TAMPERE UNIVERSITY OF TECHNOLOGY

Seyed Alireza Razavi

On Statistical Modelling and Hypothesis Testing by Information Theoretic Methods

Julkaisu 979 • Publication 979

Tampere 2011

Tampereen teknillinen yliopisto. Julkaisu 979
Tampere University of Technology. Publication 979

Seyed Alireza Razavi

On Statistical Modelling and Hypothesis Testing by Information Theoretic Methods

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 12th of August 2011, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2011

Supervisors:

Dr. Ciprian Doru Giurcăneanu,
Department of Signal Processing,
Tampere University of Technology,
Tampere, Finland.

Prof. Ioan Tăbuş (Custos),
Department of Signal Processing,
Tampere University of Technology,
Tampere, Finland.

Pre-examiners:

Dr. Peter Harremöes,
Copenhagen Business College,
Copenhagen, Denmark.

Dr. Teemu Roos,
Helsinki Institute of Information Technology (HIIT),
University of Helsinki,
Helsinki, Finland.

Opponents:

Prof. Petri Myllymäki,
Helsinki Institute of Information Technology (HIIT),
University of Helsinki,
Helsinki, Finland.

Dr. Peter Harremöes,
Copenhagen Business College,
Copenhagen, Denmark.

ISBN 978-952-15-2619-0 (printed)
ISBN 978-952-15-3618-2 (PDF)
ISSN 1459-2045

Abstract

The main objective of this thesis is to study various information theoretic methods and criteria in the context of statistical model selection. The focus in this research is on Rissanen's Minimum Description Length (MDL) principle and its variants, with a special emphasis on the Normalized Maximum Likelihood (NML).

We extend the Rissanen methodology for coping with infinite parametric complexity and discuss two particular cases. This is applied for deriving four NML-criteria and investigate their performance. Furthermore, we find the connection between Stochastic Complexity (SC), defined as minus logarithm of NML, and other model selection criteria.

We also study the use of information theoretic criteria (ITC) for selecting the order of autoregressive (AR) models in the presence of nonstationarity. In particular, we give a modified version of Sequentially NML (SNML) when the model parameters are estimated by forgetting factor LS algorithm.

Another contribution of the thesis is in connection with the new approach for composite hypothesis testing using Optimally Distinguishable Distributions (ODD). The ODD-detector for subspace signals in Gaussian noise is introduced and its performance is evaluated.

Additionally, we exploit the Kolmogorov Structure Function (KSF) to derive a new criterion for cepstral nulling, which has been recently applied to the problem of periodogram smoothing.

Finally, the problem of fairness in multiaccess communication systems is investigated and a new method is proposed. The new approach is based on partitioning the network into subnetworks and employing two different multiple-access schemes within and across subnetworks. It is also introduced an algorithm for selecting optimally the subnetworks such that to achieve the max-min fairness.

Preface

The research work presented in this Thesis has been accomplished in the Department of Signal Processing at Tampere University of Technology during the years 2007-2011.

I wish to express my deepest gratitude to my supervisors Dr. Ciprian Doru Giurcăneanu and Prof. Ioan Tabus for their continuous support and encouragement during last four years. Especially, I wish to thank Ciprian for his wise and highly professional guidance, and his effort in all steps of elaboration of this Thesis.

I am thankful to the pre-examiners of this Thesis, Dr. Teemu Roos and Dr. Peter Harremöes for their constructive feedback which significantly improved the quality of this work.

I would like to thank Prof. Ulla Ruotsalainen, the Dean of the Faculty of Computing and Electrical Engineering, Prof. Ari Visa, the Head of the Department of Signal Processing, and other professors and personnel of the Department for creating such an enjoyable academic environment. I am also thankful to Prof. Moncef Gabbouj, the former head of the Department, and Prof. Jaakko Astola, the Director of Tampere International Center for Signal Processing (TICSP), for helping me to join the Department. I am grateful to Ms. Ulla Siltaloppi, the Coordinator of International Education, Ms. Virve Larmila, the Secretary of the Department of Signal Processing, and Ms. Elina Orava, the Coordinator of International Education of the Faculty of Computing and Electrical Engineering, for the smooth administrative arrangements.

I would like to thank the Academy of Finland for supporting my Thesis via project numbers 113572, 118355, 134767 and 213462, and the Nokia Foundation for granting me two scholarships in 2009 and 2010.

I am thankful to my friends Payman Aflaki, Hamid Dadkhahi, and Mehdi Rezaei for giving me a lot of help and joy during last four years.

I wish to express my sincere gratitude to my beloved wife, Zahra, for her endless love, and to the sun of our life, our son, Amir Shahryar. During last seven years, I, Zahra and Amir Shahryar have shared all the sad and happy moments together. Nothing was possible without their patience, support and love.

Last, but by no means least, I would like to thank my parents, Havva and Hossein, for their unconditional sacrifices and for their continuous encouragement, and my brothers, Reza, Mohammad, and Morteza for their support.

This work is dedcated to my family and my parents.

Tampere, June 2011,
Seyed Alireza Razavi.

Contents

List of Figures

List of Tables

List of Publications

This thesis consists of the following publications. In the text, these publications are referred to as [**P1**],...,[**P7**].

P1 C. D. Giurcăneanu, S. A. Razavi, and A. Liski, "Variable selection in linear regression: several approaches based on normalized maximum likelihood," *Signal Processing*, vol. 91, Issue 8, pp. 1671–1692, Aug. 2011.

P2 C. D. Giurcăneanu, S. A. Razavi, "New insights on stochastic complexity," in *Proc. European Signal Processing Conference, 2009 (EUSIPCO 2009)*, Glasgow, Scotland, August 24-28, 2009, pp. 2475–2479.

P3 C. D. Giurcăneanu, S. A. Razavi, "AR order selection in the case when the model parameters are estimated by forgetting factor least-squares algorithms," *Signal Processing*, vol. 90, Issue 2, pp. 451–466, Feb. 2010.

P4 S. A. Razavi and C. D. Giurcăneanu, "Optimally distinguishable distributions: a new approach to composite hypothesis testing with applications to the classical linear models," *IEEE Trans. on Signal Processing*, vol. 57, no. 7, pp. 2445–2455, Jul. 2009.

P5 S. A. Razavi and C. D. Giurcăneanu,"Composite hypothesis testing by optimally distinguishable distributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2008 (ICASSP 2008)*, Las Vegas, Nevada, USA, March 31-April 4, 2008, pp. 3897–3900.

P6 C. D. Giurcăneanu and S. A. Razavi, "On the use of Kolmogorov structure function for periodogram smoothing," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2010 (ICASSP 2010)*, Dallas, Texas, USA, March 14-19 2010, pp. 3966 – 3969.

P7 S. A. Razavi and C. D. Giurcăneanu, "A novel method for improving fairness over multiaccess channels," *EURASIP Journal on Wireless Communications and Networking, vol. 2010, Article ID 395763, 10 pages, 2010. doi:10.1155/2010/395763.*

List of Abbreviations

ACT	Algorithmic Complexity Theory
AIC	Akaike Information Criterion
AME	Asymptotic Multiuser Efficiency
AR	AutoRegressive
ARMA	AutoRegressive Moving Average
BER	Bit Error Rate
BIC	Bayesian Information Criterion
CLT	Central Limit Theorem
CME	Conditional Model Estimator
DOF	Degree Of Freedom
ENR	Energy-to-Noise Ratio
FIM	Fisher Information Matrix
GIC	Generalized Information Criterion
GLRT	Generalized Likelihood Ratio Test
ITC	Information Theoretic Criteria
KC	Kolmogorov Complexity
KL	Kullback-Leibler
KSF	Kolmogorov Structure Function
KSFM	Modified KSF
LS	Least Squares
MA	Moving Average
MDL	Minimum Description Length
ML	Maximum Likelihood
MML	Minimum Message Length
NML	Normalized Maximum Likelihood
ODD	Optimally Distinguishable Distributions
OMA	Orthogonal Multiple Access
PDC	Predictive Densities Criterion
PDF	Probability Density Function
PLS	Predictive Least Squares
QoS	Quality of Service
RRM	Radio Resource Management

SC	Stochastic Complexity
SIC	Successive Interference Cancellation
SNLS	Sequentially Normalized Least-Squares
SNML	Sequentially Normalized Maximum Likelihood
TS	Time-Sharing
TV	Total Variance
UMPUT	Uniformly Most Powerful Unbiased Test

Mathematical Notations

$\mathrm{sgn}(\cdot)$	Sign of a real-valued variable
$\exp(\cdot)$	Exponential function
e	$\exp(1)$
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^+$	Set of positive real numbers
$\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{C})$	p-variate Gaussian distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\mathbf{C} \in \mathbb{R}^{p \times p}$
$\Gamma(\cdot)$	Euler integral of second kind
$(\cdot)^\top$	Transpose
$\|\| \cdot \|\|$	Euclidean norm
$\mathbb{E}_f[\cdot]$	Expected value with respect to distribution function $f(\cdot)$. Sometimes the index is dropped.
$o(\cdot)$	"Little-o", we say that $f(x) = o(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) = 0$
$O(\cdot)$	"Big-o", we say that $f(x) = O(g(x))$ if there are constants $M, x_0 > 0$ such that $\|f(x)\| \leq M\|g(x)\|$, for all $x \geq x_0$
P_{FA}	Probability of false alarm
P_{D}	Probability of detection
$\lfloor m \rfloor$	Largest integer less than or equal to the real-valued argument m
$\ln(\cdot)$	Natural logarithm
$\log_2(\cdot)$	Logarithm to base two
$\log(\cdot)$	Natural and base-two logarithm are both valid. The base determines the unit of information.
$D(f\|\|g)$	KL divergence between distributions $f(\cdot)$ and $g(\cdot)$
$\mathbf{I}$	Identity matrix of appropriate dimension
$\mathbf{0}$	Zero vector or matrix of appropriate dimension
$\|\cdot\|$	Depending on the context, the significance is the following: (I) absolute value of a scalar, (II) determinant of a matrix; (III) cardinality of a set; (IV) volume of a box

Chapter 1

Introduction

A principled method used to select a particular model from a class of models is based on the evaluation of the stochastic complexity (SC) for which the very first formula was introduced in [43]. The method is rooted in information theory and relies on a coding scenario for transmitting the available measurements from an hypothesized encoder to a decoder. The selection procedure chooses the model that allows the data to be encoded with the shortest code length, or equivalently, to minimize SC. This thesis is focused on new developments in both the theory and the applications of the main concepts related to SC.

The most recent advances in the theoretical basis of the SC emerge from a happy union between the algorithmic complexity theory (ACT) [36] and the coding theory [9]. As the central notions from the ACT, namely Kolmogorov complexity (KC), universal distribution and the structure function are non computable, their use in practical applications poses troubles. To circumvent such difficulties, Rissanen extended all these notions to statical models by replacing the set of programs from the algebraic theory of complexity with classes of parametric models. With the understanding that each model class is a likelihood function, the role of the universal model is played by the normalized maximum likelihood (NML) density function [5, 45]. Furthermore the KC is replaced by the SC that is defined as the minus logarithm of the NML. To construct the Kolmogorov structure function (KSF), the parameter space is partitioned into rectangles such that the Kullback-Leibler (KL) divergence between any two adjacent models is constant [48]. To the center of each rectangle it is assigned a special probability distribution function which takes value zero outside the rectangle and inside the rectangle coincides with the likelihood function properly normalized such that to integrate to one. This introduces a set of so-called optimally distinguishable distributions (ODD). Note that the idea of distinguishability is borrowed from the differential geometry [4]. Moreover, the distance between the "real" models and the distinguishable models depends on a parameter d. By applying the Central Limit Theorem (CLT), it can be shown that

there exists a unique d which minimizes this distance [48].

This PhD thesis is centered on the following lines of research:

- For most of the models which are commonly used in signal processing, the NML does not have a finite value because the term which corresponds to the parametric complexity is not finite. This problem was addressed in [47], where its was proposed an elegant solution based on a particular constraint of the parameter space. We investigate some of the properties of the resulting SC-formula [**P2**], and we also extend the approach from [47] for a general family of constraints [**P1**].

- We also consider the problem of order estimation in the case of piecewise autoregressive (AR) models. For such applications, we alter the sequentially normalized maximum likelihood (SNML) criterion from [51] such that to be compatible with the forgetting factor least-squares algorithms [25]. In [**P3**], we provide a solid analysis for the performance of the modified criterion as well as the performance of other criteria which have been introduced in the previous literature.

- The findings on ODD from [48,49] can be applied almost straightforwardly for composite hypothesis testing and, more importantly, they define a totally new framework for this problem. In the new paradigm, the decisions are based on how well the models can be fitted to the data, and it is not necessary to resort to the *level of the test*, which is generally taken to be 0.05. The approach is very promising, but because it is so new, it was utilized only for very few examples in order to illustrate the concept [46, 48]. In this Ph.D. thesis, it is shown how this paradigm can be applied to the detection of subspace signals in Gaussian noise [**P4**]-[**P5**].

- We also demonstrate in [**P6**] how the KSF can be used in cepstral analysis [68].

- Another result included in the thesis and which was published in [**P7**] concerns a novel approach for improving fairness in multiaccess communication systems.

The thesis comprises two parts. The first one is an introductory part, while the second one consists of seven publications. The structure of the first part is as follows. Chapter 2 introduces the fundamental notions like NML, KC, KSF. In Chapter 3, we elaborate on the computation of the NML such that to help the understanding of the content within publications [**P1**, **P2**, **P3**]. Chapter 4 is devoted to the ODD detector: We outline briefly the results from [**P4**], and we also include the complete proofs for some of the results from [**P5**]. The cepstral nulling problem is addressed in Chapter 5, where we discuss a modified version of the criterion used in [**P6**] for threshold selection. The performance of various criteria are illustrated by numerical examples. The aim of Chapter 6 is to gain more insight into the capabilities of

the methods from [**P7**]. Chapter 7 presents the summary of the publications and highlights the contributions of the author, while Chapter 8 concludes the thesis.

Chapter 2

A happy union: Algorithmic Complexity Theory and Coding Theory

Algorithmic Complexity Theory (ACT) emerged as a new branch of computer science in early sixties due to the seminal works of Kolmogorov, Solomonov and Chaitin [7, 29, 63, 64]. Because the central notions from ACT, namely Kolmogorov Complexity (KC), universal model, and the Kolmogorov Structure Function (KSF) are noncomputable, their use in practice poses troubles. To circumvent such difficulties, Rissanen extended all the notions from ACT to statistical models, which led to a novel method of inference [48]. The most important concepts of the newly introduced methodology are presented in this chapter.

2.1 Key definitions from ACT

Kolmogorov complexity and universal model: The algorithmic complexity of a string written with letters from a given alphabet is defined as the shortest program that can produce the string. When there exist infinite many programs that can produce any given string, one program will always be the shortest. To clarify the ideas, let us start with a simple example.

Example 2.1. Consider the following three strings:

- **STRING1**:
  ABABABABABABABABABABABABABABABABABABABABABAB-
  ABABABABABABABABABABABABABABABABABABABABABAB
- **STRING2**:
  RFGNAUWXTFXPQZMKIEUTGYQETQRBCTYFYBXMETDFAF-
  OICGHDWLBYARIBHXGQTREXJWEOVJMVXBYWSTXF
- **STRING3**:
  ABABABABABABABABABABABABABABABABABABABABABCZ-
  AUVWCKGULXEGDDWPODWQJNKBGDEGKBXYMMIXJC

STRING1 has a nice pattern and can be easily generated with the following Matlab program:

```
txt ='';
for i=1:40
txt=[txt,'AB'];
end
disp(txt)
```

STRING2 is quite random and, apparently, the easiest way to describe it is to write it down. Equivalently, the string is the output of the following Matlab program:

```
txt='RFGNAUWXTFXPQZMKIEUTGYQETQRBCTYFYBXMETDFAFOICGHDWLBYAR
IBHXGQTREXJWEOVJMVXBYWSTXF';
disp(txt)
```

The length of the program above equals the length of string itself plus an overhead which is independent of the length of string. Therefore, for long strings, the overhead is negligible in comparison with the string length.

STRING3 comprises two parts: the first one can be programmed in an easy way and therefore can be seen as a modelable string, but the second part is random and is treated as noise. This string can be seen as the output of the following Matlab program:

```
txt ='';
for i=1:20
txt=[txt,'AB'];
end
txt=[txt,'CZAUVWCKGULXEGDDWPODWQJNKBGDEGKBXYMMIXJC'];
disp(txt)
```

Example 2.1 provides intuition on how the notions from computer programming can be exploited to find regularities in arbitrary strings. In our examples, we use Matlab code because it is known that the length of the shortest program is language-

independent up to a constant [9].

Now we are prepared to introduce the KC [49]:

Definition 2.1. Kolmogorov complexity $K(x)$ of an object x is the length of the shortest program p (in some universal programming language U) which generates x as an output and terminates. More formally:

$$K(x) = \min\{|p| : U(p) = x\}. \tag{2.1}$$

Suppose that we are interested in the class of binary programs such that binary program p is analogous to a codeword for data x. The assumption that a program terminates after generating the string x as output implies that the countable set of data generating programs has the prefix property, i.e. no program can be a prefix of another one. In other words, if the programs are placed in a binary tree, then each program is a leaf of the tree. This implies that the set of Kolmogorov complexities (which is the KC of the elements of the set of all finite strings $\mathcal{X}$) satisfy the Kraft inequality [9]:

$$\sum_{x\in\mathcal{X}} 2^{-K(x)} \leq 1. \tag{2.2}$$

By normalizing we can get the following universal model for the set of finite binary strings:

$$P_U(x) = \frac{2^{-K(x)}}{\sum_{y\in\mathcal{X}} 2^{-K(y)}}. \tag{2.3}$$

Kolmogorov structure function: Consider any finite set S that includes x. Note that S defines a model for x in the following sense: All the members of S share some of the properties of x, but not necessarily all of its properties. In general, the cardinality $|S|$ is an inverse measure of the amount of properties shared by the members of S. The smaller is the amount of properties, more objects have those properties and, therefore, the larger is $|S|$. On contrary, the smaller is $|S|$, more properties are likely to be common for the few elements of S. In the very extreme case of $|S| = 1$, the string x is the only member of the set S, which means that the model S describes all properties of x. In general, for every finite set $S \ni x$ we have [73, 74]

$$K(x) \leq K(S) + \log|S| + O(1). \tag{2.4}$$

Kolmogorov proposed the following KSF for a given data x [73, 74]:

$$h_x(\alpha) = \min_{S\in\mathcal{S}}\{\log|S| : S \ni x, K(S) \leq \alpha\}, \tag{2.5}$$

where α is a parameter which controls the complexity of model S, and $\mathcal{S}$ is a set restricting the desirable properties. The structure function $h_x(\alpha)$ can be seen as a

measure of noise that is not modeled by a finite set of complexity not exceeding α.

2.2 From Algorithmic Complexity to Stochastic Complexity

Because KC is not computable, the KSF cannot be used for determining the optimal model. In this section we follow the approach of Rissanen to show how statistical models can be utilized to define computable structure functions as well as universal models. As a prepatory step, we give some definitions which are taken from [9, 22].

2.2.1 Key definitions

Consider a random variable X with possible outcomes $x_1, x_2, ..., x_m$ and corresponding probabilities $p(x_1), p(x_2), ..., p(x_m)$. A binary code C is a mapping from $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$ to a set of finite-length strings of bits. Let $C(x_i)$ denote the codeword corresponding to x_i, and let $l(x_i)$ denote the length of $C(x_i)$.

A code is called a prefix code or an instantaneous code if no codeword is a prefix of any other codeword. For any prefix code over a binary alphabet, the codeword lengths $l_1, l_2, \ldots, l_m$ must satisfy

$$\sum_{i=1}^{m} 2^{-l_i} \leq 1.$$

This is the Kraft inequality which we have already given in (2.2) for Kolmogorov complexities of finite strings. An *optimal code* is a prefix code whose expected code length $\sum_{i=1}^{m} p(x_i) l_i$ is minimum, where $p(x_i)$ is the probability of occurance of word i. It can be shown that the optimal code length is

$$l_i^* = -\log p(x_i). \tag{2.6}$$

Since l_i^* is not necessarily an integer, we cannot always set the optimal code lengths as in (2.6). Instead, we should apply the Huffman algorithm.

However, in this thesis we are not focused on efficiently compressing the data, but rather in the code length interpretation of the probability distributions. Thus, we will refer to $-\log p(x_i)$ as the code length assigned to the probability of x_i. Moreover, we assume that the available measurements are not symbols from a finite alphabet, but instead they are real numbers. Therefore, the probability distributions involved are not discrete, but continuous. One can apply a quantization process (with a certain precision) so as to transform the continuous distribution to a discrete one. This aspect will be discussed next in the context of the modelling problem.

2.2.2 Universal models

Let us consider the parametric models

$$\mathcal{M}_k = \{f(x^n;\boldsymbol{\theta},k) : \boldsymbol{\theta} = [\theta_1,\ldots,\theta_k]^\top \in \Omega\}, \tag{2.7}$$

$$\mathcal{M} = \{\mathcal{M}_k : k \geq 0\}, \tag{2.8}$$

where $x^n = x_1,\ldots,x_n$ are the observations and k denotes the number of parameters of the model. We assume that Ω is a subset of $\mathbb{R}^k$. Note that given observations x^n, each model defines a likelihood function, $f(x^n;\boldsymbol{\theta},k)$, that is here considered a function of the parameter vector $\boldsymbol{\theta}$.

By definition, $\hat{f}(x^n;k)$ is a *universal model* for $\mathcal{M}_k$ if

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{f(x^n;\boldsymbol{\theta},k)}{\hat{f}(x^n;k)} \tag{2.9}$$

for all parameters $\boldsymbol{\theta} \in \Omega$ [10]. The convergence is in a probabilistic sense, either in the mean taken with respect to $f(y^n;\boldsymbol{\theta},k)$, in probability or almost surely. We consider next two well-known universal models.

2.2.3 Mixture Model

One possibility to define a universal model is to take the mixture model [44]

$$f_\omega(x^n;k) = \int_\Omega f(x^n;\boldsymbol{\theta},k)\omega(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{2.10}$$

where $\omega(\boldsymbol{\theta})$ is a prior distribution. Note that $f_\omega(x^n;k)$ is a solution of the following optimization problem [8]

$$\min_q \int \omega(\boldsymbol{\theta}) D(f(x^n;\boldsymbol{\theta},k)\|q(x^n))\mathrm{d}\boldsymbol{\theta}. \tag{2.11}$$

Let

$$\mathbf{J}(\boldsymbol{\theta}) = \lim_{n\to\infty} \mathbf{J}_n(\boldsymbol{\theta}), \tag{2.12}$$

where

$$\mathbf{J}_n(\boldsymbol{\theta}) = -\frac{1}{n}\mathbb{E}\left[\frac{\partial^2 \ln f(x^n;\boldsymbol{\theta},k)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right] \tag{2.13}$$

is the Fisher information matrix (FIM). Then

$$\omega(\boldsymbol{\theta}) = \frac{|\mathbf{J}(\boldsymbol{\theta})|^{1/2}}{\int_\Omega |\mathbf{J}(\boldsymbol{\eta})|^{1/2}\mathrm{d}\boldsymbol{\eta}} \tag{2.14}$$

is the Jeffreys' prior, and it can be shown that asymptotically we have [8]

$$\mathbb{E}_{\boldsymbol{\theta}} \log \frac{f(x^n; \hat{\boldsymbol{\theta}}(x^n), k)}{f_\omega(x^n; k)} = \frac{k}{2} \log \frac{n}{2\pi} + \log \int_\Omega |\mathbf{J}(\boldsymbol{\eta})|^{1/2} \mathrm{d}\boldsymbol{\eta} + o(1). \tag{2.15}$$

2.2.4 Normalized maximum likelihood

Let $\hat{\boldsymbol{\theta}}(x^n)$ be the maximum likelihood (ML) estimate which minimizes the ideal code length $-\log f(x^n; \boldsymbol{\theta}, k)$ for a fixed k. The normalized maximum likelihood (NML) is expressed as

$$\hat{f}(x^n; k) = \frac{f(x^n; \hat{\boldsymbol{\theta}}(x^n), k)}{C_{n,k}} \tag{2.16}$$

where

$$C_{n,k} = \int_{y^n : \hat{\boldsymbol{\theta}}(y^n) \in \Omega} f(y^n; \hat{\boldsymbol{\theta}}(y^n), k) \mathrm{d}y^n. \tag{2.17}$$

NML was originally obtained as the solution to Shtarkov's minmax problem [61]:

$$\min_q \max_{x^n} \log \frac{f(x^n; \hat{\boldsymbol{\theta}}(x^n), k)}{q(x^n)}. \tag{2.18}$$

More recently, it was proved that NML is also the unique solution of the following maxmin problem [48]:

$$\max_g \min_q \mathbb{E}_g \log \frac{f(x^n; \hat{\boldsymbol{\theta}}(x^n), k)}{q(x^n)} = \max_g \min_q D(g\|q) - D(g\|\hat{f}(x^n; k)) + \log C_{n,k}. \tag{2.19}$$

The quantity

$$-\log \hat{f}(x^n; k) = -\log f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) + \log C_{n,k} \tag{2.20}$$

is called *stochastic complexity* (SC) of the data x^n given the model $\mathcal{M}_k$. The term $\log C_{n,k}$ is called *parametric complexity* and is a measure of learnable information [5].

In the case when the ML estimate satisfies the CLT, i.e. $\sqrt{n}(\hat{\boldsymbol{\theta}}(x^n) - \boldsymbol{\theta})$ converges in distribution to $\mathcal{N}_k(\mathbf{0}, \mathbf{J}^{-1}(\boldsymbol{\theta}))$, we have the following asymptotic formula for SC [45]:

$$-\log \hat{f}(x^n; k) = -\log f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) + \frac{k}{2} \log \frac{n}{2\pi} + \log \int_\Omega |\mathbf{J}(\boldsymbol{\eta})|^{1/2} \mathrm{d}\boldsymbol{\eta} + o(1), \tag{2.21}$$

where $\mathbf{J}(\cdot)$ is defined in (2.12).

By comparing the formulas in (2.15) and (2.21) one can notice immediately that

$$\mathbb{E}_{\boldsymbol{\theta}} \log \frac{f(x^n; \hat{\boldsymbol{\theta}}(x^n), k)}{f_\omega(x^n; k)}$$

coincides with

$$-\log \frac{\hat{f}(x^n;k)}{f(x^n;\hat{\boldsymbol{\theta}}(x^n),k)}$$

up to a constant which goes to zero as $n \to \infty$. We refer to [48, Section 5.2.2] for a much more elaborated analysis which clarifies the relationship between (2.15) and (2.21).

It is not straightforward to apply the approximation above for all model classes. For example, it has been proved in [45] that the NML of the Markov models can be evaluated with the formula in (2.21), but the very first results were published only in 2007. More precisely, the case of order-1 Markov chains for binary strings was investigated in [18], where it was shown that the integral term in (2.21) equals $4G$. We mention that G is the Catalan constant defined by $G = \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1)^2}$ and is usually approximated by $G \approx 0.915965594177$ [6]. Additionally, by utilizing results from [3], it was shown in [18] that the evaluation of the integral is very difficult for models whose order is larger than one. Hence, as it was pointed out in [17], for Markov models, it is preferable to resort to the approaches from [30] or [52] instead of using (2.21).

Note that, for many models used in signal processing, the value of the integral in (2.21) is not finite if the domain of integration is the entire parameter space. This problem is well-known and some of the proposed solutions involve specifically chosen restrictions for the ranges of the parameters. A comprehensive discussion on this issue can be found in [24]. For the important case of AR models, the reference [19] and the recently published article [57] are the only attempts to work out the NML-approximation from (2.21).

In the rest of the thesis, we employ methods for the calculation of the NML which do not involve this asymptotic approximation. However, publication [**P1**] investigates shortly how (2.21) can be used in the particular case of Gaussian linear regression.

2.2.5 Kolmogorov structure function

This section is based on [46,48,49] and contains only the most important definitions and notations. A more comprehensive presentation of the topic can be found in the aforementioned references.

Consider the finite partition $\Lambda_n = \{B_i : i = 1, 2, \ldots, N_n\}$ of the compact parameter space Ω for the model $\mathcal{M}_k$. Let $\boldsymbol{\theta}^i$ be the representative of the equivalent class B_i for all $i \in \{1, 2, \ldots, N_n\}$. Let $f_i = f(\cdot; \boldsymbol{\theta}^i, k)$ denote the quantized model. Additionally, assume that the Kullback-Leibler (KL) divergence $D(f_i \| f_{i+1})$ is constant for the adjacent models. This partitioning can be done by the following strategy:

Take $B_i = B_{d/n}(\boldsymbol{\theta}^i)$ as the maximal rectangle within the hyperellipsoid

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^i)^\top \mathbf{J}(\boldsymbol{\theta}^i)(\boldsymbol{\theta} - \boldsymbol{\theta}^i) = d/n, \tag{2.22}$$

centered at $\boldsymbol{\theta}^i$. The volume of equivalence class $B_{d/n}(\boldsymbol{\theta}^i)$ is given by

$$|B_{d/n}(\boldsymbol{\theta}^i)| = \Big(\frac{4d}{kn}\Big)^{k/2} |\mathbf{J}(\boldsymbol{\theta}^i)|^{-1/2}. \tag{2.23}$$

After observing x^n, we first obtain the ML estimate $\hat{\boldsymbol{\theta}}(x^n)$ and then take the center of the rectangle in which $\hat{\boldsymbol{\theta}}$ lies as the quantized version of $\hat{\boldsymbol{\theta}}$. Remark that due to (2.22) the centers of rectangles depend on the size parameter d.
To define the structure function, it is necessary to recast the notions from ACT in terms of stochastic models. So,

- A set of programs is to be replaced by a model class.
- A set S is to be replaced by a quantized model $f(x^n; \boldsymbol{\theta}^i, k)$.
- KC is to be replaced by SC.
- $K(S)$ is to be replaced by the shortest code length for $\boldsymbol{\theta}^i$ which is denoted by $L(\boldsymbol{\theta}^i)$.
- $\log |S|$, which is the maximum code length of $y \in S$ is to be replaced by the maximum or the mean code length of observations x^n for which $\hat{\boldsymbol{\theta}}(x^n) \in B_{d/N}(\boldsymbol{\theta}^i)$.

For an arbitrary $B_{d/n}(\boldsymbol{\theta}^i)$, we define

$$Q_{d/N}(i) \;=\; \int_{x^n : \hat{\boldsymbol{\theta}}(x^n) \in B_{d/n}(\boldsymbol{\theta}^i)} f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) \mathrm{d}x^n. \tag{2.24}$$

The calculation of the integral above is easier in the case when $f(x^n; \boldsymbol{\theta}, k)$ can be factored as

$$f(x^n; \boldsymbol{\theta}, k) \;=\; f(x^n \mid \hat{\boldsymbol{\theta}}) g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}), \tag{2.25}$$

where $f(x^n \mid \hat{\boldsymbol{\theta}})$ is the conditional density of x^n which does not depend on the unknown parameter vector $\boldsymbol{\theta}$ and $g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta})$ is the marginal density of $\hat{\boldsymbol{\theta}}$. It follows from (2.24) and (2.25) that

$$Q_{d/n}(i) \;=\; \int_{\hat{\boldsymbol{\theta}} \in B_{d/n}(\boldsymbol{\theta}^i)} \left[\int_{x^n : \hat{\boldsymbol{\theta}}(x^n) = \hat{\boldsymbol{\theta}}} f(x^n \mid \hat{\boldsymbol{\theta}}) \mathrm{d}x^n \right] g(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}) \mathrm{d}\hat{\boldsymbol{\theta}} \tag{2.26}$$

$$\;=\; \int_{\hat{\boldsymbol{\theta}} \in B_{d/n}(\boldsymbol{\theta}^i)} g(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\theta}}) \mathrm{d}\hat{\boldsymbol{\theta}}, \tag{2.27}$$

where the last identity is a consequence of the fact that the inner integral in (2.26) equals one. From [48, Section 6.2], we have

$$\lim_{n\to\infty} Q_{d/n}(i) \to \left(\frac{2d}{\pi k}\right)^{k/2}. \tag{2.28}$$

More importantly, it is possible to define a discrete prior for the rectangles,

$$W(\boldsymbol{\theta}^i) = \frac{Q_{d/n}(\boldsymbol{\theta}_i)}{C_{n,k}}, \quad i = 1, \ldots, N_n, \tag{2.29}$$

where $C_{n,k}$ is defined in (2.17). This makes the code length to be $L_d(\boldsymbol{\theta}^i) = -\ln W(\boldsymbol{\theta}^i)$.

Based on these findings, Rissanen introduced two structure functions [48, 49]:

- For the first one, the amount of unexplained noise is taken to be the maximum code length for the data sequences x^n, having the property that $\hat{\boldsymbol{\theta}}(x^n) \in B_{d/n}(\boldsymbol{\theta}^i)$. This happens when the ML estimate falls in a corner of the rectangle. Applying Taylor's expansion to $-\ln f(x^n; \boldsymbol{\theta}^i, k)$ about the point $\hat{\boldsymbol{\theta}}(x^n)$ and truncating after the third term yields

$$-\ln f(x^n; \boldsymbol{\theta}^i, k) \approx -\ln f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) + d/2. \tag{2.30}$$

  In the equation above we have used the fact that the second term of the Taylor's series equals zero, while the third term equals $d/2$ due to (2.22).

  The approximation from (2.30) leads to the structure function

$$h^{(1)}_{x^n}(\alpha) = \min_d \{-\ln f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) + d/2 : L_d(\boldsymbol{\theta}^i) \leq \alpha\}. \tag{2.31}$$

  Relying on the minimum description length (MDL) principle, the value of the parameter d is chosen such that to minimize the two-part code length

$$-\ln f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) + \frac{d}{2} + L_d(\boldsymbol{\theta}^i). \tag{2.32}$$

  After some algebra it can be shown that asymptotically the optimum value of d is $\hat{d} = k$. Remark in (2.31) that $L_d(\boldsymbol{\theta}^i)$ represents the amount of learnable information in data.

- The second structure function is defined by taking the unexplained noise to be the average code length for the data x^n which satisfies $\hat{\boldsymbol{\theta}}(x^n) \in B_{d/n}(\boldsymbol{\theta}^i)$.

  Let $\bar{L}_i(d)$ denote

$$\frac{1}{|B_{d/n}(\boldsymbol{\theta}^i)|} \int_{\hat{\boldsymbol{\theta}}(x^n) \in B_{d/n}(\boldsymbol{\theta}^i)} \ln \frac{f(x^n; \hat{\boldsymbol{\theta}}(x^n), k)}{f(x^n; \boldsymbol{\theta}^i)} \mathrm{d}x^n.$$

Then we have

$$h^{(2)}_{x^n}(\alpha) = \min_d \{-\ln f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) + \bar{L}_i(d) : L_d(\boldsymbol{\theta}^i) \leq \alpha\}. \tag{2.33}$$

Asymptotically, we get $\hat{d} = 3k$, which makes the amount of learnable information to be given by

$$\ln C_{n,k} + \frac{k}{2} \ln \frac{\pi}{6},$$

while the amount of unexplained noise is

$$-\ln f(x^n; \hat{\boldsymbol{\theta}}(x^n), k) + \frac{k}{2}.$$

Remark that the difference between the two above-mentioned formulas comes from the fact that in the second, instead of calculating the structure function for the worst code length in the set $B_{d/n}(\boldsymbol{\theta}^i)$, we calculate it for the average code length over the same set.

Chapter 3

NML and its applications in signal modelling and detection

As already told in the previous chapter, the computation of the NML with formula (2.21) poses troubles because, in most of the cases, it is difficult to evaluate the integral term. Another drawback of (2.21) is its asymptotic nature which does not recommend it to be used where only a small number of measurements are available.

In this chapter, we focus on the calculation of the NML, as well as on its performance as an yardstick for model selection. First, we study the problem of variable selection in linear regression using the NML. A major challenge in this case is to cope with infinite parametric complexity.

We address also a topic which was seldom treated in the previous literature, namely the calculation of the code length for the model structure. Then, we investigate the relationship between NML and the generalized likelihood ratio test (GLRT). The last section of the chapter is devoted to a particular form of the NML, which is dubbed SNML (sequentially normalized maximum likelihood) [50, 54] or SNLS (sequentially normalized least squares) [51].

3.1 NML when the parametric complexity is not finite

To circumvent the difficulties related to the fact that the parametric complexity is not finite when parameter space is unbounded, Rissanen proposed the following solution [47, 48]: First restrict the parameter space to a subset $\Theta_0 \subset \Theta$ to achieve a well-defined universal model as a function of Θ_0 boundaries, and then treat the boundaries of Θ_0 as hyperparameters and compute a new universal model by another round of normalization. This approach was applied to variable selection in Gaussian linear regression. To fix the ideas, let the measurements $\mathbf{y} \in \mathbb{R}^{n\times 1}$ be modeled by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the regressor matrix having more rows than columns ($n > m$), $\boldsymbol{\beta} \in \mathbb{R}^{m\times 1}$ is the vector of unknown parameters, and $\boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \tau\mathbf{I})$.

Because in most of the practical applications, not all the parameters $\beta_1, \ldots, \beta_m$ are equally important in modelling $\mathbf{y}$, one wants to eliminate those that are deemed to be irrelevant. This reduces to choosing a subset of the regressor variables indexed by $\gamma \subseteq \{1, \ldots, m\}$.

Let $\boldsymbol{\beta}_\gamma \in \mathbb{R}^{k\times 1}$ be the vector of the unknown regression coefficients within the γ-subset. The matrix $\mathbf{X}_\gamma$ is given by the columns of $\mathbf{X}$ that correspond to the γ-subset. Similarly to (3.1), we have:

$$\mathbf{y} = \mathbf{X}_\gamma \boldsymbol{\beta}_\gamma + \boldsymbol{\epsilon}_\gamma, \tag{3.2}$$

where $\boldsymbol{\epsilon}_\gamma \sim \mathcal{N}_n(0, \tau_\gamma \mathbf{I})$ and the variance τ_γ is unknown. Under the hypothesis that $\mathbf{X}_\gamma$ has full-rank, the ML estimates are [59]: $\hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}) = \left(\mathbf{X}_\gamma^\top \mathbf{X}_\gamma\right)^{-1} \mathbf{X}_\gamma^\top \mathbf{y}$ and $\hat{\tau}_\gamma(\mathbf{y}) = \|\mathbf{y} - \mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma\|^2/n$. It is customary to select γ by using either the Akaike Information Criterion (AIC) [1]

$$\mathrm{AIC}(\mathbf{y}; \gamma) = \frac{n}{2} \ln \hat{\tau}_\gamma + k, \tag{3.3}$$

or the Bayesian Information Criterion (BIC) [58]:

$$\mathrm{BIC}(\mathbf{y}; \gamma) = \frac{n}{2} \ln \hat{\tau}_\gamma + \frac{k}{2} \ln n. \tag{3.4}$$

We denote the cardinality of γ by k, and we make the assumption that $k > 0$. Remark that BIC coincides with a crude form of the SC, which was originally introduced in [43]. The main difficulty in computing SC comes from the fact that the parametric complexity is not finite when taking the data space to be

$$\{\mathbf{y} : (\hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}), \hat{\tau}_\gamma(\mathbf{y})) \in \Theta\},$$

where $\Theta = \{(\hat{\boldsymbol{\beta}}_\gamma, \hat{\tau}_\gamma) : \hat{\boldsymbol{\beta}}_\gamma \in \mathbb{R}^k, \hat{\tau}_\gamma > 0\}$. To solve the problem, Rissanen considered the data space such that

$$\{\mathbf{y} : (\hat{\boldsymbol{\beta}}_\gamma(\mathbf{y}), \hat{\tau}_\gamma(\mathbf{y})) \in \Theta_0\},$$

where $\Theta_0 = \{(\hat{\boldsymbol{\beta}}_\gamma, \hat{\tau}_\gamma) : \hat{\boldsymbol{\beta}}_\gamma^\top \boldsymbol{\Sigma}_\gamma \hat{\boldsymbol{\beta}}_\gamma < R, \hat{\tau}_\gamma > \tau_0\} \subset \Theta$, $\boldsymbol{\Sigma} = \mathbf{X}_\gamma^\top \mathbf{X}_\gamma/n$, and the strictly positive constants R and τ_0 are chosen arbitrarily.

After two rounds of normalization and by taking the negative logarithm of the resulting NML, the following SC-formula is obtained:

$$\mathrm{SC}_\gamma(\mathbf{y}) = \frac{n-k}{2} \ln \hat{\tau}_\gamma + \frac{k}{2} \ln \frac{||\mathbf{X}_\gamma \hat{\boldsymbol{\beta}}_\gamma||^2}{n} - \ln \Gamma\left(\frac{n-k}{2}\right) - \ln \Gamma\left(\frac{k}{2}\right) + \frac{n}{2} \ln(n\pi).$$

The result was extended in [37] by utilizing two other constraints. In [**P1**] we generalized Rissanen's methodology for the case when the data constraint is given by $\{\mathbf{y} : \rho(\hat{\boldsymbol{\beta}}) \leq R, \hat{\tau} \geq \tau_0\}$, where R and τ_0 are strictly positive, and the mapping $\rho : \mathbb{R}^k \to \mathbb{R}$ is chosen so that the set $\mathcal{B}_\rho(R) = \{\hat{\boldsymbol{\beta}} : \rho(\hat{\boldsymbol{\beta}}) \leq R\}$ is convex and its volume $V_\rho(R) = \int_{\mathcal{B}_\rho(R)} \mathrm{d}\hat{\boldsymbol{\beta}}$ has the expression

$$V_\rho(R) \;=\; \eta R^{\zeta k}, \tag{3.5}$$

where the constants η and ζ are strictly positive. In some cases, η might depend on $\mathbf{X}_\gamma$. Note that ζ depends on the shape of $\mathcal{B}_\rho(R)$, but is independent of the number of parameters.

It was shown in [**P1**] that the Rissanen constraint as well as the two constraints from [37] can be interpreted as particular cases of the general framework. We have also compared in [**P1**] the performance of the NML (with three different constraints), AIC, BIC, Conditional Model Estimator (CME) [26] and two variants of the Minimum Massage Length (MML) [56] by running experiments with simulated and real-life data sets. The summary of this comparison can be seen in Table 1 and Table 2 from [**P1**].

3.2 Description length for the γ-structure

The complete SC-formula should also include the description length, or equivalently, the code length for the γ-structure, which is conventionally denoted by L_γ. This term has a marginal effect in most of the applications, but not in all of them (see [53] for a more elaborated discussion). Next we derive the expression of L_γ by following the main lines from [48]. To this end, we have to consider some possible scenarios for transmitting from an hypothetized encoder to a decoder the entries of the γ-set. This amounts to inform the decoder which entries of the vector $\boldsymbol{\beta}$ are nonzero. For the sake of simplicity, we assume that the value of m is apriori known by the decoder. Additionally, $\gamma \neq \emptyset$, which implies that $k \in \{1, \ldots, m\}$.

Scenario $\mathcal{A}$ Let $j \in \{1, \ldots, m\}$. We send a zero to the decoder if the j-th entry of $\boldsymbol{\beta}$ equals zero. Otherwise, we send a one. Hence, if we want to let the decoder know which are the non-zero entries of $\boldsymbol{\beta}$, we have to transmit a binary string whose length is m. Because $k > 0$, the string cannot contain only zeros. Therefore, the current string is one out of $2^m - 1$ possible strings. By assuming that all strings are equally probable, we get the code length:

$$L_\gamma^{\mathcal{A}} = -\ln \frac{1}{2^m - 1} = \ln(2^m - 1). \tag{3.6}$$

Scenario $\mathcal{B}$ First we need to transmit to the decoder the value of k. Because it is already known at the decoder site that $k \in \{1, \ldots, m\}$, we resort to an encoding

method which has been introduced in [48, Section 2.1]. More precisely, we define $w(k) = k^{-1}/\Xi$, where $\Xi = \sum_{j=1}^{m} j^{-1}$. The code length for k is given by $-\ln w(k)$, but the result can be approximated by using the following upper bound [41, Section 3.1]: $\Xi < 1 + \ln m$. Like in [48, Section 2.1], the expression of the code length for k is:

$$L_k^{\mathcal{B}} = \ln k + \ln(1 + \ln m). \tag{3.7}$$

It remains to inform the decoder which are the indexes of the k non-zero entries. The code length for transmitting this information is $\ln \binom{m}{k}$ [53]. By considering the result from (3.7), the code length for γ has the expression:

$$L_\gamma^{\mathcal{B}} = \ln \binom{m}{k} + \ln k + \ln(1 + \ln m). \tag{3.8}$$

The strategy of the encoder is to select either scenario $\mathcal{A}$ or scenario $\mathcal{B}$ such that to minimize the code length. By employing (3.6) and (3.8), we readily obtain

$$L_\gamma = \min \left\{ \ln(2^m - 1), \left[\ln \binom{m}{k} + \ln k + \ln(1 + \ln m) \right] \right\}. \tag{3.9}$$

Note that the part of the code which tells to the decoder the name of the current scenario can be ignored since it only adds a constant to all code lengths.

Remark 1 Under the hypothesis that, in Scenario $\mathcal{B}$, all possible values of k are equally probable, the code length for k is $\ln m$. If, addittionally, Scenario $\mathcal{A}$ is never applied by the encoder, then the expression of the description length for γ becomes $\ln \binom{m}{k} + \ln m$. Obviously, this is equivalent, up to a constant term, to the formula $L_\gamma = \ln \binom{m}{k}$ that was introduced in [53].

Remark 2 To reduce the computational burden in (3.9), we notice for $k < m$ that

$$\ln \binom{m}{k} = \ln \frac{m\Gamma(m)}{[k\Gamma(k)]\,[(m-k)\Gamma(m-k)]}.$$

Then we apply the Stirling approximation [2, 48, 53],

$$\ln \Gamma(r) \;=\; \left(r - \frac{1}{2}\right) \ln r - r + \frac{1}{2} \ln(2\pi),$$

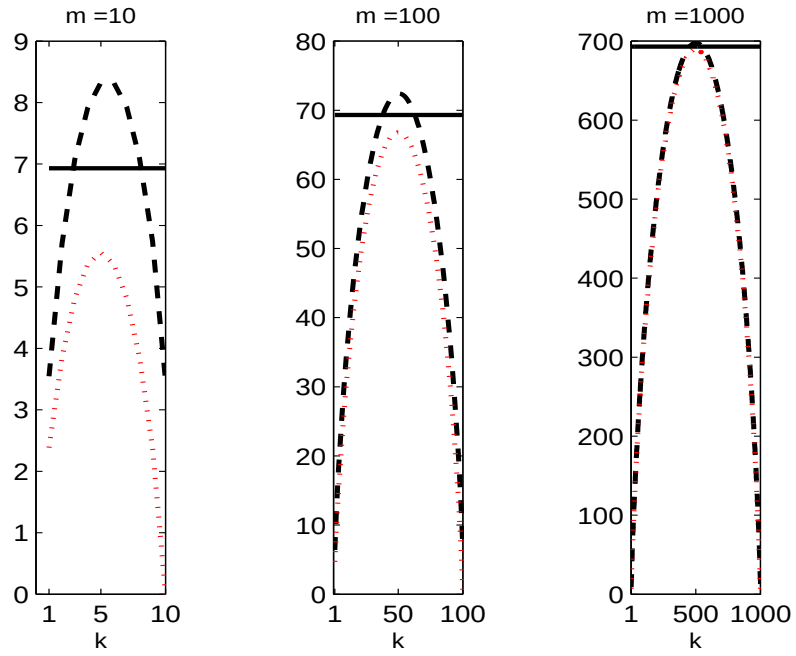

Figure 3.1: Comparison between $L_\gamma^{\mathcal{A}}$ (solid line) and $L_\gamma^{\mathcal{B}}$ (dashed line) for various values of k and m. The red dotted line represents the term $\ln \binom{m}{k}$ computed with the formula from (3.10).

and after some algebra we get

$$\ln \binom{m}{k} \approx (m+\frac{1}{2})\ln m - (k+\frac{1}{2})\ln k - (m-k+\frac{1}{2})\ln(m-k) - \frac{\ln(2\pi)}{2}. \quad (3.10)$$

Remark 3 To gain more insight on how the code length for Scenario $\mathcal{A}$ compares with the one for Scenario $\mathcal{B}$, we plot in Fig. 3.1 $L_\gamma^{\mathcal{A}}$ and $L_\gamma^{\mathcal{B}}$ versus k when $k \in \{1,\ldots,m\}$ and $m \in \{10, 100, 1000\}$. Note that $L_\gamma^{\mathcal{B}}$ is computed by applying the approximation from (3.10).

3.3 Selecting between two nested models

In this section, we address the problem of selection between two nested models, $\mathcal{M}_p$ and $\mathcal{M}_q$ by using the measurements $\mathbf{y}$. With the notations from (2.7) and (2.8), we have $\mathcal{M} = \{\mathcal{M}_p, \mathcal{M}_q\}$ and conventionally we take $p < q$. One possible approach is

to decide between the two models by applying the GLRT [28]:

$$2\ln\frac{f(\mathbf{y};\hat{\boldsymbol{\theta}}_1,\mathcal{M}_p)}{f(\mathbf{y};\hat{\boldsymbol{\theta}}_0,\mathcal{M}_q)}\underset{\mathcal{M}_q}{\overset{\mathcal{M}_p}{\lessgtr}}\eta, \tag{3.11}$$

where the two-way inequality denotes that we select the model $\mathcal{M}_p$ if the term in the left-hand side is smaller than η. Otherwise, we select the model $\mathcal{M}_q$. For $k\in\{p,q\}$, $f(\mathbf{y};\hat{\boldsymbol{\theta}}_k,\mathcal{M}_k)$ is the likelihood function when the model is $\mathcal{M}_k$, and $\hat{\boldsymbol{\theta}}_k$ denotes the ML estimate. The threshold η determines the level of performance, which is expressed in terms of probability of detection (P_{D}) and probability of false alarm (P_{FA}) with a terminology borrowed from detection theory.

Another possibility is to choose the model by solving the minimization problem [69]:

$$\min_{k\in\{p,q\}}[\mathrm{GIC}(k)=-2\ln f(\mathbf{y};\hat{\boldsymbol{\theta}}_k,\mathcal{M}_k)+\zeta k], \tag{3.12}$$

where GIC is a generalized information criterion. A simple comparison of (3.3) and (3.12) shows that GIC is equivalent with AIC when $\zeta=2$. Similarly, GIC reduces to BIC for $\zeta=\ln n$ (see (3.4)). From (3.11) and (3.12), we can notice the connection between GLRT and GIC, which was carefully investigated in [62] for the particular case when GIC coincides with AIC. More recently, the equivalence between GLRT and GIC was analyzed by assuming that the number of competing models is larger than two (see [69] and references therein).

In the previous literature, the relationship between GLRT and SC has been studied for deciding whether a Poisson or a Geometric model better fits the available measurements [11,34]. Apparently, the only published attempt to relate GLRT and SC in the context of Gaussian linear regression is the paper [**P2**]. More precisely, we have investigated the selection between model $\mathcal{M}_0$ which corresponds to $\gamma=\emptyset$ in (3.2) and the model $\mathcal{M}_m$ which is the equivalent of $\gamma=\{1,\ldots,m\}$ in the same equation.

One of the most interesting results from [**P2**, Proposition 3.1] is that SC is equivalent with GLRT only if the F-statistic $\frac{\|\mathbf{P}_{\mathbf{X}}\mathbf{y}\|^2/m}{\|\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{y}\|^2/(n-m)}$ is larger than one. Note that $\mathbf{X}$ and $\mathbf{y}$ are like in (3.1), $\mathbf{P}_{\mathbf{X}}$ denotes the orthogonal projection onto the column space of $\mathbf{X}$, and $\mathbf{P}_{\mathbf{X}}^{\perp}=\mathbf{I}-\mathbf{P}_{\mathbf{X}}$.

In the same framework of linear regression, we have compared SC and BIC when selecting between two nested models [**P2**]. In our study, we have assumed that the matrix $\mathbf{X}$ in (3.1) is deterministic and its number of columns (m) is finite. In another research on connection between various model selection criteria, Hansen and Yu have proven that an information theoretic criterion which is akin to SC combines the strength of AIC and BIC [23]. The main difference between our result and the one from [23] is that Hansen and Yu assumed the entries of $\mathbf{X}$ to be random and $m\to\infty$.

3.4 Sequentially Normalized Universal Models

As we have already pointed out, the calculation of the normalizing coefficient in (2.17) is a non-trivial task for most commonly used models. An alternative solution is the SNML [50,51,54]. We prefer to use the notation $f(y^n;\boldsymbol{\theta})$ instead of $f(y^n;\boldsymbol{\theta},k)$ for writing the equations more compactly. For each $t \in \{1,\ldots,n\}$, let $y^t = y_1,\ldots,y_t$ denote the observations up to time moment t. Additionally, $\hat{\boldsymbol{\theta}}(y^t)$ denotes the ML estimate.

With the convention that m' is such that $\hat{\boldsymbol{\theta}}(y^t)$ can be computed for all $t \in \{m'+1,\ldots,n\}$, the expression of SNML is given by

$$f_{\mathrm{SNML}} = f^{m'}(y^{m'}) \prod_{t=m'+1}^{n} \hat{f}(y_t|y^{t-1}), \tag{3.13}$$

$$\hat{f}(y_t|y^{t-1}) = \frac{f(y^t;\hat{\boldsymbol{\theta}}(y^t))}{K_t(y^{t-1})}, \tag{3.14}$$

$$K_t(y^{t-1}) = \int f(y^{t-1},y;\hat{\boldsymbol{\theta}}(y^{t-1},y))\mathrm{d}y, \tag{3.15}$$

where $f^{m'}(\cdot)$ is a suitably chosen initial distribution. We refer to [**P3**, Section 2] for a discussion on how m' can be selected.

The simplification in computing the complexity term is evident from (3.15), where the integration is over one data point only. For the linear regression problem in (3.1), the integral can be expressed in closed form [51]. To give the formula of the universal model which in this case is dubbed SNLS, we need some supplemetary notations. Let $\mathbf{x}_t$ be the column vector obtained by transposing the t-th row of $\mathbf{X}$. For $t \geq m'+1$ we define:

$$\begin{aligned}
\mathbf{X}_t &= [\mathbf{x}_t,\ldots,\mathbf{x}_1],\\
\mathbf{y}_t &= [y_t,\ldots,y_1]^\top,\\
\mathbf{V}_t &= (\mathbf{X}_t\mathbf{X}_t^\top)^{-1},\\
\hat{\boldsymbol{\beta}}_t &= \mathbf{V}_t\mathbf{X}_t\mathbf{y}_t, && \text{(ML estimate which belongs to } \mathbb{R}^{k\times 1})\\
\hat{e}_t &= y_t - \hat{\boldsymbol{\beta}}_t^\top\mathbf{x}_t, && \text{(forward a posteriori prediction error)}\\
c_t &= \mathbf{x}_t^\top\mathbf{V}_{t-1}\mathbf{x}_t.
\end{aligned}$$

Then according to [51, Eq. (15)], we have

$$\begin{aligned} \mathrm{SNLS}(k) &= \frac{n-m'}{2}\ln\Big(\frac{2\pi e}{n-m'}\sum_{t=m'+1}^{n}\hat{e}_t^2\Big) \\ &+ \sum_{t=m'+1}^{n}\ln(1+c_t)+\frac{1}{2}\ln n+O(1). \end{aligned} \tag{3.16}$$

For the selection of m', we refer to [51, Section 1]. More interestingly, the formula above can be applied not only when the matrix $\mathbf{X}$ is fixed as in (3.1), but also when it is random. This makes it possible to employ the SNLS criterion for choosing the order of autoregressions, as it was clearly shown in [51].

The results have been extended further in [**P3**] by modifying the criterion so that it can be applied when the coefficients of the AR models are estimated by forgetting factor least-squares (LS) algorithms [25]. With the convention that the forgetting factor λ is positive and less than one, the following criterion was introduced in [**P3**]:

$$\begin{aligned} \mathrm{SNML}_\lambda(k) &= \frac{n_{\mathrm{ef}}}{2}\ln\Big(\frac{1}{n_{\mathrm{ef}}}\sum_{t=m'+1}^{n}\lambda^{n-t}\hat{e}_{\lambda,t}^2\Big) \\ &+ \sum_{t=m'+1}^{n}\ln[(1+c_{\lambda,t})\lambda^k]+\frac{1}{2}\ln n_{\mathrm{ef}} \end{aligned} \tag{3.17}$$

where

$$\begin{aligned} n_{\mathrm{ef}} &= \sum_{t=0}^{n-1}\lambda^t, && \text{(effective number of samples)} \\ \mathbf{V}_{\lambda,t} &= (\sum_{i=1}^{t}\lambda^{t-i}\mathbf{x}_i\mathbf{x}_i^\top)^{-1}, \\ \hat{\boldsymbol{\beta}}_{\lambda,t} &= \mathbf{V}_{\lambda,t}\sum_{i=1}^{t}\lambda^{t-i}\mathbf{x}_i y_i, && \text{(weighted LS estimate which belongs to } \mathbb{R}^{k\times 1}) \\ \hat{e}_{\lambda,t} &= y_t-\hat{\boldsymbol{\beta}}_{\lambda,t}^\top\mathbf{x}_t, \\ c_{\lambda,t} &= \lambda^{-1}\mathbf{x}_t^\top\mathbf{V}_{\lambda,t-1}\mathbf{x}_t. \end{aligned}$$

The performance of SNML_λ was compared with that of other five criteria and the conclusions are outlined in [**P3**, Section 6].

Chapter 4

Application of optimally distinguishable distributions to the detection of subspace signals in Gaussian noise

In Section 2.2.5 the concept of optimally distinguishable distributions (ODD) was already mentioned in connection with the partition of the parameter space which allows to define KSF.

In this chapter, we show how the ODD methodology from [48] can be applied to solve the following detection problem [28, 55]. Let the vector of measurements be $\mathbf{x} = [x_0, \ldots, x_{N-1}]^\top$, where $x_0, \ldots, x_{N-1} \in \mathbb{R}$ are samples from a time series. Based on these data, one selects between the hypotheses specified by the model classes:

$$\begin{cases} \mathcal{M}_0 = \{f(\mathbf{x}; \boldsymbol{\theta}, \tau) : \boldsymbol{\theta} = \mathbf{0}\}, \\ \mathcal{M}_1 = \{f(\mathbf{x}; \boldsymbol{\theta}, \tau) : \boldsymbol{\theta} \neq \mathbf{0}\}. \end{cases} \tag{4.1}$$

We adopt the convention that $\mathbf{0}$ is a null vector of appropriate dimension. The Gaussian density function $f(\mathbf{x}; \boldsymbol{\theta}, \tau)$ is given by

$$f(\mathbf{x}; \boldsymbol{\theta}, \tau) = \frac{1}{(2\pi\tau)^{N/2}} \exp\left(-\frac{1}{2\tau}\|\mathbf{x} - \mathbf{H}\boldsymbol{\theta}\|^2\right), \tag{4.2}$$

where $\mathbf{H} \in \mathbb{R}^{N \times k}$ is a *known* matrix of rank k ($N > k + 1$), and $\boldsymbol{\theta} \in \mathbb{R}^{k \times 1}$ is the vector of linear parameters which are *unknown*.

Remark that, in this chapter, we use some notations which are slightly different by those from Chapter 2 and Chapter 3. For instance, to be in line with the previous notations, we should employ $\mathcal{M}_k$ instead of $\mathcal{M}_1$ in (4.1). We operate these small modifications for the sake of simplicity.

In [**P4**], it is shown how the detection problem in (4.1) can be solved by optimal distiguishability. However, the results from [**P4**] are restricted to the case when the noise variance τ is *known*. Here, as a preparatory step, we first outline briefly the results from [**P4**], and then move to the case of unknown variance. The treatment of this case is based on the solution which we have proposed in [**P5**, Theorem 3.1].

4.1 ODD detector when the noise variance is known

Partition of parameter space: In the case of the model class $\mathcal{M}_1$, the FIM in (2.13) takes the particular form [27]:

$$\mathbf{J}_N(\boldsymbol{\theta}) = \frac{1}{N\tau}\mathbf{H}^\top\mathbf{H}. \tag{4.3}$$

Obviously, FIM does not depend on the values of the linear parameters, which allows us to use the notation $\mathbf{J}_N$ instead of $\mathbf{J}_N(\boldsymbol{\theta})$. Consider the hyper-ellipsoid centered at $\boldsymbol{\theta}^0 = \mathbf{0}$ and defined by $(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^\top\mathbf{J}_N(\boldsymbol{\theta}-\boldsymbol{\theta}^0) = d/N$, where d is a parameter whose optimal value we will find next. Let $B^\tau_{d/N}(0)$ be the largest rectangle within the hyper-ellipsoid. This rectangle lies parallel to the eigenvectors of $\mathbf{J}_N$, and its side lengths are $2\mu_1, \ldots, 2\mu_k$. For $i \in \{1, \ldots, k\}$, we have $\mu_i = \left(\frac{d}{Nk\lambda_i}\right)^{1/2}$, where λ_i is the i-th eigenvalue of the matrix $\mathbf{J}_N$ [48]. We assume that the parameter space Θ is a bounded closed subset of $\mathbb{R}^k$. Furthermore, Θ is partitioned by filling it up with adjacent copies of $B^\tau_{d/N}(0)$. The construction is done in such a way that, for any two adjacent rectangles, the straight line connecting their centers is parallel to one of the eigenvectors of $\mathbf{J}_N$.

Optimum value of parameter d: Let $\mathfrak{N}_{d/N}$ be the number of rectangles within the partition of Θ. We denote by $\boldsymbol{\theta}^j$ the center of the j-th rectangle $B^\tau_{d/N}(j)$. For all $j \in \{0, \ldots, \mathfrak{N}_{d/N} - 1\}$, the probability density $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j)$ is defined by [48]

$$\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j) \;=\; \begin{cases} f(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{x}))/Q^\tau_{d/N}(j), & \hat{\boldsymbol{\theta}}(\mathbf{x}) \in B^\tau_{d/N}(j) \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

where $Q^\tau_{d/N}(j) = \displaystyle\int_{\mathbf{x}:\hat{\boldsymbol{\theta}}(\mathbf{x})\in B^\tau_{d/N}(j)} f(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{x}))\mathrm{d}\mathbf{x}$ and

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{x} \tag{4.5}$$

are the ML estimates of the linear parameters.

The theoretical results from [48, 49] lead to the conclusion that $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j)$, $j \in \{0, \ldots, \mathfrak{N}_{d/N} - 1\}$, are *optimally distinguishable distributions*. The parameter d is selected so as to minimize the KL divergence $D(\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j) \parallel f(\mathbf{x}; \boldsymbol{\theta}^j))$

between the "artificial" model $\hat{f}(\mathbf{x} \mid \boldsymbol{\theta}^j)$ and the "natural" model $f(\mathbf{x}; \boldsymbol{\theta}^j)$ for all $j \in \{0, \ldots, \mathfrak{N}_{d/N} - 1\}$. In [**P4**], it is proved that the optimum value of d is $\hat{d} = 3k$, which is in perfect agreement with the findings from [48, 49].

Detection strategy: The ODD criterion selects the model class $\mathcal{M}_0$ whenever $\hat{\boldsymbol{\theta}}(\mathbf{x}) \in B^{\tau}_{\hat{d}/N}(0)$ [48]. This is equivalent to choose $\mathcal{M}_0$ if

$$\max\left(|z_1|, \ldots, |z_k|\right) < \sqrt{3}, \tag{4.6}$$

where

$$z_j = \frac{(\mathbf{v}_j^{\top}\mathbf{H}^{\top}\mathbf{x})/\sqrt{\ell_j}}{\sqrt{\tau}} \quad \forall j \in \{1, \ldots, k\}, \tag{4.7}$$

with the convention that $\ell_1, \ldots, \ell_k$ are the eigenvalues of the matrix $\mathbf{H}^{\top}\mathbf{H}$, and $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are the corresponding eigenvectors. Remark that we can write

$$\mathbf{H}^{\top}\mathbf{H} = \mathbf{V}\mathbf{D}\mathbf{V}^{\top}, \tag{4.8}$$

with

$$\begin{aligned} \mathbf{V} &= [\mathbf{v}_1, \ldots, \mathbf{v}_k], \\ \mathbf{D} &= \begin{bmatrix} \ell_1 & & \\ & \ddots & \\ & & \ell_k \end{bmatrix}. \end{aligned} \tag{4.9}$$

Confidence indices: The outcome of the ODD method is assessed by two indices conventionally denoted $E1$ and $E2$. According to the definition from [49], we have $E1 = 1 - P_{0|0}$. Notice that, for an arbitrary pair (i, j), $P_{i|j}$ is the probability mass function of $B^{\tau}_{\hat{d}/N}(i)$ induced by the model $f(\mathbf{x}; \boldsymbol{\theta}^j)$. Additionally, for $j \neq 0$, $E2$ is defined as $E2 = P_{0|j}$. To be in line with the definitions above, we have computed in [P4,P5] the indices $E1$ and $E2$.

Even if there are some problems with the interpretation of $E1$ and $E2$, the ODD partition is an important concept, and the decision rule implied by ODD makes perfect sense. This is why in both [P4] and [P5], we assessed the outcome of the ODD method by computing the probability of detection (P_{D}) and the probability of false alarm (P_{FA}). For the sake of clarity, we mention that, with the notations from (4.1), P_{D} is the probability to choose $\mathcal{M}_1$ when $\mathcal{M}_1$ is true, whereas P_{FA} is the probability to choose $\mathcal{M}_1$ when $\mathcal{M}_0$ is true [28]. This terminology is the one used in the engineering literature, and we refer to [28, Table 3.1] for the equivalent terms from the statistical literature. Apparently, the only published results on the P_{D} and the P_{FA} for the ODD-based detector are those from [P4,P5]. It is interesting to remark that $P_{\mathrm{FA}} = E1$, even if $E1$ was introduced by Rissanen from a different

interpretation. We also note that ODD method does not seek to maximize the P_{D} for a given P_{FA} like in the Neyman-Pearson methodology [28].

The rest of the chapter is focused on solving the hypothesis testing problem (4.1) by using the ODD method when the variance τ is unknown. The problem was studied in [**P5**], where we did not give the complete proofs because of the limited typographic space.

4.2 ODD detector when the noise variance is unknown

We treat τ as a nuisance parameter, and we assume $0 < \tau_1 < \tau < \tau_2$, where τ_1 and τ_2 are arbitrary. It will become clear from the results outlined below that τ_1 and τ_2 do not have any influence on the outcome of ODD detector.

As already told in Section 4.1, the first step in the ODD methodology is to use the FIM that corresponds to the model class $\mathcal{M}_1$ for defining a partition of the parameter space. In the case of the detection problem which we discuss, FIM has the well-known expression [27]:

$$\mathbf{J}_N(\boldsymbol{\psi}) = \begin{bmatrix} (\mathbf{H}^\top\mathbf{H})/(N\tau) & \mathbf{0} \\ \mathbf{0} & 1/(2\tau^2) \end{bmatrix}, \tag{4.10}$$

where $\boldsymbol{\psi} = [\boldsymbol{\theta}^\top \tau]^\top$. By comparing (4.3) and (4.10), we remark that $\mathbf{J}_N(\boldsymbol{\theta})$ does not depend on the parameters $\boldsymbol{\theta}$, whereas $\mathbf{J}_N(\boldsymbol{\psi})$ depends on $\boldsymbol{\psi}$. More precisely, $\mathbf{J}_N(\boldsymbol{\psi})$ depends explicitly on the noise variance τ. This makes the construction of the partition to be much more difficult than the one from [**P4**].

To circumvent the difficulties, we have proposed the following solution [**P5**]. With the notations from Section 4.1, $B^{\tau}_{d/N}(0)$ is the rectangle corresponding to the null hypothesis when the value of τ is fixed. Because in the case of interest for us the noise variance τ is a nuisance parameter, we assign to the model class $\mathcal{M}_0$ the region $B_{d/N}(0)$ of the parameter space which is given by the union of all $B^{\tau}_{d/N}(0)$ with $\tau \in (\tau_1, \tau_2)$. More formally,

$$B_{d/N}(0) = \bigcup_{\tau_1<\tau<\tau_2} B^{\tau}_{d/N}(0) = \left\{(\boldsymbol{\theta}, \tau) : \boldsymbol{\theta} \in B^{\tau}_{d/N}(0), \tau \in (\tau_1, \tau_2)\right\}.$$

The definition can be extended for all $j \neq 0$ by taking $B_{d/N}(j) = \bigcup_{\tau_1<\tau<\tau_2} B^{\tau}_{d/N}(j)$, where the significance of $B^{\tau}_{d/N}(j)$ is the same as in Section 4.1. The center of the rectangle $B^{\tau}_{d/N}(j)$ is $\boldsymbol{\theta}^j(\tau) = \sqrt{\tau}\mathbf{V}\mathbf{D}^{-1/2}(2\sqrt{d/k}\mathbf{m}^j)$, where all entries of the vector $\mathbf{m}^j = [m^j_1, \ldots, m^j_k]^\top$ are integers. We emphasize that $\boldsymbol{\theta}^j(\tau)/\sqrt{\tau}$ does not depend on the value of τ. To enhance intuition, we show in Fig. 4.1 a graphical representation of $B_{d/N}(0)$ together with the equivalence classes which are located in its vicinity.

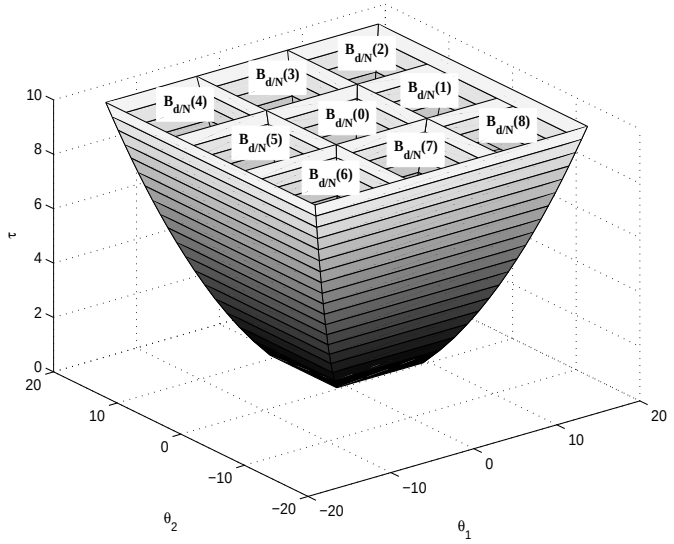

Figure 4.1: Partition of the parameter space: equivalence classes that correspond to $B_{d/N}(0)$ and its neighbors. Note that $\boldsymbol{\theta}^0 = \mathbf{0}$ and the number of linear parameters is $k = 2$. The matrix $\mathbf{H}$ is assumed to have orthonormal colums, and the parameter d equals $3k$. The noise variance bounds are conventionally taken to be $\tau_1 = 1$ and $\tau_2 = 10$. For $\tau \in [\tau_1, \tau_2]$ and $i \in \{0, 1, \ldots, 8\}$, $B_{d/N}(i)$ is a square with side length $2\sqrt{3\tau}$.

By analogy with (4.4), we define the PDF:

$$\hat{f}_j(\mathbf{x}) \;=\; \begin{cases} f(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{x}), \hat{\tau}(\mathbf{x}))/Q_{d/N}(j), & \hat{\boldsymbol{\theta}}(\mathbf{x}) \in B^{\hat{\tau}}_{d/N}(j), \hat{\tau}(\mathbf{x}) \in (\tau_1, \tau_2) \\ 0, & \text{otherwise} \end{cases}$$

where $j \in \{0, \ldots, \mathfrak{N}_{d/N} - 1\}$ and

$$Q_{d/N}(j) = \int_{\mathbf{x}: \hat{\boldsymbol{\theta}}(\mathbf{x}) \in B^{\hat{\tau}}_{d/N}(j), \hat{\tau}(\mathbf{x}) \in (\tau_1, \tau_2)} f(\mathbf{x}; \hat{\boldsymbol{\theta}}(\mathbf{x}), \hat{\tau}(\mathbf{x})) \mathrm{d}\mathbf{x}. \tag{4.11}$$

The formula for $\hat{\boldsymbol{\theta}}(\mathbf{x})$ is given in (4.5) and

$$\hat{\tau}(\mathbf{x}) = \frac{1}{N} \|\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}(\mathbf{x})\|^2 \tag{4.12}$$

is the ML estimate of the variance when the model class is $\mathcal{M}_1$.

Next we need to choose d such as to minimize the KL divergence between the

"artificial" model $\hat{f}_0(\mathbf{x})$ and the "natural" model which corresponds to $\mathcal{M}_0$. Recall that, for known τ, $\mathcal{M}_0$ contains a single density function and it is straightforward to select it as the "natural" model. However, when τ is unknown, $\mathcal{M}_0$ contains an infinite number of density functions, and it is not so obvious how to decide which one is the "natural" model. As it was suggested in [46], one possibility is to take the "natural" model for $\mathcal{M}_0$ to be the NML defined by

$$\begin{aligned} \tilde{f}_0(\mathbf{x}) &= f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0(\mathbf{x}))/C_0, \\ C_0 &= \int_{\mathbf{x}:\hat{\tau}_0(\mathbf{x})\in(\tau_1,\tau_2)} f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0(\mathbf{x}))\mathrm{d}\mathbf{x}. \end{aligned} \tag{4.13}$$

We will prove in Appendix 4.4 that choosing the ML function $f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0(\mathbf{x}))$ as the "natural" model leads to the same optimum value for the parameter d as in the case when the "natural" model is the NML function. Note that, for $\mathcal{M}_0$ class, $\hat{\tau}_0(\mathbf{x}) = ||\mathbf{x}||^2/N$ is the ML estimate of the variance. Whenever it is clear from the context which measurements are used for estimation, the simpler notation $\hat{\tau}_0$ is preferred to $\hat{\tau}_0(\mathbf{x})$. The same applies for $\hat{\boldsymbol{\theta}}(\mathbf{x})$ and $\hat{\tau}(\mathbf{x})$ which are defined in (4.5) and (4.12), respectively.

After these preparations, we are ready to formulate the principal theorem.

Theorem 4.2.1. For the data sequence $\mathbf{x} = [x_1,\ldots,x_N]^\top$, consider the ODD testing between the hypotheses specified by the model classes

$$\begin{aligned} \mathcal{M}_0 &= \{f(\mathbf{x};\boldsymbol{\theta},\tau) : \boldsymbol{\theta} = \mathbf{0}, \tau_1 < \tau < \tau_2\}, \\ \mathcal{M}_1 &= \{f(\mathbf{x};\boldsymbol{\theta},\tau) : \boldsymbol{\theta} \neq \mathbf{0}, \tau_1 < \tau < \tau_2\}, \end{aligned}$$

where $f(\mathbf{x};\boldsymbol{\theta},\tau)$ is given in (4.2). We have the following results:
a) Optimum value of parameter d: Let $\boldsymbol{\theta}^0 = \mathbf{0}$. If $d \ll N$, then the KL divergence $D(\hat{f}_0(\mathbf{x}) \parallel \tilde{f}_0(\mathbf{x}))$ is minimized by $\hat{d} = 3k$.
b) Detection strategy: After observing $\mathbf{x}$, select $\mathcal{M}_0$ if

$$\max\left(|t_1|,\ldots,|t_k|\right) < \sqrt{3}, \tag{4.14}$$

where $t_j = \dfrac{(\mathbf{v}_j^\top \mathbf{H}^\top \mathbf{x})/\sqrt{\ell_j}}{\sqrt{\hat{\tau}}}$ $\forall j \in \{1,\ldots,k\}$. The role of $\ell_1,\ldots,\ell_k$, $\mathbf{v}_1,\ldots,\mathbf{v}_k$ is analogous to that in (4.7), and $\hat{\tau}$ is defined in (4.12).
c) Confidence indices: Let $\mathbf{t} = [t_1,\ldots,t_k]^\top$, and let $\varphi(\mathbf{t};\boldsymbol{\delta})$ be the PDF given by

$$\varphi(\mathbf{t};\boldsymbol{\delta}) = \frac{\exp\left(-||\boldsymbol{\delta}||^2/2\right)}{(\pi N)^{k/2}\Gamma\left((N-k)/2\right)} \sum_{\alpha=0}^{\infty} \frac{2^{\alpha/2}\left(\mathbf{t}^\top\boldsymbol{\delta}\right)^\alpha \Gamma\left((N+\alpha)/2\right)}{N^{\alpha/2}\ \alpha!\ \left(1+\|\mathbf{t}\|^2/N\right)^{(N+\alpha)/2}}, \tag{4.15}$$

with parameter vector $\boldsymbol{\delta} \in \mathbb{R}^{k\times 1}$.

If the condition in (4.14) is satisfied, then

$$E1 = 1 - \int_{(-\boldsymbol{\xi},\boldsymbol{\xi})} \varphi(\mathbf{t};\mathbf{0})\mathrm{d}\mathbf{t}. \tag{4.16}$$

Otherwise,

$$E2 \;\; = \;\; \int_{(-\boldsymbol{\xi},\boldsymbol{\xi})} \varphi(\mathbf{t};2\sqrt{3}\mathbf{m})\mathrm{d}\mathbf{t}, \tag{4.17}$$

where $\boldsymbol{\xi} = \sqrt{3}[1,\ldots,1]^\top$ and $(-\boldsymbol{\xi},\boldsymbol{\xi}) = \{\mathbf{t} : -\xi_j < t_j < \xi_j, 1 \leq j \leq k\}$ in which ξ_j is the j-th component of vector $\boldsymbol{\xi}$. The vector $\mathbf{m} = [m_1,\ldots,m_k]^\top$ is such that $m_j = \left\lfloor \frac{t_j+\sqrt{3}}{2\sqrt{3}} \right\rfloor \; \forall j \in \{1,\ldots,k\}$.

The proof is deferred to Appendix 4.4. □

In the same Appendix, the meaning of the parameter $\boldsymbol{\delta}$ from (4.15) is clarified. To gain more insight on the condition $d \ll N$, we refer to the numerical examples in Section 4.3.

Remark 1 It is straightforward to derive the expressions of P_{FA} and P_{D} for the decision rule (4.14). Obviously, we have

$$P_{\mathrm{FA}} = E1. \tag{4.18}$$

Under the hypothesis that $\mathbf{x} \sim \mathcal{N}_N(\mathbf{H}\underline{\boldsymbol{\theta}}, \underline{\tau})$, where $\underline{\boldsymbol{\theta}}$ has at least one entry which is nonzero and $\underline{\tau} > 0$, Result 4 from Appendix 4.4 implies

$$P_{\mathrm{D}} = 1 - \int_{(-\boldsymbol{\xi},\boldsymbol{\xi})} \varphi\left(\mathbf{t};\mathbf{D}^{1/2}\mathbf{V}^\top\underline{\boldsymbol{\theta}}/\sqrt{\underline{\tau}}\right)\mathrm{d}\mathbf{t}. \tag{4.19}$$

Remark 2 The condition in (4.14) can be obtained from (4.6) by using the ML estimate $\hat{\tau}$ instead of the unknown variance. A slightly different decision rule can be devised by replacing in (4.6) the unknown value of τ with the unbiased estimate $\hat{\nu} = \frac{N}{N-k}\hat{\tau}$. Based on this approach, the $\mathcal{M}_0$-class is selected if

$$\max\left(|\bar{t}_1|,\ldots,|\bar{t}_k|\right) < \sqrt{3}, \tag{4.20}$$

where $\bar{t}_j = \sqrt{\frac{N-k}{N}}t_j$ for all $j \in \{1,\ldots,k\}$. Note that the random vector $\bar{\mathbf{t}} = [\bar{t}_1,\ldots,\bar{t}_k]^\top$ is related to the random vector $\mathbf{t}$ by the linear transformation $\bar{\mathbf{t}} = c\mathbf{t}$, where

$$c = \sqrt{\frac{N-k}{N}}. \tag{4.21}$$

Therefore, it is easy to compute the confidence indices which are analogous to $E1$

and $E2$ from Theorem 4.2.1. When the condition in (4.20) is satisfied, we have

$$\bar{E}1 = 1 - \int_{(\boldsymbol{\xi},\boldsymbol{\xi})} \bar{\varphi}(\bar{\mathbf{t}};\mathbf{0})\mathrm{d}\bar{\mathbf{t}},$$

where

$$\begin{aligned}\bar{\varphi}(\bar{\mathbf{t}};\boldsymbol{\delta}) &= \frac{\exp(-||\boldsymbol{\delta}||^2/2)}{(\pi(N-k))^{k/2}\Gamma((N-k)/2)} \\ &\quad \times \sum_{\alpha=0}^{\infty} \frac{2^{\alpha/2}\left(\bar{\mathbf{t}}^\top\boldsymbol{\delta}\right)^\alpha \Gamma((N+\alpha)/2)}{(N-k)^{\alpha/2}\alpha!\,(1+\|\bar{\mathbf{t}}\|^2/(N-k))^{(N+\alpha)/2}}. \end{aligned} \tag{4.22}$$

We refer to Appendix 4.4 for more details on $\bar{\varphi}(\bar{\mathbf{t}};\boldsymbol{\delta})$. If (4.20) is not satisfied, then we get

$$\bar{E}2 = \int_{(-\boldsymbol{\xi},\boldsymbol{\xi})} \bar{\varphi}(\bar{\mathbf{t}};2\sqrt{3}\bar{\mathbf{m}})\mathrm{d}\bar{\mathbf{t}}, \tag{4.23}$$

where $\bar{\mathbf{m}} = [\bar{m}_1,\ldots,\bar{m}_k]^\top$ with the convention that $\bar{m}_j = \left\lfloor \frac{\bar{t}_j+\sqrt{3}}{2\sqrt{3}} \right\rfloor \ \forall j \in \{1,\ldots,k\}$. The closed-form expressions for P_{D} and P_{FA} can be also obtained without difficulties.

4.3 Numerical aspects

4.3.1 Calculation of the confidence indices

Because we want to apply a similar methodology for both $E1$ and $\bar{E}1$, let us observe that

$$E1 = 1 - \int_{(-c\boldsymbol{\xi},c\boldsymbol{\xi})} \bar{\varphi}(\bar{\mathbf{t}};\mathbf{0})\mathrm{d}\bar{\mathbf{t}}, \tag{4.24}$$

where c is given in (4.21). It can be also easily verified that $\bar{\varphi}(\bar{\mathbf{t}};\mathbf{0})$ coincides with the central multivariate t-distribution having $N-K$ degrees of freedom [31]. This makes the calculation of $E1$ and $\bar{E}1$ to become a simple exercise of using the Matlab function `mvtcdf`.

Similarly with (4.24), we can write

$$E2 = \int_{(-c\boldsymbol{\xi},c\boldsymbol{\xi})} \bar{\varphi}(\bar{\mathbf{t}};2\sqrt{3}\mathbf{m})\mathrm{d}\bar{\mathbf{t}}, \tag{4.25}$$

where $\mathbf{m}$ is defined in Theorem 4.2.1. However, the computation of $E2$ and $\bar{E}2$ is much more difficult than evaluating $E1$. For providing some numerical examples, we

consider the particular case $k = 2$. Then we apply the following formula from [65]:

$$\int_{(-\mathbf{a},\mathbf{a})} \bar{\varphi}\left(\bar{\mathbf{t}}; [\delta_1\ \delta_2]^\top\right) \mathrm{d}\bar{\mathbf{t}}$$

$$= 2\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{\left(\exp(-\delta_1^2/2)(\delta_1^2/2)^i/i!\right)\left(\exp(-\delta_2^2/2)(\delta_2^2/2)^j/j!\right)}{B(i+1/2, j+1/2)} \tag{4.26}$$

$$\times \int_0^{\pi/4} (\sin^{2i} v \cos^{2j} v + \sin^{2j} v \cos^{2i} v) I_{V_1}(i+j+1, N/2)\mathrm{d}v, \tag{4.27}$$

where $\mathbf{a} = [A\ A]^\top$ with $A > 0$, $B(\cdot,\cdot)$ is the complete beta function, $I_{V_1}(\cdot,\cdot)$ is the incomplete beta function ratio, and $V_1 = A^2 \sec^2 v/(N + A^2 \sec^2 v)$.

In our implementation, the double series from (4.26) is truncated by constraining i and j to take values from zero to one hundred. The integral from (4.27) is computed by using the `quadgk` Matlab function, which is based on an adaptive quadrature algorithm [60]. The results obtained for various sample sizes are shown in Table 4.1. In the same table, we reproduce from [**P4**] the values of the confidence indices for the case when τ is known. As it was already pointed out in [**P4**], if τ is known, then the confidence indices do not depend on the sample size N. On contrary, if τ is unknown, $E1$ and $E2$ as well as $\bar{E}1$ and $\bar{E}2$ depend on N, and for $N \gg 1$, they approach the values corresponding to known variance.

Observe in Table 4.1 that almost all the indices decrease when N increases. There exists one single exception, namely $E2$ when $(m_1, m_2) \in \mathcal{P}_1$. This erratic behavior can be explained by the fact that $\hat{\tau}$ is a biased estimate of τ, which makes the limits of the integration domain in (4.25) to vary with N. The integration domain expands from $(-0.949\boldsymbol{\xi}, 0.949\boldsymbol{\xi})$ to $(-0.999\boldsymbol{\xi}, 0.999\boldsymbol{\xi})$ when N increases from 20 to 1000. It is evident that the integration is done on the same domains also when evaluating $E2$ for $(m_1, m_2) \in \mathcal{P}_2$. However, the main difference between computing $E2$ for $(m_1, m_2) \in \mathcal{P}_1$ and $(m_1, m_2) \in \mathcal{P}_2$ is the following: $||2\sqrt{3}\mathbf{m}||^2 = 12$ if $(m_1, m_2) \in \mathcal{P}_1$, whereas $||2\sqrt{3}\mathbf{m}||^2 = 24$ if $(m_1, m_2) \in \mathcal{P}_2$. So, for the points within $\mathcal{P}_2$, the effect produced by the dependence of the integration domain on N is diminished because $2\sqrt{3}\mathbf{m}$ is located at larger distance from the null vector $\boldsymbol{\theta}^0$.

4.3.2 Calculation of P_{D} and P_{FA}

We focus on the case $k = 2$. From (4.19), we have:

$$P_D = 1 - \int_{(-c\boldsymbol{\xi}, c\boldsymbol{\xi})} \bar{\varphi}\left(\bar{\mathbf{t}}; \boldsymbol{\zeta}\right) \mathrm{d}\bar{\mathbf{t}},$$

where c is given in (4.21) and

$$\boldsymbol{\zeta} = \mathbf{D}^{1/2}\mathbf{V}^\top \underline{\boldsymbol{\theta}}/\sqrt{\underline{\tau}}. \tag{4.28}$$

N		20	50	100	500	1000	**Known variance**
E_1		0.2177	0.1819	0.1706	0.1618	0.1607	**0.1596**
E_2	$(m_1, m_2) \in \mathcal{P}_1$	0.0345	0.0367	0.0374	0.0380	0.0381	**0.0382**
	$(m_1, m_2) \in \mathcal{P}_2$	0.0020	0.0018	0.0018	0.0017	0.0017	**0.0017**
$\bar{E}_1$		0.1873	0.1702	0.1648	0.1606	0.1601	**0.1596**
$\bar{E}_2$	$(m_1, m_2) \in \mathcal{P}_1$	0.0426	0.0398	0.0390	0.0383	0.0382	**0.0382**
	$(m_1, m_2) \in \mathcal{P}_2$	0.0029	0.0021	0.0019	0.0018	0.0018	**0.0017**

Table 4.1: Confidence indices for the detection rules from (4.14) and (4.20) when the number of linear parameters is $k = 2$. The results for $E2$ and $\bar{E}2$ correspond to the equivalence classes $B_{\hat{d}/N}(i)$ which are located in the vicinity of $B_{\hat{d}/N}(0)$. With the notations from Fig. 4.1, we consider $B_{\hat{d}/N}(i)$ for which $i \in \{1, 3, 5, 7\}$. This is equivalent with selecting, in (4.23) and (4.25), the vectors $\bar{\mathbf{m}} = \mathbf{m} = [m_1\ m_2]^\top$ such that $(m_1, m_2) \in \mathcal{P}_1$, where $\mathcal{P}_1 = \{(-1, 0), (0, -1), (1, 0), (0, 1)\}$. Similarly, we choose $\bar{\mathbf{m}} = \mathbf{m}$ to have the entries $(m_1, m_2) \in \mathcal{P}_2$ with $\mathcal{P}_2 = \{(-1, -1), (1, -1), (-1, 1), (1, 1)\}$ for calculating $E2$ and $\bar{E}2$ when the equivalence classes are $B_{\hat{d}/N}(2)$, $B_{\hat{d}/N}(4)$, $B_{\hat{d}/N}(6)$ and $B_{\hat{d}/N}(8)$ (see Fig. 4.1).

Remark that $\|\boldsymbol{\zeta}\|^2 = ||\mathbf{H}\underline{\boldsymbol{\theta}}||^2/\underline{\tau}$ is the energy-to-noise ratio (ENR) [28]. A closer look to the formula in (4.26)-(4.27) reveals that $P_{\rm D}$ depends on the entries of $\boldsymbol{\zeta}$ and not only on $||\boldsymbol{\zeta}||^2$. A similar property has been demonstrated to hold true for the ODD detector when τ is known [**P4**]: $P_{\rm D}$ is maximized when either $\zeta_1^2 = \text{ENR}$ or $\zeta_2^2 = \text{ENR}$, and it is minimized when $\zeta_1^2 = \zeta_2^2 = \text{ENR}/2$. When τ is unknown, $P_{\rm D}$ depends also on the sample size N, as we can see from (4.27).

To illustrate the performance of the ODD detector introduced in Theorem 4.2.1, we take $N = 50$ and d equals its optimum value, namely $d = 6$. Then we compute $P_{\rm D}$ when ENR increases from 0 dB to 20 dB. To be in line with the experiments from [**P4**]-[**P5**], for each value of ENR, we evaluate $P_{\rm D}$ when $\zeta_1^2 = \text{ENR}$ and $\zeta_1^2 = \text{ENR}/2$, respectively. By relying on (4.18), we have from Table 4.1 that $P_{\rm FA} = 0.1819$. The results are shown in Fig. 4.2, where we also plot the $P_{\rm D}$ achieved by the GLRT when $P_{\rm FA} = 0.1819$. The results for GLRT are produced by applying Theorem 9.1 from [28]. From the same theorem, one can notice that the performance of GLRT depends on ENR, but it does not depend on how ENR is "distributed" between the entries of the vector $\boldsymbol{\zeta}$.

Next we consider the detection rule which is obtained by re-writing the condition (4.14) for an arbitrary d instead of the optimum $\hat{d}$ provided at the point a) of the Theorem 4.2.1. Therefore, after observing $\mathbf{x}$, $\mathcal{M}_0$ is selected if

$$\max\left(|t_1|, \ldots, |t_k|\right) < \sqrt{d/k}. \tag{4.29}$$

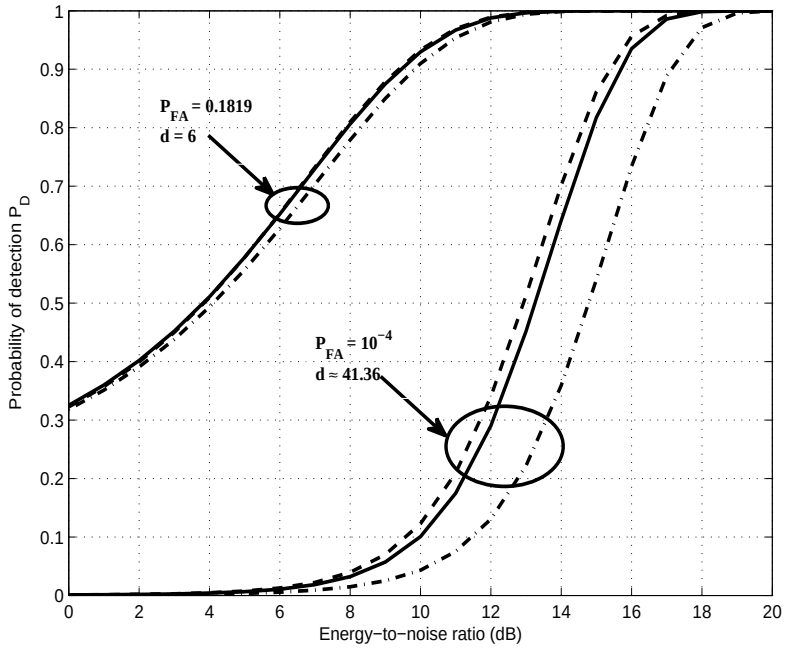

Figure 4.2: Comparison of the GLRT with the ODD detector in (4.29) for two cases: (i) d equals the optimum value given by Theorem 4.2.1; (ii) d is chosen such that $P_{\mathrm{FA}} = 10^{-4}$. The sample size is $N = 50$ and the number of linear parameters is $k = 2$. Graphical conventions: solid line - GLRT, dashed line - ODD when $\zeta_1^2 = \mathrm{ENR}$, and dash-dotted line - ODD when $\zeta_1^2 = \mathrm{ENR}/2$. Note that ζ_1 is the first entry of the vector $\boldsymbol{\zeta}$ which is defined in (4.28), and ENR is the acronym for the energy-to-noise ratio ($||\boldsymbol{\zeta}||^2 = \mathrm{ENR}$). Remark for $d = 6$ that the P_{D} curves of GLRT and ODD when $\zeta_1^2 = \mathrm{ENR}$ are very close and one cannot easily distinguish between them.

It is straightforward to prove that the P_{FA} for (4.29) is given by

$$1 - \int_{(-c\boldsymbol{\omega}, c\boldsymbol{\omega})} \bar{\varphi}(\bar{\mathbf{t}}; \mathbf{0}) \mathrm{d}\bar{\mathbf{t}},$$

where $\boldsymbol{\omega} = \left[\sqrt{d/k} \ \sqrt{d/k}\right]^{\top}$. Then we fix the P_{FA} to be 10^{-4}, and by applying a grid search method we find out that the corresponding value of d is approximately 41.36. By using this value of d, we compute the P_{D} for (4.29) with the same experimental settings that have been employed for the ODD detector with $d = \hat{d} = 3k$. The results are plotted in Fig. 4.2, together with the P_{D} for GLRT when the P_{FA} takes the predefined value of 10^{-4}.

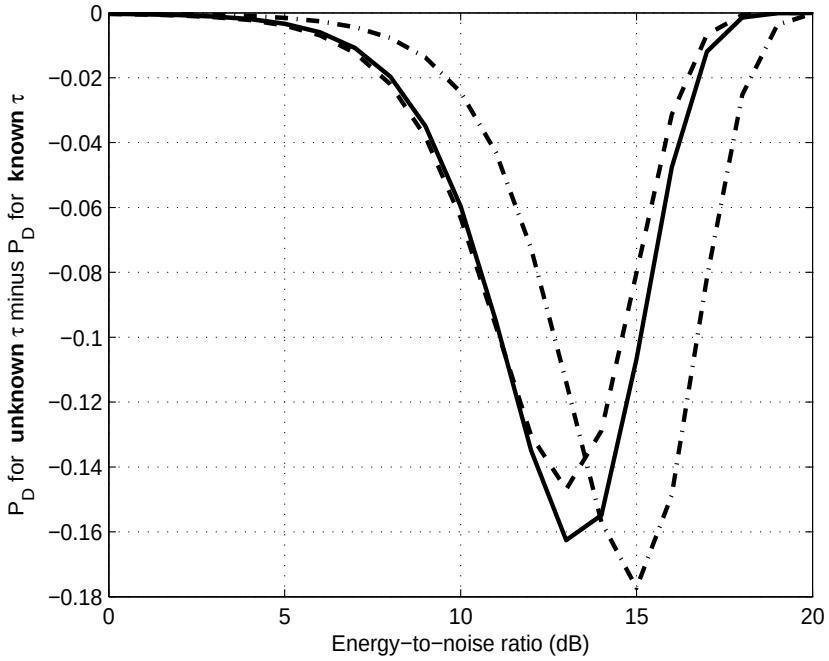

Figure 4.3: GLRT and ODD for which P_{FA} equals 10^{-4}: decrease of P_{D} when τ is unknown. All the experimental settings as well as the graphical conventions are the same like in Fig. 4.2.

These numerical examples suggest that the conclusions which have been drawn in [**P4**] for the case when the variance is known can be extended to the case with unknown variance. So, the selection $d = \hat{d}$ leads to $P_{\mathrm{FA}} = 0.1819$, which is too large for many practical applications. For this P_{FA}, the ODD and GLRT detectors have similar performance. If the P_{FA} is fixed to a much smaller value like for example 10^{-4}, then ODD is superior to GLRT when the whole ENR is "concentrated" only in ζ_1. However, ODD becomes clearly inferior to GLRT when ENR is equally "distributed" between ζ_1 and ζ_2.

All that remains is to clarify how the performance is lowered due to lack of knowledge on the noise variance. We illustrate this aspect by comparing in Fig. 4.3 the results of (4.29) with those of the modified ODD from [**P4**], which selects $\mathcal{M}_0$ whenever

$$\max\left(|z_1|, \ldots, |z_k|\right) < \sqrt{d/k},$$

where the definition of z_j is given in (4.7) for all $j \in \{1, \ldots, k\}$. We show in the same figure how the performance of the GLRT changes when τ is unknown. For the

GLRT, the comparison reduces to subtract from the $P_{\rm D}$ computed with [28, Theorem 9.1] the $P_{\rm D}$ given by [28, Theorem 7.1]. Remark that, for both GLRT and ODD, the performance is the same as in the known case when ENR is either very small or very large.

4.4 Appendix: Proof of Theorem 4.2.1

We briefly outline some results which are instrumental in our proof. The notation χ^2_f is used for the chi-square distribution with f degrees of freedom.

Result 1 [59, Theorem 3.5] If $\mathbf{x} \sim \mathcal{N}_N(\mathbf{H}\boldsymbol{\theta}, \tau\mathbf{I})$ then (1) $\hat{\boldsymbol{\theta}} \sim \mathcal{N}_k(\boldsymbol{\theta}, \tau(\mathbf{H}^\top\mathbf{H})^{-1})$, (2) $(N\hat{\tau})/\tau \sim \chi^2_{N-k}$, and (3) $\hat{\boldsymbol{\theta}}$ is independent of $\hat{\tau}$. Note that $\hat{\boldsymbol{\theta}}$ and $\hat{\tau}$ are defined in (4.5) and (4.12), respectively.

Result 2 [38, 48] The Gaussian density function whose formula is given in (4.2) can be factored as follows:

$$f(\mathbf{x};\boldsymbol{\theta},\tau) = f(\mathbf{x}|\hat{\boldsymbol{\theta}},\hat{\tau})g(\hat{\boldsymbol{\theta}},\hat{\tau}|\boldsymbol{\theta},\tau) \tag{4.30}$$

where the conditional density $f(\mathbf{x}|\hat{\boldsymbol{\theta}},\hat{\tau})$ does not depend on the unknown parameters $\boldsymbol{\theta}$ and τ, and

$$\begin{aligned} g(\hat{\boldsymbol{\theta}},\hat{\tau};\boldsymbol{\theta},\tau) &= g_1(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})g_2(\hat{\tau};\tau), \\ g_1(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) &= \frac{|\mathbf{H}^\top\mathbf{H}|^{1/2}}{(2\pi\tau)^{k/2}} \exp\left(-\frac{1}{2\tau}\|\mathbf{H}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\|^2\right), \\ g_2(\hat{\tau};\tau) &= \frac{(N/2)^{(N-k)/2}}{\Gamma\left((N-k)/2\right)} \left(\frac{\hat{\tau}}{\tau}\right)^{(N-k)/2} \frac{1}{\hat{\tau}} \exp\left(-\frac{N}{2}\frac{\hat{\tau}}{\tau}\right). \end{aligned} \tag{4.31}$$

Result 3 [31, 32] (Kshirsagar noncentral multivariate t-distribution) Let $\mathbf{x}' \sim \mathcal{N}_p(\boldsymbol{\mu}', \sigma^2\mathbf{I})$ and let $u = fs^2/\sigma^2$ be distributed independently of $\mathbf{x}'$ according to a χ^2_f distribution, so that s^2 is an estimate of σ^2. Then the distribution of $\mathbf{t}' = \mathbf{x}'/s$ is given by

$$\varphi_{\rm ksh}(\mathbf{t}';\boldsymbol{\delta}',f) = \frac{\exp\left(-||\boldsymbol{\delta}'||^2/2\right)}{(\pi f)^{p/2}\Gamma(f/2)} \sum_{\alpha=0}^{\infty} \frac{2^{\alpha/2}\left(\mathbf{t}'^\top\boldsymbol{\delta}'\right)^\alpha \Gamma\left((f+p+\alpha)/2\right)}{f^{\alpha/2}\alpha!\left(1+||\mathbf{t}'||^2/f\right)^{(f+p+\alpha)/2}}, \tag{4.32}$$

where $\boldsymbol{\delta}' = \boldsymbol{\mu}'/\sigma$.

a) Optimum value of parameter d

With the convention that $\mathcal{X}_0 = \{\mathbf{x} : (\hat{\boldsymbol{\theta}}(\mathbf{x}), \hat{\tau}(\mathbf{x})) \in B_{d/N}(0)\}$, the KL divergence

can be computed as follows:

$$\begin{aligned} & D(\hat{f}_0(\mathbf{x}) \parallel \tilde{f}_0(\mathbf{x})) \\ &= \frac{1}{Q_{d/N}(0)} \int_{\mathcal{X}_0} f(\mathbf{x}; \hat{\boldsymbol{\theta}}, \hat{\tau}) \ln \frac{f(\mathbf{x}; \hat{\boldsymbol{\theta}}, \hat{\tau})/Q_{d/N}(0)}{f(\mathbf{x}; \boldsymbol{\theta}^0, \hat{\tau}_0)/C_0} \mathrm{d}\mathbf{x} \\ &= \ln \frac{C_0}{Q_{d/N}(0)} + \frac{1}{Q_{d/N}(0)} \int_{\mathcal{X}_0} f(\mathbf{x}; \hat{\boldsymbol{\theta}}, \hat{\tau}) \ln \frac{f(\mathbf{x}; \hat{\boldsymbol{\theta}}, \hat{\tau})}{f(\mathbf{x}; \boldsymbol{\theta}^0, \hat{\tau}_0)} \mathrm{d}\mathbf{x}. \end{aligned} \tag{4.33}$$

First we find closed-form expressions for C_0 and $Q_{d/N}(0)$. Then we evaluate the integral term in (4.33). Similarly with the calculations from [46], we have:

$$C_0 = \int_{\hat{\tau}_0 \in (\tau_1, \tau_2)} \left[\int_{\mathbf{x}: \hat{\tau}_0(\mathbf{x}) = \tau_0} f(\mathbf{x} | \boldsymbol{\theta}^0, \hat{\tau}_0) \mathrm{d}\mathbf{x} \right] g_2(\hat{\tau}_0, \hat{\tau}_0) \mathrm{d}\hat{\tau}_0 \tag{4.34}$$

$$= \int_{\tau_1}^{\tau_2} g_2(\hat{\tau}_0; \hat{\tau}_0) \mathrm{d}\hat{\tau}_0 \tag{4.35}$$

$$= \frac{(N/2)^{N/2} \exp(-N/2)}{\Gamma(N/2)} \ln \frac{\tau_2}{\tau_1}. \tag{4.36}$$

The identity in (4.34) is obtained by applying Result 2 (for $k = 0$) to the definition of C_0 from (4.13). Then we get (4.35) by noticing that the inner integral in (4.34) gives unity.

For an arbitrary pair $(\hat{\boldsymbol{\theta}}, \hat{\tau})$ with $\tau_1 < \hat{\tau} < \tau_2$, we denote $\mathcal{X}_{\hat{\boldsymbol{\theta}}, \hat{\tau}} = \{\mathbf{x} : (\hat{\boldsymbol{\theta}}(\mathbf{x}), \hat{\tau}(\mathbf{x})) = (\hat{\boldsymbol{\theta}}, \hat{\tau})\}$. The definition from (4.11) together with Result 2 produce the following chain of identities:

$$\begin{aligned} Q_{d/N}(0) &= \int_{(\hat{\boldsymbol{\theta}}, \hat{\tau}) \in B_{d/N}(0)} \left[\int_{\mathcal{X}_{\hat{\boldsymbol{\theta}}, \hat{\tau}}} f(\mathbf{x} | \hat{\boldsymbol{\theta}}, \hat{\tau}) \mathrm{d}\mathbf{x} \right] g(\hat{\boldsymbol{\theta}}, \hat{\tau}; \hat{\boldsymbol{\theta}}, \hat{\tau}) \mathrm{d}\hat{\boldsymbol{\theta}} \mathrm{d}\hat{\tau} \\ &= \int_{B_{d/N}(0)} g(\hat{\boldsymbol{\theta}}, \hat{\tau}; \hat{\boldsymbol{\theta}}, \hat{\tau}) \mathrm{d}\hat{\boldsymbol{\theta}} \mathrm{d}\hat{\tau} \\ &= h_{\mathbf{H},N,k} \int_{B_{d/N}(0)} \frac{1}{\hat{\tau}^{k/2+1}} \mathrm{d}\hat{\boldsymbol{\theta}} \mathrm{d}\hat{\tau} \\ &= h_{\mathbf{H},N,k} \int_{\tau_1}^{\tau_2} \frac{2^k}{\hat{\tau}^{k/2+1}} \left[\int_{\mathfrak{M}_{d/N}(0)} \mathrm{d}\hat{\boldsymbol{\eta}} \right] \mathrm{d}\hat{\tau} \qquad (4.37) \\ &= d^{k/2} h'_{\mathbf{H},N,k} \ln \frac{\tau_2}{\tau_1}, \qquad (4.38) \end{aligned}$$

where

$$h_{\mathbf{H},N,k} = \frac{|\mathbf{H}^\top \mathbf{H}|^{1/2}(N/2)^{(N-k)/2}\exp(-N/2)}{(2\pi)^{k/2}\Gamma\left((N-k)/2\right)},$$
$$h'_{\mathbf{H},N,k} = \left(\frac{2}{k\pi}\right)^{k/2}\frac{(N/2)^{(N-k)/2}\exp(-N/2)}{\Gamma\left((N-k)/2\right)}.$$

Remark that (4.37) is obtained by changing the variables $(\hat{\boldsymbol{\theta}}, \hat{\tau})$ to $(\hat{\boldsymbol{\eta}}, \hat{\tau})$, where $\hat{\boldsymbol{\eta}} = \mathbf{V}^\top \hat{\boldsymbol{\theta}}$ and $\mathbf{V}$ is the matrix from (4.9). In (4.37) we use also the notation $\mathfrak{M}_{d/N}(0) = [0, \hat{\mu}_1] \times \cdots \times [0, \hat{\mu}_k]$, where

$$\hat{\mu}_j = \left(\frac{d\hat{\tau}}{k\ell_j}\right)^{1/2} \quad \forall j \in \{1, \ldots, k\}. \tag{4.39}$$

Now we focus on the evaluation of the integral term from (4.33):

$$\begin{aligned}
&\int_{\mathcal{X}_0} f(\mathbf{x}; \hat{\boldsymbol{\theta}}, \hat{\tau}) \ln \frac{f(\mathbf{x}; \hat{\boldsymbol{\theta}}, \hat{\tau})}{f(\mathbf{x}; \boldsymbol{\theta}^0, \hat{\tau}_0)} \mathrm{d}\mathbf{x} \\
&= \int_{B_{d/N}(0)} \left[\int_{\mathcal{X}_{\hat{\boldsymbol{\theta}},\hat{\tau}}} f(\mathbf{x} \mid \hat{\boldsymbol{\theta}}, \hat{\tau}) \mathrm{d}\mathbf{x}\right] g(\hat{\boldsymbol{\theta}}, \hat{\tau}; \hat{\boldsymbol{\theta}}, \hat{\tau}) \ln\left(\frac{\hat{\tau}_0}{\hat{\tau}}\right)^{N/2} \mathrm{d}\hat{\boldsymbol{\theta}}\mathrm{d}\hat{\tau} \\
&= \frac{N}{2}\int_{B_{d/N}(0)} g(\hat{\boldsymbol{\theta}}, \hat{\tau}; \hat{\boldsymbol{\theta}}, \hat{\tau}) \ln\left(1 + \frac{\|\mathbf{H}\hat{\boldsymbol{\theta}}\|^2}{N\hat{\tau}}\right) \mathrm{d}\hat{\boldsymbol{\theta}}\mathrm{d}\hat{\tau} \\
&= \frac{Nh_{\mathbf{H},N,k}}{2}\int_{\tau_1}^{\tau_2} \frac{2^k}{\hat{\tau}^{k/2+1}} \left[\int_{\mathfrak{M}_{d/N}(0)} \ln\left(1 + \frac{\sum_{j=1}^k \hat{\eta}_j^2 \ell_j}{N\hat{\tau}}\right) \mathrm{d}\hat{\boldsymbol{\eta}}\right] \mathrm{d}\hat{\tau}
\end{aligned}$$

Most of the calculations above are straightforward. The last result is obtained after applying the same change of variables like in (4.37).

We take

$$\varepsilon_N = \sum_{j=1}^k \frac{\hat{\eta}_j^2 \ell_j}{N\hat{\tau}},$$

where $|\hat{\eta}_j| \leq \hat{\mu}_j$ for all $j \in \{1, \ldots, k\}$. By using (4.39), we get

$$\varepsilon_N \leq \sum_{j=1}^k \frac{\hat{\mu}_j^2 \ell_j}{N\hat{\tau}} = \frac{d}{N}.$$

As $d \ll N$, we employ the approximation $\ln(1 + \varepsilon_N) \approx \varepsilon_N$, which leads to

$$\begin{aligned}
&\int_{\mathcal{X}_0} f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})\ln\frac{f(\mathbf{x};\hat{\boldsymbol{\theta}},\hat{\tau})}{f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0)}\mathrm{d}\mathbf{x}\\
&\quad\approx\quad \frac{Nh_{\mathbf{H},N,k}}{2}\int_{\tau_1}^{\tau_2}\frac{2^k}{\hat{\tau}^{k/2+1}}\left[\int_{\mathfrak{M}_{d/N}(0)}\frac{\sum_{j=1}^k\hat{\eta}_j^2\ell_j}{N\hat{\tau}}\mathrm{d}\hat{\boldsymbol{\eta}}\right]\mathrm{d}\hat{\tau}\\
&\quad=\quad \frac{d^{k/2+1}}{6}h'_{\mathbf{H},N,k}\ln\frac{\tau_2}{\tau_1}\\
&\quad=\quad \frac{d}{6}Q_{d/N}(0). \qquad (4.40)
\end{aligned}$$

Combining the results from (4.33), (4.36), (4.38) and (4.40), we get

$$D(\hat{f}_0(\mathbf{x})\parallel\tilde{f}_0(\mathbf{x}))\quad\approx\quad\frac{k}{2}\ln\frac{k\pi}{2d}+\frac{d}{6}+\ln\left(\frac{(N/2)^{k/2}\Gamma\left((N-k)/2\right)}{\Gamma(N/2)}\right). \qquad (4.41)$$

It is clear from (4.41) that $D(\hat{f}_0(\mathbf{x})\parallel\tilde{f}_0(\mathbf{x}))$ is minimized for $\hat{d}=3k$.

The calculations above also show that selecting the "natural" model for $\mathcal{M}_0$ to be the ML function $f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0)$ instead of the NML function $f(\mathbf{x};\boldsymbol{\theta}^0,\hat{\tau}_0)/C_0$ leads to the same optimum $\hat{d}=3k$. The KL distance between the "artificial" and the "natural" models will be different when choosing ML instead of the NML, but this is less important for our detection problem.

b) Detection strategy: After observing $\mathbf{x}$, the model class $\mathcal{M}_0$ is selected when $(\hat{\boldsymbol{\theta}},\hat{\tau})\in B_{\hat{d}/N}(0)$. This reduces to verifying for all $j\in\{1,\ldots,k\}$ that

$$|\mathbf{v}_j^\top\hat{\boldsymbol{\theta}}|<\hat{\mu}_j.$$

By using (4.5), (4.8) and (4.9), one can easily prove that the condition above is equivalent to

$$\frac{|\mathbf{v}_j^\top\mathbf{H}^\top\mathbf{x}|}{\ell_j}<\hat{\mu}_j,$$

which becomes

$$\frac{|\mathbf{v}_j^\top\mathbf{H}^\top\mathbf{x}|}{\ell_j}<\left(\frac{3\hat{\tau}}{\ell_j}\right)^{1/2} \qquad (4.42)$$

by employing (4.39) for $d=\hat{d}=3k$. The condition in (4.14) is obtained straightforwardly from (4.42).

c) Confidence indices

First we prove the following result on the distributional properties of the vector $\mathbf{t} = [t_1, \ldots, t_k]^\top$.

Result 4 If $\mathbf{x} \sim \mathcal{N}_N(\mathbf{H}\boldsymbol{\theta}, \tau\mathbf{I})$, then the PDF of the random vector $\mathbf{t}$ is the one from (4.15), where

$$\boldsymbol{\delta} = \mathbf{D}^{1/2}\mathbf{V}^\top\boldsymbol{\theta}/\sqrt{\tau}. \tag{4.43}$$

Proof Let $\bar{\mathbf{t}} = \frac{\mathbf{D}^{1/2}\mathbf{V}^\top\hat{\boldsymbol{\theta}}}{\sqrt{\hat{\nu}}}$, where $\hat{\nu} = \frac{N}{N-k}\hat{\tau}$ is the unbiased estimator of the unknown variance τ. Based on Result 1, we have $\hat{\boldsymbol{\theta}} \sim \mathcal{N}_k(\boldsymbol{\theta}, \tau(\mathbf{H}^\top\mathbf{H})^{-1})$, which leads to $\mathbf{D}^{1/2}\mathbf{V}^\top\hat{\boldsymbol{\theta}} \sim \mathcal{N}_k(\mathbf{D}^{1/2}\mathbf{V}^\top\boldsymbol{\theta}, \tau\mathbf{I})$ (see, for example, [40, Theorem 3.2.1]). We have also from Result 1 that $(N-k)\hat{\nu}/\tau \sim \chi^2_{N-k}$, and $\hat{\boldsymbol{\theta}}$ is independent of $\hat{\nu}$. Then we operate the following substitutions in Result 3: $p = k$, $\boldsymbol{\mu}' = \mathbf{D}^{1/2}\mathbf{V}^\top\boldsymbol{\theta}$, $\sigma^2 = \tau$, $s^2 = \hat{\nu}$ and $f = N - k$. So, $u = (N-k)\hat{\nu}/\tau$, and the distribution of $\bar{\mathbf{t}}$ is given by

$$\bar{\varphi}(\bar{\mathbf{t}}; \boldsymbol{\delta}) = \varphi_{\text{ksh}}(\bar{\mathbf{t}}; \boldsymbol{\delta}, N-k),$$

where $\boldsymbol{\delta}$ is defined in (4.43). The expression of $\bar{\varphi}(\bar{\mathbf{t}}; \boldsymbol{\delta})$ is presented in (4.22). The distribution of $\mathbf{t}$ is readily obtained by applying the change of variables $\mathbf{t} = \sqrt{\frac{N}{N-k}}\bar{\mathbf{t}}$. □

By definition, $B_{\hat{d}/N}(i) = \bigcup_{\tau_1<\tau<\tau_2} B^\tau_{\hat{d}/N}(i)$ for an arbitrary index i. Given $\tau \in (\tau_1, \tau_2)$, the center of the rectangle $B^\tau_{d/N}(i)$ is $\boldsymbol{\theta}^i(\tau) = \sqrt{\tau}\mathbf{V}\mathbf{D}^{-1/2}(2\sqrt{3}\mathbf{m}^i)$, where all entries of the vector $\mathbf{m}^i = [m^i_1, \ldots, m^i_k]^\top$ are integers. So, $\boldsymbol{\theta}^i(\tau)/\sqrt{\tau}$ does not depend on the value of τ. Based on Result 4, we have that, for all $\tau \in (\tau_1, \tau_2)$, the PDF of $\mathbf{t}$ is $\varphi(\mathbf{t}; 2\sqrt{3}\mathbf{m}^i)$ if $\mathbf{x} \sim \mathcal{N}_N\left(\mathbf{H}\boldsymbol{\theta}^i(\tau), \tau\right)$.

When the condition in (4.14) is satisfied, we have $(\hat{\boldsymbol{\theta}}, \hat{\tau}) \in B_{\hat{d}/N}(0)$, and by applying the result above for $i = 0$, we obtain:

$$\begin{aligned} E1 &= 1 - P_{0|0} \\ &= 1 - P\left(B_{\hat{d}/N}(0) \mid \boldsymbol{\theta}/\sqrt{\tau} = \mathbf{0}\right) \\ &= 1 - \text{Prob}\{-\boldsymbol{\xi} < \mathbf{t} < \boldsymbol{\xi}; \boldsymbol{\theta}/\sqrt{\tau} = \mathbf{0}\} \\ &= 1 - \int_{(-\boldsymbol{\xi}, \boldsymbol{\xi})} \varphi(\mathbf{t}; \mathbf{0}) \mathrm{d}\mathbf{t}. \end{aligned}$$

If the condition in (4.14) is not satisfied, then $(\hat{\boldsymbol{\theta}}, \hat{\tau})$ falls within $B_{\hat{d}/N}(i)$ which has the property that $\mathbf{m}^i = \mathbf{m}$. It is evident that at least one entry of $\mathbf{m}^i$ is nonzero.

So,

$$\begin{aligned} E2 &= P_{0|i} \\ &= P\left(B_{\hat{d}/N}(0) \mid \boldsymbol{\theta}/\sqrt{\tau} = \mathbf{V}\mathbf{D}^{-1/2}(2\sqrt{3}\mathbf{m})\right) \\ &= \text{Prob}\left\{-\boldsymbol{\xi} < \mathbf{t} < \boldsymbol{\xi}; \boldsymbol{\theta}/\sqrt{\tau} = \mathbf{V}\mathbf{D}^{-1/2}(2\sqrt{3}\mathbf{m})\right\} \\ &= \int_{(-\boldsymbol{\xi},\boldsymbol{\xi})} \varphi(\mathbf{t}; 2\sqrt{3}\mathbf{m})\mathrm{d}\mathbf{t}, \end{aligned}$$

which concludes the proof. □

Chapter 5

Cepstral nulling: selection of the threshold via Kolmogorov structure function

One of the classical problems in statistics is the following. Consider a vector of independent Gaussian random variables with unknown means but known variances. A possible approach for reducing the total variance (TV) of these random variables is to exploit the (a priori) information that most of them have "small" means [68]. A similar problem occurs in cepstral analysis. More precisely, under the assumption that the observed signal is stationary, the estimated cepstral coefficients have neat distributional properties which allow to recast the reduction of their TV in the mathematical framework which was shortly discussed above [67].

This line of thinking led to the thresholding procedure from [68], where two different schemes have been applied for the selection of the threshold: the first one is based on a carefully designed most powerful unbiased test (UMPUT), while the second one uses a modified form of BIC. In [**P6**], we have introduced a threshold selection method which is based on the KSF. The estimated cepstrum resulting after thresholding can be further utilized to smooth the periodogram.

We note that KSF and BIC are fully automatic procedures, whereas UMPUT requires supplementary information provided by the user. Next we briefly outline the three selection methods and compare their performance.

5.1 Cepstral nulling

Periodogram For a stationary, discrete time, real-valued signal, consider the estimation of the spectrum $\Phi(\omega)$ from the measurements $y_0, \ldots, y_{N-1}$. With the convention that $\omega_p = (2\pi p)/N$, $p \in \{0, \ldots, N-1\}$, are the Fourier frequency grid points, we use the notation Φ_p for $\Phi(\omega_p)$. The estimate of the spectrum at point ω_p

is [66]:

$$\hat{\Phi}_p = \frac{1}{N}\left|\sum_{t=0}^{N-1} y_t \exp(-i\omega_p t)\right|^2,$$

where $i = \sqrt{-1}$.

Cepstral coefficients We assume that N, the number of samples, is even and take $M = N/2 + 1$. Under the hypothesis that $\min\{\Phi_p, \hat{\Phi}_p\} > 0$ for all p, the first M cepstral coefficients and their estimates are given by [68]:

$$\begin{aligned} c_j &= \frac{1}{N}\sum_{p=0}^{N-1} \ln(\Phi_p)\exp(i\omega_j p), \\ \hat{c}_j &= \frac{1}{N}\sum_{p=0}^{N-1} \ln(\hat{\Phi}_p)\exp(i\omega_j p) + \gamma\delta_{j,0}, \end{aligned}$$

where $j \in \{0, \ldots, M-1\}$ and $\gamma = 0.577216\ldots$ is the Euler-Mascheroni constant. The Kronecker indicator $\delta_{j,0}$ takes value one if $j = 0$, and otherwise takes value zero. The rest of the coefficients can be obtained without difficulties because $c_{N-j} = c_j$ and $\hat{c}_{N-j} = \hat{c}_j$ for $j \in \{1, \ldots, M-2\}$.

Theorem 5.1.1. [13, 21, 67] When $N \gg 1$, the normalized vector of estimation errors $\sqrt{N}(\hat{\mathbf{c}} - \mathbf{c})$ converges in distribution to the normal distribution of mean zero and covariance matrix

$$\mathbf{C} = \frac{\pi^2}{6}\begin{bmatrix} 2 & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ 0 & & & & 2 \end{bmatrix}. \tag{5.1}$$

For clarifying the meaning of $N \gg 1$, we mention that, according to [68], the distribution of the estimation errors may differ significantly from the one in Theorem 5.1.1 if $N < 128$. It was also pointed out in [68] that, in most of the practical applications, the number of measurements is not larger than 2048. Hence, we focus on the cases when

$$128 \leq N \leq 2048. \tag{5.2}$$

Thresholding schemes Based on the distributional properties of

$$\hat{\mathbf{c}} = [\hat{c}_0 \ldots \hat{c}_{M-1}]^\top,$$

the following thresholding scheme was introduced [20, 68]:

$$\check{c}_j = \begin{cases} 0, & |\hat{c}_j| < \mu \left[\mathbf{C}(j+1, j+1)/N\right]^{1/2}, \\ \hat{c}_j, & \text{otherwise}, \end{cases} \tag{5.3}$$

where $j \in \{0, \ldots, M-1\}$ and $\mathbf{C}(j+1, j+1)$ is the $(j+1)$-th diagonal element of the matrix $\mathbf{C}$ which is defined in (5.1). The parameter $\mu > 0$ determines the threshold level.

The statistical properties of the thresholding scheme have been investigated in [16]. Based on Theorem 5.1.1, the following method for smoothing the periodogram was proposed in [33]. First, the empirical cepstrum is plotted by tracing the points $\{(k, \hat{c}_k)\}_{k=0}^{M-1}$ in the plane. Then a so-called grid-transformation is applied which re-scales the coordinates so that the horizontal distance between $\hat{c}_k$ and $\hat{c}_{k+1}$ becomes larger for small values of k, and smaller for large k. After transformation, the coefficients are smoothed with local linear regression. In comparison with thresholding, the solution from [33] has a higher degree of complexity, both conceptually and computationally.

Threshold selection We present below two different methods which have been proposed for the selection of parameter μ.

UMPUT [68] - The formula is derived by combining the UMPUT test [35] with some empirical evidence, and it is given by:

$$\mu_{\text{UMPUT}} = (5 - \mathrm{I}_{\text{st}}) + \frac{N - 128}{1920}. \tag{5.4}$$

I_{st} is a parameter to be chosen according to the signal type:

- $\mathrm{I}_{\text{st}} = 1$ for broadband signal with small dynamic range;
- $\mathrm{I}_{\text{st}} = 2$ for broadband signal with medium dynamic range;
- $\mathrm{I}_{\text{st}} = 3$ for narrowband signal with large dynamic range.

The application of μ_{UMPUT} is restricted to data sets for which N is an integer power of two and satisfies the condition in (5.2).

BIC [68] - In this case, the choice of the threshold relies on BIC [58]. The expression of μ is:

$$\mu_{\text{BIC}} = 1 + (\ln M)^{1/2}. \tag{5.5}$$

5.2 An approach based on KSF

The key point is to recast the thresholding as a protocol for transmitting $\{\hat{c}_j\}_{j=0}^{M-1}$ from an hypothesized encoder to a decoder. However, the methods based on the

MDL principle [47, 53] cannot be applied straightforwardly because we do not want to identify the coefficients which are deemed to be noninformative, but rather to test the statistical hypothesis [68]:

$$c_j^2 \leq \mathbf{C}(j+1, j+1)/N,\ j \in \{0, \ldots, M-1\}. \tag{5.6}$$

Note that nulling some of the coefficients $\{\hat{c}_j\}_{j=0}^{M-1}$ is similar to a quantization process. This observation makes the connection between the thresholding procedure defined in (5.3) and the formula from (2.32). However, for re-deriving (2.32) in the particular case of our problem, we need to consider the following steps: (1) Define a partition for the space of the cepstral coefficients; (2) Quantize the estimated cepstral coefficients to the values given by the centroids of the equivalence classes to which they belong; (3) Evaluate the code length for the quantized coefficients, as well as the distortion produced by quantization.

For the sake of brevity, we do not outline all the derivations which can be found in [**P6**]. It was also pointed out in [**P6**] that it is more convenient to apply the transformation

$$x_j = \hat{c}_j \left[N/\mathbf{C}(j+1, j+1)\right]^{1/2},\ j \in \{0, \ldots, M-1\},$$

and then to perform all the calculations by using $\{x_j\}_{j=0}^{M-1}$ instead of $\{\hat{c}_j\}_{j=0}^{M-1}$. The outcome of Steps (1) and (2) is the operator $\mathcal{Q}(\cdot)$ which gives, for and arbitrary $\mathbf{x} \in \mathbb{R}^M$, the centroid of the hyper-cube to which $\mathbf{x}$ belongs. According to [**P6**], $\mathcal{Q}(\mathbf{x}) = [q(x_0) \ldots q(x_{M-1})]^\top$, where

$$q(x_j) = \begin{cases} 0, & |x_j| < 1 + \ell \\ \operatorname{sgn}(x_j)\left(1 + 2\ell + 2\ell \left\lfloor \frac{|x_j| - 1 - \ell}{2\ell} \right\rfloor\right), & \text{otherwise.} \end{cases} \tag{5.7}$$

The nonnegative number $\ell \geq 0$ is a parameter. To draw a parallel to the parametrization from Section 2.2.5, we mention that $\ell = \sqrt{d/M}$. Note from the equation (5.7) that the volume of the equivalence class centered at zero is larger than the volumes of all other equivalence classes. This makes us to apply a special strategy in the evaluation of the term $L_d(\boldsymbol{\theta}^i)$ from (2.32). Recall that $L_d(\boldsymbol{\theta}^i)$ is the code length for the center of the equivalence class where the ML estimate falls.

It is helpful to introduce some definitions. Let $\eta = \{j : 0 \leq j \leq M-1, q(x_j) \neq 0\}$. The cardinality of η is denoted by k, and we assume that $0 < k < M$. Additionally, we have:

$$\begin{aligned} \eta &= \{j_0, \ldots, j_{k-1}\}, \\ \mathbf{z} &= [x_{j_0}, \ldots, x_{j_{k-1}}]^\top, \\ \tilde{\mathbf{z}} &= [q(x_{j_0}), \ldots, q(x_{j_{k-1}})]^\top. \end{aligned}$$

The decoder will have full information on the quantized values of $\{x_j\}_{j=0}^{M-1}$ if the encoder transmits the entries of η and $\tilde{\mathbf{z}}$.

The code length for η, which is denoted $L_\eta(\ell)$, can be evaluated by applying the same method which was used to derive (3.9). The major difference is that we force at least one of the values $\{x_j\}_{j=0}^{M-1}$ to be turned to zero. Hence, we get:

$$L_\eta(\ell) = \min\left\{L_\eta^{\mathcal{A}}, L_\eta^{\mathcal{B}}(\ell)\right\}, \tag{5.8}$$

$$L_\eta^{\mathcal{A}} = \ln(2^M - 2), \tag{5.9}$$

$$L_\eta^{\mathcal{B}}(\ell) = \ln\binom{M}{k} + \ln k + \ln[1 + \ln(M-1)]. \tag{5.10}$$

Obviously, the approximation from (3.10) can be applied when evaluating $L_\eta^{\mathcal{B}}(\ell)$. Notice also that the expression of $L_\eta(\ell)$ is slightly different from the formula employed in [**P6**]. We mention that the formula from [**P6**] is based on [48, Eq. 9.41].

From [**P6**], we have that the code length for $\tilde{\mathbf{z}}$ is

$$L_{\tilde{\mathbf{z}}}(\ell) = \frac{k}{2}\ln\left(\frac{||\tilde{\mathbf{z}}||^2/k}{\ell^2}\frac{\pi e}{2}\right) + \frac{1}{2}\ln k. \tag{5.11}$$

Moreover, instead of measuring the distortion by considering the worst case scenario which corresponds to the term $d/2$ in (2.32), we take the distortion to be given by

$$D_{\mathbf{x}}(\ell) = \frac{1}{2}\sum_{j=0}^{M-1}\left[x_j - q(x_j)\right]^2. \tag{5.12}$$

Because in our settings the first term within (2.32) is a constant, we obtain the following KSF-based criterion:

$$L_{\mathbf{c}}(\ell) = L_\eta(\ell) + L_{\tilde{\mathbf{z}}}(\ell) + D_{\mathbf{x}}(\ell). \tag{5.13}$$

If one increases the value of the parameter ℓ, then it is likely that k decreases, $L_{\tilde{\mathbf{z}}}(\ell)$ decreases as well, whereas $D_{\mathbf{x}}(\ell)$ increases. The optimum value ℓ^* is chosen from a pre-defined set of nonnegative numbers such as to minimize the expression in (5.13), or equivalently,

$$\ell^* = \arg\min_{\ell} L_{\mathbf{c}}(\ell).$$

This result together with (5.7) lead to the following threshold to be used in cepstral nulling:

$$\mu_{\text{KSF}} = 1 + \ell^*. \tag{5.14}$$

5.3 A modified KSF criterion

The assumption for the derivation of the criterion in (5.13) is that all cepstral coefficients are quantized, and not only those which are turned to zero. Because this hypothesis is not in total agreement with (5.3), we propose to modify the KSF criterion such that $\tilde{\mathbf{z}}$ is replaced by $\mathbf{z}$. More precisely, the formula in (5.11) becomes:

$$\bar{L}_{\mathbf{z}}(\ell) = \frac{k}{2} \ln \left(\frac{||\mathbf{z}||^2/k}{\ell^2} \frac{\pi e}{2} \right) + \frac{1}{2} \ln k.$$

In this situation, it is not longer necessary to consider the distortion produced to the coefficients which are not turned to zero. Hence, instead of (5.12), we have

$$\bar{D}_{\mathbf{x}}(\ell) = \frac{1}{2} \sum_{j \in \{0,\ldots,M-1\} \setminus \eta} x_j^2,$$

and the modified KSF criterion (KSFM) is given by

$$\bar{L}_{\mathbf{c}}(\ell) = L_{\eta}(\ell) + \bar{L}_{\mathbf{z}}(\ell) + \bar{D}_{\mathbf{x}}(\ell). \tag{5.15}$$

To gain more insight, let us assume that

$$|x_{(M-1)}| > |x_{(M-2)}| > \cdots > |x_{(0)}|. \tag{5.16}$$

If parameter ℓ satisfies the condition

$$|x_{(M-k)}| \geq 1 + \ell > |x_{(M-k-1)}|, \tag{5.17}$$

then KSFM has the expression:

$$\bar{L}_{\mathbf{c}}(\ell) = L_{\eta}(\ell) + \frac{k}{2} \ln \left(\frac{S_k/k}{\ell^2} \frac{\pi e}{2} \right) + \frac{1}{2} \ln k + \frac{1}{2} \sum_{j=k+1}^{M} x_{(M-j)}^2. \tag{5.18}$$

The formula for $L_{\eta}(\ell)$ is given in (5.8)-(5.10) and, for all $k \in \{1, \ldots, M-1\}$, we denote

$$S_k = \sum_{j=1}^{k} x_{(M-j)}^2. \tag{5.19}$$

It is easy to observe that among all ℓ which satisfy (5.17), the one which minimizes $\bar{L}_{\mathbf{c}}(\ell)$ is $\ell = |x_{(M-k)}| - 1$. This property allows us to re-write $\bar{L}_{\mathbf{c}}(\ell)$ as a function of

k:

$$\bar{L}_{\mathbf{c}}(k) = L_{\eta}(k) + \frac{1}{2}\ln k \tag{5.20}$$

$$+\frac{k}{2}\ln\left[\frac{S_k/k}{(|x_{(M-k)}|-1)^2}\frac{\pi e}{2}\right] \tag{5.21}$$

$$+\frac{1}{2}(S_M - S_k), \tag{5.22}$$

where we re-denoted the formula in (5.8) as $L_{\eta}(k)$. Additionally, $S_M = \sum_{j=0}^{M-1} x_j^2$.

Therefore, finding the optimum μ by using KSFM reduces to select from $\{1,\ldots,M-1\}$ the value of k which minimizes $\bar{L}_{\mathbf{c}}(k)$. Because we want to draw a parallel between KSFM and BIC, we discuss briefly the following interpretation of the BIC-based thresholding scheme. The use of (5.3) when μ has the expression from (5.5) is equivalent with choosing $k \in \{1,\ldots,M-1\}$ so as to minimize:

$$\begin{aligned}\mathrm{BIC}(k) &= \frac{1}{2}\sum_{j=0}^{M-1} x_j^2 - \frac{1}{2}\sum_{j=1}^{k}\left[x_{(M-j)}^2 - \mu_{\mathrm{BIC}}^2\right] \\ &= \frac{1}{2}(S_M - S_k) + \frac{k}{2}\mu_{\mathrm{BIC}}^2 \\ &= \frac{1}{2}(S_M - S_k) + \frac{k}{2}\left(1+\sqrt{\ln M}\right)^2. \end{aligned} \tag{5.23}$$

The most important difference between $\mathrm{BIC}(k)$ and $\bar{L}_{\mathbf{c}}(k)$ is that, in (5.23), the term $\frac{k}{2}\left(1+\sqrt{\ln M}\right)^2$ depends only on k and the number of samples, whereas in (5.21) the term $\frac{k}{2}\ln\left[\frac{S_k/k}{(|x_{(M-k)}|-1)^2}\frac{\pi e}{2}\right]$ depends on k and $x_{(M-1)},\ldots,x_{(M-k)}$. This makes the analysis of $\bar{L}_{\mathbf{c}}(\cdot)$ to be much more complicated than that of $\mathrm{BIC}(\cdot)$.

5.4 Numerical examples

The criterion for evaluating the performance of the thresholding-based scheme defined in (5.3) is the ratio $\rho = \mathrm{TV}(\hat{\mathbf{c}})/\mathrm{TV}(\check{\mathbf{c}})$, where

$$\mathrm{TV}(\hat{\mathbf{c}}) = \sum_{j=0}^{N-1}\mathbb{E}\left[(\hat{c}_j - c_j)^2\right], \tag{5.24}$$

$$\mathrm{TV}(\check{\mathbf{c}}) = \sum_{j=0}^{N-1}\mathbb{E}\left[(\check{c}_j - c_j)^2\right]. \tag{5.25}$$

It is clear that a large value of ρ means a significant reduction of the TV. Like in [68], we calculate TV($\hat{\mathbf{c}}$) and TV($\check{\mathbf{c}}$) by replacing in (5.24)-(5.25) the expectation operator with an average over 1000 Monte Carlo simulations.

For $N \in \{128, 256, 512, 1024, 2048\}$, we generate data according to the following models:

- **Example 1** - broadband MA with a small dynamic range of the log-spectrum [68]:

$$y_t = e_t + 0.55e_{t-1} + 0.15e_{t-2}; \tag{5.26}$$

- **Example 2** - broadband MA with a medium dynamic range of the log-spectrum [33, 68]:

$$y_t = e_t + 0.4574e_{t-1} + 0.2157e_{t-2} + 0.3951e_{t-3} + 0.1383e_{t-4}; \tag{5.27}$$

- **Example 3** - narrowband ARMA with a large dynamic range of the log-spectrum [33, 68]:

$$y_t = 1.55y_{t-1} - 0.95y_{t-2} + e_t + 0.75e_{t-1} + 0.35e_{t-2}; \tag{5.28}$$

In Eqs. (5.26)-(5.28), e_t is zero-mean white Gaussian noise with variance one, and $t \in \{0, \ldots, N-1\}$. The interested reader can find in [33,68] plots with the log-spectra of the three models outlined above.

Note that the same models have been also used for the experimental results reported in [**P6**]. The main reason for which we consider them also here is because we want to illustrate the effect of the modifications operated on the KSF criterion from [**P6**].

The thresholds μ_{UMPUT}, μ_{BIC} and μ_{KSF} are computed with formulas from (5.4), (5.5) and (5.14), respectively. The value of ℓ^* in (5.14) is chosen from the set $\{0.0, 0.1, \ldots, 9.0\}$ so as to minimize the KSF in (5.13). Note that $L_{\mathbf{c}}(\ell) \to \infty$ when $\ell = 0$. The following property is worth mentioning: if $k = 0$ for a particular value $\ell_0 > 0$, then $k = 0$ for all $\ell \in \{\ell_0 + 0.1, \ldots, 9.0\}$. This observation can be used to make the algorithm faster. The parameter ℓ is initialized with value 0.1, then it is increased at each step with 0.1, and the algorithm is stopped when either $k = 0$ or $\ell = 9$.

For the sake of comparison, we compute ρ also for the case when one knows the values of the true cepstral coefficients and selects $\mu \in \{1.0, 1.1, \ldots, 10.0\}$ such that to minimize TV($\check{\mathbf{c}}$). The outcome of this procedure is named μ_{genie} because we have assumed knowledge of the ground truth. Remark that, for a given model, μ_{KSF} changes from one realization to another, whereas μ_{UMPUT}, μ_{BIC} and μ_{genie} are the same for all realizations.

The results of the experiments are shown in Fig. 5.1. For completeness, we plot, in the same figure, the average number of the retained coefficients (ν). Note that ν represents the average number of the cepstral coefficients $\{\hat{c}_j\}_{j=0}^{N-1}$ which have not been set to zero by the thresholding scheme.

For Example 1, the results obtained when thresholding with $\mu = \mu_{\text{BIC}}$ are very modest, whereas for $\mu = \mu_{\text{UMPUT}}$ and $\mu = \mu_{\text{KSF}}$ the TV reduction is almost the same as the one obtained by $\mu = \mu_{\text{genie}}$. Observe that BIC performs so poorly because it retains too many coefficients. For Example 2, all threshold-selection methods lead to similar results, as we can notice from Fig. 5.1. It is interesting that BIC is nearly optimal for the Example 3. For Examples 2 and 3, the average number of coefficients retained by KSF tends to be smaller than the values of ν for the other methods.

Remark how the maximum value of ρ varies from one example to another. It is about 500 for Example 1, becomes about 50 for Example 2, and is as small as 7 for Example 3.

Furthermore, we use the same examples to compare the capabilities of KSF and KSFM. As in the previous experiments, we take μ_{KSF} to be the one given by (5.14), and its selection is done as it was already explained before. We define $\mu_{\text{KSFM}} = 1+\bar{\ell}^*$, where $\bar{\ell}^*$ is chosen from the set $\{0.0, 0.1, \ldots, 9.0\}$ so as to minimize $\bar{L}_{\mathbf{c}}(\ell)$ (see Eq. (5.15)). The results are plotted in Fig. 5.2. Observe that, for all sample sizes, KSFM retains more coefficients than KSF when the signal is broadband with a small dynamic range of the log-spectrum. This leads to a decrease of the performance in the case of Example 1, where the measurements are outcomes from an MA process. We also note that, in Example 1, KSFM is inferior to KSF, but it is much better than BIC.

Remark that the threshold for the KSFM-based cepstral nulling was selected by picking-up from $\{0.0, 0.1, \ldots, 9.0\}$ the value $\bar{\ell}^*$ which minimizes the formula in (5.15). As we already know from Section 5.3, there exists an alternative procedure, namely the value of k is chosen from $\{1, \ldots, M-1\}$ so as to minimize the criterion in (5.20)-(5.22). For the numerical examples considered in this chapter, both methods lead to similar reduction of the TV. However, in most practical applications, M is much larger than the number of points on the ℓ-grid, which makes the method based on the selection of k to have higher computational complexity than the method based on the selection of ℓ.

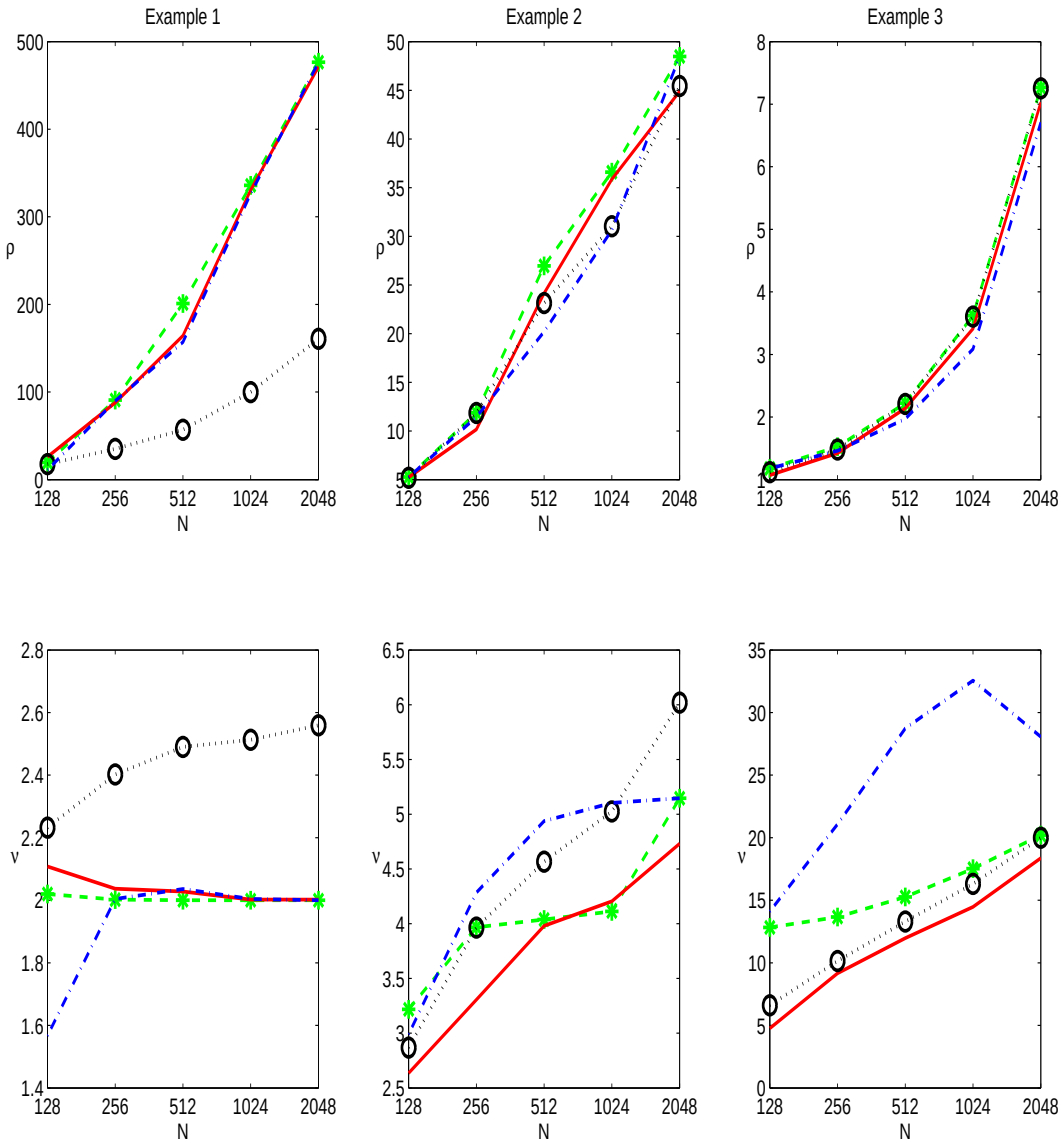

Figure 5.1: Experimental results for the Examples 1-3. First row: the ratio $\rho = \mathrm{TV}(\hat{\mathbf{c}})/\mathrm{TV}(\check{\mathbf{c}})$ versus the sample size N for various selections of the threshold μ. Second row: the average number of retained cepstral coefficients ν versus the sample size. The following values of the threshold are employed in experiments (we indicate in parentheses the color, the line type and, in some cases, the marker symbol used in plots): μ_{genie} (green-dashed line-asterisk), μ_{KSF} (red-solid line), μ_{BIC} (black-dotted line-circle), μ_{UMPUT} (blue-dashdot line).

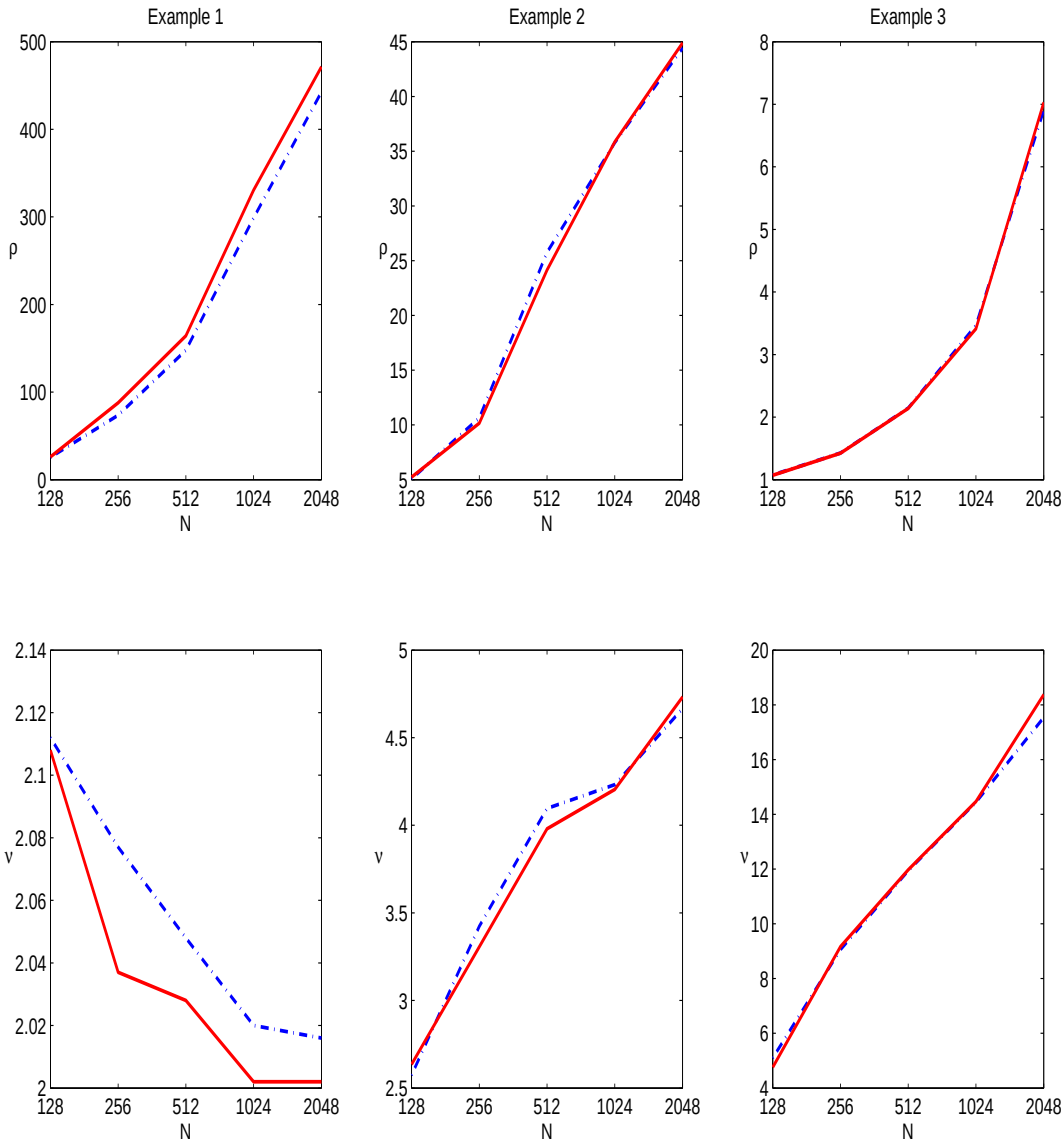

Figure 5.2: Examples 1-3: comparison of the results obtained with KSF (red-solid line) and KSFM (blue-dashdot line). First row: the ratio $\rho = \mathrm{TV}(\hat{\mathbf{c}})/\mathrm{TV}(\check{\mathbf{c}})$ versus the sample size N. Second row: the average number of retained cepstral coefficients ν versus the sample size. Remark in Example 3 that the graphs for KSF and KSFM almost coincide.

Chapter 6

Fairness in multiaccess communication systems

Radio resource management (RRM) is a key function in wireless communication systems which involves strategies and algorithms for admission control, scheduling, subcarrier allocation, rate control, transmit power allocation, choosing the modulation scheme, etc. RRM is essential for utilizing the limited spectrum resources as efficiently as possible and providing quality of service (QoS) in wireless networks.

It is known that, in many cases, the performance measures (e.g., overall throughput) can be optimized if opportunistic algorithms (e.g, opportunistic beamforming [70, Chapter 6]) are employed. However opportunistic RRM techniques always disfavor the users with poor channel conditions or high level of interference, which leads to unfair allocation of resources (e.g., rate) especially under low mobility conditions.

In this chapter, we focus on fairness in rate allocation. We investigate the problem in an information theoretic framework.

6.1 Capacity region of multiaccess channels

6.1.1 Key definitions and concepts

The *capacity* of a channel is the maximum rate of communication over the channel for which arbitrary small error probability can be achieved. Denoting the input variable to a channel by X and output by Y, the capacity of the channel is given by [9]

$$C = \max_{p(x)} I(X;Y),$$

where $p(x)$ is the distribution of X and $I(X;Y)$ is the *mutual information* between X and Y. It is well-known that the capacity of a real Gaussian channel with received

signal power P and noise power N_0 is [9]

$$C = \frac{1}{2}\log_2(1+\frac{P}{N_0}), \quad \text{bits/sec/Hz}.$$

This means that over a Gaussian channel, the reliable rate cannot exceed this value, regardless of the decoding method used. By *decoding* we mean the process which translates the received message to the codewords of a given code. For the complex Gaussian channel the capacity formula becomes $C = \log_2(1+\frac{P}{N_0})$ due to the fact that we can independently transmit on real and imaginary dimensions.

The capacity region of a *multiuser system* includes all reliable rates that the users can achieve. Consider the uplink of a multiuser system when K users are transmitting to a base station. Denoting the received power of k-th user by P_k, the capacity region includes all the rate vectors $\mathbf{r} = [R_1, R_2, \ldots, R_K]^\top$ which satisfy the condition [9, 70]:

$$\sum_{i=1}^{K} R_k \leq C\Big(\frac{\sum_{i=1}^{K} P_i}{N_0}\Big) \quad \text{bits/sec/Hz}, \tag{6.1}$$

where the function $C(\cdot)$ is defined as $C(x) = \log_2(1+x)$. Like in the previous definitions, N_0 is the variance of additive white Gaussian noise. The term in the right-hand side of (6.1) is the *sum-capacity* of multiuser system.

We refer to [9, Chapter 14] for more details on multiuser information theory.

6.1.2 Capacity region as a polymatroid

Polymatroid structure has been used in some resource allocation problems to obtain greedy optimization algorithms (see e.g. [14]). Later on, Tse and Hanley exploited it to characterize the capacity region of multiaccess systems [71].

Here, we first define a polymatroid structure and then show that the capacity region of a multiuser system can be described by a polymatroid.

Definition 6.1. [12,71] Let $E = \{1, 2, \ldots, K\}$ and $f : 2^E \to \mathbb{R}^+$ be a set function. The polyhedron

$$\mathcal{B}(f) \equiv \{(x_1, \ldots, x_K) : \mathbf{x}(S) \leq f(S) \;\; \forall S \subseteq E, \;\; x_i \geq 0 \;\; \forall i\}, \tag{6.2}$$

where $\mathbf{x}(S) = \sum_{i\in S} x_i$, is a polymatroid if $f(\cdot)$ satisfies the following three conditions:

(1) $f(\emptyset) = 0$ (normalized);

(2) $f(S) \leq f(T)$ if $S \subset T$ (nondecreasing);

(3) $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$ (submodular).

To show that the capacity region is a polymatroid structure, we introduce the function $f_c(S) = C\Big(\sum_{i\in S}\frac{P_i}{N_0}\Big)$, where S is an arbitrary subset of E, and $E = \{1, 2, \ldots, K\}$ is the index set of users. Now, we should prove that the function $f_c(\cdot)$ satisfies the three conditions in Definition 6.1. Conditions (1) and (2) are trivial to check. Let us denote $t_i = \frac{P_i}{N_0}$ and $g(S) = 1 + \sum_{i\in S} t_i$ for an arbitrary $S \subseteq E$. Consider two subsets $S, T \subseteq E$ and, without loosing the generality, assume $g(S) \leq g(T)$. It is easy to see that $g(S\cap T) \leq g(S) \leq g(T) \leq g(S\cup T)$ and $g(S) + g(T) = g(S\cap T) + g(S\cup T)$. Therefore, there exists z such that $0 \leq z \leq g(S) \leq g(T)$, $g(S\cup T) = g(T) + z$ and $g(S\cap T) = g(S) - z$. From this, we have

$$\begin{aligned} g(S\cup T)g(S\cap T) &= \Big(g(T)+z\Big)\Big(g(S)-z\Big) \\ &= g(T)g(S) - \Big(g(T)-g(S)\Big)z - z^2 \\ &\leq g(T)g(S), \end{aligned} \tag{6.3}$$

where the last inequality is a consequence of the fact that $g(T) - g(S)$ and z are both non-negative. Applying $\log_2(\cdot)$ to both sides of (6.3) we get

$$\log_2 g(T) + \log_2 g(S) \geq \log_2 g(S\cup T) + \log_2 g(S\cap T). \tag{6.4}$$

The proof that $f_c(\cdot)$ satisfies Condition (3) is concluded by substituting $f_c(\cdot) = \log_2(g(\cdot))$ in the inequality above.

To gain more insight, we give an example of a function which satisfies conditions (1) and (2) but does not satisfy condition (3). For simplicity, we keep the framework from the proof for capacity function and we define $h(S) = (\sum_{i\in S}\frac{P_i}{N_0})^2$. It is a simple exercise to verify that

$$h(S) + h(T) \leq h(S\cup T) + h(S\cap T), \tag{6.5}$$

where S and T are arbitrary subsets of E. For the sake of completeness, we mention that the set functions which satisfy conditions (1) and (2) and the inequality in (6.5) are called *supermodular* [75].

6.2 Multiple-access techniques and their achievable rates

Successive Interference Cancellation: SIC is known to achieve the sum-capacity of a multiuser system [9,70]. The decoding is done in as many stages as the number of users, where in each stage one user is decoded. Let us denote the decoding order by $\sigma(1) \to \sigma(2) \to \ldots \to \sigma(K)$, where $\sigma(\cdot)$ is a permutation on the set of users E. In the first stage, the receiver decodes the signal of user $\sigma(1)$ treating all the other users as Gaussian noise. This means that the aggregate noise power seen by the first

decoded user is $N_{\sigma(1)} = N_0 + \sum_{j=2}^{K} P_{\sigma(j)}$ which gives the rate $R_{\sigma(1)} = C(\frac{P_{\sigma(1)}}{N_{\sigma(1)}})$. The decoder *cancels* the successfully decoded signal from the composite received signal and continues the procedure by decoding the next user and so on.

It is clear that the achievable rate for a specific user in SIC depends on its decoding order. The rate of user $\sigma(i) \in \{1, 2, \ldots, K\}$ is expressed by the following formula [70]:

$$R_{\sigma(i)} = C\Big(\frac{P_{\sigma(i)}}{N_0 + \sum_{j>i} P_{\sigma(j)}}\Big). \tag{6.6}$$

The $K!$ rate vectors obtained by $K!$ different decoding orders of K users in SIC method, correspond to $K!$ corner points of sum-capacity facet. The other points within sum-capacity facet cannot be achieved by SIC.

Time-sharing and rate-splitting: TS is a method for achieving the other points on sum-capacity facet via sharing the whole transmission time between different decoding orders corresponding to different corners of the facet [70]. Referring to Fig. 6.1, all the points on the sum-capacity facet can be achieved by time sharing between the two SIC points. Assuming that $K = 2$ and $0 < \alpha < 1$, time sharing means that in α fraction of time the SIC decoder first decodes user 2 and in $1 - \alpha$ fraction of time the SIC decoder first decodes user 1.

In general, when there are K users, the sum-capacity facet has $K!$ corner points each corresponding to a decoding order. Then TS means to choose $K!$ positive numbers $\{\alpha_i\}_{i=1}^{K!}$, each for a corner point, such that $\sum_i^{K!} \alpha_i = 1$, and then to allocate α_i fraction of time to the i-th corner point.

Another equivalent way to achieve the points on sum-capacity facet is *rate splitting* [42]. Consider a simple multiaccess system with $K = 2$ users, having powers P_1 and P_2 and rates R_1 and R_2, where $R_1 + R_2 = C(\frac{P_1+P_2}{N_0})$. The idea behind rate-splitting is that either of users or both of them split their powers (and therefore rates) into two or more parts, where each part is treated as a virtual user and is decoded using SIC in the receiver. Suppose that user 1 is transformed into two virtual users $1a$ and $1b$ with powers P_{1a} and $P_{1b} = P - P_{1a}$, but user 2 remains unchanged, and also suppose that the decoding order is $1a \to 2 \to 1b$. Then the rates of three users are $r_{1a} = C(\frac{P_{1a}}{N_0+P_2+P_{1b}})$, $r_b = C(\frac{P_2}{N_0+P_{1b}})$, $r_{1b} = C(\frac{P_{1b}}{N_0})$, and the total rate of user 1 is $r_1 = r_{1a} + r_{1b}$. It is easy to see that $r_1 + r_2 = R_1 + R_2$, which means that the new rate vector (r_1, r_2) is on sum-capacity facet.

Orthogonal Multiple Access: OMA is another modality for accessing the channel by K users. In OMA, user i gets a fraction α_i of degrees of freedom (DOF), where $\sum_i^K \alpha_i = 1$. Note that it is irrelevant for the capacity analysis whether the partitioning in DOF is accross time or frequency. This means that the maximum rate

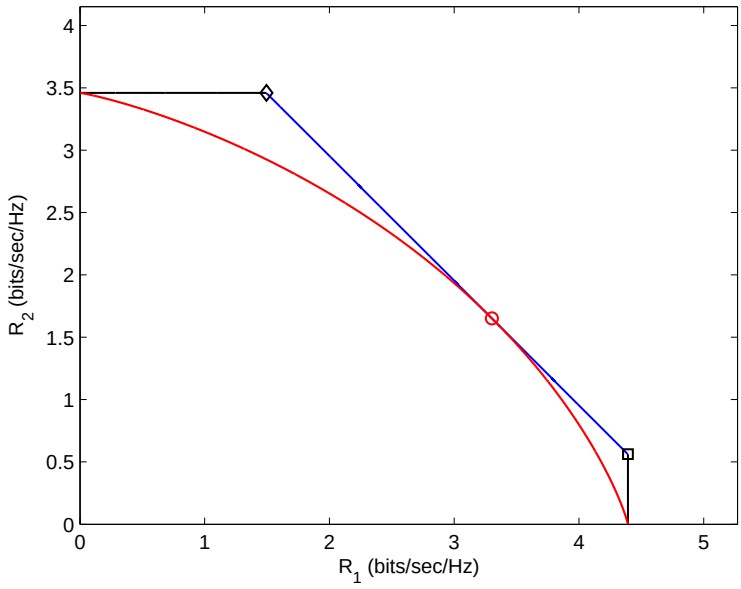

Figure 6.1: Capacity region in the case when $K = 2$. Note that the capacity region includes all the points inside the pentagon area whose sides are the two axes, the two black lines and the blue line. This area corresponds to a polymatroid structure. All points inside this region are reliably achievable. Blue line shows the sum-capacity facet, i.e., the points with maximum achievable sum-rate. The two extreme points of blue line are points achievable by SIC, where the black square is obtained if user 2 is decoded first and black diamond is obtained if user 1 is decoded first. The red curve shows the OMA achievable rates. As it can be seen, OMA curve intersects with the sum-capacity facet only in one point (red circle), namely the point in which the DOF allocated to users are proportional to their received powers. All other points on sum-capacity facet can be achieved by time-sharing of corner points (SIC points) or by using rate splitting. In our settings $\frac{P_1}{N_0} = 20$ and $\frac{P_2}{N_0} = 10$. The figure is adapted from [70].

the user i can achieve is [70]:

$$R_i^{(\mathrm{OMA})} = \alpha_i C\left(\frac{P_i}{\alpha_i N_0}\right) \quad \text{bits/sec/Hz}. \tag{6.7}$$

OMA is suboptimal in rate, meaning that $\sum_{i=1}^{K} R_i^{(\mathrm{OMA})} < C\left(\frac{\sum_{i=1}^{K} P_i}{N_0}\right)$, except for one point: When the amount of DOF allocated to each user is proportional to its received power, i.e. $\alpha_i = \frac{P_i}{\sum_{j=1}^{K} P_j}$.

The achievable rates of above-mentioned multiaccess schemes have been shown

in Fig. 6.1.

6.3 Fairness, efficiency and heterogeneity

It is worth mentioning that, our aim is to improve fairness without sacrificing the overall throughput. Equivalently, we are looking for *a point on sum-capacity facet* which guarantees fairness and, at the same time, satisfies a certain level of performance.

Max-min as a measure of fairness: Among different notions of fairness, we consider the max-min fairness which is defined as follows:

Definition 6.2. [39] A vector of quantities (e.g., rates) is called max-min fair if and only if an increase in any component makes a decrease in at least one other component with smaller or equal value.

Based on optimization of submodular functions [15] a recursive algorithm for finding the time-sharing coefficients of the fairest rate vector (in terms of max-min fairness) inside sum-capacity facet was derived in [39]. We call the method proposed in [39] *fairest TS*.

Asymptotic multiuser efficiency as a measure of performance [72]: The main performance measure in communication systems is the bit-error-rate (BER). Here, we discuss an alternative measure of performance which is well-known in the context of multiuser detection.

The *effective energy* of user k in the presence of background noise N_0, denoted by $e_k(N_0)$, is defined as the energy that this user would require to achieve BER equal to the BER in a single-user Gaussian channel with the same background noise level. It is clear that the effective energy is always upper bounded by the actual energy:

$$e_k(N_0) \leq P_k T_s, \tag{6.8}$$

where T_s is the signal duration.

The *multiuser efficiency* (ME) defined as the ratio between the effective and actual energies, $\frac{e_k(N_0)}{P_k T_s}$, is an alternative measure of performance in multiuser systems. ME depends on the correlation between signature waveforms, background noise and the detector employed. Signature waveform of a user is the waveform which is used for spreading the signals of that user along time axis.

To remove the effect of background noise, which is of less interest in the study of multiuser systems, Verdu defined the *asymptotic multiuser efficiency* (AME) as

$$\eta_k = \lim_{N_0 \to 0} \frac{e_k(N_0)}{P_k T_s}. \tag{6.9}$$

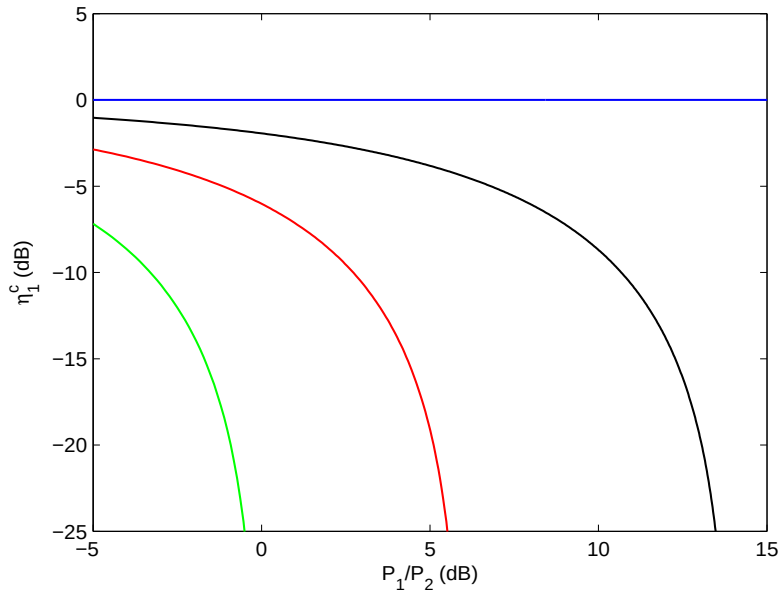

Figure 6.2: AME versus interference-to-signal ratio in a multiuser system with $K = 2$ users (blue: $\rho_{1,2} = 0$, black: $\rho_{1,2} = 0.2$, red: $\rho_{1,2} = 0.5$, green: $\rho_{1,2} = 1$). The figure is a slightly modified variant of [72, Fig. 3.17], where only the case $\rho_{1,2} = 0.2$ was shown.

AME is usually very close to ME except for low signal-to-noise ratios. In fact, AME quantifies the performance when the interferer users are present and the background noise vanishes. It is a function of users' signatures and the detector employed. When the conventional (matched-filter) decoder is used, the AME can be written in the following way

$$\eta_k^c = \max^2\Big\{0, 1 - \sum_{j\neq k} \sqrt{\frac{P_j}{P_k}}|\rho_{j,k}|\Big\}, \tag{6.10}$$

where $\rho_{j,k}$ is the cross-correlation between signature waveforms of users k and j which satisfies $0 \leq \rho_{j,k} \leq 1$. For OMA we have $\rho_{j,k} = 0,\ \forall j \neq k$. In SIC the signals of all users are transmitted and superposed over the same DOF, therefore users' signal are fully correlated, i.e. $\rho_{j,k} = 1,\ \forall j, k$.

Effect of heterogeneity on fairness The fairness of multiaccess methods depends on the disparity of users' received powers. If multiuser system is homogeneous or almost homogeneous, in the sense that the users' received powers are very close, then the OMA method provides higher degree of fairness than SIC. If the system is heterogeneous, i.e. the received powers are very disparate, then SIC (when the stronger user is decoded first) outperforms OMA [70, Chapter 6].

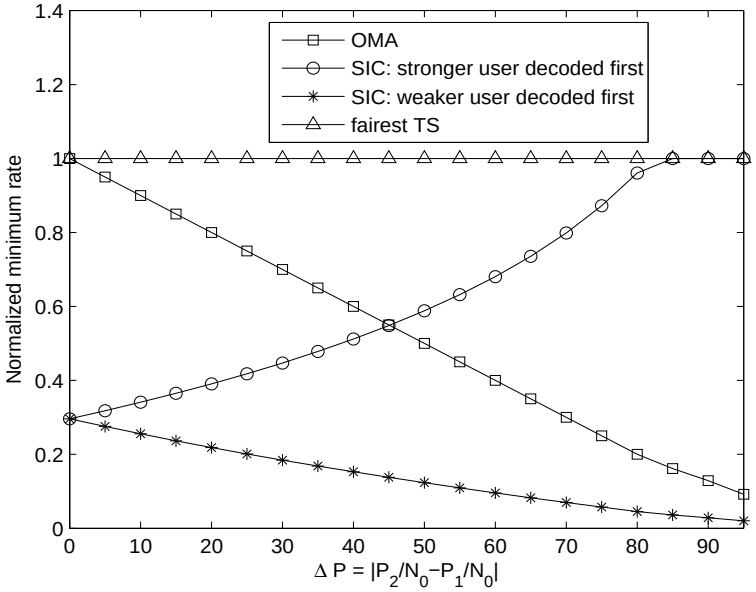

Figure 6.3: Fairness versus heterogeneity in a multiuser system with $K = 2$ users for the following multiaccess methods: OMA, SIC when the stronger user is decoded first, SIC when the weaker user is decoded first, and fairest TS. The normalized minimum rate is considered as fairness measure and the SNR difference $|\frac{P_2}{N_0} - \frac{P_1}{N_0}|$ is considered as heterogeneity measure : the larger is the difference, the more heterogeneous is the network. The total SNR is kept fixed: $\frac{P_1}{N_0} + \frac{P_2}{N_0} = 100$. Normalization has been done with respect to fairest TS.

For SIC, it is known that among all decoding orders, the fairest scenario in terms of max-min (or, in other words, the fairest corner point of sum-capacity facet) happens when in each decoding step the strongest user is decoded [39]. This can be seen from Fig. 6.3 in the particular case when $K = 2$ users.

Effect of heterogeneity on performance In OMA scenario, because users are sending in non-overlapping fractions of DOF, maximum value of AME is always achieved. But the situation is not similar in SIC scenario, where the AME usually shows poor performance especially when the network is homogeneous or the number of users is high.

When the network is heterogeneous and the number of users is small, the situation improves slightly if the decoding order is from stronger users to weaker users. The performance of fairest TS is almost the same as SIC.

Fig. 6.4 presents the average performance of different multiaccess scenarios versus the heterogeneity of network.

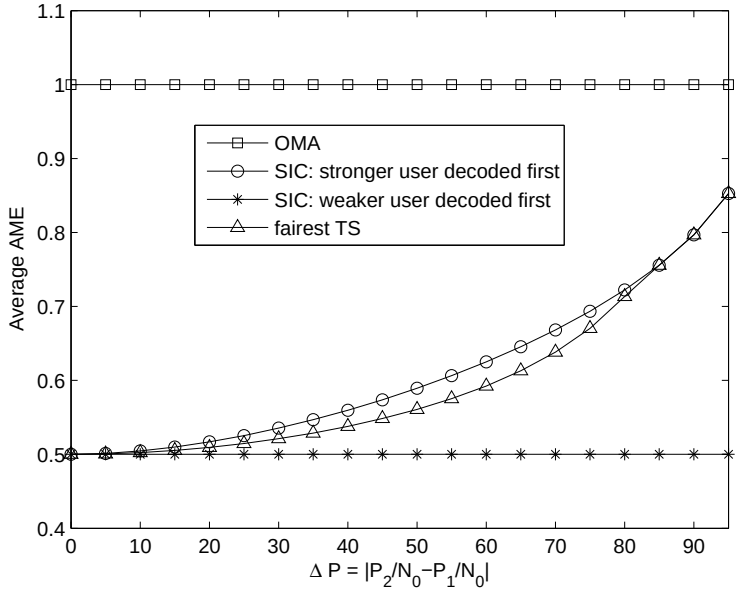

Figure 6.4: Average AME versus heterogeneity in a multiuser system with $K = 2$ users for the following multiaccess methods: OMA, SIC when the stronger user is decoded first, SIC when the weaker user is decoded first, and fairest TS. The total SNR is kept fixed: $\frac{P_1}{N_0} + \frac{P_2}{N_0} = 100$.

New method for achieving a trade-off between fairness and AME Based on the above-mentioned observations, in [**P7**] we proposed a new multiaccess scheme which exploits and combines the beneficial aspect of both OMA and SIC methods such that to provide a reasonable fairness over both homogeneous and heterogeneous networks. At the same time, the novel method is superior in performance when compared with fairest TS [39], which represents the state-of-the-art of fairness in multiuser systems with polymatroid capacity regions.

In [**P7**], the key idea is to partition the network into (almost) homogeneous subnetworks such that the users within each subnetwork employ OMA, which provides high degree of fairness and ideal performance within each subnetwork, and then utilize SIC across subnetworks which have disparate total powers.

Given that the number of subnetworks is $T \in \{1, \ldots, K\}$, the newly-proposed scheme is equivalent to partition the K users into T ordered groups. Note that the order matters because it corresponds to the order in which the groups are decoded. Remark for $T = 1$ that the grouping method is the same with OMA. Moreover, the grouping method is identical with SIC for $T = K$. Similarly to conventional SIC, the max-min rate achieved in this case depends on the order in which the groups are decoded. We consider the family of all *ordered partitions* of the K users

into T non-empty groups. Then we pick-up the ordered partition for which the minimum rate is maximized, and we name it $\mathrm{BORG}_{K/T}$ (*basic ordered grouping of K users into T groups*). Conventionally, $\mathrm{BORG}_{K/1}$ coincides with OMA, and we write $\mathrm{BORG}_{K/1} \equiv \mathrm{OMA}$. Obviously, $\mathrm{BORG}_{K/K} \equiv \mathrm{SIC}$.

Furthermore, one can select again from $\mathrm{BORG}_{K/1}, \mathrm{BORG}_{K/2}, \ldots, \mathrm{BORG}_{K/K}$ the ordered partition which maximizes the minimum rate. The new selection is dubbed BORG_K^*. Remark that the rate-vector which corresponds to BORG_K^* is not necessarily the same with the max-min fair rate-vector that was defined in Section 6.3. However, BORG_K^* is guaranteed to be max-min fair among all possible user groupings for which the sum-capacity is achieved.

In [**P7**], we investigated how the fairness can be evaluated for OMA, SIC and BORG. In this context we demonstrated for $\mathrm{BORG}_{K/T}$ a fundamental property, which allowed us to introduce a low-complexity search method for choosing $\mathrm{BORG}_{K/T}$ from all ordered partitions of K users into T groups [**P7**, Theorem 1].

We gave also a geometrical interpretation for the rate-vector yield by our algorithm. More exactly, we pointed out the connections between the outcome of the proposed method and the polymatroid structure of the capacity region.

We compared the proposed method against multiaccess methods discussed in previous section by considering four different heterogeneity models. Simulation results show that the novel method provides a good trade-off between fairness and performance [**P7**, Figs. 2-5].

Chapter 7

Summary of publications and author's contribution

This thesis consists of seven publications including four published journal papers ([**P1**, **P3**, **P4**, **P7**]) and three conference papers ([**P2**, **P5**, **P6**]).

In this chapter, we first summarize each paper and then explain the author's contribution to them.

7.1 Summary of publications

Publication [**P1**] addresses the problem of variable selection in Gaussian linear regression using NML. In this article, we extend the Rissanen methodology for computing the parametric complexity and discussed two particular cases, namely the rhomboidal and the ellipsoidal constraints. The new findings are used to derive four NML-based criteria. For three of them which have been already introduced in the previous literature, we provide a rigorous analysis. We also compare them against five state-of-the-art selection rules by conducting Monte Carlo simulations for families of models commonly used in signal processing. Additionally, for the eight criteria which are tested in [**P1**], we report results on their predictive capabilities for real-world data sets.

In [**P2**], we investigate the problem of selection between nested models using the SC. For better understanding of the properties of the SC, we relate it to the GLRT. We also compare SC with BIC.

Publication [**P3**] studies the use of information theoretic criteria (ITC) for selecting the order of AR models when the model parameters are estimated by forgetting factor LS algorithms. Because the ITC are derived under the strong assumption that the measured signals are stationary, it is not straightforward to employ them in combination with the forgetting factor LS algorithms. In the previous literature, the attempts for solving the problem were focused on the AIC, the BIC and the predic-

tive least squares (PLS). In connection with PLS, an ad hoc criterion called SRM was also introduced. In [**P3**], we modify the predictive densities criterion (PDC) and the SNML criterion such that to be compatible with the forgetting factor least-squares algorithms. Additionally, we provide rigorous proofs concerning the asymptotic approximations of four modified ITC, namely PLS, SRM, PDC and SNML. Then, the four criteria are compared by simulations with the modified variants of BIC and AIC.

In [**P4**], we derive the ODD detector for the classical linear model. In this framework, we provide answers to the the following problems that have not been previously investigated in the literature: (i) the relationship between ODD and GLRT; (ii) the connection between ODD and ITC applied in model selection. We point out the strengths and the weaknesses of the ODD method in detecting subspace signals in broadband noise. Effects of subspace interference are also evaluated. All the derivations in [**P4**] are based on the assumption that the level of Gaussian noise is *known*.

Publication [**P5**] consists of two main parts. The first one is a preliminary version of [**P4**], and the second one is devoted to the extension of the results to the case when the noise variance is *unknown*. The solution provided in [**P4**] was explained with more details in Chapter 4 of this thesis.

In [68], it was shown how the periodogram can be smoothed by thresholding the estimated cepstral coefficients. In [**P6**], we use the KSF to derive a new criterion for selecting the threshold. For the numerical examples taken from the previous literature, the KSF selection rule compares favorably with the existing schemes. Some possible extensions were discussed in Chapter 5 of the thesis.

Publication [**P7**] studies a different problem. In this paper, a novel approach for improving fairness over (possibly heterogeneous) multiaccess channels is introduced. It is known that the Orthogonal Multiple-Access (OMA) guarantees for homogeneous networks, where all users have almost the same received power, a higher degree of fairness (in rate) than that provided by Successive Interference Cancellation (SIC). The situation changes in heterogeneous networks, where the received powers are very disparate, and SIC becomes superior to OMA. In [**P7**], we propose to partition the network into (almost) homogeneous subnetworks such that the users within each subnetwork employ OMA and SIC is utilized across subnetworks. The newly-proposed scheme is equivalent to partition the users into ordered groups. The main contribution is a practical algorithm for finding the ordered partition that maximizes the minimum rate. We also give a geometrical interpretation for the rate-vector yield by our algorithm. Experimental results show that the proposed strategy leads to a good trade-off between fairness and the asymptotic multiuser efficiency.

7.2 Author's contribution

The author's contribution to [**P1**] is threefold: (I) He proposed the use of rhomboidal constraint and worked out the criterion in Section 2.3; (II) He also contributed to the proofs of Proposition 3.1, Proposition 3.2, Proposition 3.3 as well as Lemma 3.1 and Lemma 3.2; (III) He implemented part of the Matlab code used in numerical examples within Section 4.

For [**P2**], the author brought the idea of using SC for the detection of interferer in multiuser communication systems. However, it was not possible to explain the outcome of the experimental results obtained by the author in this context. This made it necessary to perform the theoretical analysis from [**P2**]. The author has also assisted in the derivations of all results within the paper.

For [**P3**], the author implemented a preliminary version of the Matlab code used in simulation examples. He also assisted in the derivations of the main results.

Regarding [**P4**], the author contributed substantially to all the theoretical results within the paper. He also generated all the figures included.

The author's contribution to [**P5**] can be described with the same words as in the case of [**P4**].

For the preparation of [**P6**], the author assisted in all the derivations within the paper as well as in Matlab implementation for the experimental results. He also improved the final form of the publication.

In [**P7**], the author has had the main contribution in all steps of the paper elaboration from proposing the key idea to implementation and final writing.

Chapter 8

Conclusions

The main focus in this research was on the use of information theoretic techniques for signal modelling and hypothesis testing. The following points are concluded based on the publications, which are the major outcome of this work.

ODD-based hypothesis testing: We investigated the use of the ODD detector for the linear model by emphasizing the strengths and the weaknesses of the method in [**P4**] and [**P5**]. In our work, we have obtained the expressions of P_{D} and P_{FA} for ODD, and we used them to compare ODD and GLRT by numerical examples. It is worth mentioning that the GLRT is invariant to a "natural" class of transformations, whereas the ODD detector does not share the same invariances. Furthermore, we demonstrated in [**P5**] how the ODD methodology can be extended to accommodate models with nuisance parameters.

NML in Gaussian linear regression: Because the parametric complexity is not finite, the only possibility for obtaining NML-based selection rules is to constrain the data space. Even if this was recognized more than one decade ago, the solutions proposed so far are only punctual results which treat some particular constraints. In [**P1**], we have introduced a general methodology for addressing the problem. Based on the new findings, we demonstrated how the rhomboidal constraint yields a new NML-based formula. Additionally, we used the ellipsoidal constraint to re-derive three criteria that have been introduced in the previous literature.

Comparison of SC with other rules of selecting between two nested models: In [**P2**], we investigated the relationship between SC and GLRT. This analysis has shown the importance of the hyper-parameters within SC-formula. The comparison between SC and BIC revealed the robustness of SC for families of models commonly used in signal processing.

ITC for AR order selection in the presence of nonstationarity: Transforming the ITC which have been derived for the stationary case such that to become compatible with the forgetting factor least-squares algorithms is not a trivial task. In [**P3**], we focused on five ITC which can be seen as embodiments of the MDL principle. Additionally, a modified variant of AIC was considered. For decomposing each MDL-based criterion into the goodness-of-fit term and the penalty term, we resorted to an asymptotic analysis. Both the theoretical and experimental results led to the conclusion that the modified SNML performs well in this type of application.

Cepstral nulling via KSF: In [**P6**], we focused on a thresholding-based method for TV-reduction and the main contribution was to show how the KSF can be used to derive a criterion for selecting the threshold μ. In the framework of cepstral analysis, we compared the newly proposed selection rule with other two schemes which are considered state-of-the-art. The first one chooses μ with a carefully designed UMPUT, while the second one relies on BIC for the selection of μ. It was shown experimentally that KSF is much better than BIC when the signal is broadband MA with a small dynamic range of the log-spectrum. It was also noticed that, for all numerical examples, the KSF-based criterion and UMPUT have similar performance. However, UMPUT requires a priori information on the type of the observed signal, whereas the KSF thresholding is fully automatic.

Subnetwork selection for improving fairness in multiaccess communications: In [**P7**], we investigated how OMA and SIC can be combined to improve fairness in Gaussian wireless networks. The newly-proposed method divides the network into (almost) homogeneous subnetworks such that the users within each subnetwork employ OMA, and SIC is utilized across subnetworks. Equivalently, the K users are partitioned into T ordered groups. The main theoretical result which we proved for any $T \in \{2, \ldots, K-1\}$, shows that the ordered partition which maximizes the minimum rate can be found with a low-complexity algorithm. Moreover, it was demonstrated experimentally that the user grouping strategy guarantees a good trade-off between fairness and the asymptotic multiuser efficiency.

Bibliography

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[2] E. Artin, *The Gamma Function*. Holt, Rinehart and Winston, Inc., 1964.

[3] K. Atteson, "The asymptotic redundancy of Bayes rules for Markov chains," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 2104–2109, 1999.

[4] V. Balasubramanian, "Statistical inference, Occams razor, and statistical mechanics on the space of probability distributions," *Neural Computation*, vol. 9, no. 2, pp. 349–368, 1997.

[5] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2743 –2760, 1998.

[6] D. Bradley, "Representations of Catalan's constant," Dep. of Mathematics and Statistics, Univ. of Maine, USA, http://germain.umemat.maine.edu/faculty/bradley/papers/c1.ps, Tech. Rep., 2001.

[7] G. J. Chaitin, "Algorithmic information theory," *IBM Journal of Research and Development*, vol. 21, no. 4, pp. 350–359, 1977.

[8] B. C. Clarke and A. R. Barron, "Information theoretic asymptotics of Bayes methods," *IEEE Transactions on Information Theory*, vol. 36, no. 03, pp. 453–471, 1990.

[9] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & sons, 2006.

[10] L. Davisson, "Universal noiseless coding," *Information Theory, IEEE Transactions on*, vol. 19, no. 6, pp. 783 – 795, nov 1973.

[11] S. de Rooij and P. Grünwald, "An empirical study of mdl model selection with infinite parametric complexity," *Journal of Mathematical Psychology*, vol. 50, no. 2, pp. 180–192, 2006.

[12] J. Edmonds, "Submodular functions, matroids and certain polyedra," in *Proc. Calgary Int. Conf. Combinatorial Structures and Applications*, Calgary, Canada, Jun. 1969, pp. 69–87.

[13] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 162–176, 1999.

[14] A. Federgruen and H. Groenevelt, "The greedy procedure for resource allocation problems: Necessary and sufficient conditions for optimality," *Operations Research*, vol. 34, no. 6, pp. 909–918, 1986.

[15] S. Fujishige, *Submodular Functions and Optimization*. Elsevier, 2005.

[16] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4165 –4174, 2009.

[17] C. D. Giurcăneanu, "On the use of the Fisher information in computing the stochastic complexity," in *Proc. Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, Tampere, Finland, Aug. 18-20 2008.

[18] C. D. Giurcăneanu, D. Mihalache, M. Omarjee, and M. Tetiva, "On the stochastic complexity for order-1 Markov chains and the Catalan constant," in *Proc. of the 16th International Conference on Control Systems and Computer Science*, Bucharest, Romania, May 2007, vol. 3, pp. 114–120.

[19] C. D. Giurcăneanu and J. Rissanen, "Estimation of AR and ARMA models by stochastic complexity," in *Time Series and Related Topics.*, H.-C. Ho, C.-K. Ing, and T. L. Lai, Eds. Institute of Mathematical Statistics Lecture Notes-Monograph Series, 2006, vol. 52, pp. 48–59.

[20] E. Gudmundson, N. Sandgren, and P. Stoica, "Automatic smoothing of periodograms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, 2006*, vol. 3, pp. III: 504–507.

[21] E. J. Hannan and D. F. Nicholls, "The estimation of the prediction error variance," *Journal of the American Statistical Association*, vol. 72, no. 360, pp. 834–840, 1977.

[22] M. H. Hansen and B. Yu, "Model selection and the principle of Minimum Description Length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[23] ——, "Minimum description length model selection criteria for generalized linear models," *Lecture Notes–Monograph Series, Vol. 40, Statistics and Science: A Festschrift for Terry Speed*, pp. 145–163, 2003.

[24] A. Hanson and P. C.-W. Fu, "Applications of MDL to selected families of models," in *Advances in Minimum Description Length: Theory and Applications*, P. Grünwald, I. Myung, and M. Pitt, Eds. MIT Press, 2005, ch. 5, pp. 125–150.

[25] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 1996.

[26] S. Kay, "Conditional model order estimation," *IEEE Transactions on Signal Processing*, vol. 49, no. 9, pp. 1910–1917, 2001.

[27] ——, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

[28] ——, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, 1998.

[29] A. N. Kolmogorov, "Three approaches to the definition of the quantity of information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 3–11, 1965.

[30] G. Korodi and I. Tabus, "Normalized maximum likelihood model of order-1 for the compression of DNA sequences," in *Proc. of Data Compression Conference, DCC '07*, Snowbird, UT, USA, March 2007, pp. 33–42.

[31] S. Kotz and S. Nadarajah, *Multivariate t Distributions and Their Applications*. Cambridge University Press, 2004.

[32] A. M. Kshirsagar, "Some extensions of the multivariate t-distribution and the multivariate generalization of the distribution of the regression coefficient," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 57, no. 01, pp. 80–85, 1961.

[33] R. Lai, T. Lee, R. Wong, and F. Yao, "Nonparametric cepstrum estimation via optimal risk smoothing," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1507 –1514, 2010.

[34] A. D. Lanterman, "Hypothesis testing for poisson vs. geometric distributions using stochastic complexity," in *Advances in Minimum Description Length*, P. Grünwald, I. Myung, and M. Pitt, Eds. MIT Press, 2005, ch. 4, pp. 99–124.

[35] E. L. Lehmann, *Testing Statistical Hypotheses*. John Wiley & Sons, 1970.

[36] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

[37] E. Liski and A. Liski, "Minimum Description Length model selection in Gaussian regression under data constraints," in *Statistical Inference, Econometric Analysis and Matrix Algebra, Festschrift in Honour of Götz Trenkler*, B. Schipp, Ed. Springer, 2009.

[38] E. Liski, "Normalized ML and the MDL principle for variable selection in linear regression," in *Festschrift for Tarmo Pukkila on his 60th birthday*, E. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. Styan, Eds. Univ. of Tampere, 2006, pp. 159–172.

[39] M. Maddah-Ali, A. Mobasher, and A. Khandani, "Fairness in multiuser systems with polymatroid capacity region," *"IEEE Transactions on Information Theory"*, vol. 55, no. 5, pp. 2128–2138, 2009.

[40] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. Academic Press, 1979.

[41] D. Mitrinovic and P.M.Vasic, *Analytic Inequalities*. Springer Verlag, 1970.

[42] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 364 –375, 1996.

[43] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[44] ——, "Stochastic complexity," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 49, no. 3, pp. 223–239, 1987.

[45] ——, "Fisher information and stochastic complexity," *IEEE Transactions on Information Theory*, vol. 42, no. 1, pp. 40 –47, 1996.

[46] ——, "Hypothesis selection and testing by the MDL principle," *Computer Journal*, vol. 42, no. 4, pp. 260–269, 1999.

[47] ——, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[48] ——, *Information and Complexity in Statistical Modeling*. Springer Verlag, 2007.

[49] ——, "Model selection and testing by the MDL principle," in *Information Theory and Statistical Learning*, F. Emmert-Streib and M. Dehmer, Eds. Springer, 2009.

[50] J. Rissanen and T. Roos, "Conditional NML universal models," in *Proc. Information Theory and Applications Workshop (ITA-07)*, San Diego, USA, Jan. 29-Feb. 2 2007, pp. 337–341.

[51] J. Rissanen, T. Roos, and P. Myllymäki, "Model selection by sequentially normalized least squares," *Journal of Multivariate Analysis*, vol. 101, no. 4, 2010.

[52] T. Roos, "Monte Carlo estimation of minimax regret with an application to MDL model selection," in *Proc. of IEEE Information Theory Workshop*, Porto, Portugal, May 2008.

[53] T. Roos, P. Myllymäki, and J. Rissanen, "MDL denoising revisited," *IEEE Transactions on Signal Processing*, vol. 57, no. 9, pp. 3347–3360, 2009.

[54] T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," in *Proc. Workshop on Information Theoretic Methods in Science and Engineering (WITMSE-08)*, Tampere, Finland, Aug. 18-20 2008.

[55] L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 2146–2157, 1994.

[56] D. Schmidt and E. Makalic, "MML invariant linear regression," in *Proceedings Volume 5866 of Lecture Notes in Computer Science*, A. Nicholson and X. Li, Eds. Springer, 2009.

[57] ——, "Estimating the order of an autoregressive model using normalized maximum likelihood," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 479 –487, 2011.

[58] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[59] G. Seber and A. Lee, *Linear Regression Analysis*. Wiley-Interscience, 2003.

[60] L. Shampine, "Vectorized adaptive quadrature in Matlab," *Journal of Computational and Applied Mathematics*, vol. 211, no. 2, pp. 131–140, 2008.

[61] Y. Shtarkov, "Universal sequential coding of single messages," *Prob. Inform. Transmission*, vol. 23, no. 3, pp. 175–186, 1987.

[62] T. Söderström, "On model structure testing in system identification," *Int. J. Control*, vol. 26, no. 1, pp. 1–18, 1977.

[63] R. J. Solomonoff, "A formal theory of inductive inference, part I," *Information and Control*, vol. 7, no. 1, pp. 1–22, 1964.

[64] ——, "A formal theory of inductive inference, part II," *Information and Control*, vol. 7, no. 2, pp. 224–254, 1964.

[65] F. E. Steffens, "Power of bivariate studentized maximum and minimum modulus tests," *Journal of the American Statistical Association*, vol. 65, no. 332, pp. 1639–1644, 1970.

[66] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Prentice Hall, 2005.

[67] P. Stoica and N. Sandgren, "Smoothed nonparametric spectral estimation via cepstrum thresholding - Introduction of a method for smoothed nonparametric spectral estimation," *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 34–45, 2006.

[68] ——, "Total-variance reduction via thresholding: Application to cepstral analysis," *IEEE Transactions on Signal Processing*, vol. 55, no. 1, pp. 66–72, 2007.

[69] P. Stoica, Y. Selen, and J. Li, "On information criteria and the generalized likelihood ratio test of model order selection," *IEEE Signal Processing Letters*, vol. 11, no. 10, pp. 794–797, 2004.

[70] D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*. Cambridge University Press, 2005.

[71] D. Tse and S. Hanly, "Multiaccess fading channels-Part I: Polymatroid structure, optimal resource allocation and throughput capacities," *"IEEE Transactions on Information Theory"*, vol. 44, no. 7, pp. 2796–2815, 1998.

[72] S. Verdu, *Multiuser Detection*. Cambridge University Press, 1998.

[73] N. Vereshchagin and P. Vitanyi, "Kolmogorov's structure functions and model selection," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3265 – 3290, 2004.

[74] P. Vitanyi, "Algorithmic statistics and Kolmogorov's structure function," in *Advances in Minimum Description Length*, P. Grünwald, I. Myung, and M. Pitt, Eds. MIT Press, 2005, ch. 6, pp. 151–174.

[75] X. Zhang, J. Chen, S. Wicker, and T. Berger, *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 2246 –2254, 2007.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-2619-0
ISSN 1459-2045