Contribution to MIMO-OFDM Mobile Radio Channel Estimation Using SVR

Anis Charrada

Ministry of Higher Education, Scientific Research and Information and Communication Technologies El Manar university National Engineering School of Tunis

Thesis Degree Telecommunications

Contribution to MIMO-OFDM Mobile Radio Channel Estimation Using SVR

Carried out by: Anis CHARRADA

defended on January 13th, 2015 in front of the Jury

Pr. Ammar BOUALLEGUE	President
Pr. Salem HASNAOUI	Reviewer
Pr. Iyad DAYOUB	Reviewer
Pr. Abdelaziz SAMET	Supervisor
Pr. Rabeh ATTIA	Examiner

SERCom Laboratory / Tunisia Polytechnic School, Carthage University

Academic Year: 2014-2015

Acknowledgments

I would like to express my very special gratitude to my supervisor Professor Abdelaziz Samet, whose knowledge, understanding, expertise and patience added considerably to my graduate experience.

I would like to show my gratitude to my thesis reviewers: Prof. Iyad Dayoub and Prof. Salem Hasnaoui, and to the jury members: Prof. Ammar Bouallegue and Prof. Rabeh Attia for evaluating my Ph.D. work.

I would like also to thank Dr. Bessem Ben Salah and Dr. Ines Boussnina for their guidance during the research period I spent in SERCOM Laboratory at Tunisia polytechnic School.

I would like to express my thanks to my dear parents for the support they provided me through my life.

I am grateful to my wife for her encouragement all the times.

Finally, I must acknowledge my best friends Aymen Omri for their great support.

Abstract

The performance of wireless communication systems is largely governed by the wireless channel environment. Usually, transmission that are carried out on mobile radio channel are selective in time and frequency. To overcome the channel selectivity and allow high transmission data rate, new approaches for synchronization, equalization and channel estimation are needed.

The use of Support Vector Machines (SVMs) has shown several advantages in regression, prediction and estimation over some of the classical approaches due to its improved generalization capabilities. Moreover, the introduction of complex algebra in the SVM formulation can provide us with a more natural and flexible framework when dealing with complex symbols and constellations.

The current thesis work focuses on the study and development of efficient channel estimation algorithms based on complex Support Vector Machines Regression (SVR) that are specifically adapted to pilot-aided OFDM (Orthogonal Frequency Division Multiplexing) structure and applied to Long Term Evolution (LTE) downlink system. The mathematical model of the LTE mobile radio channel is described and simulated for various scenarios based on 3GPP specifications. According to this model, a nonlinear complex SVR is proposed for SISO-OFDM system. The principle of the proposed kernel-based learning algorithm is to exploit the information provided by the reference signals to estimate the channel frequency response. The proposed approach is based on two separate phases: learning phase and estimation phase. In learning phase, we estimate first the subchannels pilot symbols and in estimation phase, frequency responses of data subchannels are determined by means of SVM interpolation mechanism. In addition, a nonlinear complex Multiple Support Vector Machines Regression (M-SVR) algorithm adapted to MIMO (Multiple Input Multiple Output) architecture is proposed to estimate the multipath fading channel in MIMO-OFDM system with both STBC (Space Time Bloc Coding) and V-BLAST (Vertical Bell Laboratories layered Space Time) schemes.

The feasibility of our approaches is confirmed by computer simulation results achieved for LTE downlink model. This experiments allow to analyze the performance of the SVR technique and the suitability of the ε-Huber cost function in the mobile radio multipath fading channel presenting non-Gaussian impulsive noise interfering with OFDM reference symbols under high mobility conditions.

Keywords: SVM, Nonlinear complex SVR, kernels, OFDM, MIMO, LTE.

Résumé

La performance des systèmes de communication sans fils dépend largement des caractéristiques du canal radio mobile. Généralement, la plupart des transmissions sont réalisés sur des canaux sans fils sélectifs en temps et en fréquence. Afin de surmonter la sélectivité du canal et permettre un débit de transmission de données élevé, de nouvelles approches pour la synchronisation, l'égalisation et l'estimation de canaux sont nécessaires. L'utilisation des machines à vecteur support (SVM) a montré plusieurs avantages dans les domaines de régression, estimation et prédiction par rapport à certaines approches classiques grâce à ses capacités de généralisation. En outre, l'introduction de l'algèbre complexe dans la formulation des SVMs a fourni un cadre plus souple et naturel lorsqu'on traite des constellations et symboles complexes.

Les travaux de recherche introduits dans cette thèse portent sur l'étude et le développement des algorithmes d'estimation de canal efficaces et robustes basés sur les machines à vecteur support pour la régression (SVR) particulièrement adaptés à la structure OFDM (Orthogonal Frequency Division Multiplexing) avec des symboles pilotes. Ces algorithmes seront ensuite appliqués à un système LTE (Long Term Evolution). Le modèle mathématique du canal radio mobile LTE est décrit et simulé pour différents scénarios basés sur les spécifications 3GPP. Selon ce modèle, l'estimateur SVR non linéaire complexe est proposé pour le système SISO-OFDM. Le principe de cet algorithme d'apprentissage basé sur les fonctions noyaux est d'exploiter les informations fournies par les symboles pilotes pour estimer la réponse fréquentielle du canal. L'approche proposée repose sur deux phases distinctes: la phase d'apprentissage et la phase d'estimation. Dans la phase d'apprentissage, nous estimons d'abord les sous-canaux des symboles pilotes, puis dans la phase d'estimation, les réponses fréquentielles des sous-canaux de données seront déterminées par le mécanisme d'interpolation SVM. En outre, un algorithme non linéaire complexe basé sur les machines à vecteur support pour la régression multiple (M-SVR) adapté à l'architecture MIMO (Multiple Input Multiple Output) est proposé pour estimer le canal dans les systèmes MIMO-OFDM exploitant les techniques STBC (Space Time Bloc Coding) et V-BLAST (Vertical Bell Laboratories layered Space Time). La faisabilité de nos approches est assistée par des résultats de simulation obtenus pour le modèle LTE. Ces expériences permettent d'analyser les performances de la technique SVR et la pertinence de la fonction coût ε-Huber dans un canal radio mobile à trajets multiples présentant un bruit impulsionnel non-Gaussian interférant avec les symboles pilots sous des conditions de haute mobilité.
Mots-clés: SVM, SVR non linéaire complexe, fonctions noyaux, OFDM, MIMO, LTE.

List of Abbreviations

3GPP : Third Generation Partnership Project,

ANN : Artificial Neural Networks,

AWGN : Additive White Gaussian Noise,

BER : Bit Error Rate,

CP : Cyclic Prefix,

DFT : Discrete Fourier Transform,

DL : Down Link,

eNB : evolved Node B (base station),

EGC : Equal Gain Combining,

EPA : Extended Pedestrian A,

ERM : Empirical Risk Minimization,

ETU : Extended Typical Urban,

EVA : Extended Vehicular A,

FFT: Fast Fourier Transform,

ICI: Inter Carrier Interference,

IDFT: Inverse Discrete Fourier Transform,

IFFT: Inverse Fast Fourier Transform,

ISI: Inter Symbol Interference,

KKT : Karush-Kuhn-Tucker,

LLM : Linear Learning Machines,

LS: Least Squares,

LTE : Long Term Evolution,

LSVC : Linear Support Vector Classification,

LSVM : Linear Support Vector Machines,

LSVR : Linear Support Vector Regression,

MIMO : Multiple-Input Multiple-Output,

ML : Maximum Likelihood,

MMSE : Minimum Mean Square Error,

MRC : Maximum Ratio Combining,

MSE : Mean Square Error,

OFDM : Orthogonal Frequency Division Multiplexing,

OSIC : Ordered Successive Interference Cancellation,

PDCCH : Physical Downlink Control Channel,

PDSCH : Physical Downlink Shared Channel,

pdf : Probability density function,

QAM : Quadrature Amplitude Modulation,

RKHS : Reproducing Kernel Hilbert Spaces,

SC : Selective Combining,

SDM : Space Division Multiplexing,

SISO : Single-Input Single-Output,

SIMO : Single-Input Multiple-Output,

SLT : Statistical Learning Theory,

SIR : Signal-to-Impulse Ratio,

SNR : Signal-to-Noise Ratio,

SRM : Structural Risk Minimization,

STBC : Space Time Block Codes,

SVC : Support Vector Classification,

SVM : Support Vector Machines,

SVR : Support Vector Regression,

UE : User Equipment,

V-BLAST : Vertical Bell Laboratories layered Space Time,

VC : Vapnik Chervonenkis,

WiMAX : Worldwide Interoperability for Microwave Access,

WLAN : Wireless Local Area Network,

WSSUS : Wide-Sense-Stationary-Uncorrelated-Scattering,

ZP : Zero Padding,

ZF : Zero Forcing.

List of Tables

List of Figures

Contents

Introduction

Wireless communications have evolved very speedily. The rapid growth in the development of wireless standards and different global technologies, the demand on the new better quality, high data rates as well as lower cost services are the major motivations for the evolution in the wireless communications.

The presence of interference, fading and mobility makes the design of wireless communication systems challenging. The appropriate design focusing on the reliability of the connection needs to mitigate fading and multipath effects. Modern wireless system design focus on the spectral efficiency gain from the rich multipath environment by means of utilizing spatial diversity through the Multiple-Input Multiple-Output (MIMO) communication systems. While high spectral efficiency can be obtained through spatial multiplexing, many other MIMO system benefits such as improved signal coverage and quality can be achieved via space time coding.

Multicarrier modulation such as Orthogonal Frequency Division Multiplexing (OFDM) is at present the most famous technology for spectrum efficient transmission, since it is enhancing system capacity and mitigating intersymbol interference (ISI). Furthermore, it facilitates using simple equalization.

By combining MIMO system with OFDM technique, we can satisfy the desired system requirements such as high peak data rates, high spectral efficiency as well as reliable transmission. Multicarrier wireless transmission has been already standardized in 3GPP Long Term Evolution (LTE) which is the key transmission technology for the future 4G wireless communication networks. By implementing this new technique in the context of mobile broadband transmission, new approaches for time and frequency synchronization, channel estimation and equalization are required.

This thesis work focuses on the study and development of new efficient mobile radio channel estimation technique based on Support Vector Machines Regression (SVR) applied to LTE systems.

In carried out literature, there has been an important activity in the theoretical development and applications of Support Vector Machines (SVMs) since 1990s [1, 2, 3, 4, 5]. The theory of SVMs is based on the combination of statistical learning theory, optimization theory, kernel theory and algorithmic. So far, machine learning has largely been affected to solving problems relating to data mining [6], biomedical problems [7], biophysical parameters estimation [8, 9], hyperspectral remote sensing images [10, 11, 12, 13, 14, 15, 16, 17], text categorization [18], magnetic resonance imaging [19, 20], time series processing [21, 22, 23], speech recognition [24, 25], linear signal processing

[26] and image processing [27, 28, 29, 30, 31, 32, 33].

Recently, support vector machines have successfully been applied to wireless communication problems [34, 35, 36], notably channel equalization [37], spread spectrum receiver design [38, 39, 40], adaptive modulation and coding [41, 42, 43] and direction of arrival estimation [44]. The learning abilities of support vectors can also be applied to solve complex optimization problems in electromagnetics in the area of microwave, radars and antennas [45].

To the best of our knowledge, nonlinear complex support vector machines regression (SVR) using kernel functions in the Reproducing Kernel Hilbert Space (RKHS) has not been studied yet in OFDM systems with comb-type pilot structure for fast fading channel. In addition, there is no existing work that deals with the channel estimation for real LTE downlink system using nonlinear complex SVR.

The aim of our work is to introduce the subject of complex SVR in its nonlinear form and then to show how it can be applied to LTE channel estimation in SISO-OFDM and MIMO-OFDM systems with the presence of nonlinear impulsive noise under high mobility conditions.

The outline of the thesis are as follows:

In the first chapter, we provide a historical overview on support vector machines. We also present the theories for which SVMs are based on, such that linear learning theory, generalization theory, optimization theory and kernel theory. Then, we give a comparison between SVMs and neural networks and precisely we outline the particular advantages of SVMs.

The second chapter is devoted to the study of support vector machines. In fact, SVMs are a system for efficiently training the linear learning machines in the kernel-induced feature space, while respecting the theory of generalization and statistical learning and exploiting the optimization theory. The linear support vector machines for classification and regression are described in details. The kernel functions necessary to project the nonlinear data in the original space into a high dimensional space where the data become represented linearly are also presented. Finally, the nonlinear support vector machines for classification and regression based on these kernel functions are described.

In the third chapter, multipath propagation characteristics of mobile radio channel for LTE communication system are analyzed and investigated. A mathematical framework of multipath Rayleigh channel is developed. An overview of the nonlinear impulsive noise which can frequently occur in communication systems is also presented. In order to evaluate the performance of the developed channel estimation techniques for LTE system, the realized simulations are based on several propagation scenarios with different mobile speed.

In the fourth chapter, we propose a complex robust support vector regression (SVR) formulation particularly adopted to a pilot-based OFDM signal for SISO system in the presence of nonlinear impulsive noise. The feasibility of this technique is validated by computer simulation results obtained for LTE channel model. We also introduce some traditional channel estimation techniques for performance evaluation purposes with the nonlinear complex SVR-OFDM approach.

The fifth chapter introduces an efficient nonlinear complex Multiple Support Vector Machines Regression (M-SVR) approach to track the frequency selective time varying multipath fading channel in mobile wireless MIMO-OFDM system with the presence of Gaussian and non-Gaussian impulsive noise. The M-SVR estimation technique is developed and applied to MIMO LTE downlink.
Different scenarios are simulated and discussed in the case of SIMO and MIMO systems with both STBC and V-BLAST schemes.
Finally, a general conclusion and perspectives are drown.

Chapter 1

Preliminaries

1.1 Introduction

Classical regression and statistical techniques of Bayesian classification are based on the strict assumption that the fundamental probability distribution must be known. However, in real life, we are often faced with problems of regression or classification with unknown distributions. The support vector machines (SVMs) are one of relatively new and promising method to learn separating functions in the case of classification, or to produce an estimation function in regression case. SVMs were derived from statistical learning theory (SLT) by Vapnik in 1995 for "learning without data distribution (distribution free)".

In this chapter, a historical overview on support vector machines was first described. Since SVMs are based on different theories such as statistical learning theory, optimization theory and kernel theory, it is therefore necessary to present these theories to understand the SVMs technique. Thus, learning methodology and linear learning machine, statistical learning theory and optimization theory are presented. Finally, we give a comparison between SVMs and neural networks and we outline the particular advantages of SVMs.

1.2 Historical Overview

Vladimir Vapnik and Alexey Chervonenkis developed the theory of Vapnik - Chervonenkis (also known as the VC theory) in 1971 [3]. VC theory is a learning theory related to the statistical learning theory without data distribution. This was the starting point for the support vector machines and the techniques of supervised learning for classification and regression. Although Vapnik introduced the topics of linear classifiers and optimum separating hyperplane in 1960, these subjects did not have that much attention until the 1992, when Bernhard Boser, Isabelle Guyon and Vapnik showed how to create nonlinear classifiers by applying kernel methods with maximum margin hyperplanes [3]. The kernel technique is a method of converting the linear classification algorithm into a nonlinear classification one by means of nonlinear function that projects the origi-

nal observations into a high dimension space, initially proposed by Aizerman in 1964 [3]. Thus, SVMs have been applied successfully in various applications of classification. In 1996, Vapnik, Harris Drucker, Chris Burges, Linda Kaufman and Alex Smola have proposed a version of SVMs for solving regression problems. This version is known by *support vector regression* (SVR).

On the other hand, support vector machines include various theories such that VC theory, statistical learning theory, optimization theory, kernel theory, etc. This is why SVMs are particularly strong compared to other methods based on minimizing the empirical risk such traditional neural networks.

1.3 Learning Methodology

The construction of a machine which is capable to learn from experience was the subject of a philosophical and technical debate for a long time. The technical aspect received a huge success with the arrival of electronic computers. They demonstrated that machines can have a significant level of learning ability.
The availability of consistent learning systems has a strategic importance because there are many tasks that can not be resolved by classical programming techniques, since no mathematical model of the problem is available.

Next, we will present the major components of the learning methodology by giving an overview of the supervised learning and linear learning machines.

1.3.1 Supervised learning

Use observations to synthesize programs is an approach known as *learning methodology*. If these observations are pairs of input/output, then this approach will be called *supervised learning*. These observations are called *training data*. When a fundamental function linking inputs to outputs exists, then it is called *target function*. The solution of the learning problem consists of estimating the target function which is learned by the *learning algorithm*. In the case of classification, this function is designated by the name of *decision function*.

The solution is selected from a set of candidate functions that map the input space to the output domain. Generally, we use a particular set of candidate functions known as *hypotheses* before we start learning the correct function. The choice of the set of assumptions (hypothesis space) is the main ingredient of the learning strategy. The algorithm that uses the training data as input to select a hypothesis in the hypothesis space is the second important ingredient. It is called the *learning algorithm*.

If the outputs of a learning problem are binary, then the problem is a *binary* classification problem. The problem is called *multi-classes* classification problem for outputs with a finite number of categories, while for the real-valued outputs the problem becomes known as *regression* problem.

The purpose of learning for the classification case consists of finding a hypothesis that performs a correct data classification. Thus, learning algorithms have been designed to

find an accurate data adaptation. Such hypothesis that provides a good adaptation is called *coherent*.

Two problems lead not to generate a coherent hypothesis:

- The function to be learned can have a complex representation.
- Frequently, the training data are noisy, therefore there is no guarantee that there is a fundamental function that adapts the training data correctly.

The ability of a hypothesis to correctly classify data outside the training set is known as *generalization*. Indeed, this property that we aim to optimize. The generalization is described in details in section 1.4.

1.3.2 Linear learning

In supervised learning, the learning machine is given by a set of training examples (or inputs) associated with outputs. Generally, the examples have the form of attributed vectors, so that the input space is a subset of $\mathbb{R}^n$. Once the attributed vectors are available, a number of hypotheses could be selected for the problem under consideration. Linear functions are the best hypotheses to learn and the simplest to apply. Traditional statistics and classical literature of neural networks have developed many methods to distinguish between two classes of examples using linear functions. These techniques include both iterative procedures and theoretical analysis of generalization properties. That is why they provide the framework for building complex systems. We will refer to learning machine using hypotheses that are linear combinations of input variables by *linear learning machines*.

We will see that in most cases, these machines can be represented as a particularly useful form called *dual representation*. In addition, the important concepts of *margin* and *margin distribution* will be introduced in the next chapter with algorithms for classification and regression.

1.4 Statistical Learning Theory

1.4.1 Generalization

The aim of modeling is to choose a model (function) from the hypothesis space, which is the closest (with some degree of error) to the fundamental function in the objective space. Thus, the learning machine can be formulated as a function $f(\boldsymbol{x})$ of the form $f : \mathbb{R}^n \longmapsto \mathbb{R}$ that associates every input vector $\boldsymbol{x}$ to a scalar y. We assume that this function is one among a set of parametric functions $f(\boldsymbol{w}, \boldsymbol{x})$. We want to estimate the parameter vector $\boldsymbol{w}$ to obtain the best estimation of the scalar y from the vector $\boldsymbol{x}$.
We assume that the pairs $\{\boldsymbol{x}, y\}$ are given with probability distribution $P(\boldsymbol{x}, y)$. Thus, we obtain the loss function $L(f(\boldsymbol{w}, \boldsymbol{x}), y)$ between the estimation $f(\boldsymbol{w}, \boldsymbol{x})$ and the desired output y. Practically, a common loss function is used, which is the quadratic loss

function *(squared loss)*. Our purpose is to find the function f that minimizes the mean error risk given by [46]

$$R[f(\boldsymbol{w})] = \int_{X\times Y} L(y, f(\boldsymbol{w}, \boldsymbol{x}))P(\boldsymbol{x}, y)d\boldsymbol{x}dy. \tag{1.1}$$

However, in practice $P(\boldsymbol{x}, y)$ is unknown. Nevertheless, it is possible to find an approximation which can be expressed by the average loss of all the learning pairs $\{\boldsymbol{x}, y\}$. This principle is often called Empirical Risk Minimization (ERM):

$$R_{emp}[f] = \frac{1}{l}\sum_{i=1}^{l} L(y^i, f(\boldsymbol{w}^i, \boldsymbol{x}^i)). \tag{1.2}$$

The empirical risk minimization is effective if the number of samples tends to infinity,

$$\lim_{l\to\infty} R_{emp}[f] = R[f]. \tag{1.3}$$

For a small set of samples, the difference between the actual error risk and the empirical risk can be large. However, the empirical risk must also satisfy:

$$\lim_{l\to\infty} \min_{f\in\mathcal{H}_n} R_{emp}[f] = \min_{f\in\mathcal{H}_n} R[f]. \tag{1.4}$$

This relationship is valid if $\mathcal{H}_n$ is small enough. This condition requires the convergence of the minimum. The following limit is obtained with probability $(1-\delta)$ [46]:

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h\ ln(\frac{2l}{h}+1) -\ ln(\frac{\delta}{4})}{l}}. \tag{1.5}$$

Note that this expression of risk is independent on the probability distribution.

As the actual risk can not be measured, the measurement of operating performance for the learning network is determined by the measurement of the error of future data outside the training set. This error is known by *generalization error*. This error can sometimes be large, such as in the case where the size of the available training set is very small compared to the size of the network parameters set.

In fact, the generalization error results from two types of error:

- *The approximation error*: is a consequence of the difference between the hypothesis and the objective function. A bad choice of the hypothesis space results in a large approximation error.
- *The estimation error*: is the error caused by the learning procedure, and results from the technique that chooses an optimal hypothesis of the hypothesis space.

In practice, it has been shown that the direct minimization of learning error, for a given training set obtained by using learning algorithm, does not necessarily imply the minimization of the generalization error.

In the literature of neural networks, this phenomenon is called *overfitting* [47]. An intuitive method to reduce overfitting is to minimize the complexity of learning machine. The rule is as follows: *Choose the simplest possible function which expresses the data in a satisfied manner*. This is the classic rule of Occam razor [48].

Thus, the optimization of our learning machine (for the case of classification or regression) implies the selection of a model that measures and minimizes the complexity of the machine. If one chooses a machine as a classifier or as a regressor, then this function will be able to adapt itself with the training data. Therefore it will reduce the empirical error. But if the data used in learning does not represent the statistics of the problem, then the solution will be bad. The solution which consists of reducing the test error (or generalization error) consists really to reduce the machine complexity, and therefore it will have less degree of freedom to express the solution. The problem of reducing the complexity of learning machine is achieved by introducing a regularization term in the function to be optimized. This term is used to penalize the complexity of the machine [49]. The particular method that introduces the penalty of complexity is given by the VC theory.

1.4.2 Dimension of Vapnik Chervonenkis (VC)

The VC dimension is a scalar value that measures the ability of a set of functions.

Definition 1.1 *(Vapnik-Chervonenkis)*
The VC dimension of a set of functions is equal to p if and only if there exists a set of points $\left\{x^i\right\}_{i=1}^{p}$ such that these points can be separated in all 2^p possible configurations, and there is no set $\left\{x^i\right\}_{i=1}^{q}$ for $q > p$ satisfying this property.■

1.4.3 Structural risk minimization (SRM)

Let S_h be a hypothesis space of VC dimension equal to h, then $S_1 \subset S_2 \subset \ldots S_\infty$.
The SRM consists of solving the following problem:

$$\min_{S_h} R_{emp}[f] + \sqrt{\frac{h\ ln(\frac{2l}{h}+1) -\ ln(\frac{\delta}{4})}{l}}. \tag{1.6}$$

Problems with multiple outputs can be reduced to a set of problems with a single output that can be considered independently. Therefore, it is appropriate to consider processes with multiple entries for which we wish provide a single output.

1.5 Optimization Theory

The function hypothesis must be chosen in order to minimize (or maximize) a certain criterion. In the case where learning machine is linear, the problem amounts to search a vector of parameters that minimizes (or maximizes) some cost function typically under some constraints.

Optimization theory is the branch of mathematics concerned with the characterization of solutions of some of these problem classes and develop efficient algorithms to search solutions. Indeed, the problem of learning machine is formulated into a form that can be analyzed within the framework of optimization theory. According to the specificity of the cost function and the nature of constraints, we can distinguish a number of well-formulated classes of optimization problems which have effective resolution methods.

We will focus on the case of convex and quadratic cost functions with linear constraints. These classes of optimization problems are known as *convex quadratic programs*. It was proved that these classes are adequate for the case of SVMs learning. Optimization theory defines the necessary and sufficient conditions for that a given function be a solution. The duality theory also produces a natural interpretation of the dual representations of learning machines. In addition, a deep understanding of mathematical structures of solutions can inspire several specific heuristic algorithms and implementation techniques.

1.5.1 Primal problem

The general form of the problem is to find the minimum (or the maximum) of a function under some constraints. The general problem of optimization can be described as follows:

Definition 1.2 *(Primal optimization problem)*
Given functions f, g_i $(i = 1, \cdots, k)$ and h_j $(j = 1, \cdots, m)$ defined in the domain $\Omega \subseteq \mathbb{R}^n$,

$$\begin{aligned} &\textit{minimize} && f(\boldsymbol{w}), && \boldsymbol{w} \in \Omega && (1.7)\\ &\textit{subject to} && g_i(\boldsymbol{w}) \leq 0, && i = 1, \cdots, k, && \\ & && h_j(\boldsymbol{w}) = 0, && j = 1, \cdots, m, && (1.8) \end{aligned}$$

where $f(\boldsymbol{w})$ is called the *objective function*, and the remaining relationships are called respectively, *the inequality and equality constraints*. The optimal value of the objective function is called the *value of the optimization problem*. To simplify the notation, we write $g(\boldsymbol{w}) \leq 0$ to indicate $g_i(\boldsymbol{w}) \leq 0,\ i = 1, \cdots, k$. The term $h(\boldsymbol{w}) = 0$ has a similar meaning as the equality constraints. As the maximization problems can be converted to minimization problems by reversing the sign of $f(\boldsymbol{w})$, the choice of the minimization is not a restriction. Similarly, any constraint can be represented by the form described above. The region of the domain where the objective function is defined and in which all constraints are satisfied is called *feasible region* that we denote by

$$R = \{\boldsymbol{w} \in \Omega : g(\boldsymbol{w}) \leq 0, h(\boldsymbol{w}) = 0\}. \tag{1.9}$$

The solution of the optimization problem is the point $\boldsymbol{w}^* \in R$ for which there is no other point $\boldsymbol{w} \in R$ such that $f(\boldsymbol{w}) < f(\boldsymbol{w}^*)$. Such a point is known as *global minimum*. The point $\boldsymbol{w}^* \in \Omega$ is called *local minimum* of $f(\boldsymbol{w})$ if $\exists\ \epsilon > 0$ such that

$f(\boldsymbol{w}) \geq f(\boldsymbol{w}^*)\ \forall\ \boldsymbol{w} \in \Omega,\ |\boldsymbol{w} - \boldsymbol{w}^* < \epsilon|$.■

Depending on the nature of the objective function and constraints, we can create different optimization problems.

Definition 1.3
An optimization problem in which the objective function, equality and inequality constraints are all linear functions, is called *linear program*. If the objective function is quadratic and constraints are all linear, the optimization problem is called *quadratic program*. An inequality constraint $g_i(\boldsymbol{w}) \leq 0$ is called *active* if the solution $\boldsymbol{w}^*$ satisfies $g_i(\boldsymbol{w}) = 0$, otherwise it is called *inactive*. In this sense, the equality constraints are all active.
Sometimes, some amounts called *slack variables* and denoted by ξ are introduced to transform inequality constraints into equality constraints as follows:

$$g_i(\boldsymbol{w}) \leq 0 \Leftrightarrow g_i(\boldsymbol{w}) + \xi_i = 0. \tag{1.10}$$

with $\xi_i \geq 0$.■

1.5.2 Lagrangian theory

The purpose of the Lagrangian theory is to characterize the solution of an optimization problem, initially when there is no inequality constraints. The basic elements of this theory are the *Lagrange multipliers* and the *Lagrangian function*. This method was developed by Lagrange in 1797 for mechanical problems. In 1951, Kuhn and Tucker extended this method for inequality constraints, which became known by the *Kuhn-Tucker theory* [3]. These three concepts produce all that is needed to develop effective solutions for SVMs optimization problems.

Theorem 1.1 *(Fermat) A necessary condition for $\boldsymbol{w}^*$ to be the minimum of $f(\boldsymbol{w}), f \in C^1$, is $\frac{\delta f(\boldsymbol{w}^*)}{\delta \boldsymbol{w}} = 0$. This condition, with the convexity of f present a sufficient conditions. In problems with constraints, we need to define a function, known as **Lagrangian**, which contains information about the objective function and the various constraints. Thus, its stationarity can be used to find solutions. Specifically, Lagrangian is defined as the objective function plus a linear combination of the constraints. The coefficients of this combination are called **Lagrange multipliers**.*

Definition 1.4 Given an optimization problem with objective function $f(\boldsymbol{w})$ and equality constraints $h_i(\boldsymbol{w}) = 0,\ i = 1, \cdots, m$. We define the Lagrangian function as

$$L(\boldsymbol{w}, \beta) = f(\boldsymbol{w}) + \sum_{i=1}^{m} \beta_i\ h_i(\boldsymbol{w}), \tag{1.11}$$

where the coefficients β_i are called *Lagrange multipliers*.■

If the point $\boldsymbol{w}^*$ is a local minimum for an optimization problem with only equality constraints, then it is possible that $\frac{\delta f(\boldsymbol{w}^*)}{\delta \boldsymbol{w}} \neq 0$, but by moving directions to reduce f, we can violate one or more constraints. To respect the equality constraints h_i, we must move perpendicular to $\frac{\delta h_i(\boldsymbol{w}^*)}{\delta \boldsymbol{w}}$. To respect all the constraints, we must move perpendicular to the subspace V spanned by $\frac{\delta h_i(\boldsymbol{w}^*)}{\delta \boldsymbol{w}}$.
This observation forms the basis of the second optimization result concerning optimization problems with equality constraints.

Theorem 1.2 *(Lagrange) A necessary condition for the point $\boldsymbol{w}^*$ to be a minimum of $f(\boldsymbol{w})$ subject to constraints $h_i(\boldsymbol{w}) = 0,\ i = 1, \cdots, m$ with $f, h_i \in C^1$ is*

$$\frac{\delta L(\boldsymbol{w}^*, \beta^*)}{\delta \boldsymbol{w}} = \mathbf{0}, \tag{1.12}$$

$$\frac{\delta L(\boldsymbol{w}^*, \beta^*)}{\delta \beta} = 0, \tag{1.13}$$

for some values β^*. These conditions are sufficient to prove that $L(\boldsymbol{w}, \beta^*)$ is a convex function of $\boldsymbol{w}$.
Note that if the constraints are zero, then the value of the Lagrangian at the optimal point is equal to the value of the objective function $L(\boldsymbol{w}^*, \beta^*) = f(\boldsymbol{w}^*)$.

Now, we consider the more general case where the optimization problem contains both equality and inequality constraints. First, we give the definition of the generalized Lagrangian.

Definition 1.5 *Consider the optimization problem in the domain $\Omega \subseteq \mathbb{R}^n$,*

$$\begin{array}{lll} \textit{minimize} & f(\boldsymbol{w}), & \boldsymbol{w} \in \Omega \qquad (1.14) \\ \textit{subject to} & g_i(\boldsymbol{w}) \leq 0, & i = 1, \cdots, k, \\ & h_j(\boldsymbol{w}) = 0, & j = 1, \cdots, m. \qquad (1.15) \end{array}$$

We define the generalized Lagrangian function as

$$\begin{aligned} L(\boldsymbol{w}, \alpha, \beta) &= f(\boldsymbol{w}) + \sum_{i=1}^{k} \alpha_i\ g_i(\boldsymbol{w}) + \sum_{j=1}^{m} \beta_j\ h_j(\boldsymbol{w}) \\ &= f(\boldsymbol{w}) + \alpha^{'} \mathbf{g}(\boldsymbol{w}) + \beta^{'} \mathbf{h}(\boldsymbol{w}). \ \blacksquare \end{aligned} \tag{1.16}$$

We can now define the *Lagrangian dual problem.*

Definition 1.6 *The Lagrangian dual problem corresponding to the primal problem is*

$$\begin{array}{ll} \textit{maximize} & \theta(\alpha, \beta) \\ \textit{subject to} & \alpha \geq 0, \end{array} \tag{1.17}$$

where $\theta(\alpha, \beta) = inf_{\boldsymbol{w} \in \Omega} L(\boldsymbol{w}, \alpha, \beta)$.

The value of the objective function at the optimal solution is called *the value of the problem.* ■

Now we quote the theorem of strong duality which guarantees that the primal and dual problems have the same value of the considered optimization problems.

Theorem 1.3 *(Strong duality theorem) Given an optimization problem with the convex domain* $\Omega \subseteq \mathbb{R}^n$*,*

$$\begin{aligned} &\textit{minimize} && f(\boldsymbol{w}), && \boldsymbol{w} \in \Omega && (1.18)\\ &\textit{subject to} && g_i(\boldsymbol{w}) \leq 0, && i = 1, \cdots, k, && \\ && h_j(\boldsymbol{w}) = 0, && j = 1, \cdots, m, && (1.19) \end{aligned}$$

where g_i *and* h_i *are affine functions, thus is* $h(\boldsymbol{w}) = \boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}$*. For some matrix* $\boldsymbol{A}$ *and vector* $\boldsymbol{b}$*, the interval of duality (duality gap) is zero.*

We are now able to give the theorem of Kuhn-Tucker giving the conditions for that a solution is optimal for a general optimization problem.

Theorem 1.4 *(Kuhn-Tucker) Given an optimization problem with the convex domain* $\Omega \subseteq \mathbb{R}^n$*,*

$$\begin{aligned} &\textit{minimize} && f(\boldsymbol{w}), && \boldsymbol{w} \in \Omega && (1.20)\\ &\textit{subject to} && g_i(\boldsymbol{w}) \leq 0, && i = 1, \cdots, k, && \\ && h_j(\boldsymbol{w}) = 0, && j = 1, \cdots, m, && (1.21) \end{aligned}$$

where $f \in C^1$ *is convex and* g_i, h_i *are affine functions. The necessary and sufficient condition for that the point* $\boldsymbol{w}^*$ *be optimal is the existence of* α^*, β^* *such that*

$$\frac{\delta L(\boldsymbol{w}^*, \alpha^*, \beta^*)}{\delta \boldsymbol{w}} = \boldsymbol{0}, \tag{1.22}$$

$$\frac{\delta L(\boldsymbol{w}^*, \alpha^*, \beta^*)}{\delta \beta} = 0, \tag{1.23}$$

$$\alpha_i{}^* g_i(\boldsymbol{w}^*) = 0, \quad i = 1, \cdots, k \tag{1.24}$$

$$g_i(\boldsymbol{w}^*) \leq 0, \quad i = 1, \cdots, k \tag{1.25}$$

$$\alpha_i{}^* \geq 0, \quad i = 1, \cdots, k. \tag{1.26}$$

The relation (1.24) is known as the KKT (Karush-Kuhn-Tucker) complementarity condition. It implies that for active constraints, we will have $\alpha_i{}^* \geq 0$. While for inactive constraints, $\alpha_i{}^* = 0$. The Lagrange multiplier represents the sensitivity of the

optimal value with respect to the constraint. Disruption of inactive constraints has no effect on the solution of the optimization problem.

On the other hand, KKT conditions state that if a constraint is active, then $g_i(\boldsymbol{w}^*) = 0$ or the correspondent multiplier will be zero ($\alpha_i = 0$). This is summarized in the equation $g_i(\boldsymbol{w}^*){\alpha_i}^* = 0$.

1.5.3 Duality

The Lagrangian treatment of convex optimization problems leads to a dual description easier to solve than the primal problems. This is due to the difficulty of handling the inequality constraints directly. The dual problem is obtained by introducing Lagrange multipliers, also called *dual variables*. The dual methods are based on the idea that only the dual variables are the basic unknown of the problem.

We can transform the primal problem into a dual problem by making simply the derivatives of the Lagrangian with respect to primal variables to zero, and substituting the obtained relations into the Lagrangian. Therefore, we eliminate the dependence on primal variables. This corresponds to explicitly calculate the function

$$\theta(\alpha, \beta) = inf_{\boldsymbol{w} \in \Omega} L(\boldsymbol{w}, \alpha, \beta). \tag{1.27}$$

The resulting function contains only dual variables and should be maximized under simpler constraints. This strategy will be adopted later as one of the standard techniques in the theory of support vector machines. The use of dual representations in the theory of SVMs allows us not only to work in spaces of higher dimensions, but also to use algorithmic techniques derived from optimization theory.

The KKT complementarity conditions imply that only the active constraints will have non-zero dual variables. This means that for some optimization problems, the real number of variables can be significantly lower than the size of the training set.
We will see in the next chapter that the *support vector* term refers to those examples for which the dual variables are non-zero.

1.6 SVMs and Neural Networks

1.6.1 Similarities between SVMs and neural networks

- SVMs and neural networks perform learning from experimental data, for which the fundamental probability distribution is often unknown .
- Both are universal approximators in the sense that they can approximate any function of any desired degree of accuracy.
- After training, they are given with the same mathematical model.

1.6.2 Differences between SVMs and neural networks

- Neural networks are based on the principle of minimization of empirical error, starting with an appropriate structure to minimize the error. SVMs are based on the principle of minimization of structural error, starting with the error to achieve an optimal structure.
- SVMs and neural networks follow different learning methods. Neural networks typically use the back-propagation algorithm or other algorithms such as gradient descent and methods based on linear algebra. By cons, SVMs are learned by solving a linear and quadratic programming problem.

1.6.3 Advantages of SVMs compared to neural networks

- SVMs can achieve a compromise between the positive errors and negative errors using asymmetric soft margin.
- SVMs converge always to the same solution for a given set of data independent on initial conditions.
- SVMs ensure convergence to the global minimum.
- SVMs avoid the problem of overfitting.

1.7 Conclusion

In this chapter, we saw a historical overview on support vector machines, especially the founding papers of Vapnik and Chervonenkis described in [3]. Then, we have presented the learning methodology and the linear learning machines which lead to a simplicity and flexibility of learning. Also, we have introduced generalization and statistical learning theory. It is this theory which limits the generalization error, not only for data from the training set, but also for future data. In addition, we have presented the optimization theory. This theory forms the basis of optimization problems, particularly the convex quadratic optimization problems which are used in the case of SVMs. Finally, we have described the differences between SVMs and neural networks. Indeed, the main advantage of SVMs is the convergence to a unique solution.

SVMs will be presented in details in the next chapter, especially the support vector machines for regression (SVR) in their linear and nonlinear version using kernel functions.

Chapter 2

State of the Art of Support Vector Machines

2.1 Introduction

The different theories presented in the previous chapter gave us the basis on which the support vector machines have been founded. Indeed, SVMs are a system for efficiently training the linear learning machines in the kernel-induced feature space, while respecting the theory of generalization and statistical learning and exploiting the optimization theory.

This chapter is devoted to the study of support vector machines. At first, we present the linear support vector classification and regression. Then, we describe the theory of kernel functions, those functions necessary to project the nonlinear data in the original space into a high dimensional space where the data become represented linearly. Finally, we present the nonlinear support vector machines for classification and regression based on these kernel functions.

2.2 Linear Support Vector Machines

We begin this chapter with a description of the basic concepts of support vector classification (SVC) and support vector regression (SVR) in their linear versions.

2.2.1 Linear support vector classification

Suppose that we have two classes of points $\boldsymbol{x}_i$ located in different areas in space, so they can be classified (or separated) using a separating hyperplane. Let as further assume that these data are located in such a way that the hyperplane passes through the origin (see Fig. 2.1). The expression of this hyperplane is then $\boldsymbol{w}^T\boldsymbol{x} = 0$. To identify the class of a point $\boldsymbol{x}_i$, we must determine the angle between this sample and the vector $\boldsymbol{w}$. If this angle is less than $\pi/2$, then the sample is classified in class A. Otherwise, the class will be B. In other terms, to determine the class of each sample, we must calculate

the dot product between the vector $\boldsymbol{w}$ and the corresponding vector of sample. If this dot product is positive (the angle is less than $\pi/2$), then the class will be A. Otherwise, the dot product will be negative and the class will be B. In addition, a sample located on the plane is orthogonal to $\boldsymbol{w}$, so the dot product is zero, consequently we can not determine its class.

If the data are not centralized around the origin, then the hyperplane must be biased from it, and its expression will be $\boldsymbol{w}^T\boldsymbol{x} + b = 0$.

2.2.1.1 Adjustment of the separating hyperplane

The main idea of the classification problem is how to determine the separating hyperplane which classifies the data optimally. The solution is to minimize the probability of misclassification in the classification of a set of samples different from that used to adjust the parameters $\boldsymbol{w}$ and b of the hyperplane. This concept is called *generalization ability*.

Indeed, classification is measured as a binary quantity, where 0 indicates a correct classification and 1 indicates a misclassification. Instead of using these quantities directly, we use the approximate quantities since the function is not differentiable [50]. Thus, the optimal separating hyperplane is a hyperplane that minimizes the MMSE (Minimum Mean Square Error) between the desired results (+1 for samples of class A and −1 for samples of class B) and actual results obtained by the classification.

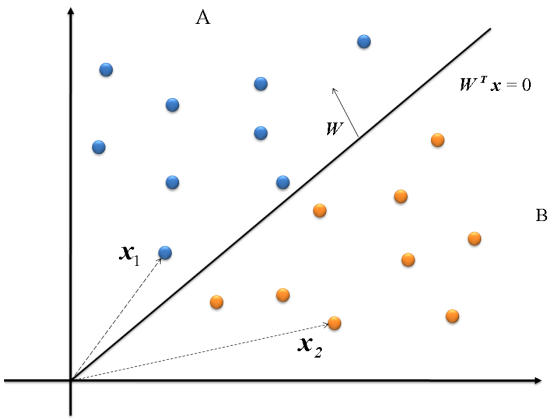


Figure 2.1: A classifier in vector space.

This criterion is optimal when the statistical properties of the data are Gaussian. But if the data are not Gaussian, the result will be biased. In addition, if the data are Gaussian and including outliers (samples far from the set), then the statistical properties of the set will be biased as well.

2.2.1.2 SVM approach

To deal with these situations, we apply a procedure that uses a linear hyperplane, while assuming that the data are linearly separable. Therefore, the optimality criterion will be: *put the hyperplane as far as possible from the nearest samples, while ensuring that all samples are in their correct side.* In other words, it corresponds to maximize the margin between the separating hyperplane and the closest samples (see Fig. 2.2). The SVM criterion can be reformulated as: *maximize the distance d between the separating hyperplane and the closest samples subject to the constraints*

$$y_i \left[\boldsymbol{w}^T \boldsymbol{x}_i + b \right] \geq 1, \tag{2.1}$$

where $y_i \in \{+1, -1\}$ is the scalar associated to the sample $\boldsymbol{x}_i$. This constraint means that the sign of the output of the classifier and the sign of the corresponding scalar must be equal in all cases.

The distance d between the separating hyperplane and the margin hyperplane is equal to $\frac{1}{\|\boldsymbol{w}\|}$ [3]. Thus, maximizing d is equivalent to minimizing the norm of the weight vector $\boldsymbol{w}$, which provides a most useful expression of the SVM criterion

$$\begin{aligned} minimize_{\boldsymbol{w},b} \quad & L_P = \tfrac{1}{2}\|\boldsymbol{w}\|^2 && (2.2) \\ s.t \quad & y_i \left[\boldsymbol{w}^T \boldsymbol{x}_i + b \right] \geq 1. && (2.3) \end{aligned}$$

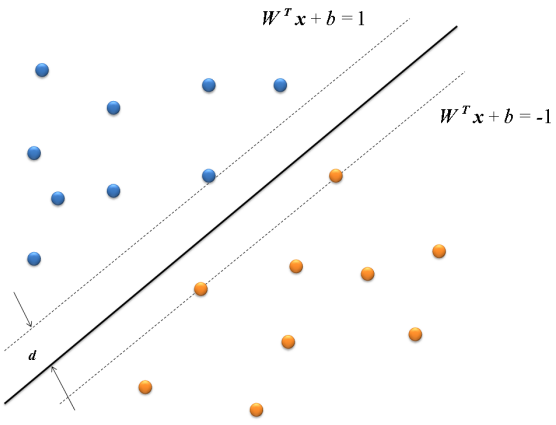


Figure 2.2: Placement of the hyperplanes.

In practice, samples are not always linearly separable, and hence the constraint (2.2) can not be satisfied. For that reason, *slack variables* should be introduced to account for the nonseparable samples. Then, the optimization criterion consists of minimizing the following primal problem [46]:

$$\begin{aligned} minimize_{\boldsymbol{w},b,\xi} \quad & L_P = \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i && (2.4) \\ s.t \quad & y_i \left(\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0. && (2.5) \end{aligned}$$

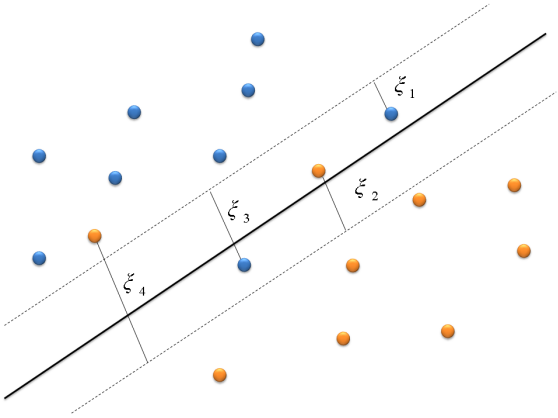


Figure 2.3: Hyperplane for the nonseparable case.

If a sample $\boldsymbol{x}_i$ is correctly classified and it is outside the margin, then its slack value ξ_i will be zero. If it is correctly classified but its position is between the margin, then $0 < \xi_i < 1$. And if it is misclassified, then $\xi_i > 1$ (see Fig. 2.3).

The term C is a tradeoff between the maximization of the margin and the minimization of the empirical error. A low value of C leads to a less complex machine and very low values lead to a poor expression of data. However, very large values of C lead to a low generalization ability. The value of C is often adjusted heuristically.

2.2.1.3 Optimization of LSVC

The optimization problem in equation (2.4) is a classic optimization problem under constraints. To solve this problem, we must exploit the Lagrange optimization procedure which consist to apply the Lagrange multipliers α_i to different constraints [51]. In addition, Lagrangian optimization procedure is based on the soft margin optimization methods which are of two types:

- 1-Norm Soft Margin - Box Constraint.
- 2-Norm Soft Margin - Weighting the Diagonal.

We used in this LSVC case the 1-Norm Soft Margin Lagrangian which is applied to the optimization problem described by expression (2.4) as follows:

$$\begin{aligned} L_{pd}(\boldsymbol{w}, b, \xi, \alpha, \beta) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b - 1 + \xi_i\right] \\ &- \sum_{i=1}^{N}\beta_i\xi_i, \end{aligned} \tag{2.6}$$

with $\alpha_i, \beta_i \geq 0$.

The corresponding dual is obtained by differentiating with respect to $\boldsymbol{w}, \xi$ and b

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \alpha, \beta)}{\delta \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{x}_i = \mathbf{0}, \tag{2.7}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \alpha, \beta)}{\delta b} = \sum_{i=1}^{N} y_i \alpha_i = 0, \tag{2.8}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \alpha, \beta)}{\delta \xi_i} = C - \alpha_i - \beta_i = 0. \tag{2.9}$$

By substituting the obtained relations in equation (2.6), we obtain the following adaptation of the dual objective function:

$$L_d(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j {\boldsymbol{x}_i}^T \boldsymbol{x}_j, \tag{2.10}$$

with constraints

$$0 \leq \alpha_i \leq C. \tag{2.11}$$

The constraints $C - \alpha_i - \beta_i = 0$ and $\beta_i \geq 0$ require that $\alpha_i \leq C$. In addition, $\xi_i \neq 0$ only if $\beta_i = 0$ yields that $\alpha_i = C$.
The KKT complementarity conditions are then given by

$$\alpha_i \left[y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b - 1 + \xi_i \right] = 0, \qquad \xi_i(\alpha_i - C) = 0, \qquad i = 1, \cdots, N. \tag{2.12}$$

The important result of this derivation is the expression of the weight vector $\boldsymbol{w}$ expressed as

$$\boldsymbol{w} = \sum_{i=1}^{N} y_i \alpha_i x_i. \tag{2.13}$$

On the other hand, the expression (2.10) can be rewritten in a matrix format as follows:

$$L_d = -\frac{1}{2} \alpha^T \boldsymbol{Y} \boldsymbol{R} \boldsymbol{Y} \alpha + \alpha, \tag{2.14}$$

where $\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix}$ is a column vector containing all Lagrange multipliers α_i,

$\boldsymbol{Y} = \begin{pmatrix} y_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & y_N \end{pmatrix}$ is a diagonal matrix with $\boldsymbol{Y}_{ii} = y_i$ and

$\boldsymbol{R} = \begin{pmatrix} R_{11} & \cdots & R_{1N} \\ \vdots & \ddots & \vdots \\ R_{N1} & \cdots & R_{NN} \end{pmatrix}$ represents the dot product matrix, where

$$R_{ij} = {\boldsymbol{x}_i}^T \boldsymbol{x}_j. \tag{2.15}$$

Note that the condition $\beta_i \xi_i = 0$ eliminates the last term of the expression (2.6). As we saw, it is evident to verify from this condition and the condition (2.9) that:

$$0 \leq \alpha_i \leq C, \tag{2.16}$$

$$\begin{array}{c} \text{if } \alpha_i \leq C, \text{ then } \xi_i = 0 \ \beta_i \neq 0 \\ \text{if } \xi_i > 0 \text{ , then } \alpha_i = C \ \beta_i = 0. \end{array}$$

Slack variables do not appear in the dual formulation because the terms $C \sum_{i=1}^{N} \xi_i$ and $C \sum_{i=1}^{N} \alpha_i \xi_i$ are null. This quadratic formulation can be solved using quadratic programming method presented in [47].

KKT conditions ensure that if $y_i \boldsymbol{w}^T \boldsymbol{x}_i \geq 1$ ($\xi_i = 0$), then ($\alpha_i = 0$). In other words, only a subset of Lagrange multipliers will have nonzero values in the solution, while other multipliers disappear. Thus, the solution will be unique. The associated samples (those for which $\alpha_i \neq 0$) are called *support vectors*.

The value of b can be calculated by considering that if $\alpha_i < C$, then $\xi_i = 0$. This leads to the following condition:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 = 0, \tag{2.17}$$

for any sample $\boldsymbol{x}_i$ for which $\alpha_i < C$.

In practice, it is numerically convenient to calculate b by averaging all samples having $0 \leq \alpha_i \leq C$.

Some eigenvectors of $\boldsymbol{R}$ may have low eigenvalues which leads to an ill-conditioned problem. This drawback is solved by adding a small identity matrix. Therefore, the dual to be solved becomes

$$L_d = -\frac{1}{2} \alpha^T \boldsymbol{Y} \left[\boldsymbol{R} + \gamma \boldsymbol{I} \right] \boldsymbol{Y} \alpha + \alpha. \tag{2.18}$$

The addition of an identity matrix is equivalent to applying a quadratic loss function to slack variables whose values are less than γC [50].

Fig. 2.4 shows two sets of linearly separable data, properly separated (classified) by a linear hyperplane. The dotted lines describe the area of the margin. Note that support vectors are all on the margin.

2.2.2 Linear support vector regression

A linear regressor is a function $\boldsymbol{w}^T \boldsymbol{x} + b$ which establishes a relationship between a set of vectors $\boldsymbol{x}$ in $\mathbb{R}^n$ and a set of scalar y in $\mathbb{R}$. Indeed, the regressor function achieves a hyperplane that provides a minimum mean square error between it and the data.

Linear regression was traditionally treated by Least Squares (LS) approaches, and it is widely used in many fields such as estimation models, linear prediction [48], equalization [52], etc. Conventional approaches use MSE criterion for processing data corrupted by Additive White Gaussian Noise (AWGN). The best approach is the Maximum Likelihood (ML) one. Nevertheless, we can not directly minimize the actual LS function because the noise distribution is unknown.

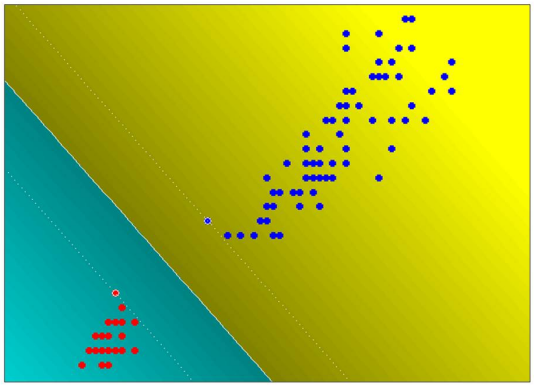


Figure 2.4: Solution of LSVC for $C = \infty$.

The basic idea of SVR is to find a function that represents pairs of data $(\boldsymbol{x}_i, y_i)$ with a small error, less than or equal to a given amount ε. This function should ensure a minimum norm $\|\boldsymbol{w}\|$. In addition, SVR does not minimize errors less than ε, but only errors above, allowing us to build a machine whose parameter vector is a linear combination of samples with an error greater than or equal to ε. This leads to a unique solution as in the case of LSVC.

The adjustment of the linear regressor can be formulated by the following primal problem:

$$\begin{aligned} minimize_{\boldsymbol{w},b,\xi,\xi^*} \quad & L_P = \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \textstyle\sum_{i=1}^{N} l\left(\xi_i, \xi_i^*\right) && (2.19) \\ s.t \quad & \langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b - y_i \leq \xi_i + \varepsilon \\ & y_i - \langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle - b \leq \xi_i^* + \varepsilon \\ & \xi_i, \xi_i^* \geq 0, && (2.20) \end{aligned}$$

where both constraints were applied to positive and negative errors, respectively.

Obviously, for each sample $\boldsymbol{x}_i$, only one slack variable is non-zero. The constraints in (2.19) imply that for every sample with a positive error (respectively negative) of absolute value greater than ε, this absolute value will be forced to be less than $\xi_i + \varepsilon$ (respectively $\xi_i^* + \varepsilon$). This is the lowest value that can take these slack variables.

So, the parameter ε is used to remove samples whose errors are inferior to ε. All Lagrange multipliers associated with samples having $\xi_i, \xi_i^* < \varepsilon$ will be zero.

Choose a cost function $l\left(\xi_i, \xi_i^*\right) = \xi_i + \xi_i^*$ in the primal problem (2.19) is equivalent to apply a linear $\varepsilon-$ insensitive function given by [52] (see Fig. 2.6)

$$l(e_i) = \begin{cases} 0, & |e_i| \leq \varepsilon \\ |e_i| - \varepsilon, & |e_i| \geq \varepsilon \end{cases} \qquad (2.21)$$

for $e_i = \xi_i + \varepsilon$ or $e_i = -{\xi_i}^* - \varepsilon$.

To solve our optimization problem under constraints, we must apply Lagrange optimization to make the problem without constraints. The optimization is performed in

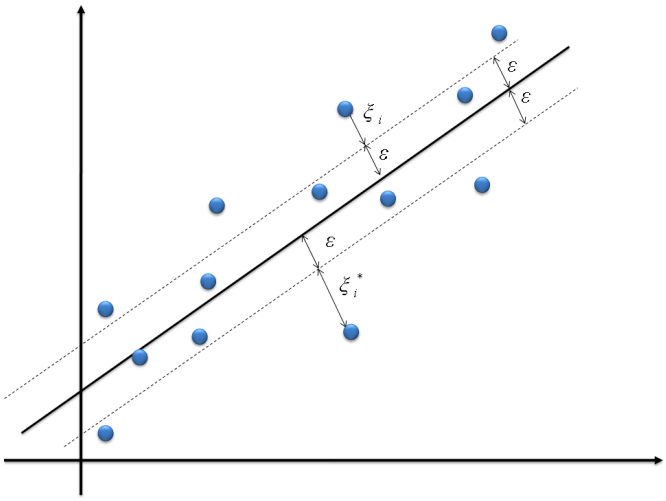


Figure 2.5: $\varepsilon-$ insensitivity concept.

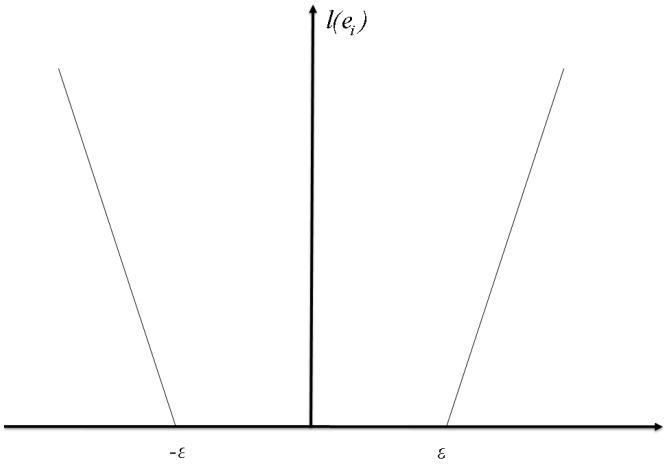


Figure 2.6: Linear $\varepsilon-$insensitive cost function.

terms of minimizing the weight vector and the norm of slack variable. We apply two techniques of loss for the $\varepsilon-$insensitive function, which are respectively:

- *Linear $\varepsilon-$ insensitive loss*: uses 1-dimensional norm of slack variables with $l\left(\xi_i, \xi_i^*\right) = \xi_i + \xi_i^*$ (see Fig. 2.6).
- *Quadratic $\varepsilon-$ insensitive loss*: uses 2-dimensional norm of slack variables with $l\left(\xi_i, \xi_i^*\right) = {\xi_i}^2 + {\xi_i^*}^2$ (see Fig. 2.7).

2.2.2.1 Optimization with linear $\varepsilon-$ insensitive loss function

The corresponding primal problem is expressed as

$$\begin{aligned} minimize_{\boldsymbol{w},b,\xi,\xi^*} \quad & \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) && (2.22)\\ s.t \quad & (\langle \boldsymbol{w}\cdot\boldsymbol{x}_i\rangle + b) - y_i \leq \varepsilon + \xi_i \\ & y_i - (\langle \boldsymbol{w}\cdot\boldsymbol{x}_i\rangle + b) \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0. && (2.23) \end{aligned}$$

The Lagrangian of this optimization problem is then

$$\begin{aligned} L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + {\xi_i}^*) \\ &+ \sum_{i=1}^{N}\alpha_i(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i - \varepsilon - \xi_i) \\ &+ \sum_{i=1}^{N}{\alpha_i}^*(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b - \varepsilon - {\xi_i}^*) \\ &- \sum_{i=1}^{N}\beta_i\xi_i - \sum_{i=1}^{N}{\beta_i}^*{\xi_i}^*, \end{aligned} \tag{2.24}$$

with $\alpha_i, {\alpha_i}^*, \beta_i, {\beta_i}^* \geq 0$.

The corresponding dual is obtained by differentiating with respect to $\boldsymbol{w}, \xi$ and b,

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*)}{\delta \boldsymbol{w}} = \boldsymbol{w} + \sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i - \sum_{i=1}^{N}{\alpha_i}^*\boldsymbol{x}_i = \boldsymbol{0}, \tag{2.25}$$

which implies that

$$\boldsymbol{w} = \sum_{i=1}^{N}({\alpha_i}^* - \alpha_i) \quad \boldsymbol{x}_i, \tag{2.26}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*)}{\delta b} = \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}{\alpha_i}^* = 0, \tag{2.27}$$

implying that

$$\sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i) = 0, \tag{2.28}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*)}{\delta \xi_i} = C - \alpha_i - \beta_i = 0, \tag{2.29}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*)}{\delta \xi_i{}^*} = C - \alpha_i{}^* - \beta_i{}^* = 0. \tag{2.30}$$

Substituting the obtained relations in the primal-dual Lagrangian, we obtain the following adaptation of the dual objective function:

$$\begin{aligned} maximize_{\alpha,\alpha^*} \quad L_d(\alpha, \alpha^*) &= \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i)y_i - \varepsilon\sum_{i=1}^{N}(\alpha_i{}^* + \alpha_i) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i{}^* - \alpha_i)(\alpha_j{}^* - \alpha_j)\boldsymbol{x}_i{}^T\boldsymbol{x}_j \qquad (2.31) \\ s.t \qquad 0 &\leq \alpha_i, \alpha_i{}^* \leq C \\ 0 &= \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i). \qquad (2.32) \end{aligned}$$

Indeed, the constraints $C - \alpha_i - \beta_i = 0$ and $C - \alpha_i{}^* - \beta_i{}^* = 0$ together with $\beta_i, \beta_i{}^* \geq 0$ required that $0 \leq \alpha_i, \alpha_i{}^* \leq C$.

The complementarity KKT conditions are as follows:

$$\alpha_i(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i - \varepsilon - \xi_i) = 0, \tag{2.33}$$

$$\alpha_i{}^*(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b - \varepsilon - \xi_i{}^*) = 0, \tag{2.34}$$

$$\xi_i\xi_i{}^* = 0 \ , \ \alpha_i\alpha_i{}^* = 0, \tag{2.35}$$

$$\xi_i(\alpha_i - C) = 0 \ , \ \xi_i{}^*(\alpha_i{}^* - C) = 0. \tag{2.36}$$

2.2.2.2 Optimization with quadratic $\varepsilon-$ insensitive loss function

The primal problem corresponding to the quadratic $\varepsilon-$ insensitive loss is

$$minimize_{\boldsymbol{w},b,\xi,\xi^*} \quad \|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}(\xi_i{}^2 + \xi_i^{*2}) \tag{2.37}$$

$$\begin{aligned} s.t \qquad & (\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) - y_i \leq \varepsilon + \xi_i \\ & y_i - (\langle \boldsymbol{w} \cdot \boldsymbol{x}_i \rangle + b) \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \qquad i = 1, \cdots, N. \end{aligned} \tag{2.38}$$

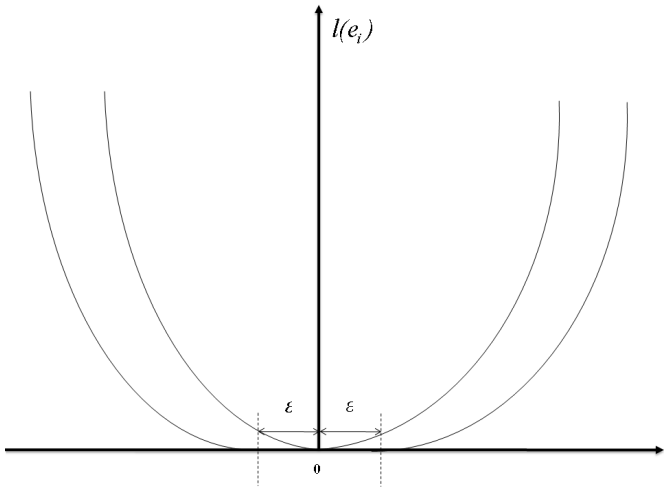


Figure 2.7: Quadratic $\varepsilon-$insensitive cost function.

The Lagrangian is calculated by

$$\begin{aligned} L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2}\sum_{i=1}^{N}({\xi_i}^2 + {\xi_i}^{*2}) \\ &+ \sum_{i=1}^{N}\alpha_i(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i - \varepsilon - \xi_i) \\ &+ \sum_{i=1}^{N}{\alpha_i}^*(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b - \varepsilon - {\xi_i}^*). \end{aligned} \tag{2.39}$$

Note that $\xi_i{\xi_i}^* = 0$ and $\alpha_i{\alpha_i}^* = 0$.

The corresponding dual is obtained by applying KKT conditions as follows

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \boldsymbol{w}} = \boldsymbol{w} + \sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i - \sum_{i=1}^{N}{\alpha_i}^*\boldsymbol{x}_i = \boldsymbol{0}, \tag{2.40}$$

implying that

$$\boldsymbol{w} = \sum_{i=1}^{N}({\alpha_i}^* - \alpha_i)\boldsymbol{x}_i, \tag{2.41}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta b} = \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}{\alpha_i}^* = \sum_{i=1}^{N}({\alpha_i}^* - \alpha_i) = 0, \tag{2.42}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \xi_i} = C\xi_i - \alpha_i = 0, \tag{2.43}$$

so

$$\xi_i = \frac{\alpha_i}{C}, \tag{2.44}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \xi_i{}^*} = C\xi_i{}^* - \alpha_i{}^* = 0, \tag{2.45}$$

implying that

$$\xi_i{}^* = \frac{\alpha_i{}^*}{C}. \tag{2.46}$$

Substituting the obtained relations in (2.39), we obtain the following dual:

$$\begin{aligned} L_d(\alpha, \alpha^*) &= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i{}^* - \alpha_i)(\alpha_j{}^* - \alpha_j)\boldsymbol{x}_i{}^T\boldsymbol{x}_j + \frac{1}{2C}\sum_{i=1}^{N}(\alpha_i{}^2 + \alpha_i{}^{*2}) \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_j(\alpha_i{}^* - \alpha_i)\boldsymbol{x}_i{}^T\boldsymbol{x}_j + b\sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i y_i - \varepsilon\sum_{i=1}^{N}\alpha_i \\ &- \frac{1}{C}\sum_{i=1}^{N}\alpha_i{}^2 + \sum_{i=1}^{N}\alpha_i{}^* y_i - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_j{}^*(\alpha_i{}^* - \alpha_i)\boldsymbol{x}_i{}^T\boldsymbol{x}_j \\ &- b\sum_{i=1}^{N}\alpha_i{}^* - \varepsilon\sum_{i=1}^{N}\alpha_i{}^* - \frac{1}{C}\sum_{i=1}^{N}\alpha_i{}^{*2} \\ L_d(\alpha, \alpha^*) &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i{}^* - \alpha_i)(\alpha_j{}^* - \alpha_j)\boldsymbol{x}_i{}^T\boldsymbol{x}_j - \varepsilon\sum_{i=1}^{N}(\alpha_i{}^* + \alpha_i) \\ &+ y_i\sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i) + b\sum_{i=1}^{N}(\alpha_i - \alpha_i{}^*) - \frac{1}{2C}\sum_{i=1}^{N}(\alpha_i{}^2 + \alpha_i{}^{*2}). \end{aligned} \tag{2.47}$$

The dual problem is then

$$\begin{aligned} maximize_{\alpha,\alpha^*} \quad L_d(\alpha, \alpha^*) &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i{}^* - \alpha_i)(\alpha_j{}^* - \alpha_j)(\boldsymbol{x}_i{}^T\boldsymbol{x}_j + \frac{1}{C}\delta_{ij}) \\ &+ \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i)y_i - \varepsilon\sum_{i=1}^{N}(\alpha_i{}^* + \alpha_i) \end{aligned} \tag{2.48}$$

$$s.t \qquad 0 = \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i) \tag{2.49}$$

$$0 \leq \alpha_i, \alpha_i{}^*.$$

The KKT complementarity conditions are resulting

$$\alpha_i(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i - \varepsilon - \xi_i) = 0, \tag{2.50}$$

$$\alpha_i{}^*(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b - \varepsilon - \xi_i{}^*) = 0, \tag{2.51}$$

$$\xi_i\xi_i{}^* = 0\ ,\ \alpha_i\alpha_i{}^* = 0 \qquad i = 1, \cdots, N. \tag{2.52}$$

2.2.2.3 Optimization with robust loss function

A *robust* loss function that combines both linear and quadratic techniques can also be applied [46]. This loss function has the following form:

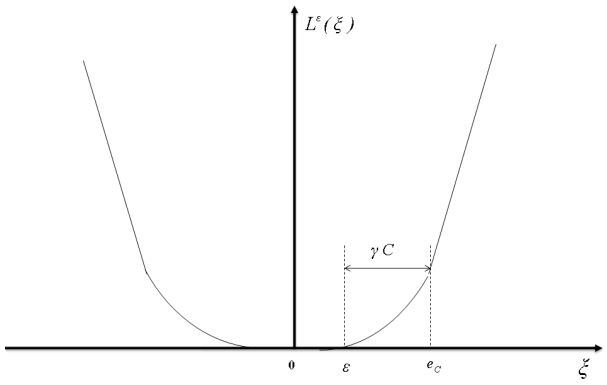


Figure 2.8: $\varepsilon-$insensitive cost function with quadratic and linear losses.

$$\mathcal{L}^{\varepsilon}(\xi) = \begin{cases} 0, & |\xi| \leq \varepsilon \\ \frac{1}{2\gamma}(|\xi| - \varepsilon)^2, & \varepsilon \leq |\xi| \leq e_C \\ C(|\xi| - \varepsilon) - \frac{1}{2}\gamma C^2, & e_C \leq |\xi| \end{cases}, \tag{2.53}$$

where $e_C = \varepsilon + \gamma C$; ε is the insensitivity parameter, and γ and C control the tradeoff between regularization and losses.

Three different regions yield to distinguish different types of noise (see Fig. 2.8):

- $\varepsilon-$ *insensitivity zone*: ignores errors less than ε.
- *zone with quadratic cost*: uses the 2-dimension norm of the errors, which is suitable for the Gaussian noise.
- *zone with linear cost*: limits the effects of the outliers.

Note that (2.53) is the Vapnik $\varepsilon-$insensitive cost function when γ is small enough. The parameter γ plays the role of numeric regulation in the quadratic problem.

The primal problem corresponding to (2.53) is as follows:

$$\begin{aligned} minimize_{\boldsymbol{w},b,\xi,\xi^*} \quad L_p &= \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2\gamma}\sum_{i\in I_1}({\xi_i}^2 + {\xi_i}^{*2}) \qquad (2.54) \\ &+ C\sum_{i\in I_2}(\xi_i + {\xi_i}^* - \frac{\gamma C}{2}) \\ s.t \quad -y_i &+ (\langle \boldsymbol{w}\cdot\boldsymbol{x}_i\rangle + b) \leq \varepsilon + \xi_i \\ y_i &- (\langle \boldsymbol{w}\cdot\boldsymbol{x}_i\rangle + b) \leq \varepsilon + \xi_i^* \qquad (2.55) \\ 0 &\leq \xi_i, \xi_i^*. \end{aligned}$$

I_1 and I_2 are the sets of samples for which the losses have quadratic and linear cost, respectively. The constraints are the same as (2.25).

The corresponding Lagrangian is then

$$\begin{aligned} L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2\gamma}\sum_{i\in I_1}(\xi_i{}^2 + \xi_i{}^{*2}) + C\sum_{i\in I_2}(\xi_i + \xi_i{}^* - \frac{\gamma C}{2}) \\ &+ \sum_{i=1}^{N}\alpha_i(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i - \xi_i - \varepsilon) \\ &+ \sum_{i=1}^{N}\alpha_i{}^*(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b - \xi_i{}^* - \varepsilon) \\ &- \sum_{i=1}^{N}\beta_i\xi_i - \sum_{i=1}^{N}\beta_i{}^*\xi_i{}^*. \end{aligned} \tag{2.56}$$

The KKT conditions are

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \boldsymbol{w}} = \boldsymbol{w} + \sum_{i=1}^{N}\alpha_i\boldsymbol{x}_i - \sum_{i=1}^{N}\alpha_i{}^*\boldsymbol{x}_i = \boldsymbol{0}, \tag{2.57}$$

so

$$\boldsymbol{w} = \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i)\boldsymbol{x}_i, \tag{2.58}$$

and

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta b} = \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i{}^* = \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i) = 0, \tag{2.59}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \xi_i} = \frac{1}{\gamma}\xi_i - \alpha_i - \beta_i = 0, \qquad i \in I_1 \tag{2.60}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \xi_i} = C - \alpha_i - \beta_i = 0, \qquad i \in I_2 \tag{2.61}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \xi_i{}^*} = \frac{1}{\gamma}\xi_i{}^* - \alpha_i{}^* - \beta_i{}^* = 0, \qquad i \in I_1 \tag{2.62}$$

$$\frac{\delta L_{pd}(\boldsymbol{w}, b, \xi, \xi^*, \alpha, \alpha^*)}{\delta \xi_i{}^*} = C - \alpha_i{}^* - \beta_i{}^* = 0, \qquad i \in I_2 \tag{2.63}$$

with $\alpha_i, \alpha_i{}^*, \beta_i, \beta_i{}^* \geq 0$.

The KKT complementarity conditions imply that

$$\alpha_i(\boldsymbol{w}^T\boldsymbol{x}_i + b - y_i - \varepsilon - \xi_i) = 0, \tag{2.64}$$

$$\alpha_i{}^*(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b - \varepsilon - \xi_i{}^*) = 0, \tag{2.65}$$

$$\beta_i\xi_i = 0 \ , \ \beta_i{}^*\xi_i{}^* = 0. \tag{2.66}$$

The result (2.58) represents the solution for the vector of parameters $\boldsymbol{w}$. In addition, β_i is null in the intervals I_1 and I_2 since $\xi_i > 0$ in these intervals.

From previous expressions, we obtain the following equality of Lagrange multipliers α_i (equivalently for ${\alpha_i}^*$):

$$\alpha_i = \begin{cases} C, & \xi_i \geq \varepsilon + \gamma C \\ \frac{\xi_i}{\gamma}, & \varepsilon \leq \xi_i \leq \varepsilon + \gamma C \\ 0, & 0 \leq \xi_i \leq \varepsilon \end{cases}. \tag{2.67}$$

We can now apply the obtained expressions to the Lagrangian (2.56) to obtain the following dual:

$$\begin{aligned} L_d(\alpha, \alpha^*) &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}({\alpha_i}^* - \alpha_i)({\alpha_j}^* - \alpha_j)\ {\boldsymbol{x}_i}^T\boldsymbol{x}_j \\ &+ \sum_{i=1}^{N}(({\alpha_i}^* - \alpha_i)y_i - ({\alpha_i}^* + \alpha_i)\varepsilon) \\ &- \sum_{i\in I_1}\frac{\gamma}{2}({\alpha_i}^2 + {\alpha_i}^{*2}) - \sum_{i\in I_2}\frac{\gamma C^2}{2}. \end{aligned} \tag{2.68}$$

We must also take into account that in the interval I_2, we have α_i and $\alpha_i^* = C$. Therefore, we can change the last term of (2.68) by $\sum_{i\in I_2}\frac{\gamma}{2}({\alpha_i}^2 + {\alpha_i}^{*2})$.
The expression becomes:

$$\begin{aligned} L_d(\alpha, \alpha^*) &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}({\alpha_i}^* - \alpha_i)({\alpha_j}^* - \alpha_j)\ {\boldsymbol{x}_i}^T\boldsymbol{x}_j \\ &+ \sum_{i=1}^{N}(({\alpha_i}^* - \alpha_i)y_i - ({\alpha_i}^* + \alpha_i)\varepsilon - \frac{\gamma}{2}({\alpha_i}^2 + {\alpha_i}^{*2})). \end{aligned} \tag{2.69}$$

In matrix notation, we obtain

$$L_d = -\frac{1}{2}(\alpha - \alpha^*)^T\ \mathbf{R}\ (\alpha - \alpha^*) + (\alpha - \alpha^*)^T\ \mathbf{y} - (\alpha + \alpha^*)\ \mathbf{1}\ \varepsilon. \tag{2.70}$$

This function can be maximized by using the same procedure as for the case of LSVC. Very small eigenvalues can possibly appear, so it is suitable to adjust them numerically by adding a diagonal matrix of low value. The function becomes:

$$L_d = -\frac{1}{2}(\alpha - \alpha^*)^T\ [\mathbf{R} + \gamma\mathbf{I}]\ (\alpha - \alpha^*) + (\alpha - \alpha^*)^T\ \mathbf{y} - (\alpha + \alpha^*)\ \mathbf{1}\ \varepsilon. \tag{2.71}$$

This numerical regularization is equivalent to apply a modified version of cost function. In this equation, the numerical regularization term is γ. The application of this term is equivalent to apply a cost function having a quadratic interval between ε and $\varepsilon + \gamma C$ (see Fig. 2.8).

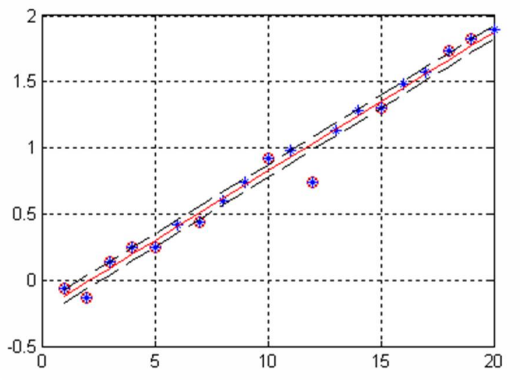


Figure 2.9: $\varepsilon-$insensitive Linear regression with $\varepsilon = .05$ and $C = 10$.

We show an example of one-dimensional regression problem which represents linear data. The solution of this LSVR for the linear $\varepsilon-$insensitive loss function with $\varepsilon = 0.05$ and $C = 10$ is shown in Fig. 2.9. The support vectors are the circled data.

After studying and analyzing the case of linear classification and regression, we examine the case of nonlinear machine. In practice, the linear solution does not perform well for classification and regression problems. Therefore, the nonlinear approach is required. Indeed, a linear machine can be built in a high dimensional space, but still nonlinear in the input space. This is possible thanks to kernel functions that can be interpreted geometrically as a products between the characteristics of learning examples in a feature space different to the input space: this is *the kernel trick*.

2.3 Kernel Functions

The limit of calculation of the linear learning machine has been highlighted in 1960 by Minsky and Papert [49]. Thus, complex real applications require hypothesis spaces more expressive than linear functions.

Kernel representations provide an alternative solution by the projection of data in a high dimension feature space. Therefore, the computational power of the linear learning machines increases. This increase can be achieved in an implicit way by using the dual representation of linear machines. The advantage of using the dual form derives from the fact that the number of parameters to be set does not depend on the number of the used attributes. Replace the dot product by a well chosen *kernel function* allows to carry out a nonlinear projection into a high dimension feature space without increasing the number of parameters to be set. In this section, we discuss the kernel technique that represents a major block of nonlinear SVM.

2.3.1 Learning in the feature space

The complexity and difficulty of learning depends on the representation of the objective function. Ideally, this representation must be chosen to express the specifics of the considered learning problem. For that reason, we use a preprocessing strategy of learning machine that transform the data representation as follows:

$$\boldsymbol{x} = (x_1, \cdots, x_n) \longmapsto \phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \cdots, \phi_N(\boldsymbol{x}))\,. \tag{2.72}$$

This step is equivalent to project the original space $\boldsymbol{X}$ into a new space $\boldsymbol{F} = \{\phi(\boldsymbol{x})/\boldsymbol{x} \in \boldsymbol{F}\}$. The projection of data into another space simplifies greatly the task of learning machine. The quantities introduced to characterize the data are called *features* while the original quantities are called *attributes*. Choose the most suitable representation is known by the name *feature selection*. The space $\boldsymbol{X}$ is the input space while the space $\boldsymbol{F}$ represents the feature space.

Fig. 2.10 shows an example of projection of a two-dimensional input space into a two dimensional feature space. The data can not be separated by a linear function in the input space, but may be in the feature space. The purpose of this section is to show how such projections can be made in high dimensional spaces where linear separation becomes easier.

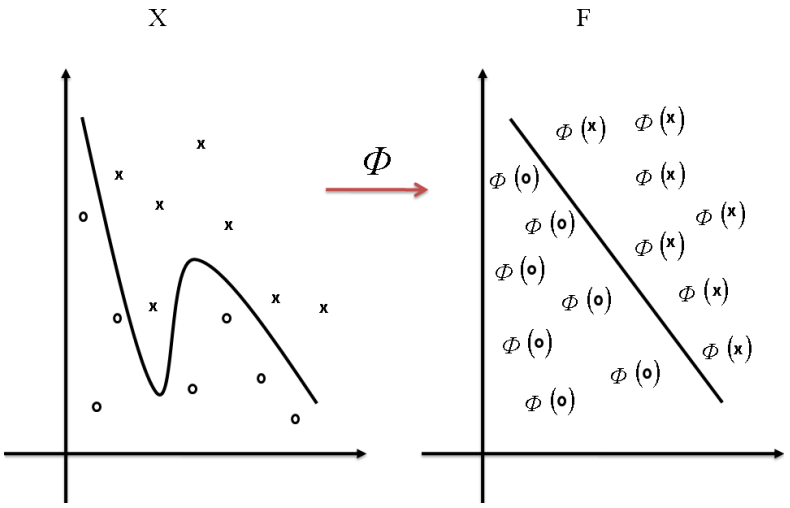


Figure 2.10: Feature space.

Several approaches to select features exist. We always try to identify the smallest set of features that keep the essential information contained in the original attributes. This is known as *dimensionality reduction*:

$$\boldsymbol{x} = (x_1, \cdots, x_n) \longmapsto \phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \cdots, \phi_d(\boldsymbol{x}))\,, \quad d < n \tag{2.73}$$

The selection of characteristics is considered as a part of the learning process itself. Indeed, theoretical learning models consider that the use of a large range of characteristics can create problems of overfitting.

So, by dimensionality reduction, problems of generalization will be avoided. Computational problems will be also avoided through the use of *implicit projection*.

2.3.2 Implicit projection in the feature space

In order to learn nonlinear relationships to learning machines, we must first select a set of nonlinear characteristics, and then rewrite the data in this new representation. This is equivalent to applying a fixed nonlinear projection of data in the feature space, wherein the linear machines may be used. Therefore, the hypothesis set will consist of functions of the form

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \boldsymbol{w}_i \phi_i(\boldsymbol{x}) + b, \tag{2.74}$$

where $\phi : \boldsymbol{X} \longmapsto \boldsymbol{F}$ is a nonlinear projection of the original space $\boldsymbol{X}$ in such a feature space $\boldsymbol{F}$.

This means that we will build nonlinear machines in two steps:

- choose a fixed nonlinear projection that transforms data into a feature space $\boldsymbol{F}$ then,
- classify these data in this feature space by a linear machine.

An important property of linear learning machine is that these machines can be represented in dual form. Implying that the hypothesis can be expressed as a linear combination of learning points. The decision rule can be then evaluated using the dot product between the test points and learning points as follows:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \; y_i < \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}) > + b. \tag{2.75}$$

If we can directly calculate the dot product $< \phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}) >$ in the feature space, it will be possible to combine these two steps to build the nonlinear learning machine. This direct calculation method is called *kernel function*.

Definition 2.1:
The kernel is a function $\boldsymbol{K}$, such that for all $\boldsymbol{x}, \boldsymbol{z} \in \boldsymbol{X}$

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) =< \phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{z}) >, \tag{2.76}$$

where ϕ is a projection of an original space $\boldsymbol{X}$ into a feature space (dot product) $\boldsymbol{F}$.■

The word kernel is derived from the theory of integral operator that describes the relationships between cores and their corresponding characteristics spaces [49]. The use

of kernels allows the projection of data in a feature space implicitly, and then learning linear machine in this space. The only information used for the training samples is the *Gram matrix*. This matrix is also called *kernel matrix* noted by $\boldsymbol{K}$.

The objective of this approach is to find a kernel function that can be evaluated efficiently. Once this function is found, the evaluation of the decision rule can be achieved by N evaluations of the kernel:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \ y_i \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{2.77}$$

An important property of the use of kernels is that we do not need to know the projection of characteristics in order to learn the linear machine in the feature space.

On the other hand, it is clear that kernel theory generalizes the standard dot product in the input space since this dot product is an example of kernel when we put the projection of characteristics equal to the identity.

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{z}) =< \boldsymbol{x}, \boldsymbol{z} > . \tag{2.78}$$

2.3.3 Kernel construction

It appears in the approach of construction of the kernels that we need first to create a complicated feature space, then calculate the dot product in this space, and finally find a direct calculation method based on original inputs. In practice, the construction of kernels occurs by defining a kernel function directly, and thus implicitly define a feature space.

Define a kernel function in the input space is more natural than create a complicated feature space. Mercer's theorem is concerned with the construction of kernels for the case of nonlinear machine. The basic idea is to project the vector $\boldsymbol{x}$ of a finite dimensional space (input space) into a high dimensional Hilbert space (possibly infinite) through a nonlinear transformation $\phi(\cdot)$. The Hilbert space is presented in details in appendix A. The linear machine can be built in a high dimensional space (feature space), but still nonlinear in the input space.

Most transformations $\phi(\cdot)$ are unknown, but their dot product in feature space can be expressed as a function of the input vectors as follows:

$$\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{2.79}$$

In the literature, these spaces are called *reproducing kernel Hilbert spaces* (RKHS), and their dot products $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ are often called *Mercer kernel*. An explicit representation of these vectors in the feature space is not necessary since the SVM formulation presents only dot products of these vectors.

Mercer's theorem gives the condition for which the kernel $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ must satisfy to be a dot product of the Hilbert space.

Theorem 2.1 *(Mercer) Let X be a compact subspace of $\mathbb{R}^n$. Suppose that K is a continuous symmetric function such that the integral operator $T_X : L_2(X) \longmapsto L_2(X)$,*

$$\left(T_X f\right)(\cdot) = \int_X K(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}, \tag{2.80}$$

is positive, then

$$\int_{X \times X} K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z} \geq 0, \tag{2.81}$$

for all $f \in L_2(X)$. One can then extend $K(\boldsymbol{x}, \boldsymbol{z})$ to a uniformly convergent series (on $X \times X$) in terms of eigen-functions $\phi_j \in L_2(X)$, normalized such that $\|\phi_j\|_{L_2(X)} = 1$, with positive associated eigenvalues $\lambda_j \geq 0$,

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{z}). \tag{2.82}$$

Note that the positivity condition:

$$\int_{X \times X} K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{z} \geq 0, \qquad \forall f \in L_2(X) \tag{2.83}$$

implies that for any subset of X, the corresponding matrix is semi-definite positive (having non-negative eigenvalues).

It is not always easy to prove the positivity condition for any function. The first proved kernels satisfying Mercer's theorem are the homogeneous polynomial kernels

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i{}^T + \boldsymbol{x}_j)^p, \tag{2.84}$$

and the heterogeneous polynomial kernels

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i{}^T + \boldsymbol{x}_j + 1)^p. \tag{2.85}$$

The Gaussian kernel is also another important type of kernels, since it is widely used in several applications. It is expressed as

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\sigma^2}}. \tag{2.86}$$

Sigmoid and Dirichlet kernels are other types of kernel expressed respectively as

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\gamma \boldsymbol{x}_i{}^T \boldsymbol{x}_j + \mu), \tag{2.87}$$

and

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\sin((n + 1/2)\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right))}{2\sin((\boldsymbol{x}_i - \boldsymbol{x}_j)/2)}. \tag{2.88}$$

After presenting the kernel functions and their main characteristics, especially their ability to transform a set of nonlinear data in the input space into a set of linear data in the feature space of higher dimension through an implicit projection, we proceed to describe the SVC and SVR in their nonlinear versions.

2.4 Nonlinear Support Vector Machines

2.4.1 Nonlinear SVC

The solution of LSVC is given by

$$\boldsymbol{w} = \sum_{i=1}^{N} y_i \alpha_i \boldsymbol{x}_i. \tag{2.89}$$

In the nonlinear case, the data are projected in the Hilbert space and the solution will be

$$\boldsymbol{w} = \sum_{i=1}^{N} y_i \alpha_i \phi(\boldsymbol{x}_i), \tag{2.90}$$

where $\phi(\cdot)$ denotes the projection function. The vector of parameters $\boldsymbol{w}$ represents a linear combination of the input vectors in the Hilbert space. However, as several transformations $\phi(\cdot)$ are unknown, we can not have an explicit form of these transformations. Yet, the problem can be solved as the machine uses just the dot products of vectors and not the explicit form of transformation. Therefore, we can not use the following expression since no explicit expression of $\boldsymbol{w}$ exists:

$$y_j = \boldsymbol{w}^T \phi(\boldsymbol{x}_j) + b. \tag{2.91}$$

However, by substituting equation (2.90) in (2.91), we obtain the following expression:

$$y_j = \sum_{i=1}^{N} y_i \alpha_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) + b = \sum_{i=1}^{N} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b. \tag{2.92}$$

The resulting machine can be expressed directly in terms of Lagrange multipliers and dot product of kernels. To determine the Lagrange multipliers by solving the corresponding dual problem, we need only the Gram matrix $\boldsymbol{K}$ such that $\boldsymbol{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. To calculate the bias b, we use the expression (2.17) which becomes for the nonlinear SVC case as follows:

$$y_j \left(\sum_{i=1}^{N} y_i \alpha_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) + b \right) - 1 = 0, \tag{2.93}$$

$$y_j \left(\sum_{i=1}^{N} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \right) - 1 = 0, \tag{2.94}$$

for all samples $\boldsymbol{x}_i$ having $\alpha_i < C$.

The SVC solution for the case of nonlinear separable data is illustrated in Fig. 2.11. This solution is obtained using the polynomial kernel of degree 2 with some tolerance of error classification for $C = 10$.

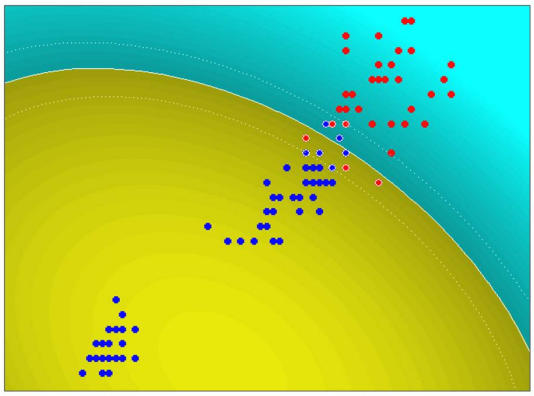


Figure 2.11: Data separation with polynomial kernel of degree 2.

To see the effect of the misclassification error tolerance on the topology of the boundaries of classification, Fig. 2.12 shows the results of the nonlinear SVC with the Gaussian RBF (Radial Basis Function) kernel for different values of C. The width σ of the RBF kernel is calculated from data by $\sigma = \sqrt{\max_{ij} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| / 2}$ [53, 54]. The values $C = 1$ and $C = 1000$ offer good solutions, depending on the topology of the most appropriate boundary (open such 2.12 (a) or closed such 2.12 (c). From that reason, we deduce that the parameter C should be chosen after knowledge of the considered problem for determining the final solution.

2.4.2 Nonlinear SVR

The solution of linear SVR is expressed by

$$\boldsymbol{w} = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \boldsymbol{x}_i. \tag{2.95}$$

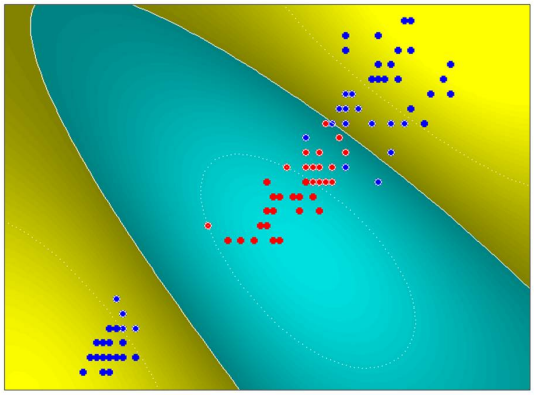
As in the case of nonlinear SVC, the solution of the nonlinear SVR has the following form:

$$\boldsymbol{w} = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \phi(\boldsymbol{x}_i). \tag{2.96}$$

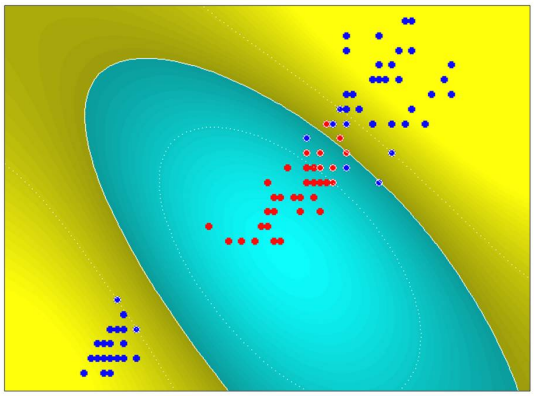
Therefore, the expression of the nonlinear SVR is as follows:

$$y_j = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) + b = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b. \tag{2.97}$$

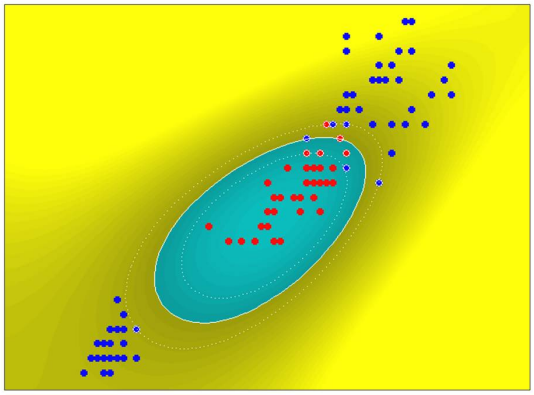
The construction of the nonlinear SVR is almost identical to the construction of the nonlinear SVC.



(a)

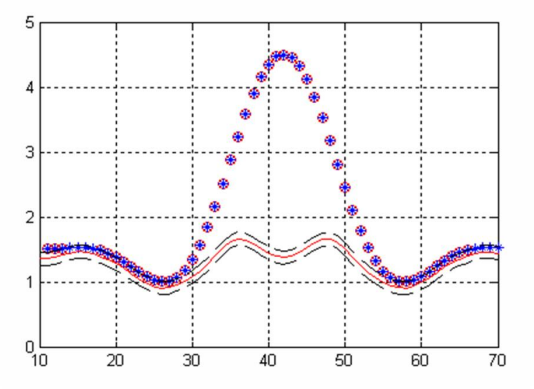


(b)

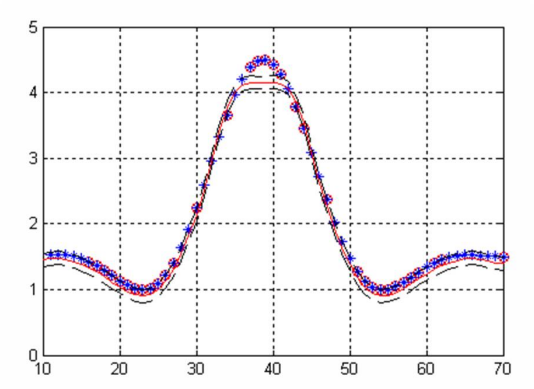


(c)

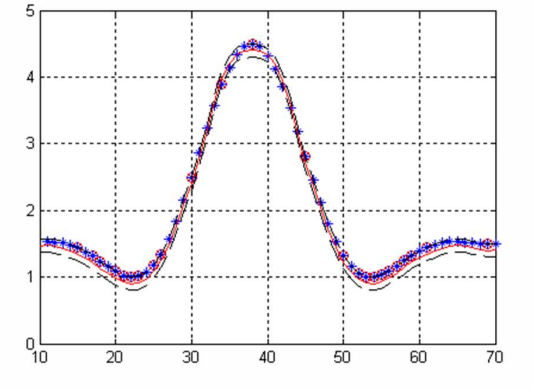
Figure 2.12: Data separation with RBF kernel for different values of C (a) $C = 1$, (b) $C = 10$ and (c) $C = 1000$.



(a)



(b)



(c)

Figure 2.13: Nonlinear regression using RBF kernel with $\varepsilon = .05$ for (a) $C = .1$, (b) $C = 1$ and (c) $C = 10$.

The dual problem for the corresponding nonlinear SVR with $\varepsilon-$insensitive loss function is equivalent to the LSVR in (2.31) except that the dot product $< \boldsymbol{x}_i \cdot \boldsymbol{x}_j >= \boldsymbol{x}_i{}^T \boldsymbol{x}_j$ is replaced by the kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$:

$$\begin{aligned} maximize_{\alpha,\alpha^*} \quad L_d(\alpha,\alpha^*) &= \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i)y_i - \varepsilon\sum_{i=1}^{N}(\alpha_i{}^* + \alpha_i) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i{}^* - \alpha_i)(\alpha_j{}^* - \alpha_j)K(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad (2.98) \\ s.t \quad 0 &\leq \alpha_i, \alpha_i{}^* \leq C \\ 0 &= \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i). \quad (2.99) \end{aligned}$$

The dual problem of the nonlinear SVR corresponding to the quadratic $\varepsilon-$insensitive loss function is as follows:

$$\begin{aligned} maximize_{\alpha,\alpha^*} \quad L_d(\alpha,\alpha^*) &= \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i)y_i - \varepsilon\sum_{i=1}^{N}(\alpha_i{}^* + \alpha_i) \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i{}^* - \alpha_i)(\alpha_j{}^* - \alpha_j)(K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &+ \frac{1}{C}\delta_{ij}) \quad (2.100) \\ s.t \quad 0 &\leq \alpha_i, \alpha_i{}^* \leq C \\ 0 &= \sum_{i=1}^{N}(\alpha_i{}^* - \alpha_i). \quad (2.101) \end{aligned}$$

The KKT complementarity conditions are the same as LSVR.

The examples given in Fig. 2.13 are illustrative examples that present a regression problem with nonlinear data. To illustrate some solutions of the nonlinear SVR, we used several values of C to model the nonlinear data. The $\varepsilon-$insensitive function ($\varepsilon = .05$) with linear loss is used with RBF kernel. Fig. 2.13(a) shows the solution for $C = .1$. Obviously the solution is unable to model data for this value of C. Fig. 2.13(b) illustrates the solution for $C = 1$. It is remarkable that the solution is unable to correctly model the data summit. Fig. 2.13(c) with $C = 10$ correctly models data and the solution is in the $\varepsilon-$insensitivity area. Typically, a selection of the capacity with methods such as cross validation is recommended. Generally, in the regression technique, it is necessary to select the representative loss function and the optimum value of capacity. These considerations should be based on the knowledge of the problem and the noise distribution. In the absence of these information, the robust loss function presented in Fig. 2.8 has shown good performance [46].

2.5 Conclusion

This chapter shows how the results of learning theory can be used to avoid the difficulties of applying linear functions in high dimension features spaces with induced kernels. We have also shown how optimization problems can be transformed under a quadratic convex dual form for each adapted approaches for classification and regression in their linear and nonlinear versions. In the case of regression, the loss function penalizes errors only above the threshold ε. Such a loss function typically leads to a fine representation of the decision rule giving significant advantages.

In this chapter, we have described the linear support vector machine for classification and regression. Also, the theory of kernel functions associated with SVM was analyzed. Finally, we have presented classification and regression support vector machine in their nonlinear version. Nonlinear support vector regression applied to channel estimation in the OFDM-LTE communication system will be developed later.

Chapter 3

Mobile Radio Channel in OFDM Communication System

3.1 Introduction

The objective of this chapter is modeling and simulation of mobile radio channel for LTE communication system. We introduce first the principle of Orthogonal Frequency Division Multiplexing (OFDM) transmission scheme. Then, we explore a mathematical framework of multipath Rayleigh channel. After that, we present an overview of the nonlinear impulsive noise which can frequently occur in communication systems. In addition, the LTE-3GPP standardized power delay profiles are presented. Then, in order to evaluate the performance of the channel estimation techniques, the mobile radio channels based on several scenarios are simulated.

3.2 OFDM Communication System

Multiple carriers can be used for high data rate transmission to overcome the frequency selectivity of the wideband channel experienced by single-carrier transmission. Indeed, the wideband signal can be considered as several narrowband signals. So that the frequency-selective wideband channel can be approximated by multiple frequency-flat narrowband channels. Thus, the complexity of the equalizer for each subchannel can be reduced by the frequency-nonselectivity of narrowband channels. As long as the orthogonality among the subchannels is maintained, the inter-carrier interference (ICI) can be suppressed, leading to distortionless transmission [55].

3.2.1 OFDM transmission scheme

Orthogonality

OFDM transmission scheme is a type of a multichannel system which employs multiple orthogonal subcarriers. These orthogonal signals are overlapped in spectrum.

In practice, Discrete Fourier Transform (DFT) and inverse DFT (IDFT) processes are useful for implementing these orthogonal signals. These DFT and IDFT can be implemented efficiently by using Fast Fourier Transform (FFT) and inverse FFT (IFFT), respectively.

Let $\left\{e^{j2\pi f_k t}\right\}_{k=0}^{N-1}$ be the time-limited complex exponential signals which represent different subcarriers at $f_k = k/T$ in the OFDM signal, where $0 \leq t \leq T$.

If the integral of the products for the common period of signals T is zero, then these signals are defined to be orthogonal, that is

$$\begin{aligned} \frac{1}{T}\int_0^T e^{j2\pi f_k t} e^{-j2\pi f_i t}\, dt &= \frac{1}{T}\int_0^T e^{j2\pi \frac{k}{T} t} e^{-j2\pi \frac{i}{T} t}\, dt \\ &= \frac{1}{T}\int_0^T e^{j2\pi \frac{(k-i)}{T} t}\, dt \\ &= \begin{cases} 1, & \forall\ k = i \\ 0, & otherwise. \end{cases} \end{aligned} \tag{3.1}$$

Equation (3.1) can be written in the discrete time domain, by tacking the discrete samples at the sampling instances $t = n\ T_s = n\ \frac{T}{N}$ for $n = 0, 1, \cdots, N-1$, where T_s denotes the subcarrier duration, as follows:

$$\begin{aligned} \frac{1}{N}\sum_{n=0}^{N-1} e^{j2\pi \frac{k}{T} n T_s} e^{-j2\pi \frac{i}{T} n T_s} &= \frac{1}{N}\sum_{n=0}^{N-1} e^{j2\pi \frac{k}{T} \frac{nT}{N}} e^{-j2\pi \frac{i}{T} \frac{nT}{N}} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} e^{j2\pi \frac{(k-i)}{N} n} \\ &= \begin{cases} 1, & \forall\ k = i \\ 0, & otherwise. \end{cases} \end{aligned} \tag{3.2}$$

So, the OFDM signal is ICI-free if the above orthogonality condition is satisfied.

OFDM modulation and demodulation techniques

The OFDM transmitter maps the message bits into a sequence of QAM symbols which will be converted into N parallel streams by serial-to-parallel (S/P) conversion. Each of these N symbols is carried out by a different subcarrier.

Let $X_l[k]$ be the l^{th} transmit symbol at the k^{th} subcarrier, where $l = 0, 1, \cdots, \infty$ and $k = 0, 1, \cdots, N-1$. The transmission duration of N symbols is extended to NT_s, which forms a single OFDM symbol with a length of T ($T = NT_s$) due to the S/P conversion.

Let $\Upsilon_{l,k}(t)$ be the l^{th} OFDM signal at the k^{th} subcarrier, expressed as

$$\Upsilon_{l,k}(t) = \begin{cases} e^{j2\pi f_k (t - lT)}, & 0 < t \leq T \\ 0, & elsewhere. \end{cases} \tag{3.3}$$

In continuous-time domain, the passband and baseband OFDM signals can be represented respectively as [56]

$$x_l(t) = \Re\left\{\frac{1}{T}\sum_{l=0}^{\infty}\left\{\sum_{k=0}^{N-1} X_l[k]\Upsilon_{l,k}(t)\right\}\right\} \tag{3.4}$$

and

$$x_l(t) = \sum_{l=0}^{\infty}\sum_{k=0}^{N-1} X_l[k]e^{j2\pi f_k(t-lT)}. \tag{3.5}$$

Equation (3.5) presents the continuous-time baseband OFDM signal which can be sampled at $t = lT + nT_s$ with $T_s = T/N$ and $f_k = k/T$ to yield the following discrete-time OFDM symbol:

$$x_l[n] = \sum_{k=0}^{N-1} X_l[k]e^{j2\pi kn/N}, \quad n = 0, 1, \cdots, N-1. \tag{3.6}$$

Equation (3.6) presents the N-point IDFT of QAM data symbols $X_l[k],\ k = 0, \cdots, N-1$ which can be well computed by the IFFT algorithm.

Consider now the received baseband OFDM symbol $y_l(t) = \sum_{k=0}^{N-1} X_l[k]e^{j2\pi f_k(t-lT)}$ with $lT < t \leq lT + nT_s$, where the noise and channel effects are not taken into account. The transmitted symbol $X_l[k]$ can be reconstructed from $y_l(t)$ using the orthogonality among the subcarriers described in equation (3.1) as follows:

$$\begin{aligned} Y_l[k] &= \frac{1}{T}\int_{-\infty}^{\infty} y_l(t)e^{-j2\pi f_k(t-lT)}\ dt \\ &= \frac{1}{T}\int_{-\infty}^{\infty}\left\{\sum_{i=0}^{N-1} X_l[i]e^{j2\pi f_i(t-lT)}\right\}\ e^{-j2\pi f_k(t-lT)}\ dt \\ &= \sum_{i=0}^{N-1} X_l[i]\left\{\frac{1}{T}\int_0^T e^{j2\pi(f_i-f_k)(t-lT)}\ dt\right\} = X_l[k]. \end{aligned} \tag{3.7}$$

Let $\{y_l[n]\}_{n=0}^{N-1}$ be the sample values of the received OFDM symbol $y_l(t)$ at $t = lT + nT_s$. So, the integration in modulation process presented in equation (3.7) can be represented in discrete time domain as

$$\begin{aligned} Y_l[k] &= \sum_{n=0}^{N-1} y_l[n]e^{-j2\pi kn/N} \\ &= \sum_{n=0}^{N-1}\left\{\frac{1}{N}\sum_{i=0}^{N-1} X_l[i]e^{j2\pi in/N}\right\}e^{-j2\pi kn/N} \\ &= \frac{1}{N}\sum_{n=0}^{N-1}\sum_{i=0}^{N-1} X_l[i]e^{j2\pi(i-k)n/N} = X_l[k]. \end{aligned} \tag{3.8}$$

Equation (3.8) is the N-point DFT of $y_l[n],\ n = 0, \cdots, N-1$ which can be well computed by the FFT algorithm.

3.2.2 Guard interval

Multipath channel effect on OFDM symbol

Let $x_l(t) = \sum_{k=0}^{N-1} X_l[k] e^{j2\pi f_k (t-lT)},\ lT < t \leq lT + nT_s$, be the l^{th} OFDM signal. The received signal for a channel with an impulse response $h_l(t)$ is given by

$$\begin{aligned} y_l(t) &= x_l(t) * h_l(t) + z_l(t) \\ &= \int_0^{\infty} h_l(\tau) x_l(t-\tau)\ dt + z_l(t),\quad lT < t \leq lT + nT_s \end{aligned} \tag{3.9}$$

where $z_l(t)$ denotes the additive white Gaussian noise (AWGN) process.

Equation (3.9) can be represented in discrete time domain by taking the samples at $nT_s = nT/N$ as follows:

$$\begin{aligned} y_l[n] &= x_l[n] * h_l[n] + z_l[n] \\ &= \sum_{m=0}^{\infty} h_l[m] x_l[n-m] + z_l[n], \end{aligned} \tag{3.10}$$

where $y_l[n] = y_l(nT_s), x_l[n] = x_l(nT_s), h_l[n] = h_l(nT_s)$ and $z_l[n] = z_l(nT_s)$.

The frequency domain symbol $X[k]$ modulates the subcarrier with a frequency $f_k = k/T$, and can be demodulated by using the orthogonality among the subcarriers at the receiver. The original symbol $X[k]$ has a duration equal to T_s, however its length has been extended to become $T = NT_s$ by transmitting N symbols in parallel since the OFDM symbol is a composite signal of N symbols transmitted in parallel form which has a duration equal to T. Let T_u denotes the effective OFDM symbol duration without guard interval. Note that the frequency spacing between subcarriers is $\triangle f = 1/T_u$. By extending the symbol duration by N times, the multipath fading channel effect can be deeply reduced on the OFDM symbol. Yet, its effect still remains as a destructive factor that may rupture the orthogonality among the subcarriers in the OFDM system.

Inter-Symbol Interference (ISI) can occur when a first received symbol is mixed up with a second received symbol. It is then evident that all subcarriers are no longer orthogonal over the duration of each OFDM symbol. So that, there must be some means to deal with the ISI effect over the multipath channel to guarantee the performance of transmission. A guard interval between two consecutive OFDM symbols will be necessary, as discussed in the sequel.

Cyclic Prefix

Two different techniques can explain the OFDM guard interval insertion. The first one is the zero padding (ZP) technique that pads the guard interval with zeros. The

other one is the cyclic extension of the OFDM symbol by a cyclic prefix (CP) for assuring some continuity. The CP consists of extending the OFDM symbol by copying its last samples into its front. Let T_{cp} be the length of the CP. Then, the duration of the extended OFDM symbol becomes $T = T_u + T_{cp}$. Note that the length of the cyclic prefix must be set equal to or longer than the maximum delay of the multipath channel. This is for assuring that the cyclic prefix confines the effect of the ISI of an OFDM symbol on the next OFDM symbol. Consequently, the orthogonality among the subcarriers will be maintained. Indeed, the orthogonality of each delayed subcarrier with all other subcarriers is maintained over T_u because its continuity has been warranted by the cyclic prefix, such that

$$\frac{1}{T_u}\int_0^{T_u} e^{j2\pi f_k(t-t_0)}e^{-j2\pi f_i(t-t_0)}\,dt = 0, \quad k \neq i \tag{3.11}$$

for the first OFDM signal that comes with a delay of t_0, and

$$\frac{1}{T_u}\int_0^{T_u} e^{j2\pi f_k(t-t_0)}e^{-j2\pi f_i(t-t_0-T)}\,dt = 0, \quad k \neq i \tag{3.12}$$

for the second OFDM signal that comes with a delay of $t_0 + T$.

Now, at the receiver side, suppose that the CP length is set longer than the maximum delay of the channel, then the FFT of the received samples $\{y_l[n]\}_{n=0}^{N-1}$ is taken by the OFDM receiver to yield

$$\begin{aligned}
Y_l[k] &= \sum_{n=0}^{N-1} y_l[n]e^{-j2\pi kn/N} \\
&= \sum_{n=0}^{N-1}\left\{\sum_{m=0}^{\infty} h_l[m]x_l[n-m] + z_l[n]\right\} e^{-j2\pi kn/N} \\
&= \sum_{n=0}^{N-1}\left\{\sum_{m=0}^{\infty} h_l[m]\left\{\frac{1}{N}\sum_{i=0}^{N-1} X_l[i]e^{j2\pi i(n-m)/N}\right\}\right\} e^{-j2\pi kn/N} + Z_l[k] \\
&= \frac{1}{N}\sum_{i=0}^{N-1}\left\{\left\{\sum_{m=0}^{\infty} h_l[m]e^{-j2\pi im/N}\right\} X_l[i]\sum_{n=0}^{N-1} e^{-j2\pi(k-i)n/N}\right\} + Z_l[k] \\
&= H_l[k]X_l[k] + Z_l[k],
\end{aligned} \tag{3.13}$$

where $Y_l[k], X_l[k], H_l[k]$ and $Z_l[k]$ are the k^{th} subcarrier frequency component of the l^{th} received symbol, transmitted symbol, channel frequency response and noise, respectively in the frequency domain.

Equation (3.13) implies that the OFDM system can be considered as a multiplication between the input symbol and the channel frequency response in the frequency domain. Therefore, it can be easily seen that the transmitted symbol can be detected by one tap equalization since $Y_l[k] = H_l[k]X_l[k]$, which just divides the received symbol by the channel response ($X_l[k] = Y_l[k]/H_l[k]$).

We notice that $Y_l[k] = H_l[k]X_l[k]$ only when $y_l[n] = x_l[n] \otimes h_l[n]$ where $\otimes$ is the operation of the circular convolution. Indeed, $Y_l[k] = H_l[k]X_l[k]$ is obtained in the receiver side thanks to CP insertion in the transmitter which makes the input samples circularly-convolved with the channel samples.

3.3 Radio Channel in Mobile Communication System

3.3.1 Multipath Rayleigh channel model

Multipath Rayleigh channel consists of a reasonable model when there are a lot of objects that scatter the radio signal in the environment before it arrives at the receiver. If there is sufficiently much scatters, the central limit theorem holds that the channel impulse response will be well-modeled as a Gaussian process respective to the distribution of the single component. Such a process will have zero mean and phase evenly distributed between 0 and 2π radians if there is no dominant component to the scatter. Therefore, the envelope of the channel response will be Rayleigh distributed. Rician fading may be applied if there is a dominant line of sight.

A mathematical representation of the real part of the channel is provided by the modeling of the transmission channel. Let $s(t)$ be a signal with carrier frequency f_c, then $s(t)$ has the general representation

$$s(t) \quad = \quad \Re\left\{s_b(t)e^{j2\pi(f_c t+\phi)}\right\}, \tag{3.14}$$

where $s_b(t)$ and ϕ represent the baseband signal of $s(t)$ and the initial phase of the local oscillator at the transmitter, respectively.

The signal $s(t)$ emitted into wireless channel undergoes reflections, diffractions and scattering on different obstacles to generate M replicas having diverse paths. Each path is diffracted into N_M under-paths. According to the law of superposition [57], the received signal $r(t)$ can be expressed as

$$r(t) \quad = \quad \Re\left\{\sum_{m=0}^{M-1}\sum_{n=0}^{N_M-1} a_{m,n}\ s_b(t-\tau_{m,n})e^{j(2\pi(f_c+f_{m,n})(t-\tau_{m,n})+\phi)}\right\}, \tag{3.15}$$

where $a_{m,n}$, $\tau_{m,n}$ and $f_{m,n}$ represent attenuation, delay and Doppler shift for the n^{th} under-path from the m^{th} path, respectively.

On the other hand, under-paths in the same path m can be considered as very similar, so that it is possible to approximate the delays $\tau_{m,n}$ by the average delay of the path m as follows:

$$\tau_m \quad = \quad \frac{1}{N_M}\sum_{n=0}^{N_M-1}\tau_{m,n}. \tag{3.16}$$

The expression of the received signal $r(t)$ becomes after using equation (3.16) as follows:

$$r(t) \quad = \quad \Re\left\{\sum_{m=0}^{M-1}\sum_{n=0}^{N_M-1} a_{m,n}\ s_b(t-\tau_m)e^{j(2\pi(f_c+f_{m,n})(t-\tau_m)+\phi)}\right\}. \tag{3.17}$$

The above expression can be rewritten as

$$r(t) = \Re\left\{\sum_{m=0}^{M-1}\left(\sum_{n=0}^{N_M-1} a_{m,n}e^{j(2\pi f_c t+2\pi f_{m,n}t-2\pi(f_c+f_{m,n})\tau_m+\phi)}\right)s_b(t-\tau_m)\right\} \quad (3.18)$$

The simplified expression of the baseband received signal is equal to

$$r_b(t) = \sum_{m=0}^{M-1} h_m(t)s_b(t-\tau_m), \quad (3.19)$$

where $h_m(t)$ is given by

$$h_m(t) = \sum_{n=0}^{N_M-1} c_{m,n}(t), \quad (3.20)$$

with $c_{m,n}(t) = a_{m,n}e^{j(2\pi f_{m,n}t+\theta_{m,n}+\phi)}$ and $\theta_{m,n} = -2\pi(f_c + f_{m,n})\tau_m$.

When the propagation environment is changing over time or when the receiver or transmitter is moving, the impulse response of the propagation channel $h(t,\tau)$ represents a function of two dimensions: time and delay spread. It can be expressed as [57]

$$h(t,\tau) = \sum_{m=0}^{M-1} h_m(t)\delta(t-\tau_m). \quad (3.21)$$

We notice that the term $h_m(t)$ can be interpreted as the channel impulse response of the m^{th} path.

3.3.2 Generation of a single path impulse response $h_m(t)$

The method of generating the impulse response $h_m(t)$ is based on the expression (3.20) where $h_m(t)$ is varying according to the following parameters:

- $\theta_{m,n}$: is a random phase. This phase is caused by rays reflection, and obtained by selection of a uniform random variables from $[0, 2\pi[$.
- $a_{m,n}$: represents attenuation, and it can be obtained by selection of a uniform random variables from $]0, \sqrt{\frac{{\sigma_m}^2}{N_M}}[$, where ${\sigma_m}^2$ is the variance of the m^{th} path.
- According to Jakes model, scatters are assumed to be uniformly distributed around a circle at angles $n\delta\theta$ with n under-paths emerging from each scatter. The Doppler shift on the n^{th} under-path of the m^{th} path is $f_{m,n} = f_d\ cos(n\delta\theta)$ where f_d represents the maximum Doppler frequency and $\delta\theta = \frac{2\pi}{N_M}$ [58].

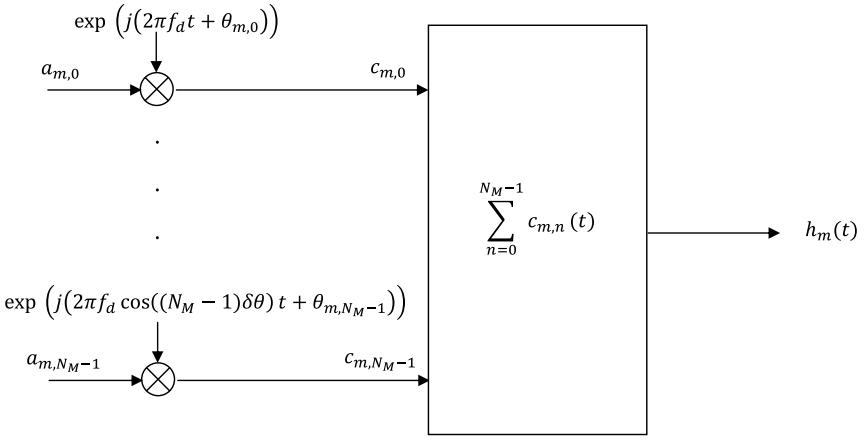


Figure 3.1: Generation of $h_m(t)$ of a single path.

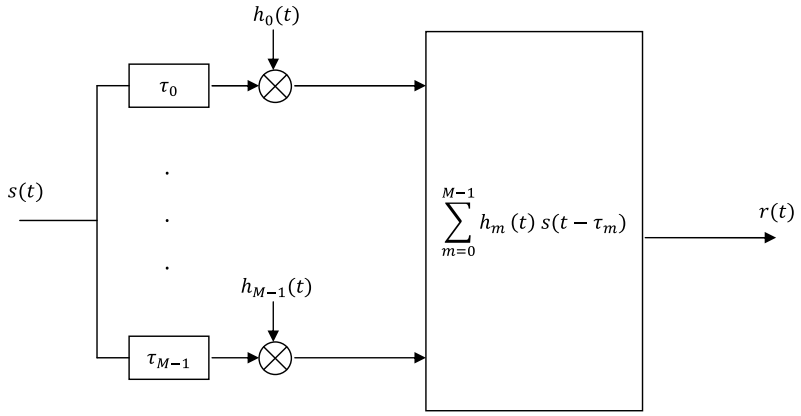


Figure 3.2: Diagram of multi-path channel simulation.

3.3.3 Generation of the channel frequency response

We can generate the equivalent (total) channel impulse response according to the expression given by [57] which is based on the impulse response $h_m(t)$ of a single path as follows:

$$h_k^{eq}(t) \quad = \quad \sum_{m=0}^{M-1} h_m(t, \tau_m) sinc(\frac{\tau_m}{T_e} - k), \tag{3.22}$$

where T_e represents the average delay related to the used channel type.

We present the diagram of a channel impulse response simulator of the k^{th} subcarrier in Fig. 3.3.
After the generation of $h_m(t)$ and its equivalent channel, we can get the channel frequency response as following, where we define $h_{k,i}^{eq}(t)$ as

$$h_{k,i}^{eq}(t) \quad = \quad h_k^{eq}(t), \quad t \in [iT_u, (i+1)T_u\,[. \tag{3.23}$$

Therefore, the channel frequency response during the i^{th} OFDM symbol can be expressed as

$$\Big(H_{0,i}^{eq} H_{1,i}^{eq} \cdots H_{N-1,i}^{eq}\Big) \quad = \quad DFT\Big(h_{0,i}^{eq} h_{1,i}^{eq} \cdots h_{N-1,i}^{eq}\Big). \tag{3.24}$$

Then, the input/output relation that summarizes the discrete model simulator for an OFDM system is given by

$$Y_{k,i} \quad = \quad H_{k,i}^{eq} X_{k,i} + Z_{k,i}. \tag{3.25}$$

Fig. 3.4 presents a simplified simulation diagram for an OFDM transmission system operating in a multipath environment.

3.4 Impulsive Noise

The noise can be considered as an unwanted signal that interferes with the communication signal. Noise can be present in different degrees in almost all environments. In digital communication, there may be several varieties of types and sources of noise such that:

- electronic noise (such as thermal noise)
- acoustic noise (caused by moving vehicle)
- electrostatic noise (caused by the presence of voltage)
- electromagnetic noise (caused by interference between transmission and reception signals over radio frequency spectrum)
- communication channel fading and signal distortion.

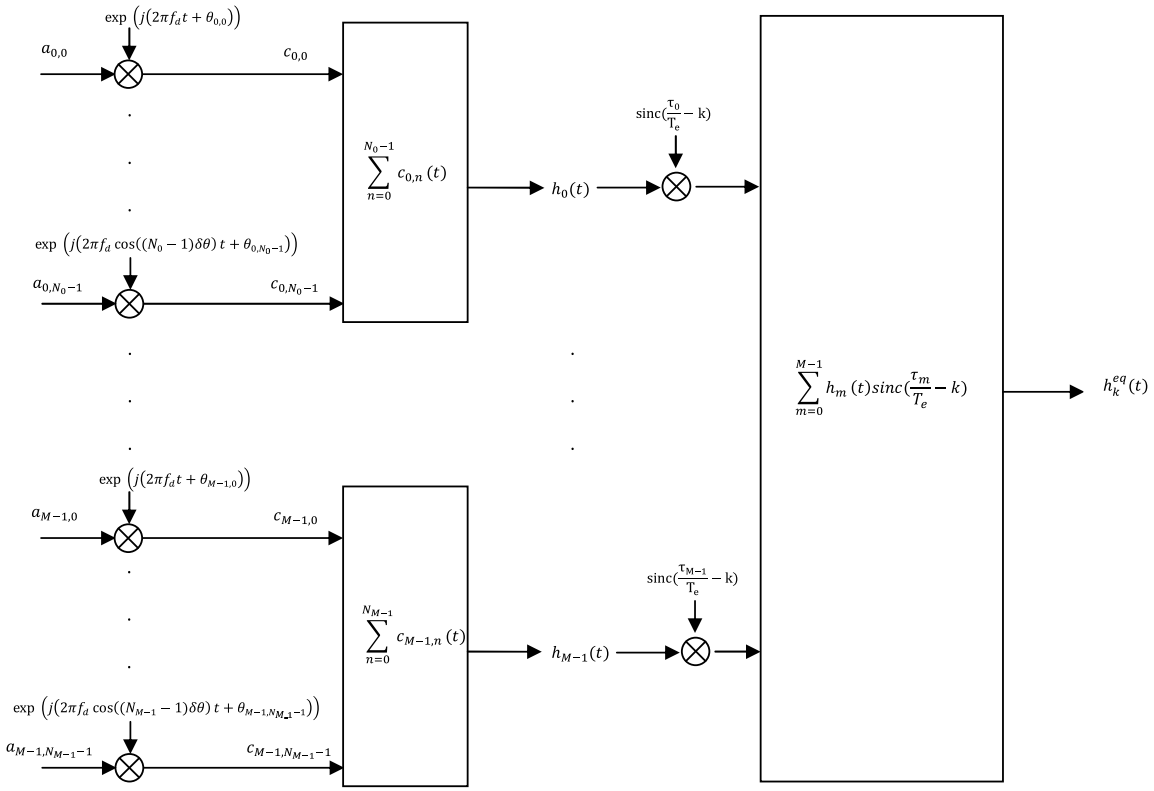


Figure 3.3: Diagram of a channel impulse response simulator of the k^{th} subcarrier.

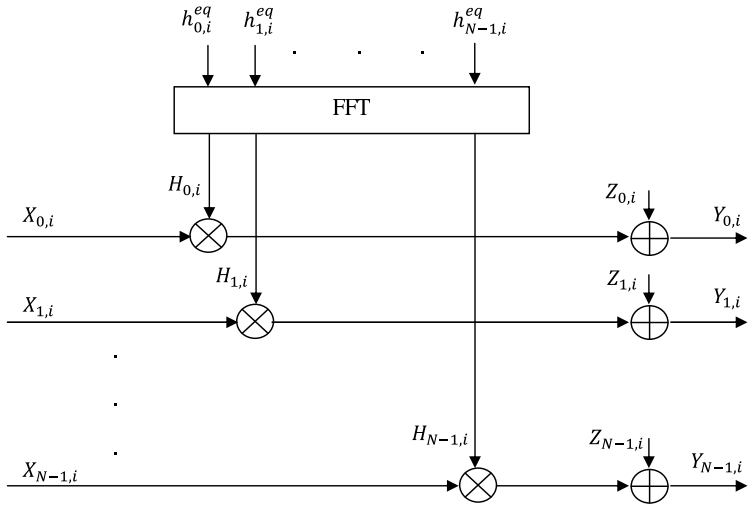


Figure 3.4: Simulation diagram for an OFDM transmission system.

Signal distortion describes the undesirable changes in a signal due to multipath effects and signal fading impact. Distortion and noise are the major factors that limit the precision of results in signal measurement systems and the capacity of data transmission. Thus, the modeling and removal of the effects of distortion and noise are important problems in many applications such as digital cellular mobile communication.

On the other hand, depending on its time characteristics or frequency spectrum, a noise process can be classified into several categories such as, white noise, narrow band noise, colored noise, impulsive noise, etc.

We will interest in this section to describe the nonlinear impulsive noise which will be exploited later in our simulations. Indeed, impulsive noise consists of random short duration noise pulses caused by a channel effects and interfering sources, such as electromagnetic interference, switching noise, adverse channel environments in a communication system, signal dropouts, data packet loss, etc [59].

Fig. 3.5 shows an example of impulsive noise combined with additive Gaussian noise represented in time domain.

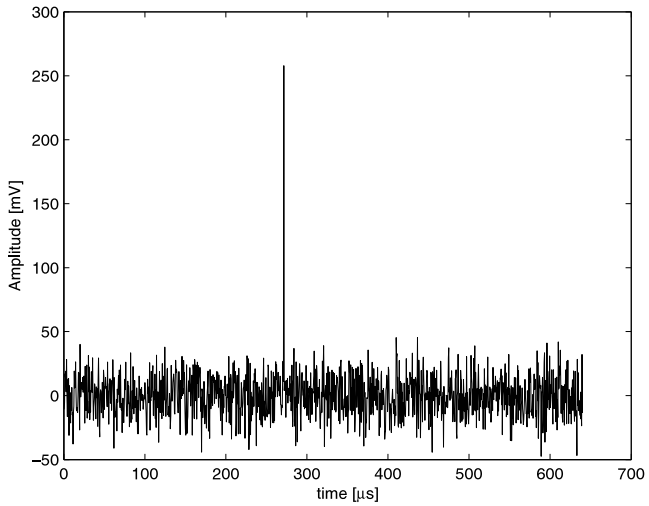


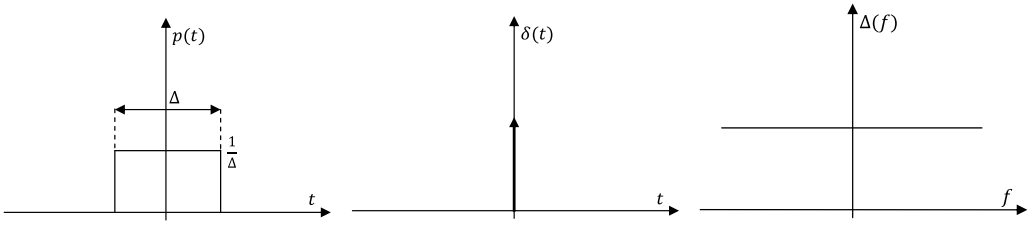
Figure 3.5: Impulsive noise with Gaussian noise in time domain [60].

3.4.1 Theoretical impulse function

In continuous time domain, the formulation of the mathematical concept of an impulse is described in Fig. 3.6.

Let $p(t)$ be a unit-area pulse. The pulse tends to an impulse with infinite power, energy and amplitude as the pulse width Δ tends to zero while the area under the pulse remains equal to unity.

$$\delta(t) = \lim_{\Delta \to 0} p(t) = \begin{cases} \frac{1}{\Delta}, & |t| \leq \Delta/2 \\ 0, & |t| > \Delta/2. \end{cases} \tag{3.26}$$



(a) (b) (c)

Figure 3.6: The impulse representation: (a) A unit-area pulse, (b) The pulse becomes an impulse as $\Delta \to 0$ and (c) The spectrum of the impulse function.

The integral of the impulse function is expressed as

$$\int_{-\infty}^{\infty} \delta(t)dt = \Delta \times \frac{1}{\Delta} = 1. \tag{3.27}$$

The Fourier transform of the impulse function is given by

$$\Delta(f) = \int_{-\infty}^{\infty} \delta(t)e^{-j2\pi ft}dt = 1. \tag{3.28}$$

In discrete-time domain, the impulse $\delta(n)$ can be represented by a signal with an 'on' duration of one sample as follows:

$$\delta(n) \quad = \quad \begin{cases} 1, & n = 0 \\ 0, & otherwise \end{cases} \tag{3.29}$$

where n is the discrete-time index. This function is known by the Kronecker delta function.

The frequency spectrum of a digital impulse can be derived using the Fourier transform relation as

$$\Delta(f) = \sum_{n=-\infty}^{\infty} \delta(n)e^{-j2\pi fn} = 1 \qquad -\infty < f < \infty. \tag{3.30}$$

3.4.2 The response of a communication system to an impulse

Because the realization of a theoretical impulse would require immeasurable power as energy would be delivered in infinitesimally small time, theoretical impulses of infinitesimal duration don't exist in real-life engineering systems. Real impulsive-type noise in communication systems has a finite non-zero duration, that is usually more than one sample long in sampled discrete-time form [59].

The impulse response is defined as the response of a system to an impulse. An impulsive noise begins at some point in space and time and then propagates through the channel to the receiver. Thus, the temporal-spectral shape of the received noise is affected by the channel. Generally, the response of a communication channel may be

stationary or time varying, linear or nonlinear. Moreover, many communication systems exhibit nonlinear characteristics in response to a large amplitude impulse.

The choice of an appropriate domain for signal representation (time domain or frequency domain) has a significant consideration in the development of a noise processing system. The choice should be determined according to the special objective of the system.

- In signal classification and parameter estimation, the objective is to compensate the average effects of the noise over a number of samples, and in many cases, it is more appropriate to use frequency domain signal processing for the impulsive noise since the mean of the power spectrum of the signal is changed by the noise effect.
- In signal restoration, the objective consists of separating the noise from the signal. Impulsive noise is usually more detectable and distinct in the time domain, so it is appropriate to process the impulsive noise in time domain.

3.4.3 Power spectrum of impulsive noise

Impulsive noise represents a non-stationary, binary-state sequence of impulses with random times of occurrence and random amplitudes. The non-stationary nature of impulsive noise can be evaluated by the power spectrum of an impulsive noise process. In fact, when an impulse is present, the noise power represents the power of the impulse, and when the noise is absent the process has zero power.
Thus, the power spectrum and the autocorrelation functions of an impulse noise are time-varying binary state processes.

An impulsive noise sequence can be modeled as a binary-state amplitude-modulated sequence, and expressed as follows:

$$i(m) = n(m)b(m), \tag{3.31}$$

where $n(m)$ is a random noise process and $b(m)$ is a state function which represents a random binary sequence of ones and zeros.

Suppose that impulsive noise is an uncorrelated random process, then the autocorrelation function of impulsive noise can be defined as a binary-state process as follows:

$$r_{ii}(k, m) = E[i(m)i(m + k)] = {\sigma_i}^2 \delta(k)b(m). \tag{3.32}$$

Since the impulsive noise is assumed to be uncorrelated, the autocorrelation is equal to zero for $k \neq 0$, thus equation (3.32) can be expressed as

$$r_{ii}(0, m) = {\sigma_i}^2 b(m). \tag{3.33}$$

By taking the Fourier transform of the autocorrelation function, the power spectrum of an impulsive noise sequence can be given by

$$P_{II}(f, m) = {\sigma_i}^2 b(m). \tag{3.34}$$

The autocorrelation and power spectrum expressions are represented as binary-state functions that at time m, depend on the 'on/off' state of impulsive noise.

3.4.4 Probability models of impulsive noise

In this section, two statistical processes for characterization and modeling of an impulsive noise as a binary-state amplitude-modulated sequence are considered. These processes are Bernoulli-Gaussian process and Poisson-Gaussian process.

3.4.4.1 Bernoulli-Gaussian model of impulsive noise

In a Bernoulli-Gaussian model of impulsive noise, the amplitude of the impulses is modeled by a Gaussian process $g(m)$ and the random time of occurrence of the impulses is modeled by a binary Bernoulli process $b(m)$. A Bernoulli process $b(m)$ represents a binary-valued process that takes a value of '0' with a probability of $1-p$ and a value of '1' with a probability p. Thus, $b(m)=0$ signals the absence of an impulse whereas $b(m)=1$ signals the presence of an impulse.
The probability function of a Bernoulli random process is represented by

$$Pr(b(m)) = \begin{cases} p, & b(m)=1 \\ 1-p, & b(m)=0. \end{cases} \tag{3.35}$$

The mean of a Bernoulli process is given by

$$\begin{aligned} E[b(m)] &= p\times 1+(1-p)\times 0 \\ &= p, \end{aligned} \tag{3.36}$$

and its variance is expressed as

$$\begin{aligned} {\sigma_b}^2 &= E\left\{[b(m)-E(b(m))]^2\right\} = E[(b(m)-p)^2] \\ &= p(1-p). \end{aligned} \tag{3.37}$$

A zero-mean Gaussian probability density function (pdf) model of the random amplitudes of an impulsive noise has the following expression:

$$f_N(i(m)) = \frac{1}{\sqrt{2\pi}\sigma_i} exp\ -\frac{i^2(m)}{2{\sigma_i}^2}, \tag{3.38}$$

where ${\sigma_i}^2$ represents the variance of the noise amplitude.

The probability density function of an impulsive noise $i(m)$ in a Bernoulli-Gaussian model is given by a mixture of two probabilities as

$$f_N^{BG}(i(m)) = (1-p)\ \delta(i(m)) + p\ f_N(i(m)), \tag{3.39}$$

where $\delta(i(m))$ denotes the Kronecker delta pdf that models the absence of noise.

Note that the function $f_N^{BG}(i(m))$ represents a combination of a discrete probability function $\delta(i(m))$ and a continuous probability density function $f_N(i(m))$.

3.4.4.2 Poisson-Gaussian model of impulsive noise

In a Poisson-Gaussian model, the distribution of the random amplitude of impulsive noise is modeled by a Gaussian process, whereas the probability of occurrence of number of impulsive noise events in a time interval of T seconds is modeled by a Poisson process.

The Poisson process is considered as a random event-counting process. The probability of occurrence of n impulsive noise in a time interval T is represented by

$$P(n,T) \quad = \quad \frac{(\lambda T)^n}{n!}\ exp(-\lambda T), \tag{3.40}$$

where the parameter λ designates a rate function with properties as follows:

$$Pr(\textit{one impulse in a small time interval } \Delta t) \quad = \lambda \Delta t \tag{3.41}$$

$$Pr(\textit{zero impulse in a small time interval } \Delta t) \quad = 1 - \lambda \Delta t. \tag{3.42}$$

It should be noted that λT is the mean value of the process that represents the average number of impulses that occurs in a time interval T.

We supposed that, in a time interval Δt, no more than one impulsive noise can occur.

The pdf of an impulsive noise $i(m)$ in a small time interval Δt with a Poisson-Gaussian model can be written as

$$f_N^{PG}(i(m)) \quad = \quad (1 - \lambda \Delta t)\ \delta(i(m)) + \lambda \Delta t\ f_N(i(m)), \tag{3.43}$$

where $f_N(i(m))$ is the Gaussian pdf presented by equation (3.38).

It can be also shown from equation (3.40) that the mean and variance of the number of impulses in a time T are as follows:

$$\textit{Expected number of impulses in } T \textit{ seconds} \quad = \lambda T \tag{3.44}$$

$$\textit{Variance of number of occurrence of impulses in } T \textit{ seconds} \quad = \lambda T. \tag{3.45}$$

3.5 Radio Channel Representation Functions

3.5.1 Bello functions

Because of the high complexity of the physical interactions, including diffractions, reflections, refractions and scattering, the most suitable model for the impulse response is, from a stochastic point of view, a multivariate random process given by $h(t,\tau)$. However, based on the physically significant parameters such as delays and Doppler shifts, a parametric approach of the random process may be obtained.

A general assumption is that $h(t,\tau)$ represents a wide-sense stationary processes, where statistical properties of the channel don't change with uncorrelated scatters, and with time. These assumptions are jointly called Wide-Sense-Stationary-Uncorrelated-Scattering (WSSUS), where time delays and Doppler shifts are uncorrelated.

In [61], Bello defines the relationships between functions of the channel characteristics. Indeed, he named $h(t,\tau)$ as the input delay spread function, defined as the response

of the channel to a unit impulse function input at some time t and some previous time τ seconds. The output of the channel $r(t)$ can be expressed by the convolution between the input signal $s(t)$ and $h(t,\tau)$ integrated over the variable τ as follows:

$$r(t) = s(t) * h(t,\tau) = \int_{-\infty}^{\infty} h(t,\tau)s(t-\tau)\ d\tau. \tag{3.46}$$

We can perform the Fourier transformation with respect to either or both of t and τ since the impulse response of a time-variant system $h(t,\tau)$ depends on these two variables. This results four equivalent representations which are represented in Fig. 3.7.

Time-varying impulse response

The impulse response $h(t,\tau)$ varies with time t and propagation delay τ and it represents the starting point for characterizing the time-variant channel.

Time-varying transfer function

The time-varying transfer function $H(t,f)$ is the Fourier transform of the impulse response $h(t,\tau)$ with respect to τ:

$$H(t,f) = \int_{-\infty}^{\infty} h(t,\tau)e^{-j2\pi f\tau}\ d\tau. \tag{3.47}$$

Delay Doppler-spread function

Fourier-transforming with respect to t results the delay Doppler-spread function $S(v,\tau)$ which describes the spreading of the impulse signal in the delay and Doppler domain:

$$S(v,\tau) = \int_{-\infty}^{\infty} h(t,\tau)e^{-j2\pi tv}\ dt. \tag{3.48}$$

Doppler-variant transfer function

The Doppler-variant transfer function $B(v,f)$ is the Fourier transform of $S(v,\tau)$ with respect to the variable τ:

$$B(v,f) = \int_{-\infty}^{\infty} S(v,\tau)e^{-j2\pi f\tau}\ d\tau. \tag{3.49}$$

This function represents the frequency variations induced by the channel due to the mobility.

It is possible to obtain the different parameters of the channel from the previous functions, such that maximum delay τ_{max}, the Doppler spread B_d, time coherence T_c and the bandwidth coherence B_c.

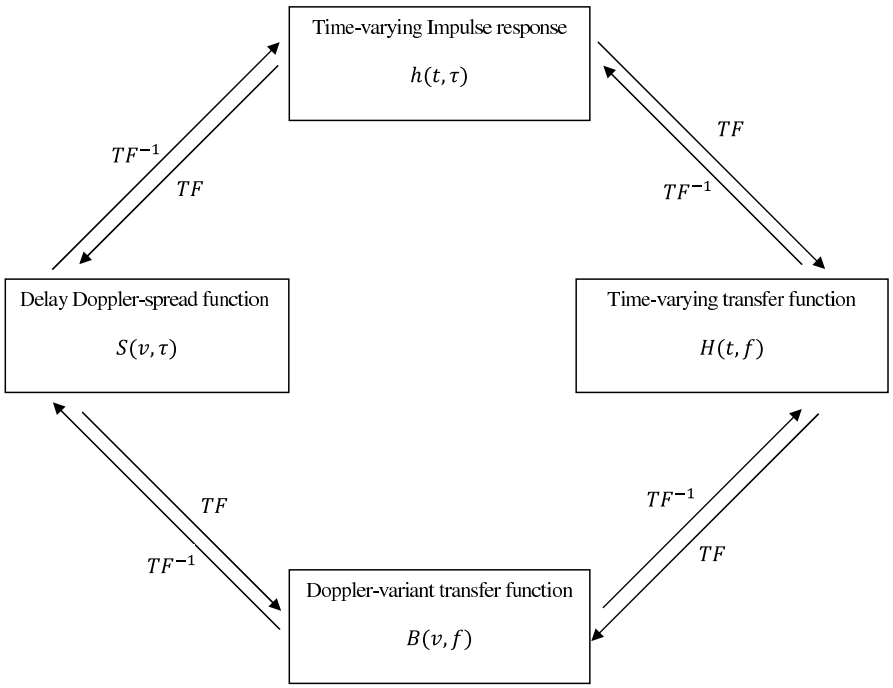


Figure 3.7: The Bello functions: Interrelations between different system functions.

3.5.2 Multipath intensity profile

The autocorrelation function can be calculated for a given channel impulse $h(t,\tau)$ as follows:

$$S(t_1,t_2;\tau) = \frac{1}{2}E\left\{h^*(t_1,\tau)h(t_2,\tau)\right\}, \tag{3.50}$$

where $h(t_1,\tau)$ and $h(t_2,\tau)$ are the channel impulse responses at instants t_1 and t_2, and $E(\cdot)$ represents the expectation operator. Assume that the channel is wide sense stationary, then equation (3.50) can be expressed as

$$S(\Delta t;\tau) = \frac{1}{2}E\left\{h^*(t,\tau)h(t+\Delta t,\tau)\right\}. \tag{3.51}$$

It can be noted that $S(\Delta t;\tau)$ gives the average power output of the channel as a function of the difference Δt in time observation and delay spread τ.

If $\Delta t = 0$, then we obtain the function $S(0;\tau) = S(\tau)$ which called the multipath intensity profile. $S(\tau)$ represents the average power output of the channel as a function of the delay spread τ.

The delay spread is defined as the signal propagation delay that exceeds the delay of the first signal arrival at the receiver. The delay spread values can be quantified by two types:

- τ_{max} (maximum excess delay): represents the time between the first and the last received path of a transmitted pulse.

- μ_τ (average delay): represents the mean of the multipath channel delay and given by

$$\mu_\tau = \frac{\int_0^{+\infty} \tau S(\tau) d\tau}{\int_0^{+\infty} S(\tau) d\tau}. \tag{3.52}$$

3.5.3 Spaced-time correlation function

Let $S(t;\tau)$ be the autocorrelation function of the channel impulse response, then the spaced-time correlation function can be expressed as [62]

$$R(\Delta t) = \int_{-\infty}^{+\infty} S(\Delta t;\tau)\ d\tau. \tag{3.53}$$

The autocorrelation function $R(\Delta t)$ gives the coherence of the channel in time domain. In addition, time coherence T_c is defined as the measure of the expected time duration over which the channel response is invariant. It should be noted that Doppler spread B_d and coherence time T_c are reciprocally related due to the Fourier relationship between spaced-time correlation function and Doppler power spectrum as follows:

$$B_d \propto \frac{1}{T_c}. \tag{3.54}$$

The coherence time and the spaced-time correlation function provide a measure of the channel variation rapidity. If the coherence time is superior to symbol duration, then the channel is said slow fading and if the coherence time is inferior to symbol duration, the channel is said fast fading. The channel fading characteristics remain almost constant during slow fading, while in fast fading the channel fading characteristics vary multiple times.

3.5.4 Spaced-frequency correlation function

The autocorrelation function $R(t_1, t_2; \Delta f)$ of the channel impulse response can be expressed as [62]

$$R(t_1, t_2; \Delta f) = \frac{1}{2} E\left\{H^*(t_1, f_1) H(t_2, f_2)\right\}, \tag{3.55}$$

with $H(t_1, f_1)$ and $H(t_2, f_2)$ are the channel frequency responses at time-frequency pairs (t_1, f_1) and (t_2, f_2) respectively, and $\Delta f = (f_2 - f_1)$ is the frequency difference. It can be assumed that the channel is wide sense stationary, so equation (3.55) can be represented in the following form:

$$R(\Delta t; \Delta f) = \frac{1}{2} E\left\{H^*(t, f_1) H(t + \Delta t, f_2)\right\}, \tag{3.56}$$

where Δt represents the time difference between channel transfer function observations, and $R(\Delta t; \Delta f)$ is called spaced-time-spaced-frequency correlation function of the channel [62]. If we take $\Delta t = 0$ in equation (3.56), then we will get the spaced-frequency

correlation function $R(0;\Delta f) = R(\Delta f)$. This function offers a correlation measure of the channel frequency response as a function of Δf. Note that the spaced-frequency function is the Fourier transform of the multipath intensity profile $S(\tau)$ and it is given by

$$R(\Delta f) = \int_{-\infty}^{+\infty} S(\tau)\ e^{-j2\pi\Delta f\tau}\ d\tau. \tag{3.57}$$

The autocorrelation function $R(\Delta f)$ gives the coherence of the channel in frequency domain. It provides a measure of the channel coherence bandwidth which represents a statistical measure of the frequency range over which all spectral components pass through the channel with equal gain and linear phases approximately.

The delay spread and the coherence bandwidth are related due to the Fourier transform relationship between $S(\tau)$ and $R(\Delta f)$ as follows:

$$B_c \propto \frac{1}{\tau_{max}}. \tag{3.58}$$

The relationship between the transmitted signal bandwidth B_s and the coherence bandwidth B_c determines the nature of the fading. The channel is flat fading or frequency non-selective fading when $B_c > B_s$, whereas it is said frequency selective when $B_c < B_s$.

3.5.5 Doppler power spectrum

The Doppler power spectrum $S(v)$ gives knowledge of spectral enlargement of the frequency impulse which passes through a channel. It can be derived using the Fourier transform of the time-spaced correlation function given in equation (3.53) as follows:

$$S(v) = \int_{-\infty}^{+\infty} R(\Delta t)\ e^{-j2\pi v\Delta t}\ d\Delta t, \tag{3.59}$$

with v represents the frequency shift relative to the carrier frequency. Note that, the Doppler spread measures the width of the Doppler power spectrum.

3.6 Simulation of Mobile Radio Channels in LTE Communication System

Basing on the standardized power delay profile given by 3GPP specifications, we simulate an LTE mobile radio channel. Indeed, 3GPP describes three discrete power delay profiles for investigating wireless communication performance at $2.15\ GHz \pm 5\ MHz$:

- Extended Pedestrian A (EPA) model
- Extended Vehicular A (EVA) model, and
- Extended Typical Urban (ETU) model.

Table 3.1: Power delay profile for LTE-3GPP channels.

EPA		**EVA**		**ETU**	
ETD [ns]	**RPT** [dB]	**ETD** [ns]	**RPT** [dB]	**ETD** [ns]	**RPT** [dB]
0	0.0	0	0.0	0	-1.0
30	-1.0	30	-1.5	50	-1.0
70	-2.0	150	-1.4	120	-1.0
90	-3.0	310	-3.6	200	0.0
110	-8.0	370	-0.6	230	0.0
190	-17.2	710	-9.1	500	0.0
410	-20.8	1090	-7.0	1600	-3.0
		1730	-12.0	2300	-5.0
		2510	-16.9	5000	-7.0

The excess tap delay (ETD) and relative power of taps (RPT) can be shown in table 3.1 [58].

In our work, we are focusing on the Extended Vehicular A (EVA) channel which provides us a possibility to investigate our contribution in highly selective mobile radio channel estimation.

3.7 Conclusion

In this chapter, we have presented the principle of OFDM transmission scheme. Then, we have presented a mathematical framework for multipath Rayleigh channel. An overview of nonlinear impulsive noise which can often appears in communication system is also presented. In addition, the main LTE-3GPP standardized power delay profiles are enumerated. Finally, we have simulated LTE channels based on 3GPP specifications according to the channel model presented in this chapter. In the next chapter, we will present the SVR channel estimation technique for SISO-LTE system.

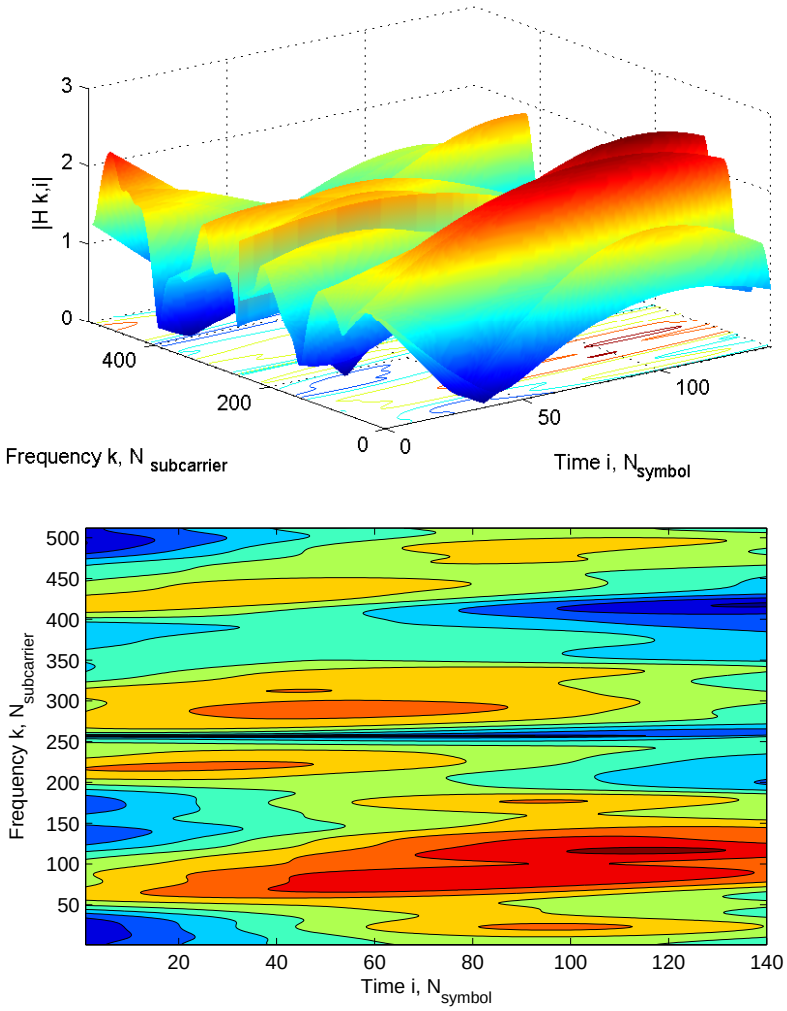


Figure 3.8: Variations in time and in frequency of the channel frequency response for mobile speed = 30 Km/h.

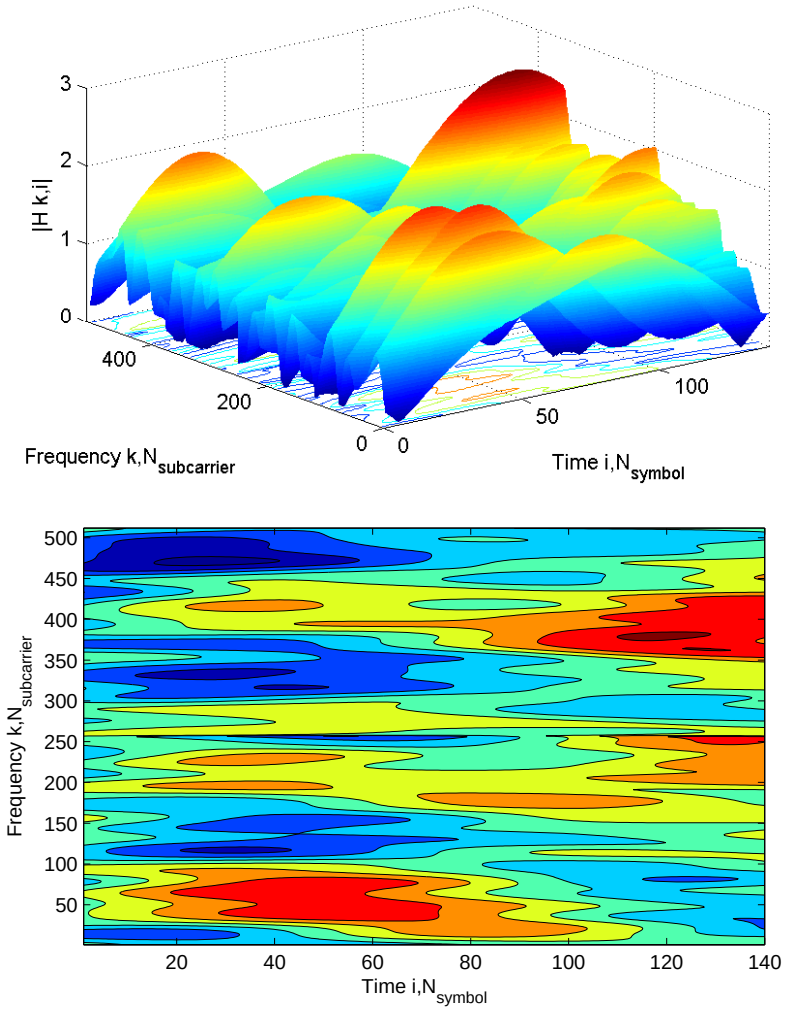


Figure 3.9: Variations in time and in frequency of the channel frequency response for mobile speed = 60 Km/h.

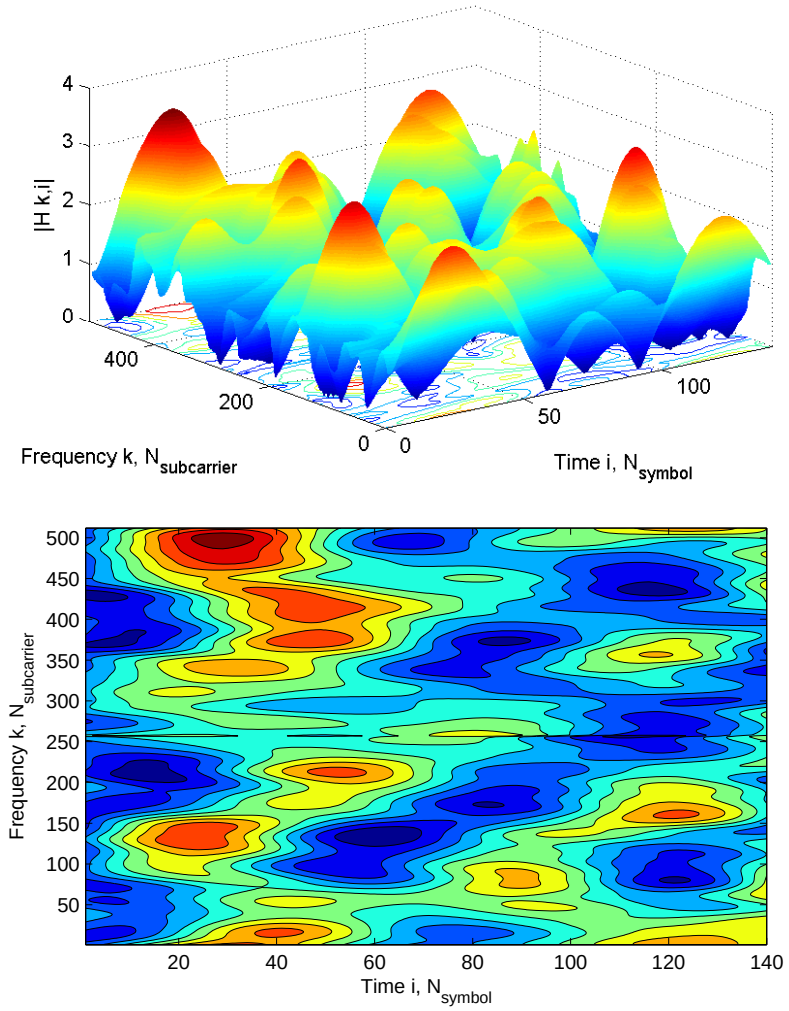


Figure 3.10: Variations in time and in frequency of the channel frequency response for mobile speed = 120 Km/h.

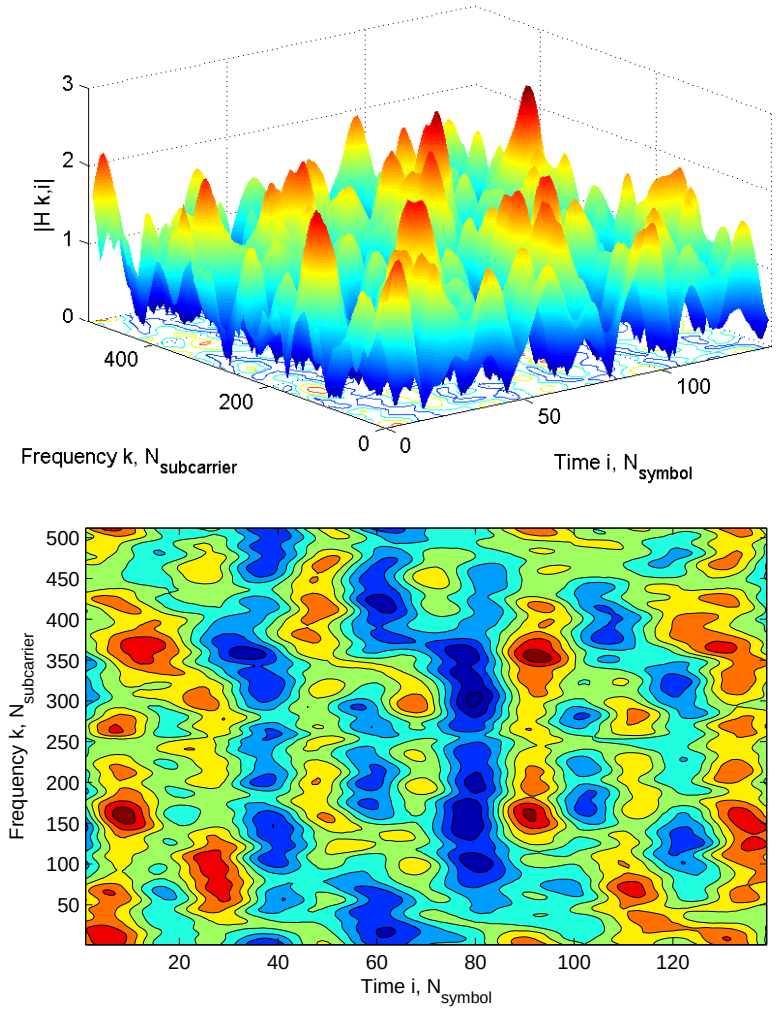


Figure 3.11: Variations in time and in frequency of the channel frequency response for mobile speed = 350 Km/h.

Chapter 4

SVR Channel Estimation in SISO-OFDM System

4.1 Introduction

Several digital communication applications using support vector machines depend on the traditional SVM classification and regression. Nevertheless, the introduction of complex algebra in the SVM algorithms can give us with a more natural and flexible framework when dealing with symbols and complex constellations.

In this chapter, a complex robust support vector regression (SVR) formulation particularly adopted to a pilot-based OFDM signal is presented. The feasibility of this technique is validated by computer simulation results obtained for Long Term Evolution (LTE) channel model. These experiments let us to analyze the performance of the SVM-OFDM system in the presence of impulsive noise interfering with OFDM reference symbols .

This chapter is organized as follows. In the second section, we present the complex SVR approach. In the third section, the OFDM system model is described. In the fourth section, some standard channel estimation techniques such as LS, MMSE and Decision Feedback are described. In the fifth section, the nonlinear complex SVR-OFDM formulation is developed. Finally, simulation results are analyzed and discussed.

4.2 Complex SVR Approach

In practice, most SVM-based solutions corresponding to complex-valued problems have been solved by reformulating complex single-dimensional vector space into real two-dimensional vector spaces. Nevertheless, an explicit complex-valued SVM algorithm can be achievable since complex-variable algebra in SVM regression gives a constrained optimization problem that is formally similar to the real-variable problem formulation.

Thus, in problems such as array processing and digital communication systems where complex numbers are often used for computation, complex-algebra representations are suitable.

In this section, a detailed version of the complex algorithm for support vector regression is presented.

4.2.1 Primal optimization functional

Consider a set of M pairs of observations $\left\{\boldsymbol{x}_m \in \mathbb{C}^K, y_m \in \mathbb{C}\right\}$ with $m = 1, \cdots, M$ that are related by a linear complex-coefficients regression model as follows:

$$y_m = \sum_{k=1}^{K} \boldsymbol{w}^k x_m^k + e_m = \mathbf{w}^H \boldsymbol{x}_m + e_m, \tag{4.1}$$

where y_m and $\boldsymbol{x}_m = \left[x_m^1, \cdots, x_m^K\right]^T$ are the m^{th} complex output and input data, respectively, $\mathbf{w} = \left[w^1, \cdots, w^K\right]^T$ represents a $K-$complex regression vector, and e_m denotes the residual noise (or complex error).

The risk of an estimator can be represented in two terms:

- empirical risk: which is measured directly on the residuals by using an appropriate cost function,
- structural risk: which includes a bound on the complexity of the resulting model.

It should be noted that minimizing the structural risk term stands for maximizing the estimator generalization capabilities when the empirical risk is reduced to zero. For linear machines, the procedure is reduced to finding the maximal margin solution by minimizing the norm of the model coefficients $\mathbf{w}$. Thus, we have to minimize in equation (4.1)

$$\frac{1}{2} \|\mathbf{w}\|^2 \tag{4.2}$$

constrained to

$$\begin{aligned} \Re(y_m - \mathbf{w}^H \boldsymbol{x}_m) &\leq \varepsilon \\ \Im(y_m - \mathbf{w}^H \boldsymbol{x}_m) &\leq \varepsilon \\ \Re(-y_m + \mathbf{w}^H \boldsymbol{x}_m) &\leq \varepsilon \\ \Im(-y_m + \mathbf{w}^H \boldsymbol{x}_m) &\leq \varepsilon, \end{aligned} \tag{4.3}$$

where ε denotes the error-tolerance parameter, and $\Re$ and $\Im$ are the real and imaginary parts of the complex number, respectively.

Note that not all of the samples will satisfy the conditions in equation (4.3), and some of them will generate errors above ε. In this case, empirical error terms have to be introduced by using a suitable cost function of the model residuals. The $\varepsilon-$ Huber cost function (4.4) is a general cost function which considers a linear and quadratic cost zones. This residual cost function is appropriate for communication environments, where additive white Gaussian noise is not the only type of noise presented in the input signals.

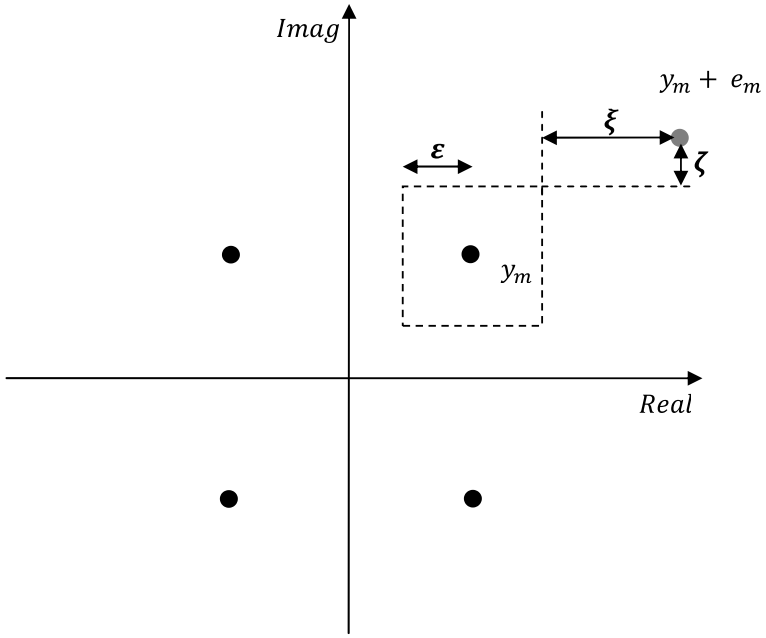


Figure 4.1: $\varepsilon-$insensitivity zone and relationship between errors and losses for the complex-valued observation y_m.

Equation (4.4) can be viewed as an approximation to the maximum likelihood cost function for sub-Gaussian noise such as impulsive or multiuser noise. At high noise level, Gaussian-noise contribution will be negligible and other noise will predominate, and then a linear cost function will perform better. At a low noise level, other noise sources will be negligible and Gaussian noise will predominate, so that the quadratic part of the function will be the optimal one. In this case, a reasonable choice for the value of $\gamma C + \varepsilon$ is to be equal to the Gaussian-noise variance.

$$\mathcal{L}^{\varepsilon}(e_m, \varepsilon, \gamma, C) = \begin{cases} 0, & |e_m| \leq \varepsilon \\ \frac{1}{2\gamma}(|e_m| - \varepsilon)^2, & \varepsilon \leq |e_m| \leq \gamma C + \varepsilon \\ C(|e_m| - \varepsilon) - \frac{1}{2}\gamma C^2, & \gamma C + \varepsilon \leq |e_m| \end{cases} . \tag{4.4}$$

Hence, when losses are considered as presented in Fig. 4.1, the problem consists of minimizing

$$\frac{1}{2}\|\mathbf{w}\|^2 + \sum_{m=1}^{M} (\mathcal{L}^{\varepsilon}(\Re(e_m), \varepsilon, \gamma, C) + \mathcal{L}^{\varepsilon}(\Im(e_m), \varepsilon, \gamma, C)). \tag{4.5}$$

Equivalently, we can state the problem as minimizing

$$\begin{aligned}\frac{1}{2}\|\mathbf{w}\|^2 &+ \frac{1}{2\gamma}\sum_{m\in I_1}(\xi_m+\xi_m^*)^2 + C\sum_{m\in I_2}(\xi_m+\xi_m^*)\\ &+ \frac{1}{2\gamma}\sum_{m\in I_3}(\zeta_m+\zeta_m^*)^2 + C\sum_{m\in I_4}(\zeta_m+\zeta_m^*)\\ &- \frac{1}{2}\sum_{m\in I_2,I_4}\gamma C^2 \end{aligned} \tag{4.6}$$

constrained to

$$\begin{aligned}\Re(y_m-\mathbf{w}^H\boldsymbol{x}_m) &\leq \varepsilon+\xi_m\\ \Im(y_m-\mathbf{w}^H\boldsymbol{x}_m) &\leq \varepsilon+\zeta_m\\ \Re(-y_m+\mathbf{w}^H\boldsymbol{x}_m) &\leq \varepsilon+\xi_m^*\\ \Im(-y_m+\mathbf{w}^H\boldsymbol{x}_m) &\leq \varepsilon+\zeta_m^*\\ \xi_m,\xi_m^*,\zeta_m,\zeta_m^* &\geq 0,\end{aligned} \tag{4.7}$$

where data have been distributed according to the sets of indices as follows:

$$\begin{aligned}I_1 &= \{m;\varepsilon\leq\xi_m,\xi_m^*\leq\varepsilon+\gamma C\}\\ I_2 &= \{m;\xi_m,\xi_m^*\geq\varepsilon+\gamma C\}\\ I_3 &= \{m;\varepsilon\leq\zeta_m,\zeta_m^*\leq\varepsilon+\gamma C\}\\ I_4 &= \{m;\zeta_m,\zeta_m^*\geq\varepsilon+\gamma C\},\end{aligned} \tag{4.8}$$

where ξ_m and ξ_m^* are slack variables which stand for positive and negative errors in the real part, respectively. ζ_m and ζ_m^* are the errors for the imaginary parts.

We notice that errors are either positive or negative, and hence at most one of the losses takes a nonzero value, so (ξ_m or ξ_m^*) either (ζ_m or ζ_m^*) is null. If we are in the $\varepsilon-$ insensitivity zone, both are null. For that reason, this constraint can be written as $\xi_m\xi_m^*=0$ ($\zeta_m\zeta_m^*=0$).

4.2.2 Primal-Dual problem

In SVM methodology, as we saw in chapter 2, it is usually possible to transform the minimization of the primal optimization functional presented in equation (4.6), subject to constraints in (4.7) into an optimization of a dual problem.

Firstly, we introduce the constraints into the primal representation by means of Lagrange multipliers.

We obtain the primal-dual functional as follows:

$$\begin{aligned} L_{Pd} &= \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2\gamma}\sum_{m\in I_1}(\xi_m+\xi_m^*)^2 + C\sum_{m\in I_2}(\xi_m+\xi_m^*) \\ &+ \frac{1}{2\gamma}\sum_{m\in I_3}(\zeta_m+\zeta_m^*)^2 + C\sum_{m\in I_4}(\zeta_m+\zeta_m^*) - \frac{1}{2}\sum_{m\in I_2,I_4}\gamma C^2 \\ &- \sum_{m=0}^{M}(\beta_m\xi_m+\beta_m^*\xi_m^*) - \sum_{m=0}^{M}(\lambda_m\zeta_m+\lambda_m^*\zeta_m^*) \\ &+ \sum_{m=0}^{M}\alpha_{R,m}[\Re(y_m-\mathbf{w}^H\boldsymbol{x}_m)-\varepsilon-\xi_m] \\ &+ \sum_{m=0}^{M}\alpha_{I,m}[\Im(y_m-\mathbf{w}^H\boldsymbol{x}_m)-j\varepsilon-j\zeta_m] \\ &+ \sum_{m=0}^{M}\alpha_{R,m}^*[\Re(-y_m+\mathbf{w}^H\boldsymbol{x}_m)-\varepsilon-\xi_m^*] \\ &+ \sum_{m=0}^{M}\alpha_{I,m}^*[\Im(-y_m+\mathbf{w}^H\boldsymbol{x}_m)-j\varepsilon-j\zeta_m^*], \end{aligned} \tag{4.9}$$

with the Lagrange multipliers constrained to $\alpha_{R,m}, \alpha_{I,m}, \beta_m, \lambda_m, \alpha_{R,m}^*, \alpha_{I,m}^*, \beta_m^*, \lambda_m^* \geq 0$ and $\xi_m, \zeta_m, \xi_m^*, \zeta_m^* \geq 0$. In addition, the Karush-Kuhn-Tucker (KKT) complementarity conditions yield that

$$\beta_m\xi_m = 0,\ \beta_m^*\xi_m^* = 0 \text{ and } \lambda_m\zeta_m = 0,\ \lambda_m^*\zeta_m^* = 0. \tag{4.10}$$

Functional presented in equation (4.9) has to be minimized with respect to the primal variables and maximized with respect to Lagrange multipliers. By making zero the primal-dual functional gradient with respect to each ω_i, we obtain an optimal solution of the weights

$$\mathbf{w} = \sum_{m=1}^{M}\psi_m\ \boldsymbol{x}_m, \tag{4.11}$$

where $\psi_m = (\alpha_{R,m} - \alpha_{R,m}^*) + j(\alpha_{I,m} - \alpha_{I,m}^*)$ with $\alpha_{R,m}, \alpha_{R,m}^*, \alpha_{I,m}, \alpha_{I,m}^*$ are Lagrange multipliers for real and imaginary parts of the residuals.

It should be noted that the solution in (4.11) is a linear combination of the input samples with complex coefficients, and certainly samples whose coefficients are nonzero will be the support vectors.

4.2.3 Dual representation

We can write the norm of the complex coefficients as

$$\begin{aligned} \|\mathbf{w}\|^2 = \mathbf{w}^H\mathbf{w} &= \sum_{n,m=1}^{M} \psi_n^* \; \psi_m \; \boldsymbol{x}_n^* \; \boldsymbol{x}_m \\ &= \sum_{k=1}^{K} \sum_{n,m=1}^{M} \psi_n^* \; \psi_m \; \boldsymbol{x}_n^{k\,*} \boldsymbol{x}_m^k. \end{aligned} \tag{4.12}$$

The Gram matrix of the dot products in equation (4.12) can be denoted by

$$\mathbf{G}(n,m) = \boldsymbol{x}_n^H \; \boldsymbol{x}_m = \sum_{k=1}^{K} \boldsymbol{x}_n^{k\,*} \boldsymbol{x}_m^k. \tag{4.13}$$

Therefore, we can write the norm of the coefficients as

$$\|\mathbf{w}\|^2 = \boldsymbol{\psi}^H \; \mathbf{G} \; \boldsymbol{\psi}, \tag{4.14}$$

where $\boldsymbol{\psi} = [\psi_0, \cdots, \psi_{M-1}]^T$. By placing equation (4.14) into equation (4.9) and re-grouping terms, the dual functional to be maximized can be written into a compact form as follows:

$$-\frac{1}{2}\boldsymbol{\psi}^H(\mathbf{G} + \gamma\mathbf{I})\boldsymbol{\psi} + \Re(\boldsymbol{\psi}^H\boldsymbol{y}) - (\boldsymbol{\alpha}_{\mathbf{R}} + \boldsymbol{\alpha}_{\mathbf{R}}^* + \boldsymbol{\alpha}_{\mathbf{I}} + \boldsymbol{\alpha}_{\mathbf{I}}^*)\mathbf{1}\varepsilon, \tag{4.15}$$

where $\mathbf{I}$ and $\mathbf{1}$ are the identity matrix and the all-ones column vector, respectively; $\boldsymbol{\alpha}_{\mathbf{R}}$ is the vector which contains the corresponding dual variables, with other subsets being similarly represented. Expression (4.15) represents a more appropriate description for complex-valued problems as it will be shown later where the complex SVR method for comb type pilot arrangement will be proposed for SISO-OFDM system.

4.3 OFDM System Model

The considered OFDM system model consists firstly of mapping binary data streams into complex symbols by means of QAM modulation. Then data are transmitted in frames by means of serial-to-parallel conversion. Some pilot symbols are inserted into each data frame which is modulated to subcarriers through IDFT. Thus, these pilot symbols are inserted for channel estimation purposes, and the IDFT is used to transform the data sequence $X(k)$ into time domain signal as follows:

$$x(n) \;=\; IDFT_N\left\{X(k)\right\} = \sum_{k=0}^{N-1} X(k) \; e^{\,j\frac{2\pi}{N}kn}, \qquad n = 0, \cdots, N-1. \tag{4.16}$$

One guard interval is inserted between every two OFDM symbols in order to eliminate inter-symbol interference (ISI). It is well known that if the channel impulse response has a maximum of L resolvable paths, then the GI must be at least equal to L [63].

Accordingly, the OFDM system comprises N subcarriers which occupy a bandwidth B. Each OFDM symbol is transmitted in time T and includes a cyclic prefix of duration T_{cp}. Therefore, the duration of each OFDM symbol is $T_u = T - T_{cp}$. Every two adjacent subcarriers are spaced by $\delta f = 1/T_u$ [64]. The output signal of the OFDM system is converted into serial signal by parallel to serial converter. A complex white Gaussian noise process $N(0, \sigma_{w_g}^2)$ with power spectral density $N_0/2$ is added through a frequency selective time varying multipath fading channel.

In a practical environment, impulsive noise can be present, and then the channel becomes nonlinear with non Gaussian impulsive noise. The impulsive noise can significantly influence the performance of the OFDM communication system for many reasons. First, the time of the arrival of an impulse is unpredictable and shapes of the impulses are not known and they vary considerably. Moreover, impulses usually have very high amplitude, and thus high energy, which can be much greater than the energy of the useful signal [65] [66].

In our thesis, we treat two models of impulsive noise as described in the previous chapter: Bernoulli-Gaussian process and Poisson-Gaussian process. The Bernoulli-Gaussian process was generated with the Bernoulli-Gaussian process function $i(n) = v(n)\mu(n)$ where $v(n)$ is a random process with Gaussian distribution and power σ_{BG}^2, and $\mu(n)$ is a Bernoulli process, as presented in equation (3.35). Whereas, the Poisson-Gaussian process was generated with the Poisson-Gaussian process function $i(n) = u(n)\lambda(n)$ where $u(n)$ is a random process with Gaussian distribution and power σ_{PG}^2, and $\lambda(n)$ is a Poisson process, as presented in equation (3.41).

At the receiver side, and after removing guard interval, the discrete-time baseband OFDM signal for the system including impulsive noise is

$$y(n) = \sum_{k=0}^{N-1} X(k)H(k)\ e^{\,j\frac{2\pi}{N}kn} + w_g(n) + i(n), \qquad n = 0, \cdots, N-1. \quad (4.17)$$

where $y(n)$ are time domain samples and $H(k) = DFT_N\{h(n)\}$ is the channel's frequency response at the k^{th} frequency. The sum of both terms of the AWGN noise and impulsive noise constitutes the total noise given by $z(n) = w_g(n) + i(n)$.

Let Ω_P be the subset of N_P pilot subcarriers and ΔP be the pilot interval in frequency domain. Over this subset, channel's frequency response can be estimated, and then interpolated over other subcarriers $(N - N_P)$. These remaining subchannels are interpolated by the nonlinear complex SVR algorithm.

The OFDM system can be expressed as [67] [68]

$$\begin{aligned} y(n) &= y^P(n) + y^D(n) + z(n) \\ &= \sum_{k\in\Omega_P} X^P(k)H(k)\ e^{\,j\frac{2\pi}{N}kn} + \sum_{k\notin\Omega_P} X^D(k)H(k)\ e^{\,j\frac{2\pi}{N}kn} + z(n), \end{aligned} \quad (4.18)$$

where $X^P(k)$ and $X^D(k)$ are complex pilot and data symbol respectively, transmitted at the k^{th} subcarrier.

After DFT transformation, $y(n)$ becomes

$$Y(k) = DFT_N\{y(n)\} = \frac{1}{N}\sum_{n=0}^{N-1} y(n)\ e^{-j\frac{2\pi}{N}kn}, \qquad k = 0, \cdots, N-1. \tag{4.19}$$

Assuming that ISI are eliminated after DFT transformation, therefore $y(n)$ becomes

$$\begin{aligned} Y(k) &= X(k)H(k) + W_G(k) + I(k) \\ &= X(k)H(k) + Z(k), \qquad k = 0, \cdots, N-1 \end{aligned} \tag{4.20}$$

where $Z(k)$ is the residual noise which represents the sum of the AWGN noise $W_G(k)$ and impulsive noise $I(k)$ in the frequency domain, respectively.
Equation (4.20) may be represented in matrix notation

$$Y = \mathbf{X}\mathbf{F}h + W_G + I = \mathbf{X}H + Z, \tag{4.21}$$

where

$$\begin{aligned} \mathbf{X} &= diag(X(0), X(1), \cdots, X(N-1)) \\ Y &= [Y(0), \cdots, Y(N-1)]^T \\ W_G &= [W_G(0), \cdots, W_G(N-1)]^T \\ I &= [I(0), \cdots, I(N-1)]^T \\ H &= [H(0), \cdots, H(N-1)]^T \\ Z &= [Z(0), \cdots, Z(N-1)]^T \\ \mathbf{F} &= \begin{bmatrix} {F_N}^{00} & \cdots & {F_N}^{0(N-1)} \\ {F_N}^{10} & \cdots & {F_N}^{1(N-1)} \\ \vdots & \ddots & \vdots \\ {F_N}^{(N-1)0} & \cdots & {F_N}^{(N-1)(N-1)} \end{bmatrix} \\ \text{and} \quad F_N^{i,k} &= (\frac{1}{\sqrt{N}})\exp(-j2\pi(\frac{ik}{N})). \end{aligned} \tag{4.22}$$

4.4 Standard Channel Estimation Techniques

4.4.1 LS channel estimation

The principal of the channel least squares estimator is minimizing the square distance between the received signal Y and the original signal X as follows:

$$\begin{aligned} \min_{H^H} J(H) &= \min_{H^H}\left\{|Y - \mathbf{X}H|^2\right\} \\ &= \min_{H^H}\left\{(Y - \mathbf{X}H)^H(Y - \mathbf{X}H)\right\}, \end{aligned} \tag{4.23}$$

where, $(\cdot)^H$ is the conjugate transpose operator. By differentiating expression (4.23) with respect to H^H and finding the minima, we obtain the LS channel estimation given by

$$\hat{H}_{LS} = \mathbf{X}^{-1}Y. \tag{4.24}$$

LS channel estimation method for OFDM system has low complexity but it suffers generally from a high mean square error.

4.4.2 MMSE channel estimation

The MMSE estimator uses the second-order statistics of the channel in order to minimize the mean square error. Let $\mathbf{R}_{hh}$, $\mathbf{R}_{HH}$ and $\mathbf{R}_{YY}$ be the autocorrelation matrix of h, H and Y respectively, and $\mathbf{R}_{hY}$ be the cross-correlation matrix between h and Y. Let σ_z^2 denotes the noise variance of $\left\{|\bar{Z}|^2\right\}$, and assume that h and Z are uncorrelated. Then, we have the following relations:

$$\mathbf{R}_{HH} = E\left\{H\ H^H\right\} = E\left\{(\mathbf{F}\ h)(\mathbf{F}\ h)^H\right\} = \mathbf{F}\ \mathbf{R}_{hh}\ \mathbf{F}^H, \tag{4.25}$$

$$\mathbf{R}_{hY} = E\left\{h\ Y^H\right\} = E\left\{h\ (\mathbf{X}\ \mathbf{F}\ h + Z)^H\right\} = \mathbf{R}_{hh}\ \mathbf{F}^H\ \mathbf{X}^H, \tag{4.26}$$

$$\mathbf{R}_{YY} = E\left\{Y\ Y^H\right\} = \mathbf{X}\ \mathbf{F}\ \mathbf{R}_{hh}\ \mathbf{F}^H\ \mathbf{X}^H + \sigma_z^2\ \mathbf{I}_N. \tag{4.27}$$

Assuming that $\mathbf{R}_{hh}$ and σ_z^2 are known at the receiver, the MMSE estimator of h can be given by [69]

$$\hat{h}_{MMSE} = \mathbf{R}_{hY}\ {\mathbf{R}_{YY}}^{-1}\ Y. \tag{4.28}$$

So, the $\hat{H}_{MMSE}$ can be obtained as [69]

$$\begin{aligned} \hat{H}_{MMSE} &= \mathbf{F}\ \hat{h}_{MMSE} = \mathbf{F}\ [(\mathbf{F}^H\ \mathbf{X}^H)^{-1}\ {\mathbf{R}_{hh}}^{-1}\ \sigma_z^2 + \mathbf{X}\ \mathbf{F}]^{-1}\ Y \\ &= \mathbf{F}\ \mathbf{R}_{hh}\ [(\mathbf{F}^H\ \mathbf{X}^H\ \mathbf{X}\ \mathbf{F})^{-1}\ \sigma_z^2 + \mathbf{R}_{hh}]\ \mathbf{F}^{-1}\ \hat{H}_{LS} \\ &= \mathbf{R}_{HH}\ [\mathbf{R}_{HH} + \sigma_z^2\ (\mathbf{X}\ \mathbf{X}^H)^{-1}]^{-1}\ \hat{H}_{LS}. \end{aligned} \tag{4.29}$$

4.4.3 Estimation with decision feedback

OFDM channel estimation with decision feedback uses the reference symbols to estimate the channel response using LS algorithm. For each coming symbol i and for each subcarrier k for $k = 0, \cdots, N-1$, the estimated transmitted symbol is found from the previous $H(i,k)$ according to the formula

$$\hat{X}(i+1,k) = Y(i+1,k)/\widehat{H}(i,k). \tag{4.30}$$

The estimated received symbols $\hat{X}(i+1,k)$ are used to make the decision about the real transmitted symbol values $\tilde{X}(i+1,k)$. The estimated channel response is updated by

$$\hat{H}(i+1,k) = Y(i+1,k)/\tilde{X}(i+1,k). \tag{4.31}$$

Therefore, $\hat{H}(i+1,k)$ is used as a reference in the next symbol for the channel equalization.

4.5 Nonlinear Complex SVR-OFDM Approach

4.5.1 Principle

Let the OFDM frame contains Ns OFDM symbols which every symbol includes N subcarriers. The transmitting pilot symbols are $\mathbf{X}^P = diag(X(i, m\Delta P))$, $m = 0, 1, \cdots N_P - 1$, where i and m are labels in time domain and frequency domain respectively, and ΔP is the pilot interval in frequency domain. Note that, pilot insertion in the subcarriers of every OFDM symbol must satisfy the demand of the sampling theory and uniform distribution [70].

The proposed channel estimation method is based on nonlinear complex SVR algorithm which has two separate phases: training phase and estimation phase. In training phase, we estimate first the subchannels pilot symbols according to LS criterion to strike $min\left[(Y^P - \mathbf{X}^P\mathbf{F}h)(Y^P - \mathbf{X}^P\mathbf{F}h)^H\right]$ [71], as

$$\widehat{H}^P = \mathbf{X}^{P^{-1}} Y^P, \tag{4.32}$$

where $Y^P = Y(i, m\Delta P)$ and $\widehat{H}^P = \widehat{H}(i, m\Delta P)$ are the received pilot symbols and the estimated frequency responses for the i^{th} OFDM symbol at pilot positions $m\Delta P$, respectively.

Then, in the estimation phase and by the interpolation mechanism, frequency responses of data subchannels can be determined. Therefore, frequency responses of all the OFDM subcarriers are [72]

$$\widehat{H}(i, q) = f(\widehat{H}^P(i, m\Delta P)), \tag{4.33}$$

where $q = 0, \cdots, N - 1$, and $f(\cdot)$ is the interpolating function, which is determined by the nonlinear SVR approach.

4.5.2 SVR estimator formulation

In high mobility environments, where the fading channels present very complicated nonlinearities especially in deep fading case, linear approaches cannot achieve high estimation precision. Therefore, we adapt a nonlinear complex SVR technique since SVM is superior in solving nonlinear, small samples and high dimensional pattern recognition [70]. Therefore, we map the input vectors to a higher dimensional feature space $\mathcal{H}$ (possibly infinity) by means of nonlinear transformation $\boldsymbol{\varphi}$. Thus, the regularization term is referred to the regression vector in the RKHS. The following linear regression function is then

$$\widehat{H}(m\Delta P) = \mathbf{w}^H \boldsymbol{\varphi}(m\Delta P) + b + e_m,\ m = 0, \cdots, N_P - 1 \tag{4.34}$$

where $\mathbf{w}$ is the weight vector, b is the bias term well known in the SVM literature and residuals $\{e_m\}$ account for the effect of both approximation errors and noise.

In the SVM framework, the optimality criterion is a regularized and constrained version of the regularized least squares criterion. In general, SVM algorithms minimize

a regularized cost function of the residuals, usually the Vapnik's ε-insensitivity cost function [73].

A robust cost function is introduced to improve the performance of the estimation algorithm which is ε-Huber robust cost function, presented by (4.4) as

$$\mathcal{L}^{\varepsilon}(e_m) = \begin{cases} 0, & |e_m| \leq \varepsilon \\ \frac{1}{2\gamma}(|e_m| - \varepsilon)^2, & \varepsilon \leq |e_m| \leq e_C \\ C(|e_m| - \varepsilon) - \frac{1}{2}\gamma C^2, & e_C \leq |e_m| \end{cases} \tag{4.35}$$

where $e_C = \varepsilon + \gamma C$, ε is the insensitive parameter which is positive scalar that represents the insensitivity to a low noise level, parameters γ and C control essentially the trade-off between the regularization and the losses, and represent the relevance of the residuals that are in the linear or in the quadratic cost zone, respectively. The cost function is linear for errors above e_C, and quadratic for errors between ε and e_C. Note that, errors lower than ε are ignored in the ε-insensitivite zone. The quadratic cost zone uses the $L_2 - norm$ of errors, which is appropriate for Gaussian noise, and the linear cost zone limits the effect of sub-Gaussian noise [34]. Therefore, the ε-Huber robust cost function can be adapted to different types of noise.

Let $\mathcal{L}^{\varepsilon}(e_m) = \mathcal{L}^{\varepsilon}(\Re(e_m)) + \mathcal{L}^{\varepsilon}(\Im(e_m))$ since $\{e_m\}$ are complex, where $\Re(\cdot)$ and $\Im(\cdot)$ represent real and imaginary parts, respectively.
Now, we can state the primal problem as minimizing

$$\begin{aligned} \frac{1}{2}\|\mathbf{w}\|^2 &+ \frac{1}{2\gamma}\sum_{m\in I_1}(\xi_m + \xi_m^*)^2 + C\sum_{m\in I_2}(\xi_m + \xi_m^*) \\ &+ \frac{1}{2\gamma}\sum_{m\in I_3}(\zeta_m + \zeta_m^*)^2 + C\sum_{m\in I_4}(\zeta_m + \zeta_m^*) \\ &- \frac{1}{2}\sum_{m\in I_2, I_4}\gamma C^2 \end{aligned} \tag{4.36}$$

constrained to

$$\begin{aligned} \Re(\hat{H}(m\Delta P) - \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) - b) &\leq \varepsilon + \xi_m \\ \Im(\hat{H}(m\Delta P) - \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) - b) &\leq \varepsilon + \zeta_m \\ \Re(-\hat{H}(m\Delta P) + \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) + b) &\leq \varepsilon + \xi_m^* \\ \Im(-\hat{H}(m\Delta P) + \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) + b) &\leq \varepsilon + \zeta_m^* \\ \xi_m^{(*)}, \zeta_m^{(*)} &\geq 0, \end{aligned} \tag{4.37}$$

for $m = 0, \cdots, N_P - 1$, where ξ_m and ξ_m^* are slack variables which stand for positive, and negative errors in the real part, respectively. ζ_m and ζ_m^* are the errors for the imaginary parts. I_1, I_2, I_3 and I_4 are the set of samples for which:
I_1 : real part of the residuals are in the quadratic zone;
I_2 : real part of the residuals are in the linear zone;
I_3 : imaginary part of the residuals are in the quadratic zone;

I_4 : imaginary part of the residuals are in the linear zone.

To transform the minimization of the primal functional (4.36) subject to constraints in (4.37), into an optimization of the dual functional, we must first introduce the constraints into the primal functional to obtain the primal-dual functional as follows:

$$\begin{aligned}
L_{Pd} &= \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2\gamma}\sum_{m\in I_1}(\xi_m+\xi_m^*)^2 + C\sum_{m\in I_2}(\xi_m+\xi_m^*) \\
&+ \frac{1}{2\gamma}\sum_{m\in I_3}(\zeta_m+\zeta_m^*)^2 + C\sum_{m\in I_4}(\zeta_m+\zeta_m^*) - \frac{1}{2}\sum_{m\in I_2,I_4}\gamma C^2 \\
&- \sum_{m=0}^{N_{P-1}}(\beta_m\xi_m+\beta_m^*\xi_m^*) - \sum_{m=0}^{N_{P-1}}(\lambda_m\zeta_m+\lambda_m^*\zeta_m^*) \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{R,m}[\Re(\widehat{H}(m\Delta P) - \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) - b) - \varepsilon - \xi_m] \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{I,m}[\Im(\widehat{H}(m\Delta P) - \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) - b) - j\varepsilon - j\zeta_m] \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{R,m}^*[\Re(-\widehat{H}(m\Delta P) + \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) + b) - \varepsilon - \xi_m^*] \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{I,m}^*[\Im(-\widehat{H}(m\Delta P) + \mathbf{w}^H\boldsymbol{\varphi}(m\Delta P) + b) - j\varepsilon - j\zeta_m^*], \qquad (4.38)
\end{aligned}$$

with the dual variables constrained to $\xi_m, \zeta_m, \xi_m^*, \zeta_m^* \geq 0$ and $\alpha_{R,m}$, $\alpha_{I,m}$, β_m, λ_m, $\alpha_{R,m}^*$, $\alpha_{I,m}^*$, β_m^*, $\lambda_m^* \geq 0$.

According to Karush-Kuhn-Tucker (KKT) complementarity conditions [74]

$$\beta_m\xi_m = 0,\ \beta_m^*\xi_m^* = 0 \text{ and } \lambda_m\zeta_m = 0,\ \lambda_m^*\zeta_m^* = 0. \qquad (4.39)$$

Then, by making zero the primal-dual functional gradient with respect to ω_i, we obtain an optimal solution for the weights

$$\mathbf{w} = \sum_{m=0}^{N_P-1}\psi_m\boldsymbol{\varphi}(m\Delta P) = \sum_{m=0}^{N_P-1}\psi_m\boldsymbol{\varphi}(P_m), \qquad (4.40)$$

where $\psi_m = (\alpha_{R,m} - \alpha_{R,m}^*) + j(\alpha_{I,m} - \alpha_{I,m}^*)$ with $\alpha_{R,m}, \alpha_{R,m}^*, \alpha_{I,m}, \alpha_{I,m}^*$ are the Lagrange multipliers for real and imaginary parts of the residuals and $P_m = (m\Delta P), m = 0, \cdots, N_P - 1$ are the pilot positions.

We define the Gram matrix as

$$\mathbf{G}(u,v) =< \boldsymbol{\varphi}(P_u), \boldsymbol{\varphi}(P_v) >= K(P_u, P_v), \tag{4.41}$$

where $K(P_u, P_v)$ is a Mercer's kernel which represents in this thesis the RBF kernel matrix which allows obviating the explicit knowledge of the nonlinear mapping $\boldsymbol{\varphi}(\cdot)$. A compact form of the functional problem can be stated in matrix format by placing optimal solution $\mathbf{w}$ into the primal dual functional and grouping terms.

Then, the dual problem consists of maximizing [72],[75]

$$-\frac{1}{2}\boldsymbol{\psi}^H(\mathbf{G}+\gamma\mathbf{I})\boldsymbol{\psi} + \Re(\boldsymbol{\psi}^H Y^P) - (\boldsymbol{\alpha}_\mathbf{R} + \boldsymbol{\alpha}_\mathbf{R}^* + \boldsymbol{\alpha}_\mathbf{I} + \boldsymbol{\alpha}_\mathbf{I}^*)\mathbf{1}\varepsilon, \tag{4.42}$$

constrained to $0 \leq \alpha_{R,m}, \alpha_{R,m}^*, \alpha_{I,m}, \alpha_{I,m}^* \leq C$, where $\boldsymbol{\psi} = [\psi_0, \cdots, \psi_{N_P-1}]^T$; $\mathbf{I}$ and $\mathbf{1}$ are the identity matrix and the all-ones column vector, respectively; $\boldsymbol{\alpha}_\mathbf{R}$ is the vector which contains the corresponding dual variables, with the other subsets being similarly represented. The weight vector can be obtained by optimizing (4.42) with respect to $\alpha_{R,m}, \alpha_{R,m}^*, \alpha_{I,m}, \alpha_{I,m}^*$ and then substituting into (4.40).

Therefore, and after training phase, frequency responses at all subcarriers in each OFDM symbol can be obtained by SVM interpolation

$$\hat{H}(k) = \sum_{m=0}^{N_p-1} \psi_m K(P_m, k) + b, \tag{4.43}$$

for $k = 1, \cdots, N$. Note that, the obtained subset of Lagrange multipliers which are nonzero will provide with a sparse solution. As usual in the SVM framework, the free parameter of the kernel and the free parameters of the cost function have to be fixed by some a priori knowledge of the problem, or by using some validation set of observations [73].

4.6 Determination of Hyperparameters

In this section, we consider some methods for determining suitable combinations of the hyperparameters for SVMs. Here, we will only consider a few aspects of how to choose such hyperparameters for regression problems.

The quality of the estimator for the unknown risk and the precision of prediction for the unknown values depend not only on the data set used for learning purposes, the loss function and the kernel but also on the choice of the hyperparameters such as kernel parameter, regularizing parameter and loss function. Choosing these hyperparameters in an optimal way usually requires computing the predictions for many combinations of the hyperparameters. In other words, it is necessary to solve not just one convex problem but a series of them, which increases the computational effort for the use of SVMs in practice [76].

The determination of a suitable combination of hyperparameters is an optimization problem. Thus, we will present some different numerical methods to solve this optimization problem.

A practical choice of the hyperparameters depends on the criteria used to measure their quality. One helpful criteria is the *accuracy*. In classification problems, the accuracy is often measured by the empirical misclassification rate. Whereas, in regression problems, the empirical risk based on a suitable calibrated loss function is often used as accuracy criteria.

In addition, as the target functions on the hyperparameters are usually unknown, the optimal parameters have to be found numerically [76]. The following six methods are often used to determine suitable hyperparameters:

- *Random search*: a random point of the parameter space is chosen, and the value of the objective function is evaluated. This operation is repeated M times and the best point is taken as the result.

- *Grid search*: after the search space is specified, each search dimension is split into P parts. The intersections of the splits which form a multi-dimensional grid are the trial points for which the objective function is evaluated. Then, the best point is taken as the result.

- *Nelder-Mead search*: the algorithm constructs a simplex of $l + 1$ points for an $l-$dimensional optimization problem. The functional values are calculated for the vertices of the simplex, and the worst point is reflected through the opposite side of the simplex. So, if this trial point is best, the new simplex is expanded further out. If the function value is worse, then the second-worst point of the simplex is contracted.

- *Cross-validation*: The data set is randomly divided into m disjoint subsets of equal size, and each subset is used once as a validation set, whereas the other $m - 1$ subsets are put together to form a training set. The average accuracy of the m validation sets is used as an estimator for the accuracy of the method.

- *Heuristic search*: This approach was proposed by Cherkassky and Ma in 2004 [76]. Their proposal is based on both theoretical considerations and empirical results. This technique is suited for all input variables scaled to the interval $[0, 1]$. The regularization parameter C should be chosen according to the range of the values of the response variable in the training data. Cherkassky and Ma propose $\varepsilon \leftarrow 3\sigma\sqrt{(\ln n)/n}$ and $C \leftarrow \max\{|\bar{y} - 3\sigma_y|, |\bar{y} + 3\sigma_y|\}$ where $\bar{y}$ and σ_y denote the mean and the standard deviation of the responses y_i in the training data set of size n, respectively. Note that σ_y is determined by accomplishing a nearest-neighbor regression with 3 to 7 neighbors, and then the noise will be estimated using the residuals of this regression [76]. The RBF kernel parameter σ is chosen as $\sigma = \sqrt{2}c^{1/q}$, where c represents some constant between .1 and .5 and q is the number of input variables of the regression problem. This heuristic technique has

the advantage that the choice of the hyperparameters can be accessed directly from the data. The heuristic choice of (C, ε, σ) is always suitable when applied to real-life data [76].

- *Pattern search*: This method was proposed by Momma and Bennett in 2002. It is a directed search technique which examines points in the parameter space that are arranged in a pattern around the actual optimal point. Pattern search works similar to the grid search, but it only makes calculations for a subset of the grid points.

To the best of our knowledge, there is currently no practical method known that chooses the hyperparameters of SVMs in an optimal manner for all data sets and is applicable for sample sizes of any size. Typically, there is no single optimal choice of the hyperparameters but a connected region of close to optimal values [76]. Note that it is often helpful to scale all input variables to increase numerical stability.

4.7 Simulation Results

We consider the channel impulse response of the frequency-selective fading multipath channel model described by (3.21). We used the specification parameters presented in table 3.1 of an extended vehicular A model (EVA) for downlink LTE system with the excess tap delay and the relative power for each path of the channel. These parameters are defined by 3GPP standard [58].

In order to demonstrate the effectiveness of our proposed technique and evaluate the performance, two objective criteria are used: the signal-to-noise ratio (SNR) and signal-to-impulse ratio (SIR). The SNR and SIR are given by [73]

$$SNR_{dB} = 10 \log_{10} \left(\frac{E\{|y(n) - w_g(n) - i(n)|^2\}}{\sigma^2_{w_g}}\right), \tag{4.44}$$

and

$$SIR_{dB} = 10 \log_{10} \left(\frac{E\{|y(n) - w_g(n) - i(n)|^2\}}{\sigma^2_{i}}\right), \tag{4.45}$$

where σ_i^2 represents the variance of the impulsive noise which takes the values of σ^2_{BG} or σ^2_{PG} for Bernoulli Gaussian model impulsive noise or Poisson Gaussian model impulsive noise, respectively.

Then, we simulate the OFDM downlink LTE system with parameters presented in table 4.1 [77],[78],[79]. The nonlinear complex SVR estimates a number of OFDM symbols in the range of 1400 symbols, corresponding to ten radio frame LTE. Note that, the LTE radio frame duration is 10 ms [77], which is divided into 10 subframes. Each subframe is further divided into two slots, each of 0.5 ms duration.

The complex SVR algorithm parameters used in the simulation are set as: $C = 100, \gamma = 10^{-5}, \varepsilon = .001$.

Table 4.1: Parameters of simulations.

Parameters	**Specifications**
OFDM System	LTE/Downlink
Constellation	16-QAM
Mobile Speed (Km/h)	120/350
$T_s(\mu s)$	72
$f_c(GHz)$	2.15
$\delta f(KHz)$	15
$B(MHz)$	5
Size of DFT/IDFT	512
Number of paths	9

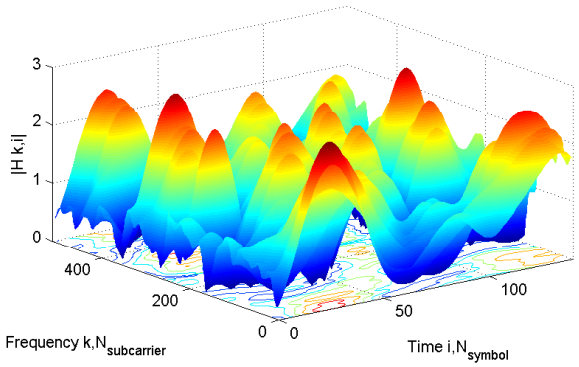


Figure 4.2: Time varying channel frequency response for mobile speed = 120 Km/h.

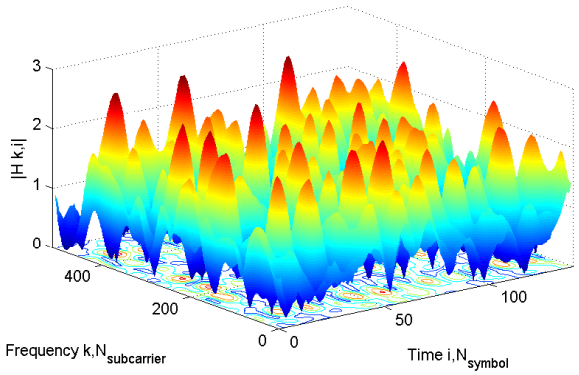


Figure 4.3: Time varying channel frequency response for mobile speed = 350 Km/h.

For the purpose of evaluation the performance of the nonlinear complex SVR algorithm under high mobility conditions, we consider a scenario for downlink LTE system for a mobile speed equal to 120 Km/h and 350 Km/h. Fig. 4.2 presents the variations in time and in frequency of the channel frequency response under a mobile speed equal to 120 Km/h, while Fig. 4.3 presents the variations in time and in frequency of the channel frequency response under a mobile speed equal to 350 Km/h. We remark from these two scenarios that the channel variations are significant in the presence of high channel selectivity and frequency shifting. Therefore, robust algorithms for channel estimation are required.

4.7.1 Simulation without impulsive noise

The performance of the proposed nonlinear complex SVR estimator is compared with other well-known estimation methods. In this part of analysis, we are interested in comparing the proposed approach with the well-defined LS [80], MMSE [69] and Decision Feedback [37].

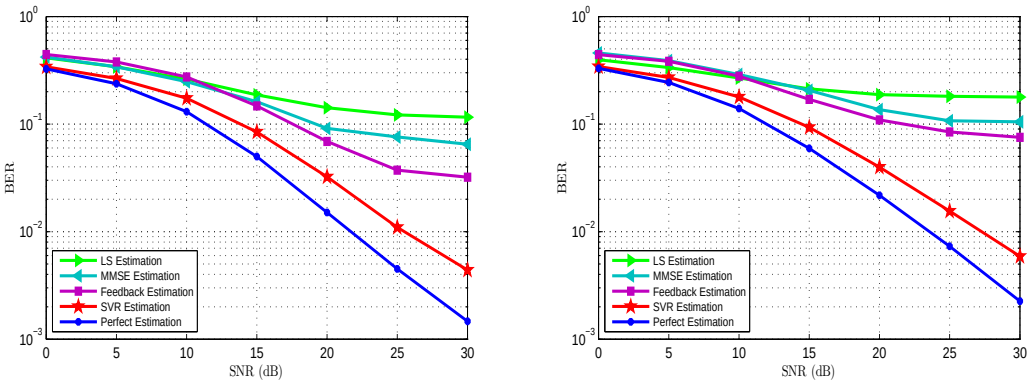
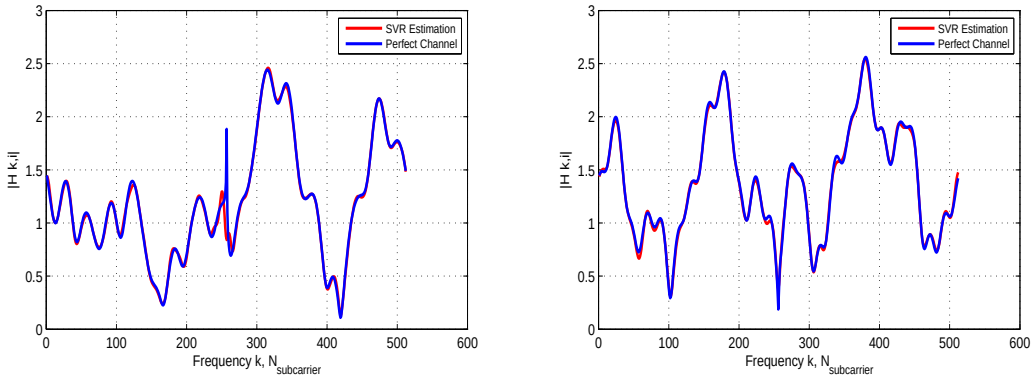


Figure 4.4: BER as a function of SNR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h without impulsive noise.

Fig. 4.4 (a) shows the variation of BER as a function of SNR. Noticeably, the proposed complex SVR method outperforms all other estimators. At high mobility, the same results are confirmed by Fig. 4.4 (b), for example at SNR = 25 dB, the complex SVR estimator achieves a BER near to 10^{-2} while the MMSE estimator achieves a BER equal to 10^{-1}.

An example of the proposed channel tracking and the nonlinear time variant channel frequency response simulated at the given multipath channel parameters is presented in Fig. 4.5. The channel response is tracked by the proposed nonlinear complex SVR method at SNR = 30 dB. Fig. 4.5 shows that the nonlinear channel response is well tracked by the proposed complex SVR method for both mobile speed 120 and 350 Km/h.



(a) (b)

Figure 4.5: An example of the proposed channel tracking and the nonlinear time variant channel frequency response simulated at SNR = 30 dB without impulsive noise for a mobile speed at (a) 120 Km/h and (b) 350 Km/h.

4.7.2 Simulation with impulsive noise

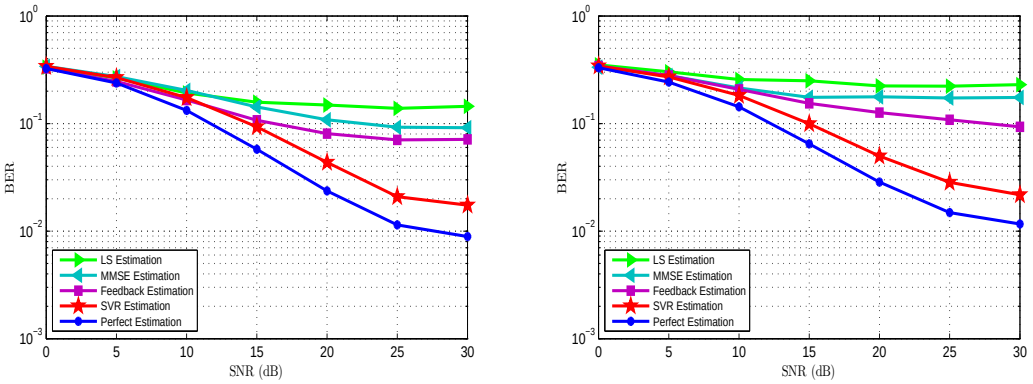
In this section, we denote by BG and PG the Bernoulli-Gaussian and Poisson-Gaussian impulsive noises, respectively.

4.7.2.1 BER vs. SNR evaluation

- BG Process

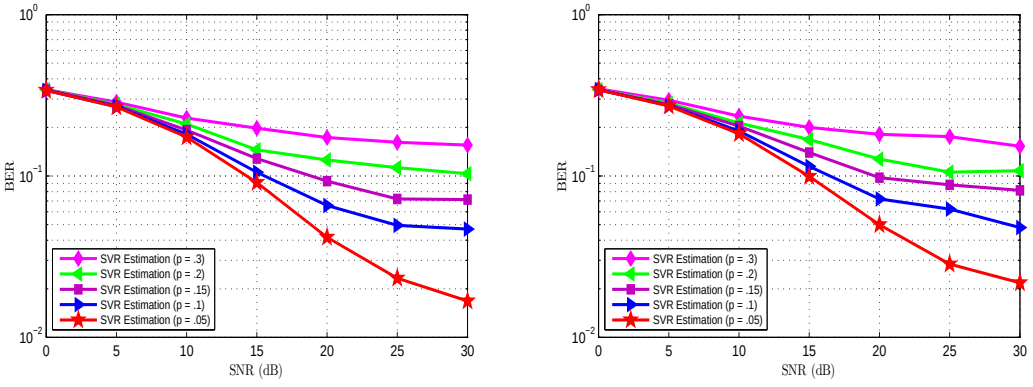
In Fig. 4.6 (a) we study the performance of the complex SVR estimator for a mobile speed at 120 Km/h in terms of BER evaluation in the presence of BG process impulsive noise with SIR = -5 dB and $p = .05$. The performance of our complex SVR estimator is compared to the aforementioned techniques for estimating the channel coefficients for different SNR values. We notice that the complex SVR approach outperforms LS, MMSE and Decision Feedback estimators especially for high SNR values.
In Fig. 4.6 (b) the complex SVR approach have almost the same behavior compared to others estimation techniques for a mobile speed at 350 Km/h.

We notice that estimators are sensitive to p variation in the case of the presence of BG impulsive noise. Fig. 4.7 shows the impact of p variation on the SVR estimation. In fact, the nonlinear complex SVR estimator achieves a significantly better estimation for small p values ($p = .05$ in our simulation case), for both higher mobile speed environments.



(a) (b)

Figure 4.6: BER as a function of SNR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h with SIR = -5 dB and $p = .05$.

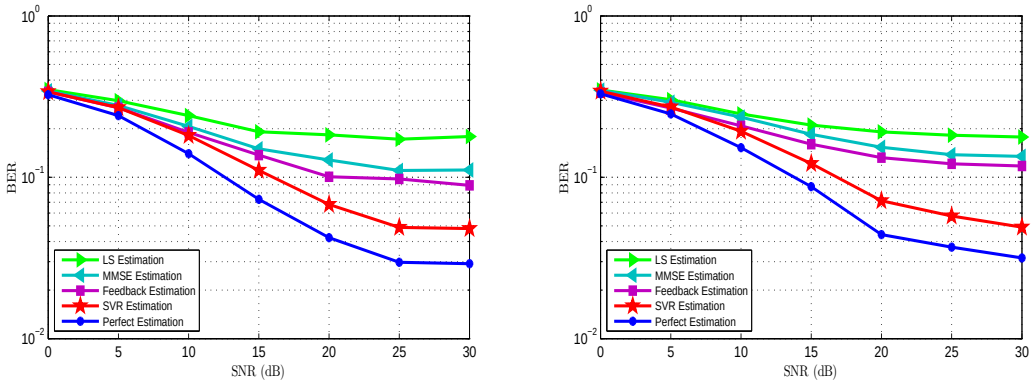


(a) (b)

Figure 4.7: BER as a function of SNR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h with SIR = - 5 dB at different p values.

- PG Process

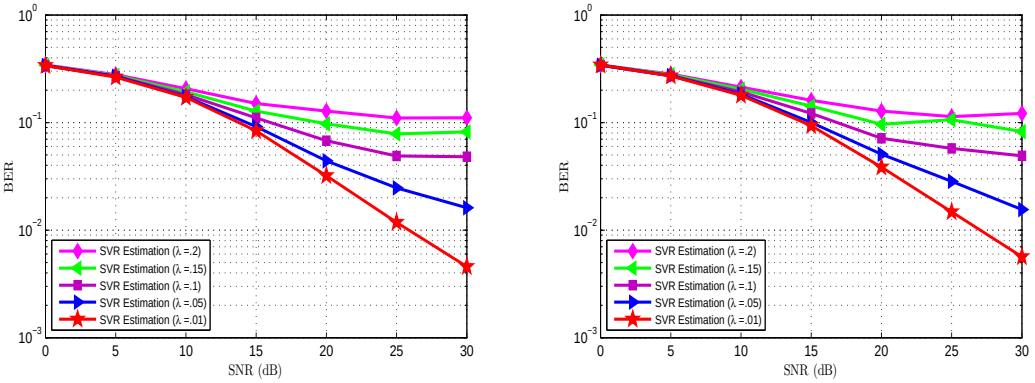
In Fig. 4.8 we compare the performance of the complex SVR estimator for a mobile speed at 120 Km/h to the LS, MMSE and Decision Feedback approaches in the presence of PG process impulsive noise with SIR = -5 dB and λ = .1. We notice here that the complex SVR estimator achieves a much better performance compared to other estimators for both high mobile speed 120 Km/h and 350 Km/h.



(a) (b)

Figure 4.8: BER as a function of SNR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h with SIR = -5 dB and λ = .1.

We notice also that estimators are sensitive to λ variation where PG impulsive noise is present. Fig. 4.9 shows the SVR estimation at different λ values.



(a) (b)

Figure 4.9: BER as a function of SNR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h with SIR = - 5 dB at different λ values.

We notice here that with PG impulsive noise model, the nonlinear complex SVR approach achieves a considerably better estimation for small values of λ (λ = .01) at

low SIR value (SIR = -5 dB) under high mobility conditions.

4.7.2.2 BER vs. SIR evaluation

- BG Process

In Fig. 4.10, we study the performance of the complex SVR estimator for a mobile speed at 120 Km/h and 350 Km/h in terms of BER evaluation for different SIR values with SNR = 20 dB and $p = .1$. The performance of our complex SVR estimator is compared to LS and Decision Feedback techniques for estimating the channel coefficients in the presence of BG impulsive noise.

We notice that our complex SVR approach has a better accuracy than LS and Decision Feedback estimators for both considered scenarios especially for low SIR values where nonlinearities improve.

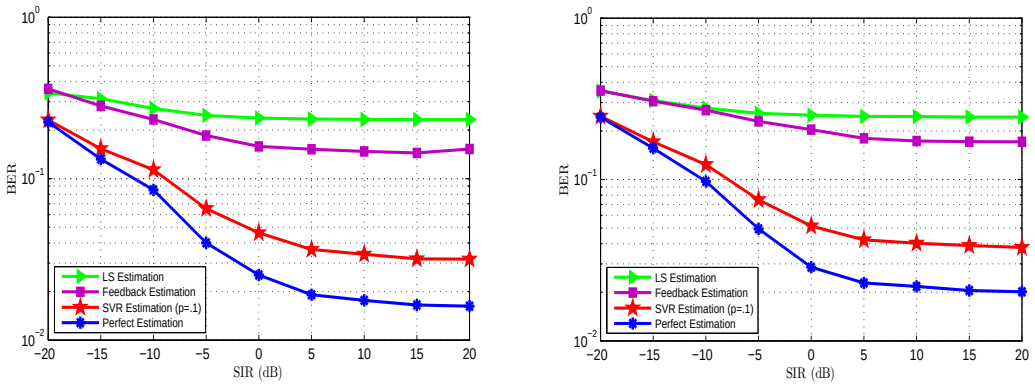


Figure 4.10: BER as a function of SIR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h for SNR = 20 dB with $p = .1$.

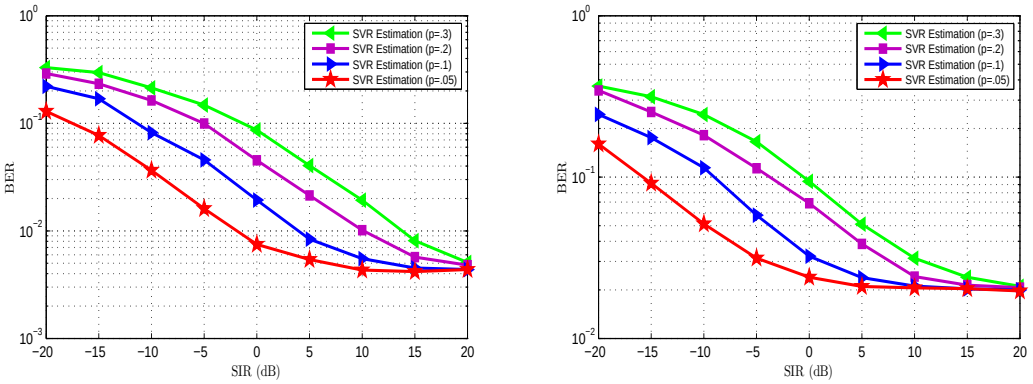
As we are mentioned above, estimators are sensitive to the variation of the parameter p of the BG process. Fig. 4.11 (a) shows the evolution of BER as a function of SIR for a mobile speed equal to 120 Km/h at different p values .

As we can see, the complex SVR method achieves a better accuracy especially for small SIR values (SIR < 0 dB) where the amount of nonlinear impulsive noise increases. Whereas for high SIR values (SIR > 0 dB) the evolution of BER still almost constant. This is due to the use of kernel-induced feature space with the exploitation of the optimization theory under nonlinear regression.

We notice that with BG impulsive noise model, the nonlinear complex SVR estimator achieves a significantly better estimation for small p values ($p = .05$ in our simulation case).

In Fig. 4.11 (b) the simulation results corroborate with the one obtained in Fig. 4.11 (a) for a mobile speed at 120 Km/h and 350 Km/h, respectively. Indeed, complex SVR

estimator still have better accuracy particularly for low SIR values where nonlinearities increase.

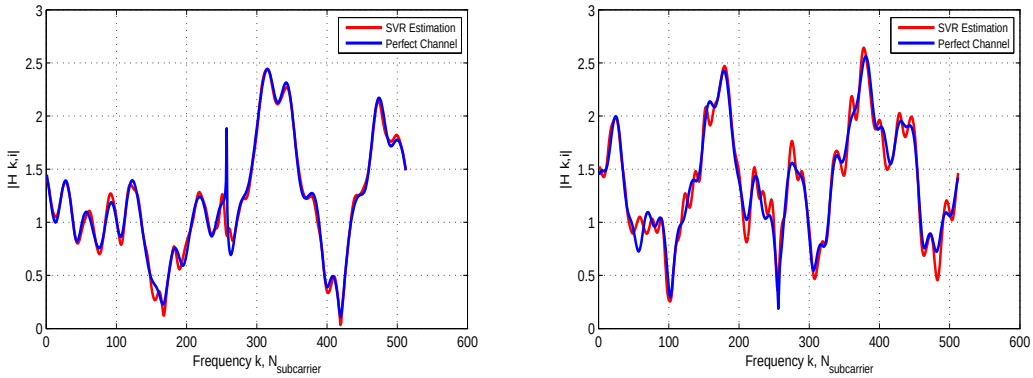


(a) (b)

Figure 4.11: BER as a function of SIR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h for SNR = 30 dB at different p values.

Fig. 4.12 shows an example of the proposed channel tracking and the nonlinear time variant channel frequency response simulated at SNR = 30 dB and SIR = -5 dB with $p = .1$ for both mobile speed at (a) 120 Km/h and (b) 350 Km/h. The channel response is tracked by the proposed nonlinear complex SVR method in the presence of BG impulsive noise.

Fig. 4.12 demonstrates that the nonlinear channel response is well tracked by the proposed complex SVR method under high mobility conditions in the presence of BG impulsive noise.



(a) (b)

Figure 4.12: An example of the proposed channel tracking and the nonlinear time variant channel frequency response simulated at SNR = 30 dB and SIR = -5dB with $p = .1$ for a mobile speed at (a) 120 Km/h and (b) 350 Km/h.

- PG Process

In Fig. 4.13, we analyze the performance of the complex SVR estimator for different SIR values with SNR = 20 dB and $\lambda = .1$. The performance of the complex SVR estimator is compared to LS and Decision Feedback techniques for estimating the channel coefficients in the presence of PG impulsive noise.

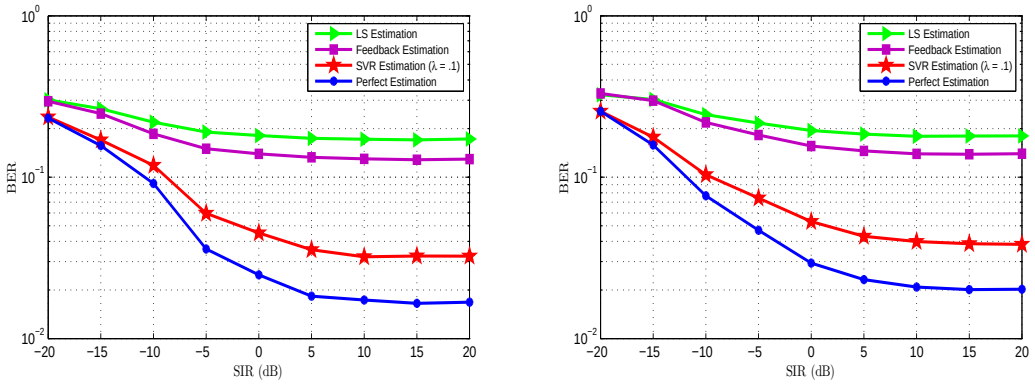


Figure 4.13: BER as a function of SIR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h for SNR = 20 dB with $\lambda = .1$.

We notice that our complex SVR approach has a better accuracy than LS and Decision Feedback estimators for both different mobile speed especially for low SIR values as in the case of BG impulsive noise. This is prove the power of the nonlinear SVR in the presence of significant nonlinearities.

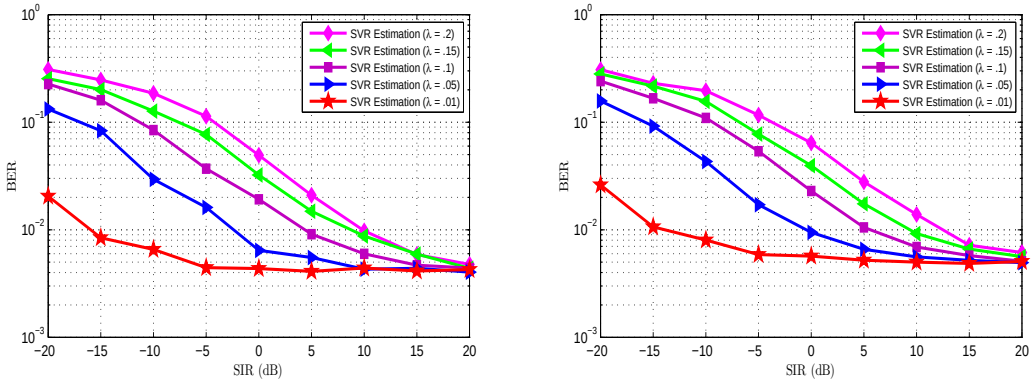
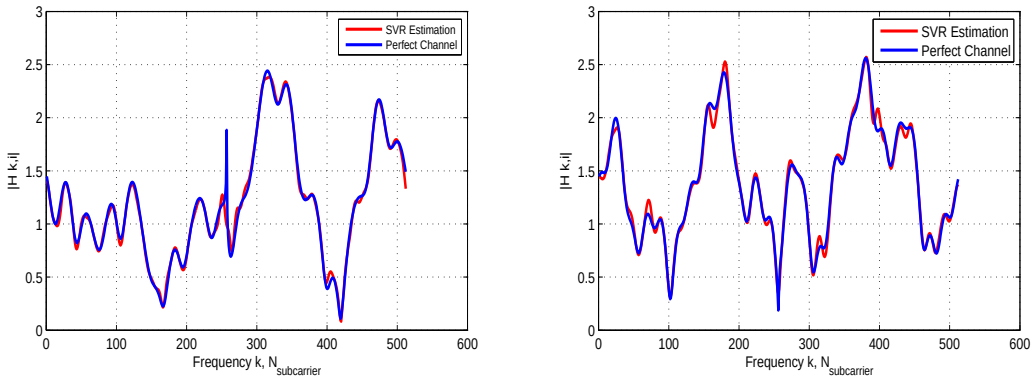


Figure 4.14: BER as a function of SIR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h for SNR = 30 dB at different λ values.

In the presence of PG impulsive noise, estimators are sensitive to the variation of the parameter λ of the PG process. Fig. 4.14 explains the evolution of BER as a function of SIR for both mobile speed 120 Km/h and 350 Km/h at different λ values for SNR = 30 dB.

We notice that the complex SVR approach achieves a slightly better estimation for different λ values for high SIR values (SIR $>$ 0 dB), whereas the accuracy of the complex SVR approach increases for smaller SIR (SIR $<$ 0 dB) . The nonlinear complex SVR estimator achieves a significantly better estimation for small λ values (λ = .01 in our simulation case) for both mobile speed cases.

Fig. 4.15 shows an example of the proposed channel tracking and the nonlinear time variant channel frequency response simulated at SNR = 30 dB and SIR = -5 dB with λ = .05 for both mobile speed at (a) 120 Km/h and (b) 350 Km/h. The channel response is tracked by the proposed nonlinear complex SVR method in the presence of PG impulsive noise. This figure confirms that the nonlinear channel response is well tracked by the proposed complex SVR method in spite of the presence of nonlinear PG impulsive noise joint with high mobility conditions.



(a) (b)

Figure 4.15: An example of the proposed channel tracking and the nonlinear time variant channel frequency response simulated at SNR = 30 dB and SIR = -5dB with λ = .05 for a mobile speed at (a) 120 Km/h and (b) 350 Km/h.

4.7.3 Simulation results with mixed impulsive noises

In the presence of both BG and PG impulsive noises, nonlinear complex SVR estimator becomes sensitive to both parameters p and λ simultaneously of the BG and PG processes, respectively.

In Fig. 4.16, we explain the evolution of BER as a function of SNR for a mobile speed at 120 Km/h in the presence of BG process impulsive noise joint with PG process impulsive noise with different SIR values (SIR = -10, -5 and 0 dB) for various values of p (p = .05 and .1) and various values of λ (λ = .01, .1, .15 and .2).

We notice that the complex SVR approach achieves a better performance for low values of p and λ especially for high SNR values.

In Fig. 4.17 the complex SVR approach has almost the same behavior compared to Fig. 4.16 for a mobile speed equal to 350 Km/h.

In Fig. 4.18, we examine the evolution of BER as a function of SIR for a mobile speed at 120 Km/h in the presence of BG process impulsive noise joint with PG process impulsive noise with different SNR values (SNR = 10, 20 and 30 dB) for various values of p (p = .05 and .1) and various values of λ (λ = .01, .1, .15 and .2).

We notice that the complex SVR approach has a better accuracy for small values of p and λ. Also, the complex SVR approach achieves a better estimation especially for smaller SIR (SIR $<$ 0 dB) where nonlinearities increase as analyzed above.

In Fig. 4.19 the simulation results corroborate with one obtained in Fig. 4.18 for a high mobile speed equal to 350 Km/h.

Our nonlinear complex SVR estimator provides an accurate estimation for the channel coefficients and gives a prediction of the unknown coefficients by means of SVR interpolation. Indeed, our complex SVR approach is efficient in the presence of impulsive noises at very high mobile speed environment for different scenarios of nonlinearities, as it has been verified and depicted in Fig. 4.16 to Fig. 4.19.

Regarding the complexity of these estimators, LS is the least complex estimator because it contains only one matrix inversion operation. However, the Decision Feedback estimator contains two operations of matrix inversion and two operations of matrix multiplication. On the other hand, our SVR estimator uses quadratic programming (*quadprog* function in Optimization MATLAB Toolbox) with the functions *Buffer* and *kron* for fast computation of kernel matrix using the Kronecker product, and thus the algorithm becomes faster.

4.8 Conclusion

In this chapter, we have presented a nonlinear complex SVR based channel estimation technique for a highly selective downlink LTE system. The proposed method is based on learning process that uses training sequence to estimate the channel variations. Our formulation is based on complex SVR specifically developed for pilot-based OFDM systems. Simulations have confirmed the capabilities of the proposed nonlinear complex SVR estimator in the presence of both Gaussian and impulsive noises interfering with the pilot symbols when compared to LS, MMSE and Decision Feedback standard methods. The proposal takes into account the temporal-spectral relationship of the OFDM signal for a highly selective channels. The Gram matrix using RBF kernel provides a natural nonlinear extension of the complex linear SVR which leads to a significant benefit for OFDM communications especially in those scenarios in which deep fading is present.

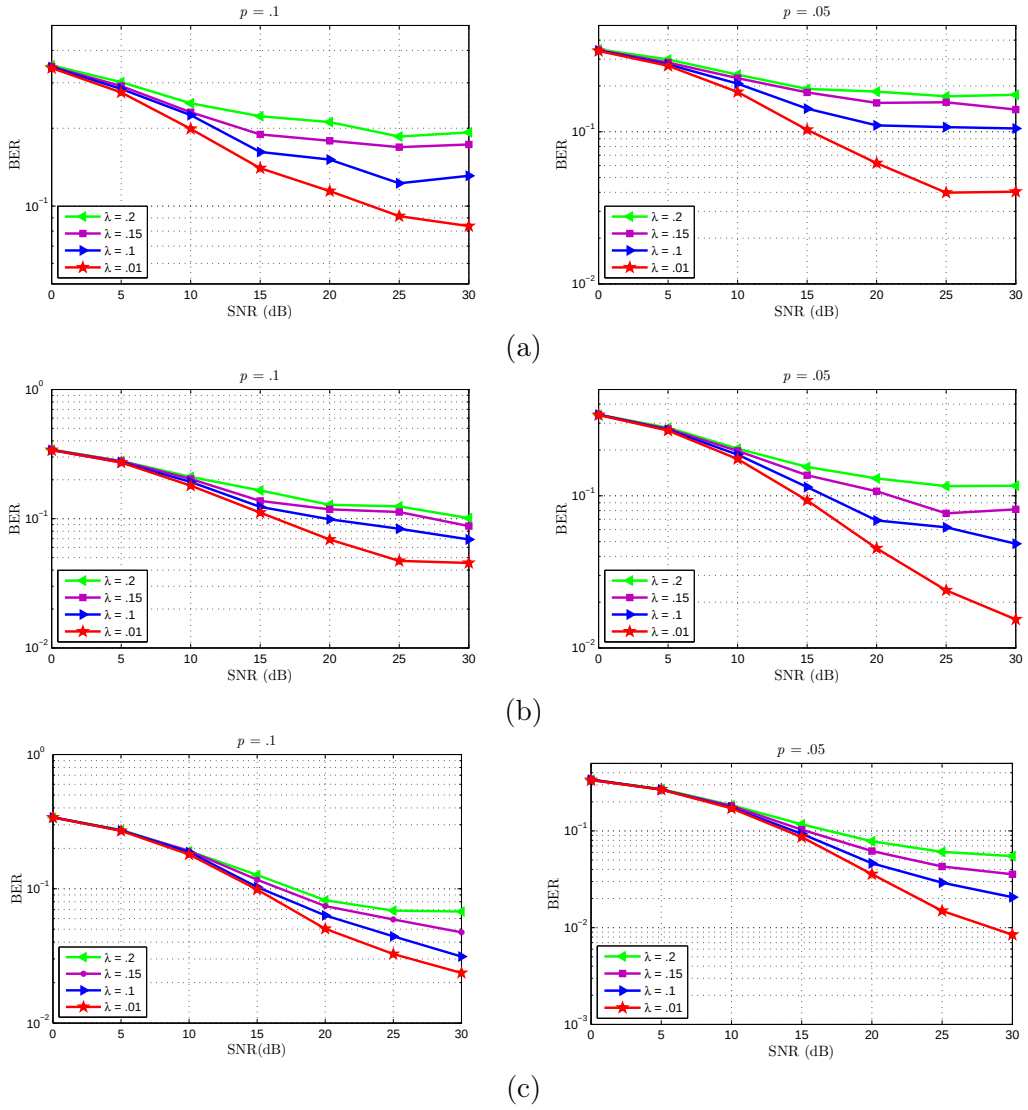
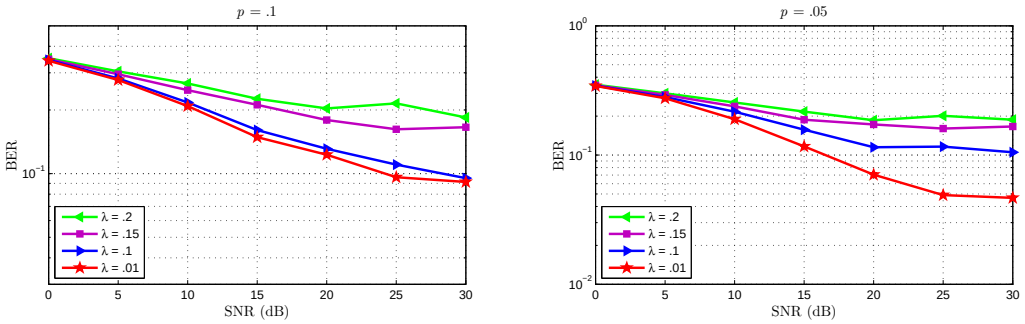
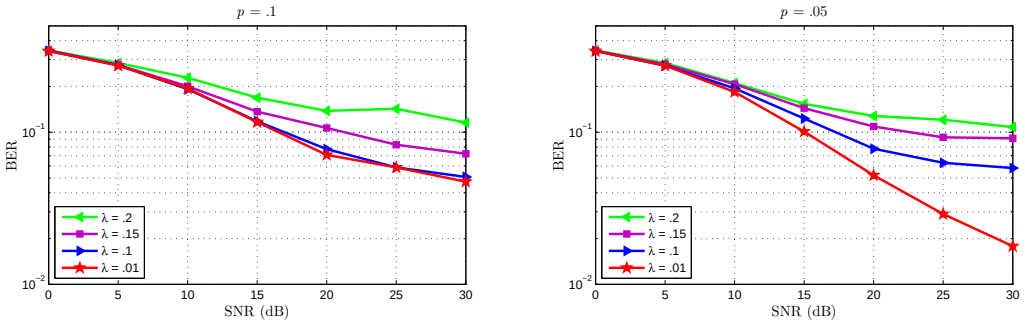


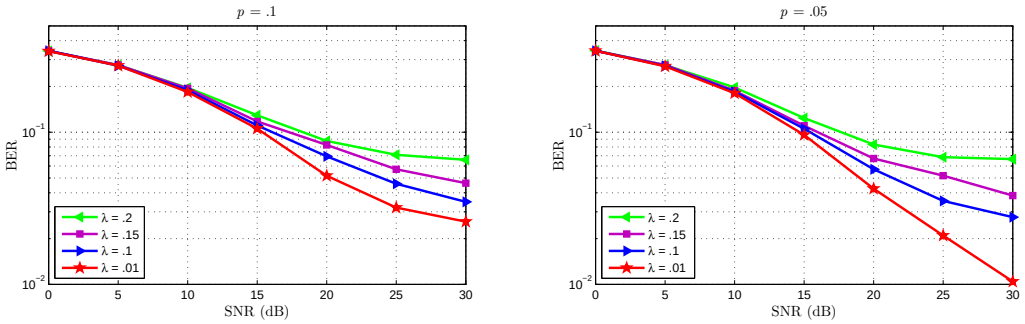
Figure 4.16: BER as a function of SNR for a mobile speed at 120 Km/h with (a) SIR = -10 dB, (b) SIR = -5 dB and (c) SIR = 0 dB with different values of p and λ.



(a)



(b)



(c)

Figure 4.17: BER as a function of SNR for a mobile speed at 350 Km/h with (a) SIR = -10 dB, (b) SIR = -5 dB and (c) SIR = 0 dB with different values of p and λ.

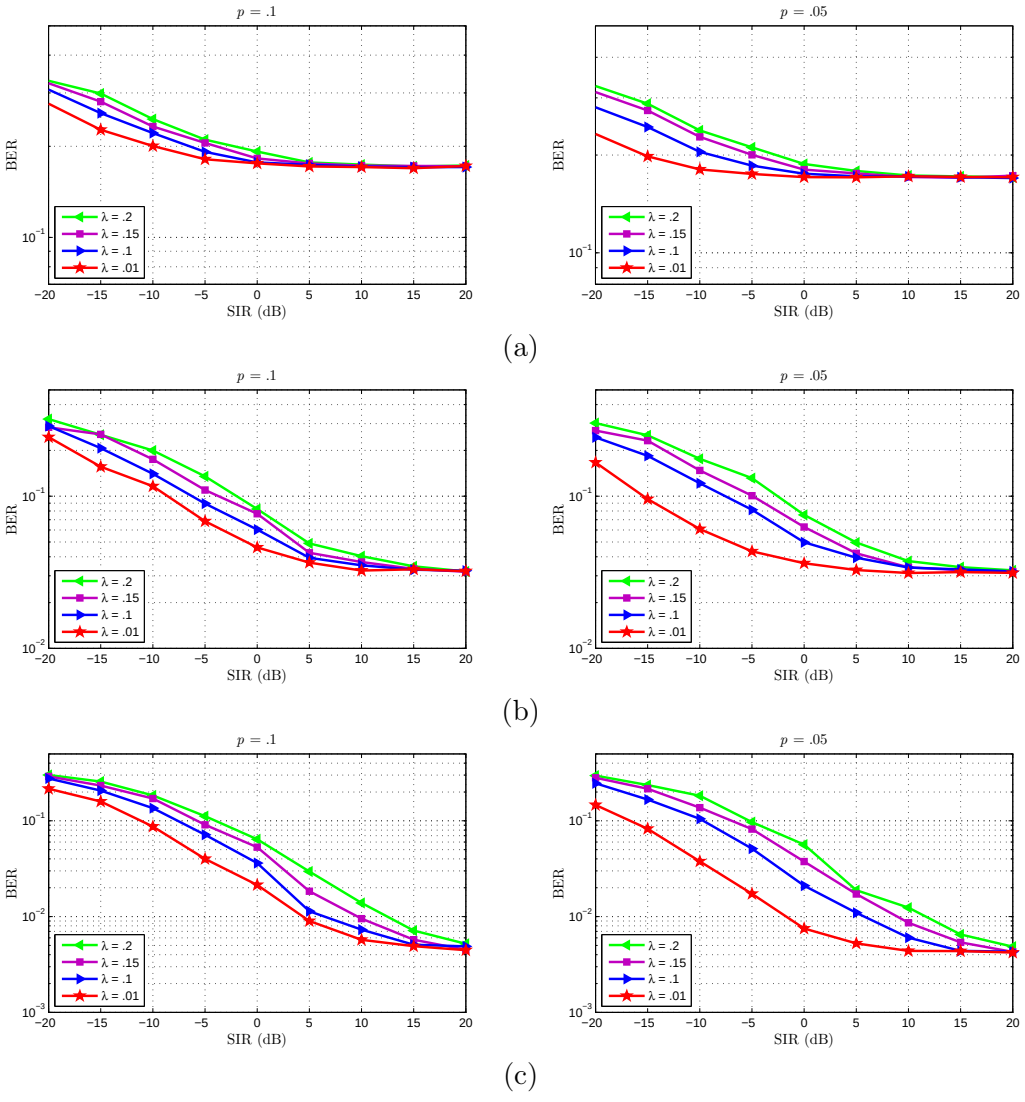


Figure 4.18: BER as a function of SIR for a mobile speed at 120 Km/h with (a) SNR = 10 dB, (b) SNR = 20 dB and (c) SNR = 30 dB with different values of p and λ.

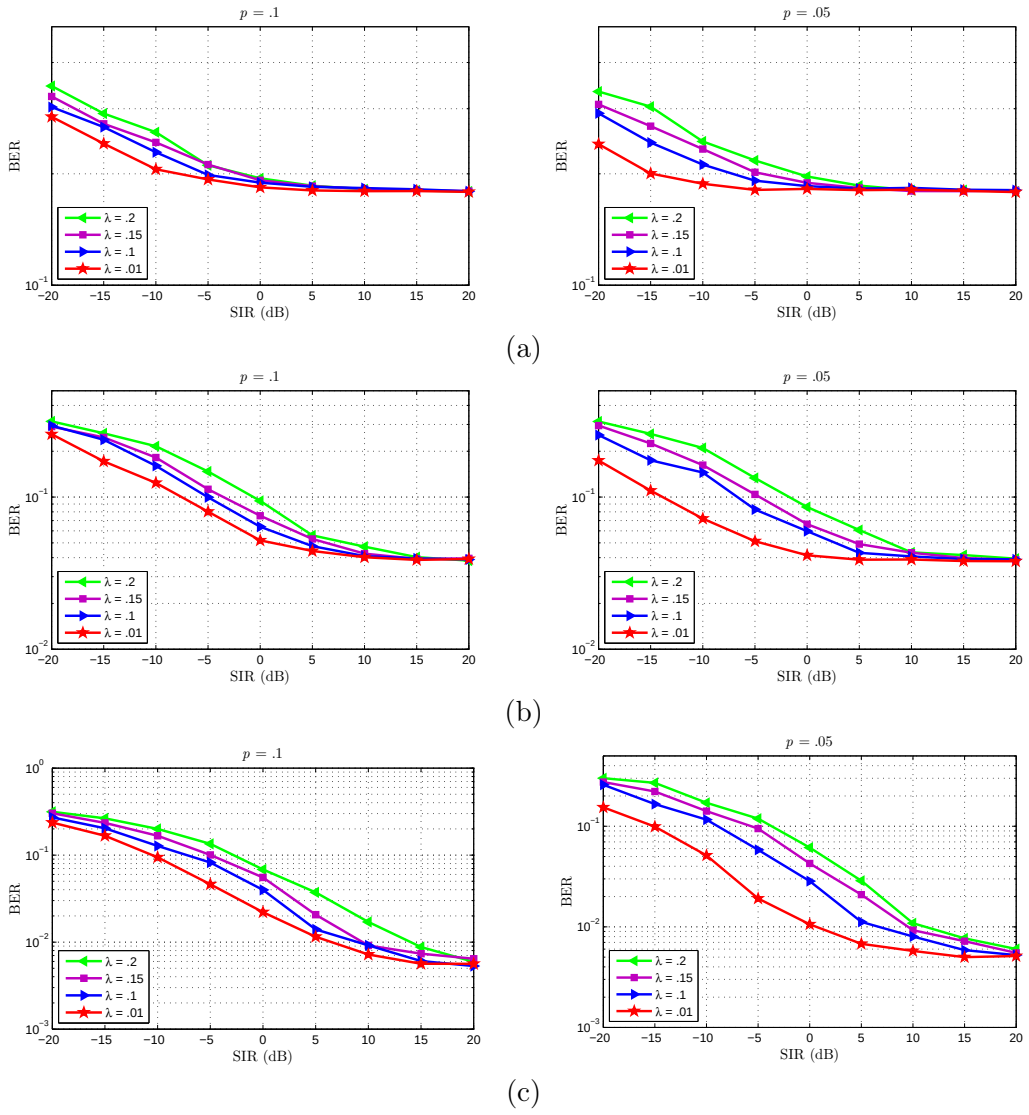


Figure 4.19: BER as a function of SIR for a mobile speed at 350 Km/h with (a) SNR = 10 dB, (b) SNR = 20 dB and (c) SNR = 30 dB with different values of p and λ.

Chapter 5

M-SVR Channel Estimation in MIMO-OFDM System

5.1 Introduction

The combination of Multiple-Input Multiple-Output (MIMO) and OFDM techniques in the same system helps to improve the link of frequency selective channels. Indeed, the transmission of multiple data streams enables multipath signals to carry more information over the same frequency channel. For that reason, MIMO-OFDM is widely adopted as the transmission scheme for almost all broadband wireless standards such as LTE, WiMAX and WLAN [81].

In fact, channel estimation for MIMO-OFDM systems have been developed by various researches intending to especially high-speed wireless transmission and diversity gain. These researches concern notably, semiblind [82, 83] and blind techniques [84], neural networks approaches [85, 86, 87, 88] and even multi-classification by support vector machines [89].

We propose in this chapter an efficient nonlinear complex Multiple Support Vector Machines Regression (M-SVR) approach to track the frequency selective time varying multipath fading channel in mobile wireless MIMO-OFDM system with the presence of Gaussian and non-Gaussian impulsive noise. The M-SVR estimation technique is developed and applied to MIMO Long Term Evolution (LTE) downlink.

This chapter is organized as follows. First, we introduce an overview of MIMO technology which is based on the diversity and spatial multiplexing. The nonlinear complex M-SVR method is presented in the third section. Finally, different scenarios are simulated and discussed in the case of SIMO and MIMO systems with both STBC and V-BLAST schemes.

5.2 An Overview of MIMO System Technology

5.2.1 Diversity

Diversity techniques are used to mitigate degradation in the error performance and combat fading presented in the unstable wireless channels. The basic principle consists of transmitting multiple replicas of the same information on multiple channels with similar power and independent fading. Thus, the probability that multiple independent fading channels simultaneously realize deep fading is very low. So, it is highly likely that at least one of the received signal replicas is not attenuated, making a good transmission quality. Two categories of diversity exist: the implied class and the explicit class. The first class consists of sending only one copy, but to rely on the multipath concept where several versions of the same signal arrives at the receiver front-end. Whereas, the second class is to transmit several replicas of the same message.
There are several ways of realizing diversity gain, including the following ones:

- Space diversity (or antenna diversity): multiple antennas are used to implement independent wireless channels. The goal is sending the same signal from various antennas separated by sufficient distance for channel decorrelation (more than 10 lambda) [90].

- Polarization diversity: independent channels are implemented using the fact that horizontally and vertically polarized paths are independent. Thus, the number of antennas required is reduced as a single dual-polarized antenna. The advantage of this technique is that one does not need much space between antennas, which is attractive to mobile equipment.

- Time diversity: the same information is repeatedly transmitted at sufficiently separated time instances (more than coherence time) . Indeed, when the transmission of the same signal is separated by the channel coherence time, it is possible to take advantage of time diversity [91].

- Frequency diversity: the same information is repeatedly transmitted at sufficiently separated frequency bands (more than coherence bandwidth). The principle of this method consists of sending the same signal at different carriers frequency which will confront different multipath fading [92]. The carriers should be separated by at least the coherence bandwidth. Note that, the required frequency spacing depends on mobile speed, channel delay spread, used frequency, etc.

- Angle diversity: this technique is largely related to the spatial diversity where multiple receive antennas with different directivity are employed to receive the same information-bearing signal at different angles. It proposes that when the beams emitted by the antennas are adequately separated angularly, it is possible that the correlation level will be low enough to benefit from diversity. This technique is largely used for domestic wireless Internet networks to increase capacity in network throughput [93].

- Space time diversity: Alamouti was the first how develops space time block coding technique in [94]. Indeed, this technique consists of sending two versions of signal delayed in time via two transmit antennas.

5.2.2 Spatial multiplexing

The major goal in developing new wireless communication systems is improving the spectrum efficiency and increasing the transmission capacity. The exploitation of the spatial dimension by using the space division multiplexing (SDM) technique is a promising solution for significant increase of bandwidth efficiency and performance under fading channels. Latest information theory researches show that the multipath wireless channel is capable of huge capacities [95, 96, 97].

Fundamentally, the SDM method transmits different data streams on different transmit antennas simultaneously, which increase the signal to noise ratio and capacity. Spatial multiplexing allows to transmit $min(N_T, N_R)$ different data streams in parallel in the case of good channel conditions [98]. The number of spatially multiplexing streams can be determined as a rank of MIMO channel matrix.

By using multiple antennas at the receiver side, these different mixed data streams can be recovered by SDM approaches like ZF [99] or V-BLAST [100].

5.2.3 MIMO with Alamouti space-time coding

The very first and well-known Space Time Block Code (STBC) is the Alamouti code which offers a simple technique for achieving spatial diversity with two transmit antenna. The Alamouti coding scheme is as follows:
Consider the transmission sequence $\{X_1, X_2, \cdots, X_n\}$. In normal transmission, we will be sending in the first time slot X_1, then X_2 in the second time slot and so on. Alamouti proposed that we group the symbols into groups of two. In the first time slot, sending X_1 and X_2 from the first and second antenna, respectively. In the second time slot, sending $-X_2^*$ and X_1^* from the first and second antenna, respectively. Note that, although we are grouping two symbols, we still need two time slots to send two symbols. Therefore, there is no change in the data rate. This represents a simple description of the transmission technique with Alamouti space time block coding.
The received signal of the first time slot can be given by

$$\begin{bmatrix} Y_1^1 \\ Y_2^1 \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} Z_1^1 \\ Z_2^1 \end{bmatrix}. \tag{5.1}$$

By assuming that the channel remains constant for the second time slot, the received signal can be obtained as

$$\begin{bmatrix} Y_1^2 \\ Y_2^2 \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix} \begin{bmatrix} -X_2^* \\ X_1^* \end{bmatrix} + \begin{bmatrix} Z_1^2 \\ Z_2^2 \end{bmatrix}, \tag{5.2}$$

where Y_1^1 and Y_2^1 are the received information at time slot 1 on the receive antenna 1 and 2, respectively. Y_1^2 and Y_2^2 represent the received information at time slot 2 on

the receive antenna 1 and 2, respectively. $H_{i,j}$ denotes the channel response from the i^{th} transmit antenna to the j^{th} receive antenna with $i, j \in \{1, 2\}$. X_1 and X_2 are the transmitted symbols. Z_1^1 and Z_2^1 represent noise at time slot 1 on the receive antennas 1 and 2, respectively, however Z_1^2 and Z_2^2 represent noise at time slot 2 on the receive antennas 1 and 2, respectively. By combining equations (5.1) and (5.2), we can derive the following expression:

$$\begin{bmatrix} Y_1^1 \\ Y_2^1 \\ Y_1^{2*} \\ Y_2^{2*} \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \\ H_{1,2}^* & -H_{1,1}^* \\ H_{2,2}^* & -H_{2,1}^* \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} Z_1^1 \\ Z_2^1 \\ Z_1^{2*} \\ Z_2^{2*} \end{bmatrix}. \tag{5.3}$$

Let us define the transmitted signal by X, the received signal by Y, the matrix of the channel response by $\mathbf{H}$ and the noise vector by Z as follows:

$$Y = \begin{bmatrix} Y_1^1 \\ Y_2^1 \\ Y_1^{2*} \\ Y_2^{2*} \end{bmatrix}, \ \mathbf{H} = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \\ H_{1,2}^* & -H_{1,1}^* \\ H_{2,2}^* & -H_{2,1}^* \end{bmatrix}, \ X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \text{ and } \quad Z = \begin{bmatrix} Z_1^1 \\ Z_2^1 \\ Z_1^{2*} \\ Z_2^{2*} \end{bmatrix}.$$

Therefore, equation (5.3) can be rewritten as

$$Y = \mathbf{H}\ X + Z. \tag{5.4}$$

To resolve the above equation for X, we need to find the inverse of $\mathbf{H}$. For that reason, we apply the following pseudo-inverse identity for a general matrix of size $(m \times n)$ [101]

$$\mathbf{H}^{\dagger} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H. \tag{5.5}$$

The term $\mathbf{H}^H \mathbf{H}$ can be expressed as

$$\mathbf{H}^H \mathbf{H} = diag_{(2\times 2)} \left[\sum_{i=1}^{2} \sum_{j=1}^{2} |H_{ij}|^2 \right], \tag{5.6}$$

where $diag_{(2\times 2)}[\cdot]$ denotes a diagonal matrix of order 2. Since the inverse of a diagonal matrix is just the inverse of the diagonal elements, we obtain

$$(\mathbf{H}^H \mathbf{H})^{-1} = diag_{(2\times 2)} \left[1/(\sum_{i=1}^{2} \sum_{j=1}^{2} |H_{ij}|^2) \right]. \tag{5.7}$$

Thus, the estimate of the transmitted symbol can be expressed as follows:

$$\begin{bmatrix} \hat{X_1} \\ X_2 \end{bmatrix} = \mathbf{H}^{\dagger} \begin{bmatrix} Y_1^1 \\ Y_2^1 \\ Y_1^{2*} \\ Y_2^{2*} \end{bmatrix} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H \begin{bmatrix} Y_1^1 \\ Y_2^1 \\ Y_1^{2*} \\ Y_2^{2*} \end{bmatrix} \tag{5.8}$$

$$\hat{X} = (\mathbf{H}^H \mathbf{H})^{-1} \mathbf{H}^H\ Y. \tag{5.9}$$

5.2.4 MIMO with ZF equalizer

Consider the transmission sequence $\{X_1, X_2, \cdots, X_n\}$. In normal transmission, we will be sending X_1 in the first time slot, X_2 in the second time slot, X_3 and so on. Now, as we have 2 transmit antennas, symbols can be grouped into groups of two. In the first time slot, sending X_1 and X_2 from the first and second antenna. In the second time slot sending X_3 and X_4 from the first and second antenna, then X_5 and X_6 in the third time slot and so on. Note that, we need only $n/2$ time slots to accomplish the transmission since we are grouping two symbols and sending them in one time slot. Let us now explain how to extract two symbols which interfered with each other. The received signal of the first time slot is given by

$$\begin{bmatrix} Y_1^1 \\ Y_2^1 \end{bmatrix} = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \begin{bmatrix} Z_1^1 \\ Z_2^1 \end{bmatrix}. \tag{5.10}$$

Let us define the transmitted signal by X, the matrix of the channel response by $\mathbf{H}$, the received signal by Y and the noise vector by Z as follows:

$$Y = \begin{bmatrix} Y_1^1 \\ Y_2^1 \end{bmatrix}, \ \mathbf{H} = \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix}, \ X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \text{ and } \quad Z = \begin{bmatrix} Z_1^1 \\ Z_2^1 \end{bmatrix}.$$

Thus, equation (5.10) can be rewritten as

$$Y = \mathbf{H}\ X + Z. \tag{5.11}$$

We assume that the receiver knows Y_1^1, Y_2^1 and $H_{i,j}$. To solve equation (5.11) for X, we need to find the matrix $\mathbf{M}$ which satisfies $\mathbf{MH} = \mathbf{I}$. This matrix represents the pseudo-inverse for a general $(m \times n)$ matrix. For satisfying this constraint, the ZF linear detector is given by [102]

$$\mathbf{M} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H. \tag{5.12}$$

The term $\mathbf{H}^H\mathbf{H}$ can be represented as

$$\begin{aligned} \mathbf{H}^H\mathbf{H} &= \begin{bmatrix} H_{1,1}^* & H_{2,1}^* \\ H_{1,2}^* & H_{2,2}^* \end{bmatrix} \begin{bmatrix} H_{1,1} & H_{1,2} \\ H_{2,1} & H_{2,2} \end{bmatrix} \\ &= \begin{bmatrix} |H_{1,1}|^2 + |H_{2,1}|^2 & H_{1,1}^*H_{1,2} + H_{2,1}^*H_{2,2} \\ H_{1,2}^*H_{1,1} + H_{2,2}^*H_{2,1} & |H_{1,2}|^2 + |H_{2,2}|^2 \end{bmatrix}. \end{aligned} \tag{5.13}$$

It should be noted that diagonal elements in the matrix $\mathbf{H}^H\mathbf{H}$ are not zero. When performing the equalization, the ZF equalizer tries to null out the interfering terms, i.e when solving for $X1$, the interference from $X2$ tries to be canceled and vice versa. As a consequence, amplification of noise can occur. Thus, ZF equalizer is not the best possible equalizer, however, it is simple and practically easy to implement.

5.2.5 MIMO with MMSE equalizer

Let us consider the same MIMO system as with ZF equalizer. The MMSE method tries to find a coefficient $\mathbf{M}$ which minimizes the following criteria:

$$E\left\{\left[\mathbf{M}Y - X\right]\left[\mathbf{M}Y - X\right]^H\right\}. \tag{5.14}$$

According to [102], solving the above equation leads to

$$\mathbf{M} = (\mathbf{H}^H\mathbf{H} + N_0\mathbf{I}_2)^{-1}\mathbf{H}^H. \tag{5.15}$$

In fact, the MMSE equalizer is reduced to ZF equalizer when the noise term $N_0\mathbf{I}_2$ is null.

5.3 Nonlinear Complex M-SVR Approach

We note first that the index i and j throughout this section denotes the i^{th} and j^{th} antenna at the transmitter and receiver side of the MIMO system, respectively.

5.3.1 Principle

The MIMO-OFDM channel estimation approach is based on nonlinear complex M-SVR algorithm adapted to MIMO systems which is a generalized form of the SISO-SVR algorithm described in the previous chapter.

The M-SVR algorithm has two separate phases: learning phase and estimation phase. In learning phase, we estimate first the subchannels pilot symbols according to LS criterion to strike $min\left[(Y_j^P - \mathbf{X}_i^P\mathbf{F}h_{i,j})(Y_j^P - \mathbf{X}_i^P\mathbf{F}h_{i,j})^H\right]$, as

$$\widehat{H}_{i,j}^P = \mathbf{X}_i^{P^{-1}}Y_j^P, \tag{5.16}$$

where ${Y_j}^P = Y_j(l, m\Delta P)$ and $\widehat{H}_{i,j}^P = \widehat{H}_{i,j}(l, m\Delta P)$ are the received pilot symbols and the estimated frequency responses for the l^{th} OFDM symbol at pilot positions $m\Delta P$, respectively.

Then, in the estimation phase and by SVM interpolation mechanism, frequency responses of data subchannels can be determined. Therefore, frequency responses of all the OFDM subcarriers can be expressed as

$$\widehat{H}_{i,j}(l,k) = f_{i,j}(\widehat{H}_{i,j}^P(l, m\Delta P)), \tag{5.17}$$

where $k = 0, \cdots, N-1$, and $f_{i,j}(\cdot)$ is the interpolating function, which is determined by the nonlinear complex M-SVR approach.

5.3.2 M-SVR estimator formulation

The M-SVR estimator formulation is designed to be adapted to the MIMO-OFDM architecture. In fact, the regression function is represented as

$$\widehat{H}_{i,j}(m\Delta P) = \mathbf{w}_{i,j}^{H}\boldsymbol{\varphi}_{i,j}(m\Delta P) + b_{i,j} + e_{i,j}^{m}, \tag{5.18}$$

for $m = 0, \cdots, N_P - 1$ and $\mathbf{w}_{i,j}$ is the weight vector, $b_{i,j}$ is the bias term and residuals $\left\{e_{i,j}^{m}\right\}$ account for the effect of both approximation errors and noise.
In addition, we used ε-Huber robust cost function

$$\mathcal{L}^{\varepsilon}(e_{i,j}^{m}) = \begin{cases} 0, & \left|e_{i,j}^{m}\right| \leq \varepsilon \\ \frac{1}{2\gamma}(|e_{i,j}^{m}| - \varepsilon)^2, & \varepsilon \leq |e_{i,j}^{m}| \leq e_C \\ C(|e_{i,j}^{m}| - \varepsilon) - \frac{1}{2}\gamma C^2, & e_C \leq |e_{i,j}^{m}| \end{cases}. \tag{5.19}$$

Let $\mathcal{L}^{\varepsilon}(e_{i,j}^{m}) = \mathcal{L}^{\varepsilon}(\Re(e_{i,j}^{m})) + \mathcal{L}^{\varepsilon}(\Im(e_{i,j}^{m}))$, the primal problem can be stated as follows:

$$\begin{aligned} min \quad & \tfrac{1}{2} \quad \|\mathbf{w}_{i,j}\|^2 + \frac{1}{2\gamma}\sum_{m\in I_1}(\xi_{i,j}^{m} + \xi_{i,j}^{m*})^2 \\ & + \quad C\sum_{m\in I_2}(\xi_{i,j}^{m} + \xi_{i,j}^{m*}) + \frac{1}{2\gamma}\sum_{m\in I_3}(\zeta_{i,j}^{m} + \zeta_{i,j}^{m*})^2 \\ & + \quad C\sum_{m\in I_4}(\zeta_{i,j}^{m} + \zeta_{i,j}^{m*}) - \frac{1}{2}\sum_{m\in I_2, I_4}\gamma C^2 \end{aligned} \tag{5.20}$$

constrained to

$$\begin{aligned} \Re(\hat{H}_{i,j}(m\Delta P) - \mathbf{w}_{i,j}^{H}\boldsymbol{\varphi}_{i,j}(m\Delta P) - b_{i,j}) &\leq \varepsilon + \xi_{i,j}^{m} \\ \Im(\hat{H}_{i,j}(m\Delta P) - \mathbf{w}_{i,j}^{H}\boldsymbol{\varphi}_{i,j}(m\Delta P) - b_{i,j}) &\leq \varepsilon + \zeta_{i,j}^{m} \\ \Re(-\hat{H}_{i,j}(m\Delta P) + \mathbf{w}_{i,j}^{H}\boldsymbol{\varphi}_{i,j}(m\Delta P) + b_{i,j}) &\leq \varepsilon + \xi_{i,j}^{m*} \\ \Im(-\hat{H}_{i,j}(m\Delta P) + \mathbf{w}_{i,j}^{H}\boldsymbol{\varphi}_{i,j}(m\Delta P) + b_{i,j}) &\leq \varepsilon + \zeta_{i,j}^{m*} \\ \xi_{i,j}^{m(*)}, \zeta_{i,j}^{m(*)} &\geq 0, \end{aligned} \tag{5.21}$$

for $m = 0, \cdots, N_P - 1$, where $\xi_{i,j}^{m}$ and $\xi_{i,j}^{m*}$ are slack variables which stand for positive, and negative errors in the real part, respectively. $\zeta_{i,j}^{m}$ and $\zeta_{i,j}^{m*}$ are the errors for the imaginary parts.

The primal-dual functional can be derived from the primal functional (5.20) subject to constraints (5.21) as follows:

$$
\begin{aligned}
L_{Pd} &= \frac{1}{2}\|\mathbf{w}_{i,j}\|^2 + \frac{1}{2\gamma}\sum_{m\in I_1}(\xi_{i,j}^m + \xi_{i,j}^{m\,*})^2 + C\sum_{m\in I_2}(\xi_{i,j}^m + \xi_{i,j}^{m\,*}) \\
&+ \frac{1}{2\gamma}\sum_{m\in I_3}(\zeta_{i,j}^m + \zeta_{i,j}^{m\,*})^2 + C\sum_{m\in I_4}(\zeta_{i,j}^m + \zeta_{i,j}^{m\,*}) - \frac{1}{2}\sum_{m\in I_2, I_4}\gamma C^2 \\
&- \sum_{m=0}^{N_{P-1}}(\beta_{i,j}^m\xi_{i,j}^m + \beta_{i,j}^{m\,*}\xi_{i,j}^{m\,*}) - \sum_{m=0}^{N_{P-1}}(\lambda_{i,j}^m\zeta_{i,j}^m + \lambda_{i,j}^{m\,*}\zeta_{i,j}^{m\,*}) \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{R,m,i,j}[\Re(\hat{H}_{i,j}(m\Delta P) - \mathbf{w}_{i,j}^H\boldsymbol{\varphi}_{i,j}(m\Delta P) - b_{i,j}) - \varepsilon - \xi_{i,j}^m] \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{I,m,i,j}[\Im(\hat{H}_{i,j}(m\Delta P) - \mathbf{w}_{i,j}^H\boldsymbol{\varphi}_{i,j}(m\Delta P) - b_{i,j}) - \sqrt{-1}\varepsilon - \sqrt{-1}\zeta_{i,j}^m] \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{R,m,i,j}^*[\Re(-\hat{H}_{i,j}(m\Delta P) + \mathbf{w}_{i,j}^H\boldsymbol{\varphi}_{i,j}(m\Delta P) + b_{i,j}) - \varepsilon - \xi_{i,j}^{m\,*}] \\
&+ \sum_{m=0}^{N_{P-1}}\alpha_{I,m,i,j}^*[\Im(-\hat{H}_{i,j}(m\Delta P) + \mathbf{w}_{i,j}^H\boldsymbol{\varphi}_{i,j}(m\Delta P) + b_{i,j}) - \sqrt{-1}\varepsilon - \sqrt{-1}\zeta_{i,j}^{m\,*}],
\end{aligned}
\tag{5.23}
$$

with the dual variables constrained to $\alpha_{R,m,i,j}$, $\alpha_{I,m,i,j}$, $\beta_{i,j}^m$, $\lambda_{i,j}^m \geq 0$, $\alpha_{R,m,i,j}^*$, $\alpha_{I,m,i,j}^*$, $\beta_{i,j}^{m\,*}$, $\lambda_{i,j}^{m\,*} \geq 0$ and $\xi_{i,j}^m$,$\zeta_{i,j}^m$, $\xi_{i,j}^{m\,*}$, $\zeta_{i,j}^{m\,*} \geq 0$.
According to KKT complementarity conditions, we have

$$
\beta_{i,j}^m\xi_{i,j}^m = 0,\ \beta_{i,j}^{m\,*}\xi_{i,j}^{m\,*} = 0 \text{ and } \lambda_{i,j}^m\zeta_{i,j}^m = 0,\ \lambda_{i,j}^{m\,*}\zeta_{i,j}^{m\,*} = 0. \tag{5.24}
$$

Then, by making zero the primal-dual functional gradient with respect to $\omega_{i,j}$, we obtain an optimal solution of the weights

$$
\mathbf{w}_{i,j} = \sum_{m=0}^{N_P-1}\psi_{i,j}^m\boldsymbol{\varphi}_{i,j}(m\Delta P) = \sum_{m=0}^{N_P-1}\psi_{i,j}^m\boldsymbol{\varphi}_{i,j}(P_m), \tag{5.25}
$$

where $\psi_{i,j}^m = (\alpha_{R,m,i,j} - \alpha_{R,m,i,j}^*) + j(\alpha_{I,m,i,j} - \alpha_{I,m,i,j}^*)$ with $\alpha_{R,m,i,j}$, $\alpha_{R,m,i,j}^*$, $\alpha_{I,m,i,j}$, $\alpha_{I,m,i,j}^*$ are the Lagrange multipliers for real and imaginary parts of the residuals and $P_m = (m\Delta P), m = 0, \cdots, N_P - 1$ are the pilot positions.
Let the Gram matrix defined by

$$
\mathbf{G}_{i,j}(u,v) =< \boldsymbol{\varphi}_{i,j}(P_u), \boldsymbol{\varphi}_{i,j}(P_v) >= K_{i,j}(P_u, P_v), \tag{5.26}
$$

where $K_{i,j}(P_u, P_v)$ is a Mercer's kernel which represents the RBF kernel matrix. A compact form of the functional problem can be stated in matrix format by placing

optimal solution $\mathbf{w}_{i,j}$ into the primal dual functional and grouping terms. Therefore, the dual problem consists of

$$\begin{aligned} max \quad & - \quad \frac{1}{2}\boldsymbol{\psi}_{i,j}^{H}(\mathbf{G}_{i,j}+\gamma\mathbf{I})\boldsymbol{\psi}_{i,j} + \Re(\boldsymbol{\psi}_{i,j}^{H}Y_{j}^{P}) \\ & - \quad (\boldsymbol{\alpha}_{\mathbf{R}i,j} + \boldsymbol{\alpha}_{\mathbf{R}i,j}^{*} + \boldsymbol{\alpha}_{\mathbf{I}i,j} + \boldsymbol{\alpha}_{\mathbf{I}i,j}^{*})\mathbf{1}\varepsilon \end{aligned} \tag{5.27}$$

constrained to

$$0 \leq \alpha_{R,m,i,j}, \alpha_{R,m,i,j}^{*}, \alpha_{I,m,i,j}, \alpha_{I,m,i,j}^{*} \leq C, \tag{5.28}$$

where $\boldsymbol{\psi}_{i,j} = [\psi_{i,j}^{0}, \cdots, \psi_{i,j}^{N_P-1}]^T$; $\mathbf{I}$ and $\mathbf{1}$ are the identity matrix and the all-ones column vector, respectively; $\boldsymbol{\alpha}_{\mathbf{R}i,j}$ is the vector which contains the corresponding dual variables, with other subsets being similarly represented. The weight vector can be obtained by optimizing (5.27) with respect to $\alpha_{R,m,i,j}$, $\alpha_{R,m,i,j}^{*}$, $\alpha_{I,m,i,j}$, $\alpha_{I,m,i,j}^{*}$ and then substituting into (5.25).

Therefore, and after learning phase, frequency responses at all subcarriers in each OFDM symbol corresponding to the i^{th} transmitter and j^{th} receiver can be obtained by SVM interpolation

$$\hat{H}_{i,j}(k) = \sum_{m=0}^{N_p-1} \psi_{i,j}^{m} K_{i,j}(P_m, k) + b_{i,j}, \tag{5.29}$$

for $k = 1, \cdots, N$. Note that, the obtained subset of dual multipliers which are nonzero will provide with a sparse solution.

The free parameter of the kernel and the cost function have to be fixed by some a priori knowledge of the problem or by using one of the hyperparameters selection techniques.

5.4 Simulation Scenarios

LTE MIMO-OFDM downlink system with parameters shown in table 5.1 is simulated. These parameters are based on downlink LTE system. Note that, in LTE system the eNBs and UEs can have 2 or 4 antennas. When two or more transmitter antennas are applied, the pilot symbols are transmitted orthogonally in space. Indeed, these orthogonality in space is obtained by letting all other antennas be silent in the resource element in which one antenna transmits a pilot symbol [103].

For the purpose of evaluation the performance of the nonlinear M-SVR algorithm, we consider three scenarios for downlink LTE system. First, we deal with a scenario for downlink LTE SIMO system for a mobile speed equal to 120 Km/h. Then we treat the MIMO case. In the first part, we discuss MIMO diversity for space time block coding including Alamouti coding scheme. Here, we assume that the channel is constant during the period of Alamouti codeword transmission, so that the mobile speed used in the simulation is equal to 30 Km/h. In the second part, we discuss the performance of a

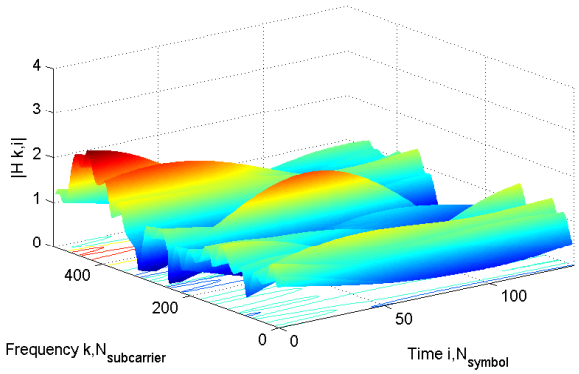


Figure 5.1: Time-varying channel frequency response for mobile speed = 30 Km/h.

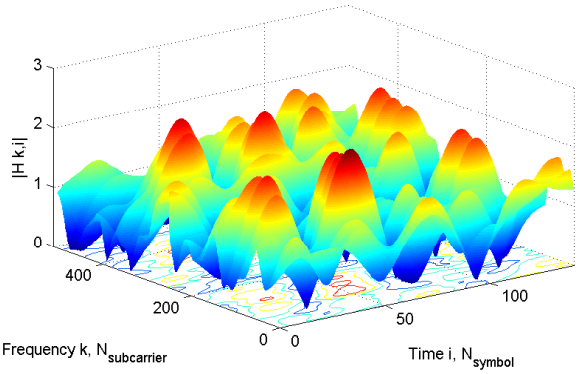


Figure 5.2: Time-varying channel frequency response for mobile speed = 120 Km/h.

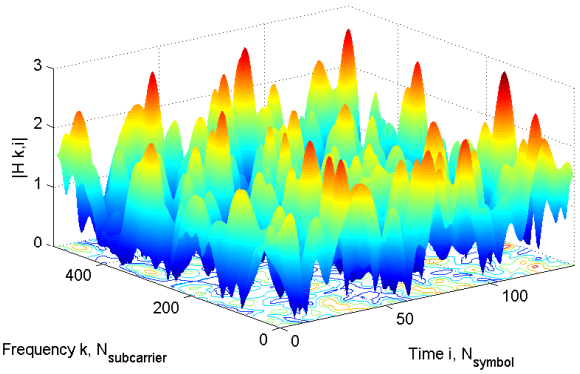


Figure 5.3: Time-varying channel frequency response for mobile speed = 350 Km/h.

Table 5.1: LTE transmission parameters.

Parameters	**Specifications**
MIMO-OFDM System	LTE/Downlink
LTE channel type	EVA
Constellation	16-QAM
$T_s(\mu s)$	72
$f_c(GHz)$	2.15
$\delta f(KHz)$	15
$B(MHz)$	5

spatially multiplexed MIMO system at higher speed (120 and 350 Km/h) by an ordered successive interference cancellation (OSIC) called also VBLAST.

The performance of the proposed M-SVR estimator is compared to other estimation techniques like LS, MMSE and Decision Feedback. Most simulations in this section are realized in the presence of nonlinear impulsive noise modeled as Bernoulli Gaussian process. Figs. 5.1, 5.2 and 5.3 present the time-varying channel frequency responses of the three different mobile speed of simulation 30, 120 and 350 Km/h.

5.5 Performance Analysis and Discussion

5.5.1 SIMO case: M-SVR-MRC evaluation

In SIMO systems, receivers extract multiple signal branches of the same signal received from different channels and apply gain combining techniques such as Maximum Ratio Combining (MRC) which gives better performance than Equal Gain Combining (EGC) or Selective Combining (SC) [65] [104].

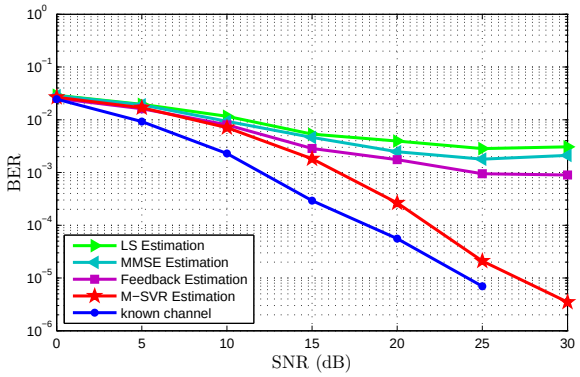


Figure 5.4: BER as a function of SNR for a SIMO system 1×2 for a mobile speed at 120 Km/h without impulsive noise.

Fig. 5.4 shows the performance of the LS, MMSE, Decision Feedback and nonlinear complex M-SVR algorithms applied on a SIMO system with two receive antennas in the presence of additive Gaussian noise as a function of SNR without impulsive noise. A poor performance is noticeably exhibited by LS, MMSE and Decision Feedback, whereas better performance is observed with our nonlinear complex M-SVR estimator.

Then, the performance of the nonlinear M-SVR with MRC scheme is compared to SVR SISO-OFDM and evaluated with the variation of the number of receive antennas.

Fig. 5.5 shows the variation of BER as a function of SNR in the presence of additive Gaussian noise with impulsive noise (SIR = -5 and -10 dB with p = .05 and .1) for a mobile speed equal to 120 Km/h.

M-SVR-MRC performs better than SVR SISO-OFDM case. For example, at BER = 10^{-3} a gain of 18 dB is obtained for $N_R = 4$ with SIR = -5 dB and p = .05. This is thanks to the increase of diversity at the receiver side. The same results are confirmed with SIR = -10 dB and p = .1.

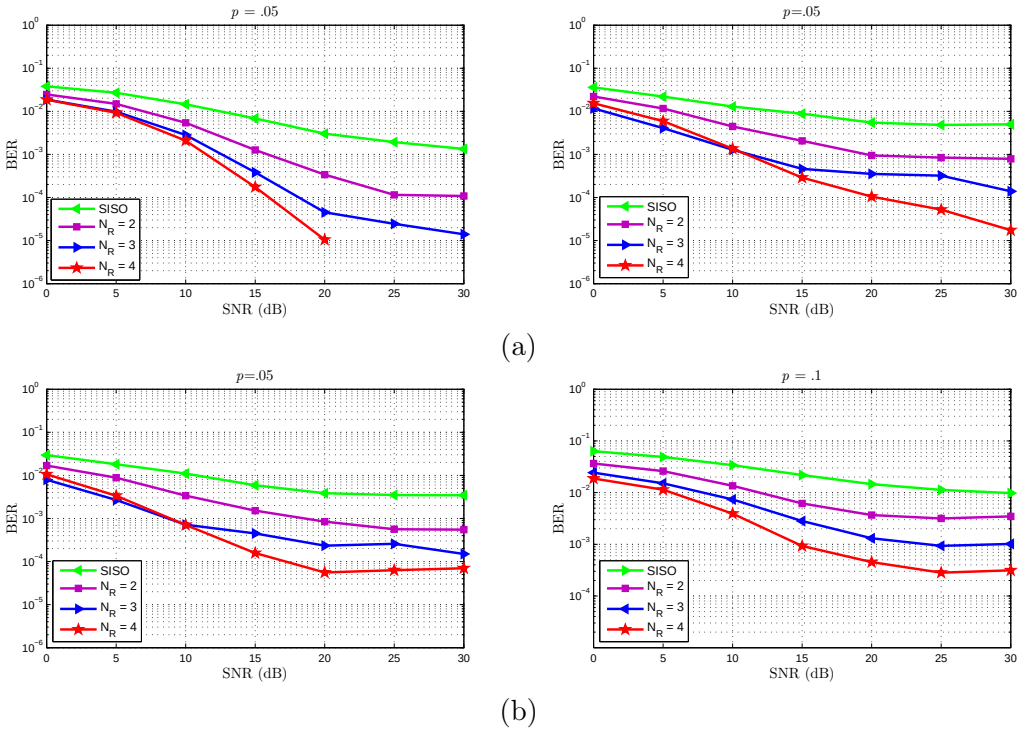


Figure 5.5: The performance of M-SVR-MRC for a mobile speed at 120 Km/h with (a) SIR = -5 dB and (b) SIR = -10 dB as a function of SNR for p = .05 and .1.

We can remark that BER decreases and good performance is obtained when the number of receive antennas increases. Fig. 5.6 presents the performance of the M-SVR-MRC as a function of SIR in the presence of AWGN and impulsive noise interfering with pilot signals for a mobile speed at 120 and 350 Km/h for SNR = 30 dB and p =.05 and .1. The comparison between M-SVR estimator in SIMO system and SVR estimator

in SISO system reveals that M-SVR outperforms SVR estimator. Indeed, for SIMO system, good performance is realized when the number of receive antennas increases which improves the receive diversity.

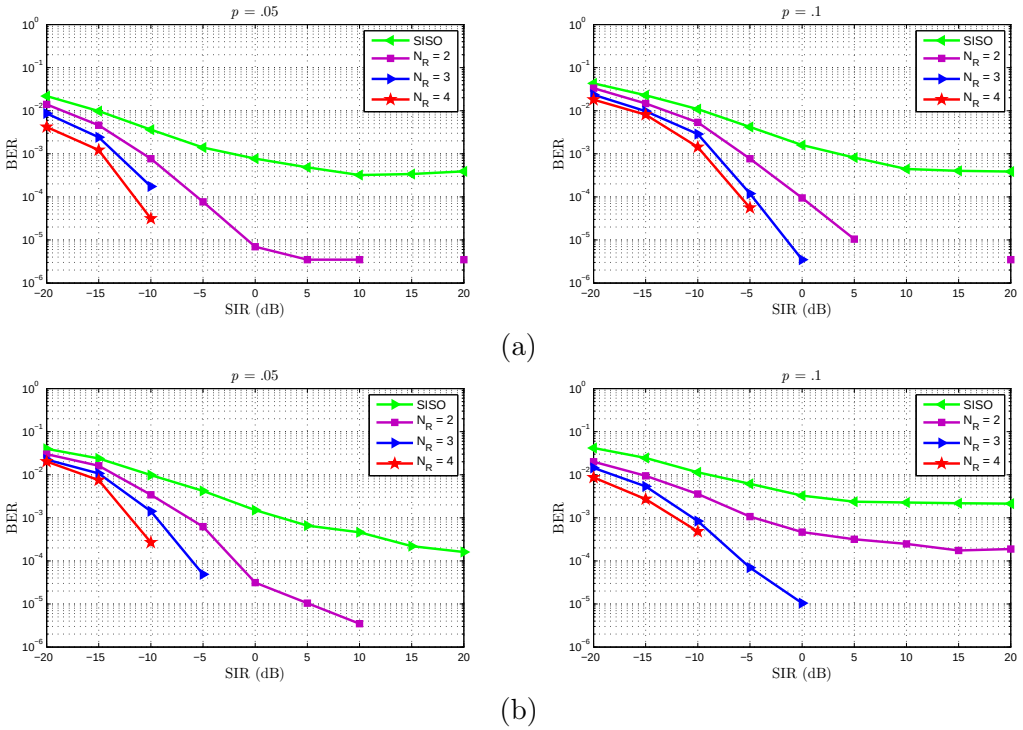


Figure 5.6: The performance of M-SVR-MRC for a mobile speed at (a) 120 Km/h and (b) 350 Km/h as a function of SIR with SNR = 30 dB and $p = .05$ and .1.

5.5.2 MIMO case

5.5.2.1 M-SVR-STBC evaluation

- BER vs. SNR evaluation

In Fig. 5.7 we study the performance of the nonlinear complex M-SVR estimator for a MIMO system with Alamouti 2×2 encoding scheme for a mobile speed at 30 Km/h in terms of BER evaluation without impulsive noise. The performance of our nonlinear complex M-SVR estimator is compared to the aforementioned techniques for estimating the channel coefficients for different SNR values. As we can see, the complex M-SVR approach outperforms LS, MMSE and Decision Feedback estimators for all SNR values.

Fig. 5.8 presents the performance of the M-SVR estimator with Alamouti encoding scheme (2×1 and 2×2) and STBC encoding scheme (3×4 and 4×4) as a function of SNR in the presence of impulsive noise (SIR = -5 and -10 dB with $p = .05$ and .1).

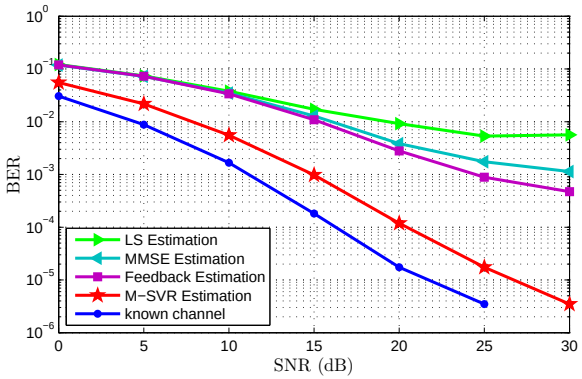


Figure 5.7: BER as a function of SNR for a MIMO system with Alamouti 2×2 encoding scheme for a mobile speed at 30 Km/h without impulsive noise.

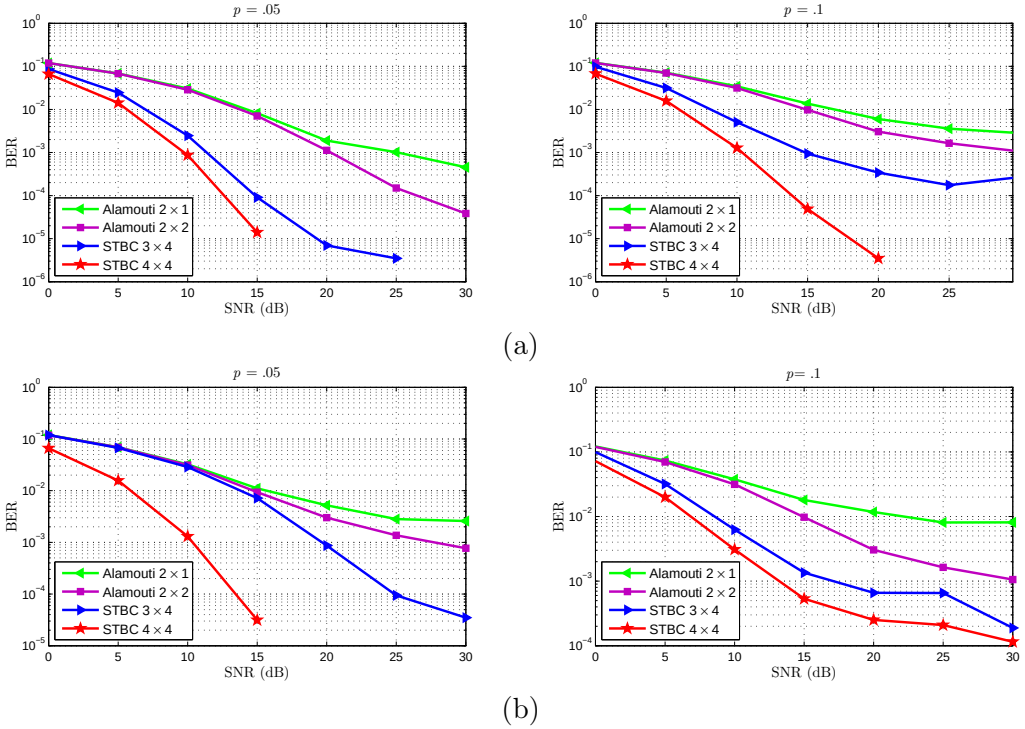


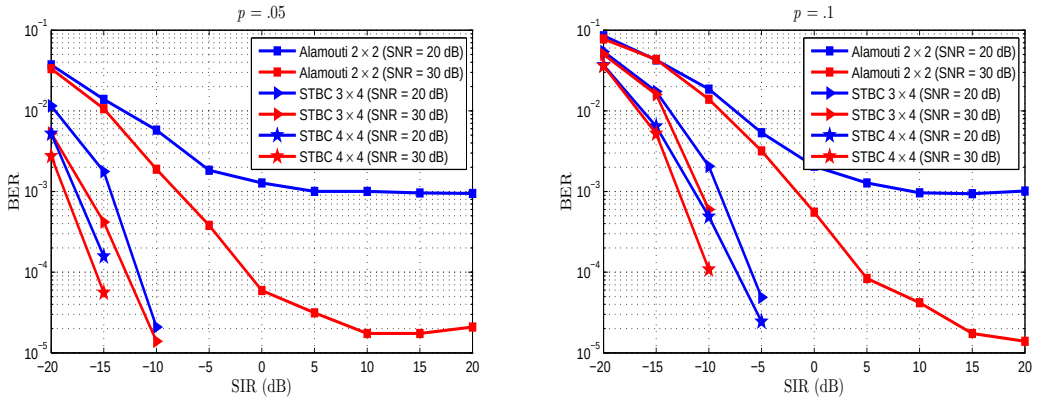
Figure 5.8: BER performance of M-SVR with Alamouti encoding and STBC encoding schemes for a mobile speed at 30 Km/h with (a) SIR = -5 dB and (b) SIR = -10 dB as a function of SNR for $p = .05$ and .1.

We notice here that with STBC (4×4) encoding scheme, the nonlinear complex M-SVR approach achieves a considerably better accuracy for all simulated cases due to the increase of transmit and receive diversity.

- BER vs. SIR evaluation

In Fig. 5.9 we study the performance of the nonlinear complex M-SVR estimator for a MIMO system with Alamouti and STBC encoding schemes for a mobile speed at 30 Km/h in the presence of impulsive noise as a function of SIR with SNR = 20 and 30 dB with $p = .05$ and .1.

As we can see, the complex M-SVR method performs better for small SIR values (SIR < 0 dB) where the amount of nonlinear impulsive noise increases. We notice that with high SNR (SNR = 30 dB in our case), the nonlinear complex M-SVR estimator achieves a significantly better estimation, especially for small value of p ($p = .05$ in our case).



(a) (b)

Figure 5.9: BER performance of M-SVR with Alamouti encoding and STBC encoding schemes for a mobile speed at 30 Km/h as a function of SIR for SNR = 20 and 30 dB with (a) $p = .05$ and (b) $p = .1$.

5.5.2.2 M-SVR-VBLAST evaluation

- BER vs. SNR evaluation

Fig. 5.10 shows the performance of M-SVR-VBLAST in terms of BER evaluation for a MIMO system (2×3) and (2×4) with V-BLAST decoding scheme in the presence of impulsive noise (SIR = -5 dB and SIR = -10 dB) with $p = .05$ and .1. The simulation is performed for a mobile speed at 120 Km/h (Fig. 5.10 (a)) and 350 Km/h (Fig. 5.10 (b)).

It is clear that the complex M-SVR scheme achieves a better accuracy with high number of receive antennas ($N_R = 4$ in our case). This result is expected since the increase of number of receive antennas implies the increase of receive diversity.

We notice that better accuracy is also achieved for low values of p.

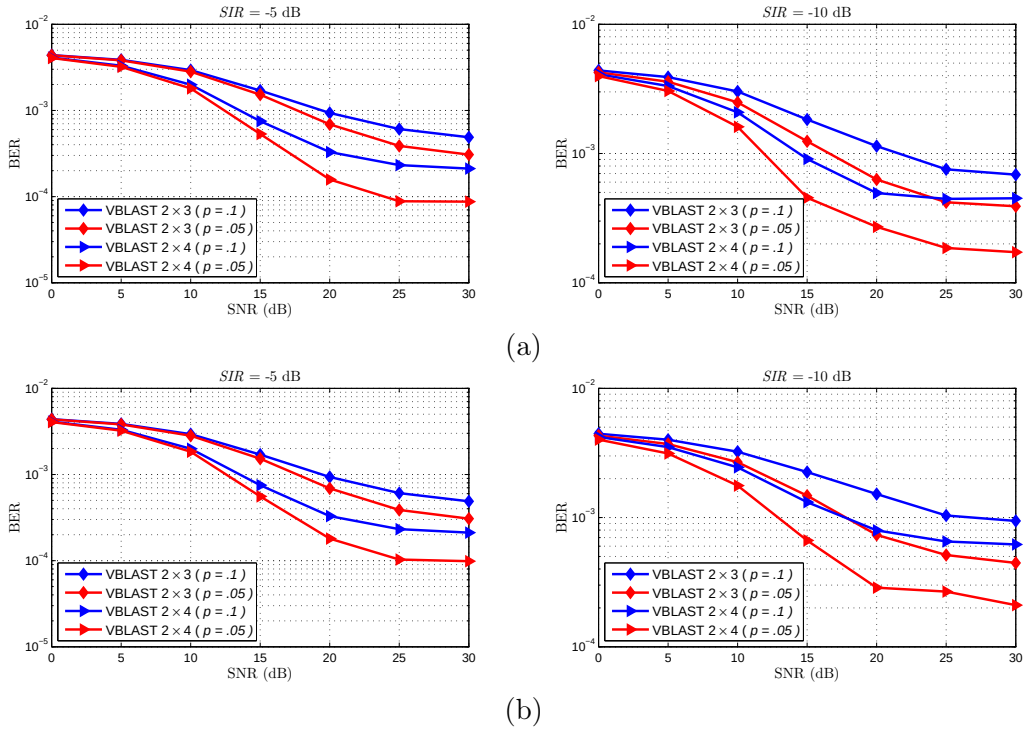


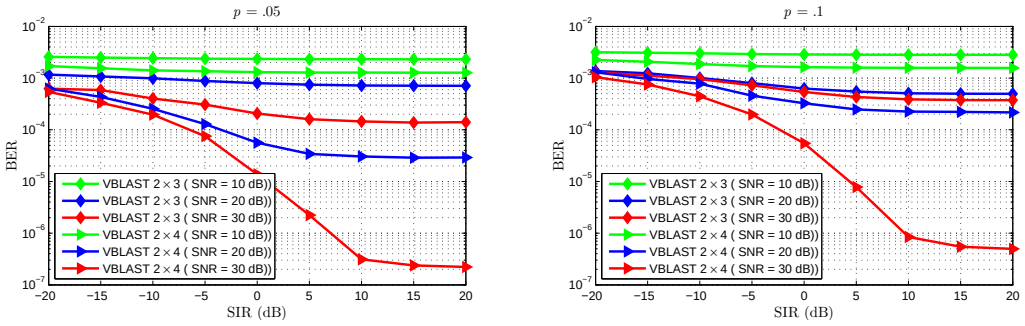
Figure 5.10: BER performance of M-SVR-VBLAST as a function of SNR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h with SIR = -5 dB and -10 dB for $p = .05$ and .1.

- BER vs. SIR evaluation

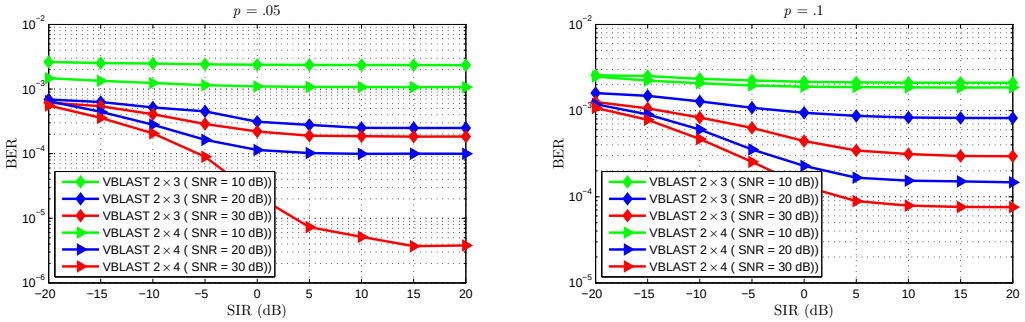
In Fig. 5.11 we study the performance of the nonlinear complex M-SVR estimator for a MIMO system with V-BLAST decoding scheme for a mobile speed at 350 Km/h in the presence of impulsive noise as a function of SIR with SNR = 10, 20 and 30 dB with $p = .05$ and .1.

As we can see, the complex M-SVR method achieves a better accuracy for high SIR values, but it is clear that for (SIR $<$ 0 dB) the M-SVR technique performs better since for (SIR $>$ 0 dB), the BER is almost constant for each case of simulation.

Note that with high number of receive antennas ($N_R = 4$) and high SNR (SNR = 30 dB) in our simulation case, the nonlinear complex M-SVR estimator achieves a significantly better estimation, especially for low value of p ($p = .05$).



(a)



(b)

Figure 5.11: BER performance of M-SVR-VBLAST as a function of SIR for a mobile speed at (a) 120 Km/h and (b) 350 Km/h with SNR = 10, 20 and 30 dB for $p = .05$ and .1.

5.6 Conclusion

In this chapter, we proposed a new approach to estimate the MIMO-OFDM mobile radio channel based on nonlinear complex M-SVR technique. Then, we applied this M-SVR channel estimator to a downlink MIMO-LTE system in the presence of impulsive noise interfering with OFDM pilot symbols in several mobility environments.

In fact, the proposed method is based on training process that uses learning sequence to estimate the channel variations. Pilot symbols are inserted into different subcarriers and transmitted over different antennas of the LTE system in order to increase the convergence rate and estimation accuracy. Through simulation, results have confirmed the capabilities of the proposed nonlinear complex M-SVR estimator when compared to other standard methods in the presence of various nonlinearities.

Conclusion

In this thesis, we developed new algorithms based on nonlinear complex Support Vector Machines Regression (SVR) to estimate mobile radio channel coefficients for pilot-aided SISO-OFDM and MIMO-OFDM LTE systems. This kernel-based algorithms involved in pattern analysis which is considered as the process of finding general relations in a set of data. Indeed, the SVR channel estimation algorithm uses knowledge of the pilot signals to estimate the total frequency response of the channel.

In highly selective multipath fading channel, where complicated nonlinearities can be present, the estimation precision can be lowed by using linear methods. So, we adapt the nonlinear complex SVR algorithm since it transforms the nonlinear estimation in low dimensional space into linear estimation in high dimensional space improving thus the estimation precision.

The 3GPP-LTE has been taken as a study case. The Ph.D thesis has been structured in five chapters:

In the first chapter, we have presented a historical overview on support vector machines, especially the founding papers of Vapnik and Chervonenkis. The learning methodology and the linear learning machine have been also presented. These linear learning machine leads to a simplicity and flexibility of learning. We have also introduced generalization and statistical learning theory. It is this theory which limits the generalization error, not only for data from the training set, but also for future data. The optimization theory has been also presented. In fact, this theory forms the basis of optimization problems, particularly the convex quadratic optimization problems which are used in the case of SVMs. We have described in this chapter the differences between SVMs and neural networks. Indeed, the main advantage of SVMs is the convergence to a unique solution.

In the second chapter, we have studied the support vector machines. We have shown how learning theory can be used to avoid the difficulties of applying linear functions in high dimension feature spaces with induced kernels. We have also presented the transformation approach of primal optimization problems into quadratic convex dual forms for classification and regression. In the case of regression, the loss function penalizes errors only above the threshold ε. Such a loss function typically leads to a fine representation of the decision rule giving significant advantages. Thus, in this chapter we have described first the linear support vector machines for classification and regression. Then, the theory of kernel functions associated with SVMs has been analyzed. Finally, we have presented classification and regression support vector machines in their

nonlinear versions.

In the third chapter, we have analyzed the multipath propagation aspects of the mobile radio channel as well as its mathematical model. An overview of nonlinear impulsive noise which can often appear in communication systems is also presented. Moreover, we have simulated LTE channels based on 3GPP specifications and several scenarios of propagation that are considered in our work to evaluate the performance of channel estimation techniques.

We have presented in the fourth chapter the nonlinear complex SVR based channel estimation technique for a highly selective downlink LTE system. The proposed channel estimation method has used pilot signals to estimate the channel variations. As the main goal of the machine learning is to estimate a function based on some training data, our method is based on a learning process that has used a training sequence for adaptation to achieve a desired performance. Comparative study with some traditional techniques such as the LS, MMSE and decision feedback have been conducted. Simulation results have shown clearly the high performance of the proposed nonlinear SVR channel estimation technique when compared to the aforementioned standard methods especially in the presence of nonlinear impulsive noise under high mobility conditions. The proposal takes into account the temporal-spectral relationship of the OFDM signal for a highly selective channels. The Gram matrix using RBF kernel provides a natural nonlinear extension of the complex linear SVR which leads to a significant benefit for OFDM communications especially in those scenarios in which deep fading is present.

In the last chapter, we have proposed an efficient nonlinear complex Multiple Support Vector Machines Regression (M-SVR) approach to track the multipath fading channel in mobile wireless MIMO-OFDM system with the presence of Gaussian and non-Gaussian impulsive noise. The M-SVR estimation technique is developed and applied to MIMO LTE downlink. Different scenarios are simulated and discussed in the case of SIMO and MIMO systems with both STBC and V-BLAST schemes. Simulation results have confirmed the capabilities of the proposed nonlinear complex M-SVR estimator when compared to other standard methods in the presence of various nonlinearities.

In future work, we can estimate other wireless parameters such as the mean Angle of Arrival (AoA), the Angle Spread (AS), the Doppler Spread (DS), the Ricean K-factor, the SNR, etc, all using SVR. Also we can apply SVR to solve mobile location problems in NLOS propagation environment. Moreover, modulation recognition can be developed for OFDM signals by means of SVM classification in order to identify modulation in multipath environments.

We can extend the introduced methods to be applied on the LTE advanced system which is an enhancement of LTE standard and evaluate them using our background in SVM.

Finally, because SVM are based on sound theoretical foundation and the solution it produces are global and unique in nature, nowadays they are the most popular prediction modeling techniques in the data mining arena. Their use and popularity will increase as the popular commercial data mining tools start to incorporate them into their modeling arsenal [6].

Appendix A

Hilbert Spaces

Definition A.1
A space $\mathcal{H}$ is *separable* if there exists a countable subset $E \subseteq \mathcal{H}$, such that every element of $\mathcal{H}$ is the limit of a sequence of elements of E. A Hilbert space is a complete separable inner product space.■

Finite dimensional vector spaces such as $\mathbb{R}^n$ are Hilbert spaces.

Theorem A.1
Let $\mathcal{H}$ be a Hilbert space, D a closed subspace of $\mathcal{H}$ and $\boldsymbol{x} \in \mathcal{H}$. There is a unique vector $\boldsymbol{d}_0 \in D$, known as the projection of $\boldsymbol{x}$ onto D, such that

$$\|\boldsymbol{x} - \boldsymbol{d}_0\| \leq inf\left\{\|\boldsymbol{x} - \boldsymbol{d}\|, \boldsymbol{d} \in D\right\}. \tag{A.1}$$

A necessary and sufficient condition for $\boldsymbol{d}_0 \in D$ to be the projection of $\boldsymbol{x}$ onto D is that the vector $\boldsymbol{x} - \boldsymbol{d}_0$ be orthogonal to vectors in D.

A consequence of this theorem is that the best approximation to $\boldsymbol{x}$ in the subspace D generated by the orthonormal vectors $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n\}$ is given by its Fourier series

$$\sum_{i=1}^{n} < \boldsymbol{x} \cdot \boldsymbol{e}_i > \boldsymbol{e}_i. \tag{A.2}$$

This leads naturally to studying the properties of series like the case of infinite bases.

Definition A.2
If S is an orthonormal set in a Hilbert space $\mathcal{H}$ and no other orthonormal set contains S as a proper subset, that is S is maximal, then S is called a complete orthonormal system or an orthonormal basis for $\mathcal{H}$.■

Theorem A.2
Every Hilbert space $\mathcal{H}$ has an orthonormal basis. Suppose that $S = \{\boldsymbol{x}_\alpha\}_{\alpha \in A}$ is an orthonormal basis for a Hilbert space $\mathcal{H}$. So, $\forall\ \boldsymbol{y} \in \mathcal{H}$,

$$\boldsymbol{y} = \sum_{\alpha \in A} < \boldsymbol{y} \cdot \boldsymbol{x}_\alpha > \boldsymbol{x}_\alpha. \tag{A.3}$$

and $\|\boldsymbol{y}\|^2 = \sum_{\alpha\in A} |< \boldsymbol{y}\cdot\boldsymbol{x}_\alpha >|^2$.

This theorem states that, as in the finite dimensional case, every element of a Hilbert space $\mathcal{H}$ can be expressed as a linear combination of (possibly infinite) basis elements.

The coefficients $\boldsymbol{y}\cdot\boldsymbol{x}_\alpha$ are often called the Fourier coefficients of $\boldsymbol{y}$ with respect to the basis $S = \{\boldsymbol{x}_\alpha\}_{\alpha\in A}$.

Example A.1
Consider countable sequences of real numbers. The Hilbert space L_2 is the set of sequences $z = \{z_1, z_2, \cdots, z_i, \cdots\}$ such that

$$\|z\|_2^2 = \sum_{i=1}^{\infty} {z_i}^2 < \infty, \tag{A.4}$$

where the inner product of sequences $\boldsymbol{x}$ and $\boldsymbol{z}$ is defined by

$$< \boldsymbol{x}\cdot\boldsymbol{z} >= \sum_{i=1}^{\infty} x_i z_i. \tag{A.5}$$

If $\boldsymbol{\mu} = \{\mu_1, \mu_2, \cdots, \mu_i, \cdots\}$ is a countable sequence of positive real numbers, the Hilbert space $L_2(\boldsymbol{\mu})$ is the set of sequences $z = \{z_1, z_2, \cdots, z_i, \cdots\}$ such that

$$\|z\|_2^2 = \sum_{i=1}^{\infty} \mu_i {z_i}^2 < \infty, \tag{A.6}$$

where the inner product of sequences $\boldsymbol{x}$ and $\boldsymbol{z}$ is defined by

$$< \boldsymbol{x}\cdot\boldsymbol{z} >= \sum_{i=1}^{\infty} \mu_i x_i z_i. \tag{A.7}$$

The normed space L_1 is the set of sequences $z = \{z_1, z_2, \cdots, z_i, \cdots\}$ for which

$$\|z\|_1 = \sum_{i=1}^{\infty} |z_i| < \infty. \tag{A.8}$$

Example A.2
Consider the set of continuous real-valued functions on a subset X of $\mathbb{R}^n$. The Hilbert space $L_2(X)$ is the set of functions f for which

$$\|f\|_{L_2} = \int_X f(x)^2 dx < \infty, \tag{A.9}$$

where the inner product of functions f and g is defined by

$$< f\cdot g >= \int_X f(x)g(x)dx. \tag{A.10}$$

The normed space $L_\infty(X)$ is the set of functions for which

$$\|f\|_{L_\infty} = \sup_{\boldsymbol{x}\in X} |f(x)| < \infty. \tag{A.11}$$

Appendix B

LTE Downlink Data Transmission

Downlink resource grid

The transmitted signal in each LTE slot is described by a resource grid of N_{BW} subcarriers and N_{sym} OFDM symbols. The bandwidth is allocated to the UEs in terms of resource blocks in order to achieve multiple access. A physical resource block, N_{RB} consists of N_{sym} consecutive OFDM symbols in the time domain and 12 consecutive subcarriers in the frequency domain.

The number of available physical resource blocks depends on the bandwidth since the resource block size is the same for all bandwidths. Each UE can be assigned one or more resource block in each transmission time interval of 1 ms depending on the required data rate. Note that the scheduling decision is done at the NodeB and the user data is carried on the Physical Downlink Shared Channel (PDSCH). The scheduling decisions are transported by the Physical Downlink Control Channel (PDCCH) to individual UEs [105].

According to Release 8, the transmitted bits are modulated using Quadrature Amplitude Modulation (QAM), and the available modulation schemes are 4-QAM, 16-QAM and 64-QAM [78].

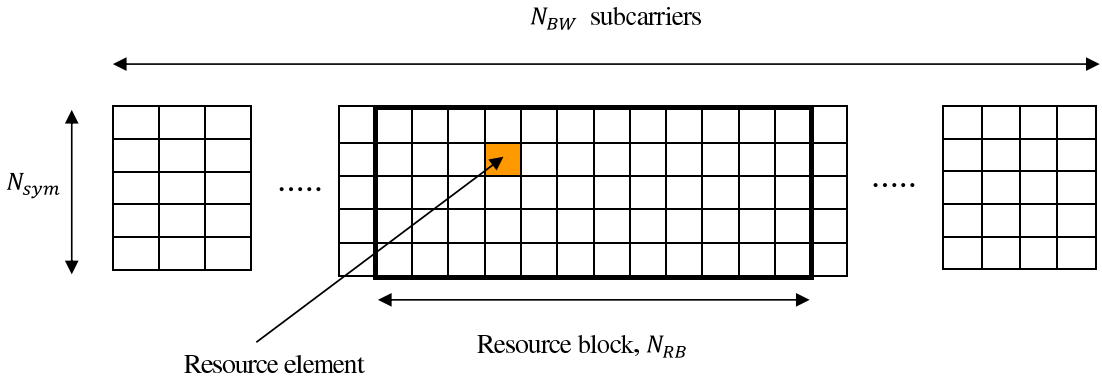


Figure B.1: Downlink resource grid.

Downlink reference signal structure

The downlink reference signal structure is important for channel estimation and cell search. Resource elements in the time-frequency domain are carrying the pilot signal sequence, which is predefined for each cell. The pilot symbols are placed in the first OFDM symbol of one slot and on the third last OFDM symbol. The spacing between the pilot symbols is 6 subcarriers.

In the LTE system, the NodeBs and UEs can have 2 or 4 antennas. When two or more transmitter antennas are applied, the pilot symbols are transmitted such that they are orthogonal in space. The orthogonality in space is obtained by letting all other antennas be silent in the resource element in which one antenna transmits a pilot symbol [106], [105]. Fig. B.2 put on view the positions of the pilot symbols for transmission with two antennas as an example. Antenna 2 is silent when antenna 1 transmits a pilot symbol and vice versa.

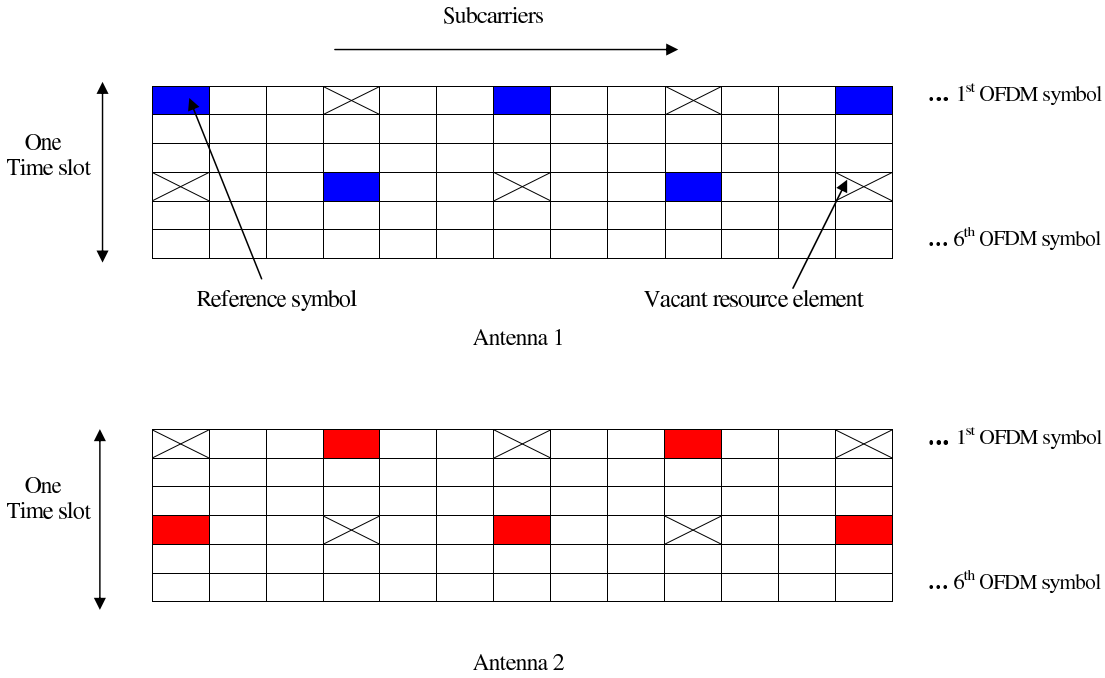


Figure B.2: The pilot symbol structure for one slot with 6 OFDM symbols using two antennas.

Publications

- A. CHARRADA and A. SAMET, "M-SVR Based MIMO-OFDM Channel Estimator Using Comb Type Pilot Arrangement," *International Conf. on electrical engineering and software applications (ICEESA)*, March 2013.
- A. CHARRADA and A. SAMET, "Support Vector Machines Regression for MIMO-OFDM Channel Estimation," *International journal of artificial intelligence (IJ-AI)*, vol. 1, no. 4, pp. 214-224, December 2012.
- A. CHARRADA and A. SAMET, "Estimation of Highly Selective Channels for OFDM System by Complex Least Squares Support Vector Machines," *International journal of electronics and communications (AËU)*, vol. 66, no. 8, pp. 687-692, 2012.
- A. CHARRADA and A. SAMET, "SIMO-OFDM Channel Estimation Based on Nonlinear Complex LS-SVM," *International journal of computer applications (IJ-CA)*, vol. 42, no. 3, pp. 1-8, March 2012.
- A. CHARRADA and A. SAMET, "Nonlinear Complex LS-SVM for Highly Selective OFDM Channel With Impulse Noise," 6^{th} *IEEE conference on sciences of electronics, technologies of information and telecommunications (SETIT)*, pp. 696-700, March 2012.
- A. CHARRADA and A. SAMET, "Complex Support Vector Machine Regression for Robust Channel Estimation in LTE Downlink System," *International journal of computer networks & communications (IJCNC)*, vol. 4, no. 1, pp. 211-223, January 2012.
- A. CHARRADA and A. SAMET, "Nonlinear Channel Estimation for OFDM System by Complex LS-SVM Under High Mobility Conditions," *International journal of wireless & mobile networks (IJWMN)*, vol. 3, no. 4, pp. 175-185, August 2011.

Bibliography

[1] J. Dias, E.L. Pinto, M.V. Ribeiro, "Time-Varying Channel Characterization based on SVM under the Impulsive Noise Presence," *IEEE international symposium on power line communications and its applications*, pp. 261–266, 2010.

[2] V. Vapnik, "The Nature of Statistical Learning Theory," *Springer-Verlag, NY, USA*, 1995.

[3] ——, "Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications and Control," *John Wiley Sons*, 1998.

[4] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," *Cambridge University Press, UK.*, 2000.

[5] K.R. Müller, S. Mika, G. Rätsch, K. Tsuda and B. Schölkopf, "An Introduction to Kernel-based Learning Algorithms," *IEEE transactions on neural networks*, vol. 12, no. 2, pp. 181–202, 2001.

[6] J.S. Taylor and N. Cristianni, "Kernel Methods for Pattern Analysis," *Cambridge University Press*, 2004.

[7] D.L. Olson and D. Delen, "Advanced Data Mining Techniques," *Springer-Verlag, Berlin*, 2008.

[8] J. Martin-Guerrero, G. Camps-Valls, E. Soria-Olivas, A.J. Serrano-Lopez, J.J. Perez-Ruixo and N.V. Jimenez-Torres, "Dosage Individualization of Erythropoietin using a Profile-Dependent Support Vector Regression," *IEEE transactions on biomedical engineering*, vol. 50, no. 10, pp. 1136–1142, 2003.

[9] G. Camps-Valls, J. Munoz-Mari, L. Gomez-Chova, K. Richter and J. Calpe-Maravilla, "Biophysical Parameter Estimation with a Semisupervised Support Vector Machine," *IEEE Geoscience and remote sensing letters*, vol. 6, no. 2, pp. 248–252, 2009.

[10] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Francs, J. Amoros-lopez and J. Calpe-Maravilla, "Retrieval of Oceanic Chlorophyll Concentration with Relevance Vector Machines," *Journal of remote sensing of environment*, vol. 105, pp. 23–33, 2006.

[11] L. Gomez-Chova, L. Alonso, L. Guanter, G. Camps-Valls, J. Calpe and J. Moreno, "Correction of Systematic Spatial Noise in Push-broom Hyperspectral Sensors: Application to CHRIS/PROBA Images," *journal on applied optics*, vol. 47, no. 28, pp. 46–60, 2008.

[12] J. Munoz-Mari, F. Bovolo, L. Gomez-Chova, L. bruzzone and G. Camps-Valls, "Semisupervised One-Class Support Vector Machines for Classification of Remote Sensing Data," *IEEE transactions on geoscience and remote sensing*, vol. 48, no. 8, pp. 3188–3197, 2010.

[13] D. Tuia, G. Camps-Valls, G. Matasci and M. Kanevski, "Learning Relevant Image Features With Multiple-Kernel Classification," *IEEE transactions on geoscience and remote sensing*, vol. 48, no. 10, pp. 3780–3791, 2010.

[14] M. Marconcini, G. Camps-Valls, and L. Bruzzone, "A Composite Semisupervised SVM for Classification of Hyperspectral Image," *IEEE Geoscience and remote sensing letters*, vol. 6, no. 2, pp. 234–238, 2009.

[15] D. Tuia and G. Camps-Valls, "Semisupervised Remote Sensing Image Classification with Cluster Kernels," *IEEE transactions on geoscience and remote sensing*, vol. 6, no. 2, pp. 224–228, 2009.

[16] G. Camps-Valls and L. Bruzzone, "Kernel-Based Methods for Hyperspectral Image Classification," *IEEE transactions on geoscience and remote sensing*, vol. 43, no. 6, pp. 1–12, 2005.

[17] F. garcia-Vilchez, J. Munoz-Mari, M. Zortea, I. Blanes, V. Gonzalez-Ruis, G. Camps-Valls, A. Plaza and J. Serra-sagrista, "On the Impact of Lossy Compression on Hyperspectral Image Classification and Unmixing," *IEEE geoscience and remote sensing letters*, vol. 8, no. 2, pp. 253–257, 2011.

[18] G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, J.D. Martin-Guerrero, E. Soria-Olivas, L. alonso-Chorda and J. moreno, "Robust Support Vector Method for Hyperspectral Data Classification and Knowledge Discovery," *IEEE transactions on geoscience and remote sensing*, vol. 42, no. 7, pp. 1530–1542, 2004.

[19] T. Joachims, "Text Categorization with Support Vector Machines: Learning with many Relevant Features," 10^{th} *european conference on machine learning*, pp. 137–142, 1998.

[20] S. Laconte, S. Strother, V. Cherkassky, J. Anderson and X. Hu, "Support Vector Machines for Temporal Classification of Block Design fmri Data," *Neuroimage*, vol. 26, pp. 317–329, 2005.

[21] M. Martínez-ramon and S. Posse, "Fast Estimation of Optimal Resolution in Classification Spaces for fmri Pattern Recognition ," *Neuroimage*, vol. 31, no. 3, pp. 1129–1141, 2006.

[22] J.L. Rojo-Alvarez, C. Figuera, C.E. Martínez-Cruz, G. Camps-Valls, F. Alonso-Atienza and M. Martínez-Ramón, "Nonuniform Interpolation of Noisy Time Series using Support Vector Machines," *IEEE transactions on signal processing*, vol. 55, no. 8, pp. 4116–4126, 2007.

[23] M. Martinez-Ramon, J.L. Rojo-Alvarez, G. Camps-Valls, J. Munoz-Mari, A. Navia-Vazquez, E. Soria-Olivas and A. R. Figueiras-Vidal, "Support Vector Machines for Nonlinear Kernel ARMA System Identification," *IEEE transactions on neural networks*, vol. 17, no. 6, pp. 1617–1622, 2006.

[24] J.L. Rojo-Alvarez, M. Martínez-Ramón, M. de Prado-Cumplido, A. Artés-Rodriguez and A.R. Figueiras-Vidal, "Support Vector Method for Robust ARMA System Identification," *IEEE transactions on signal processing*, vol. 52, no. 1, pp. 155–164, 2004.

[25] A. Ganapathiraju, J. Hamaker and J. Picone, "Applications of Support Vector Machines to Speech Recognition," *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

[26] J. Picone, "Signal Modeling Techniques in Speech Recognition," *IEEE proceedings*, vol. 81, no. 9, pp. 1215–1247, 1993.

[27] J.L. Rojo-Álvarez, G. Camps-valls, M. Martínez Ramón, A. Navia-Vázquez and A.R. Figueiras-Vidal , "A Support Vector Framework for Linear Signal Processing," *Signal processing*, vol. 85, no. 12, pp. 2316–2326, 2005.

[28] P. Pontil and A. Verri, "Support Vector Machines for 3D Object Recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 6, pp. 637–646, 1998.

[29] K.I. Kim, M.O. Franz and B. Schölkopf, "Iterative Kernel Principal Component Analysis for Image Modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 9, pp. 1351–1366, 2005.

[30] D. Cremers, "Shape Statistics in Kernel Space for Variational Image Segmentation," *Pattern recognition*, vol. 36, no. 9, pp. 1929–1943, 2003.

[31] G. Gómez-Prez, G. Camps-Valls, J. Gutirrez and J. Malo, "Perceptual Adaptive Insensitivity for Support Vector Machine Image Coding," *IEEE transactions on neural networks*, vol. 16, no. 6, pp. 1574–1581, 2005.

[32] J. Amoros Lopez, E. Izquierdo Verdiguier, L. Gomez Chova, J. Munoz Mari, J.Z.R. Barreiro, G. Camps-Valls and J. Calpe Maravilla, "Land Cover Classification of VHR Airborne Images for Citrus Grove Identification," *ISPRS journal of photogrammetry and remote sensing*, vol. 66, pp. 115–123, 2011.

[33] D. Tuia and G. Camps-Valls, "Urban Image Classification with Semisupervised Multiscale Cluster Kernels," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 4, no. 1, pp. 65–74, Mach 2011.

[34] L. Gomez-Chova, G. Camps-Valls, J. Munoz-Mari and J. calpe, "Semi-supervised Cloud Screening with Laplacian SVM," *IEEE*, pp. 1521–1524, 2007.

[35] M. J. Fernández-Getino García, J. L. Rojo-Álvarez, F. Alonso-Atienza, and M. Martínez-Ramón, "Support Vector Machines for Robust Channel Estimation in OFDM," *IEEE signal process. J.*, vol. 13, no. 7, pp. 397–400, 2006.

[36] W. Dingxue and F. Wenping, "A New Channel Estimation Method based on Pilot and MWLS-SVR," *IEEE computer society*, pp. 441–444, 2009.

[37] ——, "A New Channel Estimation Method based on Pilot-aided and Local Adaptive Least Squares Support Vector Regression in Software Radio OFDM System," *IEEE computer society*, pp. 349–352, 2009.

[38] S. Chen, S. Gunn and C.J. Harris , "Decision Feedback Equalizer Design using Support Vector Machines," *IEEE Proc. vision, image and signal processing* , vol. 147, no. 3, pp. 213–219, 2000.

[39] S. Chen, A.K. Sanmigan and L. Hanzo , "Support Vector Machine Multiuser Receiver for DS-CDMA Signals in Multipath Channels," *Neural networks* , vol. 12, no. 3, pp. 604–611, 2001.

[40] ——, "Adaptive Multiuser Receiver using a Support Vector Machine Technique," *IEEE Vehicular technology conference* , pp. 604–608, 2001.

[41] S. Chen and L. Hanzo , "Block-adaptive Kernel-based CDMA Multiuser Detector," *IEEE Conf. on communication* , pp. 682–686, 2002.

[42] R. Daniels and R.W. Heath, "Online Adaptive Modulation and Coding with Support Vector Machines," *IEEE European wireless conference*, pp. 718–724, 2010.

[43] J. Zhang and B. Li, "A New Modulation Identification Scheme for OFDM in Multipath Rayleigh Fading Channel," *IEEE Computer society*, pp. 793–796, 2008.

[44] H. Bingbing and L. Wang, "Modulation Identification for OFDM in Multipath Circumstance Based on Support Vector Machine," *IEEE ICCS conference*, pp. 1349–1353, 2008.

[45] H. Xiang, J. Bin, Z. Jingli, S. Yueguang and Z. Liu, "Direction of Arrival Estimation based on Smooth Support Vector Regression ," *2nd IEEE international conference on future computer and communication*, pp. 818–822, 2010.

[46] N. Martínez-Ramón, J. Luis Rojo-Alvarez, G. Camps-Valls and C.G. Christodoulou, "Kernel Antenna Array Processing," *IEEE transactions on antennas and propagation*, vol. 55, no. 3, pp. 642–650, 2007.

[47] V. Vapnik, "Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications and Control," *Wiley son's Inc.*, 1998.

[48] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," *MIT Press, Cambridge*, 1999.

[49] S. Haykin, "Adaptive Filter Theory," *4th edition, Prentice-Hall, Englewood Cliffs, NJ*, 2001.

[50] Steve R. Gunn, "Support Vector Machines for Classification and Regression," *Technical Report, University of Southampton*, 1998.

[51] M. Martínez-Ramon and C. Cristodoulou, "Support Vector Machines for Antenna Array Processing and Electromagnetics," *Morgan Claypool, USA*, 2006.

[52] D.P. Bersekas, "Constrained Optimization and Lagrange Multiplier Methods," *Academic Press*, 1982.

[53] A. Smola, B. Scholkopf, and K. R. Muller, "General Cost Functions for Support Vector Regression," *in Proceedings of the Ninth Australian Conference on Neural Networks, Brisbane, Australia*, pp. 79–83, 1998.

[54] Y. Zhu, C. Li, and Y. Zhang, "A Practical Parameters Selection Method for SVM," *Advances in Neural Networks, International Symposium on Neural Networks Dalian, Proceedings, Part I, China*, pp. 518–523, 2004.

[55] V. Kecman, "Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models," *The MIT Press, London, England*, 2001.

[56] H. Sari, G. Karam, and I. Jeanclaude, "Transmission Techniques for Digital Terrestrial TV Broadcasting," *IEEE Commun. Mag*, vol. 33, no. 2, pp. 100–109, 1995.

[57] Y.S. Cho, J. Kim, W.Y. Yang and C.G. Kang, "MIMO-OFDM Wireless Communications with Matlab," *John Wiley Sons*, pp. 121–131, 2010.

[58] E.K. Hlel, S. Cherif, F. Tlili and M. Siala, "Improved Estimation of Time Varying and Frequency Selective Channel for OFDM Systems," 12^{th} *IEEE conf. on electronics, circuits and systems* , pp. 1–4, 2005.

[59] 3rd Generation Partnership Project, "Technical Specification Group Radio Access Network; evolved Universal Terrestrial Radio Access (UTRA): Base Station (BS) radio transmission and reception," *TS 36.104, V8.7.0*, September 2009.

[60] S.V. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction," *John Wiley Sons, London, UK*, pp. 341–358, 2008.

[61] R. Steele and L. Hanzo, "Characterization of Mobile Radio Channels," *John Wiley Sons, UK*, pp. 91–185, 1999.

[62] J.G. Proakis, "Digital Communications," *Prentice Hall*, pp. 278–280, 2000.

[63] M. J. Fernández-Getino García, J. M. Páez-Borrallo, and S. Zazo, "DFT-based Channel Estimation in 2D-Pilot-symbol-aided OFDM Wireless Systems," *IEEE Vehicular Technology Conf.*, vol. 2, pp. 815–819, 2001.

[64] A. CHARRADA and A. SAMET, "Nonlinear Channel Estimation for OFDM System by Complex LS-SVM Under High Mobility Conditions," *International journal of wireless & mobile networks (IJWMN)*, vol. 3, no. 4, pp. 175–185, 2011.

[65] ——, "SIMO-OFDM Channel Estimation Based on Nonlinear Complex LS-SVM," *International journal of computer applications (IJ-CA)*, vol. 42, no. 3, pp. 1–8, 2012.

[66] M. Sliskovic, "Signal Processing Algorithm for OFDM Channel with Impulse Noise," *IEEE conf. on electronics, circuits and systems*, pp. 222–225, 2000.

[67] A. CHARRADA and A. SAMET, "Complex Support Vector Machine Regression for Robust Channel Estimation in LTE Downlink System," *International journal of computer networks & communications (IJCNC)*, vol. 4, no. 1, pp. 211–223, 2012.

[68] ——, "Support Vector Machines Regression for MIMO-OFDM Channel Estimation," *International journal of artificial intelligence (IJ-AI)*, vol. 1, no. 4, pp. 214–224, 2012.

[69] S. Galih, T. Adiono and A. Kumiawan, "Low Complexity MMSE Channel Estimation by Weight Matrix Elements Sampling for Downlink OFDMA Mobile WiMAX System," *International journal of computer science and network security*, vol. 10, no. 2, pp. 280–285, 2010.

[70] L. Nanping, Y. Yuan, X. Kewen, and Z. Zhiwei, "Study on Channel Estimation Technology in OFDM System," *IEEE Computer Society Conf.*, pp. 773–776, 2009.

[71] S. Coleri, M. Ergen and A. Puri, "Channel Estimation Techniques based on Pilot Arrangement in OFDM Systems," *IEEE Transaction on broadcasting*, vol. 48, no.3, pp. 223–229, 2002.

[72] A. CHARRADA and A. SAMET, "Estimation of Highly Selective Channels for OFDM System by Complex Least Squares Support Vector Machines," *International journal of electronics and communications (AËU)*, vol. 66, no. 8, pp. 687–692, 2012.

[73] J. L. Rojo-Álvarez, C. Figuera-Pozuelo, C. E. Martínez-Cruz, G. Camps-Valls, F. Alonso-Atienza, M. Martínez-Ramón, "Nonuniform Interpolation of Noisy Signals using Support Vector Machines," *IEEE Transaction on Signal process.*, vol. 55, no. 48, pp. 4116–4126, 2007.

[74] M. Martínez Ramón, N. Xu, and C. G. Christodoulou, "Beamforming using Support Vector Machines," *IEEE antennas and wireless propagation J.*, vol. 4, pp. 439–442, 2005.

[75] A. CHARRADA and A. SAMET, "Nonlinear Complex LS-SVM for Highly Selective OFDM Channel With Impulse Noise," 6^{th} *IEEE conference on sciences of electronics, technologies of information and telecommunications (SETIT)*, pp. 696–700, 2012.

[76] I. Steinwart and A. Christmann, "Support Vector Machines," *Springer Science, USA*, pp. 443–453, 2008.

[77] 3rd Generation Partnership Project, "Technical Specification Group Radio Access Network; evolved Universal Terrestrial Radio Access (UTRA): Physical Channels and Modulation layer," *TS 36.211, V8.8.0*, September 2009.

[78] ——, "Technical Specification Group Radio Access Network; Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)," *TR 25.814, V7.1.0*, September 2006.

[79] ——, "Technical Specification Group Radio Access Network; evolved Universal Terrestrial Radio Access (UTRA): Physical layer procedures," *TS 36.213, V8.8.0*, September 2009.

[80] C. Lim and D. Han, "Robust LS Channel Estimation with Phase Rotation for Single Frequency Network in OFDM," *IEEE Transactions on consumer electronics*, vol. 52, pp. 1173–1178, 2006.

[81] M. Ergen, "Mobile Broadband Including WiMAX and LTE," *Springer, Berkeley, USA*, pp. 271–307, 2009.

[82] F. Wan, W. Zhu, and M. Swamy, "A Semiblind Channel Estimation Approach for MIMO-OFDM Systems," *IEEE Transactions on signal processing*, vol. 56, no. 7, pp. 2821–2834, 2008.

[83] A. Kammoun, K. Abed-Meraim and S. Affes, "Regularized Semi-blind Estimator over MIMO-OFDM Systems," *IEEE computer society*, pp. 189–194, 2009.

[84] R. H. C. Shin and E. Powers, "Blind Channel Estimation for MIMO-OFDM Systems," *IEEE Transactions on vehicular technology*, vol. 56, no. 2, pp. 670–685, 2007.

[85] S. M. S.J. Nawaz and A. Ikram, "Neural Network Based MIMO-OFDM Channel Equalizer Using Comb-Type Pilot Arrangement," *IEEE Computer society*, pp. 36–41, 2009.

[86] C. hua and Z. Xiao-hui, "MIMO-OFDM Channel Estimation Based on PCA," *IEEE Conference on Computer, Mechatronics, Control and Electronic Engineering*, pp. 300–303, 2010.

[87] K. Sarma and A. Mitra, "ANN Based Rayleigh Multipath Fading Channel Estimation of a MIMO-OFDM System," *IEEE computer society*, 2009.

[88] C. Hua and Z. Xiao-hui, "MIMO-OFDM Channel Estimation Based on Neural Network," *IEEE computer society*, 2009.

[89] S. Yun and C. Caramanis, "Multiclass Support Vector Machines for Adaptation in MIMO-OFDM Wireless Systems," 47^{th} *IEEE Annual Allerton Conference*, pp. 1145–1152, 2009.

[90] I. Khan and P.S. Hall, "Experimental Evaluation of MIMO Capacity and Correlation for Narrowband Body-centric Wireless Channels," *IEEE Transaction on antennas and propagation* , vol. 58, pp. 195–202, 2010.

[91] W.H. Chin, "Exploiting Channel Time Selectivity in Pilot-aided Alamouti STBC Systems," *IEEE Vehicular Technology Conf.*, vol. 11, pp. 11–16, 2003.

[92] W. Shaopeng, Z. Shihua and Z. Guomei, "A Walsh-Hadamard Coded Spectral Efficient Full Frequency Diversity in OFDM System," *IEEE Transaction on communications* , vol. 58, pp. 28–34, 2010.

[93] Y.C. Lung and W.J. Chappell, "Angular Diversity Prominent for Compact Devices of Wireless Sensor Network in Indoor Environments," *Antennas and propagation society international symposium* , vol. 11, pp. 11–17, 2007.

[94] S.M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications," *IEEE journal on selected areas in communications*, vol. 16, pp. 1451–1458, 1998.

[95] G.J. Foschini and M.J. Gans, "On Limits of Wireless Communications in a Fading Environments when using Multiple Antennas," *Wireless personal communications*, vol. 6, no.3, pp. 311–335, 1998.

[96] ——, "Spatio-temporal Coding for Wireless Communication," *IEEE transaction on communications*, vol. 46, pp. 357–366, 1998.

[97] J. Salz, J.H. Winters and R.D. Gitlin, "The Impact of Antenna Diversity on the Capacity of Wireless Communication Systems," *IEEE transaction on communications*, vol. 42, pp. 1740–1751, 2002.

[98] A. Goldsmith, "Wireless Communications," *Cambridge University Press, Cambridge, UK.*, pp. 91–108, 2005.

[99] T.T. Fujita and Al., "A New Signal Detection Scheme Combining ZF and k-best Algorithms for OFDM/SDM," 15^{th} *PIMRC*, vol. 11, pp. 11–21, 2004.

[100] P.W. Wolniansky, G.J. Foschini, G.D. Golden and R.A. Valenzuela, "V-BLAST: An Architecture for Realizing Very High Data Rates over the Rich-scattering Wireless Channel," *International symposium on signal systems and electronics*, vol. 11, pp. 11–18, 1998.

[101] V.D. Blondel and A. Megretski, "Unsolved Problems in Mathematical Systems and Control Theory," *Princeton university press, Princeton, USA*, pp. 237–244, 2004.

[102] A.H. Ballard, "Performance Analysis of ZF and MMSE Equalizers for MIMO Systems: An In-depth Study of the High SNR Regime," *IEEE Transaction on information theory*, vol. 57, pp. 2008–2026, 2011.

[103] F. Khan, "LTE for 4G Mobile Broadband: Air Interface Technologies and Performance," *Cambridge university press* , pp. 110–140, 2009.

[104] R. Minkara and J.P Dubois, "Improved Root-mean-square-gain-combining for SIMO Channels," *Word academy of science, engineering and technology*, vol. 41, pp. 562–565, 2008.

[105] 3rd Generation Partnership Project, "Technical Specification Group Radio Access Network; evolved Universal Terrestrial Radio Access (UTRA): Physical channels and modulation," *TS 36.211, V0.3.1*, February 2007.

[106] L. Somasegaran, "Channel Estimation and Prediction in UMTS LTE," *Aalborg University*, pp. 16–17, 2007.