

$$\mathsf{var}_{\widehat{\theta}} \geq \frac{1}{1+\left(\sum_{\mathbb{Z}} \frac{K_{n+1}^2}{K_n} - 1\right)^{-1}} \tag{6.30}$$

where $K_n \overset{\text{def}}{=} \mathrm{DTFS}\{\kappa\}[n] = \frac{1}{2\pi}\int_{\Theta} \kappa(t) e^{-jtn} dt$.

Proof.

The equation (6.28) becomes

$$\begin{aligned}(1-e^{-jt})\cdot\kappa(t) &= \int_{\Theta} \underbrace{g(t')\kappa(-t')^{-1}}_{\widetilde{g}(t')} \kappa(t-t')dt', \\ &= (\widetilde{g} * \kappa)(t) \\ \overset{\text{DTFS}}{\longleftrightarrow} \qquad K_n - K_{n+1} &= 2\pi \cdot \widetilde{G}_n \cdot K_n, \end{aligned} \tag{6.31}$$

where $G_n = \sum_{m\in\mathbb{Z}} \widetilde{G}_m \cdot K_{-(n-m)} = \frac{1}{2\pi}\sum_{m\in\mathbb{Z}} \frac{(K_m - K_{m+1})\cdot K_{m-n}}{K_m}$.

Therefore,

$$G_0 = \frac{1}{2\pi}\sum_{m\in\mathbb{Z}} K_m - K_{m+1} = 0.$$

$$G_{-1} = -\frac{1}{2\pi}\sum_{m\in\mathbb{Z}} \frac{K_{m+1}^2}{K_m} + \frac{1}{2\pi}\underbrace{\sum_{m\in\mathbb{Z}} K_{m+1}}_{=\kappa(0)} = -\frac{1}{2\pi}\left(\sum_{m\in\mathbb{Z}} \frac{K_{m+1}^2}{K_m} - 1\right).$$

Plugging these values in (6.29) proves the result. □

Example 6.f — Single pulse estimation (ctd.)

In the continuation of the univariate example (Example 6.a), we can use Corollary 6.3 to obtain an exact computation of Barankin's bound — we showed in (6.25) that $\|\boldsymbol{s}_\theta - \boldsymbol{s}_{\theta-t}\|_2^2$ is function of t only, and so $\kappa(t,t') = \kappa(t'-t)$. Using Corollary 6.3 we obtain

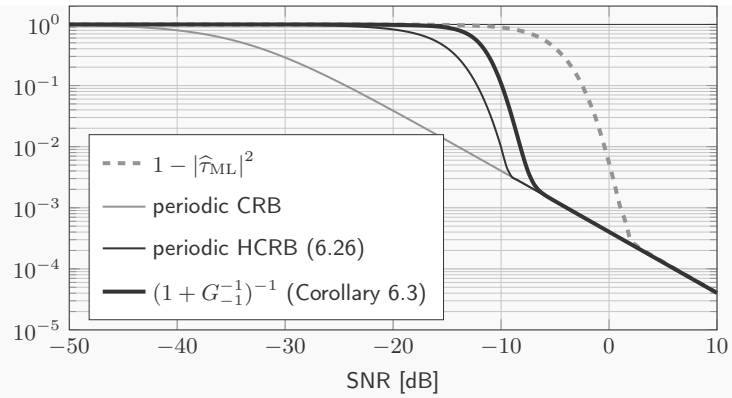


Figure 6.5: *The dominating lowerbound is the strongest lowerbound that can be achieved by replacing the derivative operator in the periodic CRB by an arbitrary LSI filtering operation. The threshold SNR is higher than the one of the HCRB, but still significantly lower than the threshold SNR of the NL estimator.*

It is an open question of whether or not the lowerbound can be achieved in practice. Nevertheless, it is shown in [131] that on this particular problem, the MSE performances between the MAP (which is equivalent to ML if the prior distribution is uniform) and the MMSE estimators with a uniform prior on the parameter are similar. It indicates that the gap is not entirely explainable by a deficiency of the ML estimator.

Example 6.g — Modulated pulse estimation

We have just seen that the Barankin bound falls short of detecting the transition between the small and large error regimes. Another type of transition can happen when the signal possess features with different resolutions. For example consider the model used in the single pulse estimation examples, but the pulse shape is now modulated with a phasor of known frequency f_0 significantly larger than the bandwidth of D_M the Dirichlet kernel of bandwidth $2M+1$

$$s_\theta[n] \;=\; e^{jf_0(\frac{2\pi}{2M+1}n-\theta)} \cdot D_M\left(\frac{2\pi}{2M+1}n-\theta\right)\;,\quad n\;=\;0,\;\ldots,\;2M\;,$$

in which case

$$\|\boldsymbol{s}_\theta - \boldsymbol{s}_{\theta+t}\|_2^2 \;=\; 2\cdot(1-\cos(f_0 t)\cdot D_M(t)).$$

Hence, the kernel κ is a convolution kernel, therefore we can use Corollary 6.3 to compute the lowerbound obtained with an optimal filter g.

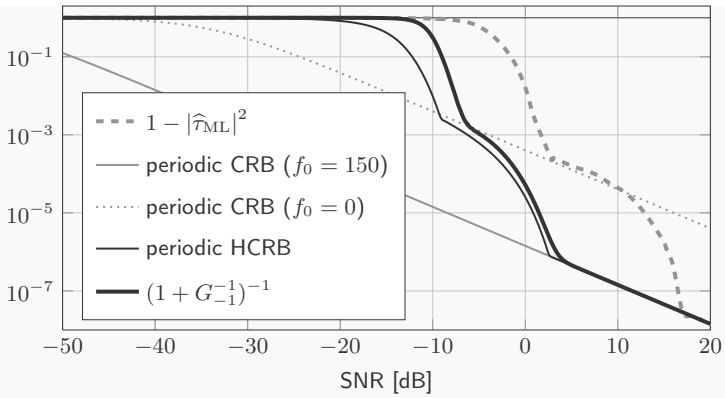


Figure 6.6: *The estimation has an additional regime at high SNR where the error depends on the frequency of the modulation.*

The intuition is that the signal now has an autocorrelation oscillating at f_0 and shaped by D_M. For a low enough noise level, the rapid oscillations of the autocorrelation function can be exploited to pin-point more accurately the exact location. As the noise power increases, errors start to be made between adjacent oscillation peaks, and the shape of D_M dictates the estimation accuracy. Finally, for a large enough noise, large spurious errors appears as in the previous example.

The CRB is based on the second derivative of the likelihood function with respect to θ around the true parameter value, and so it is mostly influenced by the fast oscillation of the cosine.

The bound proposed in Corollary 6.3 and the HCRB is not limited to this extremely local view of the signal and has the ability to detect — loosely — the aforementioned transition.

The slight difference between the HCRB and the optimal filter bound makes it relevant to consider the HCRB for signals which do not lead to a convolution equation.

6.3.3 Discussion about the "gap" between the Barankin bound and the ML estimator

The two examples illustrating the use of Corollary 6.3 seem to indicate there is a gap between the accuracy achievable in practice and lowerbounds based on the Cauchy-Schwarz inequality. A comparison is necessary with a minimum MSE (MMSE) estimator to judge definitively this hypothesis. We point to [131] for a comparison between the ML and MMSE estimators in times of arrival estimation. In [131] p.8 , the MMSE estimator is shown to have only a marginally lower MSE compared to the maximum a-posteriori (MAP) estimator[14] on the estimation of a modulated pulse as used in Example 6.g. So the gap between the Barankin bound and feasible estimators is not an artefact.

[14]The test uses a uniform prior for the parameter, therefore the MAP and ML estimators are the same. Note also that the localization operator is simply the difference, which does not cause an issue for the detection of the threshold SNR.

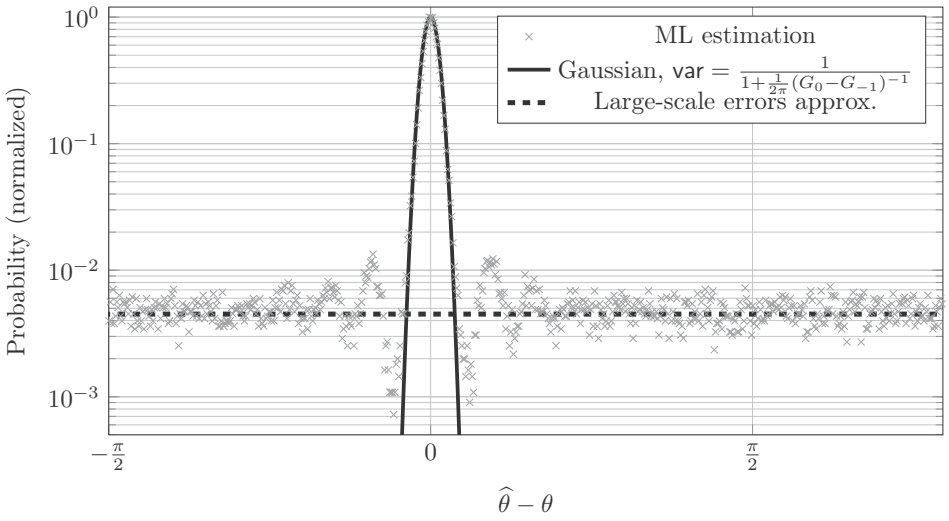


Figure 6.7: *Estimation error distribution, for the estimation of a single time of arrival θ. The SNR is $-5dB$, and corresponds to the regime where a gap is visible between the MSE of the maximum-likelihood estimator and the MSE of the CRB or the proposed Barankin bound. The fitted distributions are a gaussian distribution which variance matches the Barankin bound and a uniform distribution approximating the large scale errors. This approximation is crude, and does not capture the ripples visible on the empirical data.*

Inspecting the distribution of the ML estimator reveals that it can be approximated as a mixture of a gaussian like distribution and a uniform distribution — see Figure 6.7. The uniform distribution modelizes the large errors which are unrelated to the signal.

In an informal way the gap between the optimal filter lowerbound and the performances of the ML (or MMSE) estimator can be explained by looking at the first two DTFS coefficients of the estimator distribution[15].

Let $p_{\widehat{\theta}|\theta}$ be the distribution of an estimator $\widehat{\theta}$ with constant first angular moment and $\{P_n\}_{\mathbb{Z}}$ its DTFS coefficients. Its variance is

$$\mathsf{var}_{\widehat{\theta}} \;=\; 1-|\tau|^2 \;=\; 4\pi^2\left(|P_0|^2-|P_1|^2\right) \;\geq\; \underbrace{\frac{1}{1+\frac{1}{2\pi}\left(G_0-G_{-1}\right)^{-1}}}_{\stackrel{\text{def}}{=}\gamma}\;.$$

The coefficient P_0 of a probability distribution is $\frac{1}{2\pi}$ by definition, and the lowerbound states that

$$2\pi\,|P_1| \;\leq\; \sqrt{1-\gamma}\;,$$

[15]This is not a proof, rather an interpretation.

i.e. , that the DTFS coefficients must drop by at least $(1 - \sqrt{1-\gamma})/2\pi$ around the origin. A sharp decay of the DTFS spectrum implies a "wider" time-domain distribution, and so the lowerbound rules out distributions which are too concentrated.

Assume the distribution meets the lowerbound, *i.e.*

$$2\pi\,|P_1| \;=\; \sqrt{1-\gamma}\;,$$

then the addition of large scale errors yields a mixture with a uniform distribution. The resulting DTFS coefficients are

$$\widetilde{P}_0 \;=\; (1-\lambda)P_0 + \lambda/2\pi \;=\; \frac{1}{2\pi} \quad , \qquad \widetilde{P}_1 \;=\; (1-\lambda)P_1\;,$$

where λ controls the weight of the uniform distribution. For $\lambda \gg 0$ we have $\widetilde{P}_1 \ll P_1$, which means the lowerbound is not met.

Hence if at some point the noise is such that the large scale errors are not negligible, the distribution will not meet the lowerbound. To detect the transition, one must take into account the large scale errors as in [146] or (indirectly) as in the Ziv-Zakaï lowerbound [149] where the knowledge of an optimal binary hypothesis test error distribution is required.

6.4 Conclusion

In this chapter, we observed that

- Bounds of the Cramér-Rao family can be computed for the periodic case only using an assumption on the bias as classically done for aperiodic parameters — *i.e.* without resorting to the distribution of the MMSE estimator or Bayesian interpretations.
- Using collinearity to tighten the application of the Cauchy-Schwarz inequality did not yield a tight lowerbound on the tested signals — *i.e.* Barankin-type lowerbounds seems to not always be tight
- The simple and versatile periodic Hammersley-Chapman-Robbins bound showed to be almost as tight as the Barankin bound on the tested signals, so it may be used as a simpler alternative.
- Deterministic bounds on the variance of an estimator can be written in a form similar to Heisenberg's uncertainty principle. Sufficient requirements for this analogy are to have a constant bias and to use a "momentum" operator which kills constant signals (such as derivatives or pseudo-derivatives of various orders). It shows that the lack of commutativity between the localization operator (linked to the measure of error) and the momentum operator (chosen when designing the bound) is key.

It is an open question whether or not momentum operators other than linear filters could yield a tighter bound.

Conclusion

Part I

We have seen that a robust joint estimation of multipath channels can be achieved in a time which his superlinear with respect to the number of measurements. This property becomes decisive when the number of measurements exceeds the number of degrees of freedom in the model by at least one order of magnitude.

To apply parametric estimation techniques successfully in an ever changing environment, we proposed model selection criteria which use different properties of the signal and the noise. The selection of a particular detection method depends on the mismatch between the proposed model and the measurements E.g. we may use specific properties of an additive white gaussian noise (AWGN) if we are confident that the signal is corrupted by a similar enough noise.

We saw that highly specific models, such as the sparse common support model, have by definition a restricted but relevant range of application. So we had to reduce our playing field (in terms of bandwidth, physical distances, ...) which is not pleasant in scientific research, where universal abstractions are sought after. Nevertheless, a clear definition of the range of application of a model may be worth as much as the model itself. We obtained a preliminary validation of the proposed models and algorithms for mobile communications by testing them on CIR measurements with the addition of AWGN; further tests would be required.

Further work

Since a single class of channel models cannot fit all plausible realizations, the selection of a model among different classes — e.g. SCS models, models with sample sparsity in the time-domain and time limited models — could have an important impact on the actual performances of channel estimation. We only tried to combine the SCS channel model with the time limited channel model in Section 3.5 (if the detection criterion failed, the estimation used the time limited model). Also, heterogenous models could be formed, e.g. a SCS channel model with the addition of a residual with sparse samples in the time domain.

Part II

We studied the notion of localization for periodic phenomena, starting with the time-frequency product of periodic waveforms. We constructed analytically and numerically periodic waveforms which have a minimal time-frequency product. By doing so, we noticed that the use of the phase (angle) was troublesome for localization as it has a discontinuity, and we adopted a complex-valued definition to avoid this discontinuity. This definition was already well-known in the domain of time-frequency analysis.

Then, we showed that the Cramér-Rao bound (and bounds of the same family) and the Heisenberg uncertainty principle are formally similar[16]. We thus obtained new lowerbounds on the variance of unbiased estimators of periodic parameters by transposing the knowledge available in the time-frequency analysis literature.

The localization/momentum formulation of uncertainty principles provides much freedom in the design of the associated linear operators. For a single periodic parameter, we designed a linear shift-invariant (LSI) filter (the momentum operator) which maximizes the lowerbound among all possible LSI filters, and observed that a significant gap may still be visible between the lowerbound and the best estimator achievable in practice.

Further work

The bridge made between estimation error lowerbounds of the Cramér-Rao family and the Heisenberg uncertainty principle could lead to interesting generalizations. For example, an uncertainty principle for graphs was recently proposed in [3] and it would be of interest to see if it leads to interesting lowerbounds on the localization error on graphs[17]. Unfortunately, time did not permit to include such a study in this work.

[16] We do not say they are equivalent as they have different prerequisites.

[17] The first candidate would be a Cramér-Rao bound for graphs which would derive the momentum operator from the Laplacian of the graph.

Appendix A

Spatial correlation formula for fading channels

A.1 Azimuthal scatterers density distribution

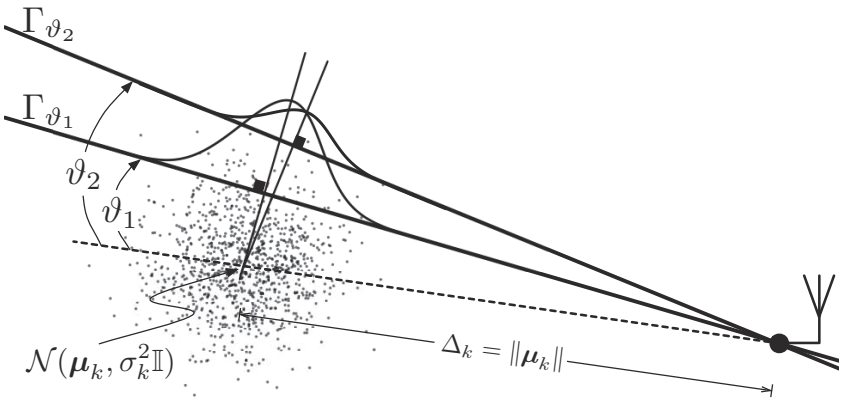


Figure A.1: *Azimuthal density of reflections at a receiving antenna*

The reflection density of each scatterer is normally distributed with mean $\boldsymbol{\mu}_k$ (its position) and covariance matrix $\sigma_k^2\mathbb{I}$ (its "girth") as seen in Figure A.1. The number of reflections within a scatterer is assumed to be large enough to warrant their approximation by their continuous probability density function. The azimuthal density is the integral of the scatterer's pdf over Γ_ϑ the straight path from the receiving antenna at an angle[1] ϑ:

$$p(\vartheta; \boldsymbol{\mu}_k, \sigma_k^2) = \int_{\mathbb{R}^2} f_{\sigma_k^2}^{(2D)}(\boldsymbol{x} - \boldsymbol{\mu}_k)\mathcal{I}_{\boldsymbol{x}\in\Gamma_\vartheta}\, d\boldsymbol{x}.$$

[1]Without loss of generality the scatterer origin is at azimuth 0, and the antenna is located at position **0**

Reparametrization in polar coordinates yield:

$$\begin{aligned} p(\vartheta;\boldsymbol{\mu}_k,\sigma_k^2) &= f_{\sigma_k^2}(\|\boldsymbol{\mu}_k\|\sin(\vartheta))\cdot\int_{\mathbb{R}_+} f_{\sigma_k^2}(r-\|\boldsymbol{\mu}_k\|\cos(\vartheta))J_{\boldsymbol{x}}(r,\vartheta)dr, \\ &= \sigma_k^{-1} f\left(\sqrt{\kappa_k'}\sin(\vartheta)\right)\cdot\int_{\sqrt{\kappa_k'}\cos(\vartheta)}^{+\infty}\sigma_k^{-1}f(r-\|\boldsymbol{\mu}_k\|\cos(\vartheta)) \\ &\qquad\cdot\left(s+\sqrt{\kappa_k'}\cos(\vartheta)\right)\sigma_k^2 ds \end{aligned}$$

such that $\kappa_k' = \|\boldsymbol{\mu}_k\|^2/\sigma_k^2$ and $J_{\boldsymbol{x}}(r,\vartheta) = r$ is the Jacobian of the cartesian to polar transformation. We performed the change of variable $s = r - \sqrt{\kappa_k'}\cos(\vartheta)$. Hence, the distribution has only one degree of freedom, and after some calculus:

$$p_{\kappa_k'}(\vartheta) = f(\sqrt{\kappa_k'}\sin\vartheta)\cdot\left[\sqrt{\kappa_k'}\cos\vartheta\cdot F(\sqrt{\kappa_k'}\cos\vartheta)f(\sqrt{\kappa_k'}\cos\vartheta)\right]. \tag{A.1}$$

The circular distribution (A.1) is well approximated by a Von-Mises distribution of scale κ_k:

$$q_{\kappa_k}(\vartheta) = \frac{e^{\kappa_k\cos\vartheta}}{2\pi I_0(\kappa_k)}.$$

where I_0 is the 0^{th} order modified Bessel function of the first kind. Asymptotically, $\kappa_k' \overset{\kappa_k'\to\infty}{=} \kappa_k$, and the approximation $\kappa_k' \approx (1-e^{-3\kappa_k/4})\kappa_k$ was found to be empirically accurate for all κ_k (K-L divergence between $p_{\kappa_k'}$ and q_{κ_k} is less than 0.02 bits).

A.2 Derivation of the correlation matrix formula

Considering the setup of Figure 1.4, and from [110]:

$$\frac{\mathbb{E}\left[C_{k,m}C_{k,n}^*\right]}{\sqrt{\mathbb{E}\left[|C_{k,m}|\right]\mathbb{E}\left[|C_{k,n}|\right]}} = \int_{-\pi}^{\pi} q_{\kappa_k}(\vartheta-\theta_{k,m,n})e^{j\frac{\omega_c}{c}d_{m,n}\sin\vartheta}d\vartheta.$$

Then, q_{κ_k} is expanded in terms of spherical harmonics via the Jacobi-Anger expansion [2](9.1):

$$\begin{aligned} q_{\kappa_k}(\vartheta\theta_{k,m,n}) &= \frac{1}{2\pi I_0(\kappa_k)}\left\{J_0(-j\kappa_k) + \sum_{l=1}^{\infty} j^l J_l(-j\kappa_k)\cos\left[l(\vartheta-\theta_{k,m,n})\right]\right\}, \\ &= \frac{1}{2\pi} + \frac{1}{\pi I_0(\kappa_k)}\sum_{l=1}^{\infty} I_l(\kappa_k)\cos\left[l(\vartheta-\theta_{k,m,n})\right], \end{aligned}$$

where the second equality is obtained with $I_l(jx) = j^l J_l(x)$ [2](9.6.3, 9.1.35).

We now have a series with l^{th} term:

$$\begin{aligned}
& \frac{I_l(\kappa_k)}{\pi I_0(\kappa_k)} \int_{-\pi}^{\pi} \cos\left[l(\vartheta - \theta_{k,m,n})\right] e^{j\frac{\omega_c}{c} d_{m,n} \sin\vartheta} d\vartheta \\
\stackrel{(a)}{=} \; & \frac{I_l(\kappa_k)}{\pi I_0(\kappa_k)} \left\{ \cos\left[l\left(\theta_{k,m,n} - \frac{\pi}{2}\right)\right] \cdot \int_{-\pi}^{\pi} \cos l\vartheta \; e^{j\frac{\omega_c}{c} d_{m,n} \cos\vartheta} d\vartheta \right. \\
& \qquad \left. + \sin\left[l\left(\theta_k - \theta_{m,n} - \frac{\pi}{2}\right)\right] \cdot \int_{-\pi}^{\pi} \sin l\vartheta \; e^{j\frac{\omega_c}{c} d_{m,n} \cos\vartheta} d\vartheta \right\} \\
\stackrel{(b)}{=} \; & \frac{2I_l(\kappa_k)}{I_0(\kappa_k)} I_l\left(j\frac{\omega_c}{c} d_{m,n}\right) \cos\left[l\left(\theta_{k,m,n} - \frac{\pi}{2}\right)\right] \\
\stackrel{(c)}{=} \; & \frac{2j^l I_l(\kappa_k)}{I_0(\kappa_k)} J_l\left(\frac{\omega_c}{c} d_{m,n}\right) \cos\left[l\left(\theta_{k,m,n} - \frac{\pi}{2}\right)\right]
\end{aligned}$$

Equality (a) is obtained with some standard trigonometric identities and a shift by $-\frac{\pi}{2}$ of the variable of integration. Equality (b) follows from the standard integral representation of I_l ([2] 9.6.19). The second integrand is antisymmetric which leads the integral over the unit-circle to vanish. Finally (c) is a consequence of $I_l(jx) = j^l J_l(x)$ again. Hence, with $\lambda_c = c/\omega_c$:

$$\begin{aligned}
\frac{\mathbb{E}\left[C_{k,m} C_{k,n}^*\right]}{\sqrt{\mathbb{E}\left[|C_{k,m}|\right] \mathbb{E}\left[|C_{k,n}|\right]}} \; = \; & J_0\left(2\pi \frac{d_{m,n}}{\lambda_c}\right) + \\
& 2\sum_{l=1}^{\infty} j^l \frac{I_l(\kappa_k)}{I_0(\kappa_k)} J_l\left(2\pi \frac{d_{m,n}}{\lambda_c}\right) \cdot \cos\left[l\left(\theta_{k,m,n} - \frac{\pi}{2}\right)\right].
\end{aligned}$$

Appendix B

Estimation algorithms

B.1 Further numerical tests

For simulations we use the fading SCS channel model. Its characteristics are listed in Table B.1. We assume 63 pilots which are uniformly spaced in frequency, one every 8. The transmitted frame is circularly padded such as to guarantee circular convolution of the transmitted signal with the CIR. Results are derived from three different experiments:

- A The medium has two paths separated by $2T$. The second path's expected amplitude is $1/10^{th}$ of the expected amplitude of the first path. The receiver possesses 1, 2, 4 or 8 uncorrelated antennas. The channels have exact SCS ($\varepsilon = 0$).
- B The medium has two paths separated by T or $2T$. Both paths have the same expected amplitude. The receiver has 4 uncorrelated antennas. The channels have either exact SCS ($\varepsilon = 0$) or non-exact SCS ($\varepsilon = T/50 = 1$ns). The discrepancy in the ToA between antennas is uniformly distributed in $[-\varepsilon\ \varepsilon]$. A time lapse of $2T/50$ corresponds to a path length difference of 60 cm.

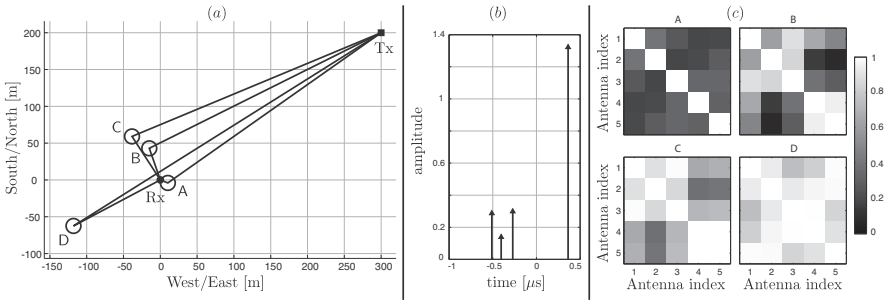


Figure B.1: *(a) The physical layout of the channel for Exp. C . The channel has four scatterers labeled A, B, C, and D. (b) The expected CIR of the channels. (c) Matrix entry (i, j) is the modulus of the fading correlation between antenna i and antenna j. Each matrix corresponds to a given scatterer.*

Table B.1: *Simulation parameters*

Parameter	Symbol	Value
Sampling step	T	50ns
Bandwidth	B	20MHz
Center frequency	f_c	2.6GHz
Frame duration (without padding)	τ	25.55μs
Samples per frame	N_{frame}	511
Pilots per frame	N	63
Pilot gap	D	8
Delay spread	Δ	1.6μs

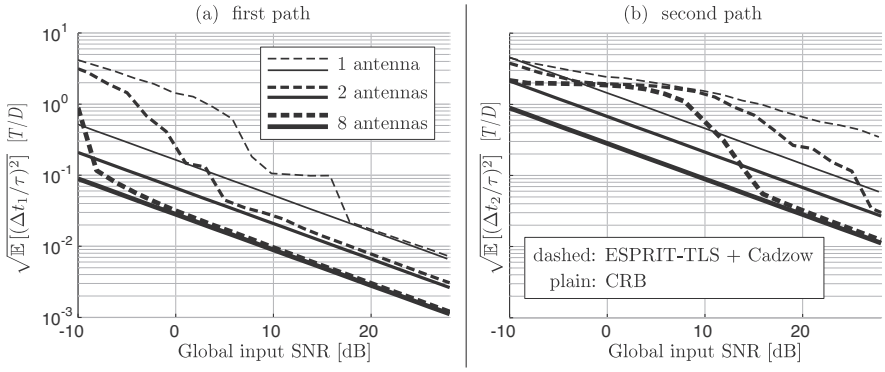


Figure B.2: *(Exp. A) For the same global input SNR, a system with more antennas estimates the ToAs more accurately and is more resilient to noise. This is a consequence of the increased receiver diversity. The second path has* $1/10^{th}$ *the amplitude of the first path and is thus quickly buried into noise as SNR decreases. The estimation reaches the Cramér-Rao bound as long as it correctly identifies the path.*

C This experiment is more realistic from a physical standpoint. The receiver has 5 antennas equi-spaced on a circle of radius 10 cm. The propagation medium contains 4 scatterers (Figure B.1.(a)). The expected CIR modulus is represented in Figure B.1.(b). We use the spatial correlation model derived in Proposition 1.2. Also the channels do not have exactly a common support, with a maximum delay of $\varepsilon = T/50 = 1$ns between channels.

Results were obtained on 400 independent noise and fading realisations.

B.1.1 Results on Exp. A

Figure B.2 shows that the SCS-FRI algorithm efficiently estimates the ToA down to a certain SNR where the recovery breaks down. This breaking point is pushed lower as spatial diversity increases, which is to be expected. Figure B.3 compares the use

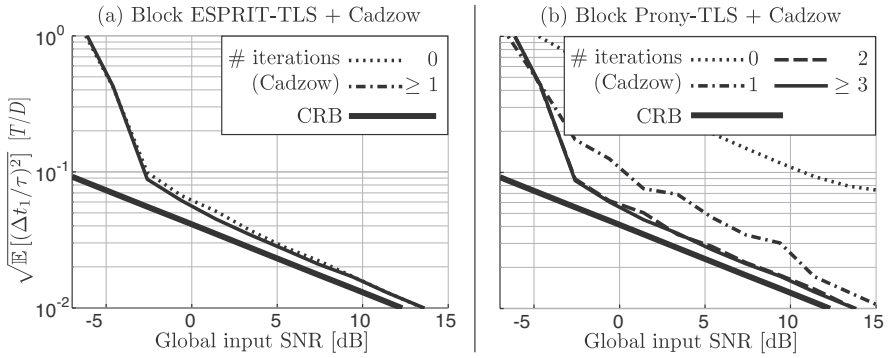


Figure B.3: *(Exp. A) Part (a) shows the performances of ESPRIT-TLS with or without Cadzow denoising. In this setup, the gain obtained with the denoising is relatively small and is achieved after one iteration. Part (b) shows the performances of annihilating filter-TLS with or without Cadzow denoising. As expected, the performance of Prony's algorithm without denoising is very poor. After 3 denoising iterations, performances of Annihilating Filter-TLS and ESPRIT-TLS are indistinguishable.*

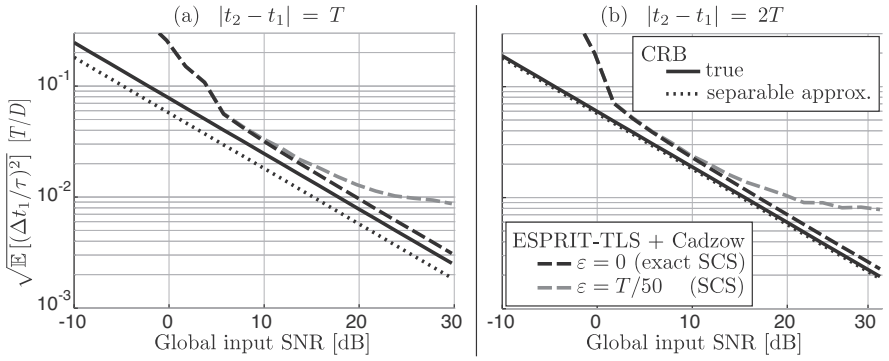


Figure B.4: *(Exp. B) This figure shows that the proposed algorithms behave as expected in the presence of ToA mismatches between antennas. Part (b) motivates the separability assumption to compute the CRB of paths located more than $2T$ apart, while Part (a) shows its inadequacy for a smaller delay T. The "true" estimate is obtained via Monte-Carlo simulations.*

and combination of the various subspace identification techniques discussed earlier. The conclusion is that the performances of Block-ESPRIT TLS or Block-Prony TLS are exactly the same on a signal denoised with the Block-Cadzow algorithm. However Block-ESPRIT TLS requires fewer to none Block-Cadzow iterations than Block-Prony TLS to reach the optimum. It is well-known that Prony TLS is not robust to noise [31; 129].

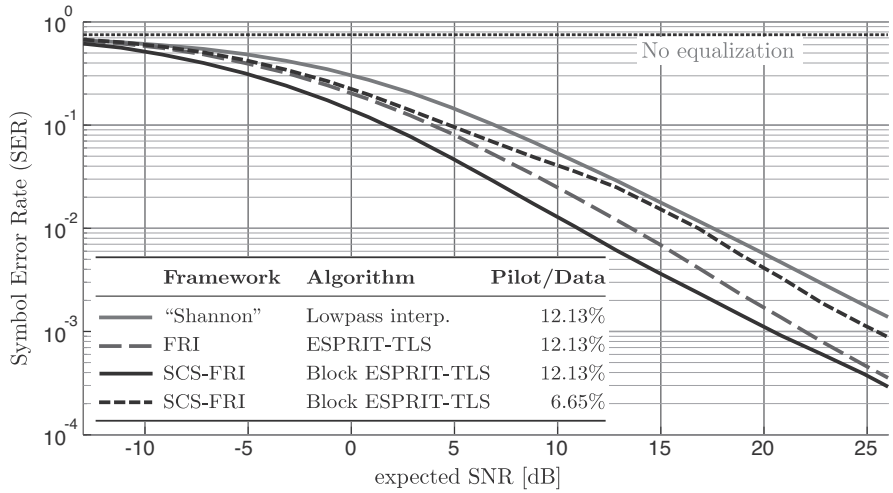


Figure B.5: *(Exp. C) Using the sparse and common support properties, the SER is decreased by a factor* 5 *above* $10dB$ *of SNR compared to the conventional non-parametric approach. Sparsity alone provides a significant SER improvement, which shall be combined with the common support property below* $30dB$ *of SNR. At very high SNR, independent channel estimation across antennas become preferable as the channels only approximately have the common support property. However, below* $15dB$ *of SNR the effect of this approximation are undetectable. Another advantage of joint sparse estimation is the reduction of pilots, it allows to halve their number while retaining performances superior to lowpass interpolation.*

B.1.2 Results on Exp. B

Figure B.4 shows that the single path CRB given in [26] (9) is a good approximation of the true bound computed via[26] (10) for multiple paths separated by more than twice the inverse bandwidth of the channel. This experiment also verifies the usefulness of the SCS assumption when ToAs are slightly perturbed from one antenna to another:

$$t_{k,p} = t_k + E_{k,p} \ , \quad E_{k,p} \sim \mathrm{U}([-\varepsilon \ \varepsilon]), \text{ i.i.d.}$$

The error caused by the random perturbation $E_{k,p}$ is of the order of the perturbation itself, and thus we may say SCS-FRI is robust on non exact SCS channels.

B.1.3 Results on Exp. C

All estimation algorithms use the fact that the delay spread is much shorter than the frame length. The difference between lowpass interpolation and other techniques is the use of the sparsity property. Using this property alone, the SER is halved at a SNR of 5dB as shown in Figure B.5. The addition of the SCS property proves to be valuable, at 5dB of SNR the SER is decreased by a factor 3. At high SNR, the SCS property provides a factor 5 of improvement over lowpass interpolation. At very high

SNR the error due to the approximate SCS nature of the channels diminishes this gain, and eventually the SCS assumption becomes detrimental.

It also shows that the number of pilots can be halved while having SER performances superior to the non-parametric approach (we retained half of the original pilots closest to the carrier frequency). For lowpass interpolation, this cannot be done without introducing aliasing. Reducing the number of pilots below "Nyqvist" is relevant at high SNR where little redundancy is required for denoising, leaving some additional spectrum for data transmission. In favorable transmission conditions, it would be possible to reduce the number of pilots down to the rate of innovation of the channel to maximize the data throughput.

B.2 Proof of Theorem 2.2

Decompose the data matrix as

$$\boldsymbol{T} = \boldsymbol{T}_{\text{sig}} + \boldsymbol{T}_{\text{noise}} \ ,$$

where $\boldsymbol{T}_{\text{noise}}$ is the data matrix obtained from the sequence $E[m]$.

Therefore the autocorrelation matrix can be seen as the noiseless autocorrelation plus a perturbation

$$\boldsymbol{T}^*\boldsymbol{T} \ = \ \boldsymbol{T}_{\text{sig}}^*\boldsymbol{T}_{\text{sig}} + \underbrace{\boldsymbol{T}_{\text{sig}}^*\boldsymbol{T}_{\text{noise}} + \boldsymbol{T}_{\text{noise}}^*\boldsymbol{T}_{\text{sig}} + \boldsymbol{T}_{\text{noise}}^*\boldsymbol{T}_{\text{noise}}}_{\stackrel{\text{def}}{=} \boldsymbol{E} \quad \text{(perturbation)}} .$$

Then, the following Lemma holds

Lemma B.1. *For all* $m \in \{1, \ldots, M+1\}$,

$$\lambda_m(\boldsymbol{T}^*\boldsymbol{T}) \leq \lambda_m(\boldsymbol{T}_{\text{sig}}^*\boldsymbol{T}_{\text{sig}}) + \|\boldsymbol{E}\| \, .$$

Proof.

One may use Corollary 6.3.4 in [68] together with *Weyl's theorem* (Theorem 4.3.1 in the same book). □

The spectrum of $\boldsymbol{T}_{\text{sig}}^*\boldsymbol{T}_{\text{sig}}$ The data matrix $\boldsymbol{T}_{\text{sig}}$ has a Vandermonde decomposition as defined in Definition 2.2, therefore

$$\boldsymbol{T}_{\text{sig}}^*\boldsymbol{T}_{\text{sig}} \ = \ \sum_p \mathcal{V}_{M+1}\boldsymbol{D}_p^*\mathcal{V}_{M+1}^*\mathcal{V}_{M+1}\boldsymbol{D}_p\mathcal{V}_{M+1}^*.$$

Moreover, the columns of the Vandermonde matrix $\mathcal{V}_{M+1}$ form asymptotically an orthogonal set as M grows. More formally, if $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are two such column vectors,

$$\lim_{M\to\infty} \frac{\langle \boldsymbol{v}_i \, , \, \boldsymbol{v}_j \rangle}{\|\boldsymbol{v}_i\| \, \|\boldsymbol{v}_j\|} \ = \ \delta_{i-j}.$$

It implies that asymptotically as $M \to \infty$, the Vandermonde decomposition of $\boldsymbol{T}_{\text{sig}}$ is a unitary diagonalization— with the proper normalization by $1/(M+1)$. Hence

$$\lim_{M\to\infty} \lambda_\ell(\boldsymbol{T}_{\text{sig}}^* \boldsymbol{T}_{\text{sig}}) = \begin{cases} (M+1)^2 \cdot \sum_p |C_{k,p}|^2 & , \ k \leq K \ , \\ 0 & , \ \text{else.} \end{cases} \tag{B.1}$$

The norm of the perturbation To prove the desired result, an upper-bound on $\|\boldsymbol{E}\|$ is enough.

The matrix norm for square matrices is sub-additive and sub-multiplicative [68], therefore

$$\begin{aligned} \|\boldsymbol{E}\| &\leq \left\|\boldsymbol{T}_{\text{sig}}^* \boldsymbol{T}_{\text{noise}}\right\| + \|\boldsymbol{T}_{\text{noise}}^* \boldsymbol{T}_{\text{sig}}\| + \|\boldsymbol{T}_{\text{noise}}^* \boldsymbol{T}_{\text{noise}}\| \,, \\ &\leq \sum_p 2 \, \|\boldsymbol{T}_{\text{sig},p}\| \, \|\boldsymbol{T}_{\text{noise},p}\| + \left\|\boldsymbol{T}_{\text{noise},p}^*\right\| \|\boldsymbol{T}_{\text{noise},p}\| \,. \end{aligned}$$

Meckes established in [88] the divergence rate of the spectral norm of square Toeplitz matrices[1] with iid subgaussian entries in the generator. Applying this result, we obtain

$$\|\boldsymbol{T}_{\text{noise},p}\| \sim \mathcal{O}(\sqrt{M} \log M) \ ,$$

which implies for P fixed

$$\|\boldsymbol{E}\| \leq \mathcal{O}(M^{3/2} \log M). \tag{B.2}$$

Plugging equations (B.1) and (B.2) in Lemma C.1 concludes the proof.

[1] The main result in [88] studies real symmetric matrices, An extension to complex and non-symmetric matrices is confirmed in Section 3 of the same paper.

Appendix C

Detection for sparse common support channels

C.1 Proof of Proposition 3.2

Lemma C.1. *Define the hermitian symmetric Toeplitz matrix*

$$\bar{\boldsymbol{E}} = \frac{1}{\sqrt{2}}(\boldsymbol{E} + \boldsymbol{E}^*)$$

Then

$$\frac{1}{\sqrt{2}} \left\|\bar{\boldsymbol{E}}\right\| \leq \|\boldsymbol{E}\| \ , \qquad \mathbb{E}\left[\|\boldsymbol{E}\|\right] \leq \mathbb{E}\left[\left\|\bar{\boldsymbol{E}}\right\|\right],$$

and $\bar{\boldsymbol{E}}$ is a random symmetric Toeplitz matrix with iid entries (on non-matching diagonals) with distribution $\mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbb{I})$.

Proof.

The lowerbound on the spectral norm is a simple consequence of the triangle inequality

$$\left\|\bar{\boldsymbol{E}}\right\| \leq \sqrt{2}\,\|\boldsymbol{E}\|\,.$$

The symmetry and the distribution of the entries of $\bar{\boldsymbol{E}}$ holds from the basic properties of normal random variables.

Recall that

$$\boldsymbol{E} = \begin{bmatrix} \varepsilon_0 & \varepsilon_{-1} & \varepsilon_{-2} & \cdots & \cdots & \varepsilon_{-n+1} \\ \varepsilon_1 & \varepsilon_0 & \varepsilon_{-1} & \ddots & & \vdots \\ \varepsilon_2 & \varepsilon_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \varepsilon_{-1} & \varepsilon_{-2} \\ \vdots & & \ddots & \varepsilon_1 & \varepsilon_0 & \varepsilon_{-1} \\ \varepsilon_{n-1} & \cdots & \cdots & \varepsilon_2 & \varepsilon_1 & \varepsilon_0 \end{bmatrix}$$

For the upper bound on the spectral norm, we embed $\boldsymbol{E}$ in a circulant matrix $\boldsymbol{C}$ of size $2n \times 2n$, as done in [112] for symmetric Toeplitz matrices

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{E} & \boldsymbol{E'} \\ \boldsymbol{E'} & \boldsymbol{E} \end{bmatrix} \ , \ \gamma \in \mathbb{C} \, ,$$

where $$\boldsymbol{E'} = \begin{bmatrix} 0 & \varepsilon_{n-1} & \varepsilon_{n-2} & \cdots & \cdots & \varepsilon_1 \\ \varepsilon_{-n+1} & 0 & \varepsilon_{n-1} & \ddots & & \vdots \\ \varepsilon_{-n+2} & \varepsilon_{-n+1} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \varepsilon_{n-1} & \varepsilon_{n-2} \\ \vdots & & \ddots & \varepsilon_{-n+1} & 0 & \varepsilon_{n-1} \\ \varepsilon_{-1} & \cdots & \cdots & \varepsilon_{-n+2} & \varepsilon_{-n+1} & 0 \end{bmatrix}$$

The matrix $\boldsymbol{C}$ is diagonalized by the unitary DFT matrix $\boldsymbol{W}$

$$\boldsymbol{C} = \boldsymbol{W}^* \boldsymbol{D} \boldsymbol{W}.$$

With the orthogonal projection matrix

$$\boldsymbol{Q} = \begin{bmatrix} \mathbb{I}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \text{one obtains} \quad \boldsymbol{QCQ} \ = \ \begin{bmatrix} \boldsymbol{E} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

which spectrum is the spectrum of $\boldsymbol{E}$ with n additional null eigenvalues, therefore the two matrices have the same spectral norm. Since the spectral norm is invariant under unitary similarities

$$\|\boldsymbol{E}\| \ = \ \|\boldsymbol{PDP}\| \ , \quad \boldsymbol{P} \stackrel{\text{def}}{=} \boldsymbol{W}^* \boldsymbol{Q} \boldsymbol{W}.$$

The next step is to split the matrix $\boldsymbol{PDP}$ into a symmetric and an antisymmetric part. This is done by taking the real and imaginary parts of

$\boldsymbol{D}$. Then, using the triangle inequality on the spectral norm, one obtains

$$\|\boldsymbol{E}\|^2 \leq 2 \cdot \max\left(\|\boldsymbol{P}\Re\{\boldsymbol{D}\}\boldsymbol{P}\|^2 \ , \ \|\boldsymbol{P}\Im\{\boldsymbol{D}\}\boldsymbol{P}\|^2\right).$$

By construction, $\Re\{\boldsymbol{D}\}$ and $\Im\{\boldsymbol{D}\}$ are identically distributed.

Both $\boldsymbol{P}\Re\{\boldsymbol{D}\}\boldsymbol{P}$ and $\boldsymbol{P}\Im\{\boldsymbol{D}\}\boldsymbol{P}$ are symmetric random circulant matrices[a] with normally iid distributed entries with variance $1/2$. Therefore, the non-null spectra of $\boldsymbol{P}\Re\{\boldsymbol{D}\}\boldsymbol{P}$ and $\boldsymbol{P}\Im\{\boldsymbol{D}\}\boldsymbol{P}$ have a distribution identical to $\bar{\boldsymbol{E}}/\sqrt{2}$. Taking the expectation proves the result. □

[a]Note that taking the real and imaginary parts of the eigenvalues of a non symmetric Toeplitz matrix does not result in Toeplitz matrices!

Since the measure of the spectrum of random Toeplitz matrices concentrates as $n \to \infty$ [39] and $\frac{\|\boldsymbol{E}\|}{\sqrt{n \log n}} \sim \mathcal{O}(1)$, we may ask to which value $\frac{\|\boldsymbol{E}\|}{\sqrt{n \log n}}$ converges to.

Such a result is known for random symmetric Toeplitz matrices

Theorem C.1. ***(Sen et al. 2011 [112])***

$$\lim_{n\to\infty} \frac{\|\bar{\boldsymbol{E}}\|}{\sqrt{n \log n}} \ = \ 2 \cdot \digamma,$$

where $\digamma \approx 0.8288 \cdots$.

Proof.

The proof is found in [112] for real valued symmetric matrices, and it applies to hermitian symmetric matrices alike. □

Using Lemma C.1, we obtain the inequalities for the non-symmetric case.

C.2 Proof of Proposition 3.3

The criterion between the path amplitude estimates and $\rho_{\max}$ is

$$(2M+1) \cdot \sum_{p=1}^{P} |\widehat{c}_{k,p}|^2 \ \geq \ \frac{\sigma^2}{2} \rho_{\max} \qquad \text{w.p. } 1-\alpha.$$

Define

$$\xi_\ell \stackrel{\text{def}}{=} \frac{2}{\sigma^2(2M+1)} \sum_{p=1}^{P} \left| \sum_{m=-M}^{M} e^{j\pi \frac{\ell m}{2M+1}} E_p[m] \right|^2 ,$$

the set $\{\xi_\ell\}_{\ell=-M,\ldots,M}$ is a set of iid random variables with

$$\xi_\ell \sim \chi^2_{2P} \ , \quad \text{and} \quad \rho'_{\max}(P, 2M+1) \ = \ \max_{\ell=-M,\ldots,M} \xi_\ell \ .$$

The extremal distribution of the maximum in a set of iid χ^2_{2P} random follows a *Gumbel distribution* [63] as the cardinality of the set tends to infinity. For our finite set size $2M+1$, computing the cumulative distribution of $\rho'_{\max}$ is easy, thanks to the independence between the random variables. The probability to have the maximum less than some value t is simply the product of the probabilities to have each random variable less than t, and so

$$\Pr\left[\rho'_{\max} \leq t\right] \;=\; \left(\frac{\gamma(P, t/2)}{(P-1)!}\right)^{2M+1} \quad , \tag{C.1}$$

which is simply the cdf of a χ^2_{2P} random variable raised to the power $2M+1$. By setting $\Pr\left[\rho'_{\max} \leq t\right] \;=\; 1-\alpha$, we solve (C.1) for t to obtain

$$\gamma(P, t/2) \;=\; (P-1)! \cdot (1-\alpha)^{\frac{1}{2M+1}} .$$

To truely characterize which correlation can be due to the noise, it is necessary to link $\rho'_{\max}$ to $\rho_{\max}$. Invoking Parseval's theorem, an intuitive relation can be found in the time domain where multiplication with a phasor of a given frequency is the energy of the correlation with the Dirichlet kernel of corresponding bandwidth and shift. For a fixed value $\rho'_{\max}$, the largest difference with $\rho_{\max}$ is obtained for the optimal ω falling at equal distance between adjacent discretized values, and so

$$\rho'_{\max} \;\leq\; \rho_{\max} \;\leq\; \frac{\pi^2}{4}\rho'_{\max} \;,$$

which is illustrated in Figure C.1. With a uniform prior over the interval $\left[-\frac{\pi}{N},\; \frac{\pi}{N}\right[$ for $\rho'_{\max}$, numerical integration yields

$$\mathbb{E}\left[\rho'_{\max}\right] \approx 0.7737 \cdot \mathbb{E}\left[\rho_{\max}\right].$$

This value has a negative bias since the maximum is more likely to correspond to a shift for which $\rho'_{\max}$ is close to $\rho_{\max}$. Therefore, we introduce a scaling factor $0.77 \;<\; c_0 \;(\approx\; 0.9) \;\leq\; 1$. Putting the elements together yields the result.

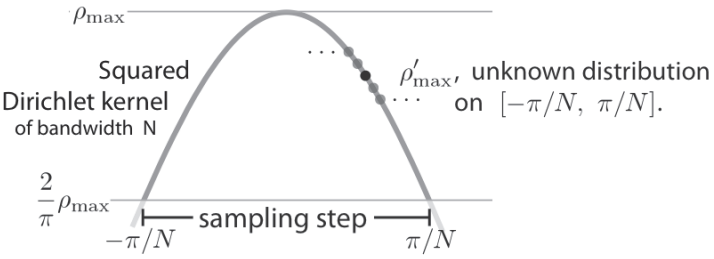


Figure C.1: *The random variable $\rho_{\max}$ is by definition greater than $\rho'_{\max}$. The distribution of the shift between the value ω yielding the maximal correlation energy $\rho_{\max}$ and the discretized value $\rho'_{\max}$ is unknown. The uniform distribution is not adequate since maximization of $\rho'_{\max}$ favors a shift closer to the middle of the interval.*

Appendix D

Tracking of SCS channels

D.1 Tracking of a single path

For $K = 1$, the parameter space is representable as the complex plane.

Proposition D.1. *Let assume that only a single channel is present, and the DFT coefficients of the signal are*

$$\boldsymbol{y} = [1,\ e^{j\omega_1},\ \ldots,\ e^{j\omega_1(N-1)}]^T.$$

In these circumstances the optimization in (4.5) is equivalent to

$$\max_{r,\omega} \boldsymbol{r}^T \boldsymbol{T}_\omega \boldsymbol{r} / \|\boldsymbol{r}\|^2 \ , \tag{D.1}$$

where $\boldsymbol{r} = [1,\ r,\ \ldots, r^{N-1}]^T$ *and* $[\boldsymbol{T}_\omega]_{m,n} = \cos[(\omega_1 - \omega)(m-n)]$.

Proof.

Let $\boldsymbol{z} \stackrel{\text{def}}{=} [1,\, re^{j\omega},\, \ldots,\, r^{N-1}e^{j\omega(N-1)}]$, where $re^{j\omega}$ is the root of the generator of the Toeplitz matrix $\boldsymbol{A}$, then

$$\begin{aligned}\left\|\text{proj}_{\text{range}(\boldsymbol{A}^*)}\boldsymbol{y}\right\|^2 &= \left\|\boldsymbol{y}\cdot\text{proj}_{\ker(\boldsymbol{A})}\boldsymbol{y}\right\|^2 ,\\ &= (\boldsymbol{y}-\boldsymbol{z}^*\boldsymbol{y}\cdot\boldsymbol{z}/\left\|\boldsymbol{z}\right\|^2)^*(\boldsymbol{y}-\boldsymbol{z}^*\boldsymbol{y}\cdot\boldsymbol{z}/\left\|\boldsymbol{z}\right\|^2),\\ &= \left\|\boldsymbol{y}\right\|^2-\frac{1}{\left\|\boldsymbol{z}\right\|^2}\left|\boldsymbol{y}^*\boldsymbol{z}\right|^2 .\end{aligned}$$

One obtains $\left|\boldsymbol{y}^*\boldsymbol{z}\right|^2 = \sum_{m,n} r^{m+n}e^{j(\omega_1-\omega)(m-n)}$. The imaginary part of the double summation cancels out by antisymmetry, and writing the result in a vectorial form yields

$$\left|\boldsymbol{y}^*\boldsymbol{z}\right|^2 = \boldsymbol{r}^T\boldsymbol{T}_\omega\boldsymbol{r}.$$

To conclude the proof, $\left\|\boldsymbol{z}\right\|^2 = \left\|\boldsymbol{r}\right\|^2$. □

The objective function in (D.1) is unfortunately not polynomial in the unknown r, and finding its critical points is not an easy task. A necessary condition for a point to be critical is that

Proposition D.2. *The derivative with respect to r of the objective function in (D.1) is 0 iff*

$$\frac{1}{\left\|\boldsymbol{r}\right\|^4}\cdot\boldsymbol{r}^T\boldsymbol{T}_\omega\left[\boldsymbol{D}\,,\,\boldsymbol{r}\boldsymbol{r}^T\right]\boldsymbol{r} = 0\,, \tag{D.2}$$

where $\boldsymbol{D} \stackrel{\text{def}}{=} \begin{bmatrix} 0 & & & & \mathbb{O} \\ 1 & 0 & & & \\ & 2 & \ddots & & \\ & & \ddots & \ddots & \\ \mathbb{O} & & & N-1 & 0 \end{bmatrix}$,

and $\left[\boldsymbol{D}\,,\,\boldsymbol{r}\boldsymbol{r}^T\right] \stackrel{\text{def}}{=} \boldsymbol{D}\boldsymbol{r}\boldsymbol{r}^T-\boldsymbol{r}\boldsymbol{r}^T\boldsymbol{D}$ *is the commutator of* $\boldsymbol{D}$ *and* $\boldsymbol{r}\boldsymbol{r}^T$.

Proof.

This is a simple development of the derivative with respect to r of (D.1). □

One can verify that for three values of r the condition (D.2) is verified

- $\boldsymbol{D}$ and $\boldsymbol{r}\boldsymbol{r}^T$ commute only for $r \to \infty$.

- For $r = 0$, $\left[\boldsymbol{D}\,,\,\boldsymbol{r}\boldsymbol{r}^T\right]\boldsymbol{r}$ vanishes.
- For $r = 1$, the vector $\boldsymbol{r}$ is the all-one vector, and $\left[\boldsymbol{D}\,,\,\boldsymbol{r}\boldsymbol{r}^T\right]\boldsymbol{r}$ is an antisymmetric vector.

  The matrix $\boldsymbol{T}_\omega$ is a rank-2 Toeplitz matrix with an antisymmetric and a symmetric eigenvectors since for $\boldsymbol{n} = [0,\ \ldots,\ N-1]^T$ it is written as

$$\begin{aligned}\boldsymbol{T}_\omega &= \Re\left[e^{j(\omega_1-\omega)\frac{N}{2}}e^{j(\omega_1-\omega)(\boldsymbol{n}-\frac{N}{2})}\left(e^{j(\omega_1-\omega)\frac{N}{2}}e^{j(\omega_1-\omega)(\boldsymbol{n}-\frac{N}{2})}\right)^*\right],\\ &= \cos\left[(\omega_1-\omega)\left(\boldsymbol{n}-\tfrac{N}{2}\right)\right]\cos\left[(\omega_1-\omega)\left(\boldsymbol{n}-\tfrac{N}{2}\right)\right]^*\\ &\quad+\sin\left[(\omega_1-\omega)\left(\boldsymbol{n}-\tfrac{N}{2}\right)\right]\sin\left[(\omega_1-\omega)\left(\boldsymbol{n}-\tfrac{N}{2}\right)\right]^*,\end{aligned}$$

  such that $\cos\left[(\omega_1-\omega)\left(\boldsymbol{n}-\frac{N}{2}\right)\right] \perp \sin\left[(\omega_1-\omega)\left(\boldsymbol{n}-\frac{N}{2}\right)\right]$.

  Note that these two vectors may not have the same norm, and so the non zero eigenvalues of $\boldsymbol{T}_\omega$ are not equal in general. We conclude that for $r = 1$, (D.2) holds since $\boldsymbol{T}_\omega$ is multiplied by a symmetric and an antisymmetric vector respectively from the left and the right.

Numerical experiments seem to indicate these are the only solutions of (D.2), though we cannot prove it.

Assuming this supposition is true, then for $r = 1$ the point $(1, \omega)$ can only be a local maximum of (D.1) if $\left(\sum_{n=0}^{N-1}\cos((\omega_1-\omega)(n-\frac{N}{2})\right)^2 \geq N$. If this condition is not met, then the objective in the maximization problem (D.1) at this point is strictly lower than the value at $r = 0$ and $r \to \infty$ and any critical point can either be a local minimum or a saddle.

D.2 Addition of a path

Let $\boldsymbol{e} \overset{\text{def}}{=} f(\widehat{\boldsymbol{a}}_{\mathsf{opt}})$.

The statistics of $\boldsymbol{e}$ depends on the statistics of the estimate $\widehat{\boldsymbol{a}}_{\mathsf{opt}}$, which is itself dependent of the algorithm used to solve (4.4).

To make the modelization of the residual $\boldsymbol{e}$ tractable, we need to make a few assumptions

---

Definition D.1. ***(Noise/Data independence)***

- *The subspace spanned by the annihilating filter is independent of the noise in the data.*
- *The energy of the residual is less than the energy of the residual obtained with perfect estimation of the tracked paths.*

The second assumption is likely to be fulfilled when the paths are accurately tracked. If it were not the case, a perfect estimation of the roots would yield a lower residual energy, and it would likely be the local minimum the iterative estimation converges to.

Hence, the residual energy observed after the estimation is completed should be inferior to the residual energy obtained after projection the noise only samples on the subspace spanned by $\widehat{\boldsymbol{A}} \stackrel{\text{def}}{=} \text{Toeplitz}(\widehat{\boldsymbol{a}})$.

The first assumption implies that the noise is independent of $\widehat{\boldsymbol{A}}$, and so, the projection of the noise in the span of $\widehat{\boldsymbol{A}}$ is the projection of a vector of 0-mean gaussian random vector in a deterministic K-dimensional subspace. The linearity of the projection operator implies the result is also a 0-mean, gaussian random vector. Namely, for a (complex-valued) AWGN of power σ^2

$$\boldsymbol{e}_{\text{noise}} \sim \mathcal{N}_{\mathbb{C}}\left(\boldsymbol{0}, \sigma^2 \mathbb{I}_P \oplus (\widehat{\boldsymbol{A}}^{\dagger}\widehat{\boldsymbol{A}})\right) \ , \tag{D.3}$$

where $\oplus$ is the *Kronecker sum*.

To decide on the absence of additional paths — under the Noise/Data independence assumptions) — one could test if the residual follows the probability law (D.3) using Pearson's χ^2 test for example.

A weaker criterion is to compare the residual energy to the expected energy of $\boldsymbol{e}_{\text{noise}}$. We know from the second assumption in Definition D.1 that the expected energy of $\boldsymbol{e}_{\text{noise}}$ should dominate the residual, which immediately provides a statistical hypothesis. Using the elementary properties of the trace,

$$\mathbb{E}\left[\|\boldsymbol{e}_{\text{noise}}\|_2^2\right] \ = \ \sigma^2 P \cdot (N-K).$$

More specifically, the energy of $\boldsymbol{e}_{\text{noise}}$ is χ^2 distributed

$$\frac{2}{\sigma^2}\|\boldsymbol{e}_{\text{noise}}\|_2^2 \sim \chi^2_{2P\cdot(N-K)} \ .$$

Knowing the distribution of the residual energy, the hypothesis of having no additional signal path in the residual can be rejected if the probability of a residual greater than or equal to the observed value falls below some threshold.

Appendix E

Time-Frequency localization in periodic domains

E.1 Proof of Lemmas 5.1 and 5.2

E.1.1 Proof of Lemma 5.1

Let x be maximally compact for a given time-spread $\Delta_n^2(x)$. Then,

$$\Delta_n^2(x) \;=\; \Delta_n^2(|x|)\,,$$

and

$$\begin{aligned}\Delta_{\omega_p}^2(|x|) \;&=\; \left|\sum_{n\in\mathbb{Z}} |x_n||x_{n+1}|\right|^{-2} - 1,\\ &\leq \left|\sum_{n\in\mathbb{Z}} x_n x_{n+1}^*\right|^{-2} - 1 \quad = \Delta_{\omega_p}^2(x). \end{aligned} \tag{E.1}$$

For maximally compact sequences, if Δ_n^2 strictly monotonically varies in function of $\Delta_{\omega_p}^2$, then fixing Δ_n^2 or $\Delta_{\omega_p}^2$ is equivalent and proves the lemma. In the following lemma we show that for maximally compact sequences Δ_n^2 changes monotonically with $\Delta_{\omega_p}^2$.

Lemma E.1.
For maximally compact sequences, Δ_n^2 is a decreasing function of $\Delta_{\omega_p}^2$.

Proof.

For proving this lemma, we use the dual formulation in (5.17). The feasible

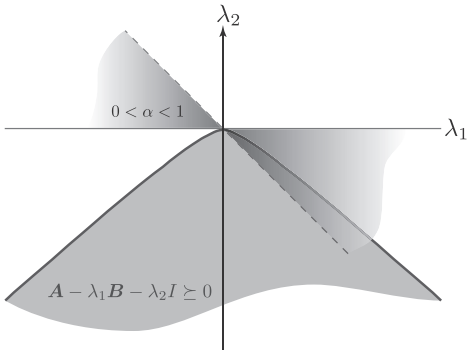


Figure E.1: ***The feasible set of the dual problem*** (5.17) ***and the range of supporting line angles for finding its maximum***. *As α increases, we need to elevate the line more to support the feasible set, which means that the optimal value of Δ_n^2 increases.*

region of the dual problem is shown in Figure E.1. We can write the dual as

$$\begin{aligned} \underset{\lambda_1,\lambda_2}{\text{maximize}} \quad & c \\ \text{subject to} \quad & \lambda_2 = c - \alpha\,\lambda_1\,, \\ & \boldsymbol{A} - \lambda_1\,\boldsymbol{B} - \lambda_2\,\boldsymbol{I} \succeq 0 \end{aligned} \tag{E.2}$$

Note that α changes between 0 and 1 (shown in Figure E.1 with the gradient region). For a fixed α, the maximum c^{opt} is found by elevating the corresponding line $\lambda_2 = c - \alpha\lambda_1$ until it supports the feasible set (it is tangent to it). Since the feasible set is convex, as α grows (which means $\Delta_{\omega_p}^2$ decreases), we need a higher elevation of the line to support the convex set, thus c^{opt} (equivalently Δ_n^2) increases. □

E.1.2 Proof of Lemma 5.2

Consider the shift operator in (5.14). Since the shift operation does not change the norm of a sequence, we will assume a unit norm sequence without loss of generality. We can show that

$$\begin{aligned} \mu_n(x_{n-\nu}) &= \frac{\left\langle j\frac{d}{d\omega}\left(e^{-j\omega\nu}X(e^{j\omega})\right), e^{-j\omega\nu}X(e^{j\omega})\right\rangle}{2\pi} \\ &= \frac{\left\langle j\frac{d}{d\omega}X(e^{j\omega}), X(e^{j\omega})\right\rangle}{2\pi} + \frac{\nu\left\langle X(e^{j\omega}), X(e^{j\omega})\right\rangle}{2\pi} \\ &= \mu_n(x) + \nu, \end{aligned} \tag{E.3}$$

where we used the DTFT domain definition of the time center [100]:

$$\mu_n(x) = \frac{\left\langle j\frac{d}{d\omega}X(e^{j\omega}), X(e^{j\omega})\right\rangle}{\|X\|^2}.$$

The proof for Lemma 5.2—trivial for $\nu \in \mathbb{Z}$—is not obvious for arbitrary shifts. Let x be maximally compact with time center $\mu_n(x)$, then according to (E.3), $x_{n-\nu}$ is centered at $\mu_n(x)+\nu$ and

$$\begin{aligned}
2\pi\Delta_n^2(x_{n-\nu}) &= 2\pi\left[\sum_{n\in\mathbb{Z}} n^2(x_{n-\nu})^2 - |\mu_n(x_{n-\nu})|^2\right] \\
&= \left\langle j\frac{d}{d\omega}e^{-j\omega\nu}X(e^{j\omega}), j\frac{d}{d\omega}e^{-j\omega\nu}X(e^{j\omega})\right\rangle - 2\pi\,|\mu_n(x)+\nu|^2 \\
&= \left\langle -j\nu X(e^{j\omega}) + X'(e^{j\omega}), -j\nu X(e^{j\omega}) + X'(e^{j\omega})\right\rangle - 2\pi\,|\mu_n(x)+\nu|^2 \\
&= 2\pi\,|\nu|^2\left\langle X(e^{j\omega}), X(e^{j\omega})\right\rangle + \left\langle X'(e^{j\omega}), X'(e^{j\omega})\right\rangle \\
&\qquad + \left\langle -j\nu X(e^{j\omega}), X'(e^{j\omega})\right\rangle + \left\langle X'(e^{j\omega}), -j\nu X(e^{j\omega})\right\rangle - 2\pi\,|\mu_n(x)+\nu|^2 \\
&= 2\pi\,|\nu|^2 + \left\langle X'(e^{j\omega}), X'(e^{j\omega})\right\rangle + 2\pi\ \mathrm{Real}[\mu_n(x)\nu^*] - 2\pi\,|\mu_n(x)+\nu|^2 \\
&= \left\langle X'(e^{j\omega}), X'(e^{j\omega})\right\rangle - 2\pi\,|\mu_n(x)|^2 \\
&= 2\pi\Delta_n^2(x)\,. \qquad \text{(E.4)}
\end{aligned}$$

This shows that time shift does not affect the time spread of a sequence. Thus, if x is a maximally compact sequence, then $x_{n-\mu_n(x)}$ is also maximally compact (note that time shift does not change the frequency characteristics of the sequence). □

E.2 Proof of Theorem 5.2

By using Lemmas 5.1 and 5.2, we can write problem (5.13) as

$$\begin{aligned}
\Delta_{n,\mathrm{opt}}^2 = \underset{x_n}{\text{minimize}} \quad & \sum_{n\in\mathbb{Z}} n^2x_n^2 \\
\text{subject to} \quad & \sum_{n\in\mathbb{Z}} x_nx_{n+1} = \frac{1}{\sqrt{1+\sigma^2}}, \qquad \text{(E.5)} \\
& \sum_{n\in\mathbb{Z}} x_n^2 = 1.
\end{aligned}$$

We can rewrite (E.5) in a matrix form as a quadratically constrained quadratic program (QCQP) [34]

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \\
\text{subject to} \quad & \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \alpha\,, \qquad \text{(E.6)} \\
& \boldsymbol{x}^T\boldsymbol{x} = 1\,,
\end{aligned}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined in Theorem 5.2 and $\alpha = 1/\sqrt{1+\sigma^2}$. This problem can be further reformulated as follows:

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & \mathrm{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T) \\
\text{subject to} \quad & \mathrm{tr}(\boldsymbol{B}\boldsymbol{x}\boldsymbol{x}^T) = \alpha \\
& \mathrm{tr}(\boldsymbol{x}\boldsymbol{x}^T) = 1\,.
\end{aligned}$$

Replacing $\boldsymbol{x}\boldsymbol{x}^T$ by $\boldsymbol{X}$, we can write equivalently

$$\begin{aligned} \underset{\boldsymbol{X}}{\text{minimize}} \quad & \operatorname{tr}(\boldsymbol{A}\boldsymbol{X}) \\ \text{subject to} \quad & \operatorname{tr}(\boldsymbol{B}\boldsymbol{X}) = \alpha \\ & \operatorname{tr}(\boldsymbol{X}) = 1 \\ & \boldsymbol{X} \succeq 0,\ \operatorname{rank}(\boldsymbol{X}) = 1\,. \end{aligned} \tag{E.7}$$

We further relax the above formulation to reach the semi-definite program

$$\begin{aligned} \underset{\boldsymbol{X}}{\text{minimize}} \quad & \operatorname{tr}(\boldsymbol{A}\boldsymbol{X}) \\ \text{subject to} \quad & \operatorname{tr}(\boldsymbol{B}\boldsymbol{X}) = \alpha \\ & \operatorname{tr}(\boldsymbol{X}) = 1, \quad \boldsymbol{X} \succeq 0\,. \end{aligned} \tag{E.8}$$

In Lemma E.2 we show that the semi-definite relaxation is tight. □

Lemma E.2. *The semi-definite relaxation (SDR) in* (5.16) *is tight.*

Proof.

Shapiro and then Barvinok and Pataki [115; 29; 98; 79] show that if the SDP in (E.7) is feasible, then

$$\operatorname{rank}(\boldsymbol{X}^{\text{opt}}) \leq \lfloor (\sqrt{8m+1} - 1)/2 \rfloor\,, \tag{E.9}$$

where m is the number of (trace-product) constraints of the SDP, and $\boldsymbol{X}^{\text{opt}}$ is its optimal solution. For our semi-definite program in (5.16), $m = 2$. Thus, (E.9) implies that the solution has rank 1. Using this fact, one can see that the semi-definite relaxation is in fact tight. Note that from the nature of the problem, (E.7) is clearly feasible; we can always find a periodic signal in the Fourier domain with a unit norm and a desired frequency spread, although not having an optimal time spread. □

E.3 Proof of Lemma 5.3

We use the following Lemma for the proof:

Lemma E.3. *[125; 128] For a semi-definite program and its dual: If the primal is feasible and the dual is strictly feasible, then strong duality holds.*

As it was already mentioned in the proof of Lemma E.2, the primal is feasible. For the dual, one can use the Gershgorin's circle theorem and show that a sufficient condition

for $\boldsymbol{A} - \lambda_1 \boldsymbol{B} - \lambda_2 \boldsymbol{I} \succ 0$ to hold is $\lambda_2 < -\lambda_1$ and $\lambda_1 > 0$. Thus, the dual problem is strictly feasible. □

E.4 Proof of Theorem 5.3

If a sequence is a solution to the dual SDP problem (5.17), the dual constraint is active. Therefore, maximally compact sequences lie on the boundary of the quadratic cone

$$\boldsymbol{A} - \lambda_1 \boldsymbol{B} - \lambda_2 \boldsymbol{I} \succeq 0.$$

A maximally compact sequence $\boldsymbol{x}$ is thus solution of the eigenvalue problem

$$(\boldsymbol{A} - \lambda_1 \boldsymbol{B})\boldsymbol{x} \;=\; \lambda_2 \boldsymbol{x}, \tag{E.10}$$

where λ_1 and λ_2 are the dual variables of the SDP problem, and λ_2 is also the minimal eigenvalue of $\boldsymbol{A} - \lambda_1 \boldsymbol{B}$ and x is the associated eigenvector (this can be also seen by forcing the derivative of the Lagrangian in (E.6) to zero).

This explicit link between the dual variables and the sequence, yields a differential equation for which the DTFT spectrum of maximally compact sequences is the solution. In the DTFT domain (E.10) becomes

$$\begin{aligned} & -X''(e^{j\omega}) - \lambda_1 \cos(\omega) X(e^{j\omega}) \;=\; \lambda_2 X(e^{j\omega}), \\ \Leftrightarrow \qquad & X''(e^{j\omega}) + (\lambda_2 + \lambda_1 \cos(\omega))\, X(e^{j\omega}) \;=\; 0\,, \end{aligned} \tag{E.11}$$

which is *Mathieu's differential equation* ([2]§ 20.1.1). The solutions of Mathieu's equation are called Mathieu functions, and they assume an odd and even form

$$\text{Mathieu's Cosine (even)} \quad \mathrm{ce}(a, q; \omega)\,, \tag{E.12}$$

$$\text{Mathieu's Sine (odd)} \quad \mathrm{se}(a, q; \omega)\,. \tag{E.13}$$

Taking into account the periodicity of (E.11), it appears not all pairs of parameters (a, q) will lead to a periodic solution. Mathieu functions can be restricted to be 2π periodic:

Definition E.1. *The solutions of Mathieu's harmonic differential equation—equation* (E.11) *with X 2π-periodic—are defined as*

$$\textit{Mathieu's harmonic Cosine (even)} \quad \mathrm{ce}_m(q\,;\,\omega) \;=\; \mathrm{ce}(a_m(q), q\,;\,\omega)\;,\; m \in \mathbb{N}. \tag{E.14}$$

$$\textit{Mathieu's harmonic Sine (odd)} \quad \mathrm{se}_m(q\,;\,\omega) \;=\; \mathrm{ce}(b_m(q), q\,;\,\omega)\;,\; m \in \mathbb{N}^+. \tag{E.15}$$

It is immediately visible that the spectrum of maximally compact sequences may only have the form

$$\begin{aligned} X(e^{j\omega}) \;&= \gamma_0 \cdot \mathrm{ce}_m(-2\lambda_1\,;\,\omega/2) + \gamma_1 \cdot \mathrm{se}_m(-2\lambda_1\,;\,\omega/2),\ \text{for } m \in \mathbb{N}^+, \\ X(e^{j\omega}) \;&= \gamma_0 \cdot \mathrm{ce}_m(-2\lambda_1\,;\,\omega/2),\ \text{for } m = 0\,, \end{aligned} \tag{E.16}$$

for any constants γ_0 and γ_1 such that $\|X(e^{j\omega})\| = 1$. More specifically, for any $\lambda_1 \geq 0$, the dual SDP problem can be posed and any solution would have the form (E.16).

Characteristic numbers of Mathieu's equation are ordered [53], such that

$$\begin{aligned} &a_0(-2\lambda_1) < a_1(-2\lambda_1) < b_1(-2\lambda_1) < b_2(-2\lambda_1) < a_2(-2\lambda_1) < \cdots \quad &,\lambda_1 > 0\,, \\ &a_0(-2\lambda_1) < b_1(-2\lambda_1) \leq a_1(-2\lambda_1) < b_2(-2\lambda_1) \leq a_2(-2\lambda_1) < \cdots \quad &,\lambda_1 \leq 0\,, \end{aligned}$$

Since $a_m(-2\lambda_1) = 4\lambda_2$ and λ_2 is the minimal eigenvalue, we conclude that $m = 0$.

Therefore we have found the real-valued maximally compact sequence up to scaling and shift. For modulation, one should simply notice that for a maximally compact sequence the following property must hold

$$x_n x_{n+1}^* \;=\; |x_n||x_{n+1}|e^{j\varphi},\ \varphi \in [0,\ 2\pi)\,,\ \forall n \in \mathbb{Z},$$

Since ce_0 is strictly positive, we assume $|x_n| > 0$ for all n. Then, it only allows to obtain complex-valued maximally compact sequences from a real-valued one as a modulation or multiplication by a complex scalar.

□

Appendix F

From uncertainty to error

F.1 Proof of Lemma 6.1

Since μ_L is constant and M kills constants ($\mu_M = 0$)

$$\langle ML p_\theta \, , \, p_\theta \rangle \; = \; M \underbrace{\langle L p_\theta \, , \, p_\theta \rangle}_{\mu_L} \; = \; 0,$$

and

$$\begin{aligned} \langle LM p_\theta \, , \, p_\theta \rangle \; &= \langle M p_\theta \, , \, L^* p_\theta \rangle^* \; = \; \langle L p_\theta \, , \, M^* p_\theta \rangle \; , \\ &= \langle (L - \mu_L) p_\theta \, , \, (M - \mu_M)^* p_\theta \rangle + \mu_L \mu_M - \mu_L \mu_M - \underbrace{\mu_L \mu_M}_{0} . \end{aligned}$$

where the first equality is obtained because L acts as a multiplier. Hence

$$\left| \langle [L \, , \, M] \, p_\theta \, , \, p_\theta \rangle \right|^2 \; = \; \left| \langle (L - \mu_L) p_\theta \, , \, (M - \mu_M)^* p_\theta \rangle \right|^2 ,$$

and one can use Cauchy-Schwarz inequality (with the adjoint of $M - \mu_M$)

$$\left\| (L - \mu_L) p_\theta \right\|^2 \; \geq \; \frac{\left| \langle (L - \mu_L) p_\theta \, , \, (M - \mu_M)^* p_\theta \rangle \right|^2}{\left\| (M - \mu_M)^* p_\theta \right\|^2} .$$

Since $\mu_M \; = \; 0$, $\left\| (M - \mu_M)^* p_\theta \right\|^2 \; = \; \left\| M p_\theta \right\|^2$, which proves the lemma.

F.2 Proof of Lemma 6.2

We cannot use Lemma 6.1 since the best possible operator found after optimization may not have $\mu_M = 0$. Therefore we go all the way back to Cauchy-Schwarz inequality

$$\mathsf{var}_{\widehat{\theta}} \; \geq \; \frac{\left| \langle (L - \mu_L) p_\theta \, , \, M p_\theta \rangle \right|^2}{\left\| M p_\theta \right\|^2}$$

With $Mf(\theta) \; = \; (g * f)(\theta)$ and $\mu_M \; = \; \int g(t) dt$, the numerator yields

$$\begin{aligned}\langle (L-\mu_L)p_\theta\ ,\ Mp_\theta\rangle \ &= \int_{\mathcal{X}} (1-e^{j(\widehat{\theta}_{\boldsymbol{x}}-\theta)}-\mu_L(\theta))\left(\int_\Theta g^*(t)p_{X|\theta-t}(\boldsymbol{x})dt\right)\boldsymbol{dx},\\ &= \underbrace{\int_\Theta g^*(t)\int_{\mathcal{X}} p_{X|\theta-t}(\boldsymbol{x})\boldsymbol{dx}dt}_{\mu_M^*} - \underbrace{\mu_L(\theta)\mu_M^*}_{(1-\tau)\mu_M^*},\\ &\qquad -\int_\Theta e^{-jt}g^*(t)\underbrace{\int_{\mathcal{X}} e^{j(\widehat{\theta}_{\boldsymbol{x}}-(\theta-t))}p_{X|\theta-t}(\boldsymbol{x})\boldsymbol{dx}}_{\tau(\theta-t)=\tau}dt\\ &= \tau\int_\Theta (1-e^{-jt})g^*(t)dt.\end{aligned}$$

so that

$$|\langle (L-\mu_L)p_\theta\ ,\ Mp_\theta\rangle|^2 \ = \ |\tau|^2\cdot\left|\int_\Theta (1-e^{jt})g(t)dt\right|^2.$$

The denominator expression follows from the definition of the inner-product.

F.3 Proof of Theorem 6.2

To prove this theorem, we use a Cauchy-Schwarz inequality for random vectors [127] which states for any two random vectors U and V of compatible dimensions and finite expected energy

$$\mathbb{E}\left[UU^*\right] \ \succeq\ \mathbb{E}\left[UV^*\right]\mathbb{E}\left[VV^*\right]^\dagger\mathbb{E}\left[VU^*\right]. \tag{F.1}$$

$$\text{By setting}\ \ U_k \ = \ \mathsf{loc}(\widehat{\theta}_{X,k},\theta_k)-\mu_k\ ,\ \text{and}\qquad V_k \ = \ p_{\boldsymbol{\theta}}^{-1}\frac{\partial}{\partial\theta_k}p_{\boldsymbol{\theta}}$$

we obtain $\mathbb{E}\left[UU^*\right] \ = \ \mathsf{cov}\left\{\mathsf{loc}(\widehat{\boldsymbol{\theta}}_X,\boldsymbol{\theta})\right\}$, $\mathbb{E}\left[VV^*\right] \ = \ \boldsymbol{J}_{\boldsymbol{\theta}}$ and

$$\begin{aligned}\mathbb{E}\left[U_kV_\ell^*\right] \ &= \int_{\mathcal{X}}\mathsf{loc}(\widehat{\theta}_{\boldsymbol{x},k},\theta_k)\frac{\partial}{\partial\theta_\ell}p_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{dx}-\mu_k\frac{\partial}{\partial\theta_\ell}\underbrace{\int_{\mathcal{X}}p_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{dx}}_{1},\\ &= \underbrace{\frac{\partial}{\partial\theta_\ell}\mu_k}_{0}-\int_{\mathcal{X}}\frac{\partial}{\partial\theta_\ell}\left[\mathsf{loc}(\widehat{\theta}_{\boldsymbol{x},k},\theta_k)\right]p_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{dx},\\ &= \delta_{k-l}\times\begin{cases}\frac{j}{\tau_{\widehat{\theta}_k}}\mathbb{E}\left[e^{j(\widehat{\theta}_k-\theta_k)}\right] \ = \ j, & \theta_k\ \text{is periodic}\\ 1, & \theta_k\ \text{is linear.}\end{cases}\end{aligned}$$

Hence $\mathbb{E}\left[UV^*\right]$ is unitary and diagonal, so we conclude from F.1 that

$$\mathsf{cov}\left\{\mathsf{loc}(\widehat{\boldsymbol{\theta}}_X,\boldsymbol{\theta})\right\} \ \succeq\ \boldsymbol{J}_{\boldsymbol{\theta}}^{-1}.$$

F.4 Proof of Lemma

The denominator from the right-hand side of the inequality (6.21) in Lemma 6.2 is for additive gaussian noise with covariance matrix $\mathbf{\Sigma}$

$$\begin{aligned}
&\mathbb{E}\left[\frac{|(g(t) * p_t)(\theta)|^2}{p_\theta}\right] \\
&= \iint g(t)g(t') \int_{\mathcal{X}} \frac{p_{\theta-t}(\boldsymbol{x})p_{\theta-t'}(\boldsymbol{x})}{p_\theta(\boldsymbol{x})} \boldsymbol{dx}dt'dt \\
&= \iint g(t)g(t') \cdot \mathrm{cst} \cdot \\
&\quad \int_{\mathcal{X}} e^{-(\|\boldsymbol{x}-\boldsymbol{s}_{\theta-t}\|^2_{\mathbf{\Sigma}^{-1}} + \|\boldsymbol{x}-\boldsymbol{s}_{\theta-t'}\|^2_{\mathbf{\Sigma}^{-1}} - \|\boldsymbol{x}-\boldsymbol{s}_{\theta}\|^2_{\mathbf{\Sigma}^{-1}})/2} \boldsymbol{dx}dtdt', \\
&= \iint g(t)g(t')\kappa\left(\theta-t,\theta\right)^{-1}\kappa\left(\theta,\theta-t'\right)^{-1}\kappa\left(\theta-t,\theta-t'\right)dtdt',
\end{aligned}$$

where cst is the normalization constant of the multivariate gaussian pdf. The last equality is obtained by a development of the norms, such that the dependence on $\boldsymbol{x}$ is factored out in a term yielding the gaussian pdf, which when integrated evaluates to cst^{-1}.

F.5 Lowerbounds with modulo-2π localization

As an example of an estimation problem, consider the *Phase Locking Problem* (PLP): a known periodic waveform is sampled at known times in the presence of noise, and from this samples the time offset of the signal is estimated. Figure F.1 sets up an instance of this problem: a periodic pulse is uniformly sampled over one period. The samples are independently corrupted by a white gaussian noise. This simple problem is very well understood [139; 104] — see [104] for the Maximum Likelihood (ML) estimator, [149; 91] for MSE lowerbounds.

F.5.1 Problem setup

As shown in Figure F.1, a time offset θ normalised on the interval $\Omega = [-1/2,\ 1/2[$, is estimated from a vector of N samples X

$$X[n] \ = \ \boldsymbol{s}_\theta[n] + E[n],\ n = 0, \ldots, N-1 \tag{F.2}$$

where $s_\theta[n] \stackrel{\text{def}}{=} s(n/N - \theta)$ are uniform samples of a periodic waveform shifted by θ, and E is a vector of iid gaussian random variables with variance N_0. Thus, for a fixed value of θ, X follows a multivariate gaussian distribution of density

$$p_X(\boldsymbol{x}|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\|\boldsymbol{x}-\boldsymbol{s}_\theta\|^2/(2N_0)}.$$

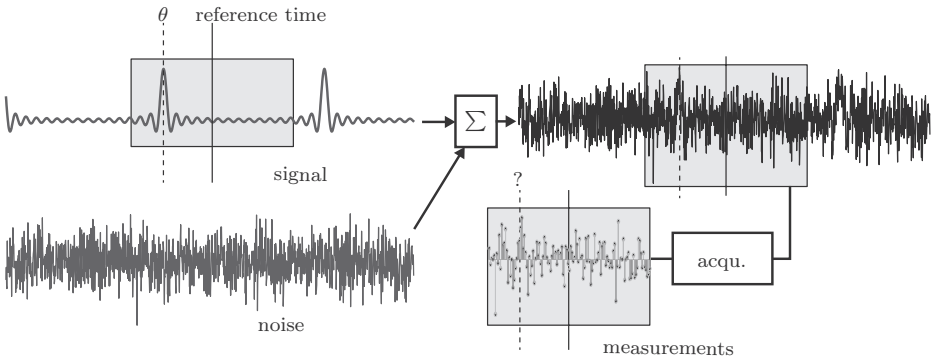


Figure F.1: *The phase locking problem (PLP) is the estimation of θ with respect to an arbitrary reference time fixed by the acquisition device (acqu.) which performs a uniform sampling. The signal waveform and its period are known. The estimation relies on noisy samples (measurements) acquired over one (or several) periods.*

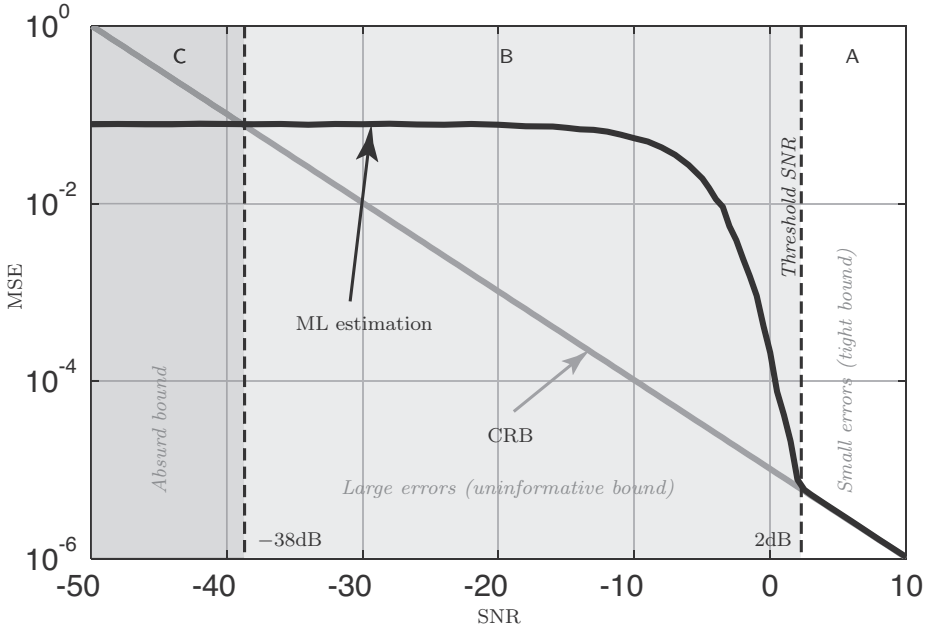


Figure F.2: *The simple phase locking problem already raises two issues about the Cramér-Rao bound (CRB): looseness and unboundedness.*

The waveform s is periodic on Ω. Therefore, the addition on the parameter space

shall be periodical

$$\theta \oplus \vartheta \stackrel{\text{def}}{=} \operatorname{mod}_\Omega \theta + \vartheta,$$
$$\theta \ominus \vartheta \stackrel{\text{def}}{=} \operatorname{mod}_\Omega \theta - \vartheta.$$

This periodic group structure induces periodical statistics for an estimator $\widehat{\theta_1}$ of θ

$$\begin{aligned} \mathsf{bias}_{\widehat{\theta_1}}(\theta) &\stackrel{\text{def}}{=} \mathbb{E}_{X|\theta}\left[\widehat{\theta_1} \ominus \theta\right] &&= \int_{\mathcal{X}} (\widehat{\theta_1} \ominus \theta) p_X(\boldsymbol{x}|\theta) \boldsymbol{dx}, \\ \mathsf{MSE}_{\widehat{\theta_1}}(\theta) &\stackrel{\text{def}}{=} \mathbb{E}_{X|\theta}\left[(\widehat{\theta_1} \ominus \theta)^2\right] &&= \int_{\mathcal{X}} (\widehat{\theta_1} \ominus \theta)^2 p_X(\boldsymbol{x}|\theta) \boldsymbol{dx}. \end{aligned}$$

Keep in mind that $\widehat{\theta_1}$ itself is a random variable, and its randomness comes from X.

Taking periodicity into account is crucial since close to the boundaries of Ω a small error ε^2 may become $(1-\varepsilon)^2$ if the periodic structure of the group is omitted.

Above, the bias and MSE were defined in term of the distribution p_X of the measurements, but they can also be defined in term of the distribution of the estimator itself

$$\begin{aligned} \mathsf{bias}_{\widehat{\theta_1}}(\theta) &= \int_\Omega (s \ominus \theta) p_{\widehat{\theta_1}}(s|\theta) ds, \\ \mathsf{MSE}_{\widehat{\theta_1}}(\theta) &= \int_\Omega (s \ominus \theta)^2 p_{\widehat{\theta_1}}(s|\theta) ds, \end{aligned}$$

both definition will be useful to derive results.

F.5.2 Computation of the MSE lower bound

Four steps are necessary to obtain the SOC constraint and the lower bound. We will need to introduce a function g, which we will call a filter since it appears within convolution products. In the third step, this undefined filter will be specifically set, such as to obtain the strongest possible conical constraint. Therefore, its symbol g is not present in the final result. For a given value of θ:

1. Convolve the bias with an indefinite filter g. We use the word "*filter*" for a function employed within a convolution product.
2. Apply Cauchy-Schwarz (CS) inequality to obtain a quadratic inequality, with the MSE as one of its sides.
3. Compute the filter g^{opt} making the CS inequality tight. The filter g^{opt} depends on the pdf of the estimator $p_{\widehat{\theta_1}|\theta}$.
4. Find the estimator pdf yielding the lowest MSE while verifying the SOC constraint — which is a second-order cone program (SOCP).

The procedure does not only provide a lower bound on the MSE, but also the distribution of the estimator achieving this bound. Note that not every estimator with a pdf lying inside the cone is feasible — but any pdf outside of the cone is infeasible. Therefore the bound is a true lower bound, but it may not be tight.

Filtering of the bias (step 1)

Filtering of the bias $(g * \mathsf{bias}_{\widehat{\theta_1}})(\theta)$ yields the equality

$$(g * h)(\theta) = \mathbb{E}_{X|\theta}\left[\underbrace{(g * p_X)(\theta)/\, p_{X|\theta}}_{u} \cdot \underbrace{(\widehat{\theta_1} \ominus \theta)}_{v}\right],$$

such that

$$h(\vartheta) \;=\; \int_{\Omega} (t \ominus \theta) p_{\widehat{\theta_1}}(t|\theta \ominus \vartheta) dt. \tag{F.3}$$

Application of Cauchy-Schwarz inequality (step 2)

Cauchy-Schwarz inequality for random variables states that

$$\mathbb{E}\left[|v|^2\right] \geqq \left|\mathbb{E}\left[uv\right]\right|^2 \Big/ \mathbb{E}\left[|u|^2\right]. \tag{F.4}$$

So,

$$\mathsf{MSE}_{\widehat{\theta}}(\theta) \geqq \frac{\left|(g * h)\,(\theta)\right|^2}{\mathbb{E}\left[\left|(g * p_X)(\theta)/\, p_{X|\theta}\right|^2\right]}. \tag{F.5}$$

Notice that if one knows the estimator's distribution $p_{\widehat{\theta_1}}$ one also knows the MSE of the estimator which seems to defeat the purpose of a lowerbound computation. Nevertheless, a useful bound can still be obtained by minimization over all admissible distributions $p_{\widehat{\theta_1}}$.

Another objection, is that the ideal estimator $\widehat{\theta_1} = \theta$ satisfies (F.5) yielding an MSE and a tight lowerbound of 0 ! This is a well-known caveat of deterministic bounds. Two solutions are possible

1. Enforce the MSE inequality for different values of θ, and use a different measure of error such as the average MSE over several values of θ for which the estimations are equally difficult (semi-bayesian approach). The trivial estimator $\widehat{\theta_1} = \theta$ is no longer optimal in this case.
2. Use the circular symmetry of the problem to affirm that $p_{\widehat{\theta_1}}(t|\theta \oplus \vartheta) = p_{\widehat{\theta_1}}(t \ominus \vartheta|\theta)$, *i.e.* that a shift of the input parameter causes the same shift in the estimator distribution. The trivial estimator $\widehat{\theta_1} = \theta$ does not verify this property.

We choose, the second solution as it simplifies the computations.
In our AWGN setup, the denominator of (F.5) can be explicitely computed yielding

$$\mathsf{MSE}_{\widehat{\theta}}(\theta) \geqq \frac{\left|(g * h)\,(\theta)\right|^2}{\left(g(\vartheta) * \left(g(\vartheta') * q(\vartheta, \vartheta')\right)(\theta)\right)(\theta)}, \tag{F.6}$$

s.t.

$$q(\vartheta, \vartheta') \;\stackrel{\text{def}}{=}\; \kappa(\vartheta, \vartheta')\kappa^{-1}(\vartheta, \theta)\kappa^{-1}(\theta, \vartheta'),$$

$$\kappa((\vartheta, \vartheta') \;\stackrel{\text{def}}{=}\; e^{-\|\boldsymbol{s}_\vartheta - \boldsymbol{s}_{\vartheta'}\|^2/2N_0}.$$

which concludes step 2.

From inequality to equality (step 3)

Leaving aside the determination of $p_{\widehat{\theta_1}}$ for the moment, a choice must be made about which filter g to use. One solution is to plug-in various candidates, which would lead to a result analogous to the Barankin bound approximation schemes [86; 147; 1; 103]. Another approach is to determine which filter g^{opt} maximizes the right-hand side of (F.6). To our knowledge this approach was pioneered by Swerling [122] and surprisingly found little echo. Swerling used it on infinitely supported parameters, and said under which conditions it provides an approximation to the finitely supported case[1]. He also focused on shift invariant signals, which allows the use of Fourier analysis.

The CS inequality is tight (equality) if and only if the functions u and v in (F.4) are collinear. Collinearity can be identified in (F.6) as

$$h(\vartheta) = \left(g^{\mathrm{opt}}(\vartheta') * q(\vartheta, \vartheta')\right)(\theta). \tag{F.7}$$

With this condition satisfied, the numerator in (F.6) is the square of the denominator; thus any estimator with pdf $p_{\widehat{\theta_1}}$ must verify

$$\mathsf{MSE}_{\widehat{\theta}} \geq \left|(g^{\mathrm{opt}} * h)(\theta)\right|, \tag{F.8}$$

where both sides of the equation depend on $p_{\widehat{\theta_1}}$.

Equation (F.7) is a *linear integral equation*, and we will detail how to solve it in general in the next section. However, we can see immediately that h must be in the convolutional range of the kernel q to have a solution. Moreover if q is almost singular, the solution may be unstable with respect to small variations in h. Since h is a function of the estimator distribution, the later must be taken into account.

With this in mind, we shall go back to solve (F.7). We assume that

$$\|s_t - s_{t+\vartheta}\|^2/2N_0 \;\approx\; \|s_0 - s_\vartheta\|^2/2N_0 \;\stackrel{\mathrm{def}}{=}\; E_p/N_0(1-\rho(\vartheta)),\ \forall t,$$

which we call *shift invariance*. The function ρ is the normalised autocorrelation – $\rho(0) = 1$ – of the sampled signal, and E_p/N_0 is the SNR. In that case, q is a convolution kernel

$$q(\vartheta, \vartheta') \;=\; q(\vartheta - \vartheta'),$$

and the integral equation (F.7) can be solved by means of Continuous Time Fourier Series (CTFS) expansion. Let $\bar{g}(\vartheta') \stackrel{\mathrm{def}}{=} g^{\mathrm{opt}}(\vartheta') \cdot \kappa^{-1}(\theta - \vartheta')$,

$$\begin{aligned} h(\vartheta) \;&=\; \left(g^{\mathrm{opt}}(\vartheta') * q(\vartheta - \vartheta')\right)(\theta) \\ \Leftrightarrow \quad h(\vartheta)\kappa(\vartheta) \;&=\; \int_\Omega \bar{g}(\vartheta')\kappa(\vartheta - \vartheta')d\vartheta' \\ &\quad\ {\scriptstyle \mathrm{CT}}\updownarrow{\scriptstyle \mathrm{FS}} \\ (H * K)[n] \;&=\; \bar{G}[n] \cdot K[n]. \end{aligned} \tag{F.9}$$

[1] Notably, Swerling's result only apply to the high SNR regime.

Then

$$\begin{aligned}\mathsf{MSE}_{\widehat{\theta}} &\geqq \left|\int_{\Omega} g^{\mathrm{opt}}(\vartheta)h(\theta-\vartheta)d\vartheta\right| \\ &= \left|\int_{\Omega} \bar{g}(\vartheta')\kappa(\theta-\vartheta)h(\theta-\vartheta)d\vartheta\right| \\ &= \sum_{\mathbb{Z}} \frac{|(H*K)[n]|^2}{|K[n]|}, \end{aligned} \tag{F.10}$$

where the last equality follows by substitution of $\bar{G}$ with (F.9).

It is immediately apparent, that if the elements specific to the estimator such as the bias and its pdf $p_{\widehat{\theta}_1}$ are neglected, H is the CTFS of a "sawtooth" function which decays at a rate $O(1/n)$.

If K decays at a faster rate, $|(H*K)[n]|/|K[n]|$ is growing with $|n|$, therefore the summation is unbounded. It shows that if periodicity is invoked to justify the use of Fourier Series formalism, its consequences must be taken into account to avoid a diverging lowerbound — and this *at any SNR*. In Figure F.6, we will show how an apparently accurate but erroneous threshold detection can be achieved, if this observation is neglected and suboptimal filters are used.

Plugging the estimator's properties in (step 4)

Because of the symmetry of the problem, the distribution of the estimator is assumed to be shift-invariant

$$p_{\widehat{\theta}_1}(t|\theta) \;=\; p_{\widehat{\theta}_1}(t\ominus\theta|0) \;\stackrel{\mathrm{def}}{=}\; p_{\widehat{\theta}_1}(t\ominus\theta).$$

This property turns the definition of h in (F.3) into a convolution product between the time-reversed sawtooth function and the estimator's distribution

$$\begin{aligned} h(\vartheta) \;&=\; \int_{\Omega} (t\ominus\theta)p_{\widehat{\theta}_1}(t\ominus\theta\oplus\vartheta)dt, \\ &\underset{\mathrm{CT}}{\overset{}{\updownarrow}}{}_{\mathrm{FS}} \\ &= \mathcal{F}\left\{-t\right\}[n]P[n]e^{-2\pi\jmath\theta n}e^{2\pi\jmath\theta n}, \\ &= \;-\frac{(-1)^n}{2\pi\jmath n}P[n]. \end{aligned} \tag{F.11}$$

Therefore, the CTFS P of the estimator's pdf acts as a damping factor. By definition, the MSE expressed in the CTFS domain is

$$\begin{aligned}\mathsf{MSE}_{\widehat{\theta}} \;&=\; \sum_{\mathbb{Z}} \mathcal{F}\{t^2\}[n]\cdot P[n], \\ &= P[0]/12 + \sum_{\mathbb{Z}^*} \frac{(-1)^n}{2n^2\pi^2}P[n]. \end{aligned} \tag{F.12}$$

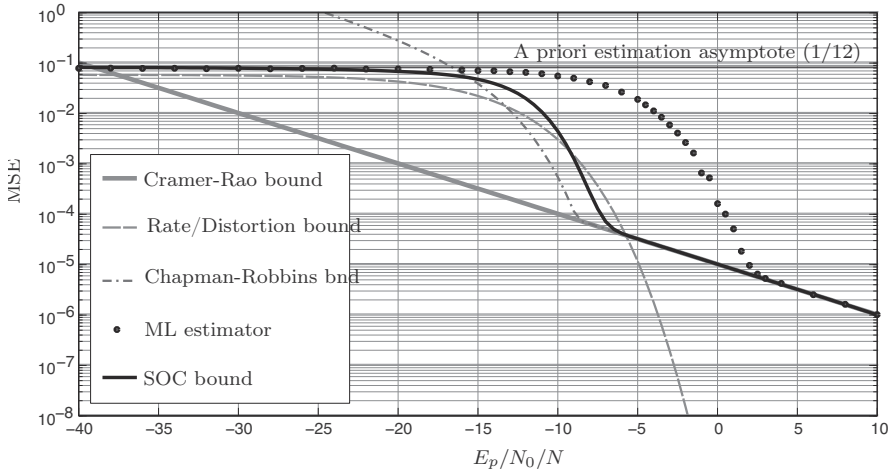


Figure F.3: *The reported bounds shall be classified in two types. The "geometrical" bounds such as the CRB, HCR and the proposed one obtained with the conical constraint (SOC). Only the third one is truly a lower-bound, and the gap with the MLE performances is due to the non-gaussian distribution of the MLE below the threshold SNR. The R/D bound is an information theoretic bounds which simulates the maximal capacity that can be obtained with a channel of similar SNR. The distribution of the corresponding estimator is in this case gaussian, which could explained the similar threshold point with the proposed geometrical bound.*

Putting together (F.10), (F.11) and (F.12), the admissible estimators have a pdf verifying

$$\frac{1}{12} + \sum_{\mathbb{Z}^*} \frac{(-1)^n}{2n^2\pi^2} P[n] \geqq \sum_{\mathbb{Z}} \frac{|((\mathcal{F}\{t\} \cdot P) * K)[n]|^2}{|K[n]|}, \tag{F.13}$$

$P[0] = 1$ so that the total probabilities sum to 1 in the time-domain. As a quick sanity check we verify the MSE is asymptotically correct: the uniform distribution, $P[n] = \delta[n]$, yields an MSE of $1/12$, and the δ distribution, $P[n] = 1$, an MSE of 0 since $\sum_{\mathbb{N}^*} (-1)^n/(n\pi)^2 = -1/12$.

This condition is *necessary* for the estimator to exist but it is *not sufficient*.

The solution of the bounding problem is thus found by solving a quadratically constrained linear problem:

$$\begin{aligned} \underset{P}{\text{minimize}} \quad & \mathsf{MSE}_{\widehat{\theta}} = \frac{1}{12} + \sum_{\mathbb{Z}^*} \frac{(-1)^n}{2n^2\pi^2} P[n] \\ \text{subject to} \quad & \mathsf{MSE}_{\widehat{\theta}} \geqq \sum_{\mathbb{Z}} \frac{|((\mathcal{F}\{t\} \cdot P) * K)[n]|^2}{|K[n]|}, \\ & P[0] = 1 \text{ , and } \mathcal{F}^{-1}\{P\} \text{ is non-negative.} \end{aligned} \tag{F.14}$$

The second line of constraints ensures P is the CTFS of a probability distribution. This problem is a second order cone optimisation (SOCP) for which fast and efficient solvers are readily available.

F.5.3 Application

Dirichlet kernel in AWGN

With the signal model (F.2), assume the number of samples is odd $N = 2M + 1$ and let the waveform be be the Dirichlet kernel with critical bandwidth M. Its uniform samples collected over Ω at rate $1/N$ are

$$s[n] \;=\; \frac{\sin(\pi n)}{\sin(\pi n/N)}, \quad |n| \leq M.$$

Its normalised autocorrelation is also a Dirichlet kernel of bandwidth M, and the CTFS of the autocorrelation sequence is

$$R[k] \;=\; 1/N, \quad \forall k \in \mathbb{Z}.$$

We could evaluate (F.11) and solve the SOCP problem (F.14) readily but first some intuitive choices for P can also be tried out.

Since the MSE and the conic constraint are related to the "compactness" of the estimator pdf in the time and frequency domain respectively, an educated guess is to choose a gaussian-like distribution since it has a small time-frequency product. We settle on the centered *wrapped gaussian* distribution with a single parameter γ. Its characteristic function is

$$P_\gamma[n] = e^{-\gamma^2 n^2/2}.$$

By design, P_γ is the CTFS of a probability distribution. The monotonicity with respect to γ in (F.13), guarantees the optimum is reached only if the conic constraint is active. Hence we solve for γ

$$\frac{1}{12} + \sum_{\mathbb{N}^*} \frac{(-1)^n e^{-\gamma^2 n^2/2}}{n^2\pi^2} \;=\; \sum_{\mathbb{Z}} \frac{\left| \left(\frac{(-1)^n e^{-\gamma^2 n^2/2}}{2\pi j n} * K \right)[n] \right|^2}{|K[n]|}. \tag{F.15}$$

The obtained MSE is an upperbound to (F.14). Even though we picked a particular family of estimator without optimisation the threshold indicated by the bound is 8dB away from the one of ML estimation as seen in Figure F.4. Since optimization can only lower the bound, we conclude that accurate threshold characerization is not achievable with this method alone.

We verified that the assumption of having a wrapped gaussian distribution is close to what can be obtained numerically using Galerkin methods (with "hat" functions) — see Figure F.5.

The gap can therefore be explained by the non-gaussian distribution of the ML estimator below the threshold SNR: large scale errors are uniformly distributed on the parameter's support. When this floor probability goes to 0 at high SNR, or becomes strongly dominant at low SNR, the ML error distribution is respectively a sharp or flat wrapped normal distribution, making the bound tight. Correct estimation of the probability of large errors is necessary to achieve good threshold estimation.

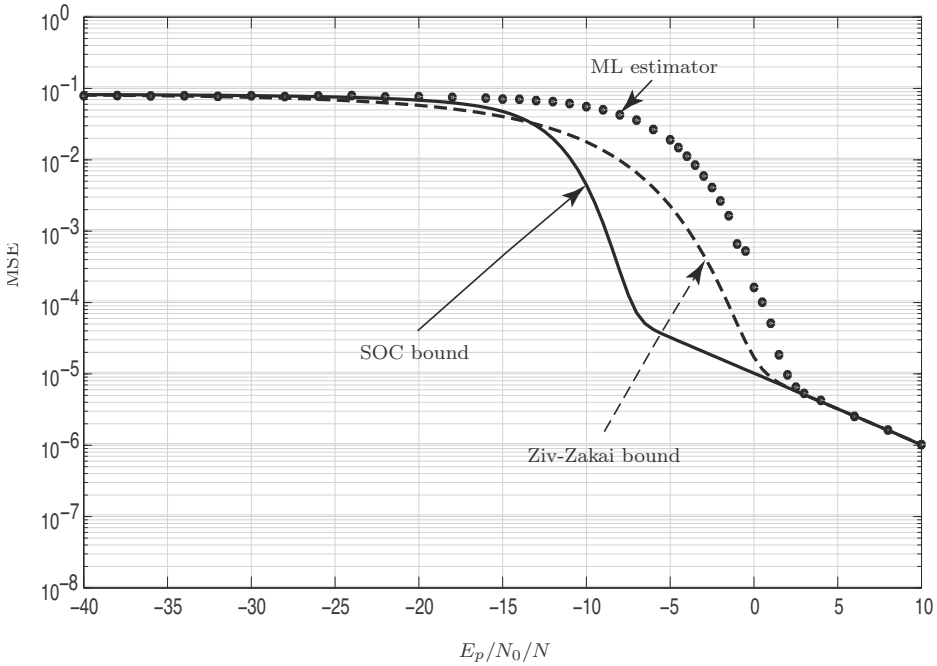


Figure F.4: *The ZZB requires the knowledge of the distribution of the optimal binary decision estimator, and can therefore infer the distribution of the optimal estimator implicitly, which gives it the best threshold prediction properties.*

The impact of periodicity

As briefly mentioned in the introduction, neglecting the contribution of the estimator's pdf may lead to misleading results. In (F.13), neglecting periodicity corresponds to consider an estimator with error distribution $P[n] = 1$ and to use the right-hand side of (F.13) as a lower bound on the MSE regardless of its admissibility.

The *incorrect* lowerbound is

$$\mathsf{MSE}_{\widehat{\theta}} \geqq \sum_{|n|<L} \frac{|(\mathcal{F}\{t\} * K)[n]|^2}{|K[n]|}.$$

As $L \to \infty$, this lowerbound diverges. For L finite, the lowerbound exhibits a threshold-like behaviour — shown in Figure F.6 — and diverges as the SNR diminishes. The conclusion is that Barankin bound approximation for threshold detection can be extremely tricky, especially if the filters are chosen in a non-optimal way; the threshold could be an artefact.

F.5.4 Conclusions

Lower bounds valid at any regime can be derived from the Cauchy-Schwarz inequality. The knowledge of the optimal estimator distribution is not strictly necessary, since

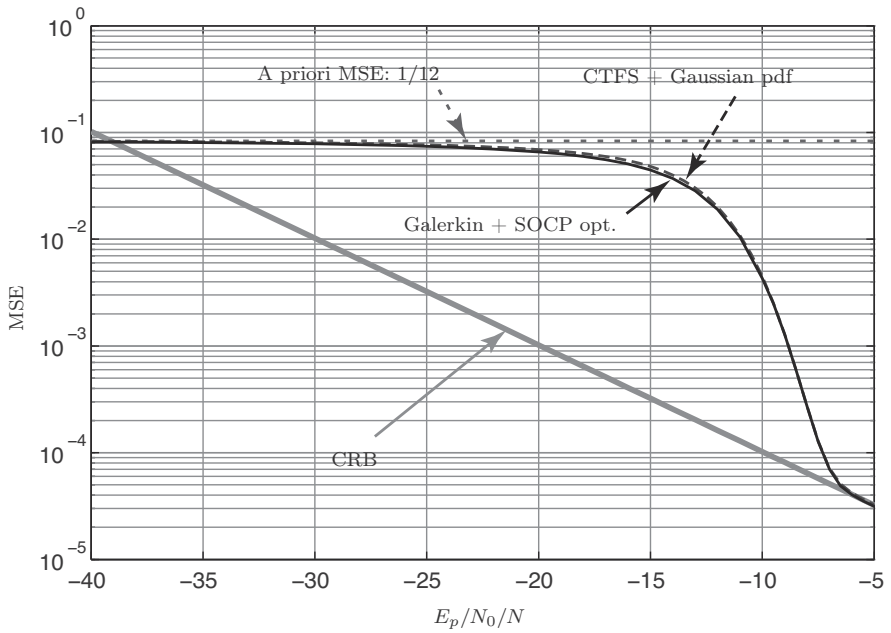


Figure F.5: *Galerkin method's followed by optimisation matches closely the result obtained with the analytical method (based on the CTFS) and the sensible choice of a wrapped-gaussian estimator distribution.*

a solution is to use the MSE inequality as a conical constraint within a (convex) minimisation problem. The caveat is that the minimal solution may be unachievable. The use of a *collinearity principle* to make the cone as small as possible indicates this limitation is not only apparent, and may require constraints on other moments. Furthermore, inadequate treatment of periodicity proved to be deceiptive : coupled with a heuristical design it can gives a false impression of threshold detection.

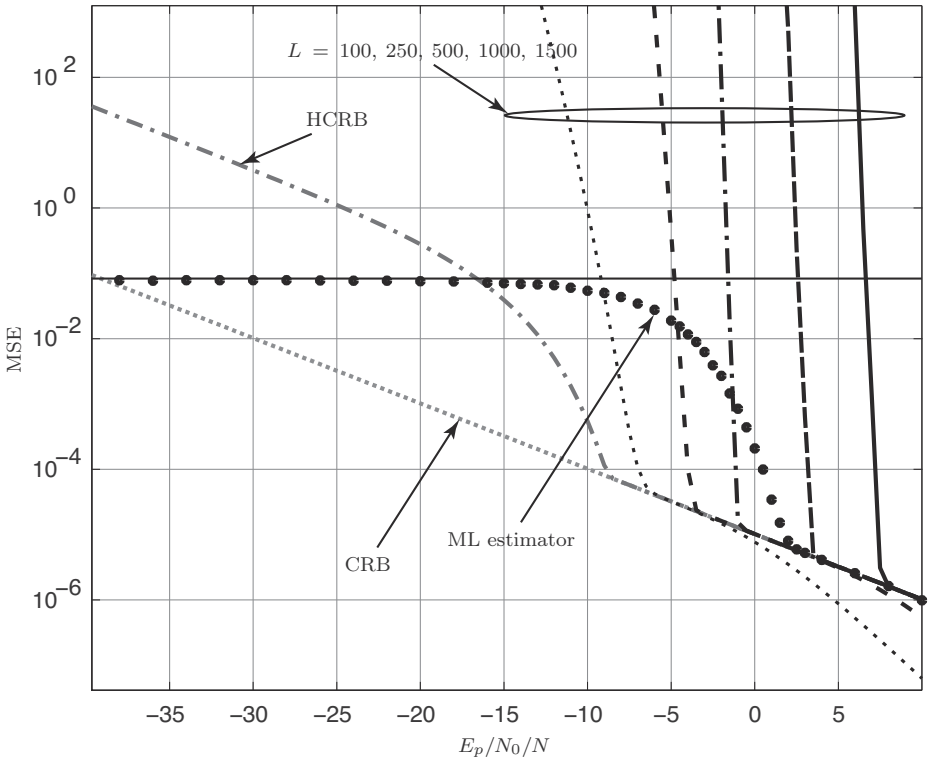


Figure F.6: *This figure shows the effect of not addressing periodicity correctly. At any SNR, there exists a filter which drives the corresponding bound to infinity. However, if the filter is not optimal, an "artificial" knee appears, which could lead to a false conclusion of threshold detection. The parameter L corresponds to the last index of the CTFS coefficients considered (highest frequency).*

Bibliography

[1] J. Abel. A bound on mean-square-estimate error. *Information Theory, IEEE Transactions on*, 39(5):1675 –1680, September 1993. ISSN 0018-9448. doi: 10.1109/18.259655. ↪ pp. 117, 169.

[2] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964. ISBN 0-486-61272-4. ↪ pp. 60, 61, 140, 141, 161.

[3] A. Agaskar and Y. Lu. Uncertainty principles for signals defined on graphs: Bounds and characterizations. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3493 –3496, march 2012. ↪ p. 138.

[4] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974. ↪ pp. 48, 50.

[5] H. Akaike. A bayesian extension of the minimum aic procedure of autoregressive model fitting. *Biometrika*, 66(2):237–242, 1979. ↪ p. 50.

[6] H. Akaike. *Prediction and entropy*. Springer, 1998. ↪ p. 50.

[7] S. M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE Journal on Selected Areas in Communications*, 16(8):1451 – 1458, 1998. doi: 10.1109/49.730453. ↪ pp. 3, 17.

[8] J. Albuquerque. The barankin bound: A geometric interpretation (Corresp.). *Information Theory, IEEE Transactions on*, 19(4):559–561, July 1973. ↪ p. 117.

[9] A. Angelow and M.-C. Batoni. About heisenberg uncertainty relation. *Bulgarian Journal of Physics*, 26(5–6):193–203, 1999. ↪ p. 101.

[10] F. Arscott. The land beyond bessel: A survey of higher special functions. In W. Everitt and B. Sleeman, editors, *Ordinary and Partial Differential Equations*, volume 846 of *Lecture Notes in Mathematics*, pages 26–45. Springer Berlin Heidelberg, 1981. ISBN 978-3-540-10569-5. doi: 10.1007/BFb0089822. URL `http://dx.doi.org/10.1007/BFb0089822`. ↪ p. 6.

[11] G. Baechler. Sensing ecg signals with variable pulse width finite rate of innovation. Master's thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2012. ↪ p. 85.

[12] J. Baglama, D. Calvetti, and L. Reichel. IRBL: An implicitly restarted block-Lanczos method for large-scale Hermitian eigenproblems. *SIAM Journal on Scientific Computing*, 24(5):1650–1677, 2003. ↪ p. 40.

[13] E. Barankin. JSTOR: the annals of mathematical statistics, vol. 20, no. 4 (Dec., 1949), pp. 477-501. *The Annals of Mathematical Statistics*, 1949. ↪ pp. 114, 119.

[14] Y. Barbotin and M. Vetterli. Fast and Robust Parametric Estimation of Jointly Sparse Channels. *JETCAS special issue on Compressed sensing*, 2012. ↪ pp. 41, 69, 73.

[15] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Estimation of Sparse MIMO Channels with Common Support. *submitted to IEEE Transactions on Communications*, 2012. ↪ pp. 32, 36.

[16] Y. Barbotin, R. Parhizkar, and M. Vetterli. Properties of maximally compact sequences. Technical report, EPFL, 2012. (no citation)

[17] Y. Barbotin. Finite Rate of Innovation sampling techniques for embedded UWB devices. Master's thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2009. (no citation)

[18] Y. Barbotin. Faster Cadzow denoising based on partial eigenvalue decomposition. Patent US20100239103, 2010. (no citation)

[19] Y. Barbotin. Equalization by the pulse shape inverse of the input to the FRI processing in pulse based communications. Patent US20100238991, 2010. (no citation)

[20] Y. Barbotin and C. Budianu. Finite rate of innovation (FRI) techniques for low power body area networks. Patent US20100240309, 2010. (no citation)

[21] Y. Barbotin and C. Budianu. Emulation of N-bits uniform quantizer from multiple incoherent and noisy one-bit measurements. Patent US20100239055, 2010. (no citation)

[22] Y. Barbotin and M. Vetterli. Fast and Robust Parametric Estimation of Jointly Sparse Channels. *Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), IEEE*, PP(99):1–11, 2012. doi: 10.1109/JETCAS.2012.2214872. URL `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6310074&isnumber=5751207`. (no citation)

[23] Y. Barbotin, D. Van De Ville, T. Blu, and M. Unser. Fast Computation of Polyharmonic B-Spline Autocorrelation Filters. *IEEE Signal Processing Letters*, 15:773–776, 2008. ISSN 1070-9908. doi: 10.1109/LSP.2008.2006714. URL `http://bigwww.epfl.ch/teaching/projects/abstracts/barbotin/`. (no citation)

[24] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Estimating Sparse MIMO Channels Having Common Support. In *2011 IEEE International Conference On Acoustics, Speech, And Signal Processing*, International Conference on Acoustics Speech and Signal Processing ICASSP, pages 2920–2923, 2011. (no citation)

[25] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Sampling of Sparse Channels with Common Support. In *9th International Conference on Sampling Theory and Applications*, 2011. invited paper. (no citation)

[26] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Estimation of Sparse MIMO Channels with Common Support. *IEEE Transactions On Communications*, 60(12):3705–3716, 2012. ISSN 0090-6778. doi: 10.1109/TCOMM.2012.091112.110439. URL `http://infoscience.epfl.ch/record/185846`. ↪ p. 146.

[27] Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Estimating sparse MIMO channels having common support. Patent US20120099435, 2012. (no citation)

[28] M. S. Bartlett. A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 296–298, 1954. ↪ p. 51.

[29] A. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(1):189–202, 1995. ↪ p. 160.

[30] C. Berger, S. Zhou, J. Preisig, and P. Willett. Sparse channel estimation for multicarrier underwater acoustic communication: From subspace methods to compressed sensing. *IEEE Transactions on Signal Processing*, 58(3):1708–1721, 2010. ↪ p. 22.

[31] T. Blu, P. L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot. Sparse Sampling of Signal Innovations. *IEEE Signal Processing Magazine*, 25(2):31–40, 2008. ↪ pp. 31, 33, 36, 37, 38, 124, 126, 145.

[32] D. L. Boley and G. H. Golub. The lanczos-arnoldi algorithm and controllability. *Systems & control letters*, 4(6):317–324, 1984. ↪ p. 40.

[33] N. Bourbaki. *Éléments de mathematique*, volume I–XI. Hermann, 1970. ↪ p. 119.

[34] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. ↪ pp. 80, 108, 159.

[35] E. Breitenberger. Uncertainty measures and uncertainty relations for angle observables. *Foundations of Physics*, 15(3):353–364, 1985. ↪ pp. 4, 97, 99, 101, 102, 105.

[36] R. Brent, F. Luk, and C. Van Loan. Computation of the singular value decomposition using mesh-connected processors. *Journal of VLSI and computer systems*, 1(3):242–270, 1985. ↪ p. 42.

[37] Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(5):1081–1089, 1986. ↪ p. 80.

[38] A. Brown, P. Halmos, and C. Pearcy. Commutators of operators on Hilbert space. *Canad. J. Math*, 17:695–708, 1965. There is no typo, the original paper lacks a "s" at the end of "space". ↪ p. 101.

[39] W. Bryc, A. Dembo, and T. Jiang. Spectral measure of large random Hankel, Markov and Toeplitz matrices. *The Annals of Probability*, 34(1):1–38, 2006. ↪ pp. 51, 151.

[40] J. Cadzow. Signal enhancement-a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1), Jan 1988. ISSN 0096-3518. doi: 10.1109/29.1488. ↪ pp. 33, 35.

[41] L. L. Campbell. Minimum coefficient rate for stationary random processes. *Information and Control*, 3(4):360–371, 1960. ↪ p. 65.

[42] D. Cassioli, M. Win, and A. Molisch. The ultra-wide bandwidth indoor channel: from statistical model to simulations. *IEEE Journal on Selected Areas in Communications*, 20(6):1247 – 1257, aug 2002. ISSN 0733-8716. doi: 10.1109/JSAC.2002.801228. ↪ p. 22.

[43] J. Cavallaro and F. Luk. CORDIC Arithmetic for an SVD Processor. *J. of parallel and dist. computing*, 5(3):271–290, 1988. ↪ p. 42.

[44] D. Chapman and H. Robbins. Minimum variance estimation without regularity assumptions. *The Annals of Mathematical Statistics*, 22(4):581–586, 1951. ↪ pp. 119, 128.

[45] E. Chaumette. A new barankin bound approximation for the prediction of the threshold region performance of maximum likelihood estimators. *Signal Processing, IEEE Transactions on*, 56(11):5319–5333, 2008. ↪ pp. 117, 120.

[46] A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, R. F. Murphy, and J. Kovacević. A multiresolution approach to automated classification of protein subcellular location images. *BMC bioinformatics*, 8:210, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-210. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933440/`. (no citation)

[47] T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2006. ↪ pp. 55, 65.

[48] H. Cramér. *Mathematical methods of statistics*. Princeton University Press, Princeton NJ, 1946. ↪ pp. 4, 114.

[49] M. Davies and Y. Eldar. Rank awareness in joint sparse recovery. *Information Theory, IEEE Transactions on*, 58(2):1135–1146, 2012. ↪ p. 71.

[50] A. Dembo. Information inequalities and uncertainty principles. Technical Report 75, Dept. Statistics, Stanford Univ., Stanford CA, July 1990. ↪ pp. 4, 118.

[51] P. Embrechts, T. Mikosch, and C. Klüppelberg. *Modelling extremal events: for insurance and finance*. Springer-Verlag, 1997. ↪ p. 61.

[52] W. Erb. Uncertainty principles on compact riemannian manifolds. *Applied and Computational Harmonic Analysis*, 29(2):182 – 197, 2010. ISSN 1063-5203. doi: 10.1016/j.acha.2009.08.012. URL `http://www.sciencedirect.com/science/article/pii/S1063520309000955`. ↪ p. 4.

[53] A. Erdélyi, W. Magnus, F. Oberhettinger, F. Tricomi, and H. Bateman. *Higher transcendental functions*, volume 3. McGraw-Hill New York, 1955. ↪ p. 162.

[54] ETSI committee for 3GPP. 3GPP TR 25.913 V9.0 – requirements for E-UTRA and E-UTRAN. ETSI, 2009. ↪ pp. 28, 38.

[55] ETSI committee for DVB. 300 744, DVB – Framing, channel coding and modulation for digital terrestrial television. ETSI, 01 2001. ↪ p. 28.

[56] R. Fletcher. Practical methods of optimization john wiley & sons. *New York*, 1987. ↪ pp. 80, 85.

[57] D. Fraser and I. Guttman. Bhattacharyya bounds without regularity assumptions. *The Annals of Mathematical Statistics*, 23(4):629–632, 1952. ↪ p. 119.

[58] D. Gabor. Theory of communication. *Journ. IEE*, 93:429–457, 1946. ↪ p. 4.

[59] M. J. Gander and G. Wanner. From euler, ritz, and galerkin to modern computing. *SIAM Review*, 54(4):627–666, 2012. ↪ p. 131.

[60] G. Golub and C. Van Loan. *Matrix Computations*. The Johns Hopkins Uni. Press, 1996. ↪ p. 79.

[61] M. Grant and S. Boyd. cvx: Matlab software for disciplined convex programming, version 1.21, 2010. ↪ p. 107.

[62] M. Gu and S. C. Eisenstat. A divide-and-conquer algorithm for the symmetric tridiagonal eigenproblem. *SIAM Journal on Matrix Analysis and Applications*, 16(1):172–191, 1995. ↪ p. 40.

[63] E. Gumbel. Les valeurs extrêmes des distributions statistiques. In *Annales de l'institut Henri Poincaré*, volume 5.2, pages 115–158. PUF, 1935. ↪ pp. 61, 152.

[64] J. Guy, B. Mangeot, and A. Sales. Solutions for fredholm equations through nonlinear iterative processes. *Journal of Physics A: Mathematical and General*, 17(7):1403, 1984. URL `http://stacks.iop.org/0305-4470/17/i=7/a=008`. ↪ p. 131.

[65] P. Hansen. Numerical tools for analysis and solution of fredholm integral equations of the first kind. *Inverse problems*, 8:849, 1992. ↪ p. 131.

[66] W. Heisenberg. The actual content of quantum theoretical kinematics and mechanics. *Physikalishche Z.*, 43:172, 1927. (Translated). ↪ pp. 4, 97.

[67] H. Hofstetter, C. Mecklenbräuker, R. Müller, H. Anegg, H. Kunczier, E. Bonek, I. Viering, and A. Molisch. MIMO dataset (Weikendorf, Austria), 2002. URL `http://measurements.ftw.at/MIMO.html`. ↪ pp. 20, 21, 22, 70, 78.

[68] R. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985. ISBN 0-521-30586-1. ↪ pp. 147, 148.

[69] E. Husserl. Die reine phänomenologie ihr forschungsgebiet und ihre methode. In *Aufsätze und Vorträge (1911–1921)*, pages 68–81. Springer, 1986. ↪ p. 1.

[70] W. C. Jakes. *Microwave mobile communications*. IEEE press, 1974. ↪ p. 77.

[71] S. Kaniel. Estimates for some computational techniques in linear algebra. *Mathematics of Computation*, 20(95):369–378, 1966. ↪ pp. 42, 43.

[72] Z. Khalid, S. Durrani, P. Sadeghi, and R. Kennedy. Concentration uncertainty principles for signals on the unit sphere. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 3717 –3720, march 2012. ↪ p. 4.

[73] J. Kunisch and J. Pamp. Measurement results and modeling aspects for the UWB radio channel. In *Ultra Wideband Systems and Technologies, 2002. Digest of Papers. 2002 IEEE Conference on*, pages 19 – 23, 2002. doi: 10.1109/UWBST.2002.1006310. ↪ p. 22.

[74] C. Lanczos. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. US Governm. Press Office, 1950. ↪ p. 40.

[75] D. N. Lawley. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika*, 43(1/2):pp. 128–136, 1956. URL `http://www.jstor.org/stable/2333586`. ↪ p. 51.

[76] R. B. Lehoucq and D. C. Sorensen. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17(4):789–821, 1996. ↪ p. 40.

[77] P. Lemmerling and S. Van Huffel. Structured total least squares. In *Total Least Squares and Errors-in-Variables Modeling*, pages 79–91. Springer, 2002. ↪ p. 79.

[78] K. Levenberg. A method for the solution of certain problems in least squares. *SIAM J Appl Math*, 11(2):431–441, 1944. ↪ p. 85.

[79] Z. Luo, W. Ma, A. M. C. So, Y. Ye, and S. Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, 27(3): 20–34, 2010. ↪ p. 160.

[80] N. W. MacLachlan. *Theory and Applications of Mathieu Functions*, page 232. Dover Publications, 1964. ↪ p. 6.

[81] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 2009. ↪ pp. 99, 105.

[82] I. Markovsky and S. Van Huffel. Overview of total least-squares methods. *Signal processing*, 87(10):2283–2302, 2007. ↪ p. 79.

[83] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, 11(2): 431–441, 1963. ↪ p. 85.

[84] S. Massar and P. Spindel. Uncertainty relation for the discrete Fourier transform. *Phys. Rev. Lett.*, 100:190401, May 2008. ↪ p. 97.

[85] J. Massey. Shift-register synthesis and BCH decoding. *Information Theory, IEEE Transactions on*, 15(1):122–127, 1969. ↪ p. 32.

[86] R. McAulay. A useful form of the barankin lower bound and its application to PPM threshold analysis. *Information Theory, IEEE Transactions on*, 15(2): 273–279, 1969. ↪ pp. 117, 169.

[87] P. McCormick and F. Elliston. *Husserl: Shorter Works*. Notre Dame University Press, 1981. ↪ p. 1.

[88] M. Meckes. On the spectral norm of a random Toeplitz matrix. *Elec. Comm. in Proba.*, 12:315–325, 2007. ↪ pp. 51, 148.

[89] A. Molisch. *Wireless Communications*. Wiley, 2005. ↪ pp. 15, 28, 79.

[90] J. R. Munkres. *Topology*, volume 2. Prentice Hall Upper Saddle River, 2000. ↪ p. 84.

[91] H. Nguyen and H. Van Trees. Comparison of performance bounds for doa estimation. In *Statistical Signal and Array Processing, 1994., IEEE Seventh SP Workshop on*, pages 313 –316, jun 1994. doi: 10.1109/SSAP.1994.572506. ↪ p. 165.

[92] J. Nocedal and S. J. Wright. *Numerical optimization*, volume 2. Springer New York, 1999. ↪ pp. 82, 85, 86.

[93] B. Ottersten, P. Stoica, and R. Roy. Covariance matching estimation techniques for array signal processing applications. *Digital Signal Processing, Transactions on*, 8(3):185–210, 1998. doi: doi:10.1006/dspr.1998.0316. ↪ p. 48.

[94] R. Parhizkar, Y. Barbotin, and M. Vetterli. Sequences with minimal time-frequency uncertainty. *CoRR*, abs/1302.2082, 2013. ↪ pp. 95, 109, 110.

[95] R. Parhizkar, Y. Barbotin, and M. Vetterli. Sequences with Minimal Time-Frequency Uncertainty. *submitted to Applied and Computational Harmonic Analysis*, 2013. ISSN 1063-5203. URL `http://infoscience.epfl.ch/record/183740`. (no citation)

[96] R. Parhizkar, Y. Barbotin, and M. Vetterli. Sequences with Minimal Time-Frequency Spreads. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013. (no citation)

[97] B. Parlett. *The symmetric eigenvalue problem*, volume 20. SIAM, 1998. ↪ pp. 39, 40.

[98] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, pages 339–358, 1998. ↪ p. 160.

[99] J. Prestin and E. Quak. Optimal functions for a periodic uncertainty principle and multiresolution analysis. *Proceedings of the Edinburgh Mathematical Society*, 42(2):225–242, 1999. ↪ pp. 4, 6, 99, 106.

[100] J. Prestin, E. Quak, H. Rauhut, and K. Selig. On the connection of uncertainty principles for functions on the circle and on the real line. *Journal of Fourier Analysis and Applications*, 9(4):387–409, 2003. ↪ pp. 4, 97, 99, 100, 110, 112, 158.

[101] R. Prony. Essai experimental et analytique. *J. Ecole Polytech. (Paris)*, 2:24–76, 1795. ↪ p. 31.

[102] T. Przebinda, V. DeBrunner, and M. Ozaydin. Using a new uncertainty measure to determine optimal bases for signal representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1999. ↪ p. 97.

[103] A. Quinlan, E. Chaumette, and P. Larzabal. 2006 IEEE international conference on acoustics speed and signal processing proceedings. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing*, page III–808–III–811. IEEE, 2006. ↪ p. 169.

[104] D. Rife and R. Boorstyn. Single tone parameter estimation from discrete-time observations. *Information Theory, IEEE Transactions on*, 20(5):591–598, 1974. ↪ p. 165.

[105] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. ↪ pp. 48, 50.

[106] T. Routtenberg and J. Tabrikian. Periodic crb for non-bayesian parameter estimation. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2448 –2451, may 2011. doi: 10.1109/ICASSP.2011.5946979. ↪ pp. 118, 122.

[107] O. Roy and M. Vetterli. The Effective Rank: A Measure of Effective Dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pages 606–610, 2007. ↪ pp. 48, 65, 66.

[108] R. Roy and T. Kailath. ESPRIT-Estimation of Signal Parameters Via Rotational Invariance Techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37:984–995, 1989. ↪ pp. 4, 34.

[109] Y. Saad. On the rates of convergence of the lanczos and the block-lanczos methods. *SIAM Journal on Numerical Analysis*, 17(5):687–706, 1980. ↪ pp. 42, 43.

[110] J. Salz and J. Winters. Effect of fading correlation on adaptive arrays in digital mobile radio. *IEEE Transactions on Vehicular Technology*, 43, 1994. ↪ pp. 18, 140.

[111] E. Schrödinger. About Heisenberg uncertainty relation. *Proceedings of The Prussian Academy of Sciences*, XIX:296–303, 1930. ↪ pp. 4, 101.

[112] A. Sen and B. Virág. The top eigenvalue of the random toeplitz matrix and the sine kernel. *arXiv preprint arXiv:1109.5494*, 2011. ↪ pp. 150, 151.

[113] S. Shamai and A. Wyner. A binary analog to the entropy-power inequality. *Information Theory, IEEE Transactions on*, 36(6):1428–1430, 1990. ISSN 0018-9448. doi: 10.1109/18.59938. ↪ p. 65.

[114] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, July, October 1948. ↪ p. 27.

[115] A. Shapiro. Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika*, 47(2):187–199, 1982. ↪ p. 160.

[116] N. Sharma, S. Das, and S. Muthukrishnan. Entropy power inequality for a family of discrete random variables. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 1945–1949, 2011. doi: 10.1109/ISIT.2011.6033891. ↪ p. 65.

[117] H. D. Simon. Analysis of the symmetric lanczos algorithm with reorthogonalization methods. *Linear Algebra and Its Applications*, 61:101–131, 1984. ↪ p. 40.

[118] D. Slepian. Some comments on Fourier analysis, uncertainty and modeling. *SIAM Rev.*, 25(3):379–393, July 1983. ↪ pp. 4, 6.

[119] P. Stoica and R. L. Moses. *Introduction to spectral analysis*, volume 1. Prentice hall New Jersey:, 1997. ↪ pp. 4, 7.

[120] P. Stoica and K. C. Sharman. Novel eigenanalysis method for direction estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 137, pages 19–26. IET, 1990. ↪ p. 81.

[121] P. Stoica, R. L. Moses, B. Friedlander, and T. Soderstrom. Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(3):378–392, 1989. ↪ p. 79.

[122] P. Swerling. Parameter estimation for waveforms in additive gaussian noise. *Journal of the Society for Industrial and Applied Mathematics*, 7(2):pp. 152–166, 1959. ISSN 03684245. URL `http://www.jstor.org/stable/2099089`. ↪ pp. 114, 118, 169.

[123] G. Szegö. *Orthogonal polynomials. Revised ed.* American Mathematical Society (AMS), 1939. ↪ pp. 39, 40.

[124] E. Telatar. Capacity of multi-antenna Gaussian channels. *European trans. on telecommunications, Wiley*, 10(6):585–595, 1999. ISSN 1541-8251. ↪ p. 3.

[125] M. J. Todd. Semidefinite optimization. *Acta Numerica*, 10, 2001. ↪ p. 160.

[126] B. Torresani. Position-frequency analyis for signals defined on spheres. *Signal Processing*, 43(3):341–346, 1995. ↪ p. 4.

[127] G. Tripathi. A matrix extension of the cauchy-schwarz inequality. *Economics Letters*, 63(1):1–3, 1999. ↪ p. 164.

[128] M. Trnovská. Strong duality conditions in semidefinite programming. *Journal of Electrical Engineering*, 56(12), 2005. ↪ p. 160.

[129] D. Tufts and R. Kumaresan. Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood. *Proceedings of the IEEE*, 70:975–989, 1982. ↪ pp. 31, 145.

[130] G. Turin. *Communication through noisy, random-multipath channels*. PhD thesis, MIT, 1956. ↪ pp. 3, 13.

[131] J. J. Van De Beek, P. Brjesson, H. Eriksson, J.-O. Gustavsson, and L. Olsson. MMSE estimation of arrival time with application to ultrasonic signals. Technical report, Lulea University of Technology, Div. of Signal Processing, Lule a University of Technology, 971 87 Lule a, Sweden, April 1993. ↪ pp. 133, 134.

[132] H. L. Van Trees. *Detection, Estimation and Modulation Theory, Part I*. John Wiley & Sons, New-York, 2001. ↪ pp. 4, 47, 50, 116, 118.

[133] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Transactions on Signal Processing*, 50(6):1417–1428, 2002. ISSN 1053-587X. ↪ p. 31.

[134] M. Vetterli, J. Kovacevic, and V. K. Goyal. *Foundations of Signal Processing*. Cambridge University Press, 2012. http://www.fourierandwavelets.org/. ↪ pp. 27, 42, 97, 98.

[135] M. Vetterli, Y. Barbotin, and A. Hormati. Methods and apparatus for estimating a sparse channel. Patent US20110103500, 2011. (no citation)

[136] R. von Mises. Über die "Ganzzahligkeit" der Atomgewichte und verwandte Fragen. *Physikalishche Z.*, 19:490–500, 1918. ↪ p. 97.

[137] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):387–392, April 1985. ↪ pp. 4, 48, 49, 50.

[138] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, 71:441–479, 1912. URL `http://eudml.org/doc/158545`. ↪ p. 53.

[139] J. Wozencraft and I. Jacobs. Principles of communication engineering. *New York*, page 49, 1965. ↪ p. 165.

[140] G. Xu and T. Kailath. Fast estimation of principal eigenspace using lanczos algorithm. *SIAM Journal on Matrix Analysis and Applications*, 15(3):974–994, 1994. ↪ pp. 4, 42, 43, 44.

[141] G. Xu, R. H. Roy, and T. Kailath. Detection of number of sources via exploitation of centro-symmetry property. *IEEE Transactions on Signal Processing*, 42 (1):102–112, 1994. ↪ pp. 38, 48.

[142] G. Xu, H. Zha, G. Golub, and T. Kailath. Fast algorithms for updating signal subspaces. *IEEE Transactions on Circuits and Systems—Part II: Express Briefs*, 41(8):537–549, 1994. ↪ pp. 38, 48.

[143] S. Yau and Y. Bresler. A compact Cramér-Rao bound expression for parametric estimation of superimposed signals. *IEEE Transactions on Signal Processing*, 40(5):1226 – 1230, 1992. ↪ p. 7.

[144] W. Yeh and C. Jen. High-speed and low-power split-radix FFT. *IEEE Transactions on Signal Processing*, 51(3):864–874, 2003. ↪ p. 42.

[145] Y. Yin and P. Krishnaiah. On some nonparametric methods for detection of the number of signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(11):1533–1538, 1987. ↪ p. 48.

[146] M. Zakai and J. Ziv. On the threshold effect in radar range estimation (Corresp.). *Information Theory, IEEE Trans. on*, 15(1):167–170, 1969. ↪ p. 136.

[147] A. Zeira and P. M. Schultheiss. Realizable lower bounds for time delay estimation. *IEEE Transactions on Signal Processing*, 41(11):3102–3113, 1993. ↪ pp. 117, 169.

[148] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai. On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis*, 20(1):1–25, 1986. ↪ p. 48.

[149] J. Ziv and M. Zakai. Some lower bounds on signal parameter estimation. *Information Theory, IEEE Trans. on*, 15(3):386–391, 1969. ↪ pp. 113, 136, 165.

La lumière de l'amour, propre à la foi, peut illuminer les questions de notre temps sur la vérité. La vérité aujourd'hui est souvent réduite à une authenticité subjective de chacun, valable seulement pour la vie individuelle. Une vérité commune nous fait peur, parce que nous l'identifions avec l'imposition intransigeante des totalitarismes. Mais si la vérité est la vérité de l'amour, si c'est la vérité qui s'entrouvre dans la rencontre personnelle avec l'Autre et avec les autres, elle reste alors libérée de la fermeture dans l'individu et peut faire partie du bien commun. Étant la vérité d'un amour, ce n'est pas une vérité qui s'impose avec violence, ce n'est pas une vérité qui écrase l'individu. Naissant de l'amour, elle peut arriver au coeur, au centre de chaque personne. Il résulte alors clairement que la foi n'est pas intransigeante, mais elle grandit dans une cohabitation qui respecte l'autre. Le croyant n'est pas arrogant ; au contraire, la vérité le rend humble, sachant que ce n'est pas lui qui la possède, mais c'est elle qui l'embrasse et le possède. Loin de le raidir, la sécurité de la foi le met en route, et rend possible le témoignage et le dialogue avec tous.

D'autre part, la lumière de la foi, dans la mesure où elle est unie à la vérité de l'amour, n'est pas étrangère au monde matériel, car l'amour se vit toujours corps et âme ; la lumière de la foi est une lumière incarnée, qui procède de la vie lumineuse de Jésus. Elle éclaire aussi la matière, se fie à son ordre, reconnaît qu'en elle s'ouvre un chemin d'harmonie et de compréhension toujours plus large. Le regard de la science tire ainsi profit de la foi : cela invite le chercheur à rester ouvert à la réalité, dans toute sa richesse inépuisable. La foi réveille le sens critique dans la mesure où elle empêche la recherche de se complaire dans ses formules et l'aide à comprendre que la nature est toujours plus grande. En invitant à l'émerveillement devant le mystère du créé, la foi élargit les horizons de la raison pour mieux éclairer le monde qui s'ouvre à la recherche scientifique.

Franciscus PP XVI, *Lumen Fidei*, §34 — 29 juin 2013.

Curiculum Vitæ

Yann Barbotin

Audiovisual Communications Laboratory (LCAV)
Swiss Federal Institute of Technology (EPFL)
1015 Lausanne, Switzerland

Personal

Date of birth:	March, 1985
Nationality:	French
Civil status:	Single

Education

2009–present	PhD candidate in School of Computer and Communication Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
2006–2009	MSc in Communication Systems, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
2003–2006	BSc in Communication Systems, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
2003	Baccalauréat Scientifique spécialité Mathématiques, *mention Très-bien*, Lycée International de Ferney-Voltaire.

Professional experience

2009–present	**Research and teaching assistant**, Audiovisual Communications Laboratory (LCAV), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
09/2008–03/2009	**Research intern**, Qualcomm Inc. (Corporate R&D), San Diego CA.
03/2007–09/2007	**Research intern**, Siemens Corporate Research, Princeton NJ.
06/2006–09/2006	**Research intern**, Center for Bioimage Informatics (CBI), Carnegie Mellon University (CMU), Pittsburgh PA.

Publications

Journal papers

1. A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, R. F. Murphy, and J. Kovacević. A multiresolution approach to automated classification of protein subcellular location images. *BMC bioinformatics*, 8:210, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-210. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933440/`

2. Y. Barbotin, D. Van De Ville, T. Blu, and M. Unser. Fast Computation of Polyharmonic B-Spline Autocorrelation Filters. *IEEE Signal Processing Letters*, 15:773–776, 2008. ISSN 1070-9908. doi: 10.1109/LSP.2008.2006714. URL `http://bigwww.epfl.ch/teaching/projects/abstracts/barbotin/`

3. Y. Barbotin and M. Vetterli. Fast and Robust Parametric Estimation of Jointly Sparse Channels. *Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), IEEE*, PP(99):1–11, 2012. doi: 10.1109/JETCAS.2012.2214872. URL `http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6310074&isnumber=5751207`

4. Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Estimation of Sparse MIMO Channels with Common Support. *IEEE Transactions On Communications*, 60(12):3705–3716, 2012. ISSN 0090-6778. doi: 10.1109/TCOMM.2012.091112.110439. URL `http://infoscience.epfl.ch/record/185846`

5. R. Parhizkar, Y. Barbotin, and M. Vetterli. Sequences with Minimal Time-Frequency Uncertainty. *submitted to Applied and Computational Harmonic Analysis*, 2013. ISSN 1063-5203. URL `http://infoscience.epfl.ch/record/183740`

Conference papers

1. Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Estimating Sparse MIMO Channels Having Common Support. In *2011 IEEE International Conference On Acoustics, Speech, And Signal Processing*, International Conference on Acoustics Speech and Signal Processing ICASSP, pages 2920–2923, 2011
2. Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Sampling of Sparse Channels with Common Support. In *9th International Conference on Sampling Theory and Applications*, 2011. invited paper
3. R. Parhizkar, Y. Barbotin, and M. Vetterli. Sequences with Minimal Time-Frequency Spreads. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013

Technical reports

1. Y. Barbotin. Finite Rate of Innovation sampling techniques for embedded UWB devices. Master's thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2009
2. Y. Barbotin, R. Parhizkar, and M. Vetterli. Properties of maximally compact sequences. Technical report, EPFL, 2012

Patents

(granted or filed)

1. Y. Barbotin and C. Budianu. Finite rate of innovation (FRI) techniques for low power body area networks. Patent US20100240309, 2010
2. Y. Barbotin. Faster Cadzow denoising based on partial eigenvalue decomposition. Patent US20100239103, 2010
3. Y. Barbotin. Equalization by the pulse shape inverse of the input to the FRI processing in pulse based communications. Patent US20100238991, 2010
4. Y. Barbotin and C. Budianu. Emulation of N-bits uniform quantizer from multiple incoherent and noisy one-bit measurements. Patent US20100239055, 2010
5. M. Vetterli, Y. Barbotin, and A. Hormati. Methods and apparatus for estimating a sparse channel. Patent US20110103500, 2011
6. Y. Barbotin, A. Hormati, S. Rangan, and M. Vetterli. Estimating sparse MIMO channels having common support. Patent US20120099435, 2012

Languages

French (fluent, native language), English (fluent, Cambridge CAE 2002)