

**République Algérienne Démocratique et Populaire**  
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique  
Université de Batna  
Faculté des Sciences de l'Ingénieur  
Département d'Electronique

**MEMOIRE**  
Présenté pour l'obtention du Diplôme de  
**MAGISTER EN ELECTRONIQUE**

Option : **Micro-Ondes pour Systèmes de Télécommunication**

Par

**Hassen BOUZGOU**  
Ingénieur d'Etat en Electronique  
*Option : Communication*

Thème

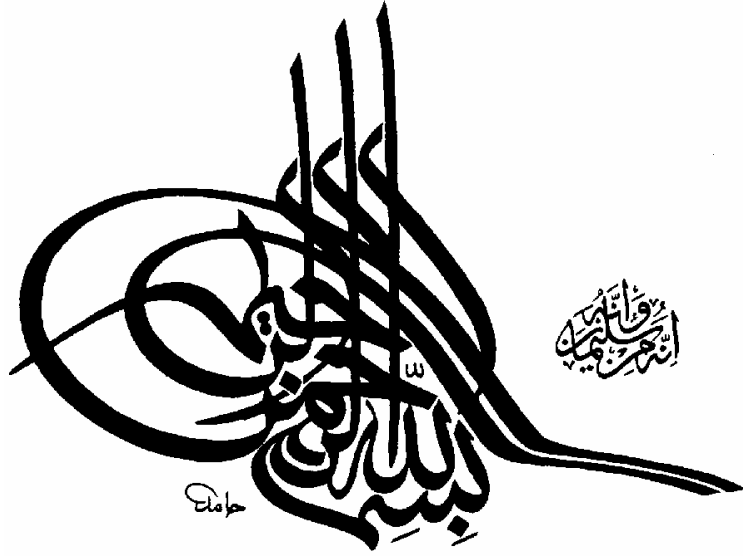
**DEVELOPMENT OF MULTIPLE REGRESSION  
SYSTEMS FOR HYPERDIMENSIONAL SPECTRAL  
SPACES**

Soutenu le :06/09/ 2006  
Devant le jury composé de:

Djamel BENATIA	Professeur	U. de Batna	Président
Nabil BENOUDJIT	Maître de Conférence	U. de Batna	Rapporteur
Farid MELGANI	Professeur	U. de Trento (Italie)	Co-rapporteur
Nourreddine GOLEA	Maître de Conférence	C. U. Oum El Bouaghi	Examineur
Lamir SAIDI	Chargé de Cours	U. de Batna	Examineur
Djemai ARAR	Chargé de Cours	U. de Batna	Examineur

*Ce travail de recherche rentre dans le cadre d'une collaboration entre  
l'Université de Batna (Algérie) et l'Université de Trento (Italie)*

Septembre 2006



# TABLE OF CONTENT

---

## CHAPTER 1: INTRODUCTION AND THESIS OVERVIEW

---

1.1 GENERAL INTRODUCTION.....	2
1.2 HIGH-DIMENSIONAL DATA .....	3
1.2.1 The Curse of Dimensionality .....	4
1.2.2 Some Geometrical Properties of High-Dimensional Spaces .....	5
1.3 INVERSE PROBLEM .....	7
1.4 SENSORS .....	8
1.4.1 Definition of a Transducer and a Sensor.....	8
1.4.2 Frequency Bands .....	8
1.4.3 Passive Sensors .....	9
1.4.4 Active Sensors.....	10
1.5 REGRESSION ANALYSIS.....	11
1.5.1 Definition of Regression .....	11
1.5.1 Parametric and Non-Parametric Regression .....	11
1.5.2 Regression in Hyperdimensional Spaces .....	12
1.6 OBJECTIVES OF THIS WORK .....	14
1.7 OVERVIEW OF DISSERTATION.....	14

---

## CHAPTER 1: THE PROPOSED APPROACH MULTIPLE ESTIMATOR SYSTEMS

---

2.1 CONCEPT OF MULTIPLE SYSTEMS .....	16
2.1.1 Multiple Classifiers .....	16
2.1.1 Multiple Estimators .....	17
2.2 THE PROPOSED APPROACH.....	18
2.2.1 Features Selection Bloc .....	20
2.2.1.1 Unsupervised grouping by sampling (UGS).....	20
2.2.1.2 Unsupervised adjacent grouping (UAG).....	20
2.2.2 Regression Bloc.....	21
2.2.2.1 Multiple linear regression (MLR) .....	21
<i>a) The coefficient of Multiple Determination.....</i>	22
<i>b) Collinearity.....</i>	23
<i>c) Features Selection.....</i>	23

2.2.2.2 Radial Basis Functions (RBFs).....	24
<i>a) Purpose</i> .....	24
<i>b) Architecture</i> .....	25
<i>c) Working Principles</i> .....	26
<i>d) Training the network</i> .....	27
2.2.3 Fusion Block .....	28
2.2.3.1 Combination-based Approach.....	28
<i>a) Average Combination Strategy (ACS)</i> .....	28
<i>b) Weighted Combination Strategy (WCS)</i> .....	28
2.2.3.2 Non-Linear Fusion Strategy (NLFS).....	29
2.2.3.3 Hybrid Fusion Strategy (HFS).....	29
2.2.3.4 Classification Based Strategy .....	30
<i>a) Selection Strategy (SS)</i> .....	30
<i>b) Dynamic Strategy (DS)</i> .....	32
2.2 CONCLUSION .....	33

---

## CHAPTER 3: SUPPORT VECTOR MACHINES

---

3.1 INTRODUCTION .....	35
3.1.1 Historic Background.....	36
3.1.2 Motivation .....	37
3.1.3 Machine Learning .....	37
3.1.3.1 Supervised learning.....	38
3.1.3.2 Unsupervised learning .....	38
3.1.4 Choosing an Hypothesis .....	39
3.1.5 Statistical Learning Theory .....	40
3.2 SUPPORT VECTOR CLASSIFICATION .....	41
3.2.1 Maximal Margin Hyperplanes.....	42
3.2.2 Kernel-Induced Feature Spaces .....	44
3.2.3 Non-Separable Data .....	46
3.3 SUPPORT VECTOR REGRESSION .....	49
3.3.1 Linear Regression.....	50
3.3.1.1 $\epsilon$ -insensitive loss function .....	50
3.3.1.2 Quadratic loss function.....	51
3.3.1.3 Huber loss function .....	52
3.3.2 Non Linear Regression.....	53
3.3.3 COMMENTS AND CONCLUSION .....	54

---

---

## CHAPTER 4: EXPERIMENTAL RESULTS

---

---

4.1 APPLICATION TO CHEMOMETRICS .....	56
4.1.1 Chemometrics and quality .....	57
4.1.2 Beer-Lambert's law .....	59
4.2 MODEL SELECTION .....	60
4.2.1 Separate training, validation and test sets .....	60
4.2.1.1 Training set .....	60
4.2.1.2 Validation set .....	61
4.2.1.3 Test set .....	61
4.3 PERFORMANCE EVALUATION CRITERIA .....	62
4.3.1 Normalized Mean Square Error .....	62
4.3.2 Computational Time .....	62
4.4 REAL-LIFE EXAMPLES .....	63
4.4.1 Wine dataset .....	63
4.4.2 Orange juice dataset .....	64
4.5 GRAPHICAL DETECTION OF OUTLIERS .....	65
4.5.1 Wine dataset .....	65
4.5.2 Orange juice dataset .....	67
4.6 RESULTS OF TRADITIONAL APPROACHES .....	69
4.6.1 Principal Component Regression .....	69
4.6.1.1 Method presentation .....	69
4.6.1.2 Datasets Results .....	71
4.6.2 Partial Least Square Regression (PLSR) .....	74
4.6.2.1 Method presentation .....	74
4.6.2.2 Datasets Results .....	76
4.6.3 Sequential forward Selection with mutual information (SFS) .....	77
4.6.3.1 Mutual information .....	77
4.6.3.2 Sequential Forward Selection .....	79
4.6.3.3 Datasets results .....	80
4.6.4 Multiple Estimator System (MES) .....	81
4.6.4.1 Multiple linear regression based .....	81
4.6.4.2 RBF regression based .....	86
4.6.5 SVM Results .....	90
4.7 SUMMARY OF BEST RESULTS .....	92
4.8. CONCLUSION .....	94

---

---

## CHAPTER 5: CONCLUSION AND PERSPECTIVES

---

---

5.1 CONTRIBUTION OF THIS DISSERTATION .....	96
5.2 PERSPECTIVES AND FUTUR WORK .....	97

### **BIBLIOGRAPHY**

## ABSTRACT

The recent advent of new generations of sensors for different application fields such those related to remote sensing and spectroscopy has shown a great interest for their improved spectral, spatial and/or temporal characteristics. In particular, hyperspectral sensors allow an accurate spectral analysis of a physical phenomenon under investigation since they provide a large number of observations (features), each coming from a very narrow spectral band. However, the automatic analysis of data acquired with such sensors is somewhat challenging since it should be carried out in hyperdimensional spectral spaces.

In the regression context where it is often desired to find a continuous relationship between the features and one or more parameters of the investigated physical phenomenon, the huge size of the feature space involves the so-called curse of dimensionality. This latter is due to the unbalancing between the number of features and the number of samples required to train the regression method.

The classical approach adopted in the literature to deal with this issue consists in reducing the hyperdimensional feature space into a subspace of smaller dimension where the curse of dimensionality disappears. The most common way to do this task is represented by applying a feature selection process which consists in selecting the most significant subset of features for the considered regression problem. This allows to define a subspace of features of reduced dimension where the risk of affecting negatively (curse of dimensionality) the adopted regression method becomes negligible. However, whatever the degree of sophistication of the feature selection technique, one can expect a loss of information when reducing the size of the original hyperdimensional feature space with a consequent negative impact on the accuracy of the regression method.

In this thesis, it is proposed to exploit the whole information available in the original hyperdimensional feature space by means of the fusion (combination) of multiple regression methods. The development of the proposed multiple regression systems will include three main steps. The first one is related to a proper partition of the original hyperdimensional feature space into subspaces of reduced dimensionality. The second step consists in training in each of the subspaces obtained in the previous step a regression method. For this purpose, it will be made use of neural regression methods

which have proved to be effective and sufficiently robust in numerous application fields. Finally, in the third and final step, the results provided by the different regression methods will be combined in order to produce a global estimate of the physical parameter of interest with an expected higher accuracy with respect to what can be achieved by the classical regression approach based on feature selection.

The proposed methodologies are general and can be applied to any of the application fields of hyperspectral sensors. In the experimental phase, the validation of the methodologies will be carried out on data acquired by near-infrared spectrometers for quantitative chemical analysis.

**Keywords:** Hyperspectral Sensors; Regression; Curse of Dimensionality; Feature Selection; Data Fusion; Neural Networks; Support Vector Machine; Pattern Recognition.



## RESUME

L'avènement récent d'une nouvelle génération de capteurs pour différents champs d'applications tels que ceux liés à la télédétection et à la spectroscopie a montré un grand intérêt pour leurs meilleures caractéristiques spectrales, spatiales et/ou temporelles. En particulier, les capteurs hyperspectraux qui permettent une analyse précise des phénomènes de recherche, puisqu'ils fournissent un grand nombre d'observations, chacune venant d'une bande spectrale très étroite. Cependant, l'analyse automatique des données acquises avec de tels capteurs est en quelque sorte défiante, puisqu'elle devrait être effectuée dans des espaces spectraux hyper-dimensionnels.

Dans le contexte de régression où on a souvent le désir de trouver une relation continue entre les observations et un ou plusieurs paramètres du phénomène physique étudié, la taille énorme de l'espace d'observations implique la prétendue "malédiction" de la dimensionnalité. Cette dernière est due au déséquilibre entre le nombre d'observations et le nombre d'échantillons exigés pour exécuter la méthode de régression.

L'approche classique adoptée dans la littérature pour traiter ce problème, consiste à réduire l'espace d'observations hyper-dimensionnelles à des sous-espaces de petite dimension, où la malédiction de la dimensionnalité disparaît. La façon la plus utilisée de faire cette tâche est d'appliquer un procédé de sélection de variables qui consiste à choisir le sous-ensemble d'observations le plus significatif pour le problème de régression considéré. Ceci laisse définir un sous-espace d'observations de dimension réduite où le risque d'affecter négativement (malédiction de la dimensionnalité) la méthode de régression adoptée devient négligeable. Cependant, quelque soit le degré d'efficacité de la technique de sélection, on peut s'attendre à une perte d'information en réduisant la taille de l'espace original d'observations hyper-dimensionnelles avec un impact négatif conséquent sur l'exactitude de la méthode de régression.

Dans ce mémoire, on propose d'exploiter l'information complète disponible dans l'espace de variables hyper-dimensionnelles par le biais de la fusion (combinaison) des méthodes de régression multiple. Le développement des systèmes proposés inclura trois étapes principales. La première concerne la répartition de l'espace original des observations hyper-dimensionnelles à des sous-espaces de dimensionnalité réduite. La deuxième étape consiste à appliquer à chacun des sous-espaces obtenus en étape précédente une méthode de régression. À ce propos, on utilise une des méthodes

neuronales de régression qui se sont avérées efficaces et suffisamment robustes dans de nombreux domaines d'application. Dans la troisième et finale étape, les résultats fournis par les différentes méthodes de régression seront combinés afin de produire une évaluation globale du paramètre physique d'intérêt avec une exactitude prévue plus élevée en ce qui concerne ce qui peut être réalisé par l'approche classique de régression basée sur la sélection des variables.

Les méthodologies proposées sont générales et peuvent être appliquées à n'importe quel champ d'application des capteurs hyper-spectraux. Dans la phase expérimentale, la validation des méthodologies sera effectuée sur des données acquises par les spectromètres proche-infrarouges pour l'analyse quantitative.

**Mots-Clés:** capteurs hyper-spectraux; régression; problème de dimensionnalité; sélection des variables; fusion des données; réseaux de neurones; reconnaissance des formes.