DOTTORATO DI RICERCA
in
SCIENZE COMPUTAZIONALI ED INFORMATICHE
Ciclo XVIII
Consorzio tra Università di Catania, Università di Napoli Federico II,
Seconda Università degli studi di Napoli, Università di Palermo, Università di Salerno
SEDE AMMINISTRATIVA: UNIVERSITÀ DI NAPOLI FEDERICO II

AUTORE
LUISA CUTILLO

CLASSIFICATION, MULTIPLE HYPOTHESIS TESTING AND WAVELET THRESHOLDING PROCEDURES WITH APPLICATIONS

TESI DI DOTTORATO DI RICERCA

UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

Autore: **Luisa Cutillo**

Titolo: **Classification, Multiple Hypothesis Testing and Wavelet Thresholding procedures with Applications**

Dipartimento: **Matematica e Applicazioni ’R.Caccioppoli’**

Tesi di **Dottorato** Anno: **2005**

Firma dell’autore

To Luigi and Angela

Table of Contents

Table of Contents v

Acknowledgements ix

Introduzione 1

Introduction 5

1 Classification Theory 9

1.1 General framework 10
1.2 Statistical Decision theory 12
1.3 Parametric discriminant analysis 15
1.3.1 Estimation of the parameters 16
1.3.2 Linear discriminant analysis 17
1.3.3 Quadratic discriminant Analysis 18
1.4 Non parametric discriminant analysis 19
1.4.1 Principal component discriminant analysis 24
1.4.2 Independent components discriminant analysis 26
1.5 Local Discriminant methods for image classification 28
1.6 Notations and Assumptions 31
1.7 Local priors 34
1.7.1 Local voting priors 35
1.7.2 Local frequency priors 36
1.7.3 Local integrated priors 36
1.7.4 Local nested priors 37
1.8 Asymptotics 38
1.8.1 Local voting priors 38

1.8.2 Local frequency priors 39
1.8.3 Local integrated priors 39
1.8.4 Local nested priors 40
1.8.5 Iterations 40

2 Multiple Hypothesis Testing 41
2.1 General Framework 42
2.2 Type I Error Rates 44
2.2.1 Strong and Weak control 46
2.3 Error Controlling Procedures 47
2.3.1 $FWER$ and FDR Controlling Procedures 48
2.4 Bootstrap estimation of the null distribution 49
2.5 Bayesian testing 50
2.6 Bayesian multiple hypothesis testing 52
2.6.1 MAP multple testing procedure 53
2.6.2 Connections with frequentist procedures and choice of the priors . 55
2.6.3 Custom normal model 56
2.6.4 Possible extension and future work 58

3 Wavelet Filtering of Noisy Signals 59
3.1 Mathematical background 60
3.2 Advantages of wavelets 65
3.3 The statistical problem 67
3.4 Bayes rules and wavelets 69
3.5 Larger posterior mode (LPM) wavelet thresholding 71
3.5.1 Derivation of the thresholding rule 72
3.5.2 Exact risk properties of LPM rules 76
3.6 Generalizations 77
3.6.1 Model 1: exponential prior on unknown variance. 78
3.6.2 Model 2: inverse gamma prior on unknown variance. 80
3.7 Conclusions and future works 81

4 Numerical Experiments for Discriminant Analysis 83
4.1 Multispectral cloud detection: general problem 83
4.1.1 Classification methods considered 85
4.1.2 Case studies 86
4.1.3 Results 87

4.2 Simulation of Local discriminant methods 88
4.2.1 Real data 91
4.2.2 Conclusions 92

5 Application of multiple testing procedures to DNA microarray data 105
5.1 What is a microarray ? 106
5.2 Experimental Design 108
5.3 Expression Ratio 109
5.4 Normalization 110
5.4.1 Single slide data displays 110
5.4.2 Within slide normalization 111
5.4.3 Global normalization 111
5.4.4 Intensity dependent normalization 112
5.4.5 Within print tip group normalization 113
5.4.6 Scale normalization 113
5.4.7 Multiple slide normalization 114
5.5 Yeast experiment 114
5.5.1 Experiment description 114
5.5.2 Methods 116

6 LPM: simulations, comparisons and real life example 119
6.1 Selection of hyperparameters 119
6.2 Simulations and comparisons 126
6.3 An example in atomic force microscopy 129

A 133
A.1 Proof of Lemma (3.6.1) 133
A.2 Proof of lemma (3.6.2) 135
A.3 Selection of hyperparameters α and β in Model 2. 138

Bibliography 139

Acknowledgements

I would like to thank Umberto Amato, my supervisor, for his many suggestions and constant support during this research. I am also thankful to Claudia Angelini for her guidance through the early years of confusion.

I am gratefull to Professor Brani Vidakovic, who expressed his interest in my work and supplied me with some of his recent works. Yoon Young shared with me her knowledge of statistical analysis and provided many useful references and friendly encouragement. I should also mention the support from University "Frederico II" that, under the program *mobilit di breve durata*, gave me the opportunity of visiting the Georgia Institute of Thecnology, Atlanta.

Of course, I am grateful to my husband Luigi for his patience and *love*. Without him this work would never have come into existence (literally).

Finally, I wish to thank the following friends: Annamaria for her sincere friendship, Francesca and Italia for bearing me in their work place, Monica for her cakes, Alba for her encouragements, Angela, Claudia, Mariarosaria and Woulla for having lunch with me everyday and my sister Angela.... because she asked me to.

Napoli, Italia Luisa Cutillo
November 30, 2005

Introduzione

Lo sviluppo delle metodologie statitistiche per l'analisi dati è generalmente collegato a progressi ottenuti in altri campi scientifici. Da un lato l'analisi statistica è spesso indirizzata a problemi reali, di conseguenza, il miglioramento delle metodologie nasce dall'esigenza di fornire una soluzione sempre più accurata ed efficiente a problemi specifici. D'altro canto accade anche che le procedure statistiche siano prima esplorate in ambito teorico e successivamente testate prima in simulazione e quindi su dati reali. In quest'ottica, lo scopo di questo lavoro è quello di mostrare sia come problemi reali possano essere efficientemente risolti mediante tecniche statistiche, sia come modelli statistici teorici possano essere adatti a descrivere problemi reali.

Nella pratica spesso ci troviamo a dover analizzare grandi moli di dati con molte dimensioni. Conseguentemente siamo costretti ad affrontare il problema della dimensionalità. Esistono differenti approcci statistici per fronteggiare questa difficoltà. Ad esempio, data una immagine satellitare dell'Europa ad una risoluzione di 800x600 pixel, consideriamo un insieme di dati costituito dalle radianze, misurate su 15 canali, associate ad ogni pixel. Supponiamo lo scopo sia quello di classificare ciascun pixel come appartenete ad una di due classi predefinite, come ad esempio nuvoloso e sereno. Prendendo a modello il processo cognitivo della nostra mente, abbiamo bisogno di estrarre le informazioni dai dati, cioè dobbiamo poter individuare strutture dati significative e di piccole dimensioni, nello spazio delle ossevazioni. Esiste un'ampia classe di tecniche statistiche mediante le quali il problema della dimensionalità può essere gestito, come ad esempio l'analisi delle componenti principali utilizzate in combinazione con i metodi kernel, o l'analisi delle componenti indipendenti. Illustreremo la teoria statistica della classificazione supervisionata e discuteremo alcuni aspetti riguardanti la classificazione localizzata di immagini.

Un altro esempio di dati di grandi dimensioni proviene dal recente e interessante avvento della tecnologia dei microarray. Negli ultimi anni i DNA microarray sono diventati uno strumento base per la ricerca biologica. Il diffondersi di questa tecnologia ha potenziato la ricerca nell'ambito della genomica funzionale, consentendo il monitoraggio dei profili di espressione di migliaia di geni (anche dell'intero genoma) contemporaneamente. La grande mole di dati generati da questo tipo di esperimenti ha permesso lo sviluppo di nuove interessanti metodologie statistiche. Conseguentemente l'analisi di dati da DNA microarray costituisce un'applicazione Biostatistica e Bioinformatica di crescente interesse. Oggetto della nostra analisi sarà lo sviluppo di una tecnica per l'individuazione dei pochi geni differenzialmente espressi in un particolare contesto sperimentale. Il problema verrà formulato in termini di test di ipotesi multipla e verranno anche illustrate tutte le fasi dell'analisi dei dati da DNA microarray.

Come ultimo esempio, consideriamo un esperimento in cui lo scopo è analizzare un segnale digitale proveniente da una strumentazione elettronica a partire da migliaia di misurazioni empiriche. Anche in condizioni sperimentali ottimali, le misurazioni di cui disponiamo saranno affette da errore. L'analisi di tali segnali è riconducibile al problema di ricostruzione di un segnale contaminato da rumore. Questo probelma è noto sotto diversi nomi (*denoising*, *filtering*, *smoothing*, *regression* etc.) a seconda del campo scientifico in cui è affrontato. In letteratura sono state proposte differenti soluzioni mediante splines, funzioni kerel, serie di Fourier e wavelet. In questa sede affronteremo il problema nell'ottica della regressione non parametrica e presenteremo come soluzione alcune regole di wavelet thresholding. La scelta di utilizzare la teoria delle wavelet deriva principalmente dalla possibilità di ottenere una ricostruzione ottimale del segnale originale anche nel caso in cui quest'ultimo sia fortemente irregolare. Questo risultato non può essere ottenuto mediante nessun altro stimatore lineare e deriva dalla proprietà delle basi wavelet di approssimare un vasto insieme di spazi funzionali.

La tesi è organizzata come segue. Nel Capitolo 1 viene affrontato il problema della classificazione supervisionata con lo scopo di risolvere il problema della classificazione di immagini. Vengono passati in rassegna alcuni metodi standard ed in particolare è descritto il problema della classificazione di immagini mediante tecniche locali. I risultati dell'applicazione delle metodologie proposte a dati reali e simulati verranno poi presentati nel Capitolo 4. Nel Capitolo 2 viene introdotto il problema dei test di ipotesi multipla con l'obiettivo di fornire uno strumento di analisi di dati da cDNA microarray. Viene fornita una prospettiva critica dell'impostazione Bayesiana e frequentista del problema e sono descritti punti di forza, di debolezza e di contatto tra le due filosofie. L'applicazione a dati reali da cDNA microarray delle metodologie discusse sarà presentata nel Capitolo 6. Nel Capitolo 3 sono analizzate nel dominio wavelet alcune regole di thresholding indotte da una variazione del principio bayesiano del *Maximum A Posteriori* (MAP). Le regole MAP sono azioni Bayesiane che massimizzano la probabilità a posteriori. La metodologia proposta risulta essere di tipo thersholding ed è caratterizzata dalla proprietà di selezionare la moda della probabilità a posteriori che risulta essere più grande in valore assoluto, da cui il nome *Larger Posterior Mode* (LPM). Forniamo un'analisi del rischio associato alla regola LPM e mostriamo come le sue prestazioni della regola LPM sono competitive con quelle di tecniche di letteratura. Il Capitolo 6 presenta infine una discussone sulla scelta degli iperparametri, uno studio in simulazione della rregola LPM ed una sua applicazione ad un problema reale.

Questo lavoro è stato svolto durante la mia attività di ricerca presso l'Istituto per le Applicazioni del Calcolo Mauro Picone (IAC) , sezione di Napoli. L'interesse all'analisi dei dati da DNA microarray è nato da una collaborazione con il Telethon Institute of Genetic and Medicine (TIGEM) e con il Policlinico di Napoli, dove sono stati fisicamente effettuati gli esperimenti sui DNA microarray .

La parte finale della tesi è stata svolta durante il mio periodo di ricerca presso il Georgia Institute of Technology, Atlanta, Georgia.

Introduction

Development in the field of statistical data analysis is often related to advancements in other fields to which statistical methods are fruitfully applied. In fact statistical analysis is often addressed to real problems and methodological improvements are consequently motivated by the search for the solution of a specific problem. The other way round, sometimes statistical concepts are first theoretically investigated and then applied to simulated or real data for the development of new techniques. The aim of this work is to show how different real world problems can be solved efficiently by statistical techniques, and simultaneously to show how theoretical statistical models can fit real data problems.

In real world problems we frequently face with large sets of high-dimensional data, and as a consequence, with the problem of dimensionality. This problem can be approached in different ways. As an example, consider a satellite image of Europe made of 800 x 600 pixel, and suppose we have radiance measures from 15 channels associated to each pixel. Suppose our purpose is to classify each pixel as coming from two different predefined classes, e.g. cloudy or non cloudy. As the human brain does in everyday perception, we need then to find meaningful low-dimensional structures hidden in the high-dimensional observation space. There is a wide class of statistical techniques, by which this problem can be handled, as principal component analysis, in combination with Kernel methods, or independent components discriminant analysis. We will illustrate the statistical theory of supervised classification and discuss some features regarding localized classification of images.

Another example of very fashionable high dimensional dataset is microarray data. In a few years, DNA microarray technology has become a basic tool in biological research. The growth of this technology has empowered researchers in functional genomics to monitor gene expression profiles, thousands of genes (even the entire genome) at a time. As a consequence, the large volume of data generated by these experiments has created an opportunity for some very interesting statistical works. For this reasons DNA microarray data analysis is one of the fastest growing area of applications in Biostatistics and Bioinformatics. We will focus on the problem of finding differentially expressed genes, formulating it in terms of multiple hypothesis testing. We will illustrate the statistical issues involved at the various stages of the analysis on real datasets from DNA microarray experiments.

As last example of high dimensional real dataset suppose we have thousands of empirical measurements of a signal. Even in the best experimental conditions the measurements will be contaminated by noise, nevertheless the aim is to recover the underlaying unknown signal. This problem is known under different names (*denoising*, *filtering*, *smoothing*, *regression* etc.) according to the scientific field where it is formulated. Different solutions have been formulated in terms of spline smoothing, kernel estimation, Fourier or wavelet expansion. We will state the problem in the context of non-parametric regression and will discuss solutions provided by wavelet thresholding rules. It can be shown that when the underlaying signal is regular and spatially homogeneous, all these methods are asymptotically equivalent but, for an irregular non homogeneous signal, the wavelet non linear estimation is asymptotically optimal and similar results cannot be achieved by any other linear method. This happens because wavelet basis can characterize a wide range of functional spaces.

The present thesis is organized as follows. In Chapter 1 we deal with the problem of supervised classification having in mind the problem of image classification. We review some of the classical statistical methods for pattern recognition, introduce the problem of localized classification of images and propose new localized discriminant analysis methods. Applications of the proposed methodology to simulated and real data, will be provided in Chapter 4. In Chapter 2 we introduce the statistical problem of multiple hypothesis testing with the target of analyzing cDNA microarray data. We review the guiding lines of frequentist and Bayesian approach to multiple hypothesis testing, describing strength and weakness of the two philosophies and trying to find some connections between them. The application of the described methods to a genetic microarray data experiment is provided in Chapter 6. In Chapter 3 we explore the thresholding rules in the wavelet domain induced by a variation of the Bayesian *Maximum A Posteriori* (MAP) principle. The MAP rules are Bayes actions that maximize the posterior. The proposed rule is thresholding and always picks the mode of the posterior larger in absolute value, thus the name *Larger Posterior Mode* (LPM). We show that the introduced shrinkage performs comparably to several popular shrinkage techniques. The exact risk properties of the thresholding rule are explored. Comprehensive simulations and comparisons are provided in Chapter 6 which also contains discussion on the selection of hyperparameters and a real-life application of the introduced shrinkage.

The present work was done during my research activity at the Istituto per le Applicazioni del Calcolo ”Mauro Picone” (IAC) in Naples. The interest on microarrays data was motivated by a collaboration with the Telethon Institute of Genetic and Medicine (TIGEM) and the Policlinico of Naples, where the biological experiments were carried out.
The last part of this work was done during a visiting period at the Georgia Institute of Technology (GATECH), in Atlanta, U.S.A.

Chapter 1

Classification Theory

Introduction

In this Chapter we deal with the problem of supervised classification. We review some of the classical statistical methods for pattern recognition, introduce the problem of local classification of images and propose new local discriminant methods. Application of the proposed methodology to simulated and real data, along with suggestions for future work, would be provided in Chapter 4. Some of the results showed in this Chapter were presented at the *IEEE Gold* conference (Naples, 2004) and at the *CLADAG* meeting (Parma, 2005).

The Chapter is organized as follows. The first two sections are a brief introduction to the statistical problem of pattern classification. Sections 3 and 4 describe respectively some parametric and non parametric approach to supervised classification. Sections 5 and 6 are devoted to the problem of local discriminant analysis and proposals for new local discriminant methods are discussed in Sections 7 and 8.

1.1 General framework

Building pattern recognition systems would be very useful in solving myriad of nowadays problems like fingerprint identification, speech recognition and DNA sequence identification. It is amazing to think that humans are used to classify data received from senses quite immediately and unconsciously. For example most of humans can recognize shapes by touching, foods by tasting, faces by watching, can detect a specific illness or identify different types of car. Of course it is crucial for science progress to automatize the human decision making process so to perform some of these tasks faster, more cheaply or accurately. One characteristic of human pattern recognition is that it is learnt but learning involves a teacher. If we try different unlabelled cups of tea we could discover that there are different groupings and that one group has a green color, but again we need a teacher to tell us that the common factor is that they were made by the same tea leaves. When the target of pattern recognition is the discovering of new groupings, it is called unsupervised. Otherwise, learning from a given set of labelled examples, the training set, in order to classify future examples is called supervised pattern recognition. We will be only concerned with supervised pattern recognition. We will assume we are given a finite set of classes and that a teacher can tell us the correct class label for each pattern in a training set. We could imagine a pattern recognition system like a machine, called classifier, that takes in input some measurements of the data, known as features, and tells in output whether the example is from one of the known classes or not. In statistical pattern recognition, there isn't any assumption about the structure of the classifier but it is learnt from data. The training set is regarded as a sample from a population of possible examples and it is used to make statistical inference for each class. The traditional model for the feature pdf from each class can be parametric, non-parametric or semi-parametric.

In the parametric approach, a general formula for the probability distribution of observation vectors for each class is assigned. The free parameters contained in the formula are estimated by the classifier during the learning stage. For example it can be specified that the observation vectors in each class follow a multivariate normal distribution with a common covariance matrix, and the class means and covariance matrix are estimated from the traing set. As we will see in the next Section this is a classical pattern recognition technique known as linear discriminant analysis (Johnson and Wichern, 1998).

The non-parametric approach does not require any assumption on the formula of probability distribution in advance (Hollander and Wolfe, 1999). There are several types of non-parametric methods and in particular Section 1.4 will focus on the procedures for estimating the conditional pdf from sample patterns. Other approaches consist in procedures for directly estimating the class each feature vector belongs to, bypassing probability estimation, like the nearest neighbor approach.

Recently, there has been interest in what might be called semi-parametric methods (Ripley, 1996). These methods are in between parametric methods, in which the underlying probability distributions are completely specified, and non-parametric methods in which they are completely free. Examples of such a method are neural networks which are characterized from a large number of parameters which can be optimized to fit different possible input configurations.

The three approaches have their own advantages and disadvantages, and each one is most appropriate in its own set of circumstances. Parametric approaches work best when it is possible to specify an accurate formula for the input distributions. However, some parametric approaches may still work well even if the parametric model only approximately fits the true distribution. Such approaches are said to be robust (Huber, 1986). Non-parametric methods have the advantage of not requiring a model to be specified but, because of the increased flexibility of non-parametric methods, they require larger quantities of training data. This is particularly a problem when the dimensionality of the feature space is large. This problem is known as the *curse of dimensionality*. Semi-parametric methods give a compromise between these two extremes.

1.2 Statistical Decision theory

The theory of statistical classification deals with the problem of assigning one or more individuals to one of several possible groups or populations on the basis of a set of characteristics observed on them. Thus, the problem of classification can be considered as a special case of multivariate decision theory. This Section introduces some fundamentals of this theory for classification problems with predefined classes. Given a set of objects to be classified, let K be the finite number of classes we are going to consider. The vector X of the measurements of each object is called the *feature vector*; the feature space χ is typically a subset of $\mathbb{R}^p$. Suppose there exists an *a priori probability* π_k that an object belongs to a specific class k; π_k represents the proportion of class k cases in the population under study and it can be known or unknown. Suppose we are forced to make a decision about the class the object we are observing belongs to without measuring it and the only information we are allowed to know are the prior probabilities. In this case it seems logical to use this simple decision rule: decide k if $\pi_k \geq \pi_j \; \forall j = 1, ...K$. Of course this rule will always bring the same decision if there exists any prior probability greater than the others. Fortunately in most circumstances we are given observations of the feature vector X to improve our classifier. We consider X to be a random variable whose distribution depends on the specific class. Let $p_k(x)$ indicate the density according to which feature vectors from class k are distributed. This is the *class conditional probability density function* $p(x|k)$. In this framework classifying an object, on the bases of an observed value $X = x$, means making one of the K possible decisions $1, 2, \ldots, K$. Thus a classifier can be defined as a procedure $c : x \in \chi \mapsto \hat{k} \in \{1, 2, \ldots, K\}$. The usual way to determine the goodness of this procedure is in term of a *loss function* $L(\hat{k}, k)$ that is the loss incurred by making the decision $\hat{k}$ while the true labelling is k. A very commonly used loss function in classification theory is the *0-1 loss*

$$L(\hat{k}, k) = 1 - \delta(k, \hat{k}), \tag{1.2.1}$$

where $\delta(\cdot, \cdot)$ is the Kronecker symbol. As we can see from (1.2.1), the $0-1$ loss is a reasonable choice if every misclassification is equally serious and we will always employ the $0-1$ in the following. Given an observation x, the *conditional risk* $R(\hat{k}|x)$ associated with the action $\hat{k} = c(x)$ characterizes the performance of the rule $c(\cdot)$. Let C indicate the true and unknown class label of the observed vector x, the conditional risk is usually defined in terms of the underlying loss function $L(\hat{k}, k)$ as

$$R(\hat{k}|x) = E[L(c(x), C)|x] = \sum_{j=1}^{K} L(\hat{k}, j)p(j|x), \tag{1.2.2}$$

where $p(j|x)$ is the posterior probability of class j given $X = x$. The posterior probability can be easily computed by the Bayes formula

$$p(j|x) = Pr(C = j|X = x) = \frac{\pi_j p_j(x)}{\sum_{i=1}^{K} \pi_i p_i(x)}, \tag{1.2.3}$$

thus the conditional risk (1.2.2) can be expressed as

$$R(\hat{k}|x) = \frac{\sum_{j=1}^{K} L(\hat{k}, j) p_j(x) \pi_j}{\sum_{i=1}^{K} \pi_i p_i(x)}. \tag{1.2.4}$$

The *total risk* is the expected loss associated with a given decision rule $c(x)$ and it is given by

$$R(c) = E_x(R(c(x)|x)) = \int_{\chi} R(c(x)|x) p(x) dx \tag{1.2.5}$$

where $p(x) = \sum_{i=1}^{K} \pi_i p_i(x)$. Let D be the collection of all measurable decision rules. According to the definition of Lehman (1986) the *Bayes decision rule* is the rule $c \in D$ that minimizes the total risk (1.2.5) and this minimum value is called *Bayes risk.* In practice a Bayes classifier $c(x)$ is built up associating at each observed vector x the label $\hat{k}$ that minimizes the conditional risk

$$c(x) = \hat{k} = argmin_{k=1,\ldots,K} R(k|x),$$

thus the overall risk results minimized. The classification rules based on the minimization of the risk result in minimum error rate classifications. For the $0-1$ loss case, that we are considering in this chapter, the Bayes rule is

$$c(x) = \hat{k} = argmax_{k=1,\ldots,K} \{p_k(x)\pi_k\}. \tag{1.2.6}$$

One of the most useful way to represent a classification rule is in terms of a set of discriminant functions $g_i(x),\ i = 1, \ldots, K$ such that the classifier $c(x)$ will assign the feature vector x to the class corresponding to the largest discriminant

$$c(x) = k \ \Leftrightarrow\ g_k(x) > g_i(x) \ \ \forall i \neq k.$$

For the minimum error rate case, the discrimination functions (df) correspond to the posterior probabilities $g_i(x) = p(i|x)$. Clearly the choice of discriminant functions is not unique as we get the same classification result if we compose each df with a monotonically increasing function, in the sense that if G is a monotonically increasing function we have

$$c(x) = k \;\Leftrightarrow\; g_k(x) > g_i(x) \;\; \forall i \neq k \;\Leftrightarrow\; G(g_k(x)) > G(g_i(x)) \;\; \forall i \neq k.$$

Thus sometimes in our case it could be easier to compute the df as

$$g_k(x) = p_k(x)\pi_k \;\; \forall k = 1, \ldots, K,$$

or as

$$g_k(x) = \log p_k(x) + \log \pi_k \;\; \forall k = 1, \ldots, K. \tag{1.2.7}$$

1.3 Parametric discriminant analysis

In the parametric approach, a general formula for the probability distribution of observation vectors for each class is assigned. The free parameters contained in the formula are estimated by the classifier during the learning stage. In the present Section we will assume that the observation vectors in each class follow a multivariate normal distribution

$$p(x|k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)\right] \;\; , k = 1, \ldots, K \tag{1.3.1}$$

where we are considering x as a p - component vector, μ is the p component mean vector, Σ is the (p,p) covariance matrix and the operators $|\cdot|$ and $(\cdot)^{-1}$ are respectively the determinant and the inverse. Furthermore, If not indicated explicitly, each vector will be considered as a column vector. In the multivariate normal case the discriminant functions (1.2.7) are

$$g_k(x) = -\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k) - \frac{p}{2}\,ln 2\pi - \frac{1}{2}\,ln|\Sigma_k| + \; ln\pi_k \;\; k = 1, \ldots, K. \tag{1.3.2}$$

In the following subsections we will show some special cases.

1.3.1 Estimation of the parameters

The parametric approach to pattern recognition is characterized by a learning stage before classification. As said in Section 1.1, the set of data used for learning, that is to estimate the parameters of the assigned distributions and the prior classes probabilities, is called training set. The parameters of the class conditional density are usually estimated via Maximum Likelihood (ML) criterion. If we explicit the dependence on the unknown vector of parameters θ, we have

$$\begin{aligned} p_k(x) &= p_k(x, \theta), \\ \pi_k &= \pi_k(\theta). \end{aligned}$$

Let $\{x_{ki}, i = 1, \dots, n_k\}$ be the training set of observations from class k, $k = 1, \dots, K$. The likelihood function of the whole training set is

$$L(\theta) = \prod_{k=1}^{K} \prod_{j=1}^{n_k} p_k(x_{kj}, \theta) \pi_k(\theta).$$

If the classes prior probabilities are completely known, they can be dropped from the likelihood function, otherwise they are retained and considered as parameters to be estimated. The maximum likelihood estimators of $(\theta, \pi_1, \dots, \pi_K)$ are the maximizers of the log likelihood

$$\log L(\theta, \pi_1, \dots, \pi_K) = \sum_k \sum_j \log p_k(x_{kj}, \theta) + \sum_k n_k \log \pi_k.$$

Considering the constraint

$$\sum_k \pi_k = 1, \tag{1.3.3}$$

we get that the ML estimates of the classes prior probabilities are

$$\hat{\pi}_k = \frac{n_k}{\sum_j n_j}, k = 1, \dots, K,$$

that are the proportion of training samples from class k over the whole training set observations. The ML estimates of the remaining parameters are then obtained maximizing the function

$$\log L(\theta) = \sum_k \sum_j \log p_k(x_{kj}, \theta) + constant,$$

over θ. More often the parameters to be estimated divide into separate vectors θ_k specific for each class k, thus the ML estimators are obtained maximizing each class specific log likelihood

$$\log L_k(\theta_k) = \sum_j \log p_k(x_{kj}, \theta_k) + constant, \quad k = 1, \ldots, K$$

over θ_k. As example if we assume that the class conditional pdf are p-variate normal $N(\mu_k, \Sigma_k)$, the ML estimates of the mean vector μ_k and of the variance matrix Σ_k are given by their empirical analogs

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}, \\ \hat{\Sigma}_k &= \frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ik} - \hat{\mu}_k)(x_{ik} - \hat{\mu}_k)' \ , \end{aligned}$$

for every $k = 1, \ldots, K$.

1.3.2 Linear discriminant analysis

Suppose the feature vector components are statistically independent with the same variance σ^2. In this simple case the covariance matrix is equal for each class k, $\; k = 1, \ldots, K$ and diagonal $\Sigma_k = \Sigma = \mathbf{I}\sigma^2$, where $\mathbf{I}$ is the identity matrix. Observing that $|\Sigma| = \sigma^{2p}$ and $|\Sigma|^{-1} = \mathbf{I}(1/\sigma^2)$ and dropping the terms that are not class dependent, the (1.3.2) can be rewritten as

$$g_k(x) = -\frac{\|x - \mu_k\|^2}{2\sigma^2} + \ ln\, \pi_k \;\; k = 1, \ldots, K$$

where $\|\cdot\|$ is the *euclidean norm*. Expanding the quadratic form $\|x-\mu_k\|^2 = (x-\mu_k)^t(x-\mu_k) = x^tx - 2\mu_k^tx + \mu_k^t\mu$, and ignoring the additive constant x^tx leads to the equivalent *linear discriminant functions*

$$g_k(x) = \frac{2\mu_k^t}{2\sigma^2}x - \frac{\mu_k^t\mu}{2\sigma^2} + \ ln\ \pi_k \ \ k = 1, \ldots, K.$$

Consider now another simple case. Suppose again the covariance matrices for all the classes identical $\Sigma_k = \Sigma, \ \ k = 1, \ldots, K$ but arbitrary. In this case the simplification of the (1.3.2) leads to

$$g_k(x) = (\Sigma^{-1}\mu_k)^tx - \frac{\mu_k^t\Sigma^{-1}\mu_k}{2} + \ ln\ \pi_k \ \ k = 1, \ldots, K$$

thus the resulting discriminant functions are again linear. Geometrically if the discriminant functions are linear, the *decision surface* that separates the decision regions are subsets of the hyperplanes defined by the linear equations $g_h(x) = g_k(x)$.

1.3.3 Quadratic discriminant Analysis

In the general case the covariance matrix Σ_k is a totally arbitrary symmetric and positive definite matrix for each class k, thus the quadratic form $x^t\Sigma_kx$ in the (1.3.2) can not be ignored and the resulting discriminant functions are quadratic

$$g_k(x) = -\frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\ ln|\Sigma_k| + \ ln\ \pi_k \ \ k = 1, \ldots, K$$

Geometrically if the discriminant functions are quadratic the decision surfaces can assume any general hyperquadratic form (hyperhypeboloids, hyperparaboloids, pair of hyperplanes, etc).

1.4 Non parametric discriminant analysis

The common parametric forms rarely fit the actual underlying class densities. When no distribution assumptions within each class is made, nonparametric methods can be used to estimate the class specific densities $p_k(x),\ k = 1, \dots, K$. Non parametric discriminant analysis (NPDA) consists in classification criteria based on nonparametric estimates of class specific pdf. In NPDA, the class membership of each observed x can be evaluated plugging in the Bayes classification rule the class specific densities estimated from the training set and their prior probabilities. A popular non parametric estimation of the density function is given by this is the case of *kernel methods*.

In order to introduce the kernel approach, we start considering the univariate case. Assume we have a random sample $x_1, \dots, x_n$ taken from a univariate continuous density f. The kernel density estimator $\hat{f}$ of f is defined as

$$\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^{n} \mathsf{K} \left\{ \frac{(x - X_i)}{h} \right\}, \tag{1.4.1}$$

where h is a positive number called *bandwidth* or *smoothing* parameter and K is a function called *kernel* satisfying

$$\int_{\mathbb{R}} \mathsf{K}(x) dx = 1 \ .$$

As we can see from equation (1.4.1), the kernel estimate at some point x is the average of the n kernel centered at each observation x_i and scaled by h. It can be shown that the choice of the kernel function is not particularly important in the sense that the ”goodness” of the estimation slightly depend on the shape of K but it is strongly influenced by the choice of the smoothing parameter h. In the classical parametric statistics the goodness of an estimator, that is its closeness to the parameter of interest, is measured in terms of MSE. In our case we are considering $\hat{f}(x,h)$ as an estimator of the density function $f(x)$ at each fixed point $x \in \mathbb{R}$, thus we need an error measure that globally measure the distance between $f(\hat{\cdot},h)$ and $f(\cdot)$ over $\mathbb{R}$. An error rate that satisfy this request is the Integrated Square Error

$$ISE\{\hat{f}(\cdot,h)\} = \int_{\mathbb{R}} \{\hat{f}(x,h) - f(x)\}^2 dx.$$

Actually the ISE so defined is implicitly specific for the dataset $x_1, \dots, x_n$ by witch we constructed $\hat{f}$ thus, in order to take into account all possible sets of data, we use the Mean Integrated Squared Error

$$MISE = E[ISE\{\hat{f}(.,h)\}] = E\int_{\mathbb{R}}\{\hat{f}(x,h)-f(x)\}^2 = \int_{\mathbb{R}} E\{\hat{f}(x,h)-f(x)\}^2 = \int MSE\{\hat{f}(x,h)\}d$$

In estimation theory a very important concept is the rate of convergence that is a measure of how ”quickly” an estimator approaches its target as the sample size n increases. Using the MISE criterion, in the hypothesis that the density function to be estimated belongs to the Sobolev space $H^s(R), s \in \mathbb{N}$, it can be shown that

$$\inf_{h>0} MISE\hat{f}(\cdot,h) = O\{n^{-\frac{2s}{2s+1}}\},$$

and the h that realizes this limit is the *optimal bandwidth*. We notice that $O\{n^{-\frac{2s}{2s+1}}\}$ is the best error rate in the *minimax* sense (see Robbins, 1951), thus plugging in our estimator the optimal h we gain asymptotical optimality properties. An example of univariate kernel function is the Epanechnikov kernel

$$\mathsf{K}(x) = \frac{3}{4}(1-x^2)\mathbf{1}_{\{|x|<1\}}.$$

Consider now the multivariate case. A p-variate kernel K is a function from $\mathbb{R}^p$ to $\mathbb{R}$ satisfying

$$\int \mathsf{K}(x)dx = 1, \;\; x \in \mathbb{R}^p$$

In the most general form, the p-dimensional kernel estimator is

$$\hat{f}(x; H) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{K}_H(x - x_i) \tag{1.4.2}$$

where H is a positive definite symmetric (p, p) matrix called *bandwidth matrix* and its elements are called *smoothing parameters*; furthermore

$$\mathsf{K}_H(x) = |H|^{-1/2}\mathsf{K}(H^{-1/2}x).$$

As in the univariate case, (see Wand and Jones, 1995) if the density function to be estimated belongs to the Sobolev space $H^s(R^p), s \in \mathbb{N}$, it can be shown that

$$\inf_{H \in S_p} MISE\hat{f}(\cdot, H) = O\{n^{-\frac{2s}{2s+p}}\}, \tag{1.4.3}$$

where S_p is the set of symmetric and positive definite (p, p) matices. The H that realizes this limit is the *optimal bandwidth matrix*. We notice again that $O\{n^{-\frac{2s}{2s+p}}\}$ is the best error rate in the *minimax* sense (see Robbins, 1951), thus again plugging in our estimator the optimal H we gain an asymptotical optimality properties.

Unfortunately in the multivariate case the rate of convergence of any asymptotical optimal density estimator becomes slower as the dimension p increases. This slower rate is a manifestation of the *course of dimensionality* or *empty space phenomenon* (Scott at *al.*, 1992). Multivariate density estimation is in fact very difficult, and usually not practically applied, in more than about five dimensions due to the sparseness of data in higher dimensional spaces (see Wand and Jones, 1995).

In order to circumvent the problem of the slow convergence of density estimators at high dimensions, we could think (see Amato et al. 2003) to transform the data so to be able to factorize the density in the product of univariate densities, one for each dimension

$$f(x) = \prod_{j=1}^{p} f_j(x_j).$$

Estimating each dimension pdf by a generic optimal univariate density estimator $\hat{f}_j$, we would obtain that the multivariate estimator

$$\tilde{f}(x) = \prod_{j=1}^{p} \hat{f}_j(x_j), \tag{1.4.4}$$

and then we would have the same convergence order as in the univariate case $O\{n^{-\frac{2s}{2s+1}}\}$.

Using a univariate kernel estimator (1.4.1) in the (1.4.4) we would obtain the multivariate kernel estimator

$$\begin{aligned} \hat{f}(x, h_1, \ldots, h_p) &= \left(\prod_{d=1}^{p} h_d\right)^{-1} \frac{1}{n} \prod_{j=1}^{p} \sum_{l=1}^{n} \mathsf{k}\left(\frac{x_j - x_{lj}}{h_j}\right) \\ &= \left(\prod_{d=1}^{p} h_d\right)^{-1} \frac{1}{n} \sum_{l=1}^{n} \prod_{j=1}^{p} \mathsf{k}\left(\frac{x_j - x_{lj}}{h_j}\right) \end{aligned}$$

This leads to use a *product kernel* estimator, that is the product of symmetric univariate kernels κ

$$\mathsf{K}(x) = \prod_{j=1}^{p} \kappa(x_j).$$

In conclusion if we were able to factorize the underling pdf of the data in the product of univariate pdf (one for each dimension) we could circumvent the curse of dimensionality, using for each univariate pdf an asymptotically optimal univariate density estimator. In the next sections we will face the problem of finding a transformation of the original data such that the transformed variables cumulative pdf could be estimated trough a product kernel. The transformed variables should be the underlying factors or components that describe the essential structure of the data. It is hoped that these components correspond to some physical causes that were involved in the process that generated the data in the first place. We will consider linear transformations only, because then the interpretation of the representation is simpler, and so is its computation. Thus we will express the transformed variables as a linear combination of the observed variables. In matrix representation we have

$$y = Tx \tag{1.4.5}$$

where T is not necessarily a square matrix. In the following sections, we discuss some statistical properties that could be used to determine the transformation matrix T.

In order to relate the multivariate kernel density estimation to our classification problem, we remember we want to estimate the probability density function of each class j, using the N_j observations from the training sample ($j \in \{1, \ldots, K\}$). The general form of the class k (p-dimensional) kernel density estimator is

$$\hat{f}_k(x; H_k) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathsf{K}_{H_k}(x - x_{ki}) \tag{1.4.6}$$

where H_k is the bandwidth matrix specific for class K, and $\{x_{ki},\ i = 1, \ldots, N_k\}$ is the training sample from the population k. For a full discussion on the choice of *smoothing parameters*, see Silverman, 1986.

Choosing a diagonal bandwidth matrix and the univariate κ as the normal density function leads to the class densities estimates

$$\hat{f}_k(x) = (2\pi)^{-p/2}\left(\prod_{d=1}^{p} h_{kd}\right)^{-1}\frac{1}{N_k}\sum_{l=1}^{N_k}\prod_{j=1}^{p} exp\left\{-\frac{(x_j - x_{lkj})^2}{2h_{kj}^2}\right\} \tag{1.4.7}$$

called *gaussian product* kernel estimators. Equation (1.4.7) is very popular in multivariate kernel density estimation.

1.4.1 Principal component discriminant analysis

One statistical principle for choosing the transformation matrix T in (1.4.5) is to limit the number of components y_i to be quite small so that they contain as much information on the data as possible. This leads to a family of techniques called principal component analysis or factor analysis. Given a vector x of a large number p of interrelated random variables, the main idea of *principal components analysis* (PCA) is to look for a fewer number ($<< p$) of derived variables that retains the variation present in the component of x as much as possible. The PCA procedure consists in transforming the original set of variables to the principal components variables (PCs), which are uncorrelated and ordered so that most of the original set variation is concentrated in the first few. The choice of the most important PCs number is more or less an heuristic decision, and it may depend on the application. We will briefly show the derivation of PCA using the covariance method. We can interpret PCA as a linear transformation that chooses a new coordinate system to represent the data in order to have that by any projection of the data set, the greatest variance comes to lie on the first axis (called the first principal component), the second greatest variance on the second axis, and so on. Therefore, assuming x has zero empirical mean, we want to find an orthonormal projection matrix P such that the transformed data $y = P^t x$ are uncorrelated. This results in finding an orthogonal matrix P such that y covariance matrix D is diagonal

$$D = cov(y) = diag(a_1, \dots, a_p)$$

where $a_i = var(y_i), \;\; i = 1, \dots, p$. It's easy to see that

$$D = P^t cov(x) P$$

and thus

$$PD = cov(x)P$$

Indicating each column of P as p_i, we get

$$a_i p_i = cov(x) p_i$$

This last expression reveals a simple way to calculate the PCs that consists in finding the eigenvectors of x covariance matrix. It turns out that the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the data set. The original measurements are finally projected onto the reduced vector space. Note that the eigenvectors of $cov(x)$ are actually the columns of the matrix V, where cov(x)=ULV' is the singular value decomposition (SVD) of $cov(x)$. For a detailed description of PCA see Jolliffe, 2002.

PCA is a popular technique in pattern recognition and we will use it only to linearly transform the data in order to decorrelate them before the classification step. In practice, the training data from each class k are used to compute the sample mean and the sample covariance matrix of the centered data from class k. The projection matrix P_k for each class k is then evaluated via SVD procedure. Using the N_k observation from class k training set,the transformed data $y_{lk} = P_k^t x_{lk}$ are calculated and so for each class k and for each dimension j, the univariate pdf g_{kj} of the transformed variable y_k j-th component is estimated via univariate kernel density estimator. The resulting product kernel density estimator for y_k is $\hat{g}_k(y_k) = \prod_{j=1}^p g_{kj}(y_{kj})$. A change of variables $x = P_k y$ allows to go back to the original data domain and get the estimation of the class k pdf

$$\hat{f}_k(x) = \hat{g}_k(P_k^t x)|det(P_k)| = \prod_{j=1}^{p} g_{kj}((P_k^t x)_j). \tag{1.4.8}$$

However, PCA just decorrelates the data without making them independent. Thus the factorization in (1.4.10) can be supported only under the assumption of independence that is not valid unless the data are Gaussian. In fact if we assume that class k pdf is a p-variate normal $N_p(\mu_k, \Sigma_k)$ of mean μ_k and covariance Σ_k, then the random vector $y = P_k^t x$ has a p-variate normal pdf $N_p(P_k^t \mu_k, D_k)$ with covariance matrix D_k diagonal and so has independent components, thus in this case the factorization in (1.4.10) holds because for gaussian data uncorrelated components are always independent.

1.4.2 Independent components discriminant analysis

Another principle that has been used for determining the matrix T in (1.4.5) is independence: the components y_i should be statistically independent. This means that the value of any one of the components gives no information on the values of the other components. As seen in the previous Section, in PCA the transformed variables are assumed to be independent, but this is only true when the data are assumed to be gaussian. In reality, however, the data often do not follow a gaussian distribution. Independent Component Analysis (ICA) is a statistical method whose main target is to find statistically independent components, in the general case where the data are non gaussian. We could define ICA as a linear transformation given by a matrix as in (1.4.5), so that the transformed random variables are as independent as possible. A very intuitive and important principle of ICA estimation is maximum non gaussianity. The idea is that according to the central limit theorem, sums of non gaussian independent random variables are closer to gaussian than the original ones. Therefore, if we take a linear combination of the observed variables, this will be maximally non gaussian if it equals one of the independent components. This is because according to the central limit theorem, if it was a real mixture of two or more components, it would be closer to a gaussian distribution.

A very important measure of non gaussianity is given by negentropy. Negentropy is based on the differential entropy. The differential entropy H of a random vector y with density $f(y)$ is defined as $H(y) = -\int f(y) log f(y) dy$, (Cover and Thomas, 1991). Negentropy J is defined as $J(y) = H(y_{gauss}) - H(y)$, where y_{gauss} is a Gaussian random variable of the same covariance matrix as y. The estimation of negentropy is difficult and in practice, some approximation have to be used. Here we introduce the approximation proposed by Hyv$\ddot{a}$rinen (1997) that has very promising properties, and which will be used in the following to derive an efficient method for ICA. Hyv$\ddot{a}$rinen approximates the negentropy $J(y)$ as

$$J_G(y) = \sum_{i=1}^{p} \{\mathbb{E}(G(y_i) - \mathbb{E}(G(Z))\}^2 \tag{1.4.9}$$

where Z is a zero-mean standard normal random variable and the function G, called contrast function, is usually the power three transform. We refer to Hyv$\ddot{a}$rinen(1997) for the details of the derivation of the ICA transform $y = Tx$ and of its statistical properties.

ICA is can be used in pattern recognition to linearly transform the data in order to make them approximately independent before the classification step. This approach was first introduced by Amato *et al* (2003). In practice, the training data from each class k are used to compute the sample mean and the centered data are then used to derive the ICA transformation matrix T_k for class k sample. Using the N_k observations from class k training set,the transformed data $y_{lk} = T_k x_{lk}$ are calculated and so for each class k and for each dimension j, the univariate pdf g_{kj} of the transformed variable y_k j-th component is estimated via univariate kernel density estimator. The resulting product kernel density estimator for y_k is $\hat{g}_k(y_k) = \prod_{j=1}^{p} g_{kj}(y_{kj})$. A change of variables $x = T_k^{-1}y$ allows to go back to the original data domain and get the estimation of the class k pdf

$$\hat{f}_k(x) = \hat{g}_k(A_k x)|det(A_k)| = \prod_{j=1}^{p} g_{kj}((T_k x)_j)|det(A_k)| \tag{1.4.10}$$

where A_k is the pseudo inverse of T_k. Amato *et al.* (2003) showed that the decision rule resulting substituting the estimated class pdf $\hat{f}_k$ in the 1.2.6 converges uniformly in probability to the Bayes classification rule and is asymptotically optimal.

1.5 Local Discriminant methods for image classification

In the present and following sections we shall deal with supervised classification of bidimensional images. The general problem can be formulated as follows. A continuous two dimensional region is partitioned into a finite number of sites called pixels (pictures elements), each pixel belonging to one of a predefined finite set of classes $\{1, \dots, K\}$. The set can represent, e.g., land cover categories, cloudy or clear sky conditions, etc.. The true labelling of the region is unknown but associated with each pixel there is a multivariate (actually, multispectral) value which provides information about its label. Bayesian discriminant analysis consists in choosing the class $\hat{k}$ from the set $\{1, \dots, K\}$ according to the Bayes decision rule (1.2.6). Several discriminant analysis methods have been proposed in the literature, according to the choice of the class conditional densities $p_k(x)$ parametrically (e.g., Gaussian) or nonparametrically (e.g., Kernel density estimation) and to the way the multidimensionality of x is faced. It is out of the scope of the present chapter to review the methods developed in the framework of discriminant analysis. Rather, we shall focus on the observation that traditional image classification approaches often neglect the information about spatial relationships between adjacent pixels. In other words, classification through Eq. (1.2.6) is performed pixel-wise and no information on other pixels, neither the surrounding ones, is used. However, pixels belonging to a same class tend often to cluster together in many applications, and remote sensing is just one of these. Referring to the above mentioned examples, land cover and cloud fields usually extend over regions of several pixels, depending on the spatial resolution of the sensor. Also note that strict application of Bayes decision rule gives rise to typical ‘pseudo-noisy’ reconstructed label fields, where often isolated labels are present that are not physically feasible (that is, a pixel belongs to a certain class, and all surrounding pixels belong to other, different classes). This effect is disturbing especially in the analysis of medical images, where sometimes these isolated pixels refer to tissues that cannot be present in the corresponding locations. This effect is intrinsic to the discriminant analysis and is due to the uncertainty of the decision rule coming from the overlap of the probability density functions among different classes: the more such density functions are overlapped, the bigger the effect of ‘pseudo-noise’. To overcome this problem, the procedure usually used is an empirical post-processing of the retrieved label field, where a sort of smoothing of the label field obtained by discriminant analysis is accomplished in a remote sensing application (see Ju, Gopal,and Kolaczyk, 2005).

An attempt to incorporate pixel context in image classification goes back to the Iterated Conditional Modes (ICM) (Besag, 1986). Basically, this method assumes that the true label set of an image is a realization of a locally dependent Markov Random field so that the posterior class probability for a specific data point also depends on the labelling of its neighborhood. After obtaining a first class estimate for each pixel using any non local method, local (i.e., depending on the location in the image) priori probabilities of classes are computed from the estimate, considering a neighbor of each pixel; then new labels are assigned to the pixels maximizing the class posterior probability and relying on the prior probabilities just estimated. The procedure is iterated until convergence. ICM method has been applied successfully in the field of remote sensing (Khedam et al., 2004) and compared to Maximum Likelihood classification (Keuchel et al., 2003).

In the following we first formalize discriminant analysis in a framework that focuses on how much a class can be visible or nonvisible, then we introduce some discriminant analysis methods that use spatial information around each pixel in order to localize the methods. We have the twofold objective of a) improving local label estimates by increasing the number of pixels (i.e., information) involved in the decision rule; b) reducing of the ‘pseudo-nuisance’ present in pixel-wise discriminant analysis. These methods will be best suited for visible and nonvisible classes. Numerical experiments will be performed. In particular, methods will be applied to the problem of retrieving cloud mask from remote sensed images.

1.6 Notations and Assumptions

Let us consider a general case where an object has to be classified as coming from one of a fixed number of predefined classes, say $1, \ldots, K$. Associated with this object there is possibly a multivariate record $x = (x_1, \ldots, x_D)$ belonging to a subset χ of $\mathbb{R}^D$ and it is interpreted as a particular realization of a random vector $X = (X_1, \ldots, X_D)$. In our case, without any loss of generality, an object is a pixel of an image and it is usually identified by a couple of coordinates. With a slight abuse of notation, we will identify an observation or pixel with its measurement x when no ambiguity arises.

In the present work we shall consider the univariate case. This does not restrict applicability of the methods we are going to consider, since extension to the D-dimensional case is straightforward. In addition this assumption is particularly suited for those applications where only univariate measurements ($D = 1$) are available, or one covariate is already able to give good classification rates with respect to the multivariate case; then improvement of the univariate classification could give classification rates comparable with those of the (more expensive) multivariate case.

Let us now consider first the case where the random variable X is discrete, so $\chi = \{1, \ldots, N\} \subseteq \mathbb{N}$.

For the purpose of the present paper, we introduce the following definitions.

Definition 1. $x \in \chi$ is called **dominant** for the class k with respect to a Bayes classification rule γ if and only if $p(k|x) \geq p(i|x)$, $i = 1, \ldots, K$.

Definition 2. For $k = 1, \ldots, K$ we define **dominant set**, $\mathcal{D}_k^\gamma$, for class k with respect to the Bayes rule γ, the set

$$\mathcal{D}_k^\gamma := \{x \in \chi : x \text{ is dominant for the class } k \text{ and the rule } \gamma\}.$$

Definition 3. For $k = 1, \ldots, K$ we define **dominance index** of class k with respect to the Bayes classification rule γ, $\delta^{\gamma}(k)$, the quantity

$$\delta^{\gamma}(k) := \sum_{x \in \mathcal{D}_k^{\gamma}} p_k(x),\ k = 1, \ldots, K.$$

Definition 3 assumes that a dominant yields only one class k. In the general case this is not the rule; then let us give the following

Definition 4. A target class of an observation $x \in \chi$, $\kappa(x)$, with respect to the Bayes rule γ is the set of classes for which x is dominant:

$$\kappa(x) := \{k : x \text{ is dominant for } k,\ 1 \leq k \leq K\}.$$

Let

$$w_k(x) = \frac{1}{|\kappa(x)|},$$

with $|\cdot|$ being cardinality of the set. Then Definition 3 can be generalized as

$$\delta^{\gamma}(k) := \sum_{x \in \mathcal{D}_k} w_k(x) p_k(x),\ k = 1, \ldots, K$$

that for each $x \in \chi$ corresponds to assign equal probability of occurrence to all classes for which x is dominant. In the following we assume for simplicity's sake that a dominant yields only one class, so that $w(k) = 1$.

The above formalism can be applied to probability density functions provided that Definition 3 is changed as

Definition 5. For $k = 1, \ldots, K$ we define **dominance index** of class k, $\delta^{\gamma}(k)$, with respect to the Bayes classification rule γ, the quantity

$$\delta^{\gamma}(k) := \int_{\mathcal{D}_k^{\gamma}} p_k(x) dx,\ k = 1, \ldots, K.$$

Table 1.1 shows an example of discrete probabilities with corresponding dominance index and class of dominance supposing, e.g., constant priori class probabilities. We want

	1	2	3	4	$\dots$	N	Dominance index
$C=1$	$p_1(1)$	$p_1(2)$	$p_1(3)$	$p_1(4)$	$\dots$	$p_1(N)$	$\delta(1)$
$C=2$	$p_2(1)$	$p_2(2)$	$p_2(3)$	$p_2(4)$	$\dots$	$p_2(N)$	$\delta(2)$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$C=K$	$p_K(1)$	$p_K(2)$	$p_K(3)$	$p_k(4)$	$\dots$	$p_K(N)$	$\delta(K)$
Class of dominance	2	2	1	3	$\dots$	2	

Table 1.1: Example of discrete distributions of K classes and corresponding dominants and dominance index.

to pay particular attention on the dominance index for each class. Actually visibility of a class (that is, capability of a discriminant analysis method to predict that class) depends on how many x predict that class and on the probability of occurrence of those x, both contributing to the dominance index $\delta(k)$. It can happen that a class is least visible or it is even very easy to build examples where a class is not visible at all (that is $\delta(k) = 0$), which means that it won't ever be predicted by the method. In these cases visibility of the classes has to be improved by increasing as much as possible its class priori probability, with the risk to make less visible the other classes and then to degrade capability of their correct prediction.

In the classical discriminant analysis the class prior probabilities do not depend on each single pixel x of an image and they are generally estimated from the training set as the fraction $\hat{\pi}_k$ of training set pixels belonging to each class:

$$\pi_k = \hat{\pi}_k,\ k = 1, \dots, K.$$

In the practice it is very common also to consider uniformly distributed classes, so that each class has the same prior probability:

$$\pi_k = \frac{1}{K},\ k = 1, \ldots, K.$$

These assumptions give rise to *naîve* classification rules, $\gamma^{\text{naîve}}$, that globally take best account of the occurrence probability of the classes or do not privilege the classes a priori at all.

1.7 Local priors

Nonlocal priors of Section 1.6 take account of the global occurrence of the classes over an image and do not take any account of spatial correlation or of local features in an image. In particular when an image has a wide homogeneous region labelled by the same class k, the spatial correlation is maximum and it appears natural to be more confident about the presence of class k in pixels belonging to that region. Note that this is the rule in most applications of image classification. Moreover, location of these homogeneous regions cannot be predicted in advance in several applications, especially when different images have to be classified. In particular we cannot rely in general on the training images to this purpose, nor we can assume that the naîve global priors represent the true local probability of occurrence of classes accurately. For these reasons local priors are prone to improve accuracy of classification in homogeneous regions, provided that a good estimation of prior classes probabilities can be given.

In this Section we propose some methods that exploit information contained in the neighborhood of each single pixel; they modify the posterior probability estimates given by Eq. (1.2.3) introducing a set of local prior probabilities, $\{\pi_k(x)\}_{k=1,\ldots,K}$, specific for each single pixel x of an image:

$$p(k|x) = Pr(C = k|X = x) = \frac{\pi_k(x)p_k(x)}{\sum_{i=1}^{K} \pi_i(x)p_i(x)}. \tag{1.7.1}$$

Let us consider for each pixel x_c a neighborhood region $\mathcal{B}(x_c)$. This region can have any shape and size. Furthermore let

$$\mathcal{B}_k(x_c) := \{x \in \mathcal{B}(x_c) : \gamma(x) = k\}$$

be the set of pixels labelled as k in the neighborhood $\mathcal{B}(x_c)$. Let us indicate as $\mathcal{L}(\mathcal{B}(x_c); \gamma)$ the set of labels associated to all pixels of an image by any discriminant rule γ.

In the following sections we introduce some classification Bayes-like rules relying on Eqs. (1.7.1) and (1.2.6), differing in the way of estimating class prior local probabilities.

1.7.1 Local voting priors

Suppose an estimate of class label is available for all pixels through a classical Bayes rule (1.2.6) with some a priori class probabilities. Let

$$\varphi_k(x_c) = \frac{|\mathcal{B}_k(x_c)|}{|\mathcal{B}(x_c)|}, \; k = 1, \dots, K, \tag{1.7.2}$$

be the relative frequency of labels k in $\mathcal{B}(x_c)$. Intuitively , for any pixel x_c, we would estimate the set of prior probabilities $\{\pi_k(x_c)\}_{k=1,\dots,K}$ in order to enhance the class with the highest relative frequency in $\mathcal{B}(x_c)$. Thus a first attempt to estimate the generic class k prior probability $\pi_k(x_c)$ is :

$$\pi_k^{LV}(x_c) = \begin{cases} 1 & \text{if } k = \text{argmax}_{j=1,\dots,K}\{\varphi_j(x_c)\}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.7.3}$$

The local priors just defined naturally satisfy the constraint (1.3.3). By using initial (first-guess) a priori class probabilities and by an iterative application of the above mentioned procedure until convergence, we obtain final class prior probabilities and corresponding class labels. We call these priors *Local Voting priors*. The arising classification method is well suited for those classes that are visible, since it is able to maximize their visibility.

1.7.2 Local frequency priors

Under the same hypothesis of the previous subsection, we propose now to estimate the class k prior probability $\pi_k(x_c)$ by the relative frequency of k labels in $\mathcal{B}(x_c)$:

$$\pi_k^{\mathrm{LF}}(x_c) = \varphi_k(x_c),\ k = 1, \ldots, K.$$

They naturally satisfy the constraint (1.3.3). As with Local Voting priors, the procedure is iterated until convergence thus obtaining the final class prior probabilities and corresponding class labels. We call these priors *Local Frequency priors*. Since priors are based on the relative occurrence frequencies resulting from some discriminant analysis method, φ_k, then the resulting classification method is again well suited for those classes that are visible, since it is able to enhance their visibility, but does not penalize the less visible classes as much as the Local Voting method.

1.7.3 Local integrated priors

Given a pixel x_c and its neighbor $\mathcal{B}(x_c)$ we estimate the prior probability $\pi_k(x_c)$ for the generic class k by summing the probability density functions $p_k(x)$ over the neighborhood region $\mathcal{B}(x_c)$:

$$\pi_k^{\mathrm{LI}}(x_c) \propto \sum_{x \in \mathcal{B}(x_c)} p_k(x),\ k = 1, \ldots, K.$$

We define these priors *Local Integrated priors*. If we consider the normalization

$$\pi_k^{\mathrm{LI}}(x_c) = \frac{\sum_{x \in \mathcal{B}(x_c)} p_k(x)}{\sum_{j=1}^{K} \sum_{x \in \mathcal{B}(x_c)} p_j(x)}, \; k = 1, \ldots, K, \tag{1.7.4}$$

then the set of local priors $\pi_k^{\mathrm{LI}}(x_c)$ satisfies constraint (1.3.3). Each data point label is then estimated through Eq. (1.7.1) using the local class prior probabilities (1.7.4). Notice that this method is not iterative. Moreover these priors are not related to the class prediction obtained by some discriminant analysis method, so that they naturally tend to be not sensitive to visibility or nonvisibility of a class; in practice they are suited for nonvisible classes.

1.7.4 Local nested priors

Suppose again, as in subsections (1.7.1) and (1.7.2), that a first-guess estimate of each data point label is obtained by the classical Bayes rule (1.2.6) with some a priori class probabilities. Given a pixel x_c and its neighbor $\mathcal{B}(x_c)$, we estimate the prior probability $\pi_k(x_c)$ for the generic class k by summing its posterior probability $p(k|x)$ over the region $B(x_c)$:

$$\pi_k^{\mathrm{LN}}(x_c) \propto \sum_{x \in \mathcal{B}(x_c)} p(k|x), \; k = 1, \ldots, K.$$

To satisfy the constraint (1.3.3), priors π_k^{LN} are normalized as

$$\pi_k^{\mathrm{LN}}(x_c) = \frac{1}{|\mathcal{B}(x_c)|} \sum_{x \in \mathcal{B}(x_c)} p(k|x), \; k = 1, \ldots, K. \tag{1.7.5}$$

The procedure is iterated until convergence. We define these priors *Local Nested priors*. As far as visibility of classes is concerned, these priors are a sort of trade-off between LF (suited for visible classes) and LI (suited for nonvisible classes), since they depend on the class label of some discriminant analysis method, but anyway potentially include the contribution of probability density functions all over their domain.

1.8 Asymptotics

We now discuss asymptotic behavior of the local priors considered in Section 1.7. To this purpose let the neighborhood region $\mathcal{B}(x_c)$ of each pixel x_c tend to infinity and assume that $\mathcal{B}(x_c)$ is a homogeneous region of class ℓ, $1 \leq \ell \leq K$.

1.8.1 Local voting priors

Let's consider the asymptotic behavior of the local frequency priors $[\pi_k^{\mathrm{LV}}]$, $k = 1, \ldots, K$, defined in section (1.7.1). Even if the classification process they generate is iterative and local, the dependence from the iteration is lost asimptotically. In fact if we let the neighborhood region of the generic pixel tend to infinity we get that the local voting prior for the generic class k behaves as

$$\pi_k^{LV} = \delta(k, \hat{k}) = \begin{cases} 1 & \text{if } k = \hat{k}, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\hat{k} = \operatorname*{argmax}_{k=1,\ldots,K} P(x \in \mathcal{D}_k \mid C = \ell). \tag{1.8.1}$$

We point out that $\mathcal{D}_k$ is the dominant set defined in (2) at the first step of the iterative process described in subsection 1.7.1. More explicitly in (1.8.1) we have

$$P(x \in [\mathcal{D}_k] \mid C = \ell) = \sum_{x \in D_k} p_\ell(x),\ k = 1, \ldots, K$$

in the discrete case, and

$$P(x \in [\mathcal{D}_k] \mid C = \ell) = \int_{x \in \mathcal{D}_k} p_\ell(x),\ k = 1, \ldots, K$$

in the continuous case.

1.8.2 Local frequency priors

It is easy to see that at each iteration ν, the local frequency priors $[\pi_k^{\mathrm{LF}}]^\nu$, $k = 1, \ldots, K$, asymptotically behave as the probability $P(x \in [\mathcal{D}_k]^{\nu-1} \mid C = \ell)$, where $[\mathcal{D}_k]^{\nu-1}$ is the dominant set at the step $\nu - 1$ of the iterative process described in subsection 1.7.2. It follows

$$[\pi_k^{\mathrm{LF}}]^\nu \to \sum_{x \in [D_k]^{\nu-1}} p_\ell(x),\ k = 1, \ldots, K$$

in the discrete case, and

$$[\pi_k^{\mathrm{LF}}]^\nu \to \int_{x \in [\mathcal{D}_k]^{\nu-1}} p_\ell(x),\ k = 1, \ldots, K$$

in the continuous case.

1.8.3 Local integrated priors

Equation (1.7.4) can be rewritten as

$$\pi_k^{\mathrm{LI}}(x_c) = \frac{\sum_{x \in \mathcal{B}(x_c)} p_k(x)}{|\mathcal{B}(x_c)|} \frac{1}{\frac{\sum_{j=1}^K \sum_{x \in \mathcal{B}(x_c)} p_j(x)}{|\mathcal{B}(x_c)|}},\ k = 1, \ldots, K,$$

so that we can say

$$\pi_k^{\mathrm{LI}}(x_c) \propto \frac{\sum_{x \in B(x_c)} p_k(x)}{|\mathcal{B}(x_c)|},\ k = 1, \ldots, K. \tag{1.8.2}$$

Equation (1.8.2) tells us that asymptotically the local integrated priors tend to be proportional to

$$\sum_{x \in \chi} p_k(x) p_\ell(x),\ k = 1, \ldots, K.$$

in the discrete case, and to

$$\int_\chi p_k(x) p_\ell(x) dx,\ k = 1, \ldots, K.$$

in the continuous case.

1.8.4 Local nested priors

Consider iteration ν of the iterative process described in the Section 1.7.4. From Eq. (1.7.5) it follows that asymptotically the local nested priors $[\pi_k^{\mathrm{LN}}]^\nu$, $k = 1, \ldots, K$, tend to

$$\sum_{x \in \chi} p^{\nu-1}(k|x) p_\ell(x) dx, \; k = 1, \ldots, K \tag{1.8.3}$$

in the discrete case, and to

$$\int_\chi p^{\nu-1}(k|x) p_\ell(x) dx, \; k = 1, \ldots, K \tag{1.8.4}$$

in the continuous case. More explicitly Eq. (1.8.3) can be rewritten as

$$[\pi_k]^{\nu-1} \sum_{x \in \chi} \frac{p_k(x) p_\ell(x)}{\sum_{j=1}^K [\pi_j]^{\nu-1} p_j(x)}, \; k = 1, \ldots, K$$

and Eq. (1.8.4) as

$$[\pi_k]^{\nu-1} \int_\chi \frac{p_k(x) p_\ell(x)}{\sum_{j=1}^K [\pi_j]^{\nu-1} p_j(x)} dx, \; k = 1, \ldots, K$$

1.8.5 Iterations

In LF and NF methods priori class probabilities are defined in terms of an iterative procedure. Therefore the natural question arises about the presence of more solutions. It is easy to see that both methods surely admits several solutions. In particular we can see that, e.g., $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, 0, 0, \ldots, 1)$ are all solutions (that we call *trivial*) of the local classification methods. These solutions are obtained starting iterations with the same final values. In the general case we found that final solutions are very robust with respect to the first-guess chosen and that a few iterations are sufficient to get convergence. As a practical rule it is possible to start from constant class priori probabilities over the classes.

Chapter 2

Multiple Hypothesis Testing

Introduction

In this chapter we introduce the statistical problem of multiple hypothesis testing and review the guiding lines of frequentist and Bayesian approach to it, describing strength and weakness of the two philosophies and trying to find some connections between them. We describe a specific multiple hypothesis testing problem, and propose a new testing procedure that represents a sort of ”empirical” approach. The application of the described methods to a genetic microarray data experiment is provided in Chapter 6.

The Chapter is organized as follows. The first section is a brief overview of the multiple hypothesis testing (MHT) problem. In Section 2 some recent error measures for MHT are introduced and multiple testing error controlling procedures (MTP) are described in Section 3. In Section 4 bootstrap methods are presented. MHT in the Bayesian framework is introduced in Section 5. In Section 6, MAP multiple testing procedure is described and a new MHT procedure is proposed.

2.1 General Framework

In the general framework of single hypothesis testing, we want to test the hypothesis H_0 that an unknown parameter θ of a certain distribution belongs to some subspace $\Theta_0 \in \mathbb{R}^q$ ($q \in \mathbb{N}$), against the alternative hypothesis H_1 that θ belongs to $\Theta_1 \subseteq \mathbb{R}^q$, where $\Theta_0 \bigcap \Theta_1 = \emptyset$. The solution of this problem is in terms of a rejection region Rt which is a set of values in the sample space which leads to the decision of rejecting the null hypothesis H_0 in favor of the alternative H_1. Usually an hypothesis is formulated in order to be rejected, so that we interpret a rejection as a discovery or positive result. In general the rejection region Rt is constructed in order to control at some level α the size of Type I error, i.e. the probability of rejecting the null hypothesis when it is true, while looking for a procedure that possibly minimize the probability of observing a false negative, i.e. the Type II error. A standard approach is to specify an acceptable level α for the Type I error rate and derive testing procedures, i.e., rejection region, that aims to minimize the Type II error rate, i.e., maximize the power, within the class of tests with Type I error rate at most α. For single hypothesis testing, optimality results are available for particular types of data generating distributions, null and alternative hypotheses, and test statistics. In a multiple testing context we need a generalization of Type I and Type II error. Simultaneous testing of multiple hypotheses has always attracted the attention of statisticians. Folks (1984) gives a first introduction to multiple hypothesis testing. When thousands of hypotheses need to be tested simultaneously, the traditional methods are not sensible because of loss of specificity and power. To illustrate the problem, consider the gene expression example. Assume that a chip reveals the expression level of $m = 10.000$ genes relatively to two different biological conditions and we know that not a single gene is differentially expressed. We want to test, simultaneously for each gene, the null hypothesis that the gene is not differentially expressed against the alternative that it is. If we test each of the m hypothesis at level $\alpha = 0.01$, we would expect 100 of the tests would have p-value less then α (i.e. we expect about 100 false positive) and the probability that al least one p-value is less than α is about 1. To illustrate the general procedure, consider the problem of testing simultaneously m null hypotheses. Suppose we have m independent vectors of observations $X_j, \;\; j = 1, \dots, m$, of size n_j and the distribution of each $X_j, \;\; f_j(X_j \mid \theta_j)$, depends on a vector of parameters $\theta_j \in \Omega_j \subseteq \mathbb{R}^{d_j}$. Without loss of generality assume $n_j = n, \; j = 1, \dots, m$. Usually in the applications n is much smaller than m. We want to test simultaneously each of the m non nested hypotheses

$$H_{0j} = \theta_j \in \Theta_{0j} \;\; vs \;\; H_{1j} = \theta_j \in \Theta_{1j} \; j = 1, \dots, m, \;\; \Theta_{0j} \cup \Theta_{1j} = \Omega_j, \;\; \Theta_{0j} \cap \Theta_{1j} = \emptyset. \tag{2.1.1}$$

This general representation covers tests of means, differences in means, parameters in linear models, generalized linear models, and so on.

The decisions to reject or not the null hypotheses are based on test statistics, i.e., functions of the data, $T_j = T(X_{j1}, \dots, X_{jn})$. The testing procedure provides rejection regions, Rt_j, i.e., sets of values for the test statistics T_j that lead to the decisions to reject the null hypotheses H_{0j} if $T_j \in Rt_j$. Suppose that m_0 null hypotheses are true and R is the number of hypotheses rejected. Let U and V be the numbers of the true null hypotheses respectively accepted and rejected and let T and S be the numbers of the non true null hypothesis respectively accepted and refused. This situation is summarized in Table 2.1.

While R is an observable random variable, U, V, S and T are not.

	ACCEPTED	REJECTED	TOTAL
TRUE NULL HP	U	V	m_0
FALSE NULL HP	T	S	$m - m_0$
TOTAL	$m - R$	R	m

Table 2.1: Number of errors testing m null hypotheses

2.2 Type I Error Rates

When many statistical tests are conducted simultaneously, the probability of making a false discovery, grows with the number of statistical tests performed, and becomes much larger than the nominal value at which each test is performed. Usually Type I error measures involve the distribution of V, because the frequentist target is to minimize the number of false negative while rejecting the maximum number of hypothesis. One classical error measure is the Family Wise Error Rate ($FWER$) that is the probability of at least one Type I error: $Pr(V \geq 1)$. The control of $FWER$ is very conservative (see Hochberg and Tamhane (1987)). In some cases $FWER$ control is needed, for example when a conclusion from the individual inferences is likely to be erroneous when at least one of them is. In other cases it can be inappropriate, for example microarray analysis do not require a protection against even a single Type I error, so that a $FWER$ control is not justified. Other kind of measures are the per-comparison error rate (PCER), or expected proportion of Type I errors among the m tests, $PCER = \frac{E[V]}{m}$, and the per-family error rate (PFER), or expected number of Type I errors, $PFER = E[V]$. As we can see these are error rates are defined as parameters of the distribution of the Type I error rate V. In general procedures that control the $PFER$ are more conservative than those that control the $FWER$ or the $PCER$, in the sense that they lead to fewer rejections. At the same time procedures that control the $FWER$ are more conservative than those that control the $PCER$. Actually it can be easily shown that $PCER \leq FWER \leq PFER$ (the order is reversed with respect to the number of rejections). Benjamini and Hockberg (1995) suggested that in many multiplicity problems the number of erroneous rejections should be taken into account and not only the question wether any error was made. From this point of view they proposed, as new error measure, the expected proportion of errors among the rejected hypotheses: the False Discovery Rate (FDR). The proportion of null hypotheses that are erroneously rejected, among all the rejected hypotheses, is a random variable Q that express the proportion of errors committed by falsely rejecting null hypotheses. When $R = V + S = 0$, the random variable Q should be set to zero as no error of false rejection can be committed. Therefore if we define the False Discovery Proportion (FDP) as

$$FDP = \begin{cases} Q = \frac{V}{R} & \text{if R} > \text{0,} \\ 0 & \text{if R = 0} \end{cases}$$

the FDR can be defined as the expected value of Q:

$$FDR = E(FDP) = E\left(\frac{V}{R}|R > 0\right) Pr(R > 0). \tag{2.2.1}$$

FDR has become a popular tool for controlling the error in microarray analysis. In fact the purpose of this kind of analysis is to individuate genes that are potential candidate for further investigation. Thus few erroneous rejections will not distort the conclusions at this stage of the analysis, as long as their proportion is small.

Note that the control of FDR is implicitly a control of $FWER$ when all the null hypothesis are true, in fact when all the null hypotheses are true $m = m_0$, $V = R$, $FDP = 1$ and consequently $FWER = Pr(V \geq 1) = Pr(R > 0) = FDR$. The other way round, it's easy to see that when $m_0 < m$ any procedure that controls the $FWER$ also controls the

FDR. As we will see the controls of Type I error when $m = m_0$ and $m_0 < m$ are referred to as *weak* and *strong* control. Storey (2001) suggested to control another quantity called Positive False Discovery Rate ($pFDR$)

$$pFDR = E\left(\frac{V}{R}|R > 0\right). \tag{2.2.2}$$

The term *positive* describes the fact that we have conditioned on at least one positive finding having occurred. We argue that when $m_0 = m$, one would want the false discovery rate to be 1, and that one is not interested in cases where no test is significant. These considerations lead Storey to propose definition (2.2.2) as an error rate alternative to (2.2.1). In Storey (2003) the $pFDR$ is used to define the q-value, which is a natural Bayesian version of the $pFDR$ analogue to the p-value.

2.2.1 Strong and Weak control

It is important to note that the error rates described above depend upon which specific subset of null hypotheses is true for the (unknown) data generating distribution. A very important distinction is that between strong and weak control of Type I error rate. This distinction is pointed out in Westfal and Young (1993). Strong control relates to the control of Type I error under any combination of true and false hypotheses. Weak control relates to control of the Type I error rate when all the null hypotheses are true. Note that the concept of strong and weak control applies to any of the Type I error rated defined above.

2.3 Error Controlling Procedures

Consider m independent statistics $T_1, T_2, \ldots, T_m$ for the null hypotheses $H_{01}, H_{02}, \ldots, H_{0m}$. We define the j-th p-value to be

$$p_j = Pr(|T_j| > |t_j| \ | \ H_{0j} \ is \ true), \ j = 1, \ldots, m$$

where t_j is the observed value of the test statistic T_j. It is well known that each p-value is a random variable uniformly distributed in $[0, 1]$ under any simple null hypothesis. Genovese and Wesserman (2002) define a marginal distribution for the p-values when the null hypotheses are composite. We will deal only with the case of simple null hypotheses.

A multiple testing procedure (MTP) aims to produce a set of rejected hypothesis $Rt(T_1, ..., T_m, \alpha) = \{j \in \{1, \ldots, m\} \ : \ H_{0j} \ \ is \ \ rejected\}$ to estimate the set of the false null hypothesis. The set Rt will depend on the data through the test statistics T_j and on the level α fixed as upper bound for a suitably defined Type I error measure. In this dissertation the dependence on the data will be carried out through the p-values. Usually each single hypothesis j is rejected if $p_j \leq Tr$ where Tr is a data dependent critical value (cut-off) value, and the different MTP techniques specify a way to determine a well set cut off value. Some authors, see as example Dudoit *et al.* (2003), prefer instead to define a new kind of p-values, called ***adjusted*** p-values, that are function of the classical concept of p-value, and leave the cut off fixed to a certain α. It can be shown that consider an adjustment of the critical value is equivalent to consider an adjustment of the p-value for any given MTP.

The MTP are categorized as single-step procedures and stepwise procedures. In the single step approach the p-values are compared to a predetermined cut off level that is a function of the level α and of the number of hypotheses m. As for the stepwise procedures, there are two different approaches: the step-down and the step-up. The step-down (Holm, 1979) starts with the most significant hypothesis and, as soon as one fails to reject a null hypothesis, no further rejections are made. The step-up procedure (Hochberg, 1988) starts with the least significant hypotheses and, as soon as one rejects a null hypothesis, rejects all the hypotheses that are more significant. More explicitly, consider the ordered p-values: $p_{(1)}, \ldots, p_{(m)}$. The step-down procedures start examining $p_{(1)}$ and continue rejecting until the first acceptance, while the step-up procedures starts with $p_{(m)}$ and continue accepting until the first rejection. It follows that the step-down procedures are more conservative than the step-up

2.3.1 $FWER$ and FDR Controlling Procedures

The most classical example of $FWER$ controlling procedure is the Bonferroni correction which is a single-step procedure fixing a universal critical value $Tr = \frac{\alpha}{m}$. This means that one would reject only the hypotheses for which the correspondent p-value is less than $\frac{\alpha}{m}$. Other $FWER$ controls are the one-step and the step-down proposed in Sidak (1967 and 1971), the step-down Holm (Holm, 1979), the step-up in Hochberg and Benjamini (1990), and the step-down $minP$ (Van der Laan *et al.*, 2003).

The FDR defined in (2.2.1) is an error measure that provides less strict control on the number of false positives so that the FDR controlling procedures has a gain in power with respect to the $FWER$ ones. The Benjamini-Hochberg (BH) is the most classical FDR controlling procedure, it consists in a step-up procedure that can be defined as follows:

$$reject\ all\ H_{0(i)}\ s.t.\ i \leq k = argmax_{j=1,\ldots,m} \left\{ p_{(j)} \leq \frac{j}{m}\, q \right\}$$

where $p_{(1)}, \ldots, p_{(m)}$ are the ordered p-values, $H_{0(i)}$ is the null hypothesis corresponding to $p_{(i)}$ and q is the chosen FDR upper bound. Benjamini and Hochberg (1995) showed that the above procedure controls FDR at level q. Benjamini and Yekutieli (2001) proved that the BH procedure controls FDR at level $\frac{m_0}{m}\,q$ when the number of true null hypotheses m_0 is smaller than m and the test statistics are continuous. Consequently if we knew m_0, the BH procedure could be improved by using as controlling level $q' = \frac{m}{m_0}\,q$. In practice m_0 is unknown and many adaptive procedures which estimate this factor have been constructed; see Efron et al. (2001), Storey (2002) and Benjamini and Yekutieli (2003) for a complete review.

2.4 Bootstrap estimation of the null distribution

In principle we may make parametric assumption on the joint distribution of the test statistics but these assumptions would seldom be reliable. In many practical situations the joint and marginal distributions of the test statistics are unknown and so the true joint distribution G, for the test statistics $T_j, j = 1, \ldots, m$, is estimated by a null joint distribution G_0 in order to derive the resulting p-values. The choice of a joint null distribution G_0 is crucial to ensure that the control of the Type I error rate actually provides the required control under the true distribution G. Resampling methods such as bootstrap and permutation are used to estimate G_0. The name bootstrap alludes to pulling yourself up by your own boot strap. In statistics Bootstrapping is a method for estimating the sampling distribution of interest by resampling with replacement from the original sample of data. This means that one available sample gives rise to many others by resampling. Bootstrap technique was invented by Bradley Efron (1979) and further developed by Efron and Tibshirani (1993). We describe in the followings a generic bootstrap estimation procedure of the null distribution . The first step is to generate B bootstrap samples starting from the data. Then for each bootstrap sample, compute an m vector of test statistics, $\hat{T}(.,b) = (\hat{T}(j,b) : j = 1,...,m)$, which can be arranged in an (m, B) matrix, $\hat{T}$, with rows corresponding to the m hypotheses and columns to the B bootstrap samples.The bootstrap estimate of the joint null distribution G_0 is the empirical distribution of the columns $\hat{T}(.,b)$ of the matrix $\hat{T}$. As example for two sided alternative hypotheses, the empirical p-value for the hypothesis H_j is

$$\hat{p}_j = \frac{\sum_{b=1}^{B} I(|\hat{T}(j,b)| \geq |t_j|)}{B}$$

For a discussion of resampling based methods see Pollard and Van der Laan (2003).

2.5 Bayesian testing

Given a model on an observable variable X

$$X \mid \theta \ \sim \ f(X|\,\theta), \ \ \theta \in \mathbb{R}^p,$$

consider again the general task of deciding between a null hypothesis

$$H_0 \ : \ \theta \in \Theta_0 \subseteq \mathbb{R}^d$$

and an alternative hypothesis

$$H_1 \ : \ \theta \in \Theta_1 \subseteq \mathbb{R}^d$$

where $\Theta_0 \bigcap \Theta_1 = \emptyset$. Using a Bayesian formulation, θ is considered to be a random variable described by a prior distribution

$$\theta \ \sim \ \pi(\theta).$$

The hypothesis testing problem thus reduces to determining the posterior probabilities

$$\alpha_0 = P(\Theta_0 \mid x) \quad \text{and} \quad \alpha_1 = P(\Theta_1 \mid x) \ ,$$

and deciding between H_0 and H_1 accordingly. More explicitly, if we use π_0 and π_1 to denote the prior probabilities of Θ_0 and Θ_1 respectively, the posterior probabilities of the hypotheses can be rewritten as

$$\alpha_0 = \frac{\pi_0 f(X|H_0)}{f(X)} \quad \text{and} \quad \alpha_1 = \frac{\pi_1 f(X|H_1)}{f(X)}.$$

One of the main objections made to Bayesian methods is the subjective input required in specifying the prior $\pi(\theta)$. In order to relate the information given by the data to the prior confidence on the two hypotheses, often *Bayes Factor* is computed. The Bayes Factor (BF) in favor of Θ_0 is defined as

$$BF = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}.$$

The quantities π_0/π_1 and α_0/α_1 are respectively named prior and posterior odds in favor of H_0. It ca be easily shown that

$$BF = \frac{f(X|H_0)}{f(X|H_1)}.$$

This last definition shows that the BF can be interpreted as the likelyhood ratio of H_0 and H_1, similar in spirt to the frequentist approach. Intuitively a BF greater than one reveals evidence in favor of H_0. More rigorously, let a_0 and a_1 be the actions denoting respectively to accept H_0 and H_1, and let $L(\theta, a_i), i = 0, 1$ be the corresponding losses. The decision problem is solved in terms of a $Bayes\ action$ that is the action that minimize the expected losses $E^{\pi(\theta|x)}[L(\theta, a_i)]$, under the posterior distribution $\pi(\theta|x),\ (i = 0, 1)$. It can be easily shown that under the $0-1$ loss, the Bayes decision is the hypothesis with the larger posterior probability α_i, $(i = 0, 1)$, that is

$$H_0\ is\ rejected \iff \frac{\alpha_0}{\alpha_1} < 1 \iff BF < \frac{\pi_1}{\pi_0}.$$

Under a more general $0 - L_i, (i = 0, 1)$ loss, this result can be extended to

$$H_0 \, is \, rejected \iff BF < \frac{L_0}{L_1}\frac{\pi_1}{\pi_0}.$$

2.6 Bayesian multiple hypothesis testing

In the first part of this Chapter, we tried to explain some of the difficulties of testing multiple hypotheses. Unfortunately in a standard Bayesian perspective, the multiplicity effect is ignored (see Berger, 1985). Indeed it is possible to show that under additive loss and independent priors, one simply computes the Bayes factor for each single test and applies the decision rule described in Section 2.5 independently. Of course the number of false positives would be large if we are testing many hypotheses simultaneously. Recently Sakar and Chen (2004) propose a new method to account for multeplicity within a Bayesian setup. Suppose we have m independent vectors of observations $X_j, \;\; j = 1, \ldots, m$, of size n and each X_j distribution, $\;f_j(X_j \mid \theta_j)$, depends on a vector of parameters $\theta_j \in \Omega_j \subseteq \mathbb{R}^{d_j}$ and consider the general multiple hypotheses testing problem (2.1.1). The key idea is to consider the θ_i dependent from each other a priori. This allows the posterior distributions of θ_i to depend on all the X_i. Here we consider a simpler solution proposed from Abramovich and Angelini (2005). Where a hierarchical prior model is obtained by eliciting a prior distribution on the number of the false null hypotheses. Then the most likely configuration of true and false hypotheses is chosen using the MAP rule. We will adapt this proposal to a specific problem and will describe a frequentist procedure based on this idea.

2.6.1 MAP multple testing procedure

The first need in a Bayesian multiple testing procedure, is to give a prior odds in favor of each single null hypothesis. In most of the cases, we might have an idea about the number k of false null hypotheses and thus we could impose a prior distribution $\pi(k)$ on it. Let the configuration of true and false hypotheses be determined by a m-dimensional vector $(y_1, \ldots, y_m)$ where

$$y_i = \begin{cases} 1 & \text{if } \theta_i \in \Theta_{1i} \\ 0 & \text{if } \theta_i \in \Theta_{0i} \end{cases} \qquad i = 1, \ldots, m\ . \tag{2.6.1}$$

Let $k = y_1 + \ldots + y_m$ be the number of false null hypotheses and elicit a prior $\pi(k)$ on it. It is reasonable that given k, the $\binom{m}{k}$ possible configurations of **y** are equally likely a priori, thus

$$P(\mathbf{y} \mid \sum_{i=1}^{m} y_i = k) \ = \ \binom{n}{k}^{-1} . \tag{2.6.2}$$

Furthermore assume

$$(\theta_i \mid y_i \ = \ 0) \sim p_{0i}(\theta_i) \ \ and \ \ (\theta_i \mid y_i \ = \ 1) \sim p_{1i}(\theta_i), \tag{2.6.3}$$

where $p_{0i}(\theta_i)$ and $p_{1i}(\theta_i)$ are densities respectively on Θ_{0i} and Θ_{1i}.

The posterior is then

$$\pi(\mathbf{y}, k, |X_1, \ldots, X_m) \propto \binom{n}{k} \pi(k) I\{\sum_{i=1}^{m} y_i = k\} \prod_{i=1}^{n} (B_i^{-1})^{y_i} \tag{2.6.4}$$

where B_i is the Bayes factor in favor of H_{0i}. The maximization of the $\pi(\mathbf{y}, k, |\mathbf{X})$ may be very expansive as in the general case we should maximize over the 2^m possible configurations of true and null hypotheses. However, for model (2.6.4), the number of possible configurations to be considered in the maximization reduces to $m+1$. Indeed, for each given $k \in \{0,1,\ldots,m\}$, the obvious maximizer of (2.6.4) is $\hat{\mathbf{y}}(k)$ such that

$$\hat{y}_i = \begin{cases} 0 & \text{if } B_i \text{ is one of the } k \text{ smallest Bayes factors },\\ 1 & \text{otherwise}. \end{cases}$$

Hence the Bayesian MAP multiple testing procedure can be summarized in three steps:

1. Compute the m Bayes factors B_i in favor of each single null hypothesis $H_{0i}, i = 1,\ldots,m$ and order them from the smallest to the largest as $B_{(1)},\ldots,B_{(m)}$.

2. Find the $\hat{k}$ that maximizes $\hat{\pi}_k = \pi(\hat{\mathbf{y}}(k), k|X_1,\ldots,X_n) \propto \binom{n}{k}^{-1}\pi(k)\prod_{i=0}^{k} B_{(i)}^{-1}$.

3. Reject all null hypotheses corresponding to $B_{(1)},\ldots,B_{(\hat{k})}$ and accept others.

In order to reduce the computational cost, step-down and step-up versions of the described MAP procedure can be implemented. The step-down procedure consists in starting with the most significant hypothesis, corresponding to the smaller Bayes Factor $B_{(1)}$, and reject the null hypotheses as long as

$$\frac{\hat{\pi}_k}{\hat{\pi}_{k-1}} > 1,$$

thus as long as

$$B_{(k)} < \frac{k}{n-k+1}\frac{\pi(k)}{\pi(k-1)}. \tag{2.6.5}$$

The step-up procedure, starts with the less significative hypothesis, corresponding to $B_{(n)}$, and accept the null hypotheses until the (2.6.5) holds. Evidently the step-up and step-down procedures will lead to different solutions only if the sequence $\{\hat{\pi}_k\}_{k=0,\ldots,n}$ presents local maxima. Furthermore, equation (2.6.5) reveals that different priors $\pi(k)$ will lead to different decisions.

2.6.2 Connections with frequentist procedures and choice of the priors

Throughout this chapter we reviewed frequentist and Bayesian guide lines to approach the multiple hypothesis testing problem. It is clear that frequentist tests procedures are based on p-values and Bayesian tests procedures are based on BF. In general there is no connection between these two values unless we consider a specific model. Suppose each X_i has a symmetric location distribution

$$X_i \sim f_i(|x_i - \theta_i|), i = 1, \ldots, m \ \ ,$$

consider one sided simultaneous tests

$$H_{0i} : \theta_i \leq \theta_{0i} \ \ vs \ \ H_{1i} : \theta_i > \theta_{0i} \quad i = 1, \ldots, m$$

and assume non informative priors on θ

$$p_{0i} = 1_{(-\infty, \theta_{0i})} \ \ and \ \ p_{1i} = 1_{(\theta_{0i}, +\infty)}.$$

Then we have

$$B_i = \frac{p_i}{1 - p_i}, i = 1, \ldots, m.$$

In particular we observe that constraint (2.6.5) on the BF reveals in the following constraint on the p-values

$$p_{(i)} < \frac{c_i}{1 + c_i} \ \text{ where } \ ci = \frac{i}{n - i + 1} \frac{\pi(i)}{\pi(i-1)}.$$

A particular choice of the prior distribution $\pi(k)$, could lead to mimic frequentist step-wise procedures. As example, it can be shown that choosing $k \sim B(m, \alpha_m)$ yields

$$B_{(i)} < \frac{\alpha_m}{1 - \alpha_m} \Longleftrightarrow p_{(i)} < \alpha_m, \ .$$

Thus setting $\alpha_m = \alpha/m$ and the mean $m\alpha_m = \alpha$ smaller than 1, the resulting procedure mimics the Bonferroni correction with significance level α.

As a further example consider the truncated geometric prior $G^*_m(1-q)$ and set $\pi(k) = (1-q)q^k/(1-q^{m+1})$. It can be shown that in this case we get

$$B_{(i)} < \frac{q_i}{m-i+1} \Longleftrightarrow p_{(i)} < \frac{i}{m-i(1-q)+1}.$$

The obtained constraint on the p-value coincide with the critical values of the adaptive step-down procedure proposed by Benjamini *et al*, 2004.

Note that in both the examples provided the choice of the priors reflect a *sparsity* assumption, that is we are assuming only a small fraction of alternative hypotheses are true. This assumption can be plausible in many cases, as example it fits the microarray experiments, for which we believe *a priori* that only a few number of genes is differentially expressed.

2.6.3 Custom normal model

One of the advantages of Bayesian over frequentist tests is that in a Bayesian context is always possible to exchange the role of the null with alternative hypothesis. It is well known that for simple hypothesis testing

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta = \theta_1 \ ,$$

or for one sided hypothesis testing

$$H_0 : \theta \leq \theta_0 \quad vs \quad H_1 : \theta > \theta_0 \ ,$$

it is always possible to test H_1 against H_0 both in a frequentist and a Bayesian framework. On the other hand, if we consider the testing problem

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta \neq \theta_0 \tag{2.6.6}$$

it is not possible to exchange H_0 with H_1 in a classic frequentist test, but it is feasible in the Bayesian context. Of course in this case we need that the prior distribution on the parameter on interest θ is discontinuous at least in θ_0 in order to have the prior, and hence the posterior, probability that $\theta = \theta_0$ greater than zero.

In the light of the idea in Abramovich and Angelini (2005), summarized in the previous subsection, we consider now a specific problem. Suppose we have m independent random samples $X_i, i = 1, \ldots, m$ of size n and that each X_i is from a normal population $N(\mu_i, \sigma^2)$ with unknown mean μ_i and known variance σ^2. Consider the multiple testing problem

$$H_{0i} : \ \mu_i \neq \mu_{0i} \quad vs \quad H_{1i} : \ \mu_i = \mu_{0i} \quad i = 1, \ldots, m. \tag{2.6.7}$$

We assume that

$$\pi_i(\ \mu_i\ ) = \pi_{0i} p_{0i}(\mu_i) + (\ 1\ -\ \pi_{0i}\ )\delta_{0i}(\mu_i), \quad \pi_{0i} > 0, \quad i = 1, \ldots, m$$

where δ_{0i} is a point mass at μ_{i0} and the non zero part $p_{0i}(\mu_i)$ is a generic density function. Setting $p_{0i}(\mu_i) = N(\mu_{0i}, \tau^2)$ yields to the Bayes factor

$$B_i = \frac{1}{\sqrt{1+\gamma}} \exp\left\{ \frac{Z_i^2}{2(1+1/\gamma)} \right\} \tag{2.6.8}$$

where $Z_i = \sqrt{n}(\bar{X} - \mu_{0i})/\sigma$ and $\gamma = n\tau^2/\sigma^2$ (see Berger, 1985).

Note that equation (2.6.8) allows to express the statistic Z in close form as a function of the BF, i.e.

$$|Z_i| = \left[2\left(1 + \frac{1}{\gamma}\right) (\log(B_i) + \log(\sqrt{1+\gamma}) \right] \quad i = 1, \ldots, m. \tag{2.6.9}$$

Recall now that in the Bayesian MAP multiple testing procedure, the decision whether or not to reject a null hypothesis H_{0i} is carried out by comparing the corresponding Bayes factor B_i to a specific critical value $c_i = \frac{i}{n-i+1}\frac{\pi(i)}{\pi(i-1)}$. From the (2.6.9) we get

$$B(i) < c_i \iff |Z_i| < \hat{c}_i = \left[2\left(1+\frac{1}{\gamma}\right)(\log(c_i) + \log(\sqrt{1+\gamma})\right] \quad i = 1, \ldots, m.$$

This last equation gives the insight to deal with the model (2.6.7) in an ”empirical” framework. In fact, given the threshold c_i and the testing statistic $|Z_i|$, we could reject the null hypothesis H_{0i} in (2.6.7) if the value of the statistic is more extreme than c_i. Moreover the threshold choice is connected to the the prior choice trough a the close formula

$$\hat{\lambda}_i = \left[2\left(1+\frac{1}{\gamma}\right)\left(\log\left(\frac{i}{n-i+1}\frac{\pi(i)}{\pi(i-1)}\right) + \log(\sqrt{1+\gamma}\right)\right] \quad i = 1, \ldots, m,$$

thus different cut off on the Z statistic can be regarded as different prior distributions $\pi(k)$.

2.6.4 Possible extension and future work

The Bayesian MAP procedure described in section 2.6.1 can be used regardless the distribution of X_i and θ_i. Hence testing problem (2.6.7) can be analyzed under different combination of prior $p_{0i}(\mu_i)$ and error model $f_i(X_i|\mu_i)$. In particular Johnstone and Silverman (2005) consider in a different context the combination of normal error with heavier tailed prior on μ_i. Similarly Pensky and Sapatinas (2005) consider different models. Their results are preliminary to the applications of Bayesian multiple hypothesis testing for several type of real data applications and provide an interesting starting point for future research.

Chapter 3

Wavelet Filtering of Noisy Signals

Introduction

In this Chapter we first provide a smart introduction to wavelet and thresholding methods, then we explore the thresholding rules, in the wavelet domain, induced by a variation of the Bayesian *Maximum A Posteriori* (MAP) principle. The proposed rule is thresholding and always picks the mode of the posterior larger in absolute value, thus the name *Larger Posterior Mode* (LPM). We demonstrate that the introduced shrinkage performs comparably to several popular shrinkage techniques. The exact risk properties of the thresholding rule are explored, as well. The chapter is organized as follows. The first two sections are a short review of the basic mathematical background behind wavelets theory and some of their statistical property. In Section 3 wavelet thresholding rules are presented and in Section 4 Bayes rules in he wavelet domain are introduced. In Section 5 a basic Bayesian model is described, the LPM rule is derived, and the exact risk properties of the LPM rule are discussed. Section 6 discusses two models that generalize the model from Section 5 by relaxing the assumption of known variance. Derivations of LPM rules corresponding to these two models are deferred to the Appendix. Comprehensive simulations and comparisons are provided in Chapter 6 which also contains discussion on the selection of hyperparameters and a real-life application of the introduced shrinkage. We conclude the chapter with an outline on some possible directions for future research.

The work described in the present Chapter was done during a visiting period to the Georgia Institute of Technology. The results achieved have been submitted to an international journal and are now under review.

3.1 Mathematical background

The word wavelet is due to Morlet and Grossman in the early 1980s. They used the French word *ondelette* meaning *small wave*. The key idea of wavelets is to express functions or signals as sums of these little waves and of their translations and dilations. Wavelets play the role of sines and cosines in ordinary Fourier series. The idea of approximation using superposition of functions has in fact existed since the early 1800's, when Joseph Fourier discovered that sines and cosines could be used to represent other functions. We will not attempt a full review neither of the wide field of non parametric function approximation nor of wavelet theory but we will just give some basic concepts. According to the definition given in 1980 by Grossman and Morlet, a physicist and an engineer, a wavelet is a function $\Psi \in L^2(\mathbb{R})$ with zero mean

$$\int \Psi(t)dt = 0$$

that satisfies the so called *admissibility condition*

$$\int_0^{-\infty} \frac{\hat{\Psi}(\omega)^2}{\omega} d\omega < +\infty$$

with $\hat{\Psi}(\omega)$ denoting the fourier transform of Ψ. These properties ensure that wavelets are bounded functions with a fast decay to zero. Starting from the basic wavelet Ψ, called the *mother wavelet*, a whole family of wavelets $\{\Psi_{ab} = a^{-1/2}\Psi(\frac{t-b}{a})\ ,\ a > 0, b \in \mathbb{R}\}$ can be generated by dilation and translation in time. In the following we will deal with discrete wavelets thus the parameters a and b are restricted to a discrete set, usually $a = 2^{-j}$ and $b = ka = 2^{-j}$, where j and k are integers. Under suitable assumption on Ψ(see Vidakovic, 1999) an orthonormal basis of $L_2(\mathbb{R})$ is constituted by the set

$$\{\Psi_{jk}(t) = 2^{j/2}\Psi(2^j t - k), j, k \in \mathbb{Z}\} \tag{3.1.1}$$

so that any function $f \in L_2(\mathbb{R})$ can be expressed as

$$f = \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} \beta_{jk}\Psi_{jk}$$

where the wavelet coefficients β_{jk} are given by

$$\beta_{jk} = \int f(t)\Psi_{jk}(t)dt \quad \forall j, k \in \mathbb{Z}. \tag{3.1.2}$$

The crucial feature of wavelet theory is the concept of Mallat's *multiresolution analysis* (Mallat, 1989). In the multiresolution analysis framework, we see that details or fluctuations at different levels of resolution are represented by the superposition of wavelets associated with the appropriate dilation. One of the most considerable advantage is ability to zoom in on details in order to visualize complex data. On the other hand, details can also be suppressed easily and thus wavelets can be used for data smoothing. More explicitly a multiresolution analysis of $L_2(\mathbb{R})$ is a nested sequence of its closed subspaces $\{V_j\}_{j\in\mathbb{Z}}$ such that

- $V_j \subset V_{j+1} \quad \forall j \in \mathbb{Z}$,

- $\bigcap_{j\in\mathbb{Z}} V_j = \{0\} \quad and \quad \overline{\bigcup_{j\in\mathbb{Z}} V_j} = L_2(\mathbb{R}),$
- $f(t) \in V_0 \Leftrightarrow f(t+k) \in V_0 \quad \forall k \in \mathbb{Z}$,
- $f(2^j t) \in V_j \Leftrightarrow f(t) \in V_0 \quad \forall j \in \mathbb{Z}$
- $There\ exists\ a\ function\ \varphi \in V_0,$ *called the* scaling function *or* father wavelet, *with mean* 1 *and such that* $\{\varphi(t-k),\ k \in \mathbb{Z}\}$ *is an orthonormal basis of* V_0.

Because of the inclusion $V_0 \subset V_1$, the function φ can be represented as a linear combination of functions from V_1, i.e.,

$$\varphi(t) = \sum_{k\in\mathbb{Z}} h_k \sqrt{2}\varphi(2t-k) \tag{3.1.3}$$

for some coefficients h_k ,$k \in \mathbb{Z}$. Equation (3.1.3) called the *scaling equation* (or *two-scale equation*) is fundamental in constructing, exploring, and utilizing wavelets.

Given a multiresolution analysis of $L_2(\mathbb{R})$, for each of the subspaces V_j we can define its detail space W_j that is the orthogonal complement of V_j in V_{j+1} thus

$$\begin{aligned} V_{j+1} &= V_j \bigoplus W_j \ \forall j \in \mathbb{Z}, \\ W_j &\perp V_j \ \forall j \in \mathbb{Z}, \\ L_2(\mathbb{R}) &= \bigoplus_{j\in\mathbb{Z}} W_j = V_J \bigoplus_{j\geq J} W_j, \end{aligned}$$

and it can be shown (Daubechies, 1992) that there exists a wavelet function Ψ s.t. the set $\{\Psi_{jk}(t) = 2^{j/2}\Psi(2^j t - k), k \in \mathbb{Z}\}$ constitutes an orthonormal basis for W_j, for every $j \in \mathbb{Z}$. The wavelet function Ψ can be derived from the scaling function φ. Since $\Psi(t) = \Psi_{00}(t) \in W_0 \subset V_1$, it can be represented as

$$\Psi(t) = \sum_{k\in\mathbb{Z}} g_k \sqrt{2}\varphi(2t-k) \tag{3.1.4}$$

for some g_k, $k \in \mathbb{Z}$. The coefficient h_k in (3.1.3) and g_k in (3.1.4) are known respectively as low pass and hight pass filters and it can be shown that they can uniquely define a multiresolution analysis.

Suppose we want to approximate f with its projection

$$P_N[f(t)] = f_N(t) = \sum_{k\in\mathbb{Z}} \alpha_{Nk}\varphi_{Nk}(t) \tag{3.1.5}$$

in the space V_N, for a given level N. It can be shown that

$$f(t) = \lim_{N \rightarrowtail \infty} P_N[f(t)] = \lim_{N \rightarrowtail \infty} \sum_{k\in\mathbb{Z}} \alpha_{N,k}\varphi_{N,k}(t)$$

where the convergence is in the L_2 space.

We want now underline that, for any resolution level $J \in \mathbb{Z}$, the exposed properties of the multiresolution analysis enables to express the approximation function $f_N = \sum_{k\in\mathbb{Z}} \alpha_{Nk}\varphi_{Nk}$ as

$$f_N = f_J + \sum_{j=J}^{N-1} d_j = \sum_{k\in\mathbb{Z}} \alpha_{Jk}\varphi_{Jk} + \sum_{j=J}^{N-1}\sum_{k\in\mathbb{Z}} \beta_{jk}\Psi_{jk} \tag{3.1.6}$$

where $\alpha_{jk} = \int f(t)\varphi_{jk}(t)dt$ are the *scaling* coefficients and β_{jk} are the wavelet coefficients (3.1.2). The first part of the expansion (3.1.6) is the coarse representation of f in V_J, that is the projection of the function f in the space V_J, and the second part is the projection of f in the remaining details spaces. It is easy to see that the direct calculation of the wavelet expansion (3.1.6) is computationally expansive and moreover the scaling and the wavelet functions could not have an analytical close form. Therefore the procedure of wavelet estimation is based on a fast algorithm introduced by Mallat to perform the discrete wavelet transform (DWT). This algorithm relates the wavelet coefficients from different levels with the wavelet filters. Let's summarize the basic idea of this algorithm. From equations (3.1.6) and (3.1.5) it comes that

$$f_N \text{ is known} \Longleftrightarrow (\alpha_{Nk}) \text{ are known } \Longleftrightarrow (\alpha_{Jk}, \beta_{jk}) \text{ are known}$$

$\forall k \in \mathbb{Z}$ and $\forall j \in \{J, \ldots, N-1\}$. The process of calculating the wavelet transform starting from the the approximation domain is called *Forward Wavelet Transform* (FWT). It is carried out trough the recursive application of an algorithm that starting from the scaling coefficients of a generic level j, turns out both the scaling and wavelet coefficients at the lower resolution level $j-1$. The recursion formulae for the FWT are:

$$\begin{aligned}
\alpha_{j-1,k} &= \langle f, \varphi_{j-1,k} \rangle = \sum_{l \in \mathbb{Z}} h_{l-2k} \alpha_{j,l} \;\; \forall k \in \mathbb{Z}, \\
\beta_{j-1,k} &= \langle f, \Psi_{j-1,k} \rangle = \sum_{l \in \mathbb{Z}} g_{l-2k} \alpha_{j,l} \;\; \forall k \in \mathbb{Z}
\end{aligned}$$

On the other hand the process of reconstructing the approximation function given the wavelet transform is called *Inverse Wavelet Transform* (IWT). It is carried out trough the recursive application of an algorithm that starting from the scaling and wavelet coefficients at the generic level $j-1$, turns out both the scaling coefficients at the higher resolution level j. The recursion formulae are:

$$\alpha_{j,k} = \langle f, \varphi_{j,k} \rangle = \sum_{l \in \mathbb{Z}} h_{l-2k} \alpha_{j-1,l} + \sum_{l \in \mathbb{Z}} g_{l-2k} \alpha_{j-1,l} \;\; \forall k \in \mathbb{Z}.$$

In statistical settings we are more usually concerned with discretely sampled, rather then continuous, functions. The extension to the discrete case is straightforward (Vidakovic, 1999).

Discrete wavelet transforms are applied to discrete data sets and produce discrete outputs and map data from the time domain (the original input data vector) to the wavelet domain. The result is a vector of the same size. Wavelet transforms are linear and they can be defined by matrices of dimension nxn if they are applied to inputs of size n, level by level. With the proper boundary conditions, such matrices are orthogonal and the corresponding transform is a rotation in $\mathbb{R}^n$. The coordinates of the point in the rotated space comprise the discrete wavelet transform of the original coordinates. More explicitly, given a vector of observed data **y** $= (y_1, \ldots, y_n)^t = (y(t_1), \ldots, y(t_n))^t$, at equally spaced points t_i, the *Discrete Wavelet Transform* (DWT) of **y** is given by

$$\mathbf{d} = W\mathbf{y} \tag{3.1.7}$$

where W is the nxn DWT matrix associated with the orthonormal wavelet basis chosen and **d** is an nx1 vector comprising both discrete scaling and the discrete wavelet coefficients. Note that, by orthogonality of W, the inverse transform ($IDWT$) is simply given by

$$\mathbf{y} = W^t\mathbf{d}.$$

Under the assumption $n = 2^J$ for some integer J, the DWT and the $IDWT$ may be performed using Mallat's fast algorithm ($O(n)$ operations). In this case for a given j_0 and under periodic boundary conditions, the n-dimensional vector **d** consists of the discrete scaling coefficients $c_{j_0,k}$, $k = 1, \ldots, 2^{j_0} - 1$ and the discrete wavelet coefficients d_{jk}, $j = j_0, \ldots, J-1$, $k = 0, \ldots, 2^J - 1$. We refer to Mallat (1989) for a full description.

3.2 Advantages of wavelets

Wavelet transforms are now being adopted for a vast number of different applications often replacing the conventional Fourier transform. Many areas of physics have seen this paradigm shift, including molecular dynamics, astrophysics, density-matrix localization, seismic geophysics, optics, turbulence and quantum mechanics. Other areas seeing this change have been image processing, blood-pressure, heart-rate and ECG analysis, DNA analysis, protein analysis, climatology, general signal processing, speech recognition, computer graphics and multifractal analysis. In contrast with standard Fourier sine and cosine series, wavelets are local both in scale (frequency), via dilatation, and in time, via translation. This localization allows a parsimonious (*sparse*) representations of different functions in the wavelet domain, i.e. the energy of the transformed signal is concentrated in few wavelet coefficients. This property enables the localization of events and singularity of the signal under study; furthermore this property also implies that by choosing a sufficiently regular mother wavelet (Cohen *et al.* 1993), the wavelet system constitutes an unconditional base for a wide set of function spaces, such as Besov spaces. It can be shown that the Fourier and Wavelet linear approximation are asymptotically equivalent for homogeneously regular functions. This is the case of functions belonging to Besov spaces $B^s_{p,q}$ where $p \geq 2$, as Holder and Sobolev spaces. Advantages of wavelets are more evident when non homogeneously regular functions have to be approximated, in fact it can be shown that in this case the best non linear wavelet approximation is asymptosically optimal and similar results cannot be achieved via Fourier series. Examples of non homogeneous classes are Besov spaces $B^s_{p,q}$ where $1 \leq p < 2$. This result is due to the fact that wavelet bases can characterize a much wider range of spaces than Fourier bases. We may conclude stating that the main advantage of wavelet basis is their *universality* , in the sense that functions from a wide set of function spaces have a parsimonious representation in wavelet series and fast algorithms are available.

3.3 The statistical problem

One statistical task in which the wavelets are successfully applied is recovery of an unknown signal **f** imbedded in Gaussian noise η. In practice, given a vector of observed data **y** $= (y_1, \ldots, y_n)^t = (f(t_1), \ldots, f(t_n))^t$, at equally spaced points t_i, consider the standard non parametric regression problem

$$y_i = f_i + \eta_i, \quad i = 1, \ldots, n \tag{3.3.1}$$

where η_i are independent normal variables with zero mean and variance σ^2. We want to recover the unknown signal **f** from the observed noisy data **y** without assuming any parametric form. In the literature there are many approaches to non parametric estimation of the unknown signal **f** (e.g. kernel estimation, spline smoothing, Fourier series expansion and of course wavelet series expansion). In this Chapter we present a wavelet based estimator of **f**.

Given the noisy measurements **y** of model (3.3.1), let **d** be its discrete wavelet coefficients obtained by performing a DWT (3.1.7) on it. In the following we will assume $n = 2^J$, for some positive index J. This assumption, together with the hypothesis of equispaced observations, allow us to perform the fast Mallat algorithm (1989). The linearity of transformation W implies that the transformed vector $\mathbf{d} = W\mathbf{y}$ is the sum of the transformed signal $\theta = W\mathbf{f}$ and transformed noise $\epsilon = W\eta$. For the discrete scaling and wavelet coefficients holds the model

$$c_{j_0k} = \theta_{j_0k} + \epsilon_{j_0k}, \quad k = 1, \ldots, 2^{j_0} - 1 \tag{3.3.2}$$

$$d_{jk} = \theta_{jk} + \epsilon_{jk}, \quad j = j_0, \ldots, J-1, \quad k = 0, \ldots, 2^J - 1. \tag{3.3.3}$$

Due to orthogonality of W, the DWT of white noise is also a vector ϵ_{jk} of independent

$N(0, \sigma^2)$. Due to the decorrelation property of wavelet transforms, the coefficients are modelled individually and independently. White noise obviously contaminates each coefficient d_{jk} but, for the sparseness of the wavelet representation, we expect that only few large d_{jk} contains information about the signal, while all the rest are close to zero and can be attributed to the noise. In order to obtain an approximate wavelet representation of the underling signal **f**, we need to find a rule to decide which are the significant large wavelet coefficients, retain them and set all the others to zero. It is also reasonable to keep the scaling coefficients $c_{j_0 k}$ at the lower *coarse* level intact, as they are low frequency terms that may contain important components about the function **f**. The procedure to estimate the non zero wavelet coefficients and set to zero the others is called *thresholding*. In the exposition that follows the double index jk representing scale/shift indices is omitted and a typical wavelet coefficient, d, is considered.

There exist two kind of thresholding called *hard* and *soft* that can be summarized as follows

$$\begin{aligned} \hat{d} &= d\mathbf{1}_{|d| \geq \lambda}, \text{ is hard thresholding} \\ \hat{d} &= sign(d)(|d| - \lambda)\mathbf{1}_{|d| \geq \lambda}, \text{ is soft thresholding} \end{aligned}$$

for a certain *threshold* λ. The soft rule is also know as *shrinkage* as it clearly shrinks the wavelets coefficients. The *Universal threshold* T_U (Donoho and Johnnstone, 1994)

$$T_U = \sqrt{2 \log(n)}\sigma, \tag{3.3.4}$$

is one of the most common choices. This T_U has the property to be *global*, as it is identical for each resolution level j, and to be *universal*, in the sense that it does not depend on the underling function and has several optimal properties.

After the thresholding step the final estimator of the signal is reconstructed through the fast algorithm of the inverse DWT ($IDWT$).

3.4 Bayes rules and wavelets

The Bayesian paradigm has become very popular in wavelet-based data processing (Vidakovic, 1999). The Bayes rules allow the incorporation of prior information about the unknown and are usually shrinkers. For example, in location models the Bayesian estimator shrinks toward the prior mean (usually 0). This shrinkage property holds in general, although examples of Bayes rules that expand can be constructed, see Vidakovic and Ruggeri (1999). The Bayes rules can be constructed to mimic the traditional wavelet thresholding rules: to shrink the large coefficients slightly and shrink the small coefficients heavily. Furthermore, most practicable Bayes rules should be easily computed by simulation or expressed in a closed form.

Bayesian estimation is applied in the wavelet domain, i.e., after the data have been transformed. The wavelet coefficients can be modeled block-wise, as a single vector, or individually, due to the decorrelating property of wavelet transforms. In this Chapter we model wavelet coefficients individually, i.e., elicit a model on a typical wavelet coefficient. Recall that it is advisable to keep the scaling coefficients $\{c_{j_0k}\ k = 0, \ldots, 2^{j_0} - 1\}$ intact. Thus, according to (3.3.2), $\{\theta_{j_0k}\ k = 0, \ldots, 2^{j_0} - 1\}$ are assumed to be mutual independent random variables and vague priors are placed on them

$$c_{j_0k} \sim N(0, \nu), \quad \nu \rightarrowtail \infty. \tag{3.4.1}$$

The (3.3.2) and (3.4.2) yields the θ_{j_0k} are *a posteriori* conditionally independent

$$\theta_{j_0k}|c_{j_0k}, \sigma^2 \sim N(c_{j_0k}, \sigma^2) \quad k = 0, \ldots, 2^{j_0} - 1 \tag{3.4.2}$$

Consequently, through the use of Bayesian *Maximum A Posteriori* (MAP) rule, that we will discuss later in this section, the θ_{j_0k} would be estimated by the corresponding c_{j_0k}, (see Antoniadis and Sapatinas, 2001).
In the following we will model the wavelet coefficients d_{jk}. The double index jk representing scale/shift indices is omitted and a ”typical” wavelet coefficient, d, is considered.

Thus we concentrate on the model: $d = \theta + \epsilon$. Bayesian methods are applied to estimate the location parameter θ, which will be, in the sequel, retained as the shrunk wavelet coefficient and back transformed to the data domain. Various Bayesian models have been proposed in the literature. Some models have been driven by empirical justifications, others by pure mathematical considerations; some models lead to simple and explicit rules, others require extensive Markov Chain Monte Carlo simulations. Reviews of some early Bayesian approaches can be found in Abramovich and Sapatinas (1999), Vidakovic (1998, 1999) and Ruggeri and Vidakovic (2005). Müller and Vidakovic (1999) provide an edited volume on various aspects of Bayesian modeling in the wavelet domain.

In this chapter we explore thresholding rules induced by a variation of the Bayesian MAP principle. MAP rules are Bayes actions that maximize the posterior. In all models considered in this paper the posterior is infinite at zero, i.e., zero is trivially the mode of the posterior. If no other modes exist, zero is the Bayes action. If the second, non-zero mode of the posterior exists, this mode is taken as the Bayes action. Such a rule is thresholding and always picks the mode larger in absolute value if such local mode exists. This is the motivation for the name LPM - Larger (in absolute value) Posterior Mode. We show in Chapter that the thresholding rule induced by replacing the wavelet coefficient with the larger posterior mode of the corresponding posterior, performs well compared to several popular shrinkage techniques.

3.5 Larger posterior mode (LPM) wavelet thresholding

As is commonly done in Bayesian wavelet shrinkage, a Bayesian model is proposed on observed wavelet coefficients. As mentioned in Section 3.4, a prior vague model will be set on the scaling coefficients in order to let them intact and, due to the decorrelation property of wavelet transforms, a "typical" wavelet coefficient d will be modelled. Therefore, our model starts with

$$d = \theta + \epsilon, \tag{3.5.1}$$

where we are interested in the location θ corresponding to the signal part contained in the observation d. Bayes rules under the squared error loss and regular models often result in shrinkage rules resembling thresholding rules, but they are never thresholding rules. In many applications rules of the thresholding type are preferable to smooth shrinkage rules. Examples include model selection, data compression, dimension reduction, and related statistical tasks in which it is desirable to replace by zero a majority of the processed coefficients. In order to have thresholding rules, different loss functions have to be considered. For example the posterior median (see Abramovich, Sapatinas and Silverman, 1998) minimizes the L_1 loss while the Bayes factor (see Vidakovic, 1998) minimize the $0 - 1$ loss.

This Chapter considers construction of bona fide thresholding rules via selection of a larger (in absolute value) posterior mode (LPM) in a properly set Bayesian model. The models considered in this chapter produce posteriors with no more than two modes. The selected mode is either zero (a single mode – thus trivially the larger) or non-zero mode if the posterior is bimodal.

3.5.1 Derivation of the thresholding rule

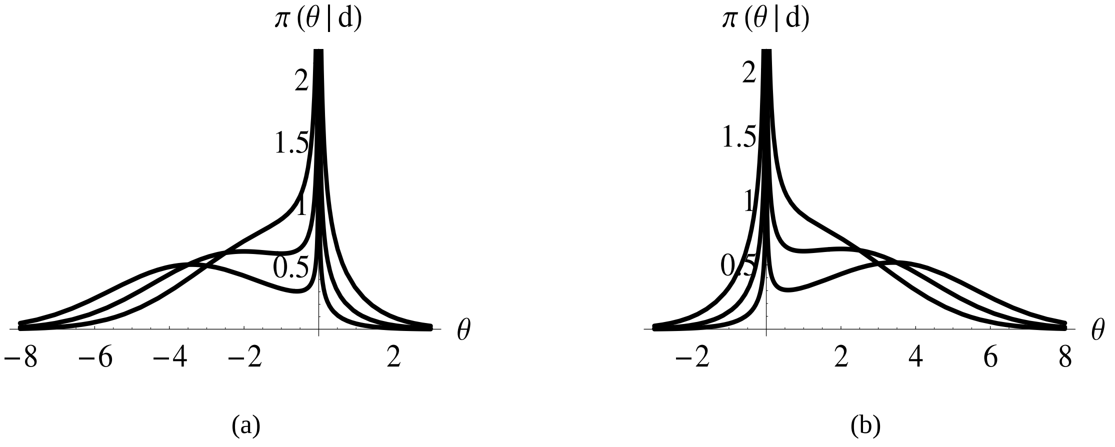
We consider several versions of the model, under the assumption of Gaussian noise. In the basic version, discussed in this Section, the variance of the noise σ^2 is assumed to be known and the prior is elicited only on the unknown θ location. In the generalized versions discussed in the following section, the variance of the noise is assumed unknown and will be modelled by (i) inverse-gamma and (ii) exponential priors which are independent from the location parameter.

Consider the following *basic* model

$$\begin{aligned} d|\theta &\sim \mathcal{N}(\theta, \sigma^2), \\ \theta|\tau^2 &\sim \mathcal{N}(0, \tau^2), \\ \tau^2 &\sim (\tau^2)^{-k},\ k > \frac{1}{2}, \end{aligned} \tag{3.5.2}$$

where the variance σ^2 is assumed known and in practice estimated from the data and plugged in the model. We seek a MAP solution, i.e., an estimator of θ that (locally) maximizes the posterior, $p(\theta|d)$. To find the extrema of the posterior on θ we note that the posterior is proportional to the joint distribution of d,θ and τ^2, so the value of θ maximizing the joint distribution maximizes the posterior, as well. The joint distribution is such that

$$
\begin{aligned}
p(d,\theta) &= \int p(d|\theta)p(\theta|\tau^2)p(\tau^2)d\tau^2 \\
&= \int \frac{1}{\sqrt{2\pi}\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\frac{1}{\sqrt{2\pi\tau^2}}e^{-\theta^2/(2\tau^2)}\frac{1}{(\tau^2)^k}d\tau^2 \\
&= \frac{1}{2\pi\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\int(\tau^2)^{-(k+1/2)}e^{-\theta^2/(2\tau^2)}d\tau^2 \\
&= \frac{1}{2\pi\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\int y^{(k-1/2)-1}e^{-\theta^2 y/2}dy \\
&= \frac{1}{2\pi\sigma}e^{-(d-\theta)^2/(2\sigma^2)}\frac{\Gamma(k-1/2)}{(\theta^2/2)^{k-1/2}}, k > 1/2
\end{aligned}
$$



(a)

(b)

Figure 3.1: Posterior distribution for $k = 3/4$ and $\sigma^2 = 2^2$; (a) $d = -4, -3, -2$; (b) $d = 2, 3, 4$. The unimodal density graphs in panels (a) and (b) correspond to $k = -2, 2$, respectively.

This leads to posterior

$$
p(\theta|d) \propto p(d,\theta) \propto e^{-(d-\theta)^2/(2\sigma^2)}\,|\theta|^{-2k+1}. \tag{3.5.3}
$$

Figure 3.1 (a,b) depicts the posterior distribution for $k = 3/4$, $\sigma^2 = 2^2$, and various values of d. Note that if d is small in absolute value compared to σ^2, the posterior is unimodal with (infinite) mode at zero. For $|d|$ large, the posterior is bimodal with non-zero mode sharing the same sign as the observation d.

The logarithm of the posterior is proportional to

$$\ell = \log p(\theta|d) \propto -\frac{(d-\theta)^2}{2\sigma^2} + (1-2k)\log\theta,$$

and has extrema at the solutions of a quadratic equation,

$$\theta^2 - d\theta + \sigma^2(2k-1) = 0,$$

$$\theta_{1,2} = \frac{d \pm \sqrt{d^2 - 4\sigma^2(2k-1)}}{2}.$$

The roots $\theta_{1,2}$ are real if and only if $d^2 \geq 4\sigma^2(2k-1)$, i.e., if $|d| \geq 2\sigma\sqrt{2k-1} = \lambda(\sigma) = \lambda$. If this condition is not satisfied, then the likelihood is decreasing in $|\theta|$ and the MAP is given by $\hat{\theta} = 0$.

The value of the posterior at zero is infinite, thus zero is always a mode of the posterior. When this is the only mode, the resulting rule takes value zero. If the second, non-zero mode exists, then this mode is taken as the Bayes action.

We assume, without loss of generality, $d > 0$. Since $k > 1/2$, $\sqrt{d^2 - 4\sigma^2(2k-1)} < d$ and both roots are positive and smaller than d, we have shrinkage. Then the LPM is $\frac{d+\sqrt{d^2-4\sigma^2(2k-1)}}{2}$, since the posterior is decreasing from 0 to smaller root, increasing between the two roots and decreasing after the larger root. For arbitrary d, and $\lambda = 2\sigma\sqrt{2k-1}$, the LPM rule is

$$\hat{\theta} = \frac{d + \text{ sign}(d)\ \sqrt{d^2 - 4\sigma^2(2k-1)}}{2}\mathbf{1}(|d| \geq \lambda). \tag{3.5.4}$$

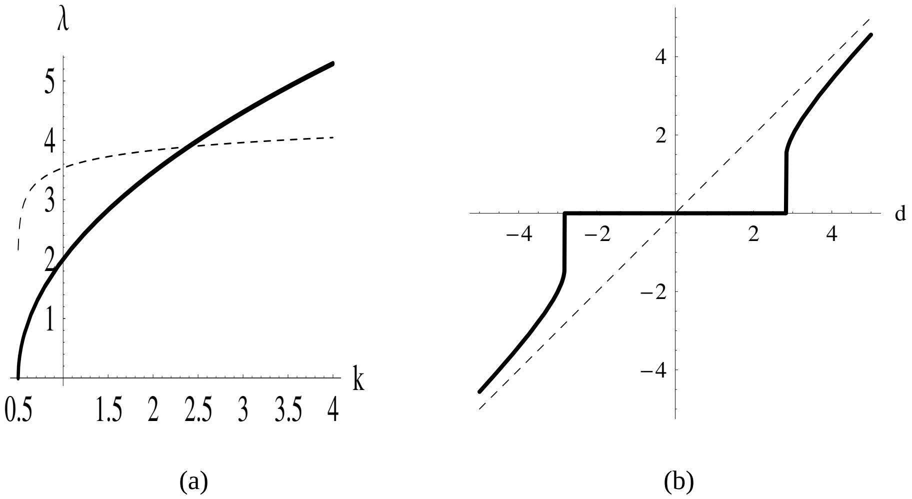


Figure 3.2: (a) Influence on the threshold λ by power parameter k; (b) LPM thresholding rule.

Figure 3.2 (a) compares values of threshold λ to properly scaled universal threshold (Donoho and Johnstone, 1994). In both cases the variance $\sigma^2 = 1$. The dotted line represents the values of universal threshold rescaled by $n = (k-1/2)\cdot 2^{10}$. This sample size n is selected only for comparison reasons. As depicted in Figure 3.2 (b), the thresholding rule looks like a compromise between hard and soft thresholding, The rule generally remains close to d for intermediate and large values of d.

Note that the posterior (3.5.3) is proper (integrable at 0) if and only if $2k - 1 < 1$, i.e., when $k < 1$. The existence of a finite second mode does not require the posterior to be proper and we will consider all $k > 1/2$.

Remark. If the square root in (3.5.4) is approximated by Taylor expansion of the first order, $(1-u)^\alpha \approx 1 - \alpha u$, the LPM rule mimics James-Stein estimator,

$$\hat{\theta} \approx \left(1 - \frac{\sigma^2(2k-1)}{d^2}\right)_+ d,$$

which is considered extensively in the wavelet shrinkage literature.

3.5.2 Exact risk properties of LPM rules

The exact risk analysis of any proposed shrinkage rule has received considerable attention in the wavelet literature since it allows for comparison of different wavelet-based smoothing methods. When the rule is given in a simple form, the exact risk analysis can be carried out explicitly. For instance, Donoho and Johnstone (1994) and Bruce and Gao (1996) provide exact risk analyses for hard and soft thresholding under squared error loss. Gao and Bruce (1997) give a rationale for introducing the “firm” or “semi-soft” thresholding utilizing exact risk arguments. The goal of exact risk analysis is to explore robustness in risk, bias, and variance when the model parameters and hyper-parameters change.

For our model the analytic form of LPM rule (3.5.4) is more complex and the exact risk analysis was carried out numerically. Computations performed in the software package MATHEMATICA produced Figure 3.3. We briefly describe the properties inferred from Figure 3.3.

In Figure 3.3(a) the risks of rule (3.5.4) for $k = 0.6, 0.75,$ and 0.9, are presented. These risks are partitioned to variances and biases-squared given in panels Figure 3.3(b) and Figure 3.3(c). The shapes of risks are typical for hard thresholding rules. The risk is

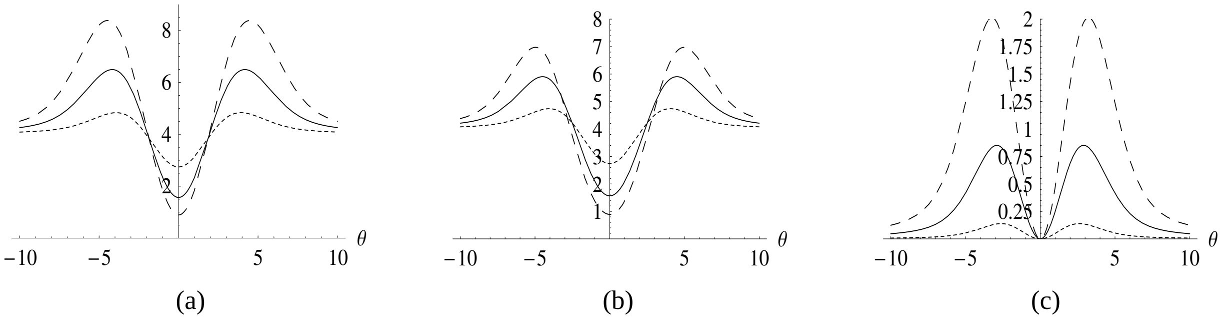


Figure 3.3: Exact risk plots for LPM rule, for $k = 0.6$ (short dash), $k = 0.75$ (solid), and $k = 0.9$ (lomg dash). For all three cases $\sigma^2 = 2^2$. (a) Risk; (b) Variance, and (c) Bias squared.

minimal at $\theta = 0$ and it stabilizes about the variance for $|\theta|$ large. For values of θ that are comparable to the threshold λ the risk is maximized. This signifies that largest contribution to the MSE is for the values of θ close to the threshold. This is to be expected since for θ's close to threshold, given that the noise averages to 0, the largest errors are made by the "keep-or-kill" policy. The variance plots Figure 3.3(b) generally resemble the plots for the risk. As is typical for hard thresholding rules, the squared bias Figure 3.3(c) is small in magnitude compared to variance and risk. This is a desirable property when the users are concerned about the bias of the rule and ultimately, the estimator $\hat{\mathbf{f}}$.

We note that the role of k in the shapes of risk, variance, and bias-squared is linked to the role of sample size and increased variance in standard shrinkage situations. This link will be discussed further in Section 4.

3.6 Generalizations

In the previous section we assumed that the variance of the noise, σ^2 was known. In applications, this variance can be estimated from the data (usually using the finest level of detail in the wavelet decomposition) and the estimate is then plugged in the shrinkage rule. In this section we generalize the methodology by eliciting a prior distribution on the variance.

We consider two generalizations of the model in (3.5.3). In the first, the variance is assigned an exponential prior, leading to a double exponential marginal likelihood, while in the second, the variance is assigned an inverse gamma prior, leading to a t marginal likelihood.

3.6.1 Model 1: exponential prior on unknown variance.

Assume that for a typical wavelet coefficient d the following model holds.

$$\begin{aligned} d|\theta,\sigma^2 &\sim \mathcal{N}(\theta,\sigma^2), \\ \sigma^2|\mu &\sim \mathcal{E}\left(\frac{1}{\mu}\right) \text{ with density } p(\sigma^2|\mu)=\mu e^{-\mu\sigma^2}, \mu>0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k>\frac{1}{2}. \end{aligned}$$

It is well known that an exponential scale mixture of normals results in a double exponential distribution. Thus this model is equivalent to

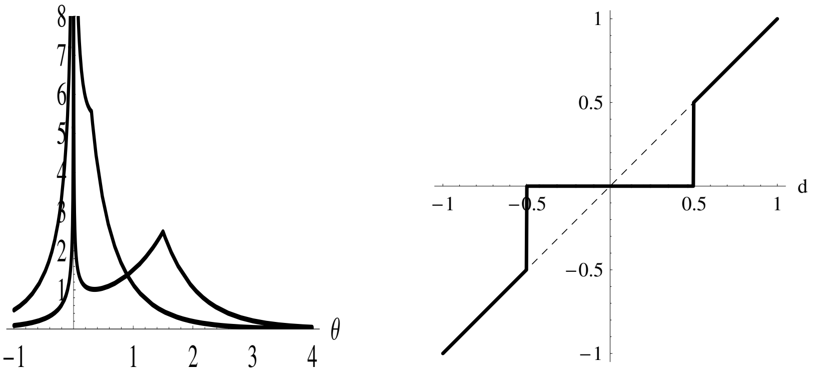
$$\begin{aligned} d|\theta,\mu &\sim \mathcal{DE}\left(\theta,\frac{1}{\sqrt{2\mu}}\right), \quad \text{with density } f(d|\theta)=\frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k>\frac{1}{2}. \end{aligned}$$

Lemma 3.6.1. *The resulting LPM rule turns out to be hard-thresholding,*

$$\hat{\theta}=d\ \boldsymbol{1}(|d|\geq\lambda) \tag{3.6.1}$$

where $\lambda=\frac{2k-1}{\sqrt{2\mu}}$.

Proof. Proof of Lemma 3.6.1 is deferred to the Appendix □



(a)

(b)

Figure 3.4: (a) Influence on the posterior in Model 1 by two different values of d; (b) LPM rule for Model 1.

Figure 3.4(a) shows the posterior distribution for Model 1 for values of d leading to unimodal (infinite at mode 0) and bimodal cases. The values are $d = 0.3$ and $d = 1.5$, $k = 0.75$ and $\mu = 1$. The LPM rule (3.6.1) is shown in Figure 3.4(b) for $k = 0.75$ and $\mu = 1/2$.

The double exponential marginal likelihood is a realistic model for wavelet coefficients. In fact, if a histogram of wavelet coefficients for many standard signals is plotted, it resembles a double exponential distribution. This observation first explicitly stated by Mallat (1989), is used in many Bayesian models in the wavelet domain, examples are BAMS wavelet shrinkage (Vidakovic and Ruggeri, 2001) or the wavelet image processing methodology of Simoncelli and coauthors (e.g., Simoncelli and Adelson, 1996).

3.6.2 Model 2: inverse gamma prior on unknown variance.

The inverse gamma prior on the unknown variance of a normal likelihood is the most common and well understood prior. The resulting marginal likelihood on the wavelet coefficients is t-distributed, which models heavy tails of empirical distributions of wavelet coefficients well. Model 2 with an inverse gamma prior will not realistically model the behavior of wavelet coefficients in the neighborhood of 0, but will account for heavy tails encountered in empirical distributions of wavelet coefficients. Model 2 is given by

$$\begin{aligned} d|\theta,\sigma^2 &\sim \mathcal{N}(\theta,\sigma^2), \\ \sigma^2 &\sim \mathcal{IG}(\alpha,\beta) \text{ with density } p(\sigma^2|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-1-\alpha} e^{\frac{-\beta}{\sigma^2}}, \alpha > 0, \beta > 0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k > \frac{1}{2}. \end{aligned}$$

Lemma 3.6.2. *The resulting LPM rule is*

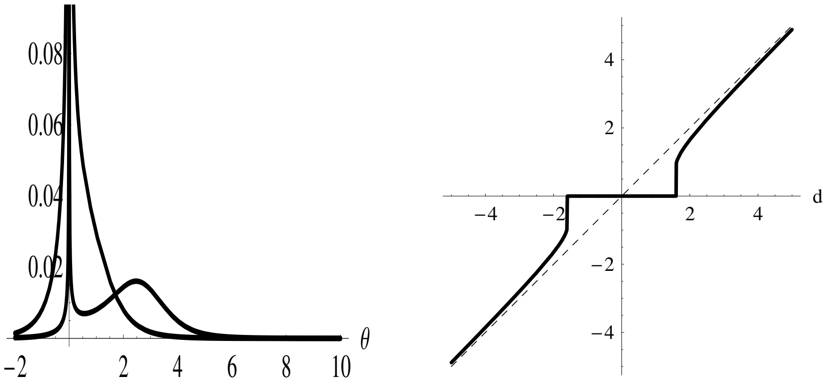
$$\hat{\theta} = \frac{(2\alpha + 4k - 1)d + \textit{ sign}(d)\sqrt{(2\alpha+1)^2 d^2 + 16(1-2k)(k+\alpha)\beta}}{4(k+\alpha)} \boldsymbol{1}(|d| \geq \lambda), \tag{3.6.2}$$

where

$$\lambda = \frac{2}{2\alpha - 1}\sqrt{(2k-1)(k+\alpha)\beta}\,.$$

Proof. Proof of Lemma 3.6.2 is deferred to the appendix. □

Figure 3.5(a) shows the posterior distribution for the Model 2 for values of d leading to unimodal and bimodal cases. The values are: $d = 0.7$ and $d = 2.7$, $k = 0.85$, $\alpha = 2.5$ and $\beta = 2$. The LPM rule (3.6.2) is shown in Figure 3.5(b) for $k = 0.85$, $\alpha = 2.5$ and $\beta = 2$.



(a)

(b)

Figure 3.5: (a) Influence on the posterior in Model 2 by two different values of d; (b) LPM rule for Model 2.

3.7 Conclusions and future works

In this chapter we developed a method for wavelet-filtering of noisy signals based on larger (in absolute value) posterior mode when the posterior is bimodal. Three variants of the model are considered. The resulting shrinkage rules are thresholding. As we will see in Chapter 6, the LPM is a global method, i.e., the model parameters/hyperparameters are common across the scales in wavelet decompositions. Models for which the parameters/hyperparameters are level-dependent are called adaptive.

We envision several avenues for future research. The LPM thresholding could possibly be improved by level-based specification of model hyperparameters. Such level adaptive formulations are more appropriate for signals and noises that exhibit scale-dependent heterogeneity.

In generalizing the basic model to account for unknown variance we considered only exponential and inverse gamma scale mixtures of normals. Scale mixtures of normals comprise a rich family of models and it would be possible to find an optimal mixing distribution. Specifically, an exponential power distribution (EPD) that contains as special cases normal and double exponential distributions can be obtained as a scale mixture of normals with positive stable distribution as a mixing distribution.

Chapter 4

Numerical Experiments for Discriminant Analysis

Introduction

In this chapter we apply some of the classification techniques introduced in chapter 1. In section 4.1 Cloud detection from satellite multispectral images through statistical discriminant analysis is investigated. Validation on case studies from the AVHRR sensor is performed. In section 4.2 some local discriminant methods are tested on synthetic data, performance of the different method are compared and suggestions for future work are provided.

4.1 Multispectral cloud detection: general problem

Cloud detection is a preliminary important step in most algorithms for processing radiance data measured from satellites. For this reason a cloud mask endows radiance data coming from most last generation sensors onboard satellites. Practically all operative cloud mask algorithms are physically based: cloud models are introduced in radiative transfer models and their influence on the radiance emitted from the Earth surface is estimated with respect to clear sky conditions at spectral regions that simulate the spectral channels of a sensor. Then generally single bands or couples of bands are considered and thresholds on the value of radiances at the bands (or their differences or ratios, e.g.) are empirically chosen able to discriminate between the clear and cloudy sky conditions. Such a procedure is very consolidated and robust, especially when the type of clouds present in the atmosphere matches the clouds that were simulated in the radiative models. Physical methodologies suffer from three main drawbacks: the variability of clouds in the sky is much larger than that resulting from simulations by radiative transfer models; the dependence of radiance on the emissivity of the surface, which is very difficult to estimate accurately over land; the increase of the number of spectral channels of sensors, that makes more difficult the choice of suitable bands for the decision rules. For this reason there was in the recent years interest towards classification methods that approach the problem of cloud detection through statistical methods: the classification methods learn the statistical features of the cloudy and clear sky conditions “on-field”, that is starting from “truth” images where the sky conditions are *certainly* known; then sky conditions on other “new” images are inferred from these by relying on some of the statistical properties learned. However there is a main drawback that limits evolution of this methodology into an operative algorithm. *Supervised* classification methods rely on a “truth” cloud mask as a training set. Indeed at the same time this is the strongest link of the (statistical) discriminant analysis to cloud physics. Actually to develop such a cloud mask is a hard task. Therefore a common procedure is to use as a training set the result of classification obtained by a different methodology, in general non based on pure statistical arguments. This weakens the role of the classification methods as competitor of the physically based ones, since formally the final target of classification is moved to reproduce the classification results of another methodology. Nevertheless, these statistical approaches can shed more light on misclassifications of physical methods, by looking deeply at the pixels classified differently between statistical and physical methods; moreover they allow one to perform an assessment of the cloud detection methodology with respect to the spectral bands, aimed at estimating *on field* the role of each spectral band and decision rule in the physical cloud detection methodology.

4.1.1 Classification methods considered

Three nonparametric (NPDA, PCDA, ICDA) and one parametric (LDA) discriminant analysis methods for multispectral cloud detection have been considered. These methods were described in Chapter 1 and we briefly summarize them in the following. LDA (Linear Discriminant Analysis) is based on Gaussian density functions with common variance among classes; in NPDA (NonParametric Discriminant Analysis) a nonparametric estimate of the density functions is made for each component separately; in PCDA (Principal Component Discriminant Analysis) original components are transformed into principal components prior to nonparametric density estimation; in ICDA (Independent Component Discriminant Analysis) original components are transformed into independent components prior to nonparametric density estimation.

4.1.2 Case studies

We consider the NOAA/NASA Pathfinder AVHRR Land data set available at the NASA Distributed Active Archive Center (DAAC) Web site `http://daac.gsfc.nasa.gov`. The Pathfinder AVHRR Land data sets are terrestrial data sets produced from 20 years of archived data from the five-channel AVHRR sensor aboard the NOAA satellites. AVHRR has five channels in the order: 0.58–0.68 μm (visible), 0.725–1.10 μm (near infrared), 3.55–3.93 μm (middle infrared), 10.3–11.3 and 11.5–12.5 μm (thermal).

We consider the Daily Data Set product, available at the resolution of 8 Km $\times$ 8 Km that contains reflectances and radiances for the five channels and the Clouds from AVHRR (CLAVR) product, see *Agbu and James* (1994), for a detailed description of the products. In particular CLAVR uses the five-channel multispectral information in a series of sequential decision-tree type tests to identify the cloud-free, mixed (variably cloudy), and cloudy pixels. In *Stowe at al.* (1991) it is claimed that the CLAVR technique is based on the following differences between the radiative and physical properties of clouds and the underlying surface: magnitudes of reflected and emitted radiation (contrast), wavelength dependence, and spatial variability.

We consider two datasets (Jun 21, 2000 and Jun 21, 2001) and two experiments. In the first one, data of June 21, 2000 are considered as a training and testing dataset. In the second experiment data of June 21, 2001 are used as a training set, whereas data of June 21, 2000 are actually classified. In both cases training set is taken from the CLAVR product. An area covering Mediterranean is considered (30^o–50^o latitude and 0^o–40^o longitude range) that includes over 40000 useful pixels.

To rank the effectiveness of the discriminant analysis methods in detecting clouds, the observed percentage of agreement, S, for the whole volume of data (the percentage of correctly classified pixels) has been considered.

For each experiment, classification is performed by the four discriminant analysis methods discussed in Chapter 1. In addition, assessing of spectral bands is made considering all their possible combinations (that is, cloud detection using just one band, two bands, and so on). Therefore the total number of experiments for all combinations of bands is 31.

4.1.3 Results

First of all we show as an example in Fig. 4.1 typical probability density functions of clear and cloudy sky reflectances in the spectral band 0.58–0.68μm over different zones in the Mediterranean area on June 21, 2000. It is clear that hypothesis of Gaussianity of distributions is not met at all. The same result holds also for the other spectral bands. Therefore it makes sense to consider discriminant analysis methods that estimate probability density functions nonparametrically.

Table 4.1 shows performance of the discriminant analysis methods in terms of success percentage (S) for the case when both training and testing datasets are given by the day June 21, 2000.

We see that the thermal channels 4 and 5 (10.3–11.3 μm and 11.5–12.5 μm), taken independently, give the best prediction, whereas band in the middle infrared (3.55–3.93 μm) has the poorest performance. It is also clear that nonparametric methods significantly improve cloud detection capability. When bands are put together, we notice that detection capabilities improve, as expected. In particular, the middle infrared spectral channel is always present in all top performance combinations, that is it is the best companion spectral band.

Also noteworthy is that top performances are reached already with only two bands and that differences of performance among the various methodologies tend to be small when the number of channels increases — in other words multispectrality fixes the departure of data from the theoretical assumptions of the methodology.

In order to test robustness of the methodology, we considered again June 21, 2000 as a training set, but test was made on June 21, 2001. Table 4.2 shows performance of the Independent Component Discriminant Analysis (the best performing). Findings of the previous analysis were substantially confirmed and top performance is reduced only slightly.

4.2 Simulation of Local discriminant methods

This section includes results obtained applying on synthetic datasets the local discriminant methods introduced in Chapter 1. We consider an image formed by a 100x100 array of pixels. The image contains three distinct regions, each one populated according to a specific probability density function, so that class labels are homogeneous inside each region. The image is shown in Fig. 4.2. We assume that probability density function inside each region is Gaussian with specified mean μ_k and variance σ_k^2, $k = 1, 2, 3$.

We shall compare the four local classification methods proposed in chapter 1 (LV, LF, LI and LN) with proper nonlocal methods and with ICM (Besag, 1986). To this purpose we recall that ICM method assumes that the true label set of an image is a realization of a locally dependent Markov Random field so that the posterior class probability for a specific data point also depends on the labelling of its neighborhood. After obtaining a first class estimates for each pixel using the classical Bayes rule (1.2.6) with constant a priori class probabilities, at every iteration step and for each pixel x_c the classes a priori probabilities are estimated as follows:

$$\pi_k^{\mathrm{ICM}}(x_c) = \frac{\exp(\beta\varphi_k(x_c))}{\sum_{j=1}^{K}\exp(\beta\varphi_j(x_c)}),\ k = 1,\ldots,K, \tag{4.2.1}$$

where $\varphi_k(x_c)$ are defined by Eq. (1.7.2), and β is a fixed parameter that, when positive, encourages the central pixel to have the same class as the dominant one in the neighborhood. In our experiments we will set $\beta = 1.5$ (for more details see Besag 1986). Notice that by its very definition, namely the presence of the exponential term in Eq. (4.2.1), ICM strongly privileges the most probable class according to the class labels estimated by some discriminant analysis method; therefore it strongly enhances visibility of already visible classes.

Example 1

In the first example we choose ($\mu_1 = 1$, $\sigma_1^2 = 1$), ($\mu_2 = 4$, $\sigma_2^2 = 1$) and ($\mu_3 = 7$, $\sigma_3^2 = 1$). As we can see from Figure 4.3, in this situation the three distributions overlap just on small intervals so that all of them are very *visible*, in the sense that their dominance index is high for all classes (0.9332, 0.8664 and 0.9332 for the three classes, respectively). As a consequence all five local classification methods show a high global percentage of success rate (see Table 4.3; digits are averages over 50 different realizations). We also compare performance with a nonlocal method (Linear Discriminant Analysis, LDA), that is well suited in this case, due to the Gaussian distributions and to equal variance for the three classes. Constant priori class probabilities for all the three classes were chosen for LV, LF, LN, LDA and ICM. Convergence was reached in less than 10 iterations. Moreover final solution was quite robust with respect to the choice of the first-guess priori class probabilities. The region $\mathcal{B}$ was a square 3x3 around each pixel. Figure 4.4 shows the reconstructed label field for one realization through (nonlocal) LDA. It is possible to note the presence of the 'pseudo-nuisance' discussed in the paper. As a comparison, Figure 4.4 shows the reconstructed label field obtained through LV. In this case the 'pseudo-nuisance' practically disappeared. Also note that boundaries between regions of different classes are well represented, which means that perfromance is excellent even when $\mathcal{B}$ includes pixels belonging to different classes.

Example 2

In the second example we choose ($\mu_1 = 1$, $\sigma_1^2 = 1$), ($\mu_2 = 2$, $\sigma_2^2 = 4$) and ($\mu_3 = 3$, $\sigma_3^2 = 1$). As we can see from Figure 4.6, in this situation the second distribution overlaps almost everywhere with the others and it is lower, so that it is clearly *not visible*. This is confirmed also by the dominance index, which is 0.7973,0.0965 and 0.7973 for the three classes, respectively. Table 4.4 shows classification success percentage for each class and globally for the four local classification methods together with ICM and Quadratic Discriminant Analysis (QDA); the latter is well suited in this example, due to the Gaussian distributions and to different variances for the three classes. As it could have been expected, success percentage is much lower, due to the very poor capability of all methods of predicting class label 2. Figure 4.7 shows the reconstructed label field from one realization through QDA. It is possible to note that class 2 in the middle region is very poorly predicted. As a comparison, Figures 4.8 and 4.9 respectively show the reconstructed label field obtained through LN and ICM. Note that as expected our proposed local methods LF, LI and LN tend to privilege occurrence of the less visible classes, so that they perform better than ICM and LV in these cases. As in example 1, the region $\mathcal{B}$ was a square 3x3 around each pixel.

4.2.1 Real data

Clouds are a typical example where it is difficult to set values to priori class probabilities, because presence of clouds depends on the season, on the location and on the climate, all items strongly varying with images. In addition clouds do show a strong spatial correlation, just because from the physical point of view they are aggregation of water particles. Of course the degree of correlation depends on the type of cloud and on the meteorological conditions; in this respect an important role is also played by the spatial resolution of the sensor that detects images from satellite. For this reason cloud mask detection is prone to benefit from local algorithms for classification.

In this section we show an example of cloud mask retrieval for a scene over Italy of September 21st 2000. Image was taken by MODIS sensor onboard NASA EOS satellite. MODIS yields images in 36 spectral channels covering visible and near infrared spectral regions, but only two of them have the best spatial resolution of 250m. For this reason it is interesting to develop cloud mask algorithms that rely on a very limited number of spectral channels, so that the cloud mask has the same best spatial resolution as the data and no degradation to the resolution of other channels occurs.

Figure 4.10 shows the image considered in the present paper, which refers to reflectance at the spectral channel 0.465 μm. Discriminant analysis has been applied by means of the NPDA (NonParametric Discriminant Analysis) method described in Amato et al., 2003. In practice class distributions are estimated by means of Kernel density estimation, which is appropriate for this problem since probability density functions of the clear and cloudy classes cannot be approximated by Gaussian. Both the cases of nonlocal and local class priori probabilities have been considered: in the first case uniform values over the classes were considered; in the second case uniform values were used as first guess and LF method was chosen to compute final localized values. The region $\mathcal{B}$ was a square 3x3 around each pixel. Results of classification are shown in Fig. 4.11. They confirm capability of the localized discriminant analysis to reduce ‘pseudo-nuisance’ of nonlocal methods.

4.2.2 Conclusions

Some local discriminant analysis methods have been proposed for image classification aimed at exploiting spatial correlation among neighbor pixels that is natural in most applications. These methods have the twofold objective of a) reducing the ‘pseudo-nuisance’ of nonlocal methods due to the overlap of the probability density functions of the various classes; b) decrease misclassifications of even simple discriminant analysis methods, so that performance are approached of more advanced methods even using very low dimensional information for each pixel. The proposed methods are based on the choice of local priori class probabilities using information surrounding each pixel of the image. Discriminant analysis has been revisited according to the visibility or nonvisibility of classes, that is, capability of a particular classification method to retrieve a class. In these respect suitability of the proposed methods to visible or nonvisible classes has been stressed. Particular attention has been paid to the problem of detecting the cloud mask from satellite remote sensed images, which is very important in remote sensing and seems particularly suited for nonlocal discriminant analysis methods. Numerical experiments on synthetic datasets confirm performance improvement of the nonlocal methods with respect to the local ones. No degradation is detected in proximity of the boundaries between regions of different classes. Results for nonvisible classes show to be still poor. Therefore it is advisable to investigate on alternative methods suited for both visible and nonvisible classes.

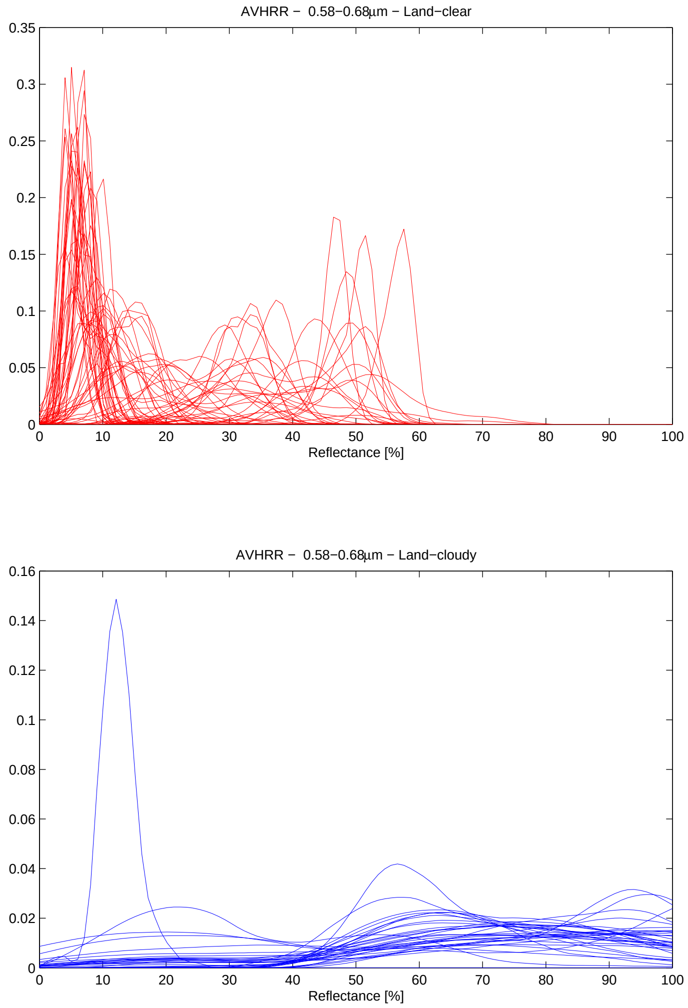


Figure 4.1: Typical density functions for clear and cloudy sky conditions.

Table 4.1: Performance (S) of Discriminant Analysis methodologies for the case of training and testing dataset June 21, 2000.

1	2	3	4	5	LDA	NPDA	PCDA	ICDA
◇					90.6	96.2	96.2	96.2
	◇				91.5	95.5	95.5	95.5
		◇			85.3	92.3	92.3	92.3
			◇		98.0	98.2	98.2	98.2
				◇	98.1	98.1	98.1	98.1
◇	◇				91.0	95.8	95.8	96.3
◇		◇			97.4	96.9	97.5	97.6
◇			◇		97.9	98.2	98.2	98.3
◇				◇	98.0	98.3	98.3	98.4
	◇	◇			96.6	96.5	96.6	96.8
	◇		◇		97.9	98.2	98.2	98.2
	◇			◇	98.0	98.0	98.0	98.0
		◇	◇		98.0	97.7	98.4	**98.5**
		◇		◇	98.2	97.7	98.3	**98.5**
			◇	◇	98.0	98.0	98.4	98.3
◇	◇	◇			97.4	96.9	96.7	96.9
◇	◇		◇		97.9	98.0	97.8	98.1
◇	◇			◇	98.0	98.1	97.7	98.0
◇		◇	◇		98.1	98.2	98.4	**98.5**
◇		◇		◇	98.2	98.2	98.4	98.4
◇			◇	◇	98.1	98.3	98.4	98.4
	◇	◇	◇		98.0	98.0	98.4	98.4
	◇	◇		◇	98.2	98.1	98.4	**98.5**
	◇		◇	◇	98.0	98.3	98.4	98.4
		◇	◇	◇	98.1	97.6	98.4	**98.5**
◇	◇	◇	◇		98.1	98.0	97.4	97.8
◇	◇	◇		◇	98.2	98.1	97.5	98.3
◇	◇		◇	◇	98.1	98.2	98.2	98.0
◇		◇	◇	◇	98.1	98.1	98.2	98.4
	◇	◇	◇	◇	98.1	98.0	98.3	**98.5**
◇	◇	◇	◇	◇	98.1	98.2	97.8	97.8

Table 4.2: Performance (S) of ICDA for the case of training and testing dataset June 21, 2001 and 2000, respectively.

1	◇					◇	◇	◇	◇
2		◇				◇			
3			◇				◇		
4				◇				◇	
5					◇				◇
	96.1	91.7	92.3	98.2	98.2	95.3	97.4	98.2	**98.3**
1							◇	◇	◇
2	◇	◇	◇				◇	◇	◇
3	◇			◇	◇		◇		
4		◇		◇		◇		◇	
5			◇		◇	◇			◇
	95.3	97.7	98.0	98.0	98.2	**98.3**	97.2	98.0	98.1
1	◇	◇	◇					◇	◇
2				◇	◇	◇		◇	◇
3	◇	◇		◇	◇		◇	◇	◇
4	◇		◇	◇		◇	◇	◇	
5		◇	◇		◇	◇	◇		◇
	98.2	**98.3**	97.9	96.1	96.7	95.3	98.1	98.1	97.9
1	◇	◇		◇					
2	◇		◇	◇					
3		◇	◇	◇					
4	◇	◇	◇	◇					
5	◇	◇	◇	◇					
	98.0	98.1	98.0	96.9					

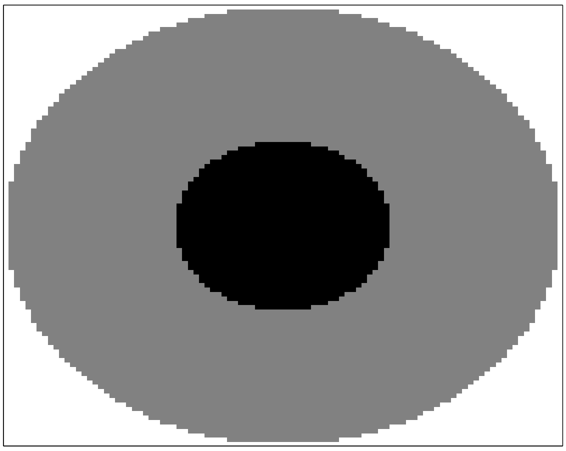


Figure 4.2: Image used for the synthetic experiments. Three regions are defined according to the color: 1 (black), 2 (gray) and 3 (white)

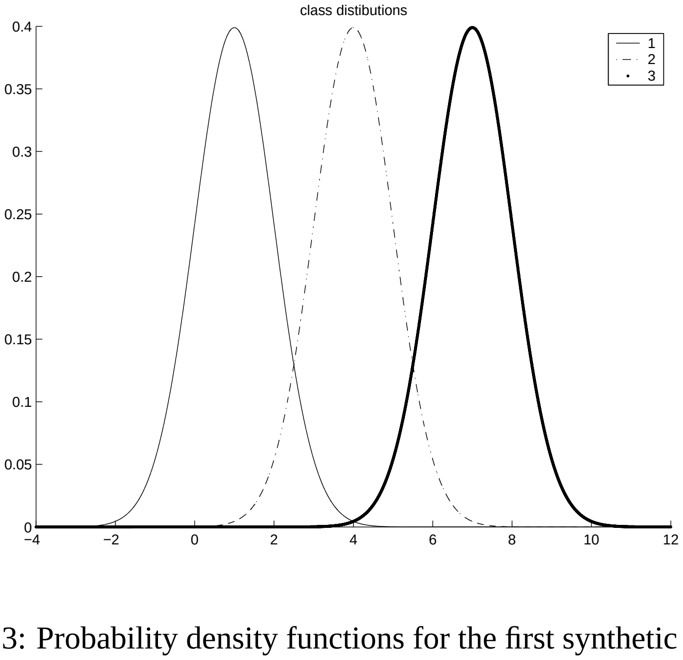


Figure 4.3: Probability density functions for the first synthetic example.

	$C = 1$	$C = 2$	$C = 3$	Global
LDA	91.9	86.9	92.1	88.69
ICM	99.5	99.8	99.4	99.65
LV	99.5	99.5	93.5	98.13
LF	97.3	95.9	97.2	96.35
LI	96.6	95.3	97.1	95.85
LN	98.6	97.5	98.0	97.77

Table 4.3: Success rate (percent) of discriminant analysis methods LF, LI, LN, ICM and LDA for the synthetic data of Example 1

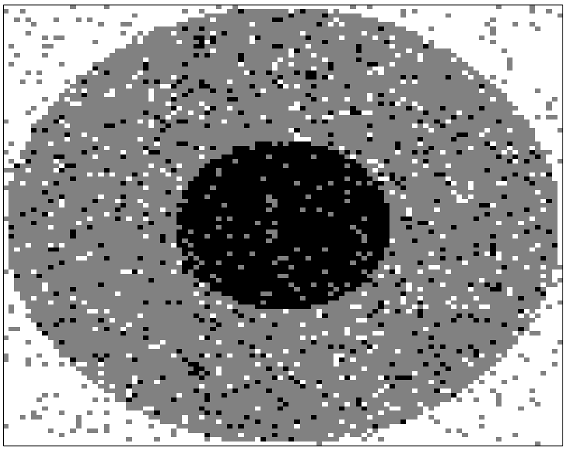


Figure 4.4: LDA classification for the example 1. Colors are black for class 1, gray for class 2 and white for class 3.

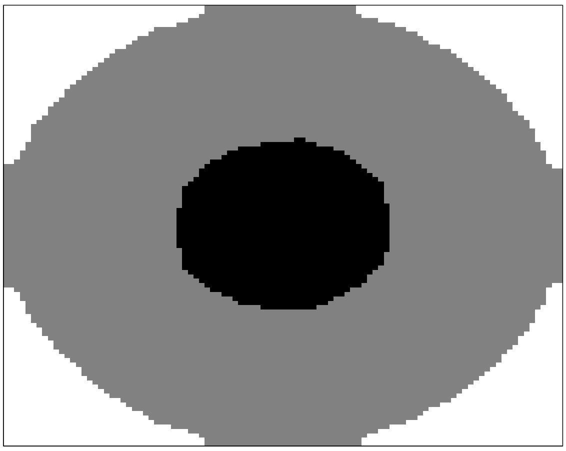


Figure 4.5: LV classification for the example 1. Colors are black for class 1, gray for class 2 and white for class 3.

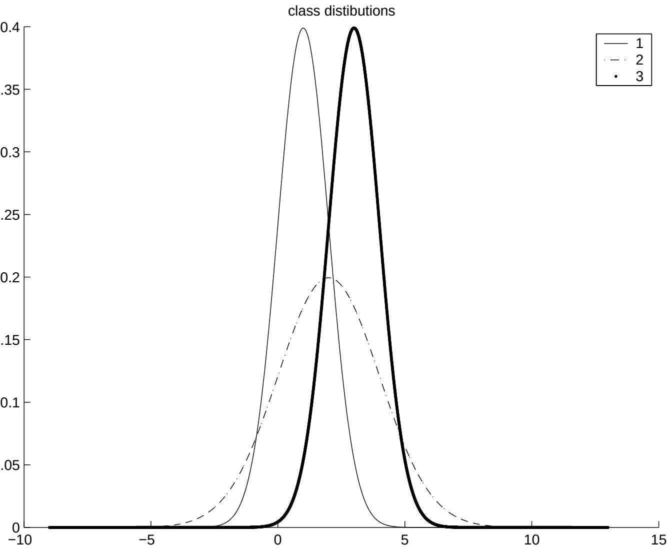


Figure 4.6: Second example distributions

	$C = 1$	$C = 2$	$C = 3$	Global
QDA	81.5	15.9	79.8	38.41
ICM	99.9	9.2	99.5	40.79
LV	100	0.4	97.7	34.66
LF	94.2	12.3	93.9	40.81
LI	89.2	15.3	88.5	40.92
LN	96.4	54.3	95.5	68.76

Table 4.4: Success rate (percent) of discriminant analysis methods QDA, ICM, LV, LF, LI and LN for the synthetic data of Example 2

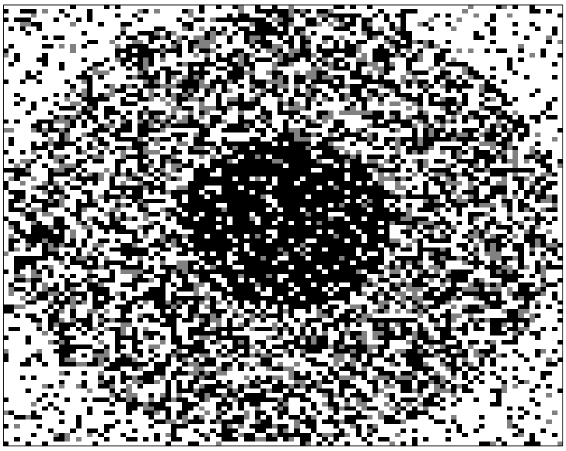


Figure 4.7: QDA classification for the example 2. Colors are black for class 1, gray for class 2 and white for class 3.

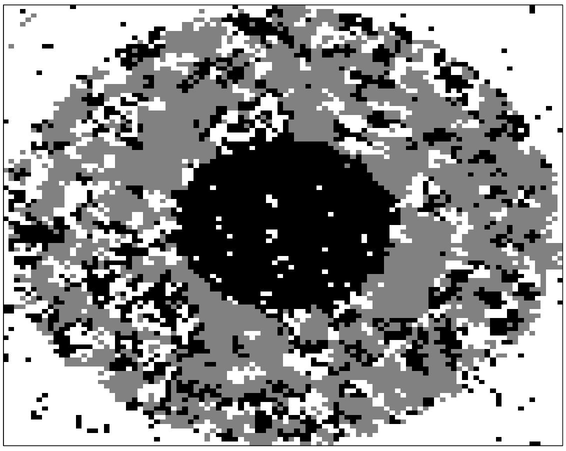


Figure 4.8: LN classification for the example 2. Colors are black for class 1, gray for class 2 and white for class 3.

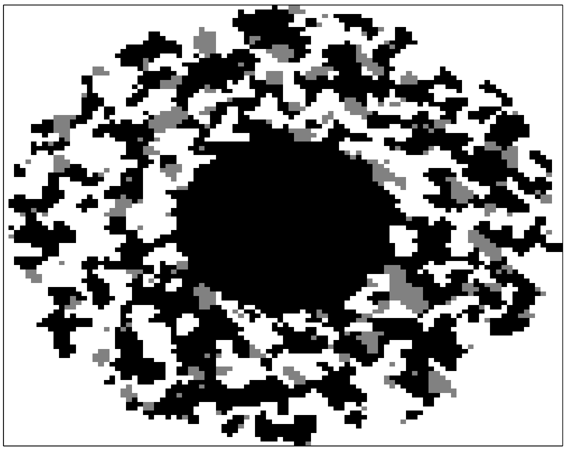


Figure 4.9: ICM classification for the example 2. Colors are black for class 1, gray for class 2 and white for class 3.

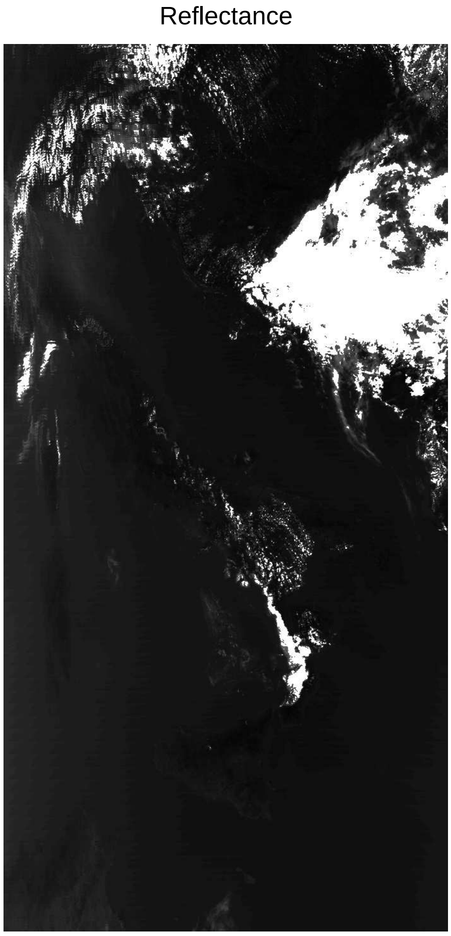


Figure 4.10: Reflectance of MODIS spectral channel 0.465 μm.

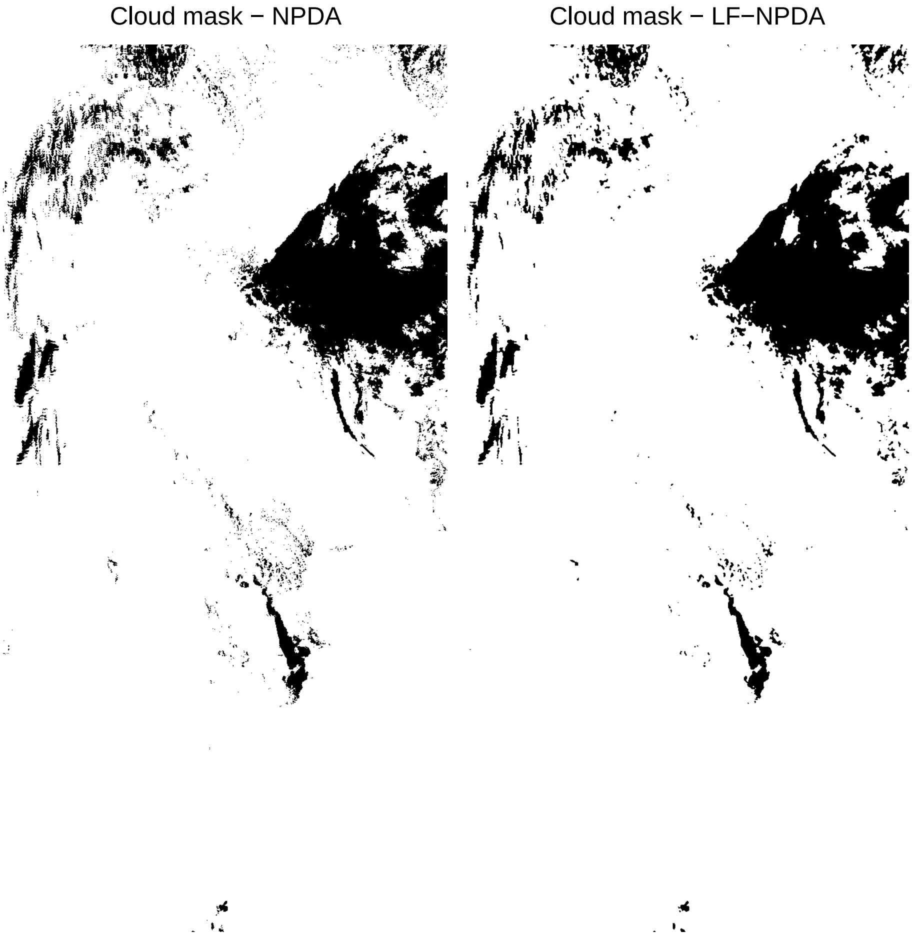


Figure 4.11: Cloud mask retrieved by NPDA (left) and NPDA with priori class probabilities chosen by LF (right).

Chapter 5

Application of multiple testing procedures to DNA microarray data

Introduction

In the present Chapter we give a brief introduction to cDNA microarray data and apply some of the multiple testing procedures (MTP) described in Chapter 2 to a specific cDNA microarray experimet . DNA microarray are part of a new class of biotechnology applications that allow the simultaneous monitoring of expression levels for thousands of genes in cells. The main question in microarray experiments is the identification of the few genes whose expression levels are associated with a response of interest, the so called *differentially expressed genes*. This biological question can be restated in terms of a multiple hypotheses testing problem where for each gene the hypothesis of no association between the expression levels and the response is tested.

The Chapter is organized as follows. Sections 1-3 provide a general background on DNA microarray data. Section 4 describes different normalization methods to remove the source of systematic variation in this kind of data. The application of MTP to a specific cDNA microarray experiment is discussed in Section 5.

5.1 What is a microarray ?

One of the main purpose of genetic is finding functions of cells, comparing tissue types, investigating the differences between healthy and diseased tissues, observing changes with respect to the application of drugs for drug discovery, or monitoring treatments. In order to reach these targets it is necessary to know the expression profiles of thousands of genes. Trough the novel technology of DNA microarray it is now possible to obtain quantitative measurements for the expression of genes present in a biological sample. The fundamental basis of DNA microarray is the process of hybridization. Two DNA strands hybridize if they are complementary to each other. One or both strands of DNA can be replaced by RNA and hybridization will still occur as long as there is complementarity. Roughly speaking a microarray is a glass slide on which oligonucleotide probes have been immobilized at micrometer distance. Small stings of RNA extracted from the sample of interest are used to hybridize to complementary fragments of DNA immobilized on the chip. The sample is usually labelled with a fluorescent dye so that the amount of hybridized mRNA can be detected by a light scanner that scans the surface of the chip. Under the assumption that the concentration of a particular messenger is a result of the expression of its corresponding gene, the fluorescent intensity detected is used to estimate its relative expression level. This application of DNA microarrays is in fact often referred to as expression analysis. Another traditional and completely different application is the so called genotyping, which consists in detecting mutation in specific genes. For expression analysis there are two major technologies available: Affymetrix genechip and spotted arrays. Affymetrix uses masks to control synthesis of oligonucleotides on the surface of a chip, divided in thousands of cells. Each oligo is about 25 nucleotides long and up to 40 oligos are used for the detection of each gene. Affymetrix chooses 11-20 oligos perfectly complementary, perfect match (PM), to the mRNA of a specific region of a gene and 11-20 oligos miss match (MM) that are identical to the PM except for the central position where one nucleotide has been changed to its complementary. The MM should be used to detect noise due to the process of hybridization itself. Messenger RNA is extracted from a single sample cell and converted to complementary DNA (cDNA). It is then amplified, labelled and so ready to undergo fragmentation and hybridization to the oligos on the surface of the chip. In spotted arrays technology a robot spots small quantities of probes in solution to the surface of a glass where they are dried. The probe can consist of cDNA, PCR (Polymerase Chain Reaction) or oligonucleotides. Each probe is complementary to a unique gene in the genome. Spotted array are often used to compare gene expression levels in two different samples, as the same cell type in two different conditions, say healthy and diseased, or two different cell types. The mRNA is extracted from the two different samples cells, converted to cDNA and labelled fluorescently with two dyes, usually red and green. After mixing they are hybridized to the probes on the glass slide. In both cases, after the hybridization step, the unhybridized material is washed away, the chip is scanned with a confocal laser and the image analyzed by computer. From a data analysis point of view the main difference between the two technologies is that in cDNA microarray two different samples labelled with two different dyes are hybridized on the same chip whereas Affymetrix chip can handle only one fluorochrome so that two chips are required to compare two samples. In the following we will focus on spotted arrays. The raw data produced from microarray experiments are the hybridized images, typically 16-bit TIFF (Tagged Information File Format), for each pair of sample to be compared. In order to obtain a final gene expression matrix, these images should be segmented, each spot identified, its intensity measured and compared to the background. This procedure is called quantization and it is done by image analysis software by which data from a set of microarray slides that constitute an experiment can be assembled into a single flat file. The main quantities of are the (R, G) fluorescence intensity pairs for each gene in each array. Note that since here R will stand for red Cy5, G for green Cy3 and the DNA sequences spotted on the array will be referred to as genes.

5.2 Experimental Design

Before a microarray experiment is performed, biological question to investigate on have to be clear and so it is necessary to draw a specific experimental design. It is important to decide whether a sample has to be considered as a biological replicate or a technical replicate (obtained from the same biological source). The design is partially imposed by the paired sample structure of two colors microarrays so that a single microarray can only be used to compare directly two samples. The simplest design for the direct comparison of two samples is the dye swap experiment. It uses two arrays to compare two samples which are called control and treatment. If in array 1 control sample is assigned to the red dye and treatment sample is assigned to the green dye, in array 2 the dye assignments are reversed. Using biological replicates this arrangement can be repeated by any even number of arrays. This is called repeated dye swap experiment and it is useful for reducing technical variation in the measurements. Using independent biological samples replicated dye swap, experiment is obtained. This design accounts for both technical and biological variation in the measurements. The most classical microarray experiments are based on the reference design that employs a special RNA sample, called the *reference*, to whom compare each test sample. Usually the reference is of no biological interest, so the number of technical replicates available for inference is half of what we could get using a different design. Even though this limitation, the reference design has the advantage of connecting different samples through their comparisons to the same reference. The reference design with dye swapping is a good design for large experiments because it is simple, robust and the distance between samples is always two. Another experimental design is the so called loop design in which samples are compared one to another in a daisy chain fashion. Small loops are a good alternative to the reference design but large loops may be inefficient. Variations on this design can be achieved combining loops with reference design or multiple loops together.

5.3 Expression Ratio

In a spotted cDNA microarray experiment, the ratio of the two fluorescent signals at each spot is commonly used to infer the ratio of the mRNA concentrations in the two samples. Lets consider an array that has N different genes and compare a sample s_1 labelled with red dye, to a sample s_2 labelled with green dye. The ratio for the i-th gene is: $D_i = \frac{R_i}{G_i}$. This is commonly called *fold change*. Intuitively the fold change gives a relative measurement of how the i-th gene is expressed in sample s_1 with respect to sample s_2 that is if D_i is greater then 1, the i-th gene is expressed more in s_1 then in s_2 and vice versa. In most experiments, for the majority of genes, the ratio should be nearly one as no differential expression is expected. Although the ratios give an intuitive way of measuring the relative genes expression level, it has the disadvantage of weighing up and down regulated genes differently. In order to treat them symmetrically with respect to the non differentially expressed genes, the logarithm of ratio is utilized to represent expression levels. Logarithm base 2 is used instead of decimal or natural logarithm because intensity is typically an integer between zero and $2^{16} - 1$.

5.4 Normalization

In every microarray experiment it is important to take into account the systematic variation in the measured gene expression levels of two cohybridized mRNA samples so to distinguish more easily biological differences and to allow comparison of expression levels across the slides. The process of removing systematic effects due to non biological sources is often referred to as normalization. There are many sources of systematic variation in microarray data, including unequal quantities of starting RNA, differences in labelling, different efficiency in the fluorescent dyes used, experimental bias in the measurements, unbalanced scanners, experiments replicated in different conditions and so on. It is necessary to normalize the fluorescent intensities before any analysis which involves comparing expression levels within or between slides. In the following we will consider different normalization methods.

5.4.1 Single slide data displays

Dudoit *at al* (2002) suggested that one of the most helpful graphical way of detecting dependence of log ratios on fluorescent intensity, for each slide, is to represent the (R,G)

data trough plotting $M = \log_2(R/G)$ versus $A = \log_2[(RG)^{1/2}]$. This graphical method, referred to as Ratio by Intensity (RI) plot or also MA-plot, is very useful for the purpose of normalization. In fact, as underlined in the previous paragraph, under the assumption that most genes are not differentially expressed and that *a priori* any differential expression is approximately symmetric with respect to up and down regulated genes, most points in an RI-plot should fall along a horizontal line. In practice this plot almost surely shows different patterns.

5.4.2 Within slide normalization

Normalization issues associated with data obtained from a single slide is called within slide normalization. The target of the within slide normalization is removing the eventual curvature from the RI plot. There are several strategies that consist in subtracting a slide specific function from the individual log ratios. Global, intensity dependent, within print tip group and scale normalization will be provided in the following.

5.4.3 Global normalization

The simplest approach to within slide normalization is the global which consists in subtracting a constant from all intensity log ratios, typically their mean or median. This method assumes that red and green intensity are related by a constant factor and shifts to zero the center of distribution of log ratio. For every gene i on the array:

$$\begin{aligned} Ri &= kG_i \ , \\ \log_2\left(\frac{R_i}{G_i}\right) &\rightarrowtail \log_2\left(\frac{R_i}{G_i}\right) - c = \log_2\left(\frac{R_i}{kG_i}\right) \quad \forall i \in \Im \ , \end{aligned}$$

where the set of index $\Im$ denotes all the genes spotted on the array.

5.4.4 Intensity dependent normalization

In almost all experiments spatial or intensity dependent biases are evident so that an intensity dependent normalization method is required. In the intensity dependent normalization a local regression line is fitted to the MA plot via locally weighted least square methods, *lowess*, and then the data are recentered along this line. The lowess function was first introduced by Cleveland (1979) and it first appeared in microarray context in Luu at al (2001) as a tool to normalize microarray data. Under the assumption that most genes are equally expressed in both channels, the overall intensity level in the array can be approximated by A, in fact for almost all $i \in \Im$:

$$\log_2(R_i) = \log_2(G_i) \rightarrowtail A_i = \log_2[(R_iG_i)^{1/2}] = \frac{\log_2(R_i) + \log_2(G_i)}{2} = \log_2(R_i) = \log_2(G_i).$$

Fitting the lowess function c(A) to the MA plot leads to:

$$\begin{aligned} R_i &= k(A_i)G_i, \\ M_i &= \log_2\left(\frac{R_i}{G_i}\right) = \log_2[k(A_i)] = c(A_i) \rightarrowtail k(A_i) = 2^{c(A_i)} = k_i. \end{aligned}$$

Then the data will be corrected as follows:

$$\log_2\left(\frac{R_i}{G_i}\right) \rightarrowtail \log_2\left(\frac{R_i}{G_i}\right) - c(A_i).$$

This is equivalent to correct both channels intensity value as:

$$\begin{cases} R_i & \rightarrow R_i, \\ G_i & \rightarrow k(A_i)G_i = k_iG_i. \end{cases}$$

While the global normalization method transforms all the genes using a unique value for every slide, the lowess normalization appears most suitable to reduce the effect of the two dyes.

5.4.5 Within print tip group normalization

Genes spotted on an array are grouped in grids. Every grid is printed using the same print tip. The printing set up depends on the design and on the target of the experiment. The print tips may be affected by systematic differences as unequal length of the tips or deformation after many hours of printing and this may cause spatial effects on the slide. Thus it could be necessary to apply a normalization depending both on intensity and on print tip group, the so called within print tip group normalization:

$$\log_2\left(\frac{R_i}{G_i}\right) \rightarrow \log_2\left(\frac{R_i}{G_i}\right) - C_\lambda(A_i), \ \ \forall\lambda \in 1,\ldots,\Lambda, \forall i \in \Im_\lambda,$$

where $C_\lambda(.)$ is the lowess fit to the MA plot for the λ-th grid, Λ represents the number of grids and $\Im_\lambda$ is the index set of the genes spotted in the λ-th grid.

5.4.6 Scale normalization

The log ratios from different grids, normalized by the within print tip group method, will result centered around zero. Even if this behavior satisfy our first request, it is also necessary that data from different print tip groups have the same spreadness. In order to reach this goal a scale adjustment is required. An example of scale normalization can be obtained under the assumption that log ratio data from the i-th grid are distributed as a Normal with mean zero and variance $a_i^2\sigma^2$ where a_i^2 is the specific scale factor of the i-th print tip group and σ^2 is the true log ratio. The scale parameter a_i can be estimated via maximum likelihood under the constraint $\sum_{=1}^{\Lambda}\log(a_i^2) = 0$. Let n_i be the number of genes in the i-th print tip group and let $M_{ij} = \log(\frac{R_i}{G_i})$, $i = 1,\ldots,\Lambda$ and $j = 1,\ldots,n_i$. It results that

$$a_i^2 = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{(\Pi_{i=1}^{\Lambda}\sum_{j=1}^{n_i} M_{ij}^2)^{1/2}}.$$

5.4.7 Multiple slide normalization

Reguardless the normalization method used, in the within slide normalization step, the normalized log ratio will turn out to be centered around zero. After this first step it is necessary to make all the data from different slide comparable, in the sense that log ratios from different slide should have similar spread. The target of multiple slide normalization is just to allow comparisons between experiments. This kind of normalization may also be performed using the method described in section 5.4.6.

5.5 Yeast experiment

5.5.1 Experiment description

In this section we describe a real data cDNA microarray experiment on the whole Yeast genome. The aim of biologist's study is to identify target genes, whose transcription activation is dependent on Cdk1. The starting point was to test Cdk 1's role in recruiting the SRB/Mediator and Pol II to the whole yeast genome. Wild type GAL-CDC20 and cdc28-13 GAL-CDC20 mutant cells were arrested in metaphase by incubating cells at the permissive temperature (25 C) in medium lacking galactose. The cultures were then shifted to 37 to inactivate Cdk 1 and 20 minutes later were induced to enter G1 by the re-addiction of galactose. The cdc28-13 mutants remained as unbudded cells failing to re-enter S phase. A similar experiment was performed using SRB4-Myc GAL-CDC20 and SRB4-Myc cdc28-13 GAL-CDC20. Samples were taken each 10-20 minutes starting from time point zero (arrested cells) until 110 minutes after release. Cells were formaldehyde crosslinked and distrupted with glass beads. After shearing of the chromatin by sonication, the crosslinked DNA-protein complexes were immunoprecipitated either with the monoclonal antibody anti Myc (9E11) to recover the SRB4's associated chromatin fragments or with the mouse monoclonal antibody 8WG16 against the C terminal domain of Rbp1 to recover PolII 's associated chromatin fragments. The crosslinks were reversed by incubation at 65 C and the recovered chromatin fragment purified. To control the efficiency of the experiments, PCR amplification on the purified chromatin was performed to verify the association of known target genes. The biologists want to analyze the genome wide association of metaphase arrested cells (time point 0) of G1 cells (time point 40) and of cells either still arrested in G1 in the cdc28-13 mutant or already re-entered in the next cell cycle (time point 60 and 90). The chromatin purified was analyzed from the time point 40. Chromatin fragments from wild type GAL-CDC20 and cdc28-13 GAL-CDC20 and from SRB4-Myc GAL-CDC20 and SRB4- Myc cdc28-13 GAL-CDC20 were randomly amplified. Chromatin was first amplified with degenerated tagged oligonucleotides by the use of the sequenase enzyme. A PCR was then performed onto the obtained template using oligonucleotides complementary to the tag. Fragments ranging between 500 and 1000 bp were clearly visible onto agarose gel after PCR amplification. To test the non-repetitively abundance of some PCR amplified target genes with respect to others, we cloned the SRB4's associated amplified fragments in a T vector. Millions of colonies were obtained and we only 100 of them were analyzed by sequence. Finally, the amplified chromatin samples were labelled with Cy3 and Cy5 fluorophores and combined two by two: samples from Srb4 IP together with samples from Srb4 cdc28-13 IP, and samples from Pol II IP together with samples from Pol II cdc28-13 IP. The purified probes were hybridized to arrays containing 6400 yeast ORFs. The main target of this experiment is to analyze the arrays and individuate genes which are dependent or independent of Cdk1.

5.5.2 Methods

In this section we briefly describe the processing of the Yeast experiment data. For our analysis purpose, we consider this experiment as consisting of two related sets of data: SRB and POL II. The number of arrays employed to test the Cdk 1's role is 6 for SRB and 9 for Pol II. Each array is a kind of treatment vs control experiment, where the treatment is the cell at the mutant stage, while the control is the cell at the wild type stage. The number of technical replicates of each gene on each array is 2. The number of gene spotted on each array is m=6400. After image processing and normalization step (see previous sections), the gene expression data were summarized by a bi-dimensional array of log intensities ratios **X** of components,

$$X_{ji} = \log_2 \left(\frac{WT_{ji}}{MTji} \right) \ \ , \ \ j = 1, \ldots, m \ \ i = 1, \ldots, n_j,$$

where WT_{ji} and MT_{ji} are respectively the fluorescence intensities measuring the expression level of the gene j spot i, in the wild type condition and in the mutant condition. We measure each gene spot expression level, both in WT and MT, by the average background corrected fluorescence intensity.

We want to find out the genes that are not differentially expressed, between WT and MT, in SRB and that, at the same time, are over expressed in WT with respect to MT in POL II. We formulate this problem in terms of multiple hypothesis testing and we compare different strategies to reach our goal. The basic idea is to:

1. Consider a *t-statistic* T_j for each gene j

$$T_j = \sqrt{n} \, \frac{\bar{X}_j - \theta_j}{S_j},$$

where the gene standard deviation, S_j, is opportunely estimated from the data (different corrections were also considered.

2.Choose a multiple testing procedure (MTP) to test on the SRB dataset, simultaneously for each gene j, the following assumption

$$H_{0j} : \theta_j \neq 0 \quad \text{vs} \quad H_{1j} : \theta_j = 0\ , \tag{5.5.1}$$

where θ_j denotes the expression level of gene j. Let G be the set of gene selected by the choosen MTP.

3. Choose a MTP to test on the POL II dataset, simultaneously for each gene j in G, the following assumption

$$H_{0j} : \theta_j = 0 \quad \text{vs} \quad H_{1j} : \theta_j > 0. \tag{5.5.2}$$

The resulting selected genes are then considered the target genes of the described Yeast experiment. We tried different way to implement these three steps. In the step 1, three different estimates of S_j were considered: the sample standard deviation, a SAM like estimate (see Tusher *et al*, 2001) and a percentile estimate (see Efron *et al*, 2001). For the SRB dataset we used the MAP and the *empirical* MTP, described in Chapter 2. For the POL II dataset we adopted the MAP MTP. Note that the Bayesian testing approach assumes the normal distribution on the error term. Such assumption is often criticized when microarray data are considered. In order to avoid this assumption we consider also an empirical approach on SRB, which is similar in spirit to the MAP MTP; i.e. we select as non differentially expressed in SRB those genes whose $|T_i|$ is lower then a user selected threshold. As a second stage a frequentist adaptive FDR controlling procedure is used. P-values are estimated by resampling methods (see Section 2.4). For each choice lists of genes where obtained and the common genes were selected. These different strategies were chosen indifferently and the resulting lists of genes were compared. The common genes were selected. The results were almost stable in a number of about fifty genes. Positive controls in the literature database were found. The biologists are now analyzing the final list of genes in order to confirm the findings.

Chapter 6

LPM: simulations, comparisons and real life example

Introduction

In this chapter we apply the thresholding rules proposed in chapter three. In the first Section we discuss the selection of the hyperparameters for each model. This is important for an automatic application of the methodology. In the first Section we compare performance of the proposed rules to eight other commonly used methods (both global and adaptive). In the last Section we apply the shrinkage methodology to a real-life example involving Atomic Force Microscopy.

6.1 Selection of hyperparameters

In any Bayesian modeling task the selection of the hyperparameters is instrumental for good performance of the model. It is also desirable to have an automatic way to select the hyperparameters, thus making the shrinkage procedure automatic, i.e., free of subjective user intervention. More about specification of hyperparameters in Bayesian models in the wavelet denoising context can be found in Chipman, Kolaczyk, and and McCulloch (1997), Vidakovic and Ruggeri (2001), and Angelini and Sapatinas (2004), among others.

The hyperparameters should be selected so that the resulting methodology is robust with respect to a wide range of input signals (sample sizes, signal regularity, size of noise, etc). In contemporary wavelet practice the values of the hyperparameters are usually assessed by empirical Bayes arguments due to enormous variability of potential input data (Clyde and George, 1999; 2000). Straightforward Empirical Bayes techniques such as predictive moment matching, or MLII method, are most commonly used efficient methods for hyperparameter specification. In this paper we determine hyperparameters by moment-matching.

In chapter three we considered three Bayesian models on wavelet coefficients, (i) the basic model with σ^2 assumed known, and two generalizations in which the variance σ^2 is modeled by (ii) exponential and (iii) inverse-gamma priors. In this section the elicitation of corresponding hyperparameters is discussed for each case.

The Basic Model. In the basic model the only hyperparameter is the *power parameter* k. Even though the proper posterior is obtained for $k < 1$, the existence of the second, non-zero mode does not depend on the "properness" of the posterior. Thus we will consider all $k > 1/2$. Note that the condition $k > 1/2$ is needed to ensure that Gamma function $\Gamma(2k-1)$ is finite and non-negative.

The sample size of the input signal should influence our selection of k. Figure 6.1 shows the Bumps signal at SNR =5 and sample sizes $n = 512, 1024, 2048,$ and $4096,$ smoothed by LPM thresholding rule (3.5.4) for various values of k. The minimum average mean square error ($AMSE$) is achieved at $k = 1.0, 1.2, 1.4,$ and 1.6 respectively. Thus the increasing of the sample size increases the optimal k.

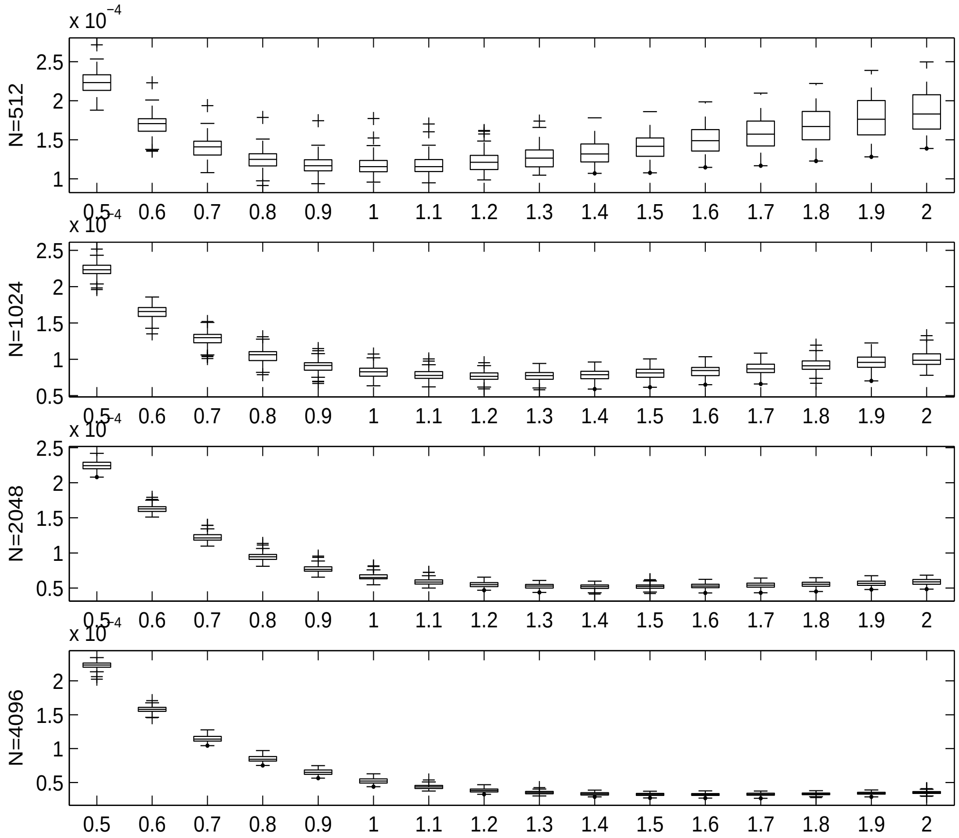


Figure 6.1: The AMSE for the Bumps function for four different sample sizes $n = 512$ (top), $n = 1024$, $n = 2048$, $n = 4096$ (bottom), evaluated at different values of power parameter k. The level of noise is such that SNR=5. The thresholding rule used was (3.5.4).

Another feature of the signal is also important for specifying k - signal regularity. The power parameter k is small if the signal (to be estimated) is irregular. Figure 6.2 illustrates this relationship. Four standard test signals, Bumps, Blocks, HeaviSine and Doppler of size $n = 1024$ are considered at SNR=5. Bumps is an irregular signal. The optimal k was 1.2.

HeaviSine is the most regular signal with optimal value 1.8. Blocks and Doppler exhibit irregularities of different nature (Blocks is a piecewise constant, but discontinuous, while Doppler is smooth but with time varying frequency). For both the optimal value of k was 1.6.

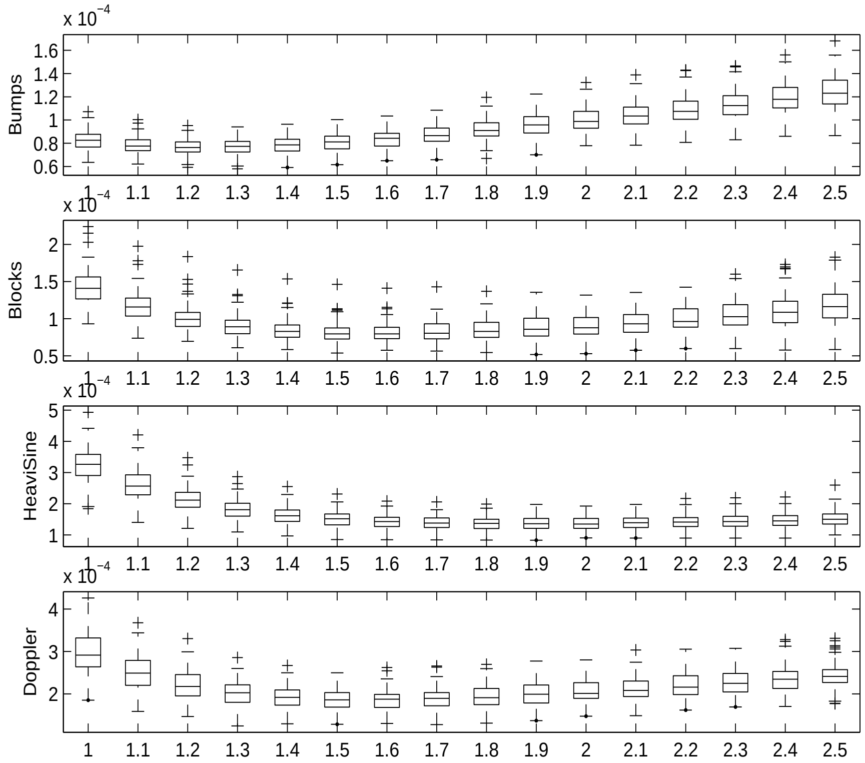
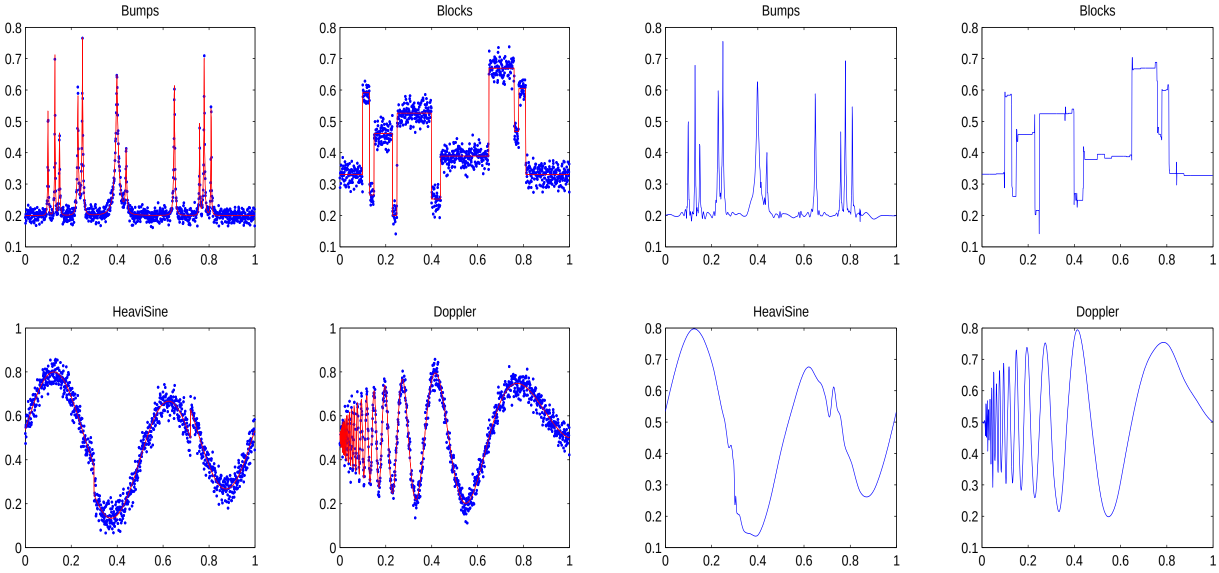


Figure 6.2: Boxplots of the AMSE for the various values of the power parameter k for four test signals Bumps, Blocks, HeaviSine, and Doppler. Sample size was $n = 1024$ and SNR = 5.

Taking into account the above analysis, a single universal value of k for an automatic use of the rule (3.5.4) should be chosen from the interval $(1, 2)$.

Figure 6.3 shows the true signals and the noisy signals based on $n = 1024$ design points at SNR=5 along with the reconstruction obtained after thresholding the coefficients



(a) (b)

Figure 6.3: (a) Test signals with superimposed noisy versions at SNR=5. The sample size is $n = 1024$. (b) Estimates obtained by using the LPM method with optimal k.

by LPM method with optimal k. we can see that LPM method does a very good job at removing the noise. From Figure 6.4 we can see the change in smoothness of recovered signals with the change of k.

Model 1. In the model with an exponential prior on σ^2 in addition to the power parameter k we also have the hyperparameter μ which is the reciprocal of the scale parameter. Given an estimator $\hat{\xi}$ of the noise variance, a moment-matching choice for μ would be $\hat{\mu} = \frac{1}{\hat{\xi}}$. Donoho and Johnstone (1994) suggested to estimate the noise level σ by the median absolute deviation (MAD) of the wavelet coefficients at the finest level adjusted by 1/0.6745, our choice is to consider $\hat{\xi} = \text{MAD}^2$. In this model we will consider $k > 1/2$ with no upper bounds because, even if the posterior distribution is not proper, the choice of $k > 1$

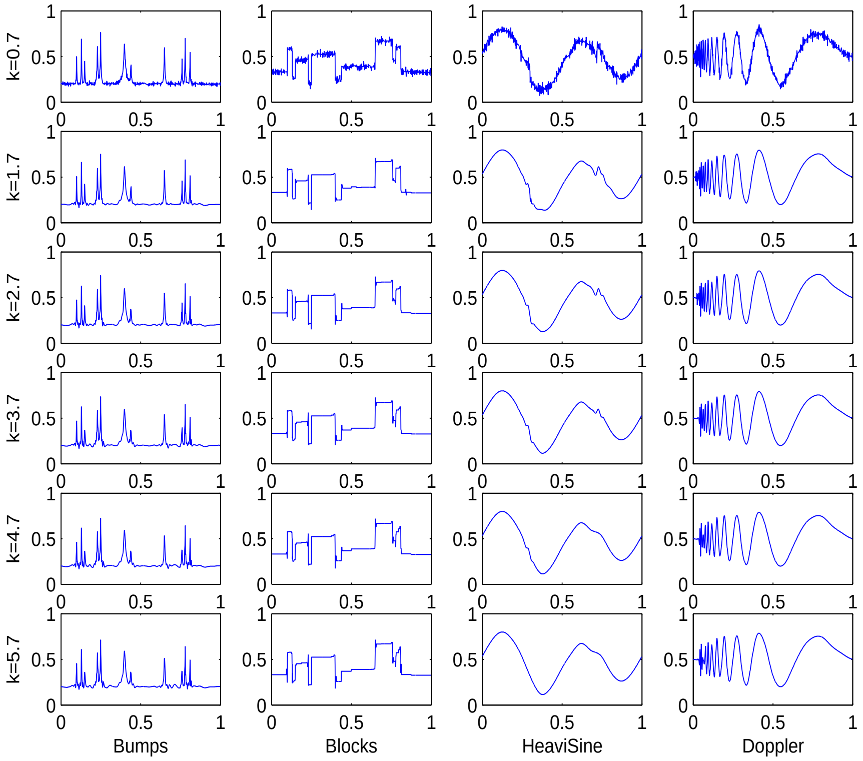


Figure 6.4: Estimates obtained using LPM method for roughly selected k, based on n=1024 points at SNR=5.

does not affect the existence of the non-zero mode.

As could be seen from Figure 3.4(b) the rule resulting from this model coincides with a hard-thresholding rule and clearly differs from the basic model rule displayed in Figure 3.2(b). Therefore, we anticipate different behavior of the optimal k as the sample size increases. Our simulations reveal that the optimal power parameter is dependent on the sample size and on the regularity of the signal in this model, as well, but the minimum AMSE is achieved at larger values of k compared to the basic model under the same test conditions. For instance, for the Bumps signal at SNR=5 and $n = 512, 1024, 2048$ and 4096

the optimal values of k are $2.1, 2.4, 2.6$ and 2.8, respectively, while for the four standard test signals Bumps, Blocks, HeaviSine and Doppler of size $n = 1024$ at SNR=5 the optimal values of k are $2.4, 2.7, 3.0$ and $2.7.$ Therefore, for an automatic use of the thresholding rule in the exponential model, a single universal value of k should be selected from the interval $(2, 3)$.

Model 2. In the model with an inverse gamma prior on σ^2 in addition to the power parameter k we also have two new hyperparameters α and β which specify the prior. As in Model 1 we will match the prior moments with the observed moments in order to specify the hyperparameters. The n-th moment of an inverse gamma random variable $X \sim \mathcal{IG}(\alpha, \beta)$ is

$$EX^n = \frac{\beta^n}{(\alpha - 1) \ldots (\alpha - n)}.$$

Thus, the first two moments matched with the corresponding empirical moments of wavelet coefficients from the finest level of detail will "estimate" α and β. This consideration and Gaussianity of the noise yields $\alpha = 2.5$ and $\beta = 1.5\, \hat{\xi}$, where $\hat{\xi}$ is some estimator of the variance of the noise. As in the previous models we use the robust $(MAD)^2$ estimator. An argument for the specification of α and β are given in the Appendix.

As in the previous two cases, we anticipate different behavior of the optimal k with respect to the sample size and regularity of test functions. For instance, for $\alpha = 2.5$ and β determined using α and an estimator of the variance of the noise for Bumps signal with SNR = 5 the AMSE minimizing values of k are 1.3, 1.6, 1.8, and 2.0 for sample sizes 512, 1024, 2048, and 4096, respectively. For the four standard test signals Bumps, Blocks, HeaviSine and Doppler of size $n = 1024$ at SNR=5 the optimal values of k are $1.6, 2.0, 2.3$

and 2.0. Therefore, for an automatic use of the thresholding rule in the inverse gamma model, a single universal value of k should be selected from the interval $(1, 3)$.

6.2 Simulations and comparisons

We present a simulation study of the performance of LPM method for the three models. The simulation is done with the "known truth", that is with test functions specified, and controlled signal-to-noise ratio. We also compare the average mean square error (AMSE) performance with several popular methods.

For our simulation study, four standard test functions (`Bumps, Blocks, HeaviSine` and `Doppler`) were added rescaled normal noise to produce a preassigned signal-to-noise ratio (SNR). For each method, test functions were simulated at $n = 512, 1024,$ and 2048 points equally spaced on the unit interval. Three commonly used SNR's were selected: SNR=3 (weak signal), 5 (moderate signal), and 7 (strong signal). The wavelet bases are also standard for the above test functions: Symmlet 8 for `HeaviSine` and `Doppler`, Daubechies 6 for `Bumps` and Haar for `Blocks`.

Closeness of the reconstruction to the theoretical signal of each method was measured by an average mean-square error (AMSE), calculated over 1000 simulation runs. In each case, the optimal power parameter k (minimizing AMSE) was used. All computations are carried out using MATLAB, with the WaveLab toolbox (see Buckheit, Chen, Donoho, Johnstone, and Scargle, 1995) and the GaussWaveDen toolbox (see Antoniadis, Bigot, and Sapatinas, 2001).

The results are summarized in two tables. Table 6.1 gives minimum AMSE for the three introduced models at three SNR levels and for four standard test functions, while Table 6.2

presents the corresponding optimal value of the power parameter k.

Final Results × 10^{-3}										
Function	n	SNR=3			SNR=5			SNR=7		
Bumps	512	0.2825	0.3116	0.2875	0.1079	0.1180	0.1095	0.0570	0.0621	0.0577
	1024	0.1953	0.2150	0.1993	0.0733	0.0802	0.0745	0.0373	0.0401	0.0379
	2048	0.1254	0.1371	0.1282	0.0469	0.0509	0.0477	0.0240	0.0257	0.0244
Blocks	512	0.3820	0.4111	0.3876	0.1202	0.1265	0.1213	0.0553	0.0568	0.0554
	1024	0.2752	0.3004	0.2790	0.0802	0.0827	0.0800	0.0359	0.0364	0.0357
	2048	0.1584	0.1692	0.1601	0.0480	0.0502	0.0483	0.0201	0.0204	0.0200
HeaviSine	512	0.4066	0.4305	0.4155	0.2243	0.2441	0.2300	0.1432	0.1575	0.1472
	1024	0.2769	0.2966	0.2835	0.1353	0.1443	0.1379	0.0890	0.0964	0.0914
	2048	0.1711	0.1786	0.1734	0.0950	0.1007	0.0969	0.0604	0.0666	0.0622
Doppler	512	0.7046	0.7706	0.7187	0.2725	0.2959	0.2767	0.1449	0.1557	0.1470
	1024	0.4491	0.4879	0.4590	0.1896	0.2062	0.1931	0.1032	0.1125	0.1054
	2048	0.2514	0.2649	0.2540	0.1064	0.1135	0.1081	0.0596	0.0639	0.0607

Table 6.1: AMSE for the Basic Model (left), Model 1 (center), and Model 2 (right) at different SNR levels and for the four standard test functions.

We also compare LPM method with several established wavelet-based estimators for reconstructing noisy signals. In particular we consider the term-by-term Bayesian estimator *BAMS* of Vidakovic and Ruggeri (2001), the classical term-by-term estimators *VisuShrink* of Donoho and Johnstone (1994) and *Hybrid-SureShrink* of Donoho and Johnstone (1995), the scale invariant term-by-term Bayesian *ABE* method of Figueiredo and Nowak (2001), the “leave-out-half” version of the *Cross-Validation* method of Nason (1996), the term-by-term False Discovery Rate (*FDR*) method of Abramovich and Benjamini (1995), and finally *NeighCoeff* of Cai and Silverman (2001) and *BlockJS* of Cai (1999) which represent classical estimators that incorporate the blocking procedure to achieve a better performance. Note that, for excellent numerical performance, we consider the *VisuShrink* and the “leave-out-half” version of the *CrossValidation* methods with the hard threshold and the *BlockJS* with the option ’Augment’ (see Antoniadis, Bigot, and Sapatinas, 2001).

The LPM is a global method, i.e., the model parameters/hyperparameters are common

Final Results Optimal k										
Function	n	SNR=3			SNR=5			SNR=7		
Bumps	512	1.0	2.1	1.4	1.0	2.1	1.3	1.0	2.0	1.3
	1024	1.2	2.4	1.6	1.2	2.4	1.6	1.2	2.4	1.6
	2048	1.4	2.6	1.8	1.4	2.6	1.8	1.4	2.6	1.8
Blocks	512	1.4	2.5	1.8	1.4	2.6	1.8	1.5	2.7	1.9
	1024	1.5	2.6	1.9	1.6	2.7	2.0	1.6	2.8	2.1
	2048	1.6	2.8	2.1	1.7	2.9	2.2	1.8	2.9	2.2
HeaviSine	512	1.9	3.4	2.4	1.7	2.8	2.1	1.5	2.8	2.0
	1024	2.0	3.2	2.4	1.8	3.0	2.3	1.7	3.0	2.2
	2048	2.1	3.2	2.6	2.0	3.2	2.4	1.8	2.9	2.2
Doppler	512	1.4	2.6	1.8	1.4	2.6	1.8	1.4	2.5	1.8
	1024	1.6	2.8	2.1	1.6	2.7	2.0	1.5	2.7	1.9
	2048	1.8	3.0	2.3	1.8	3.0	2.2	1.7	2.9	2.2

Table 6.2: Values of optimal k for the Basic model (left), Model 1 (center), and Model 2 (right) at different SNR levels and for the four standard test functions.

across the scales in wavelet decompositions. Models for which the parameters/hyperparameters are level-dependent are called adaptive. To avoid confusion, we note that term adaptive is also used in large sample theory for parameters/methods that do not affect the convergence rates. Four of the methods contrasted to LPM are global (*VisuShrink, ABE, CrossValidation* and *FDR*), while the four remaining methods (*BAMS, Hybrid-SureShrink, NeighCoeff* and *BlockJS*) are adaptive .

Figure 6.5 presents the boxplots of the AMSE computed for the above 9 methods based on $n = 1024$ design points at SNR=5. It is clear that LPM method outperforms well-known methods such as VisuShrink, Cross-Validation, FDR and BlockJS methods, and often performs comparably to (sometimes even better than) BAMS, Hybrid-SureShrink, ABE and NeighCoeff methods.

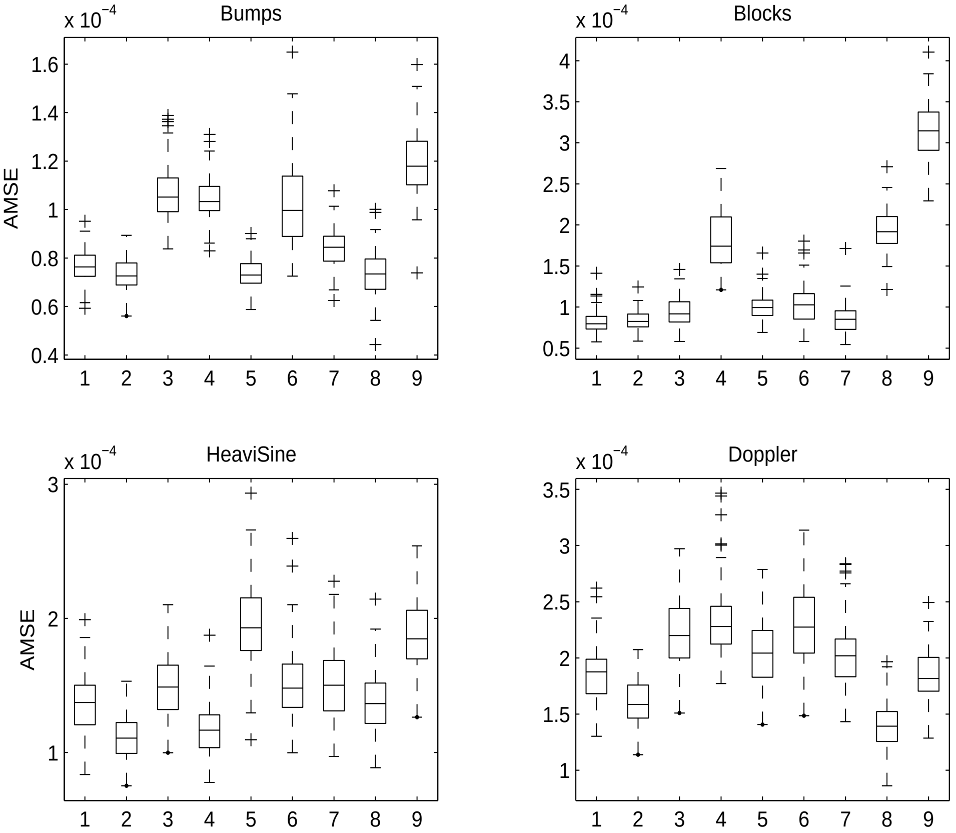


Figure 6.5: Boxplots of the AMSE for the various methods (1) LPM, (2) BAMS, (3) VisuShrink, (4) Hybrid, (5) ABE, (6) CV, (7) FDR, (8) NC, (9) BJS, based on $n = 1024$ points at SNR=5.

6.3 An example in atomic force microscopy

To illustrate the performance of the LPM thresholding method proposed here, we estimate an underlying smooth function in the noisy measurements from an atomic force microscopy (AFM) experiment.

AFM is a type of scanned proximity probe microscopy (SPM) that can measure the adhesion strength between two materials at the nanonewton scale (Binnig, Quate and Gerber, 1986). In AFM, a cantilever beam is adjusted until it bonds with the surface of a sample, and then the force required to separate the beam and sample is measured from the beam deflection. Beam vibration can be caused by factors such as thermal energy of the surrounding air or the footsteps of someone outside the laboratory. The vibration of a beam acts as noise on the deflection signal; in order for the data to be useful this noise must be removed.

The AFM data from the adhesion measurements between carbohydrate and the cell adhesion molecule (CAM) E-Selectin was collected by Bryan Marshal from the BME Department at Georgia Institute of Technology. The detailed technical description is provided in Marshall, McEver, and Zhu (2001).

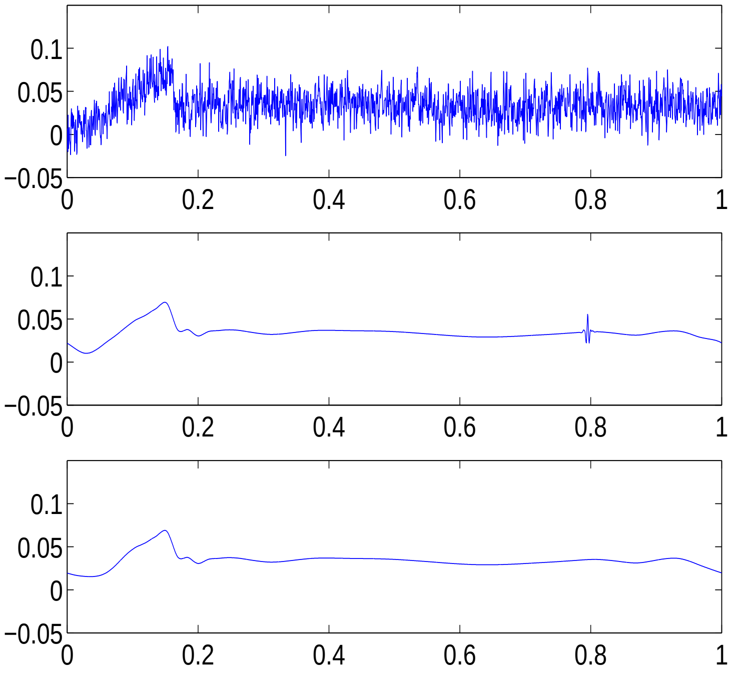


Figure 6.6: Original AFM measurements (top), LPM estimator with the default parameter $k = 1$ (middle), LMP estimator with the parameter $k = 1.4$ (bottom).

In Figure 6.6 the top panel shows the original noisy data. The middle panel shows the LPM estimate with the default parameter $k = 1$, while the bottom panel shows LPM estimate with the parameter $k = 1.4$. The sample size was $n = 2^{11}$ and Symmlet 8-tap filter was used to obtain the estimate. We observe that the latter estimate exhibits slightly smoother behavior, especially in the long-middle part without oversmoothing the "ramp-like" structure which is the feature of interest here.

We adhere to the concept of reproducible research (Buckheit and Donoho, 1995). The m-files used for calculations and figures in this work can be downloaded from Jacket's Wavelets page `http://www.isye.gatech.edu/~brani/wavelet.html`.

Appendix A

A.1 Proof of Lemma (3.6.1)

Assume that for a typical wavelet coefficient d the following model holds.

$$\begin{array}{rcl} d|\theta,\sigma^2 & \sim & \mathcal{N}(\theta,\sigma^2), \\ \sigma^2 & \sim & \mathcal{E}\left(\frac{1}{\mu}\right) \text{ with density } p(\sigma^2|\mu)=\mu e^{-\mu\sigma^2},\ \mu>0, \\ \theta|\tau^2 & \sim & \mathcal{N}(0,\tau^2), \\ \tau^2 & \sim & (\tau^2)^{-k}, k>\frac{1}{2}. \end{array}$$

It well known that the marginal likelihood, as a scale mixture of normals, is

$$d|\theta \sim \mathcal{DE}\left(\theta,\frac{1}{\sqrt{2\mu}}\right), \quad \text{with density } f(d|\theta)=\frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}.$$

Therefore the model can be rewritten as

$$\begin{array}{rcl} d|\theta & \sim & \frac{1}{2}\sqrt{2\mu}e^{-\sqrt{2\mu}|d-\theta|}, \\ \theta|\tau^2 & \sim & \mathcal{N}(0,\tau^2), \\ \tau^2 & \sim & (\tau^2)^{-k}, \frac{1}{2}. \end{array}$$

The joint distribution of d and θ is proportional to

$$\begin{aligned} p(d,\theta) &\propto \int_0^\infty p(d|\theta)p(\theta|\tau^2)p(\tau^2)d\tau^2 \\ &= \frac{1}{2}\sqrt{\frac{\mu}{\pi}}e^{-\sqrt{2\mu}|d-\theta|}\int_0^\infty e^{-\theta^2/(2\tau^2)}\frac{1}{(\tau^2)^k}d\tau^2 \\ &= \frac{1}{2}\sqrt{\frac{\mu}{\pi}}e^{-\sqrt{2\mu}|d-\theta|}\int_0^\infty y^{(k-1/2)-1}e^{-\theta^2 y/2}dy \\ &= \frac{1}{2}\sqrt{\frac{\mu}{\pi}}e^{-\sqrt{2\mu}|d-\theta|}\Gamma\left(k-\frac{1}{2}\right)\left(\frac{\theta^2}{2}\right)^{1/2-k}, k > 1/2. \end{aligned}$$

Furthermore we have

$$p(\theta|d) \propto p(d,\theta) \propto e^{-\sqrt{2\mu}|d-\theta|}\left(\theta^2\right)^{1/2-k}.$$

The likelihood of θ

$$l(\theta) = e^{-\sqrt{2\mu}|d-\theta|}\left(\theta^2\right)^{1/2-k} \tag{A.1.1}$$

is integrable if and only if $k < 1$.

The eventual modes of the posterior $p(\theta|d)$ exist if and only if they maximize the function (A.1.1), that is if and only if they maximize $L(\theta) = \log[l(\theta)]$. More explicitly

$$L(\theta) = \log[l(\theta)] = -\sqrt{2\mu}|d-\theta| + 1 - 2k\log\theta. \tag{A.1.2}$$

Consider its derivative

$$L' = \sqrt{2\mu}\,\text{sign}(d-\theta) + \frac{1-2k}{|\theta|}\,\text{sign}(\theta) = \sqrt{2\mu}\,\text{sign}(d-\theta) + \frac{1-2k}{\theta}, \tag{A.1.3}$$

and WLOG, suppose $d > 0$. Observe that the critical points of (A.1.3) are $\hat{\theta}_1 = 0$ and $\hat{\theta}_2 = \lambda = \frac{2k-1}{\sqrt{2\mu}}$.When $d < \lambda$ there exists only one mode in zero. When $d > \lambda$ there exists two modes, the smaller is zero and the larger is d; in fact the function (A.1.2) is decreasing between zero and lambda, increasing between lambda and d and decreasing after d.

A.2 Proof of lemma (3.6.2)

The model considered was

$$\begin{aligned} d|\theta,\sigma^2 &\sim \mathcal{N}(\theta,\sigma^2), \\ \sigma^2 &\sim \mathcal{IG}(\alpha,\beta) \text{ with density } p(\sigma^2|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-1-\alpha}e^{\frac{-\beta}{\sigma^2}}, \alpha>0,\beta>0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, k>\frac{1}{2}. \end{aligned}$$

It is well known that t distribution is a scale mixture of normals, with mixing distribution being an inverse gamma.

$$d|\theta \sim \frac{1}{\sqrt{2\beta}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}, \quad \text{where } \mathcal{B}\left(\frac{1}{2},\alpha\right)=\frac{\Gamma(\frac{1}{2})\Gamma(\alpha)}{\Gamma(\frac{1}{2}+\alpha)}.$$

Therefore the model can be rewritten as

$$\begin{aligned} d|\theta &\sim \frac{1}{\sqrt{2\beta}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}, \alpha>0,\beta>0, \\ \theta|\tau^2 &\sim \mathcal{N}(0,\tau^2), \\ \tau^2 &\sim (\tau^2)^{-k}, \frac{1}{2}. \end{aligned}$$

The joint distribution of d and θ is proportional to

$$\begin{aligned}
p(d,\theta) &\propto \int_0^\infty p(d|\theta)p(\theta|\tau^2)p(\tau^2)d\tau^2 \\
&= \int_0^\infty \frac{1}{\sqrt{2\beta}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}\frac{1}{\sqrt{2\pi\tau^2}}e^{-\theta^2/(2\tau^2)}\frac{1}{(\tau^2)^k}d\tau^2 \\
&= \frac{1}{2\sqrt{\beta\pi}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}\int_0^\infty (\tau^2)^{-(k+1/2)}e^{-\theta^2/(2\tau^2)}d\tau^2 \\
&= \frac{1}{2\sqrt{\beta\pi}\mathcal{B}(\frac{1}{2},\alpha)}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}\int_0^\infty y^{(k-1/2)-1}e^{-\theta^2 y/2}dy \\
&= \frac{1}{2\sqrt{\beta\pi}\mathcal{B}(\frac{1}{2},\alpha)}\Gamma\left(k-\frac{1}{2}\right)\left(\frac{\theta^2}{2}\right)^{1/2-k}\left[\frac{(d-\theta)^2}{2\beta}+1\right]^{-\alpha-\frac{1}{2}}, k>1/2
\end{aligned}$$

Furthermore, we have

$$p(\theta|d)\propto p(d,\theta)\propto|\theta|^{1-2k}[(d-\theta)^2+2\beta]^{-\alpha-1/2}.$$

The likelihood of θ

$$l(\theta)=|\theta|^{1-2k}[(d-\theta)^2+2\beta]^{-\alpha-1/2}, \tag{A.2.1}$$

is integrable for any $k>\frac{1}{2}$.

The eventual modes of the posterior $p(\theta|d)$ exist if and only if they maximize the function (A.2.1). Since

$$\begin{aligned}
l' &= (1-2k)|\theta|^{-2k}\,\text{sign}(\theta)[(d-\theta)^2+2\beta]^{-\alpha-1/2}+|\theta|^{1-2k}(2\alpha+1)(d-\theta)[(d-\theta)^2+2\beta]^{-\alpha-3/2} \\
&= |\theta|^{-2k}\,\text{sign}(\theta)[(d-\theta)^2+2\beta]^{-\alpha-3/2}\{(1-2k)[(d-\theta)^2+2\beta]+(2\alpha+1)(d-\theta)\theta\},
\end{aligned}$$

it follows that

$$\begin{aligned}|\theta|^{-2k} &> 0, \forall\theta \in \mathcal{R} - \{0\},\\ \text{sign}(\theta) &> 0, \forall\theta > 0,\\ \left[(d-\theta)^2 + 2\beta\right]^{-\alpha-3/2} &> 0, \forall\theta \in \mathcal{R},\end{aligned}$$

and

$$l' = 0 \Leftrightarrow (1-2k)[(d-\theta)^2 + 2\beta] + (2\alpha+1)(d-\theta)\theta = 0,$$

with solutions

$$\theta_{1,2} = \frac{(2\alpha+4k-1)d \pm \sqrt{(2\alpha+1)^2 d^2 + 16(1-2k)(k+\alpha)\beta}}{4(k+\alpha)}.$$

The roots are real if and and only if $(2\alpha+1)^2 d^2 + 16(1-2k)(k+\alpha)\beta > 0$, i.e.,

$$|d| \geq \lambda = \frac{2}{2\alpha-1}\sqrt{(2k-1)(k+\alpha)\beta}\,. \tag{A.2.2}$$

If the condition (A.2.2) is not satisfied then the MAP is given by $\hat{\theta} = 0$. Now assume that (A.2.2) holds and $d > 0$. In this case both solutions $\theta_{1,2}$ are real and positive and the posterior is decreasing from zero to the smaller root, increasing between the two roots and decreasing again after the larger root. We have two posterior modes, the smaller is zero and the larger is

$$\hat{\theta} = \frac{(2\alpha+4k-1)d + \sqrt{(2\alpha+1)^2 d^2 + 16(1-2k)(k+\alpha)\beta}}{4(k+\alpha)}.$$

It is easy to see that $\hat{\theta}$ is always smaller then d, resulting in a shrinkage rule.

A.3 Selection of hyperparameters α and β in Model 2.

Note that for wavelet coefficients $(d_1, \ldots, d_m)$ from the finest level of detail the mean is close to 0, $\bar{d} \approx 0$. That means that $s_d^2 = \frac{1}{m-1}\sum(d_i - \overline{d})^2$ and $\frac{1}{m}\sum d_i^2 = \overline{d^2}$ are both comparable estimators of the variance. Also, even central empirical moments are approximately equal to the noncentral moments. The following two equations are approximately moment matching:

$$\overline{d^2} = \frac{\beta}{\alpha - 1}, \qquad \overline{d^4} = \frac{\beta^2}{(\alpha-1)(\alpha-2)},$$

where $\overline{d^4} = \frac{1}{m}\sum d_i^4$. From these equations we derive

$$\alpha = \frac{2\overline{d^4} - (\overline{d^2})^2}{\overline{d^4} - (\overline{d^2})^2},$$

which is free of the scale of wavelet coefficients. Since in the finest level of detail the contribution of signal is minimal and the wavelet coefficients are close to zero-mean normal random variables the Law of Large Numbers argument gives $\overline{d^2} \approx \sigma^2$ and $\overline{d^4} \approx 3\sigma^4$, which specifies the "shape" hyperparameter

$$\alpha = 2.5.$$

Hyperparameter β is determined from $\overline{d^2} = \frac{\beta}{\alpha-1}$, but instead of $\overline{d^2}$ we can use any estimator of variance of d. In simulations, we used the robust $(MAD)^2$.

Bibliography

Abramovich, F. and Angelini, C. (2005). Bayesian Maximum a Posteriori Multiple Testing Procedure. Technical Report, RP-SOR-05-01, Department of Statistics and Operations Research, Tel Aviv University.

Abramovich, F. and Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypotheses testing procedure. In *Wavelets and Statistics* (Antoniadis A. & Oppenheim G. Eds.) Lect. Notes Statist. **103**, 5-14, Springer-Verlag, New York.

Abramovich, F. and Sapatinas, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian Inference in Wavelet Based Models*, (Müller, P. & Vidakovic, B. Eds.) Lect. Notes Statist. **141**, 33–50, Springer-Verlag, New York.

Agbu, P. A. and James, M. E. (1994). NOAA/NASA Pathfinder AVHRR Land Data Set Users Manual, Goddard Distributed Active Archive Center, NASA Goddard Space Flight Center, Greenbelt.

Amato, U., Antoniadis, A. and Gregoire, G. (2003). Independent Component Discriminant Analysis. Int. J. Mathematics 3 735-753.

Angelini, C. and Sapatinas, T. (2004) Empirical Bayes approach to wavelet regression using e-contaminated priors. *Journal of Statistical Computation and Simulation*, **74**, 741 – 764.

Angelini, C. and Vidakovic, B. (2004). Γ-Minimax Wavelet Shrinkage: A Robust Incorporation of Information about Energy of a Signal in Denoising Applications, *Statistica Sinica,* **14**, 103–125.

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. J. Statist. Soft. **6**, 1–83.

Benjamini, Y. and Hochberg, Y. (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing*. Journal of the Royal Statistical Society (Ser. B), 57, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). *The control of the false discovery rate in multiple testing under dependency*. Annals of Statistics, 29, 11651188.

Benjamini, Y. Reiner,A. and Yekutieli, Y. (2003). *Identifying differentially expressed genes using the FDR* . Bioinformatics 19(3) 368-375.

Berger, J. O. (1985).*Statistical Decision Theory and Bayesian Analysis*. Second Edition. Springer, New York.

Besag J. (1996). *On Statistical Analysis of Dirty Pictures*. J. R. Statistical Society Series B 48 (3).

Binnig, G., Quate, C.F., and Gerber, Ch. (1986). Atomic force microscope. *Phys. Rev. Lett.* **56**, 930–933.

Bruce, A.G., and Gao, H-Y., (1996). Understanding WaveShrink: Variance and bias estimation. Biometrika **83**, 727–745.

Buckheit, J.B., Chen, S., Donoho, D.L., Johnstone, I.M. and Scargle, J. (1995). About WaveLab. Technical Report, Department of Statistics, Stanford University, USA.

Buckheit, J. and Donoho, D. L. (1995). Wavelab and reproducible research, in *Wavelets and Statistics*, A. Antoniadis and G. Oppenheim eds., LNS 103, Springer-Verlag, New York.

Cai, T.T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. Ann. Statist. **27**, 898–924.

Cai, T.T. and Silverman, B.W. (2001). Incorporating information on neighbouring coefficients into wavelet estimation. Sankhy$\bar{a}$ B, **63**, 127–148.

Chen, J. and Sakar, S K. (2004). *A Bayesian Stepwise Multiple Testing Procedure*. Technical Report, Temple University.

Chipman, H.A., Kolaczyk, E.D. and McCulloch, R.E., (1997) Adaptive Bayesian wavelet shrinkage. J. Amer. Stat. Assoc. **92**, 1413-1421.

Cleveland, W. S. (1979). *Robust locally weighted regression and smoothing scatterplots*. Journal of American Statistical Association, 74, 829–836.

Clyde, M. and George, E.I. (1999). Empirical Bayes estimation in wavelet nonparametric regression. In *Bayesian Inference in Wavelet Based Models*, (Müller, P. & Vidakovic,

B. Eds.), Lect. Notes Statist., **141**, pp. 309–322, Springer-Verlag, New York.

Clyde, M. and George, E.I. (2000). Flexible empirical Bayes estimation for wavelets. J. R. Statist. Soc. B, **62**, 681–698.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. Biometrika **85**, 391–401.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley and Sons, Inc., New York.

Cutillo, L. and Amato, U. (2005).Local Discriminant Analysis for cloud detection. Proceeding from Cladag, University of Parma, italy.

Cutillo, L., Amato, U., Antoniadis, A., Cuomo, V. and Serio, C. (2004). Cloud detection from multispectral satellite images. Proceedings from IEEE Gold Conference, University Parthenope, Naples, Italy.

Cutillo, L., Jung, Y. Y. , Ruggeri, F. and Viadakovic, B. (2005). Larger Posterior Mode Wavelet Thresholding and Applications. Tecnical Report. RT 29705. Istituto per le Applicazioni del Calcolo ”Mauro Picone” (IAC), Naples, Italy.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425–455.

Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. J. Am. Statist. Assoc. **90**, 1200–1224.

Donoho, D.L., Johnstone I., Kerkyacharian G. and Picard D. (1995). Wavelet shrinkage: asymptopia J.Roy. Statist. Soc. B **57**, 301-370.

Duda, O. R., Hart, P. E. and Stork, D. G. (2001). Pattern Classification. Second Edition. Wiley-Interscience.

Dudoit, S., Shaffer, P. J. and J.C. Boldrick (2003a). *Multiple hypothesis testing in microarray experiments*. Statistical Science, 18, 71-103.

Efron, B., (1979). *Bootstrap methods: another look at the jackknife*. Ann. Statist. 7, 1-26.

Efron, B., Tibshirani, R.J., (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). *Empirical Bayes analysis of a microarray experiment*. Journal of the American Statistical Association, 96, 1151-1160.

Figueidero, M.A.T. and Nowak, R.D. (2001). Wavelet-based image estimation: an empirical Bayes approach using Jeffrey's noninformative prior. IEEE Trans. Image Process. **10**, 1322–1331.

Gao, H-Y., and Bruce, A.G., (1997) WaveShrink with firm shrinkage. Statistica Sinica **7**, 855-874.

Genovese, C. R. and Wasserman, L. (2002). *Operating characteristics and extensions of the FDR procedure*. Journal of the Royal Statistical Society (Ser. B), 64, 499518.

Ju, J., Gopal, S., and Kolaczyk, E.D. (2005). *On the choice of spatial and categorical scale in remote sensing land cover characterization*. Remote Sensing of Environment, 96(1):62-77.

Benjamini,Y. and Hochberg,Y. (1995) C*ontrolling the false discovery rate: a practical and powerful approach to multiple testing*. J. Roy. Stat. Soci. B, 57, 289300.

Hochberg, Y. (1988) *a sharper Bonferroni procedure for multiple tests of significance*. Biometrika, 75, 800-803.

Hochberg, Y. and Benjamini Y. (1990). *More powerful procedures for multiple significance testing*. Statistics in Medicine, 9, 811-818.

Hochberg Y. and Tamhane A. C. (1987). *Multiple Comparisons Procedures*. Wiley.

Hollander, M. and Wolfe, D. A. (1999). Nonparametric Statistical Methods. Second Edition. Wiley Series in Probability and Statistics.

Holm, S. (1979). *A simple sequentially rejective multiple test procedure*. Scandinavian Journal of Statistics, 6, 65-70.

Hyvärinen, A. (1997). *Independent component analysis by minimization of mutual information*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 97), 39173920.

Huber, P. J. (1981).Robust Statistics. Wiley Series in Probability and Statistics.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical BAyes selection of wavelet thresholds. The Annals of Statistics, Vol. 33, No. 4, 1700-1752.

Jolliffe, I. T. (2002). Principal Component Analysis. Second Edition. Springer Series in Statistics.

Johnson, R. A. and Wichern, D. W. (1998). Applied Multivariate Statistical Analysis, Fourth Ed., Prentice Hall, Upper Saddle River, NJ.

Keuchel, J., Naumann, S., Heiler, M., et al. (2003) *Automatic land cover analysis for Tenerife by supervised classification using remotely sensed data*. REMOTE SENS ENVIRON 86 (4): 530-541.

Khedama, R. and Belhadj-Aissaa, A. (2004). CONTEXTUAL *Classification of Renotely Sensed Data Using MAP Approach and MRF*. Proceedings from Instambul ISPRS 2004. Vol. XXXV, part B7

Mallat, S. G. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Analysis Machine Intel.*, **11**, 674–693.

Marshall, B., McEver, R. and Zhu, C. (2001). Kinetic rates and their force dependence on the P-Selectin/PSGL-1 interaction measured by atomic force microscopy. Proceedings of ASME 2001, Bioengineering Conference, BED - Vol. 50.

Müller, P. and Vidakovic, B. (1999). MCMC methods in wavelet shrinkage. In: *Bayesian Inference in Wavelet-Based Models.* Editors Müller, P. and Vidakovic, B., Springer-Verlag, Lecture Notes in Statistics **141**, 187–202.

Nason, G.P. (1996). *Wavelet shrinkage using cross-validation*. J. R. Statist. Soc. B, **58**, 463–479.

Pensky, M. and Sapatinas, T. (2005). Frequentist optimality of Bayes factor estimators in wavelet regression models. Technical Report, TR082005, Department of Mathematics and Statistics, University of Cyprus, Cyprus.

Pollard, K. S. and Van der Laan, M. J. (2003). *Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data*. Tech. Rep. 121 , Division of Biostatistics, UC Berkeley.

Ripley, B. D. (1996). *Pattern Recognition and Neural Network*. Cambridge University Press.

Robbins, H. (1951). *Asymptotically sub-minimax solutions to compound statistical decision problems*. In Proc. Second Berkeley Symposium Math. Statist. and Prob. 1, 241-251. Berkeley University of California Press.

Ruggeri, F. and Vidakovic, B. (2005), *Bayesian modeling in the wavelet domain*, to appear in *Handbook of Statistics - vol. 25 - Bayesian Thinking, Modeling and Computation*, D. Dey and C.R. Rao, Eds., North Holland.

Sidak, Z (1967). *Rectangular confidence regions for the means of multivariate normal distributions*. Journal of the American Statistical Association, 62, 626-633.

Sidak, Z. (1971). *On probabilities of rectangles in multivariate Student distributions: their dependence on correlations*. Annals of Mathematical Statistics, 42, 169-175.

Silverman, B. W. (1986). *Density estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Simoncelli, E. and Adelson, E. (1996). Noise removal via Bayesian Coring, *Proceedings of 3rd IEEE International Conference on Image Processing*, Vol. I, pp. 379–382. Lausanne, Switzerland. 16-19 September 1996.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, practice, and Visualization*. Wiley, New York.

Storey, J.D. and Tibshirani, R. (2001). E*stimating false discovery rates under dependence, with applications to DNA microarrays*. Tech. Rep. 2001-28 , Department of Statistics, Stanford University.

Stowe, L. L., McClain, L. L., Carey, R., Pellegrino, P., Gutman, G. G., Davis, P., Long, C. and Hart, S (1991). Global distribution of cloud cover derived from NOAA/AVHRR operational satellite data, Adv. Space Research 3, 51 54.

Tusher, Tibshirani and Chu (2001).*Significance analysis of microarrays applied to the ionizing radiation response*. PNAS 2001 98: 5116-5121.

Van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2003). *Multiple testing. Part II. Stepdown procedures for control of the family-wise error rate*. Tech. Rep. 139 , Division of Biostatistics, UC Berkeley.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V. Ngai, J. and Speed, T. (2002). *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation*. Nucleic Acids Research, Vol. 30, No.4 e15.

Yang, Y. H., Dudoit, S., Luu, P. and Speed T. P. (2001). *Normalization for cDNA Microarray Data*. SPIE BiOS, San Jose, California.

Vidakovic, B., (1998). Non linear wavelet shrinkage with Bayes rules and Bayes factors. J. Amer. Stat. Soc. **45** B, 173-179.

Vidakovic, B. (1999). Statistical Modeling by Wavelets. Wiley, New York.

Vidakovic, B. and Ruggeri, F. (1999). Expansion estimation by Bayes rules, *J. Stat. Plann. Infer.*, **79,** 223–235.

Vidakovic, B. and Ruggeri, F. (2001). BAMS Method: Theory and Simulations. Sankhya B **63**, 234–249.

Wand, M. P. and Jones, M. C. (1995). Kernel Smoothing. First Edition. Chapman and Hall.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley.