# A wavelet approach to dimension reduction and classification of hyperspectral data

**by**

**Jon Wickmann**

***THESIS***

*for the degree of*

***MASTER OF SCIENCE***

*(Master i Modellering og dataanalyse)*


*Faculty of Mathematics and Natural Sciences*
*University of Oslo*

*May 2007*

*Det matematisk- naturvitenskapelige fakultet*
*Universitetet i Oslo*

**Abstract**

In this thesis I will exploit the fact that the wavelet representation of hyperspectral data is sparse. Techniques from both atomic decomposition and denoising will be modified and used to make an even sparser representation. Assessment will be done on three datasets. At face value my results are better than those of a baseline study with *principal component analysis (PCA)*, however no formal test supports this claim (the variability of the studies are to high). Formal tests show some improvement in fullfilling model assumptions for my methods. This is all done under *the curse of dimensionality* (i.e. few trainings samples and many parameters).

# CONTENTS

# CHAPTER 1

# Introduction

A guide to the thesis is given in section 1.5.

## 1.1 The curse of dimensionality

> In view of all that we have said in the foregoing sections, the many obstacles we appear to have surmounted, what casts the pall over our victory celebration? It is the curse of dimensionality, a malediction that has plagued the scientist from earliest days.
>
> - Richard Bellman (see Bellman (1961) page 94)

The above quote is in a process control context. Bellman worked on aircraft control systems, solving variational minimisation problems through dynamic programming.

The curse of dimensionality is not unique to the process control setting. In all mathematical disciplines dealing with more than a few dimensions, this curse manifests itself. In multivariate calculus determination of saddle points, global minimum and maximum, volume and area and such, gets harder by increased dimensionality. This is mostly due to the increased work required to perform (analytical) differentiation and integration. The work required, does not scale linearly by dimension.

This has its parallel in the numerical approximation to the aforementioned operations. It is in the numerical approximations that *the curse of dimensionality* really takes hold. A prime example is numerical integration (quadrature). Numerical integration of the function $f(\cdot)$ over the interval $[a, b]$ requires the function to be evaluated at $c$ points in this interval. $c$ varies according to which

numerical integration method being used. Generalised to $k$-dimensions, the interval becomes $[a, b]^k$ and the number of points evaluated becomes $c^k$. For all reasonable methods $c \geq 2$. For even moderate dimensions the number of evaluations for non trivial $f()$ becomes infeasible. Besides the work required, numerical error is the superior manifestation of the curse. The numerical error increases with the work required.

In numerical integration, *Monte Carlo sampling* is a remedy to *the curse of dimensionality*. The function $f(\cdot)$ is sampled at $C$ fixed points randomly in the interval $[a, b]^k$, and an estimate of the integral is calculated. The success of *Monte Carlo sampling* over the traditional numerical integration techniques can be subscribed to the rigid way the traditional methods select where to evaluate the function. *Monte Carlo sampling*, will on average (i.e. asymptotically) select better points, and will converge faster with fewer evaluations.

*Monte Carlo sampling* is an estimation technique. Estimation techniques are unfortunately prone to *the curse of dimensionality*. In the example above the effects of dimensionality would be felt if the number of samples $C$ stays fixed, while the dimensionality $k$ was allowed to increase. In estimation, it is usually not one simple variable (e.g. the integral) that should be estimated. The case is often that a few parameters should be estimated for each dimension. This makes *the curse of dimensionality* more readily felt. The number of observations (i.e. evaluations in the *Monte Carlo* example) required for a reasonable estimation to the parameters, increases with dimensionality.

### 1.1.1 Data model

In this thesis I will consider a matrix

$$X = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix} \tag{1.1.1}$$

of spectral observations (the "hyperspectral"nature of these is described in detail in chapter 4). It will have $n$ presumed independent samples and $k$ dimensions (i.e. spectral bands). These samples are attached to a vector

$$\vec{y} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ \text{n/a} \\ 3 \\ \vdots \end{pmatrix} \tag{1.1.2}$$

of length $n$ that may or may not have knowledge of what is observed (ground truth).

The task is to fit (estimate) a model (classifier) to the spectral observations $X$ given the class knowledge in $\vec{y}$, and use this model to classify the samples for which no class knowledge is available. This is to be done under *the curse of dimensionality* (i.e. the ratio $n/k$ is low[1]).

### 1.1.2 Sparsity in high dimension

To beat *the curse of dimensionality* I will relay on the fact that $k$-dimensional data is practically always $k^* < k$ dimensional. There is always some underlying structure in the data that can be exploited.

This is by many (e.g. Scott (1992) and Jimenez & Landgrebe (1998)) illustrated by inscribing a (hyper) sphere in a (hyper) cube. Let the sides of the cube be fixed at one unit of length. The volume of the cube is always one, while the length of the main diagonal increases with the dimension. Given a sphere of radii $\frac{1}{2}$ the volume is given as

$$V_k = \frac{\pi^{\frac{k}{2}}(\frac{1}{2})^k}{\Gamma(\frac{k}{2}+1)} \tag{1.1.3}$$

where $\Gamma(\cdot)$ is the gamma function.

In figure 1.1 on the following page the dramatic drop in volume of the (hyper) sphere is seen. Combined by the increased length of the main diagonal of the (hyper) cube, this describes a very empty high dimensional space.

This emptiness should be verified with a sampling experiment. The observed data in matrix $X$ (refer to equation 1.1.1 on the previous page) is presumed to be multivariate normal distributed. For good measure the experiment is carried out with two simulated datasets and one real. Thousand independent samples were drawn from the standard multivariate normal distribution ($N_k(0, I)$), and the minimum and mean distances between the points measured. This was also done for a slightly coloured multivariate normal distribution $N_k(0, U^tU)$, where $U$ is multivariate uniformly distributed. The National Mall dataset (see section 4.5.3) has a dimension of 191. The order of these dimensions and the order of samples were randomised, and the same experiment carried out.

Results are shown in table 1.1 on page 5. Both simulated datasets show emptiness, and the real dataset shows even more emptiness. This suggests that high dimensional data has underlying structure and that this can be exploited by employing a sparser representation.

[1] How low this ratio needs to be, will be discussed in chapter 3

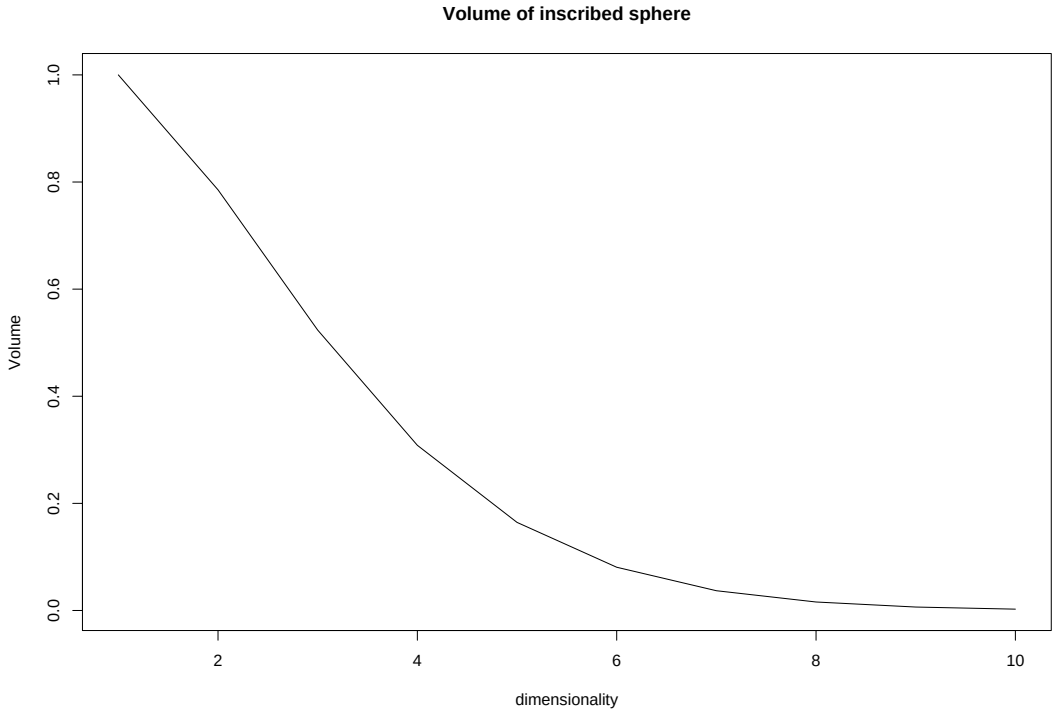


Figure 1.1: Volume of inscribed spheres with radii $\frac{1}{2}$.

### 1.1.3 Dimensionality reducing techniques

There exists many diverse methods that aim to reduce high dimensional spaces into a sparser representation. Among these are *Projection Pursuit* (Huber (1985)), *factor analysis* (Spearman (1904)) and *Principal component analysis (PCA)*. PCA will be treated in section 4.6.1.

Conceptually these techniques do the same. They project the high dimensional space into a lower dimensional space. In algebra this can be seen as the high dimensional space falling on a lower dimensional monoid. This can be seen in figure 1.2 on page 6, where a few points are projected onto a line (spiral monoid). The classical example of points on a globe (sphere), projected onto a map, is also shown.

The ability to project an "empty" high dimensional space into a more "dense" lower dimensional space in a meaningful way is often referred to as one of *the blessings of dimensionality*.

| | $N_k(0, I)$ | | $N_k(0, U^tU)$ | | National Mall | |
|---|---|---|---|---|---|---|
| dim. | min | mean | min | mean | min | mean |
| 10 | 0.89 | 4.37 | 2.32 | 14.48 | 34.48 | 3311.48 |
| 20 | 2.35 | 6.24 | 8.57 | 27.22 | 87.93 | 6402.37 |
| 30 | 3.63 | 7.73 | 17.04 | 43.51 | 104.07 | 9554.75 |
| 40 | 4.93 | 8.83 | 22.55 | 55.46 | 125.19 | 13254.48 |
| 50 | 5.60 | 9.91 | 32.35 | 72.06 | 164.13 | 16858.60 |
| 100 | 9.52 | 14.13 | 84.18 | 140.20 | 238.97 | 24476.87 |
| 150 | 13.06 | 17.27 | 143.41 | 213.10 | 239.53 | 24791.21 |
| 190 | 14.66 | 19.48 | 191.79 | 267.50 | 239.77 | 24806.48 |

Table 1.1: Distance in high dimensional data

## 1.2 Trends in statistics and data analysis

Every science changes over time. Statistics has its origin in odds and bet making from time immemorial. From the renaissance and until 1900, concepts of probability and bet making were more mathematical formalised. This was driven by formalization of mathematics and the need to facilitate global trade through shipping insurance (e.g. Lloyd's of London).

Statistics can be said to have been established as an independent discipline by the introduction of the journal *Biometrika* in 1901. I will now investigate how the changing focus in statistics has influenced the methods we have at our disposal to deal with *the curse of dimensionality*.

The paper Cox (2001) gives the history of *Biometrika*. The introduction of this journal was prompted by a disagreement with editors of a different journal, and comments like *".. more thought would show that a lot of the detailed algebra to be unnecessary"*. Karl Pearson, one of the founders had a preference for mathematical proofs, and this would colour the journal. The topics treated ranged from characteristics of criminals (anthropology/eugenics) to more mundane topics in biology (e.g. bacteria count). Applications were also considered.

In the 1930's the focus of statistics had shifted more to applications. In 1936 Pearson died, and his son Egon Pearson inherited the editorship. Egon Pearson manifested this shift in focus by insisting that articles should be accompanied with numerical illustrations.

Cramer (1945) firmly brought statistics back as a theoretical discipline, and statistic was reestablished as mathematical statistics. Cramer's book put statistics in a measure context and derived properties from this. Statistics would stay largely theoretical until 1962.

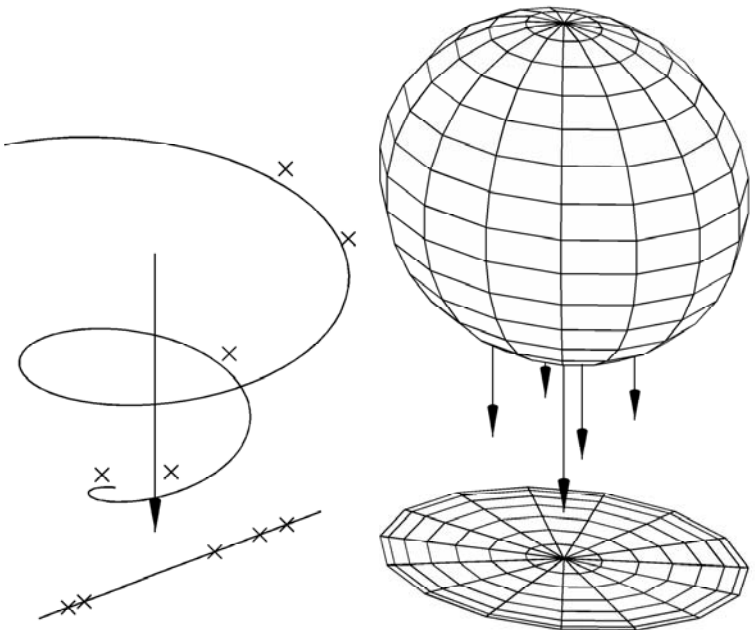

Figure 1.2: Left: A one-dimensional manifold in $\mathbb{R}^3$ (a spiral). Right: A two-dimensional manifold in $\mathbb{R}^2$ (a surface/map).

In Tukey (1962) the focus is shifted back to applications. Tukey's paper is a turning point, and has had an effect on statistics the last 45 years. Recently Mallows (2006) discusses this paper in today's light. The theme in Tukey's paper is that one should focus more on data than on theoretical models, statistics is more described as "data analysis".

In section 3.1 I describ the LDA classifier developed in the 1930's. This classifier was originally used to discern between different species of iris. In the "data analysis" setting one would at first keep the data far away from such a theoretical classifier. One would instead rely on plotting methods and summary data. The box plot in figure 1.3 on the following page, suggests that at least two species of iris can be discerned in this way. Only if the plots suggest certain model assumptions, or if the plots reveal nothing, more theoretical bound methods are brought to light.

In *Biometrika* more applied statistics are also evident. For instance one can observe that the density of delta-epsilon proofs has diminished from 1966 to today.

This change in focus "theory-applications-theory-applications" can by itself be attributed to data. In 1901 not much data was available, and the data available was univariate or simple bi-variate. This allowed for little analysis on the data itself. Instead the data was theorised to come from some distribution, and the inference was done one this distribution instead of on the data. Under certain broad assumptions this was the right ting to do.

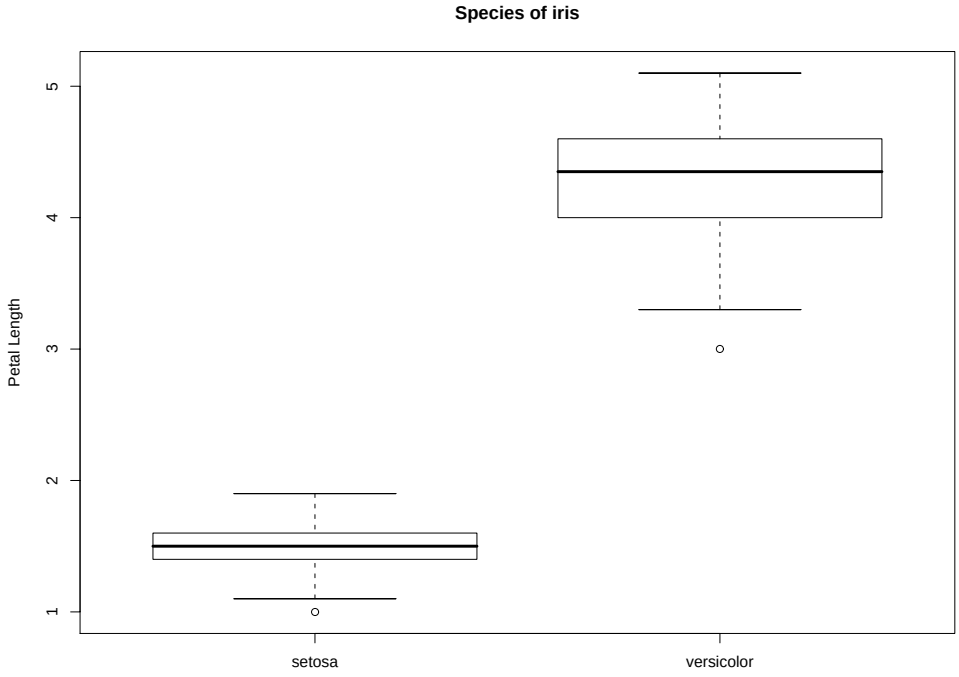


Figure 1.3: Box plot of two species of iris from the Fisher-Anderson dataset. The box represents all data between the quartiles, and the line within the box represents the median. Lines extend from the box to the farthest observations that are within 1.5 times the range between the quartiles. Any data outside this, is plotted and suspected of being outliers.

In the 1930's more data became available from designed experiments, especially in agricultural or health settings. This allowed for more inference to be made as the formerly broad assumptions could be controlled and narrowed. The shift to more theory in the 1940's could be ascribed to developments in theory, and that this theory made a uniform framework for previous developments.

The shift in 1962 was prompted by three factors. More data were available from new instruments. Computers were available to make summaries and calculations on this data,. The most significant factor was that the theory entangled mathematical statistics could no longer contribute to the sciences that needed analysis of the data. See Tukey (1962).

Today the situation is that developments in computer science and electronics make more and more data available. This data are of higher temporal and data domain resolution. An example is the *Large Hadron Collider (LHC)* of the CERN. This instrument collects about 40 million variables a second, see dos Anjos et al. (2006). The increase in available data requires more theoretical methods that can handle

the often complex interrelationship of the data.

## 1.3 Wavelets to the rescue

Chapter 2 introduces wavelets and some of their desirable properties. A brief history of wavelets will be given in section 1.6.

In Breiman (2001b) the differences between the statistics community and the data mining community are described. In the same article Breiman praises how the statistical community on grounds of theory has adopted wavelets in their models.

There is some evidence that nature behaves in a fractal wavelet form. This should be taken with a grain of salt. More concretely Field (1999) describes how the mammal visual system can be seen as doing a wavelet transform on observed light. Similar the limits of the human auditory system described in Gabor (1946) and Gabor (1947) give an incentive for how sound could better be processed, see section 2.4.3 for further explanation.

### 1.3.1 Literature review

Here I will try to give a short overview of literature where wavelets have been used for classification.

The most common way to use wavelets in classification is to use a classifier directly on wavelet decompositions at some cutoff level. By its nature the wavelet transform can be stopped at some level according to how fine details should be represented. Examples are Kaewpijit et al. (2003), Bruce et al. (2002) and Fazel-Rezai & Ramanna (2005). In section 5.1 an example of this is given.

Mallet et al. (1997) give a more innovative approach to classification with wavelets. Through different criteria new wavelets are adaptively designed to give the best representation for classification.

### 1.3.2 My approach

In this thesis I will develop several methods that use wavelets in a classification context. In figure 1.4 on the next page my methods are conceptually described in three steps. First I will transform the available data to a wavelet form. Then I will "reduce" the transformed data to a lower dimensional space. In the final step I will classify the observations by feeding a standard classifier the "reduced" transformed data.

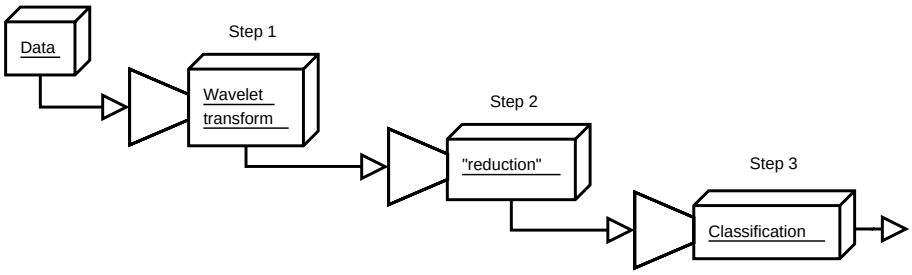


Figure 1.4: Conceptual illustration of how my methods will work.

The second step is where my contribution lies. A property of the sparseness of the wavelet transform, is that the important coefficients of the transformed data have higher magnitude than unimportant coefficients. In the second step I try to exploit this in different ways to "reduce" the number of coefficients that I will later convey to the classifier in the third step.

All this will be done under *the curse of dimensionality*, meaning that very little training data will be available to the classifier.

## 1.4 Reproducible research

> In the good old days physicists repeated each other's experiments, just to be sure. Today they stick to FORTRAN, so that they can share each other's programs, bugs included.
>
> - E.W. Dijkstra (1930-2002).

Reproducible research is important. In computer experiments, sharing of code is important. In the wavelet community this has been the standard as established in Antoniadis (1995). However as the quotation above suggests, one should not trust code blindly, even though it reproduces some results.

I cannot share the data used in my experiments. Nevertheless I will post most of my code on `http://www.purl.org/net/jwick/mastercode/`. The EMD code of section 5.4.1 is excluded because I may not sub-license it.

## 1.5 Guide to the thesis

In chapter 2 I will discuss elements of wavelet theory, especially those that make the wavelet transform suitable to my tasks. In the next chapter (ch. 3) I describe the classifier I will use, and how to assess classifier performance.

In chapter 4 I describe the datasets available and how they are created. Some emphasis is also given to why the wavelet representation is suited to represent the data. A baseline study with data reduced by the standard *principal component analysis (PCA)* is also given.

In chapter 5 atomic decomposition and best basis selection are used to "reduce" the data, and in chapter 6 wavelet denoising is used for the same. Results and discussion will be given in both chapters.

In chapter 7 I give concluding remarks and compare my methods to the PCA based baseline study.

Chapters 3 and 4 will be referred to through the thesis.

A brief history of wavelets will now follow.

## 1.6 A brief history of wavelets

This section will be based on the narration in both Meyer (1993) and Hubbard (1996).

In the late 1970's Jean Morlet of the French oil-company Elf Aquitaine (now TotalFinaElf), discovered wavelets in the field of seismic data analysis.

Morlet worked on seismic traces. Seismic traces are created by sending ('sound') impulses into the ground and recording their echoes.

Originally Morlet used the Fourier transform, and later when more computer power became available, the short time Fourier transform with coarse windows for this analysis. See section 2.4 for a comparison of the methods.

In an attempt to optimise the short time Fourier transform for speed, Morlet stretched and squeezed the window while the analysis was running. In 1981 Alex Grossman joined Morlet and together they formalised and discovered interesting properties of wavelets. This resulted in the paper Grossmann & Morlet (1984), and the Morlet wavelet in Goupillaud et al. (1984). This wavelet is shown in figure 1.5 on the following page.

This wavelet is much related to the short time Fourier transform, and is really a complex sine function "enveloped" or "localised" by a Gaussian function:

$$\phi(x,z) = (\cos 2\pi x + i \sin 2\pi x) e^{\frac{-2x^2\pi^2}{z^2}} - e^{\frac{-z^2}{2} - \frac{2x^2\pi^2}{z^2}} \tag{1.6.1}$$

The $z$ is used to control the tradeoff between time and frequency localisation. These concepts are explained fully in section 2.4.

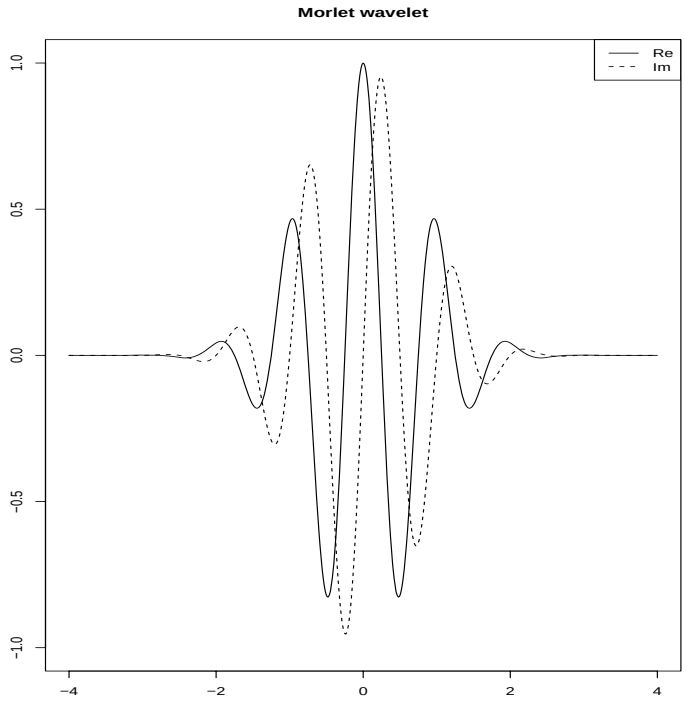


Figure 1.5: Real and imaginary parts of the Morlet wavelet with $z = 5$

This is how the history of wavelets could have started. However, the history of wavelets goes further back.

### 1.6.1 Nothing new under the sun

In 1984, at the École Polytechnique in Paris, the mathematical department shared a photocopier with the physics department. This is where Yves Meyer was introduced to the work of Morlet and Grossmann by a physicist copying the above mentioned articles.

Meyer recognised much of the wavelet theory as to be similar to findings in different mathematical disciplines. This should not be considered as plagiarism, but more as lack of interdisciplinary communication.

### 1.6.2 Ancient preliminaries to wavelets

Before 200 BCE Archimedes introduced his *method of exhaustion*. This method may be seen as the first attempt of describing a function by a series of

trigonometric "atoms". The function under consideration was a circle with a radius of 0.5, and the goal was to determine $\pi$.

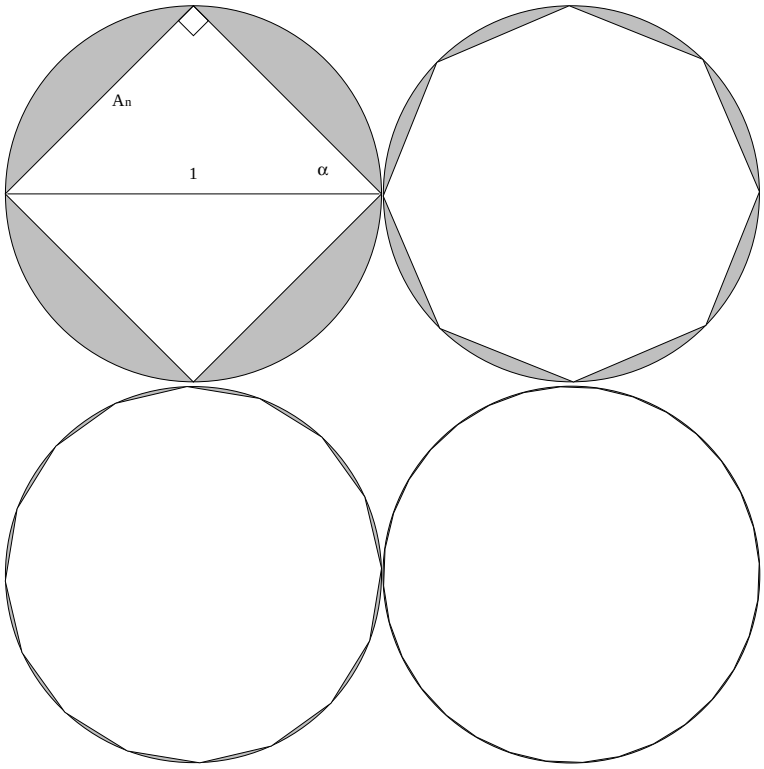


Figure 1.6: Archimedes's approach to finding $\pi$

Archimedes inscribed the circle with k-sided polygons and used trigonometric identities to find $\pi$. Details are in Heath (1921). Today we would reduce these identities to a sine.

In the four-sided polygon in figure 1.6 we know the angles, and $A_n$ can be found as

$$A_n = \sin \alpha \tag{1.6.2}$$

where $\alpha = 45°$. If the number of sides in the polygon is doubled, $\alpha$ would be halved. With a radius of 0.5, $\pi$ could be found as

$$\pi = \lim_{n \to \infty} 2^{n+1} \sin \tfrac{\alpha}{2^n} \tag{1.6.3}$$

History has it that Archimedes worked his way to a 96-sided polygon. In figure 1.6 it can be seen that the error (grey area) decreases very fast. In the bottom right corner the 32-sided polygon gives an error in the third digit of $\pi$.

### 1.6.3 Series

In the fourteenth century, Mādhava in Kerala (present day India) devised "infinite methods" for many trigonometric problems.

The circumference of a circle became:

$$\begin{aligned} c(d) &= \frac{4d}{1} - \frac{4d}{3} + \frac{4d}{5} \cdots \\ &= \sum_{k=1}^{\infty} (-1)^{k+1} \frac{4d}{2k-1} \end{aligned} \tag{1.6.4}$$

where $d$ is the diameter. Archimedes's problem, would become $\pi = c(1)$.

In 1715 Brooke Taylor introduced theorems for what would become Taylor series. Taylor series for the trigonometric functions and properties of geometric objects (i.e. circumference) were now formalised. These approaches are essentially the same as those of Mādhava some 350 years earlier.

Later, in 1747 d'Alembert proposed using superpositions of sine functions to describe the sound (oscillations) the strings of a violin make (Benedetto (1997)). Both Bernoulli and Euler found ways of doing this decomposition for special cases. This was later to become the Fourier transform.

### 1.6.4 Fourier series

The heat equation is a partial differential equation that describes heat varying over time. In 1807 Joseph Fourier generalised the above mentioned decomposition and used it to solve the heat equation. Fourier's contribution was to recognise that all functions could be decomposed to trigonometric series.

Now

$$f(x) = \frac{a_0}{2} + \sum_k \left[ a_k \cos kx + b_k \sin kx \right] \tag{1.6.5}$$

where

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx \, dx \tag{1.6.6}$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin kx \, dx \tag{1.6.7}$$

Abel, Cauchy and Dirichlet were somewhat sceptical to the convergence properties of reconstruction in 1.6.5.

In 1873 Paul David Gustav du Bois-Reymond showed that there could exist continuous functions whose Fourier series would diverge. Besides to unhinge the

very definition of a function, this gave an incentive to develop other orthonormal decompositions not plagued by the shortcomings of the Fourier decomposition.

### 1.6.5 The beginning of wavelets

In Haar (1910), Alfréd Haar uses 24 pages to discuss whether all orthogonal systems exhibit the convergence problems of the Fourier transform. He then devotes only the last nine pages to developing his own orthogonal basis. An example of a decomposition using the Haar basis is given in section 2.1.

The "atoms" (bases) in the Haar decomposition are step functions. In the 1910's Georg Faber considered triangle atoms.

(a) Haar (b) Faber & Schauder

This worked much like Archimedes's *method of exhaustion* as the function under consideration was inscribed by these atoms.

In Schauder (1927), Juliusz Schauder rediscovers Faber's atoms, and made them into a basis. Philip Franklin used the *Gram-Schmidt orthogonalisation procedure* on the Schauder basis, and for the first time moved out of the discrete nature of the Haar system. In Franklin (1928), he also worked himself back to the Haar basis.

Around the same time Littlewood-Paley theory appeared. This can roughly be described as decomposing a function by splitting the support of its Fourier transform into blocks, -working on these.

In the 1960's atomic decomposition was formalised. Atomic decomposition is the decomposition of functions (function spaces) into atoms (i.e. bases). See Calderón (1963). In Calderón (1964), Alberto Calderón proves completeness (reconstructability) under certain assumptions of such decompositions. This is what Grossmann & Morlet (1984) rediscovered. Details are in section 2.5.2.

# CHAPTER 2

# Elements of wavelet theory

What doth distinguish
Gods from us mortals ?
That they before them
See waves without number,
One infinite stream ;
But we, short-sighted,
One wavelet uplifts us,
One wavelet o'erwhelms us
In fathomless night.

– "The Limits of man", Goethe,
translation from Dwight (1839)

Wavelets may not be supernatural, but can take us near the limits imposed by the Heisenberg uncertainty principle, both in time (space) and scale (frequency).

In this chapter I will introduce wavelets and present important elements from wavelet theory. Some comparisons to the Fourier transform are given, and implementation considerations discussed.

Wavelet theory emphasised in this chapter is important as the techniques of dimension reduction of subsequent chapters are based on this theory.

This is not an attempt to give a complete treatise on wavelet theory, but rather an attempt at presenting elements of theory that are explicit or often implicit used.

Some knowledge of analysis at the undergraduate level (Davidson & Donsig (2002)) is assumed. In the discussion of implementation details, use of signal processing

concepts and terminology from Proakis & Manolakis (2007) are employed.

The standard reference in wavelet theory is Daubechies (1992) (most of this is also found in the article Daubechies (1988)). A more illustrated approach to this theory can be found in Mallat (1999).

A collection of most of the fundamental papers in wavelet theory can be found in Heil & Walnut (2006).

I will start with an example of the Haar wavelet transform, and introduce multiresolution analysis and different wavelet transforms from this context. The wavelet transform is then compared to the Fourier transform, and desirable properties are discussed. Discrete implementations are presented.

## 2.1 Basic Haar wavelet transform

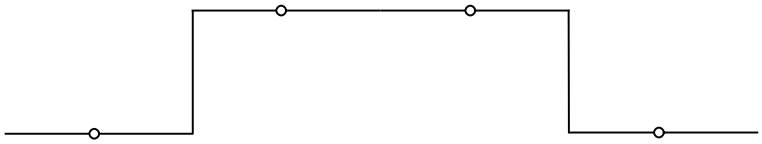

Figure 2.1: Simple function to be decomposed

Assume a simple discreet function as shown in figure 2.1. This function can be represented as a vector and as a sum of weighed coordinate vectors:

$$\begin{pmatrix}1\\2\\2\\1\end{pmatrix} = 1\begin{pmatrix}1\\0\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\1\\0\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\1\\0\end{pmatrix} + 1\begin{pmatrix}0\\0\\0\\1\end{pmatrix} \tag{2.1.1}$$

The idea is that this function can be decomposed into part means, and deviation from these means in groups of two. The mean of both $\left(\begin{smallmatrix}1\\2\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}2\\1\end{smallmatrix}\right)$ is 1.5.

In the first of these groups, the first element deviates from the mean with $-0.5$ while the second element naturally deviates with the opposite sign.

In the second group the elements also deviates by the same amount but in the opposite order.

This leads to the representation:

$$\begin{pmatrix}1\\2\\2\\1\end{pmatrix} = 1.5\begin{pmatrix}1\\1\\0\\0\end{pmatrix} + 1.5\begin{pmatrix}0\\0\\1\\1\end{pmatrix} - 0.5\begin{pmatrix}1\\-1\\0\\0\end{pmatrix} + 0.5\begin{pmatrix}0\\0\\1\\-1\end{pmatrix} \tag{2.1.2}$$

This can be done again:

$$\begin{pmatrix}1\\2\\2\\1\end{pmatrix} = 1.5\begin{pmatrix}1\\1\\1\\1\end{pmatrix} + 0\begin{pmatrix}1\\1\\-1\\-1\end{pmatrix} - 0.5\begin{pmatrix}1\\-1\\0\\0\end{pmatrix} + 0.5\begin{pmatrix}0\\0\\1\\-1\end{pmatrix} \tag{2.1.3}$$

Two observations can be done immediately:

(i) The vectors are orthogonal.

(ii) A sparser representation is possible.

Sparser in this context means that fewer vectors are needed. E.g. the second vector on the right hand side of equation 2.1.3 has a zero coefficient. The example above is not totally faithful to the Haar wavelet transform. This shall be clearer in the following sections.

The Haar-transform representation (as above) was discovered in an attempt in Haar (1910) to workaround certain convergence problems of the Fourier-transform. Haar tried to solve these problems by generating other orthogonal decompositions than the Fourier-transform.

The functions under consideration were originally in the $L^2(\mathbb{R})$ space. This is the space of functions $f$ where

$$\sqrt{\int_{\mathbb{R}} |f|^2 \, \mathrm{d}\mu} < \infty \tag{2.1.4}$$

The inner product in this space is

$$< f, g >= \int_{\mathbb{R}} f(t)\overline{g(t)}\, dt \tag{2.1.5}$$

The bar denotes complex conjugation.

Let

$$\phi(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{else} \end{cases} \tag{2.1.6}$$

be the scaling function (or the Father-wavelet) of the Haar (wavelet) transform. The first right hand side vector of equation 2.1.3 is of this type.

The last three vectors of equation 2.1.3 is scaled and dilated (translated) versions of the Haar mother-wavelet:

$$\psi(x) = \begin{cases} 1 & 0 \le x \le 1/2 \\ -1 & 1/2 \le x \le 1 \\ 0 & \text{else} \end{cases} \tag{2.1.7}$$

The mother-wavelet can also be expressed by the father-wavelet as

$$\psi(x) = \phi(2x) - \phi(2x-1) \tag{2.1.8}$$

Any function $f \in L^2(\mathbb{R})$ can now be represented by scaled and dilated (translated) versions $\psi$ and $\phi$:

$$\psi_i^j(x) = 2^{j/2}\psi(2^j x - i) \tag{2.1.9}$$
$$\phi_i^j(x) = 2^{j/2}\phi(2^j x - i) \tag{2.1.10}$$

This representation is:

$$f(x) = \sum_{i\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} c_{i,j}\psi_i^j(x) \tag{2.1.11}$$

The wavelet coefficients are given as:

$$c_{i,j} =< f, \psi_i^j > \tag{2.1.12}$$

## 2.2 Multiresolution analysis

Notice that the father-wavelet ($\phi$) does not appear in either equation 2.1.11 nor 2.1.12.

This is because the wavelet transform is best treated in a multiresolution framework. In this framework, decompositions like the final decomposition in equation 2.1.3 on the preceding page can be achieved without evaluating too many inner products (equation 2.1.12). In the real (discrete) world this becomes important as the number of such evaluations increases dramatically.

For $j \in \mathbb{Z}$, let $V_j$ be the space spanned by

$$\{\phi(2^j x + 1), \phi(2^j x), \phi(2^j x - 1), \phi(2^j x - 2)\} \tag{2.2.1}$$

This gives $V_j$ as a subspace of $V_{j+1}$:

$$V_0 \subset V_1 \subset \ldots \tag{2.2.2}$$

The orthogonal complement of $V_j$ in $V_{j+1}$, $W_j$ is

$$W_j = \{x \in V_j \,:\, \forall\, y \in V_{j+1} \;\; < x, y >= 0\} \tag{2.2.3}$$

i.e. all the functions in $V_{j+1}$ that are orthogonal to all functions in $V_j$.

Now the vector space used in decompositions like equation 2.1.3 on page 17 can be written as successive orthogonal decompositions

$$\begin{aligned} V_j &= W_{j-1} \oplus V_{j-1} \\ &\dots \\ &= W_{j-1} \oplus W_{j-2} \oplus \dots \oplus V_0 \end{aligned} \tag{2.2.4}$$

with $\psi_i^j(x) \in W_j$ and $\phi_i^j(x) \in V_j$. In equation 2.1.3 on page 17 $j = 2$.

The multiresolution representation of wavelets was given in Mallat (1989), and shall be treated further in section 2.6 on page 28. There it shall be clear that the mother wavelet $\psi$ amounts to a band-pass filter, while the father wavelet $\phi$ amounts to a low-pass filter. The father wavelet lets the reconstruction (equation 2.1.11 on the previous page) cover the whole spectrum without requiring an infinite number of coefficients.

## 2.3 Different wavelet transforms

The wavelet transform used until now is often only called the Haar transform or the discrete Haar transform.

There exists many wavelet transforms, ranging from the benign Haar transform to nearly any combination of

{complete, over-complete, incomplete, -}×
{sample, scale, -} × {discrete, continuous, packet, complex}
× {transform}

From this nomenclature I will only concentrate on the following:

(i) The Continuous Wavelet Transform (CWT)

(ii) The Discrete Wavelet Transform (DWT)

(iii) The Wavelet Packet Decomposition (WPD)

### 2.3.1 The continuous wavelet transform

The continuous wavelet transform is

$$\begin{aligned} c_{\tau,s} &= < f, \psi_s^\tau > \\ &= \int_{-\infty}^{\infty} f(t) \overline{\psi_s^\tau(t)}\, dt \end{aligned} \tag{2.3.1}$$

where the daughter wavelets are scaled and dilated versions of the mother wavelets ($\psi$), given as:

$$\psi_s^\tau(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-\tau}{s}\right) \quad \psi \in L^2(\mathbb{R}) \tag{2.3.2}$$

Besides being in $L^2(\mathbb{R})$ the mother wavelet should have certain properties that shall be clear in section 2.5.

The inverse transform is given as:

$$f(t) = \frac{1}{C_\psi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c_{\tau,s}\psi_s^\tau(t)\,d\tau\frac{ds}{s^2} \tag{2.3.3}$$

The correction divisor $C_\psi$ will be thoroughly explained in section 2.5.2 on page 26. The fraction $1/s^2$ ensures "invariant volume".

I will not detail any mother wavelet, but some are listed in appendix C.

### 2.3.2 The Discrete Wavelet Transform and the Wavelet Packet Decomposition

A realisation of the discrete wavelet transform is shown in the introductory example.

The daughter wavelets are discretised as following. The scale parameter is given as

$$s = s_0^j \tag{2.3.4}$$

where $s_0$ is some constant, typically $s_0 > 0$, and $j \in \mathbb{Z}$ is the running parameter. The dilation[1] parameter is on the form

$$\tau = i\tau_0 s_0^j \quad i \in \mathbb{Z} \tag{2.3.5}$$

where $\tau_0 > 0$ is a suitable constant so that the real-line is sampled at sufficient, or interesting intervals.

In equation 2.1.9 on page 18 the scale is dyadic ($s_0 = 2$) and the dilation is in unity steps. This is the most common, as efficient implementations exist.

The daughter wavelets become

$$\psi_i^j(x) = \frac{1}{s_0^{j/2}}\psi\left(\frac{x - i\tau_0 s_0^j}{s_0^j}\right) \tag{2.3.6}$$

[1]In mathematics dilation changes the size of an object, while the shape is preserved. This is illustrated in equation 2.3.6, where the dilation parameter is responsible for this.

The wavelet packet decomposition is discretised in the same manner as the discrete wavelet transform.

Both the wavelet packet decomposition and the discrete wavelet transform are implemented using filter banks. The only difference is the symmetry of decomposition for the wavelet packet decomposition. In the language of multiresolution analysis, the packet decomposition allows for more scaling functions (father wavelets) and ends up with symmetry between these and the daughter wavelets.

Implementation details are discussed further in section 2.6 on page 28.

## 2.4 The wavelet transform compared with the Fourier transform

Here I will show similarities and differences between the Fourier and wavelet transforms, trying to motivate the time-scale properties of the later.

### 2.4.1 The signal

Assume a signal nine seconds long, sampled at 1000 Hz. The signal is parted in three equal chunks, each consisting of none-overlapping sinusoids, with frequencies 10 Hz, 20 Hz and 25 Hz. To make the signal more interesting a discontinuity lasting 0.05 seconds is introduced in the first sinusoid. The signal is plotted in (a) of figure 2.2 on page 24.

### 2.4.2 The Fourier transform

The Fourier transform is the workhorse of the signal processing community, decomposing signals into their frequency components.

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \omega t}\, dt \tag{2.4.1}$$

If the signal $f(t)$ is sufficient smooth it can be reconstructed from its spectrum $\hat{f}(\omega)$:

$$f(t) = \int_{-\infty}^{\infty} \hat{f}(\omega) e^{2\pi i \omega t}\, d\omega \tag{2.4.2}$$

Efficient discreet implementations exist.

In (b) of figure 2.2 on page 24 the squared magnitude of the coefficients is shown for my signal.

This figure shows the three frequencies well, but fails to say anything about when they occur. The discontinuity is lost.

### 2.4.3 The short-time Fourier transform

Dennis Gabor is most known for his invention of Holography[2], for which he received the 1971 Nobel prize in physics.

However in Gabor (1946), he introduces "windowing" to the Fourier transform, showing which frequencies are present at what time.

A related but less formal approach to visualising frequency components of sound appears in Potter (1945).

The short-time Fourier transform is given as

$$\hat{f}(\omega, t) = \int_{-\infty}^{\infty} f(\tau) W(\tau - t) e^{-2\pi i \omega \tau} \, d\tau \tag{2.4.3}$$

where $W(\cdot)$ is a window that slides over the signal.

Efficient discreet algorithms exist both for decomposition and reconstruction. In (c to e) of figure 2.2 on page 24 the squared magnitude of $\hat{f}(\omega, t)$ is shown for different window sizes.

Clearly that larger windows allow more exact determination of frequency, while smaller windows allow for more exact determination of events along the time axis. The discontinuity is also visible.

Many authors relate this trade-off to the Heisenberg uncertainty principle of quantum mechanics. This uncertainty is also influenced by the shape of the window $W(\cdot)$.

In both Gabor (1946) and Gabor (1947), Gabor relates this phenomenon to the limitations of human hearing. The human ear (or hearing system), can only discern tones both in time and frequency if they have lasted longer than a certain threshold in time.

This also sheds light on an other important difference between the Fourier and short-time Fourier transform. The Fourier transform is built upon trigonometric functions (atoms), defined on the whole line ($\mathbb{R}$) and limited to given frequencies.

[2] "three dimensional" photography

In the short-time Fourier transform the atoms are "windowed" in time, and can be said to decay to zero outside the support of the signal. The question that this arises is whether a signal can be both time limited and restricted to certain frequencies (band limited). An excellent discussion of this can be found in Slepian (1976).

### 2.4.4 The wavelet time-scale analysis

In (f) of figure 2.2 on the next page my signal is decomposed by the Haar wavelet. Better wavelets can be found but this simple wavelet illustrates the power of wavelet analysis.

The frequencies are well localised. Note that the scale-frequency axis is none-linear related. The error is less than perceived. The tone change is equally well localised.

- However, most important is the exact detection of the discontinuity in time, and its correctly portrayed spread in frequency. The importance of this property shall be shown in chapter 4 on page 49.

The aliasing seen in (f) of the figure stems from oversampling. The author desires to show all the pseudo frequencies. If the plot was done by the book the sampling cutoff would have been at about scale 80.

## 2.5 Desirable properties of the wavelet transform

The wavelet transform has several desirable properties. A formal definition of what a wavelet is, does not exist.

Wavelets (or their decompositions) are taken to be functions $\psi$ that satisfy most or all of the following properties:

(i) The transform has both time and scale (frequency) localisation

(ii) The transform has orthonormal decomposition

  (a) Decorrelation

  (b) Efficient implementation

(iii) The transform has completeness of representation

  (a) $\psi$ and $\hat{\psi}$ have compact support

  (b) $\psi$ is smooth

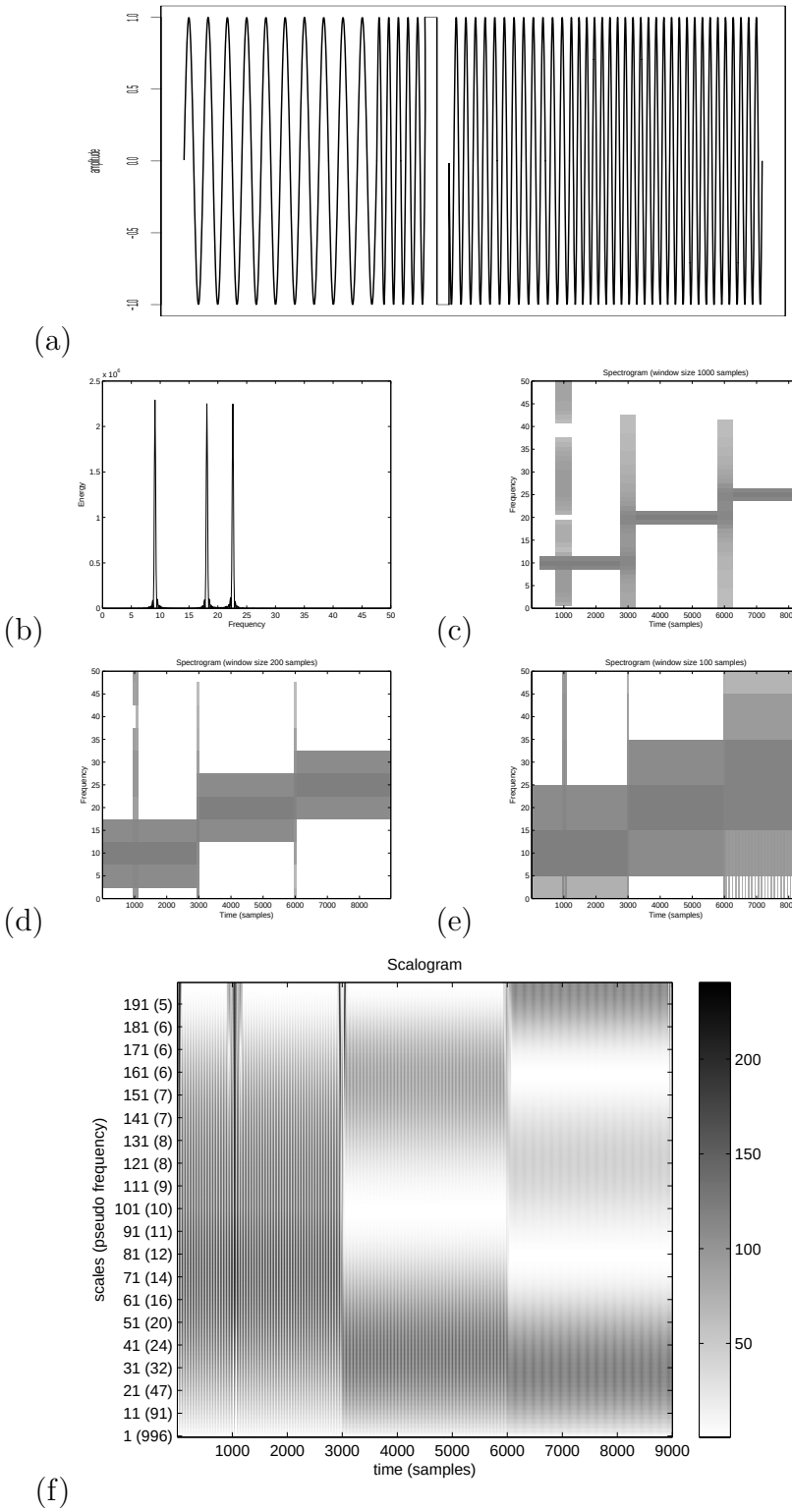


Figure 2.2: (a) Signal, (b) Fourier transform, (c)-(e) Short time Fourier transform, (f) Haar wavelet transform

(c) $\psi$ has zero average

(iv) Self similarity in decomposition

(v) Behaviour at singularities/discontinuities

The first property was illustrated in the previous section. I will show (ii) and (iii), most of the other properties partly follow from them.

Behaviour at singularities will be illustrated on data in chapter 4 on page 49, while the decorrelating properties will be showed in chapter A.1 on page 142 and in applications in following chapters.

The efficient implementation is studied in section 2.6 on page 28.

### 2.5.1 Orthonormality

For a set of functions $\{f(x-k)\}$, $f \in L^2(\mathbb{R})$ orthonormality is assured if

$$\int_{\mathbb{R}} f(x)\overline{f(x-k)}\,dx = \begin{cases} 0 & k \neq 0 \\ 1 & k = 0 \end{cases} \quad k \in \mathbb{Z} \tag{2.5.1}$$

holds.

Will obtain (2.5.1) on a more suitable form. For this I will employ Plancherel's theorem (often referred to as the Parseval relation in the signal processing literature).

First observe the Fourier transform pairs

$$f(t) = \int_{-\infty}^{\infty} \hat{f}(\omega)e^{-2\pi i\omega t}\,d\omega \tag{2.5.2}$$

$$\overline{f(t)} = \int_{-\infty}^{\infty} \overline{\hat{f}(\omega)}e^{2\pi i\omega t}\,d\omega \tag{2.5.3}$$

Now

$$\begin{aligned}
\int_{\mathbb{R}} f(x)\overline{f(x-k)}\,dx &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\hat{f}(\omega)e^{-2\pi i\omega x}\,d\omega\int_{-\infty}^{\infty}\overline{\hat{f}(\omega')}e^{-2\pi i\omega'(x-k)}\,d\omega'\right]dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{f}(\omega)\overline{\hat{f}(\omega')}e^{2\pi i(-\omega x+\omega' x-\omega' k)}\,d\omega\,d\omega'\,dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{f}(\omega)\overline{\hat{f}(\omega')}e^{2\pi(\omega' x-\omega x)-2\pi i\omega' k}\,d\omega\,d\omega'\,dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\hat{f}(\omega)\overline{\hat{f}(\omega')}e^{2\pi x(\omega'-\omega)}e^{-2\pi i\omega' k}\,dx\,d\omega\,d\omega' \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta(\omega'-\omega)\hat{f}(\omega)\overline{\hat{f}(\omega')}e^{-2\pi i\omega' k}\,d\omega\,d\omega' \\
&= \int_{-\infty}^{\infty}\hat{f}(\omega')\overline{\hat{f}(\omega')}e^{-2\pi i\omega' k}\,d\omega' \\
&= \int_{-\infty}^{\infty}|\hat{f}(\omega')|^2e^{-2\pi i\omega' k}\,d\omega' \\
&= \sum_{l=-\infty}^{\infty}\int_{2l\pi}^{2(l+1)\pi}|\hat{f}(\omega')|^2e^{-2\pi i\omega' k}\,d\omega' \\
&= \sum_{l\in\mathbb{Z}}\int_{0}^{2\pi}|\hat{f}(\omega'+2l\pi)|^2e^{-2\pi i\omega' k}\,d\omega' \\
&= \int_{0}^{2\pi}\sum_{l\in\mathbb{Z}}|\hat{f}(\omega'+2l\pi)|^2e^{-2\pi i\omega' k}\,d\omega'
\end{aligned} \tag{2.5.4}$$

Where $\delta(\cdot)$ is the Kronecker delta function.

The daughter wavelets of equation 2.1.9 on page 18:

$$\psi_i^j(x) = 2^{j/2}\psi(2^j x - i) \tag{2.5.5}$$

is on the form of equation 2.5.1 on the previous page, thus orthonormality is assured if

$$\sum_{k\in\mathbb{Z}}|\hat{\psi}(\omega+2k\pi)|^2 = 1 \quad \text{(normality)} \tag{2.5.6}$$

$$\sum_{k\in\mathbb{Z}}\hat{\psi}(2^j(\omega+2k\pi))\overline{\hat{\psi}(\omega+2k\pi)} = 0 \quad \text{(orthogonality)} \tag{2.5.7}$$

holds.

### 2.5.2 Completeness of representation

By completeness of representation it is meant that the wavelet decomposition can be recomposed without loss of information in the original signal/function.

From equation 2.3.1 and 2.3.3 on page 20:

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c_{\tau,s} \psi_s^\tau(t)\, d\tau \frac{ds}{s^2} \tag{2.5.8}$$

In physics *completeness of representation* is called *resolution of identity*, which means that there should be equal energy on both sides of any relation (principle of energy preservation). In mathematics this is related to isometry (distance-preserving isomorphism, $L^2(\mathbb{R}) \to L^2(\mathbb{R}^2)$ is injective).

Independently Calderón (1964) (mathematics) and Grossmann & Morlet (1984) (physics) proved this for what would become wavelets.

Today this is most often proved using some neat results for wavelet-frames (not covered here) see Daubechies (1992).

A proof without frames is sketched by many authors[3]. With help of the same tricks as in equation 2.5.4 on the previous page a complete proof follows. Song & Que (2006) give the Fourier transform of the daughter wavelets as

$$\hat{\psi}_s^\tau(\omega) = \sqrt{s} e^{i\tau\omega} \psi(s\omega) \tag{2.5.9}$$

then

$$\begin{aligned}
C_\psi < f, g > &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{s^2} < f, \psi_s^\tau > < g, \psi_s^\tau > \, ds\, d\tau \\
&= \int_{-\infty}^{\infty} \frac{1}{s^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \overline{\hat{f}(\omega)} \hat{\psi}_s^\tau(\omega) \hat{g}(\omega') \overline{\hat{\psi}_s^\tau(\omega')}\, d\tau\, ds\, d\omega\, d\omega' \\
&= \int_{-\infty}^{\infty} \frac{1}{s^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(\omega - \omega') |s| \hat{\psi}(s\omega) \overline{\hat{\psi}(s\omega') \hat{f}(\omega)} \hat{g}(\omega')\, d\tau\, ds\, d\omega\, d\omega' \\
&= \int_{-\infty}^{\infty} \frac{|s|}{s^2} \int_{-\infty}^{\infty} \hat{\psi}(s\omega) \overline{\hat{\psi}(s\omega) \hat{f}(\omega)} \hat{g}(\omega)\, d\omega\, ds \\
&\quad \text{let: } \Omega = s\omega \\
&= \int_{-\infty}^{\infty} \frac{|\frac{\Omega}{\omega}|}{(\frac{\Omega}{\omega})^2} \int_{-\infty}^{\infty} \hat{\psi}(\Omega) \overline{\hat{\psi}(\Omega) \hat{f}(\omega)} \hat{g}(\omega)\, d\omega\, d\frac{\Omega}{\omega} \\
&= \underbrace{\int_{-\infty}^{\infty} \frac{1}{|\Omega|} \hat{\psi}(\Omega) \overline{\hat{\psi}(\Omega)}\, d\Omega}_{C_\psi} \underbrace{\int_{-\infty}^{\infty} \overline{\hat{f}(\omega)} \hat{g}(\omega)\, d\omega}_{<f,g> \text{ (Plancherel)}}
\end{aligned} \tag{2.5.10}$$

One arrives at the reconstruction formula 2.5.8 as $\psi,\ f\ g \in L^2(\mathbb{R})$. From equation 2.5.10 several constraints on $\psi$ and $\hat{\psi}$ are incurred. Primarily the

[3] Daubechies (1992) page 24, Mallat (1999) page 81

admissibility constraint:

$$C_\psi = \int_{-\infty}^{\infty} \frac{1}{|\Omega|} \hat{\psi}(\Omega) \overline{\hat{\psi}(\Omega)}\, d\Omega \quad < \infty \tag{2.5.11}$$

which essentially says that the mother wavelet should be bandlimited. The admissibility constraint is invalidated if $\hat{\psi}(0) \neq 0$ (DC gain), which leads to the zero average constraint:

$$\hat{\psi}(0) = \int_{-\infty}^{\infty} \psi(t)\, dt = 0 \tag{2.5.12}$$

For the interchanging of variables in equation 2.5.10 on the preceding page one should also require $\psi$ to be smooth.

## 2.6 Implementation

An overview of the implementation of the discreet wavelet transform and the wavelet packet decomposition is given. The essential reference for implementing wavelet transforms (decompositions) is Strang & Nguyen (1996). It should be noted that all modern implementations have their origin in the pyramid scheme in Mallat (1989).

### 2.6.1 Discrete wavelet transform

As explained in section 2.3.2 on page 20 the daughter wavelets are sampled both in scale and dilation. Discretised:

$$\psi_i^j[x] = \frac{1}{s_0^{j/2}} \psi\left( \frac{x - i\tau_0 s_0^j}{s_0^j} \right) \tag{2.6.1}$$

Using the multiresolution framework (see chapter 5 of Daubechies (1992) for full details) the father wavelet can be written as

$$\begin{aligned} \phi &= \sum_t < \phi, \phi_{1,t} > \phi_{1,t} \\ &= \sum_t h_t \phi_{1,t} \end{aligned} \tag{2.6.2}$$

where $\phi_{1,t}$ is an orthonormal basis in $V_1$, and $\phi \in V_0 \subset V_1$. Essentially this is a convolution and $\{h_t\}$ is a filter.

Similarly the mother wavelet is given as

$$\psi = \sum_t g_t \phi_{1,t} \tag{2.6.3}$$

where $\{g_t\}$ is a related filter

$$g_t = (-1)^t \overline{h_{1-t}} \tag{2.6.4}$$

The wavelet coefficients are in the signal processing terminology parted in approximation and detail coefficients. If $\{h[t]\}$ and $\{g[t]\}$ are the discretised versions of the above filters, and $f[k]$ is the sampled signal, the coefficients can be found as:

$$c_\phi[t] = \sum_k f[k]h[2t-k] \tag{2.6.5a}$$

$$c_\psi[t] = \sum_k f[k]g[2t-k] \tag{2.6.5b}$$

This is convolution followed by a dichotomous decimation. The decimation is possible without loss of reconstructability due to the Nyquist-Shannon sampling theorem.

The sums in (2.6.5) are really infinite, but by imposing compact support on the mother wavelet, one gets away with a much shorter convolution. $g[\cdot]$ and $h[\cdot]$ are the impulse response of the low- and high-pass filters respectively.

Coefficients at different desired levels of decomposition can be found by cascading these filter banks. See figure 2.3 on the following page for an example at level three.

In the signal processing literature this approach is known as a *two-channel sub band coder with quadrature mirror filters*.

This suggests that there are additional conditions on the filters. In equation 2.6.2 on the previous page one has to choose the right orthonormal bases $\phi_{1,t} \in V_1$ that will allow perfect reconstruction later.

Arguments involved are among others that $\{g_t\}$ should compensate for aliasing in $\{h_t\}$, and vice versa. In chapter 10.3 of Strang & Nguyen (1996) this is reduced to quadratically constrained optimisation.

The discrete wavelet transform has complexity of $O(n)$. This compares favourably with the $O(n \log n)$ of the Cooley-Tukey fast Fourier transform, which is the closest competitor.

For several wavelets a speed increase in the $50 - 80\%$ range is possible if the *lifting scheme* implementation is considered. See Daubechies & Sweldens (1998) for details.

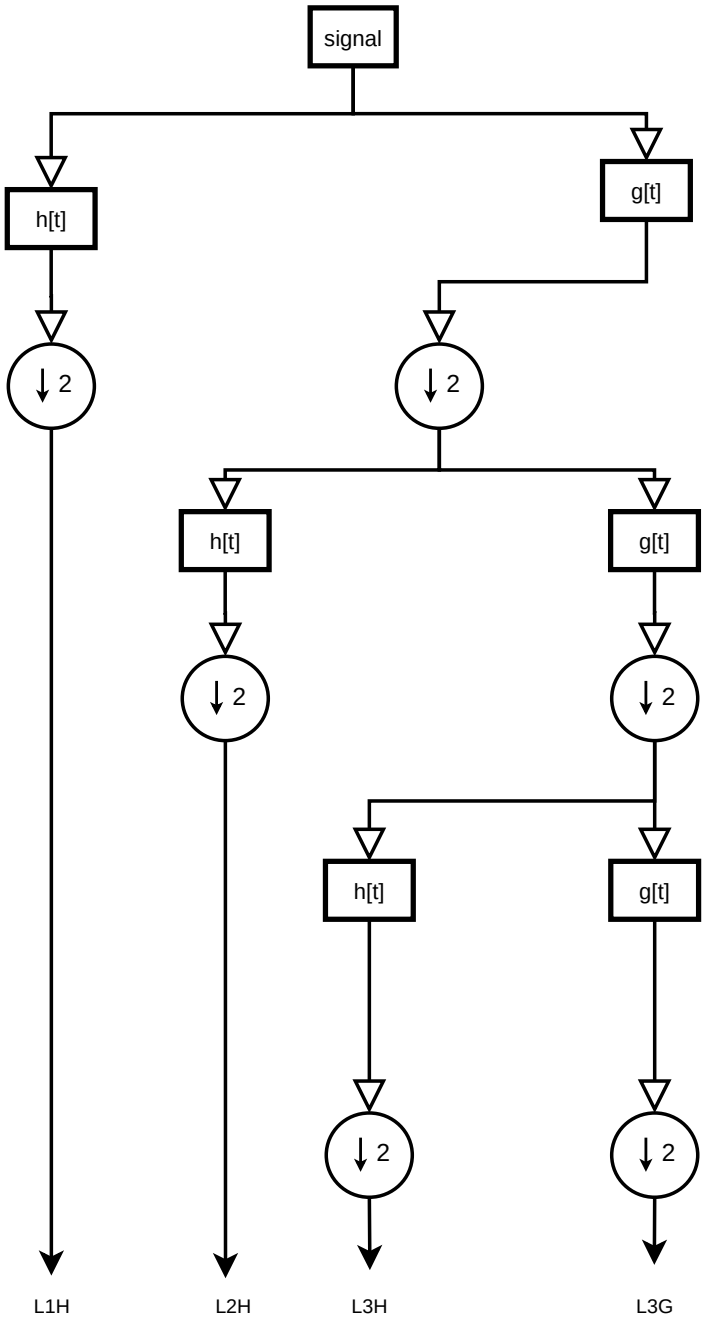


Figure 2.3: Cascading filter bank implementation of the discrete wavelet transform at level three.

**Reconstruction**

Reconstruction is done by up-sampling (inserting zeros) the coefficients and convolving by the respective inverse quadrature mirror filters, before adding.

This is illustrated in figure 2.4.

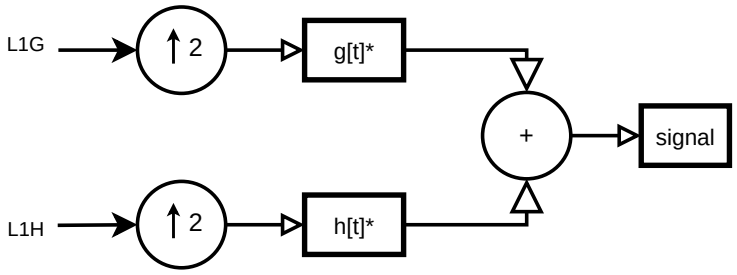


Figure 2.4: Illustration of discrete wavelet transform reconstruction, using inverse quadrature mirror filters

### 2.6.2 Wavelet packet decomposition

For $L$ levels of decomposition the discrete wavelet transform produces $(L+1)$ sets of coefficients. The wavelet packet decomposition produces $L^2$ sets of coefficients by decomposing the signal with more filters than the discrete wavelet decomposition.

See figure 2.5 on the next page.

This decomposition leads to a more symmetric decomposition tree, where the signal can be recovered in $2^{2^{L-1}}$ ways by combining coefficients at different levels. e.g. $\{L1H, L1GL2HL3H, L1GL2HL3G, L1GL2G\} = \{L1H, L1G\}$.

This will be relied upon in subsequent chapters.

## 2.7 Closing summary

The properties of wavelets outlined in this chapter, are the core properties of wavelets. These properties are by no means absolute. In chapter 3 the decorrelating properties will be investigated, and in chapter 4 on page 49 behaviour at discontinuities will be examined.

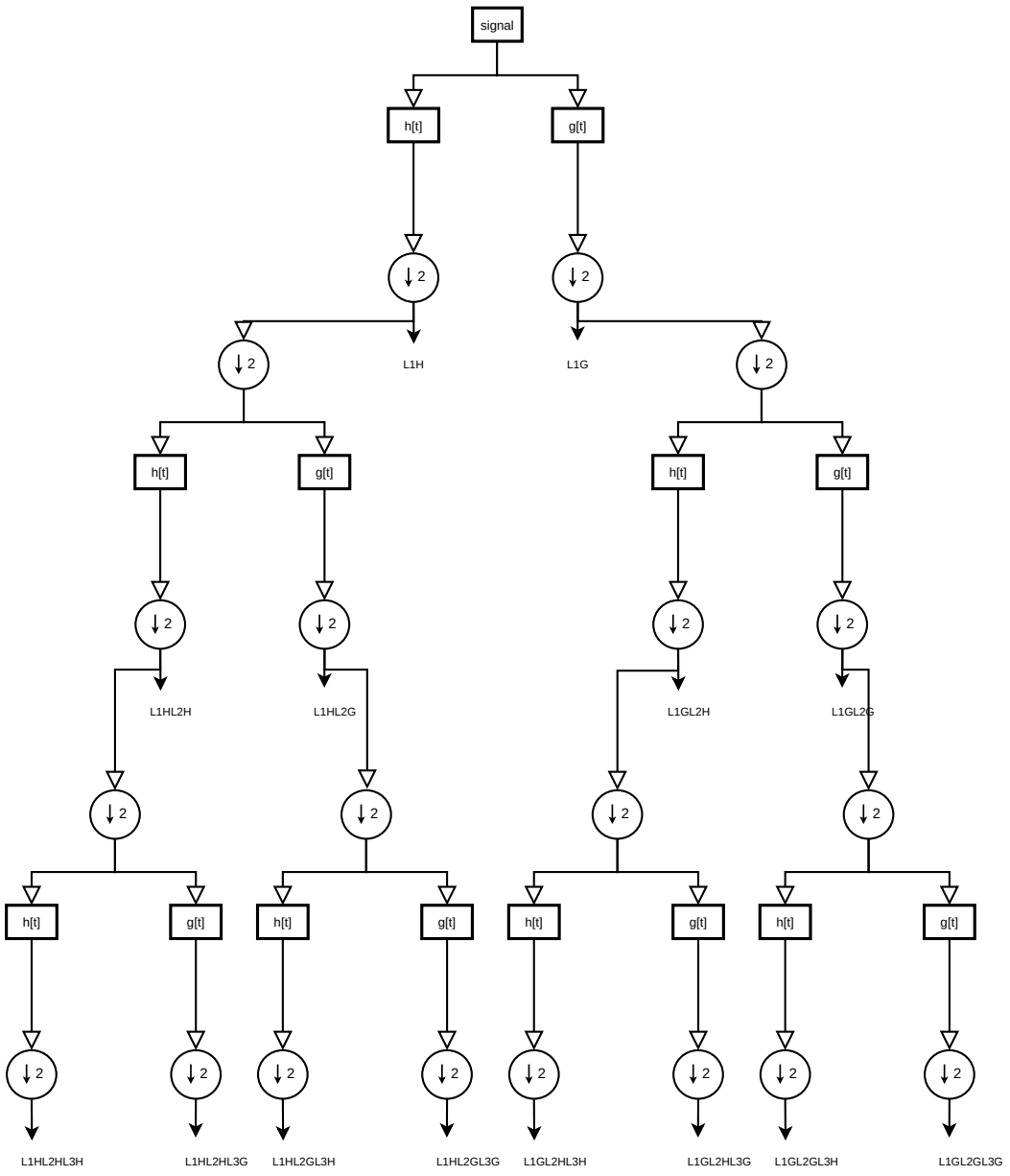


Figure 2.5: Illustration of wavelet packet decomposition

# CHAPTER 3

# Assessment

> In attempting to resolve this difficulty the forcaster may often find himself in the position of chosing to ignore the verification system or let it do the forcasting for him by "hedging" or "playing the system." This may lead the forcaster to forcast something other than what he thinks will occur, for it is often easier to analyze the effect of different possible forcasts on the verification score than it is to analyze the weather situation. It is generally agreed that this state of affairs is unsatisfactory, as one essential criterion for satisfactory verification is that the verification scheme should not influence the forcaster in no undesirable way.
>
> – Glenn W. Brier, U.S. Weather Bureau (Brier (1950))

This chapter is about assessing the quality of classifiers and their underlying models. The linear discriminant analysis classifier of Fisher (1936b) will be detailed, but generalisation to any classifier will be emphasised.

Model assessment is perhaps the single most important undertaking in any scientific endeavour. However it is often carried out without much thought and is where most undertakings tend to fail.

In Farman et al. (1985) the Antarctic ozone-hole is announced by ground measurements. This surprised the scientific community at large, as modern satellite based sensors had failed to notice the trend.

In a scientific myth the explanation is that the sensors removed any none normal trend. Although this is largely a myth (see Christie (2004)), it serves as a reminder that model assessment should not be taken lightly.

More seriously are the accusations against the father of modern genetics Mendel. He stands accused of either faking his data, or repeating his experiments several times, and in the process discarding results not consistent with his theory. See for instance Fisher (1936a).

One main inclination of Breiman (2001b) against the established statistical community, is that most papers assume certain models without verifying them.

First the classifier used will be introduced. In section 3.2 non-parametric methods for classifier assessment is given. Section 3.3 on page 45 shows how to compare classifiers while section 3.4 on page 46 contains some concluding remarks.

In appendix A a few classifier specific assessment methods, and methods to asses the classifier's assumptions (normality) are given.

Knowledge of multivariate statistics at the level Mardia et al. (1979) is assumed.

## 3.1 The LDA classifier model

In this thesis the linear discriminant analysis (LDA) classifier of Fisher (1936b) with modifications due to Rao (1948), will be the workhorse classifier employed. The reason for choosing this classifier, is although being simple, derived results for classification error at a given complexity, show high correlation with more modern classifiers at the same level of complexity. See figure 3.1 on the next page for an example.

I will first start with the multivariate normal distribution ($k$-dimensions):

$$f_k(X) = \frac{1}{|2\pi|^{k/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(X-\vec{\mu}_k)^t \Sigma_k^{-1}(X-\vec{\mu}_k)} \tag{3.1.1}$$

with log-likelihood:

$$l(X;\vec{\mu}_k,\Sigma_k) = -\frac{k}{2}\log|2\pi\Sigma_k| - \frac{1}{2}(X-\vec{\mu}_k)^t\Sigma_k^{-1}(X-\vec{\mu}_k) \tag{3.1.2}$$

The maximum likelihood (ML) estimate will only depend on:

$$\frac{k}{2}\log|\Sigma_k| + \frac{1}{2}(X-\vec{\mu}_k)^t\Sigma_k^{-1}(X-\vec{\mu}_k) \tag{3.1.3}$$

If I assume equal loss

$$L(k,\hat{k}) = \begin{cases} 0 & k = \hat{k} \\ 1 & k \neq \hat{k} \end{cases} \tag{3.1.4}$$

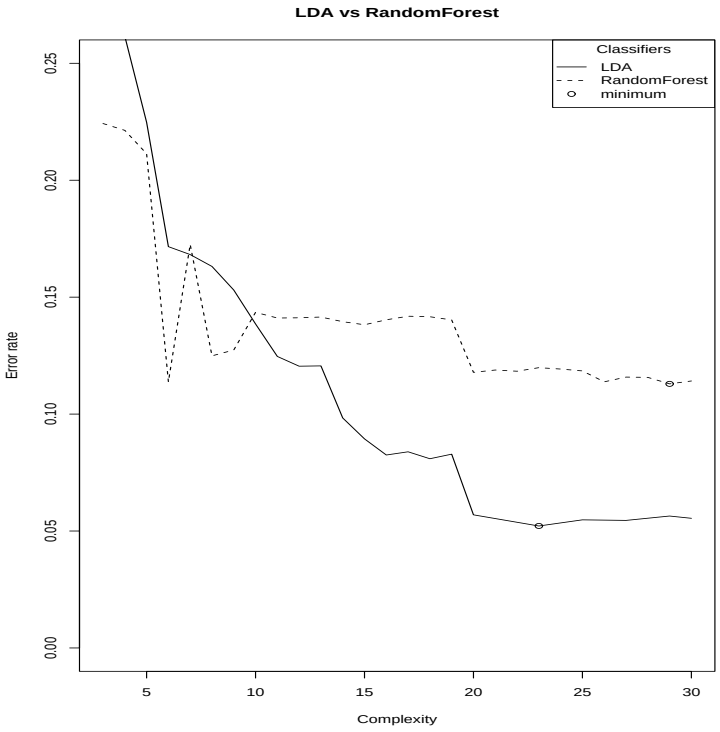


Figure 3.1: The error rate of the LDA classifier shows high correlation with the error rate of the much more complex Random Forests classifier. For details about the Random Forests classifier see Breiman (2001a).

each class can be chosen as

$$f_{\hat{k}} = \max_k f_k(X)\pi_k \tag{3.1.5}$$

where $\pi_k$ is Bayesian prior for the $k$-th class.

By equation 3.1.3 on the preceding page

$$\Lambda_k = \log|\Sigma_k| + \underbrace{(X-\vec{\mu}_k)^t\Sigma_k^{-1}(X-\vec{\mu}_k)}_{\text{squared Mahalanobis distance}} - 2\log\pi_k \tag{3.1.6}$$

Finding $\hat{k}$ by (3.1.5) is equivalent with

$$\Lambda_{\hat{k}} = \min_k \Lambda_k \tag{3.1.7}$$

Equation 3.1.6 is quadratic. LDA is a simplification of equation 3.1.6 by which a class dependant $\Sigma_k$ is replaced by the total population covariance $\Sigma$.

The function that minimises 3.1.6 have now become a squared Mahalanobis

distance and a prior term

$$\begin{aligned}\Lambda_k &= (X-\vec{\mu}_k)^t\Sigma^{-1}(X-\vec{\mu}_k) - 2\log\pi_k \\ &= -2X^t\Sigma^{-1}\vec{\mu}_k + \vec{\mu}_k^t\Sigma^{-1}\vec{\mu}_k + X^t\Sigma^{-1}X - 2\log\pi_k\end{aligned} \tag{3.1.8}$$

which is linear.

In practice the standard estimates are used. For the total population covariance matrix, the standard pooled estimate is used.

## 3.2 Non-parametric approaches

In this section I will treat the classifier as a black-box. This classifier is fed an ordered training set with attached class knowledge. *Set* is here used in the loose sense. The classifier is now trained. Subsequent classification of a test set is done without any class knowledge attached. It returns an opinion of which classes the test set samples belong to. See figure 3.2.

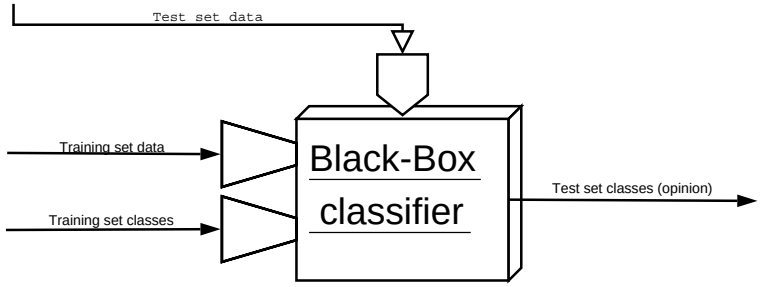


Figure 3.2: Black-box classifier

By treating all classifiers in this way, makes it easier to develop a common framework for testing different classifiers.

I will not use any of the methods in this section alone, but will use them to motivate a hybrid method given in section 3.4. The figures through this section can be compared, as they are made with the same data and settings.

Parametric methods, as mentioned in the introduction, is given in appendix A.

### 3.2.1 The apparent error rate

The apparent error rate or the training error, is the error incurred if the training set is both used in training and testing.

$$\bar{\text{err}} = \frac{\#\text{ misclassified training samples}}{\#\text{ training samples}} \tag{3.2.1}$$

More generally

$$\bar{\text{err}} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\vec{x}_i)) \tag{3.2.2}$$

where $y_i$ is the known classes, $\hat{f}(\cdot)$ the trained classifier and $\vec{x}_i$ the vector to be classified. $L(\cdot,\cdot)$ may be any loss function, but the $0-1$ loss function will be assumed.

This error rate tends to underestimate the real error rate, and is prone to overfitted classifiers.

On the positive side, it is data scarce by re-using the training set. See figure 3.3 on the next page for an example.

### 3.2.2 The holdout error rate

The holdout method tries to address the shortcomings of the apparent error rate.

The test set is chosen independently of the training set. This can be done in two ways. The sets might be chosen by random sampling without replacement, from the collected data. Alternatively the sets might be chosen by expert knowledge. This expert knowledge should lead to representative sets with typical or paired samples for the problem at hand.

It should be stressed that the benefits of the holdout method are fully assured only if proper attention is given to the set selection as outlined above. If possible, any doubt can be removed from the assessment protocol by making the process double blind.

The holdout error rate is:

$$\hat{\text{err}}_h = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\vec{x}_i)) \tag{3.2.3}$$

The only difference from the apparent error rate being the use of an independent test and training sets.

In figure 3.3 on the following page comparisons between the two error rates are shown at different levels of complexity.

If the assumption of independent and representative sets holds, the holdout error rate will be unbiased.

The problem with the holdout method, is that it is data intensive. To choose the right test set size will be a trade-off.

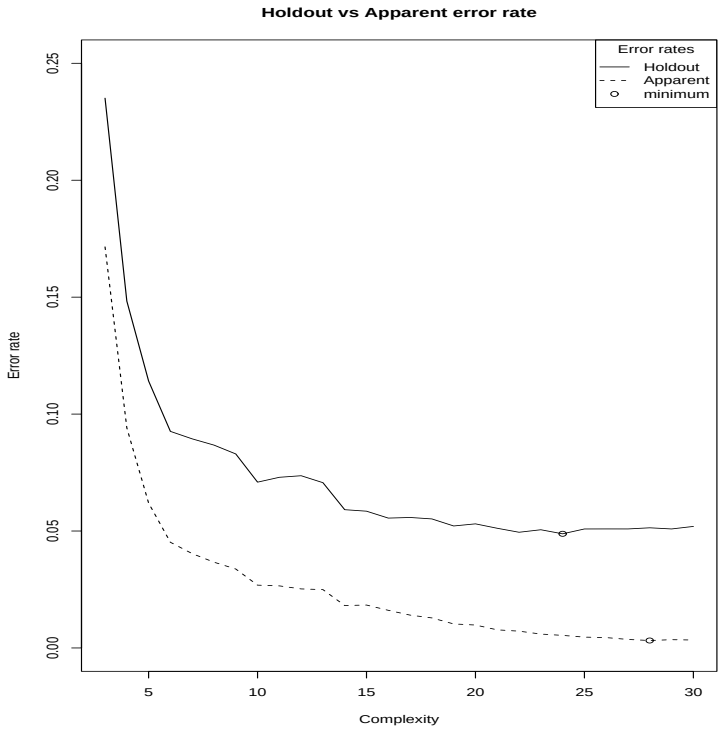


Figure 3.3: Holdout vs Apparent error rate

### Choosing a test set size

Let an estimate of the real error rate be:

$$\hat{\text{err}} = \frac{W^*}{N} \tag{3.2.4}$$

where $W^*$ is the number of erroneously classified observations, and N is the number of observations in the test set.

Clearly $W^* \sim \text{bin}(N, \text{err}\,|r_k)$, where err is the true error rate for the given classification rule $r_k$. 'bin' is the binomial distribution.

Expectation and variance:

$$\text{E}(W^*|r_k) = N \text{ err} \qquad \text{Var}(W^*|r_k) = N \text{ err}(1 - \text{err})$$

Thus:

$$\begin{aligned} \text{Var}(\hat{\text{err}}|r_k) = \text{Var}\left(\frac{W^*}{N}\middle| r_k\right) &= \frac{\text{Var}(W^*|r_k)}{N^2} \\ &= \frac{\text{err}(1-\text{err})}{N} \end{aligned} \tag{3.2.5}$$

Assuming a desired accuracy (width) $s$, a 95% confidence interval can be constructed for $\hat{\text{err}}|r_k$ using the normal approximation for the binomial. Fixing the accuracy (width) of the confidence interval $s$, yields a relation to the test set size:

$$\begin{aligned} 1.96\sqrt{\frac{\text{err}(1-\text{err})}{N}} &= s^2 \\ \frac{\text{err}(1-\text{err})}{N} &= \left(\frac{s}{1.96}\right)^2 \\ \left(\frac{s}{1.96}\right)^2 \text{err}(1-\text{err}) &= N = N(\text{err}, s) \end{aligned} \tag{3.2.6}$$

A reverse argument for this can be found in section 2.7 of Ripley (1996). The normal approximation holds roughly for $N$ err $\geq 10$.

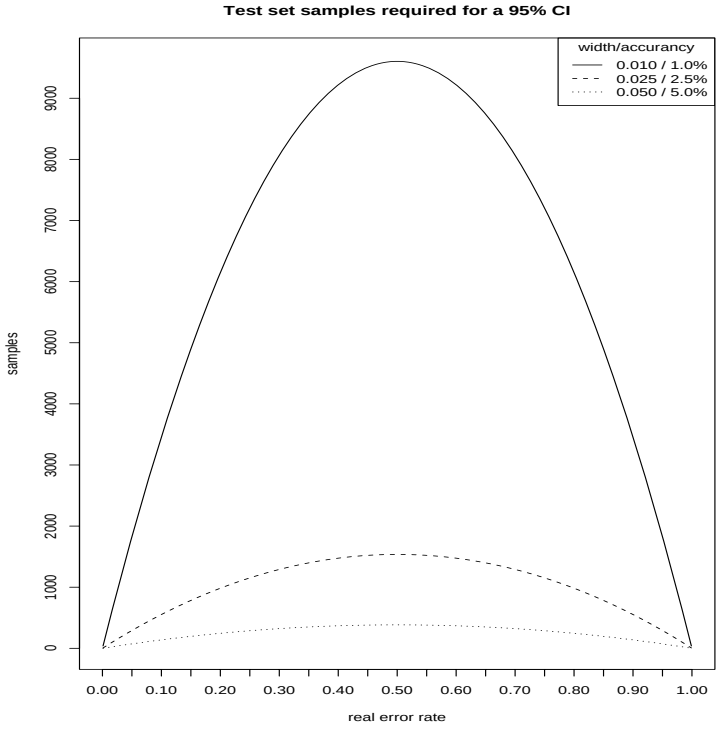


Figure 3.4: Test set samples required for a 95% confidence interval of a given width/accuracy. See equation 3.2.6

In figure 3.4 a plot reveals that many samples are required if all ranges of the real error rate are to be considered as equally likely.

In the setting of comparing different classifiers $r_k$, a width of 0.01 or smaller for the confidence interval of $\hat{\text{err}}$ is desired. This ensures reasonable overlapping and works much as *power* in the sense Neyman-Pearson hypothesis testing.

### 3.2.3 Cross-validation

Some benefits of the holdout method can be retained without sacrificing too much data.

If the training set is the only set available, one can always part this set, creating two or even $k$ independent sets. In the same manner as the holdout method, one of the sets can be withheld while the classifier is trained, only to be brought out for the test.

Without affecting the notion of independence, as established for the holdout method, the sets can be swapped and the process repeated. With a $k$- parted set, this can be done $k$ times. An average of these $k$ error rates is known as the $k$-fold cross-validation error rate. This must not be confused with cross-validation as used in smoothing, which will briefly be mentioned in section 6.2.7.

Formally

$$\hat{\text{err}}_{CV_k} \frac{1}{k} \sum_{i=1}^{k} \frac{1}{|k_i|} \sum_{j \in k_i} L[y_j, \hat{f}^{\setminus k_i}(\vec{x}_j)] \tag{3.2.7}$$

where $L(\cdot, \cdot)$ is a loss function, $k_i$ is the i'th set when the original set is parted in $k$-parts. $\setminus k_i$ is the original set without the i'th part. $|\cdot|$ is the cardinality operator. $\hat{f}(\cdot)$ being the trained classifier.

Cross-validation has some benefits of the holdout method. However, the classifier performance is generally affected by its training set size. Stone (1977) discusses this further.

The idea of cross-validation has appeared several times, but did not have any breakthrough until Stone (1974). The discussion following in Stone (1974) is worth some consideration.

Today $k \in \{5, 10\}$ is considered a reasonable bias trade-off. An extreme variant also in use is the leave-one-out cross-validation (LOOCV) where $k = n$, see figure 3.5 on the following page for examples.

The cross-validation above is known as controlled cross-validation, there exists an uncontrolled cross-validation where the $k$-folds are chosen randomly and folds may even be empty.

Uncontrolled cross-validation should be used where class clustering around certain indexes or repeated measurements are suspected.

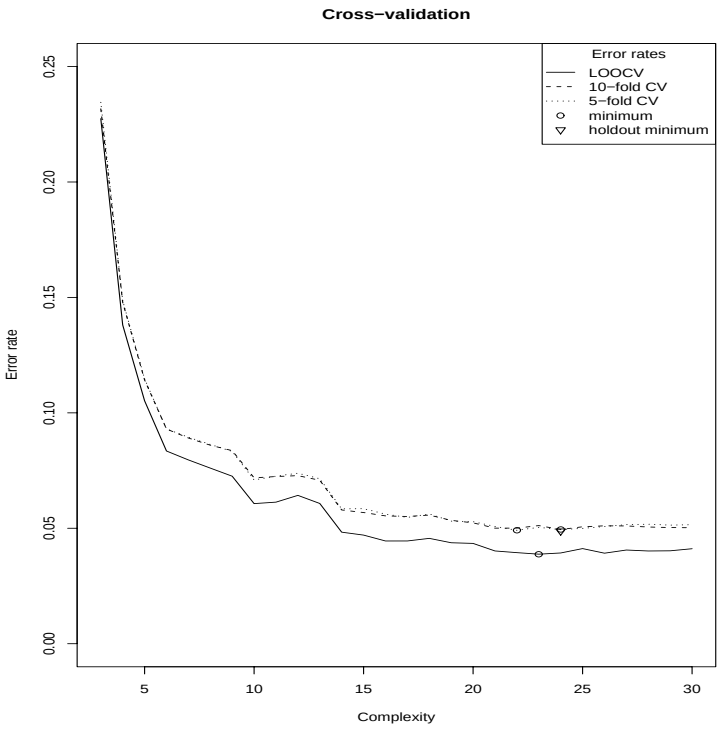


Figure 3.5: Examples of 5-fold, 10-fold and leave one out cross-validation

### 3.2.4 Optimism correction

The apparent error rate $\hat{\text{err}}$ underestimates the real error rate. I will now accept this bias. Under $0-1$ loss this optimism can be taken to be

$$\text{op} = \frac{2}{n}\sum_{i=1}^{n}\text{Cov}[\hat{f}(\vec{x}_i), y_i] \tag{3.2.8}$$

One can introduce a corrected (apparent) error rate:

$$\hat{\text{err}}^c = \hat{\text{err}} + \hat{\text{op}} \tag{3.2.9}$$

This correction is based on the notion that $\bar{\text{err}}$ only estimates the error in the data space where there exist training samples (in-sample error), while the real error can appear anywhere in the data space. The real error rate is thus compounded of in-sample and extra-sample error.

Mallows' $C_p$ (Mallows (1973)), the Akaike information criterion (Akaike (1974)) and even the Bayesian information criterion (Schwarz (1978)) can be seen in this context. All of these measures, are on the form:

$$\bar{err}^{c*} = \bar{\text{err}} + \text{penalty/correction for model complexity} \tag{3.2.10}$$

I will not use any of this directly, but only use it to clarify the next section. See section 7.5 of Hastie et al. (2001) for further discussion.

### 3.2.5 The leave-one-out bootstrap (LOOB)

The leave-one-out bootstrap (Efron (1983)) is an approach based on the same observation of in-sample and extra-sample error as in the previous section. The bootstrap (Efron (1979)) is used to try to mitigate the optimism.

$B$ non-parametric bootstraps are done where a suitable number of samples are drawn and used to train classifiers ($\hat{f}^{b_j}$). Efron (1983) reports $B = 200$ as adequate.

The leave-one-out bootstrap error rate is calculated by averaging the error committed in classifying all of the training set in each bootstrap, where they are not used to train the classifier.

$$\hat{\text{err}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{B} I_{i\notin b_j} L[y_i, \hat{f}^{b_j}(\vec{x}_i)]}{\sum_{l=1}^{B} I_{i\notin b_l}} \tag{3.2.11}$$

Here $I$ is the indicator function and $L(\cdot,\cdot)$ is a loss function.

Equation 3.2.11 can also be seen as a smoothing of the cross-validation error rate.

The problem with this error rate, is that it is depending on a "suitable number" of samples drawn in each bootstrap, and the training set size, a bias is introduced.

The overfitting problems of the apparent error rate, however have larger impact on the total performance than this bias.

### 3.2.6 The .632 and .632$^+$ error estimators

Efron (1983) shows some methods meant to improve on the cross-validation error rates. Particularly one of these, the .632 estimator stands out.

Its motivation is somewhat weak, but based on the same observations of in-sample and extra-sample error as the two previous methods. An estimate of the real error rate should allow for both these error rates.

It improves on the leave-one-out bootstrap error rate by weighing this error rate against the apparent error rate. The weights depend on the probability

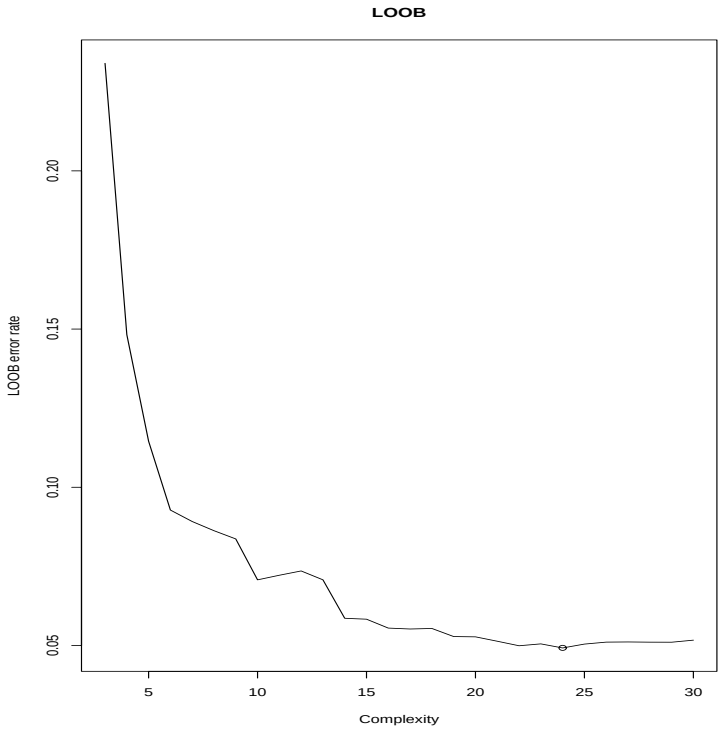


Figure 3.6: LOOB

(proportion) of the samples being in the bootstrap sample.

$$\Pr\{\vec{x}_i - \vec{x}_i^{b_j} = \vec{0}\} = 1 - (1 - \frac{1}{n})^n = 1 - e^{-1} \approx 0.632 \tag{3.2.12}$$

$$\Pr\{\vec{x}_i - \vec{x}_i^{b_j} \neq \vec{0}\} = e^{-1} \approx 0.368 \tag{3.2.13}$$

Thus

$$\hat{\text{err}}^{(0.632)} = 0.368\bar{\text{err}} + 0.632\hat{\text{err}}^{(1)} \tag{3.2.14}$$

This dampens both the overfitting bias of $\bar{\text{err}}$, and the training set size bias of $\hat{\text{err}}^{(1)}$.

$\hat{\text{err}}^{(0.632)}$ is however unreliable if the classifier is totally overfitted, yielding $\bar{\text{err}} = 0$.

Efron & Tibshirani (1997) introduced the $.632^+$ estimator to fix this. According to how overfitted it perceives the classifier to be, this estimator weigh the error terms differently.

First let the no-information error rate be

$$\hat{\gamma} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L[y_i, \hat{f}(\vec{x}_j)] \tag{3.2.15}$$

This error is the error incurred if $y$ and $\vec{x}.$ are (stochastically) independent.

From this a measure of relative overfitting is:

$$\hat{R} = \frac{\hat{\text{err}}^{(1)} - \bar{\text{err}}}{\hat{\gamma} - \bar{\text{err}}} \tag{3.2.16}$$

The new re-weighed estimator is

$$\hat{\text{err}}^{(0.632^+)} = \left(1 - \frac{0.632}{1 - 0.368\hat{R}}\right)\bar{\text{err}} + \frac{0.632}{1 - 0.368\hat{R}}\hat{\text{err}}^{(1)} \tag{3.2.17}$$

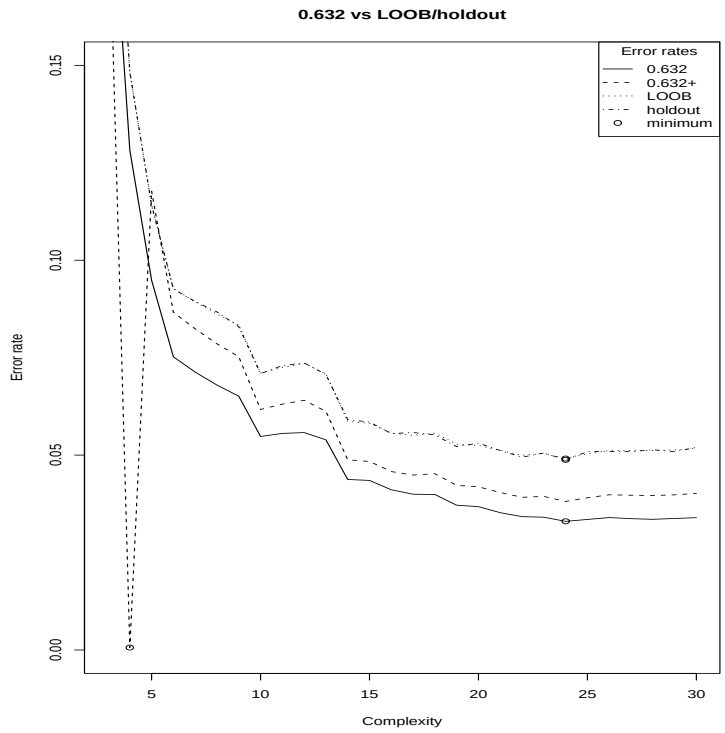


Figure 3.7: Comparison between the 0.632-family of estimators and the holdout error rate.

See figure 3.7 for a comparison between the 0.632-family of estimators and the holdout error rate. Notice how the $0.632^+$ estimator behaves erratically in one point, putting high weight on $\bar{\text{err}}$ and negative weight on $\hat{\text{err}}^{(1)}$. The holdout and LOOB errors nearly overlaps.

## 3.3 Comparing classifiers

The comparison of classifiers should not be arbitrary. Ripley (1996) notes that this really amounts to paired experiments, as the same data is used in all the classifications and in all the error rate calculations. Ripley suggests using McNemar's test to compare two classifiers.

McNemar (1947) gives the statistic

$$\frac{(n_a - n_b)^2}{n_a + n_b} \tag{3.3.1}$$

where $n_a$ and $n_b$ are the number of erroneously classified samples in the two classifiers respectively.

This test is related to the sign test, and follows a chi-square distribution.

Edwards (1948) applies a continuity correction to the above statistic so that the chi-square and the normal distribution can be used with more accuracy:

$$\Lambda = \frac{|n_a - n_b| - 1}{\sqrt{n_a + n_b}} \tag{3.3.2}$$

The null-distribution of $\Lambda$ is the standard normal distribution.

| Procedure | erroneously classified | holdout error |
|---|---|---|
| A | 106 | 0.06 |
| B | 88 | 0.05 |
| $\Lambda$ | | 1.22 |
| p-value | | 0.22 |

Table 3.1: An example of McNemar's test

Table 3.1 shows an example of McNemar's test. At a $\alpha = 0.05$ confidence level the test is rejected, and the null-hypothesis of no difference must be accepted. This also serves as a good example of the consideration of test set size, see figure 3.4 on page 39. A test set at least three times the one used would be needed to discern procedure A and B at the observed error rates.

This test works well for comparing two classifiers. If it is to be used to compare several classifiers, the Bonferroni correction must be employed to counter the total type-I error.

A related test to $\Lambda$ ( 3.3.2) is the Q-test of Cochran (1950). This test simultaneously compares $K$-procedures. - In much the same way as the F-test replaces several t-tests, the Q-test can replace several McNemar's tests.

Let

$$X_{nk} = \begin{cases} 1 & \text{if the k'th classifier correctly classifies the n'th observation} \\ 0 & \text{else} \end{cases} \tag{3.3.3}$$

Define

$$T_k = \sum_{n=1}^{N} X_{nk} \qquad \bar{T} = \frac{1}{K} \sum_{k=1}^{K} T_k \tag{3.3.4}$$

$$S_n = \sum_{k=1}^{K} X_{nk} \tag{3.3.5}$$

The Q-statistic is

$$Q = \frac{K(K-1) \sum_{k=1}^{K} (T_k - \bar{T})^2}{K \sum_{k=1}^{K} - \sum_{n=1}^{N} S_n^2} \tag{3.3.6}$$

This statistic will under the null-hypothesis follow a chi-square distribution with $K - 1$ degrees of freedom.

| Procedure | erroneously classified | holdout error |
|---|---|---|
| $P_1$ | 89 | 0.05 |
| $P_2$ | 107 | 0.06 |
| $P_3$ | 88 | 0.05 |
| $P_4$ | 88 | 0.05 |
| $P_5$ | 88 | 0.05 |
| $P_6$ | 104 | 0.06 |
| Q | | 83.75 |
| p-value | | 0.00 |

Table 3.2: An example of Cochran's Q-test

See table 3.2 for an example of Cochran's Q-test. Here the null-hypothesis of no difference between the procedures, is rejected. Confer McNemar's test in table 3.1 on the previous page, which had another conclusion for two of the procedures showed here.

## 3.4 Discussion

Non-parametric methods for performance assessment have been developed. Parametric methods both for assessing the underlying classifier assumptions

(normality) and classifier performance (tailor-made tests), can be found in appendix A.

It is not perfectly clear as to which method is the best at assessing classifier performance.

The holdout method is to be preferred if much data are available. While cross-validation and the 0.632-family could rely on somewhat less data.

The application in mind should also be considered when choosing the assessment methodology.

I will investigate how the LDA classifier with wavelet features perform on hyperspectral remote sensed data.

The objective of this exercise is to develop general procedures that can perform well under *the curse of dimensionality*.

This curse will only appear in data scarce situations. A rule of thumb for when this curse appears, is in a data size at about ten times the number of classes. This rule is partly based on the number of parameters in the LDA classifier.

I will operate with two parted sets. Either the sets will be parted by an expert or it will be parted by random sampling.

On the first set I will be free to do whatever pleases me. The second set (validation set), will be retained to verify the results found on the first.

The first data set is both larger than the training samples I have allowed myself, and hopefully representative. In this data set I will use a bootstrap estimate for the holdout error rate.

If the first set has $N$ samples and I have allowed myself $n$ training samples, this will work as following:

(i) Draw $n$ samples from the $N$ available.

(ii) Train the classifier on these samples.

(iii) Classify the $(N - n)$ remaining samples.

(iv) Repeat the above steps $B = 1000$ times.

(v) Calculate the total holdout error rate.

This bootstrap holdout error rate is related to the leave-one-out bootstrap (LOOB) error rate. It is more conservative than the 0.632-family of error-rates, and suffers from some of the shortcomings of the LOOB error rate. This is not cross-validation, although it may be confused for the uncontrolled cross-validation.

The difference in $n$ and $N$ (n$\ll$N) calls for a bootstrap estimate. - Otherwise the difference between samples (and within classes) would influence the error estimate too much. In the bootstrap estimate these *bad* samples will only be reflected in the bootstrap confidence interval.

In the introduction, I gave a conceptual drawing (page 1.4) that represents all the classifier systems I will investigate. Step (ii) and (iii) above are the system shown on the drawing. The systems I will consider are partly automatic, but requires that a few parameters (# coefficients and which mother wavelet) be selected. The steps outlined above, will first select these parameters based on minimum error, and then report an estimate for the error.

Finally the parameter choices and performance estimate, will be tested on the validation set.

# CHAPTER 4

# Datasets

This chapter describes the datasets used in this thesis, and the processes that influence them. Baseline classification results for future comparisons are established.

It is important to understand the processes influencing data acquisition and how these might be present in the data. Some of these effects make the wavelet representation desirable. A baseline study is important such that dissimilar methods can be compared to an established foundation.

Some material in this chapter build on Schowengerdt (1997). -Some knowledge of physics will be assumed. For reference an undergraduate text or outline of physics, especially electromagnetism or optics, will be convenient if available.

First influences of sensors and platforms are treated. - Then atmospheric influences are treated and the desirable properties of the wavelet transform showed. The datasets are described and a baseline study established.

## 4.1 Sensor and platform characteristics

The sensor under consideration is an optical sensor. Data from this sensor can be used in classification in two ways: - either spatial or spectral. In the spatial context objects are classified by shape. In the spectral context objects are classified by how they absorb and reflect light at different wavelengths (colour). A combination of the two approaches is also possible.

There are drawbacks with both approaches. The spatial approach is limited by the

spatial resolution of the sensor, as well as by possible obstruction and deformation of objects. Classification by spatial means will not be considered in this thesis.

Spectral classification is susceptible to attenuation by the atmosphere, which will be covered in section 4.2. The second most objectionable feature of this approach is the variability within object classes. This variability is due to:

- inhomogeneity in the object (e.g. minerals composed of varying amounts of different components)
- seasonal changes in the biosphere (e.g. leaves of deciduous trees changing colour and eventually total loss of leaves)
- water content (e.g. dry or soaked marsh)

This variability is not necessarily a bad thing. Depending on the objective of the classification study, the variability can suggest additional subclasses. Variability in crop spectra can be used to decide if the crop has acquired a disease or when it ideally should be harvested. In mineral classification, variability can be used to indicate how a mineral was formed, while in environmental inventories a marsh can be classified as either a carbon sink or carbon source (net contributor of carbon dioxide).

### 4.1.1 Separation of light into its components

A spectrometer is a device that measures light (electromagnetic-radiation) intensity, and sometimes polarisation, at different wavelengths.

Snell's law allows for the separation of (polychromatic) light into its (monochromatic) components. Snell's law has been known (under different names) for at least a millennium. Snell's law states how reflection and refraction are dependent on the angle of incidence of the light, its wavelength, and the speed of light in different media.

$$\sqrt{\frac{1}{v_{1,\lambda}^2}\frac{\mu_{1,\lambda}}{v_0}}\,\sin\theta_1 = \sqrt{\frac{1}{v_{2,\lambda}^2}\frac{\mu_{2,\lambda}}{v_0}}\,\sin\theta_2 \tag{4.1.1}$$

Here $\lambda$ is the wavelength of light under consideration, $\theta_1$ is the angle of incidence and $\theta_2$ is the angle of refraction. $v_{.,\lambda}$ is the phase velocity (i.e. effect of electric field, permittivity) of the light at wavelength $\lambda$ in the media. While $\mu_{.,\lambda}$ is the permeability (effect of magnetism) of the media at wavelength $\lambda$, $\mu_0$ is this effect in vacuum. See Hecht (1975) for details.

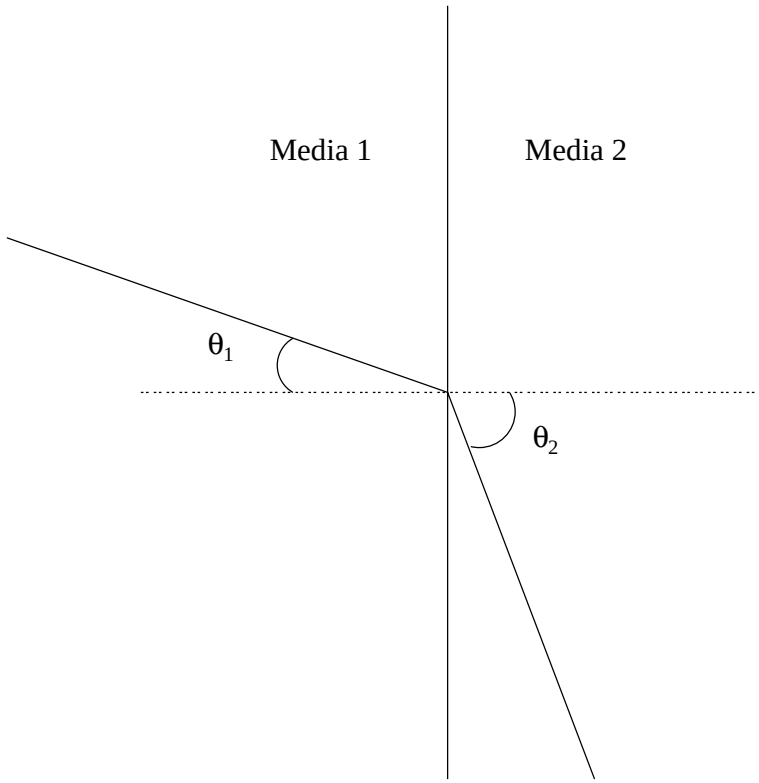


Figure 4.1: Snell's law

In a prism Snell's law spreads (spatially) polychromatic light out into its spectrum, and the light intensity can be measured at different wavelengths.

More modern spectrometers use the wave-particle duality of light, where constructive and destructive interference in a *diffraction grating filter* separates lights into its components.

### 4.1.2 Platform effects

Spectrometers can be used in laboratories where very little outside interference could be assumed. However the kind of spectrometer used in remote sensing applications, is by the very nature of remote sensing prone to problems related to the platform employed.

The most common platforms are aeroplanes and satellites. These platforms move over the scene and registers discrete points. Points (i.e. pixels) may overlap or be missing completely.

In figure 4.2 on the following page four different imaging modes are shown. The basic imager is the array imager. This instrument consists of a matrix of light

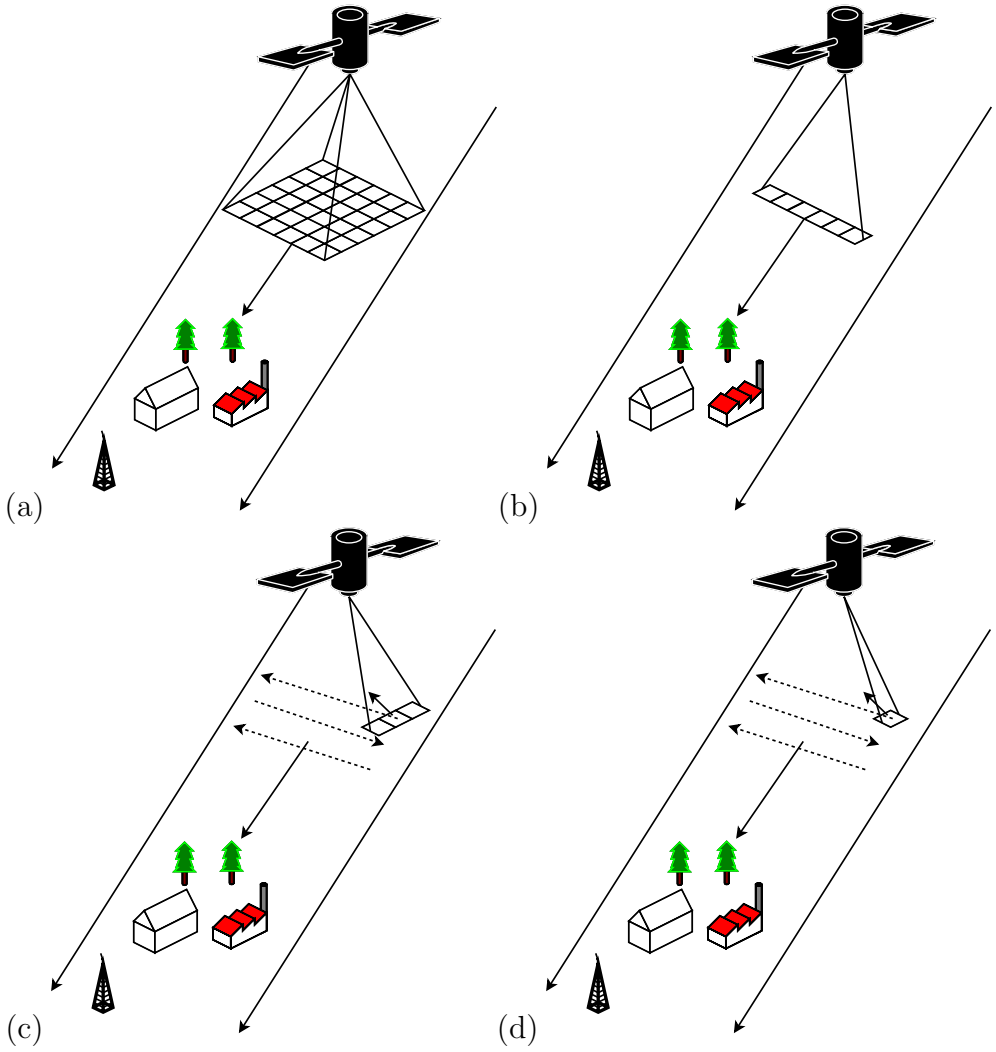


Figure 4.2: (a) Array imager, (b) pushbroom scanner, (c) whiskbroom scanner and (d) line scanner

sensors (radiometers). The most common array imager is the charge-coupled device (CCD). In this instrument the matrix consists of light sensitive capacitors. When light shines on each capacitor, it builds a charge that is measured at certain intervals. The upcoming competitor to the charge-coupled device is the active pixel sensor (APS[1]). In this sensor the matrix is constructed of photo diodes. These diodes are semiconductors that when hit by sufficient light, either exhibit the photovoltaic effect (e.g. solar or photovoltaic cell) or photoconductive behaviour (i.e. variable resistance). Active pixel sensors are to be preferred on grounds of noise control and manufacturing cost.

The array imager considered up to now, creates images instantaneously. This image is panchromatic (greyscale), which may be suitable for many applications. The platform can be made multi-spectral by adding more array imagers with different (diffraction grating) filters, or by adding some sort of filter changer to the present array imager.

This is possible for a few spectral bands. -However by employing the motion of the platform, a pushbroom scanner (b in figure 4.2 on the previous page) can attain an excessive amount of spectral bands without adding more array imagers. Hyperspectral imaging is possible as only one spatial dimension is imaged at once (cross-track); - The other spatial dimension (in-track) is imaged as the platform moves. This frees matrix elements in the in-track direction. The freed elements can now be used to collect light at different wavelengths through diffraction grating filters.

This seemingly ideal approach introduces some error. If the platform is an aeroplane, its speed, flight dynamics (pitch, roll and yaw) and its path will vary dramatically during imaging. Besides introducing erroneously placed geographical points and overlapping, this introduces both spatial correlation and spectral correlation (Doppler effect). If the sensor platform is a satellite, these problems are reduced as orbits tend to be more stable. - However physical bodies tend to rotate and if the satellite is not moving in the direction of rotation, skew is introduced in both the cross-track and in-track directions. Deviation from the direction of rotation is nonetheless the only way to get reasonable coverage. Theoretically the satellite can have perfect circular orbit, but in reality a more elliptical orbit is accepted. Near the apoapsis and periapsis some of the speed and flight dynamics problems of the aeroplane are re-introduced.

The whiskbroom scanner (c in figure 4.2 on the preceding page) can further increase the spectral resolution by dedication more matrix elements to the spectral dimension. The elements lost in the cross-track direction are replaced by a mirror that sweeps over the track. More elements should be allocated in the in-track direction to compensate for the time the mirror uses to move over the track.

[1]In commercial contexts this sensor is often referred to as a CMOS (Complementary Metal Oxide Semiconductor) sensor from one of its production processes.

Cross-track coverage can be increased by letting the mirror tilt further.

Besides inheriting the shortcomings of the pushbroom scanner, additional sources of error are introduced. The extra movement introduces more element overlapping and position error, as well as correlation (partial overlapping) in the cross-track dimension. If the mirror tilts considerably away from the nadir (straight down) direction, to increase the cross-track coverage, the on ground spatial resolution may vary substantially between the nadir and extreme tilt positions. Mechanically vibration of the whole sensors may also be a problem. This type of scanner is inheritably more complex than the two previous referred scanners. This is not only reflected in the errors introduced, but on how the scanner may fail. After four years of operation the cross-track mirror systems of the *USGS/NASA Landsat 7* satellite failed, losing 1/4 of all lines all over the scenes (Reichhardt (2003)).

The line scanner (d in figure 4.2 on page 52) is the extreme variant of the whiskbroom scanner, with only one scanning element. This scanning mode is largely historic, and can be found in old sensors that have proven their time, or in new experimental sensors where more imaging elements would be prohibitively expensive.

Some platform effects can be corrected for or flagged as susceptible if telemetry and engineering data are collected. Ideally one should be aware of such corrections to ascertain its effects on data analysis. In the Landsat 7 case above, the missing data is imputed with data from before the anomality appeared, creating potential interesting results in land change and use studies.

Besides the issues above, heat management may be a challenge as the sensors and mechanics may behave differently at different temperatures. Heat is essentially the movement of particles. These particles will have a charge and thus electromagnetic radiation will be emitted close to the sensor.

## 4.2 Atmospheric influence

In (a) of figure 4.3 on the following page laboratory spectra of two minerals can be seen. It should be fairly easy to classify (separate) these two minerals from their spectra. - However, their field spectra may look like (b) in the same figure, making it more complicated to separate the two.

This noticeable impact is due to the influence of the atmosphere. This system where sunlight can take different paths to the sensor, is illustrated in figure 4.4 on page 56.

The light source in this system is the sun. For all practical purposes the sun is a standard black-body object (i.e. none reflecting and none transparent), radiating

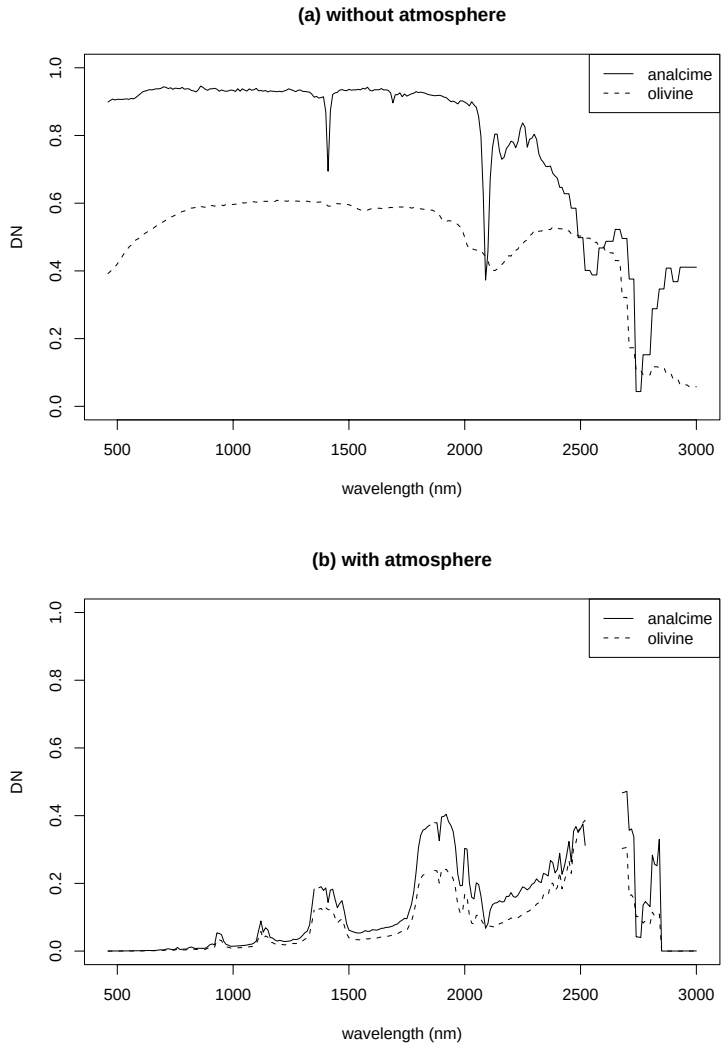


Figure 4.3: Spectra of two common minerals in (a) laboratory conditions, (b) simulated field conditions (satellite above 65km). See section 4.3 for a note on the simulation.

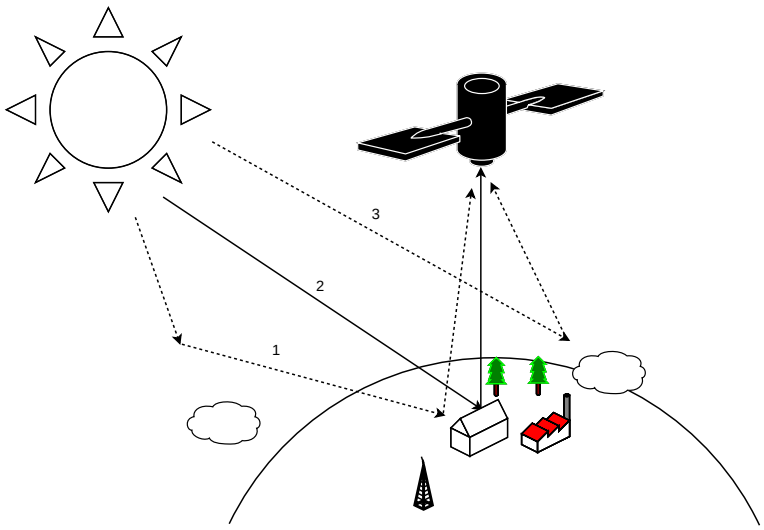


Figure 4.4: Scattering in the atmosphere

with different intensity at different wavelengths according to Planck's law.

The sunlight travels approximately $150 \times 10^6$km through space with little hindrance. Upon reaching earth, it has to travel through the atmosphere. The light can take different paths through the atmosphere, see figure 4.4. If we focus on the straight path (path 2), the atmosphere will work as a filter. The downward transmittance can be seen in (a) of figure 4.5 on the next page.

This filtering is from the principal absorbing gases: water vapour ($H_2O$), carbon dioxide ($CO_2$), oxygen ($O_2$) and ozone ($O_3$). Ozone mainly resides in the stratosphere (10-50km above sea level), while the other absorbing gases are trapped below in the exosphere. Water vapour is nearly exclusively found below 4km. Clearly the height of both the observer and the object under observation will have influence on the absorbed spectra.

If the light reflected of the observed object follow a straight path to the observation platform (e.g. satellite or aeroplane), the atmospheric filter is passed once more. - The picture is more complicated. The light that hits the sensor can follow two additional paths. Sunlight may hit particles in the atmosphere and be scattered in different directions. If the particles hit are much smaller than the wavelength of the light, this scattering is known as Rayleigh scattering. Besides wavelength, the behaviour of Rayleigh scattering depends on the angle of inclination. Rayleigh scattering is responsible for the blue skylight in the day and the crimson colours of the evening. In figure 4.4 skylight follows path (1), hits the object under consideration and travels to the sensor. This light is more blue (440-490nm) than the light taking path (2) in the figure. Path (1) also shows that light reflected of other object than the target object might hit the sensor.

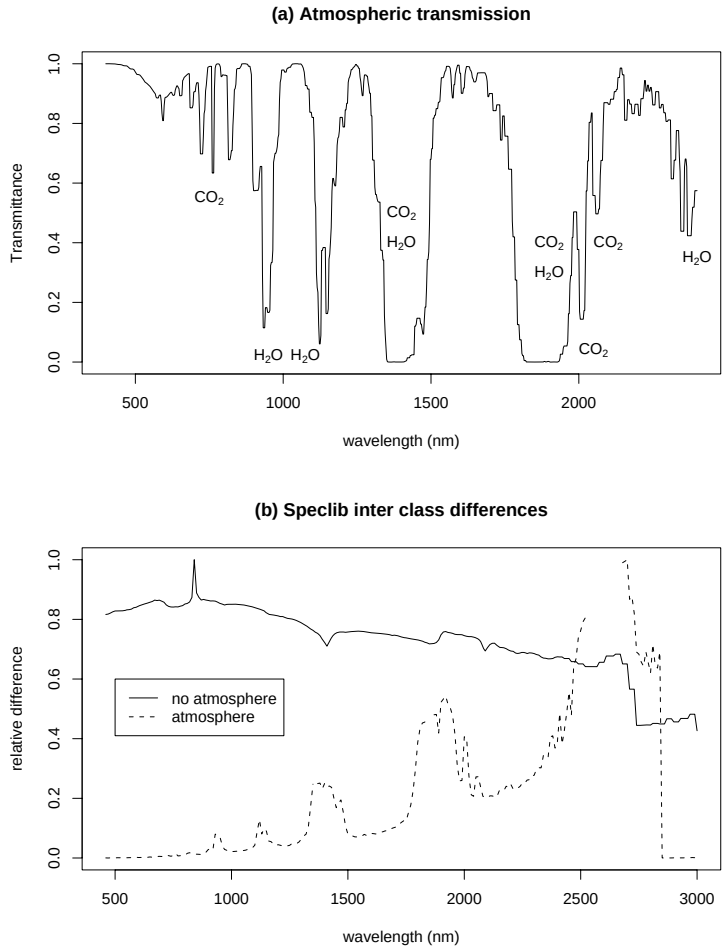


Figure 4.5: (a) Downwards atmospheric transmission. (b) Inter class differences, with and without atmosphere. See section 4.3 for a note on the simulation.

Both Snell's law (of section 4.1.1) and the Rayleigh scattering above are *"special"* cases of Maxwell's (electro-magnetic) equations. Mie solved these equations for scattering from spherical particles of any diameter. All such scattering that is not Rayleigh scattering (diameter $\ll \lambda$), is known as Mie scattering. In figure 4.4 on page 56 path (3) is known as in-path or sky-path, and is composed of light from both Mie and Rayleigh scattering.

## 4.3 A note on atmospheric simulation

Healey & Slater (1999) conduct an interesting experiment, where objects with known spectra are sensed through the atmosphere, and the MODTRAN (MODerate resolution atmospheric TRANsmission) radiative transfer code is used to simulate the same.

In the figures in this chapter, I have used the *Second Simulation of the Satellite Signal in the Solar Spectrum* or *6S* radiative transfer code of Vermote et al. (1997), along with spectral data from the USGS Spectral library (Clark et al. (1993)).

Radiative transfer codes have a myriad of options and conditions to set. Healey & Slater (1999) choose some reasonable (expert knowledge) set of these, and simulate nearly twenty-thousand combinations of these. This approach is admirable and reaches the objectives set forth.

I have elected not to follow this approach. I would like to stress that some uncertainty bounds on the interclass differences of figure 4.5 on the preceding page would be highly desirable.

If I were to do this, I would first study the variance data of the spectral library. Secondly I would not trust expert knowledge alone for the atmospheric data. Latin hypercube sampling (Mckay et al. (1979)) could be used to rapidly identify parameters that influence the outcome. This could serve as a basis for building an *emulator*. See Currin et al. (1991) and Oakley & O'Hagan (2002) for details on emulators. Once created, this emulator can be used to put realistic uncertainty bounds on both the outcome and the elicited expert knowledge.

## 4.4 Why the wavelet representation should be a good representation

Besides the agreeable theoretical properties of chapter 2, two properties of more practical importance are demonstrated. These properties should counter some of the problems described in the previous sections.

### 4.4.1 Cusps and singularities

In figure 4.3 on page 55 approximated cusps and singularities are introduced to the spectra. These can be attributed to atmospheric effects, among them the straight filter in (a) of figure 4.5 on page 57.

The USGS Spectral library provides nearly five hundred laboratory spectra. Relative differences in these were calculated with and without an atmosphere (as simulated with 6S). The results are shown in (b) of figure 4.5 on page 57. Compared to (a) of the same figure, apparently most information is present near heavily attenuate regions. These regions bear resemblance to approximated cusps and singularities.

The "time-scale localisations" (see section 2.4) properties of wavelet decompositions make them ideally suited to represent such regions in appropriate detail.

### 4.4.2 Decorrelation

In figure 4.6 on the following page the covariance matrix of the Pavia dataset (section 4.5.1) is shown. Lighter colour indicates higher correlation. By a very crude wavelet decomposition, this correlation is significantly reduced. The crude wavelet decomposition in the illustration retains only the first level of the wavelet decomposition. This experiment is done more in full in section 5.1. The aim of decorrelation is to reduce the correlation in the data while preserving other properties.

It should be noted that generally, wavelet components are not statistically independent (a stronger condition than no correlation). In certain decompositions high correlation can be anticipated. Such situations occur when the data shape differs at many scales from the wavelet atoms.

In Percival (2000) there are given several exercises, where one is asked to prove the decorrelating property of the wavelet decomposition for some special cases. Decorrelating transforms are more generally known as whitening transforms. - Among the more standard feature extraction methods, both *independent component analysis* (ICA) and *principal component analysis* (PCA) fall into this category. In comparison to the wavelet decomposition, these decorrelating transforms have no more general established decorrelating properties. -Nevertheless, they are applied in many situations where existing proofs would not warrant their use.

Flandrin (1992) proves the decorrelation property for *fractional Brownian motion*, while at least some asymptotic results for (long range dependence) Gaussian data could be gotten from Johnstone & Silverman (1997). Capobianco (2004)

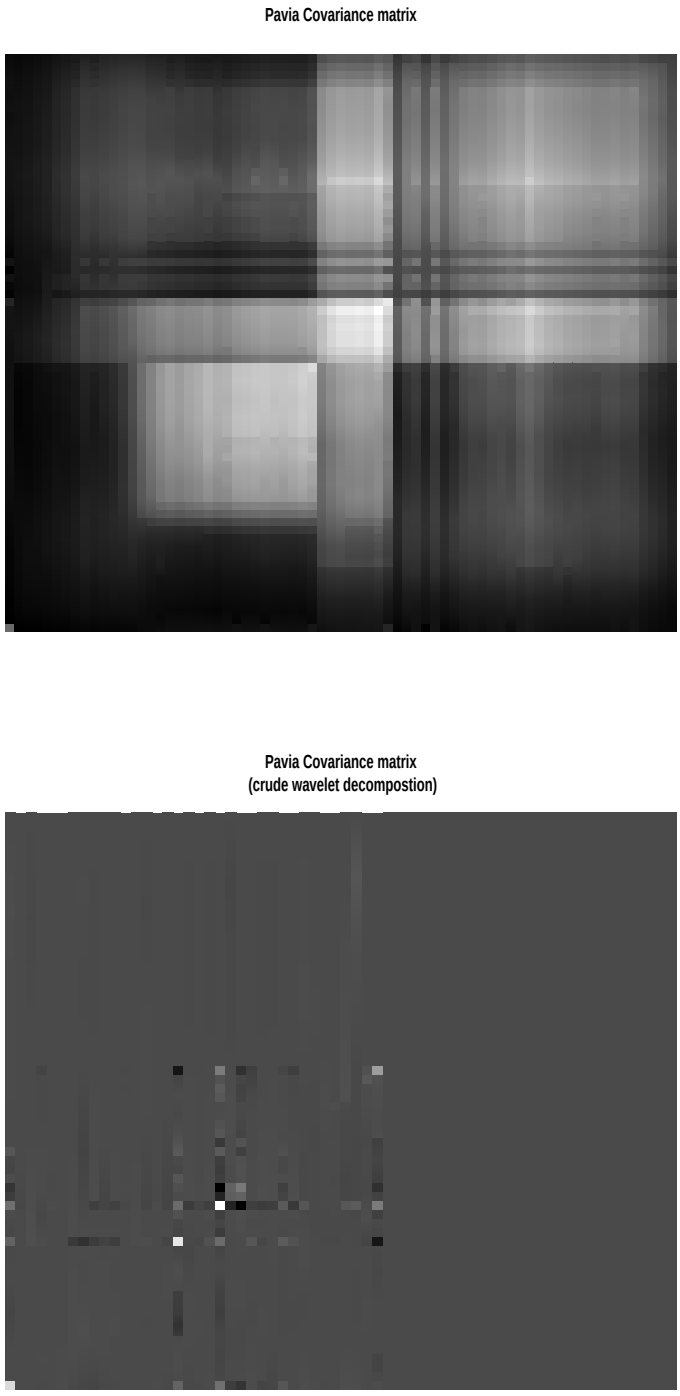


Figure 4.6: Covariance matrix of the Pavia dataset and the covariance matrix of a crude wavelet decomposition, both rotated 90 degrees. Lighter colour indicates higher value.

convincingly employs wavelet decompositions coupled with ICA to model non-Gaussian curves based on the (Gaussian) return of stock indices. Tian et al. (2000) subscribe to the view that real-world signals can be fairly decorrelated by wavelet decompositions. Instead of ascertaining this belief, they describe some shortcomings and a scheme to improve on these.

The normality test of section A.1 is suited to illustrate the improvement in whitening. In an experiment on a subset of the Pavia dataset, five columns were chosen at random, from both the raw data and from a crude wavelet decomposition. The result can be found in table 4.1. The raw data is good with regard to whitening, but the crude decomposition is even better. In figure 4.6 on the preceding page the covariance matrices of the full sets are shown.

| Form | Koziol | | Mardia | | | | | Royston | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $J_n$ | p | A | df | p | B | p | H | e | p |
| raw | 0.76 | 0.00 | 187.34 | 35 | 0.00 | 1.99 | 0.05 | 0.66 | 4.17 | 0.96 |
| crude | 1.33 | 0.00 | 226.40 | 35 | 0.00 | 3521.70 | 0.00 | 3.93 | 4.17 | 0.44 |

Table 4.1: Whitening in the raw data versus whitening in a crude wavelet decomposition

## 4.5 Characteristics of the available datasets

### 4.5.1 The Pavia dataset

Some details of this dataset can be found in Gamba (2004). The data was collected over the city of Pavia (northern Italy) in the summer of 2002. The *Digital Airborne Imaging Spectrometer* (DAIS) of the DLR (the German aerospace centre) collected 79 spectral bands. For calibration and various reasons, only 71 of these are available. The DAIS instrument has a web-page[2] where the instrument is described.

The class distribution in the training and test sets are detailed in table 4.2 on the next page. The parting of training and test sets is to the best of my knowledge done by expert knowledge. The covariance matrix of the data is illustrated in figure 4.6 on the preceding page.

[2] `http://www.op.dlr.de/dais/dais-scr.htm`

| | total | | training | | test | |
|---|---|---|---|---|---|---|
| | # | % | # | % | # | % |
| Water | 4285 | 0.28 | 202 | 0.11 | 1136 | 0.09 |
| Trees | 2507 | 0.17 | 205 | 0.11 | 1301 | 0.10 |
| Asphalt | 1342 | 0.09 | 206 | 0.11 | 2302 | 0.17 |
| Parking lot | 1506 | 0.10 | 205 | 0.11 | 1630 | 0.12 |
| Bitumen | 1834 | 0.12 | 204 | 0.11 | 2041 | 0.15 |
| Roofs | 333 | 0.02 | 201 | 0.11 | 4083 | 0.31 |
| Meadow | 2356 | 0.16 | 315 | 0.17 | 159 | 0.01 |
| Soil | 693 | 0.05 | 202 | 0.11 | 132 | 0.01 |
| Shadow | 278 | 0.02 | 119 | 0.06 | 491 | 0.04 |

Table 4.2: Class distribution in the Pavia dataset

### 4.5.2 The Fontainebleau dataset

The Fontainebleau forest is situated 60 km south-southeast of Paris, France. The original dataset consisted of six classes of trees. -Oaks of three heights, beeches of two heights and one type of pine. In the dataset at my disposal the number of classes is reduced to three by combining trees of different heights. Figure 4.7 on the next page illustrates class overlaps across the spectrum.

ROSIS the *Reflective Optics System Imaging Spectrometer* of the DLR (the German aerospace centre) collected 81 bands in the $430 - 830$nm range during the *European Multisensors Airborne Campaign* (EMAC'94). This instrument has its own web-page[3], where some details are presented. - Of notoriety to the discussion of previous sections of this chapter, the spectral bandwidth varies from 12 to 4nm at each end.

In table 4.3 the class distribution is detailed. The dataset comes without any designated training and test subsets. - For this dataset, such subsets are drawn randomly.

| | # | percent |
|---|---|---|
| Oak | 5195 | 0.64 |
| Beech | 2083 | 0.26 |
| Pine | 807 | 0.10 |

Table 4.3: Class distribution in the Fontainebleau dataset

[3] `http://www.op.dlr.de/ne-oe/fo/rosis/home.html`

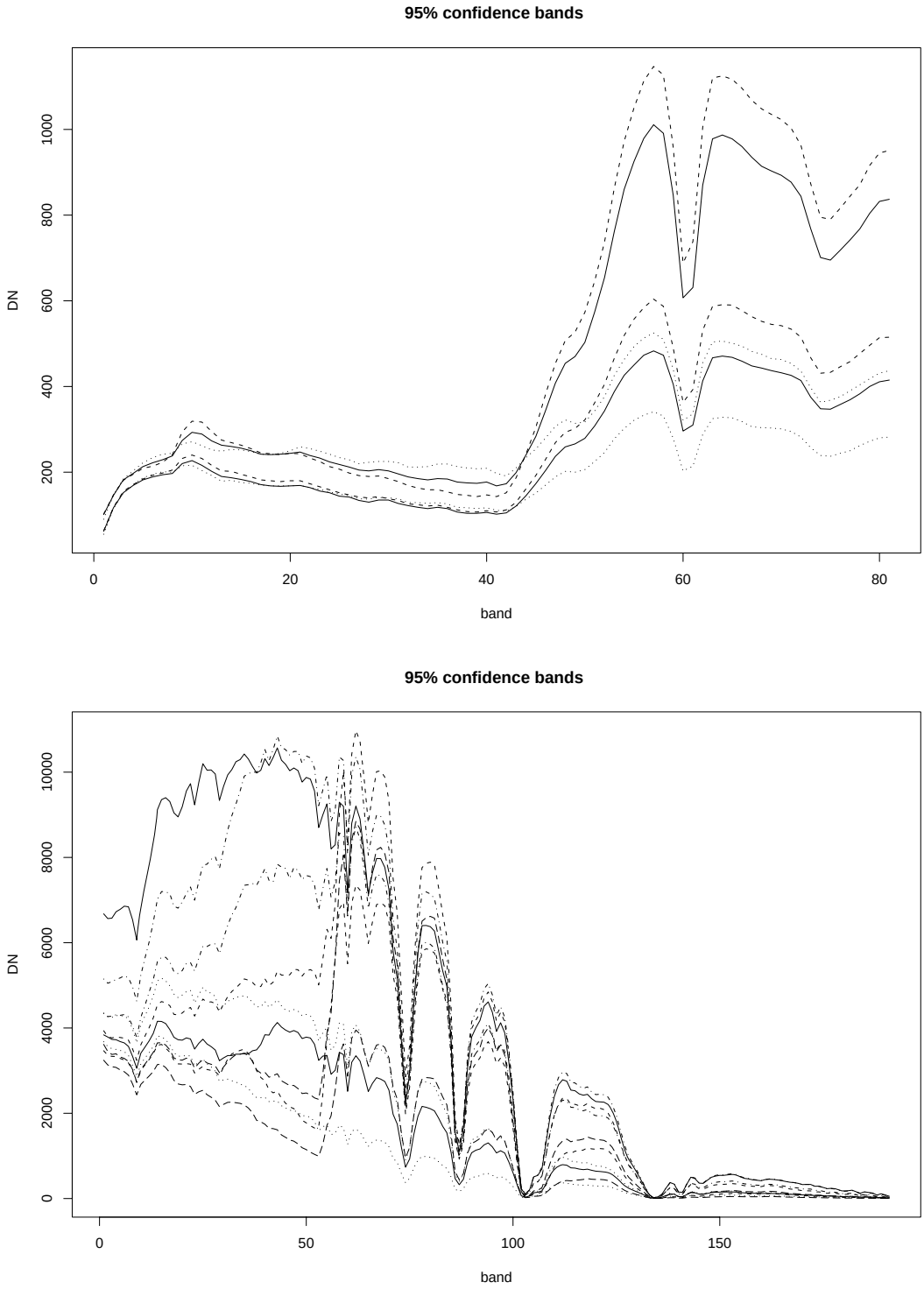


Figure 4.7: *Top:* The Fontainebleau dataset, 95% confidence bands for the class mean spectra radiance (training dataset). *Bottom*:The National Mall dataset, 95% confidence bands for the class mean spectra radiance (training dataset).

### 4.5.3 The National Mall dataset

The US National Mall is a green area stretching from the White House to the Washington monument (white obelisk), in the US capital. This scene was imaged in the autumn of 1995 by the *Hyperspectral Digital Imagery Collection Experiment* (HYDICE) sensor. This dataset can be found on the CD-ROM accompanying Landgrebe (2003). In total 210 spectral bands in the 400-2500nm range were collected with a nearly regular bandwidth of 10nm. For various undocumented and noise reasons, many bands were discarded, resulting in 191 available.

In table 4.4 the class distribution is detailed. The dataset comes without any designated training and test subsets. - For this dataset, such subsets are drawn randomly. In figure 4.7 on the preceding page class overlapping is illustrated. -Although such a comparison is not valid, notice how this dataset does not improve on the lower dimensional Fontainebleau dataset.

The HYDICE sensor was built for the US Naval Research Laboratory, based on the experience made with the AVIRIS (*Airborne Visible/Infrared Imaging*) sensor. This sensor behaves more like an ideal pushbroom scanner than both the ROSIS and DAIS sensors. Details are in Basedow et al. (1995).

| | # | percent |
|---|---|---|
| Roofs | 4902 | 0.49 |
| Streets | 895 | 0.09 |
| Grass | 638 | 0.06 |
| Trees | 2168 | 0.22 |
| Paths | 1477 | 0.15 |

Table 4.4: Class distribution in the National Mall dataset

## 4.6 Baseline results for comparison

To ascertain the quality of new methods, it is important to have an established standard solution to compare with. There exists a handful of reasonable feature extraction methods to compare with, e.g. *Decision Based Feature Extraction* (Lee & Landgrebe (1993)) and *Projection Pursuit* (Huber (1985)). Most of these lack wide support. The age-old *Principal component analysis* is an exception and will be used.

### 4.6.1 Principal component analysis

Principal component analysis (PCA) is an (orthogonal) linear transformation, through which data is projected (rotated) into a new coordinate system. This coordinate system is such that most variability is present along the first axis, the second most variability along the next axis and so forth. If all axes are retained, no reduction in dimensionality is achieved. If variability is considered an important trait in the application at hand, dimension reduction can be attained by keeping just the first few of the components (axes). In this scheme some information is lost as the original space is not spanned by the spanning set of the reduced space.

Assume the data matrix $X$, let

$$\mathcal{X} = X - \bar{X} \tag{4.6.1}$$

be a centred data matrix. Let

$$\begin{aligned} V &= \mathrm{Cov}(\mathcal{X}) \\ &= \tfrac{1}{N}\mathcal{X}^T\mathcal{X} \end{aligned} \tag{4.6.2}$$

be the corresponding covariance matrix. Covariance matrices are real and symmetric; thus the spectral theorem applies:

$$\begin{aligned} V &= QVQ^t \\ &= Q\Lambda Q^t \\ &= Q \begin{pmatrix} \lambda_1 & 0 & \dots \\ 0 & \ddots & 0 \\ \vdots & 0 & \lambda_p \end{pmatrix} Q^t \end{aligned} \tag{4.6.3}$$

where $\lambda_1 \geq \lambda_2 \geq \dots$ are the ordered eigenvalues.

Now let the principal component projection be:

$$Y = Q^t\mathcal{X} \tag{4.6.4}$$

$Q$ is orthonormal, thus $Q^{-1} = Q^t$. This leads to:

$$\begin{aligned} \mathrm{Cov}(Y) &= \tfrac{1}{N}Y^tY \\ &= \tfrac{1}{N}(Q^t\mathcal{X})^tQ^t\mathcal{X} \\ &= \tfrac{1}{N}Q^t\mathcal{X}\mathcal{X}^TQ \\ &= Q^tQ\Lambda Q^tQ \\ &= \Lambda \end{aligned} \tag{4.6.5}$$

All data transformed by the principal component projection will now have axes with ordered variability. In applications PCA is known as the discrete Karhunen-Loève transform.

If by some sensor knowledge, different spectral scale among bands are suspected, it might be beneficial to rescale the data matrix. This is essentially the same as replacing the covariance matrix in the analysis above with the correlation matrix. Further details and discussion can be found in Mardia et al. (1979).

### 4.6.2 Results

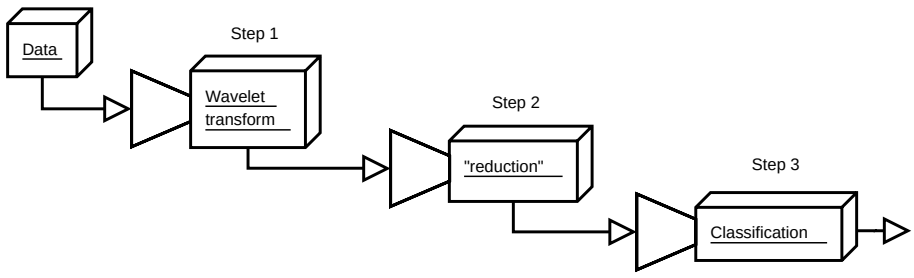


Figure 4.8: Conceptual illustration of how my methods will work.

In figure 4.8 the "reduce" step is replaced by the PCA. The experiment is conducted as described in section 3.4. I mention that I will provoke *the curse of dimensionality* by using few samples in training. Here the classifier has only been trained with 30 (Fontainbleau), 50 (National Mall) or 90 (Pavia) samples.

In figure 4.9 on the following page the response of the classifier to the different number of components is illustrated. Ideally these graphs should look like smiles, representing the variance-bias trade off. In this paradigm the error commited is decomposed into two parts. In the far-left of the smile the model is said to be under-fitted, and in the far-right over-fitted. In the under-fitted region variance dominates, and in the over-fitted region bias[4] dominates. The minima in the figures are global minima, as the classifier collapses outside the searched ranges.

In table 4.5 on page 68 the minima are detailed. For the Pavia and National Mall datasets the reduction has positive effect. If the Fontainebleau results are compared to its priors (table 4.3 on page 62), the effects of the PCA transform can be attributed to pure chance.

These results are validated in table B.2 on page 155, comments will be given in later chapters.

[4] i.e. lack of generalisation to new data

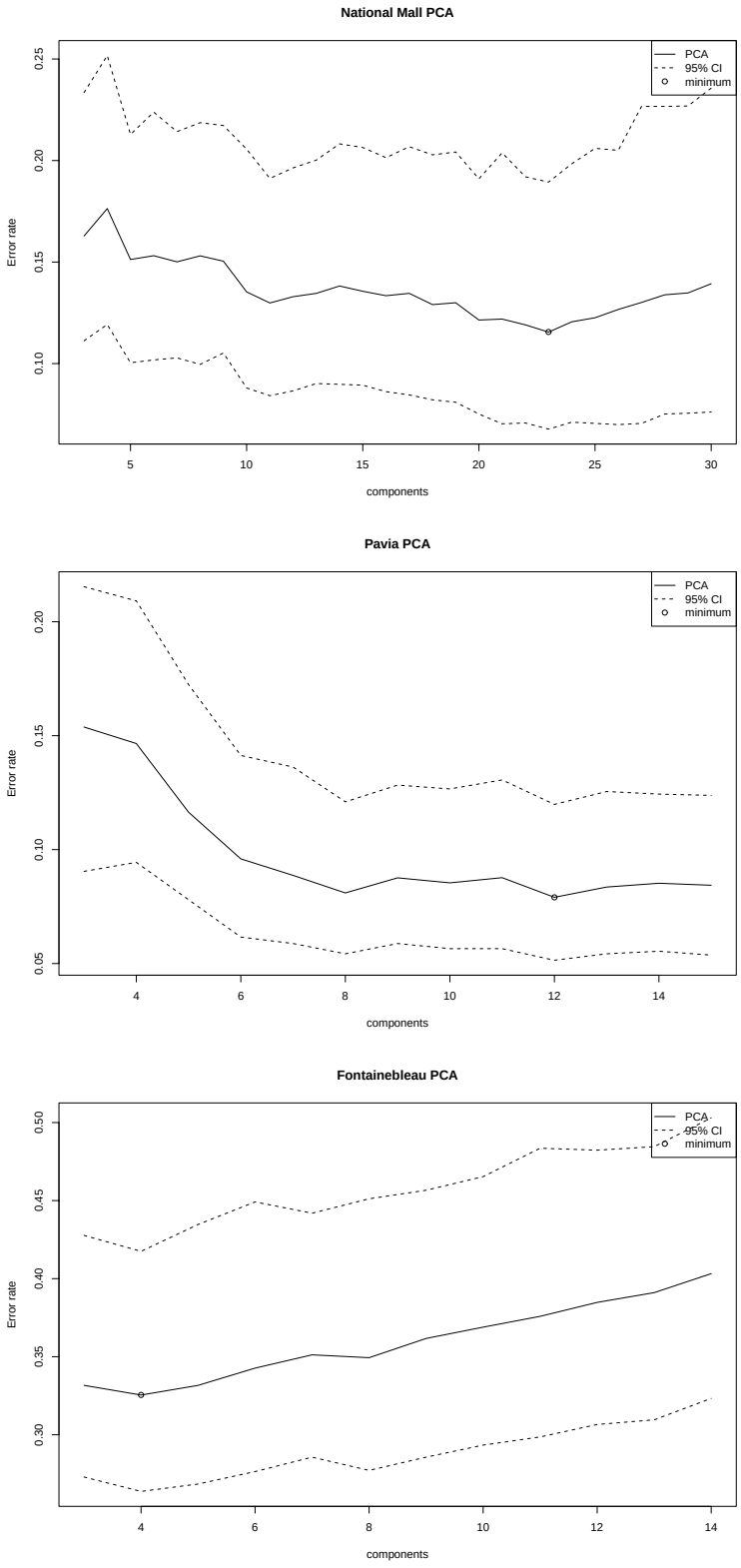
National Mall PCA
PCA
95% CI
minimum
Error rate
components
Pavia PCA
PCA
95% CI
minimum
Error rate
components
Fontainebleau PCA
PCA
95% CI
minimum
Error rate
components

Figure 4.9:

| | | | 95% CI | |
|---|---|---|---|---|
| Dataset | # components | error | low | high |
| Pavia | 12 | 0.08 | 0.05 | 0.12 |
| National Mall | 23 | 0.12 | 0.07 | 0.19 |
| Fontainebleau | 4 | 0.33 | 0.26 | 0.42 |

Table 4.5: Minimum error rate and confidence intervals for classification on principal components.

## 4.7 Short remarks

The three datasets at hand are not complete for testing the nature of classification systems. The two urban land use datasets are representative, and may give a feel for the quality in this application. The Fontainebleau dataset is challenging and may have inherent limitations.

Emulators can be built with real atmospheric data and *6S* or similar computer codes. With extensive spectral libraries and spectral mixing, wider conditions than those in the datasets above can be investigated. Better sensors and classification systems can be designed and evaluated in such systems.

The datasets considered are collected by sensors in aeroplanes. These collection campaigns should largely be regarded as preparatory experiments for anticipated wider deployment of spaceborne sensors.

# CHAPTER 5

# Atomic decomposition and best basis selection

This chapter relates atomic decomposition to the Coifman-Wickerhauser best orthogonal basis selection algorithm. The Coifman-Wickerhauser algorithm is perhaps the most used algorithm to select bases from the *wavelet packet decomposition*.

Methods to extract fewer, but important coefficients, are developed. The *Earth Movers Distance* (EMD) metric is also used in place of the standard measure in the Coifman-Wickerhauser algorithm.

First an example of classification on simple wavelet decompositions is given. Then atomic decomposition is presented. The Coifman-Wickerhauser algorithm is detailed, with and without modifications, before results are drawn and discussed.

## 5.1 Classification on simple wavelet decompositions

In many studies such as Kaewpijit et al. (2003) and Bruce et al. (2002) classification is done directly on wavelet decompositions at certain levels of decompositions, or on crude selected subsets there of. Most wavelet decomposition algorithms require data of dichotomous length. If this is not so, the data are often lengthened by zero-padding. This can create more features than the original bands, as the number of wavelet components in the decomposition is related to the length of the original data.

In the studies mentioned above, criteria such as correlation and class overlap are used to select features among the wavelet coefficients at the given level. In my

experience it is difficult to use these criteria with few observations, as the variation within each class dominates the criteria.

In table 5.1 on the following page classification is done on the father (scale) wavelet coefficients of the Pavia dataset, at different levels of decimation. In table 5.2 on page 72 the same is done for the mother (approximation) wavelet coefficients, and in table 5.3 on page 73 it is done on a combination of these. The experiment is done as outlined in section 3.4. The wavelets used are a selection of those that seem the most popular in literature. An overview of these are given in appendix C.

In section 3.4 I mention that I will provoke *the curse of dimensionality* by using few samples in training. On the dataset used here the stated thumb-rule warrants only 90 training samples. An overview of the wavelet families used is given in appendix C.

In further discussion, these result will not be heavily relied upon. One should note two things: - the many missing results, a result of the data to dimension ratio; -and the relative ease some near PCA quality results (confer table 4.5 on page 68) are obtained.

## 5.2 Atomic decomposition

In the introductory chapter, Grossman and Morlet's work on wavelets was linked to the atomic decomposition of the 1960's.

This atomic decomposition is detailed at length in Fefferman & Stein (1972) and Coifman & Weiss (1977). The language and notation may be awkward for applications, but essentially show that decomposition of functions (signals) by atoms is possible if these atoms adhere to some conditions. Feichtinger & Grochenig (1988) use the same setting and language with wavelets.

Atoms are fundamental or basic members of a family of functions. They are indexed by one or more parameters. These fundamental functions should, when combined in the appropriate way, be able to completely[1] represent their function space (family).

The most common such atomic decomposition is the Fourier transform. In this decomposition, the atoms are sines and cosines of different phases and periods. In theory any function could be an atom as long as at least some reconstructability and power of representation are retained. In figure 5.1 on page 74 (from the introductory chapter) two not too obvious candidates are shown.

-In practice previously established properties of wavelets should be suited if the

[1] completeness of representation, reconstructability, see section 2.5.2.

| | Wavelet | Level 3 | se | Level 4 | se | Level 5 | se |
|---|---|---|---|---|---|---|---|
| 1 | haar | n/a | n/a | n/a | n/a | n/a | n/a |
| 2 | d4 | 0.06 | 0.01 | 0.11 | 0.02 | 0.20 | 0.03 |
| 3 | d6 | 0.06 | 0.01 | n/a | n/a | 0.19 | 0.03 |
| 4 | d8 | n/a | n/a | n/a | n/a | n/a | n/a |
| 5 | d10 | 0.07 | 0.01 | 0.07 | 0.01 | 0.16 | 0.03 |
| 6 | d12 | 0.07 | 0.02 | 0.07 | 0.02 | 0.19 | 0.03 |
| 7 | d14 | 0.07 | 0.01 | 0.08 | 0.02 | 0.19 | 0.03 |
| 8 | d16 | 0.07 | 0.01 | 0.07 | 0.01 | 0.17 | 0.03 |
| 9 | d18 | 0.06 | 0.01 | 0.07 | 0.01 | 0.12 | 0.02 |
| 10 | d20 | 0.06 | 0.01 | 0.07 | 0.01 | 0.15 | 0.03 |
| 11 | la8 | 0.06 | 0.01 | 0.06 | 0.01 | n/a | n/a |
| 12 | la10 | 0.06 | 0.01 | 0.08 | 0.02 | 0.15 | 0.03 |
| 13 | la12 | 0.06 | 0.01 | 0.08 | 0.02 | 0.15 | 0.03 |
| 14 | la14 | 0.06 | 0.01 | 0.07 | 0.02 | 0.19 | 0.03 |
| 15 | la16 | 0.06 | 0.01 | 0.08 | 0.01 | 0.16 | 0.03 |
| 16 | la18 | 0.06 | 0.01 | 0.08 | 0.02 | 0.19 | 0.02 |
| 17 | la20 | 0.06 | 0.01 | 0.08 | 0.02 | 0.16 | 0.03 |
| 18 | bl14 | 0.06 | 0.01 | 0.07 | 0.02 | 0.17 | 0.03 |
| 19 | bl18 | 0.06 | 0.01 | 0.09 | 0.02 | 0.18 | 0.03 |
| 20 | bl20 | 0.06 | 0.01 | 0.09 | 0.02 | 0.18 | 0.03 |
| 21 | c6 | 0.07 | 0.01 | 0.07 | 0.01 | 0.19 | 0.03 |
| 22 | c12 | 0.06 | 0.01 | 0.06 | 0.01 | 0.16 | 0.03 |
| 23 | c18 | 0.06 | 0.01 | 0.07 | 0.02 | 0.17 | 0.03 |
| 24 | c24 | 0.06 | 0.01 | 0.08 | 0.02 | 0.18 | 0.03 |
| 25 | c30 | 0.06 | 0.01 | 0.08 | 0.02 | 0.13 | 0.03 |

Table 5.1: DWT father (scale) wavelet coefficients, LDA classification, 90 samples, error rate and SE.

| | Wavelet | Level 3 | se | Level 4 | se | Level 5 | se |
|---|---|---|---|---|---|---|---|
| 1 | haar | 0.08 | 0.01 | 0.10 | 0.02 | 0.27 | 0.03 |
| 2 | d4 | 0.09 | 0.02 | 0.12 | 0.02 | 0.19 | 0.03 |
| 3 | d6 | 0.08 | 0.01 | 0.13 | 0.02 | 0.28 | 0.03 |
| 4 | d8 | n/a | n/a | n/a | n/a | n/a | n/a |
| 5 | d10 | 0.09 | 0.02 | 0.13 | 0.02 | 0.26 | 0.03 |
| 6 | d12 | 0.09 | 0.02 | 0.19 | 0.02 | 0.18 | 0.02 |
| 7 | d14 | 0.11 | 0.02 | 0.13 | 0.02 | 0.27 | 0.03 |
| 8 | d16 | 0.12 | 0.02 | 0.14 | 0.02 | 0.18 | 0.02 |
| 9 | d18 | 0.12 | 0.02 | 0.13 | 0.02 | 0.17 | 0.03 |
| 10 | d20 | 0.10 | 0.02 | 0.13 | 0.02 | 0.18 | 0.02 |
| 11 | la8 | 0.09 | 0.02 | 0.13 | 0.02 | 0.19 | 0.02 |
| 12 | la10 | 0.08 | 0.02 | 0.12 | 0.02 | 0.27 | 0.03 |
| 13 | la12 | 0.08 | 0.02 | 0.10 | 0.02 | 0.15 | 0.02 |
| 14 | la14 | 0.09 | 0.02 | 0.12 | 0.02 | 0.20 | 0.02 |
| 15 | la16 | 0.08 | 0.01 | 0.14 | 0.02 | 0.23 | 0.02 |
| 16 | la18 | 0.07 | 0.01 | 0.15 | 0.03 | 0.20 | 0.03 |
| 17 | la20 | 0.07 | 0.01 | 0.14 | 0.03 | n/a | n/a |
| 18 | bl14 | 0.10 | 0.02 | n/a | n/a | 0.18 | 0.02 |
| 19 | bl18 | 0.08 | 0.01 | 0.14 | 0.03 | 0.28 | 0.03 |
| 20 | bl20 | 0.08 | 0.01 | 0.12 | 0.02 | 0.19 | 0.03 |
| 21 | c6 | n/a | n/a | n/a | n/a | n/a | n/a |
| 22 | c12 | n/a | n/a | n/a | n/a | 0.21 | 0.02 |
| 23 | c18 | n/a | n/a | n/a | n/a | n/a | n/a |
| 24 | c24 | 0.06 | 0.01 | n/a | n/a | n/a | n/a |
| 25 | c30 | n/a | n/a | n/a | n/a | n/a | n/a |

Table 5.2: DWT mother (approximation) wavelet coefficients, LDA classification, 90 samples, error rate and SE.

| | Wavelet | Level 4 | se | Level 5 | se | Level 6 | se |
|---|---|---|---|---|---|---|---|
| 1 | haar | 0.11 | 0.02 | 0.15 | 0.03 | 0.43 | 0.03 |
| 2 | d4 | 0.11 | 0.02 | 0.20 | 0.03 | 0.30 | 0.04 |
| 3 | d6 | 0.13 | 0.02 | 0.19 | 0.03 | 0.30 | 0.03 |
| 4 | d8 | n/a | n/a | n/a | n/a | n/a | n/a |
| 5 | d10 | 0.07 | 0.02 | 0.16 | 0.03 | 0.30 | 0.04 |
| 6 | d12 | 0.07 | 0.01 | 0.19 | 0.03 | 0.30 | 0.03 |
| 7 | d14 | 0.08 | 0.02 | 0.19 | 0.03 | 0.30 | 0.03 |
| 8 | d16 | 0.07 | 0.01 | 0.17 | 0.03 | 0.30 | 0.04 |
| 9 | d18 | 0.07 | 0.02 | 0.12 | 0.02 | 0.30 | 0.04 |
| 10 | d20 | 0.07 | 0.01 | 0.15 | 0.03 | 0.30 | 0.03 |
| 11 | la8 | 0.06 | 0.01 | 0.16 | 0.02 | 0.30 | 0.04 |
| 12 | la10 | 0.08 | 0.02 | 0.16 | 0.03 | 0.26 | 0.04 |
| 13 | la12 | 0.08 | 0.02 | 0.15 | 0.03 | 0.32 | 0.04 |
| 14 | la14 | 0.07 | 0.02 | 0.19 | 0.03 | 0.31 | 0.04 |
| 15 | la16 | 0.08 | 0.02 | 0.16 | 0.03 | 0.31 | 0.04 |
| 16 | la18 | 0.08 | 0.02 | 0.19 | 0.02 | 0.35 | 0.04 |
| 17 | la20 | 0.08 | 0.02 | 0.16 | 0.03 | 0.32 | 0.04 |
| 18 | bl14 | 0.07 | 0.02 | 0.17 | 0.03 | 0.31 | 0.04 |
| 19 | bl18 | 0.09 | 0.02 | 0.18 | 0.03 | 0.30 | 0.04 |
| 20 | bl20 | 0.09 | 0.02 | 0.18 | 0.02 | 0.30 | 0.03 |
| 21 | c6 | 0.07 | 0.01 | 0.19 | 0.03 | 0.30 | 0.04 |
| 22 | c12 | 0.06 | 0.01 | 0.16 | 0.03 | 0.31 | 0.04 |
| 23 | c18 | 0.07 | 0.02 | 0.17 | 0.03 | 0.31 | 0.04 |
| 24 | c24 | 0.07 | 0.02 | 0.18 | 0.02 | 0.30 | 0.04 |
| 25 | c30 | 0.08 | 0.02 | 0.13 | 0.03 | 0.31 | 0.04 |

Table 5.3: DWT mother (approximation) and father (scale) wavelet coefficients, LDA classification, 90 samples, error rate and SE.

(a) Haar (b) Faber & Schauder

Figure 5.1: Two not too obvious atoms from the introductory chapter.

scenario of the previous chapter is anticipated, namely:

(i) both time and scale (frequency) localisation

(ii) orthogonality

(iii) decorrelating power

(iv) completeness of representation

(v) well behaved at singularities/discontinuities

(vi) efficient implementation

Conceptually, atomic decomposition can be described as the projection

$$\vec{\alpha} = \Phi^{-1}\vec{s} \tag{5.2.1}$$

where $\Phi$ is a matrix containing columns of potential atoms (discrete waveforms), often called a dictionary. $\vec{s}$ is the function (signal) under consideration, and $\vec{\alpha}$ are the desired coefficients. The reconstruction is ensured as:

$$\vec{s} = \Phi\vec{\alpha} \tag{5.2.2}$$

There are two things that should be noted with the statements above. If the dictionary above for instance contains all the sines and cosines of the Fourier transform, the decomposition (5.2.1) would become prohibitive expensive. Explicit analysis (decomposition) and synthesis (reconstruction) formulae exist. These formulae are for most dictionaries faster than (5.2.1) and (5.2.2). The Fourier transform can be done in $O(n\,log\,n)$, while the wavelet transform can be done in $O(n)$ time.

- Secondly, the conceptual formulae (5.2.1-5.2.2), as most explicit approaches, only consider univariate data. By nature the classification data I will investigate are multivariate. The conceptual equations can be extended with matrices, but certain expensive constrains would apply. Alternatively the multivariate data could be

handled in the univariate sense for each observation, and techniques from data fusion used before any classification.

The coefficients found in 5.2.1 on the previous page are not necessary unique[2] or the best coefficients. There are three central approaches to handling this.

### 5.2.1 The method of frames

The theoretical foundation of the method of frames is given in Daubechies (1988). The method of frames is also integrated into the standard text book Daubechies (1992).

For the purposes of this discussion the method of frames gives the coefficients $\vec{\alpha}$ satisfying:

$$\min\|\vec{\alpha}\|_2 \qquad \text{subject to} \qquad \Phi\vec{\alpha} = \vec{s} \tag{5.2.3}$$

The coefficients found by this method, are the average of all $\vec{\alpha}^*$ satisfying $\Phi\vec{\alpha}^* = \vec{s}$, and are thus unique. The averaging involved in this method spread the magnitude (energy) over more coefficients than necessary.

In mitigating the curse of dimensionality for classification, a sparse representation is desired, where as few as possible coefficients should represent as much as possible discriminating information. This renders the method of frames unsuitable.

### 5.2.2 Basis pursuit

Basis pursuit is detailed in Chen (1995) and Chen et al. (2001). At first, the method looks very much like the method of frames. The coefficients are chosen as:

$$\min\|\vec{\alpha}\|_1 \qquad \text{subject to} \qquad \Phi\vec{\alpha} = \vec{s} \tag{5.2.4}$$

The only difference from the method of frames, being the choice of norm. The $\ell^1$ norm

$$\|\vec{x}\|_1 = \sum_i |x_i| \tag{5.2.5}$$

is known as the Manhattan distance or the taxicab norm. This is illustrated in figure 5.2 on the following page. Notice how the $\ell^2$ norm of the method of frames implies a unique route, while several routes are available at the same Manhattan distance. The averaging effect of the method of frames is avoided and more choices are available.

Solving equation 5.2.4, is however much harder than solving for equation 5.2.3 of the method of frames. The present equation conveys a convex optimisation

[2] $\Psi$ could be overcomplete, i.e. the length of the column vectors are longer than the length of $\vec{s}$.

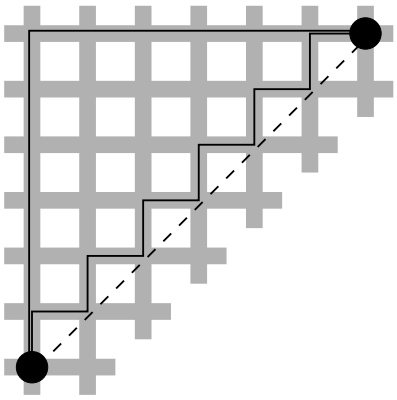

Figure 5.2: Consider a map of city-blocks. The two solid lines have the same Manhattan distance (taxicab norm), while the dashed line is the Euclidean distance.

problem with inequality constraints. This can best be handled by linear programming.

### 5.2.3 Matching pursuit

Matching pursuit was introduced in Mallat & Zhang (1993), and aims at being the projection pursuit of atomic decomposition. Projection pursuit would find interesting projections by looking after deviations from normality. Matching pursuit finds the succession of atoms that best matches the data, and its subsequent residuals.

Matching pursuit is performed iteratively:

(i) Let the initial approximation be $\vec{s}^{(0)} = \vec{0}$. Let the corresponding residual be the original data $\vec{r}^{(0)} = \vec{s}$.

(ii) Given a dictionary of atoms $\Phi = [\vec{\phi}_1, \ldots, \vec{\phi}_p]$, rank their performance against the current residual as $d_{kp} = | < \vec{r}^{(k-1)}, \vec{\phi}_p > |$ (nearly correlation).

(iii) Chose the atom $\vec{\phi}_p$, with the highest $d_{kp}$. Let $\vec{\alpha}_k =< \vec{r}^{(k-1)}, \vec{\phi}_p >$, and set

$$\vec{s}^{(k)} = \vec{s}^{(k-1)} + \vec{\alpha}_k \vec{\phi}_p \tag{5.2.6a}$$

$$\vec{r}^{(k)} = \vec{s} - \vec{s}^{(k)} \tag{5.2.6b}$$

(iv) Keep track of the coefficients $\{\vec{\alpha}_k\}$ and repeat (ii)-(iii) until the residual $\vec{r}^{(k)}$ is sufficient small.

The dictionary $\Phi$ can be large. How interesting the projections found are, depend on putting the right atoms in the dictionary. If the dictionary consists of orthogonal atoms, the residual can be made arbitrary small.

$$\lim_{k\to\infty} \vec{r}^{(k)} = \vec{0} \tag{5.2.7}$$

With this method one should reasonably expect that the representation one arrives at, is as sparse as the dictionary allows. This is not always the truth. The algorithm outlined above is shortsighted in that it iteratively chooses atoms at each step without regard to the final decomposition. Under most circumstances matching pursuit should perform well. In degenerative examples, and in some applications where the dictionary is plainly wrong, very complex solutions appear.

### 5.2.4 Discussion

The three methods considered above, are all reasonable methods with both strength and weaknesses. They are compared in Chen et al. (2001).

The most eminent shortcoming of these methods, is that they are univariate. Brown & Costen (2005) and Zhang et al. (2004) attempt multivariate approaches to basis pursuit, while Hyvärinen et al. (2001) slightly mention the method of frames for (2D) image representation.

I would like to retain the qualities vested in these methods, but have to work with multivariate data. -However, the optimisation involved in the univariate case is already huge, and an extension would further surmount this.

If limited to orthogonal atoms, all methods are related to the best orthogonal basis algorithm of Coifman & Wickerhauser (1992), in some way. The matching pursuit algorithm would benefit from having the Coifman-Wickerhauser basis as its initial value. Chen (1995) observes that basis pursuit and the best basis algorithm, with slight modification, are related. Basis pursuit can in this context be seen as a refinement of the best basis algorithm.

It is generally true for all methods, that coefficients with high magnitude (energy) are important. This is especially true for the coefficients chosen by the best orthogonal basis algorithm. Relative voting will in this case be a reasonable way of ranking and selecting desirable amount of coefficients, before classification.

## 5.3 Best orthogonal basis

Coifman & Wickerhauser (1992) introduce an algorithm for best basis selection. This algorithm aims at selecting the best basis from a tree of orthogonal

decompositions.

In section 2.6.2 the wavelet packet decomposition is described. It is remarked that there are $2^{2^{L-1}}$ ways of recomposing the data (signal) at the L'th level of decomposition. - At first, finding the best basis among such a huge collection of bases seams insurmountable. Fortunately the decomposition has a structure styled like a binary search tree. In this tree each child is a subspace of its parent node (decomposition). With the additive property of a measure, this tree can be searched in $O(n \log_2 n)$ average time and $O(n)$ worst time. $n = 2^{2^{L-1}}$ is the total number of nodes.

In the strict sense, a measure (Bartle (1995)) is a real-valued function $\mu$ defined on $X$ that satisfies:

$$\mu(\emptyset) = 0 \tag{5.3.1a}$$

$$\mu(E) \geq 0 \ \forall E \in X \tag{5.3.1b}$$

$$\mu(\cup E_n) = \sum \mu(E_n) \quad E_n \cap E_m = \emptyset \ n \neq m \tag{5.3.1c}$$

Technically the domain $X$ should be a $\sigma$-algebra, but for all practical purposes it is sufficient that the union of any $E_n \in X$ also belongs to $X$. This holds for the wavelet packet tree, since mentioned, each child node is a subspace of its parent.

### 5.3.1 Entropy

Coifman & Wickerhauser (1992) use Shannon entropy (Shannon (1948))

$$H(\vec{s}) = -\sum \frac{s_i^2}{\vec{s}^t \vec{s}} \log \frac{s_i^2}{\vec{s}^t \vec{s}} \qquad \vec{s} = \text{ signal or data} \tag{5.3.2}$$

as their measure. This measure bears reminiscence to Kullback-Leibler divergence (Kullback & Leibler (1951)).

In information theory, entropy measures information content, or uncertainty of information realisation. E.g. Assume the two vectors

$$(0, 0, 0, 1, 0, 0)^t$$
$$(3, 1, 7, 5, 11, 2)^t$$

generated by two stochastic processes. The stochastic process generating the first vector is more predictive, and should have lower entropy than the last, which seams less predictive.

Notice that variance and entropy are not the same. Entropy is a strict concave function of probability, while variance need not be. This is best illustrated with the realisation of two normal distributions. In figure 5.3 on the next page and

| | $X1$ | $X2$ |
|---|---|---|
| entropy | 6.19 | 6.21 |
| variance | 1.06 | 3.78 |

Table 5.4: Entropy and variance of the realisation of two normal distributions

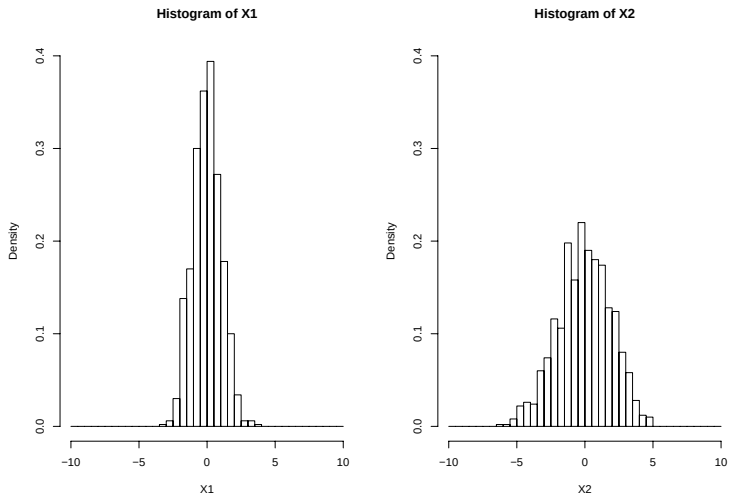


Figure 5.3: Realisation of two normal distributions with the same entropy but different variance, confer table 5.4

table 5.4, $X1$ and $X2$ have the same entropy but different variance. What entropy measure is the ability to represent the true but unknown value (here 0) of the underlying variable. The "flatness" of the distribution has little impact. Variance on the other hand, measures the apparent deviation from the underlying variable.

With this said entropy and variance are related. In Friedman (1987), the varimax (Kaiser (1958)) scheme is used in projection pursuit to find interesting projections by maximising variance. The increase in variance in these projections, would warrant an increase in entropy.

In the application at hand, entropy is more suited than variance, as it will encourage sparse representations.

### 5.3.2 Tree traversal

The Coifman-Wickerhauser algorithm starts with a wavelet packet tree. Starting at the root (top), each node is visited at most once. The nodes' entropy is compared to the sum of the entropy of its children. If the entropy of the parent node is higher, it is marked and its entropy is set to the sum of the entropy of the

children. This is done through successive induction, and the marked nodes are kept as the best basis.

This is illustrated in figure 5.4. In the left half-tree $\{HH\}$ and $\{HG\}$ are kept. In the right half-tree the measure of the $\{G\}$ node is updated, and the node kept. The final chosen basis is $\{HH, HG, G\}$.

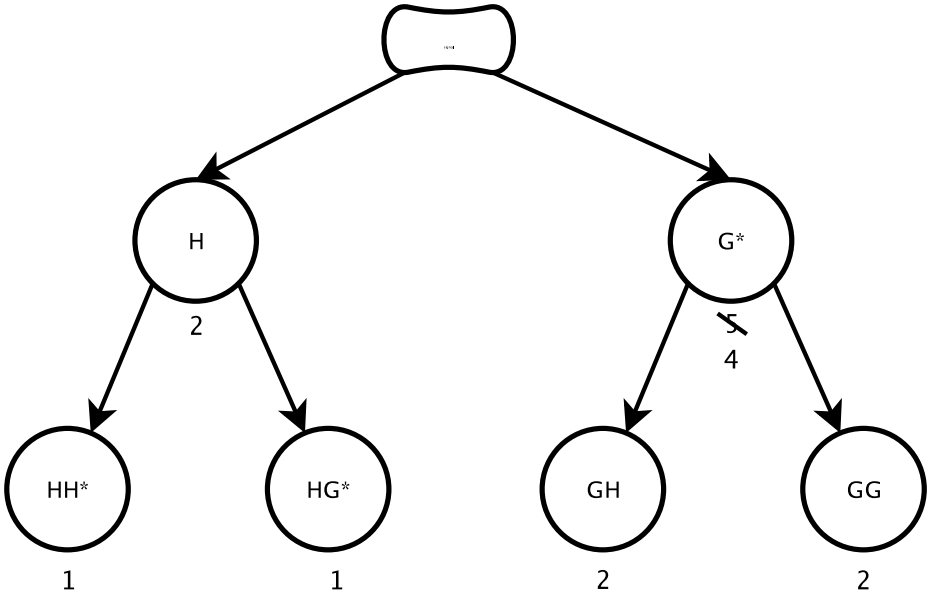


Figure 5.4: The Coifman-Wickerhauser algorithm on a small tree.

### 5.3.3 The multivariate case: ranking and voting

Coifman & Wickerhauser (1992) suggest that when working in multiple dimensions, one could move from a binary tree to a more general tree. The extra dimensions referred to are the (spatial) object dimension. E.g. a three tuple colourspace (RGB), varying over two dimensional image plane. The multivariate model considered in this thesis has a $K$-tupled "colourspace[3]" varying over some spatial dimension, and could be handled in the more general tree paradigm.

This would give a best orthogonal basis representation of the data as a whole, but would increase complexity and break the classification model. The model adopted is a model where $n$ (class) independent samples each consists of $K$-tuples. (Refer to the LDA model in section 3.1). The spatial dimension in the previous example surcomes to the independence criterion.

[3]actually k-wavelet coefficients.

If the Coifman-Wickerhauser algorithm is applied to each of the $n$ samples independently, the situation bears more resemblance to data fusion. Data fusion or sensor fusion is the science of combining data from different sensors, to describe the same object. Here the $K$-tuples stem from different sensors, and may or may not be available. Similarly the Coifman-Wickerhauser algorithm, may or may not make a coefficient or base available. Popular data fusion techniques include *Dempster-Shafer belief* (Dempster & Weisberg (1968)), the *Kalman filter* (Kalman (1960)) and *Bayesian belief networks*. An excellent book on data fusion is Goodman (1997).

Wavelets and the classification task make the "data fusion" conceptually simple. The Coifman-Wickerhauser algorithm provides a sparse and interesting representation for each of the $n$ samples. If one keeps only a few of those bases present in most of the samples, one can reconstruct the samples and see that they emphasise on regions deemed interesting in the previous chapter. These regions are the boundaries between ridges and valleys in the mineral example of figure 4.3 on page 55, which also correspond to the regions of high between class differences in figure 4.5 on page 57. Two approaches to exploit this "coincidence" are devised.

Parameter selection (which wavelet and # coefficients) is done as described in section 3.4. In figure 5.5 on the following page the conceptual system is shown. The "reduce" step now consists of the Coifman-Wickerhauser algorithm and the rank and vote methods described below.

### Method I: Above mean

The Coifman-Wickerhauser algorithm is used on the $n$ samples independently, and each of the selected nodes (bases) is given a vote. Among the bases with a none zero vote, those with an above mean vote are kept.

The basis selected has many coefficients. These coefficients are ranked after their magnitude (energy), for each sample. The $k^*$ coefficients ranked overall highest are kept. It is assumed that the classifier performance will show a variance-bias tradeoff, similar to the PCA "smile" of figure 4.9 on page 67.

### Method II: On mean

To preempt the results; -The method above performs well on all but the Fontainebleau dataset. I believe that this is due to the relative higher within class variance of this dataset.

Coifman & Wickerhauser (1992) state explicitly that averaging would increase the information cost (i.e. entropy). Consequently, I still believe that the higher within

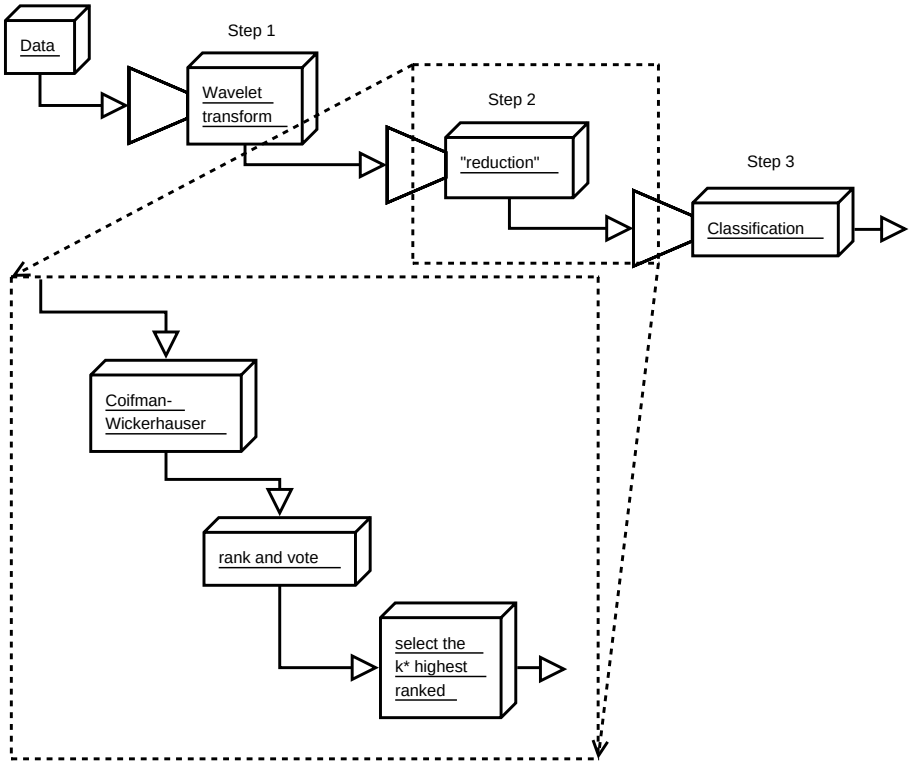


Figure 5.5: Details of the "reduce" step.

class variance of the Fontainebleau dataset warrants some averaging. Chen (1995) remarks that noise reduction can be embedded into the Coifman-Wickerhauser algorithm.

In this method, I let the algorithm select a basis on the mean of the data (i.e each spectral band is averaged over the samples). Then for each sample, each coefficient is ranked as in the previous method.

### 5.3.4 Results

The experiment is done as outlined in section 3.4. I mention that I will provoke *the curse of dimensionality* by using few samples in training. I would like to remind the reader that the classifier has only been trained with 30 (Fontainbleau), 50 (National Mall) or 90 (Pavia) samples in the results given here.

### Method I: Above mean

In table 5.5, 5.6 and 5.7, the results for the ten best parameter combinations on the three datasets are shown.

On both the Pavia and Fountainebleau datasets, the present method performs better than PCA (in table 4.5 on page 68). Performance on the National Mall dataset is better, but not that much. In figure 5.6 on page 85 the variance-bias tradeoff for this dataset is shown. It seems that the right spot has been chosen, so this is not the problem.

In the previous section the *on mean* method was motivated on possible within class variance problems of the Fontainebleau dataset. The variance-bias tradeoff of the present and proposed methods is compared in figure 5.7 on page 85. The desired theoretical smile is clearly more pronounced in the proposed *on mean* method.

| | | | | 95% CI | |
|---|---|---|---|---|---|
| | level | # comp. | error | low | high |
| 1 d8 | 5 | 19 | 0.052 | 0.030 | 0.084 |
| 2 d8 | 5 | 17 | 0.052 | 0.030 | 0.083 |
| 3 mb8 | 5 | 17 | 0.052 | 0.033 | 0.080 |
| 4 d8 | 5 | 18 | 0.052 | 0.032 | 0.083 |
| 5 d8 | 5 | 20 | 0.052 | 0.029 | 0.084 |
| 6 mb8 | 5 | 18 | 0.054 | 0.035 | 0.082 |
| 7 mb8 | 5 | 16 | 0.054 | 0.033 | 0.083 |
| 8 d16 | 5 | 19 | 0.055 | 0.034 | 0.082 |
| 9 d16 | 5 | 20 | 0.056 | 0.035 | 0.083 |
| 10 haar | 5 | 19 | 0.056 | 0.035 | 0.086 |

Table 5.5: Method I: above mean, Pavia dataset.

### Method II: On mean

In table 5.8, 5.9 and 5.10, the results for the ten best parameter combinations on the three datasets are shown. Compared to the PCA results in table 4.5 on page 68, the present method performs better all over. The relative boost in performance on the Fontainebleau dataset over the *above mean* method, seams to indicate some truth in my hypothesis. The increase in performance on the other datasets is not as firm as to draw any conclusion on any additional effects.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 haar | 7 | 11 | 0.103 | 0.063 | 0.175 |
| 2 haar | 7 | 12 | 0.104 | 0.061 | 0.172 |
| 3 fk6 | 7 | 9 | 0.105 | 0.069 | 0.168 |
| 4 haar | 7 | 13 | 0.106 | 0.064 | 0.174 |
| 5 fk6 | 7 | 13 | 0.108 | 0.070 | 0.166 |
| 6 d8 | 7 | 13 | 0.108 | 0.071 | 0.173 |
| 7 la16 | 4 | 5 | 0.108 | 0.071 | 0.176 |
| 8 la16 | 7 | 10 | 0.108 | 0.068 | 0.178 |
| 9 fk6 | 7 | 11 | 0.108 | 0.067 | 0.173 |
| 10 la16 | 3 | 5 | 0.108 | 0.070 | 0.176 |

Table 5.6: Method I: above mean, National Mall dataset.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 haar | 5 | 5 | 0.292 | 0.237 | 0.386 |
| 2 haar | 5 | 7 | 0.292 | 0.235 | 0.385 |
| 3 mb8 | 5 | 5 | 0.292 | 0.230 | 0.384 |
| 4 haar | 5 | 6 | 0.293 | 0.239 | 0.391 |
| 5 haar | 5 | 3 | 0.295 | 0.247 | 0.373 |
| 6 haar | 5 | 4 | 0.298 | 0.248 | 0.384 |
| 7 haar | 5 | 9 | 0.304 | 0.246 | 0.395 |
| 8 la10 | 4 | 3 | 0.304 | 0.259 | 0.377 |
| 9 haar | 5 | 10 | 0.305 | 0.245 | 0.401 |
| 10 haar | 5 | 8 | 0.306 | 0.244 | 0.413 |

Table 5.7: Method I: above mean, Fontainebleau dataset.

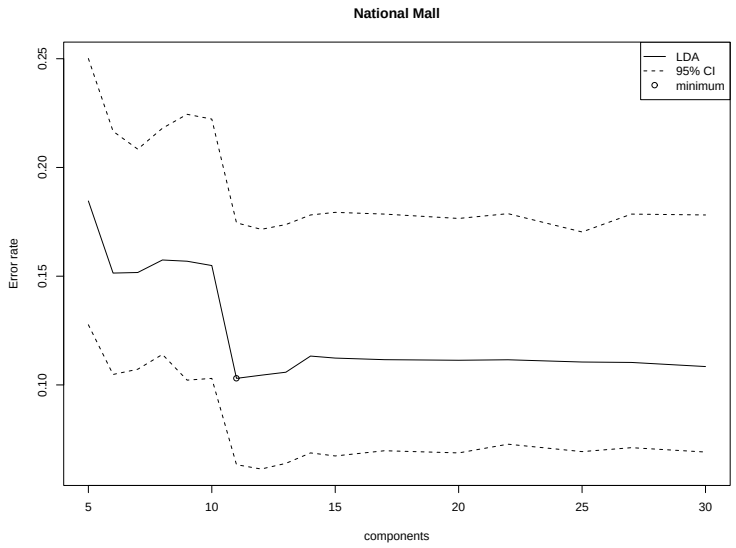


Figure 5.6: The variance-bias tradeoff smile of the high dimensional National Mall dataset.

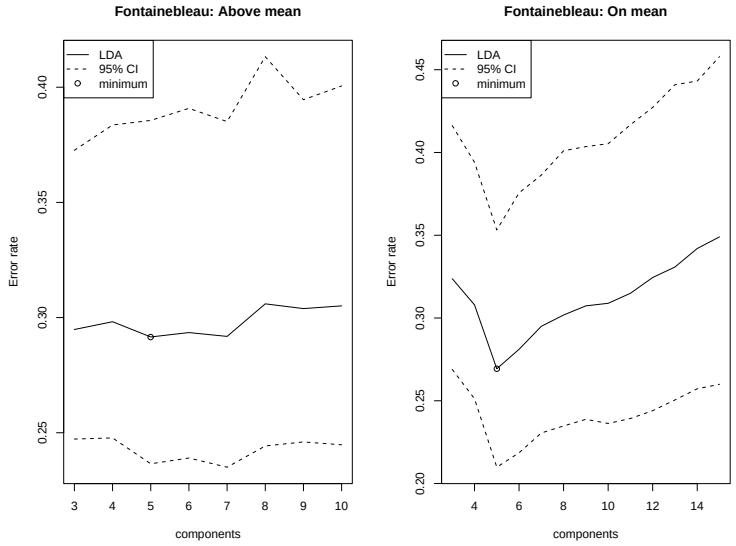


Figure 5.7: Improvement with the on mean method, both in error rate and variance-bias tradeoff smile. *Note: different scales.*

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 d20 | 7 | 23 | 0.051 | 0.029 | 0.083 |
| 2 d20 | 7 | 25 | 0.051 | 0.030 | 0.083 |
| 3 mb8 | 7 | 22 | 0.051 | 0.031 | 0.080 |
| 4 d20 | 7 | 20 | 0.051 | 0.030 | 0.081 |
| 5 fk22 | 7 | 24 | 0.051 | 0.030 | 0.081 |
| 6 d20 | 7 | 24 | 0.051 | 0.030 | 0.081 |
| 7 d20 | 7 | 19 | 0.052 | 0.032 | 0.080 |
| 8 fk22 | 7 | 22 | 0.052 | 0.032 | 0.080 |
| 9 mb8 | 7 | 23 | 0.052 | 0.031 | 0.081 |
| 10 d20 | 7 | 22 | 0.052 | 0.029 | 0.080 |

Table 5.8: Method II: on mean, Pavia dataset.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 fk6 | 7 | 11 | 0.099 | 0.069 | 0.159 |
| 2 fk6 | 7 | 13 | 0.101 | 0.064 | 0.168 |
| 3 fk6 | 7 | 8 | 0.101 | 0.067 | 0.164 |
| 4 fk6 | 7 | 6 | 0.101 | 0.067 | 0.163 |
| 5 fk6 | 7 | 12 | 0.102 | 0.070 | 0.168 |
| 6 haar | 6 | 11 | 0.103 | 0.062 | 0.178 |
| 7 fk6 | 7 | 16 | 0.103 | 0.066 | 0.163 |
| 8 fk6 | 7 | 15 | 0.104 | 0.067 | 0.164 |
| 9 haar | 6 | 25 | 0.104 | 0.067 | 0.172 |
| 10 haar | 6 | 10 | 0.105 | 0.061 | 0.174 |

Table 5.9: Method II: on mean, National Mall dataset.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 d16 | 7 | 5 | 0.269 | 0.210 | 0.353 |
| 2 haar | 5 | 4 | 0.279 | 0.235 | 0.357 |
| 3 d16 | 7 | 6 | 0.281 | 0.219 | 0.376 |
| 4 mb16 | 7 | 7 | 0.288 | 0.223 | 0.375 |
| 5 haar | 6 | 4 | 0.289 | 0.240 | 0.383 |
| 6 d8 | 6 | 5 | 0.291 | 0.231 | 0.393 |
| 7 haar | 6 | 11 | 0.292 | 0.230 | 0.378 |
| 8 mb16 | 7 | 8 | 0.292 | 0.224 | 0.392 |
| 9 haar | 5 | 10 | 0.292 | 0.230 | 0.384 |
| 10 haar | 6 | 6 | 0.293 | 0.239 | 0.380 |

Table 5.10: Method II: on mean, Fontainebleau dataset.

## 5.4 Modifications to the best orthogonal basis algorithm

The Coifman-Wickerhauser algorithm is possible to modify in several ways. Beside the multidimensional modification discussed on page 80, the obvious candidate for modification is the measure.

Kreutz-Delgado & Rao (1998) discuss several measures and proposes Schur-concavity as a good property for measures.

It should also be mentioned that Pesquet et al. (1996a) and Pesquet et al. (1996b) consider the Coifman-Wickerhauser algorithm in the Bayesian framework, and relate the *minimum description length* (MDL) metric to the basis selection problem.

### 5.4.1 Earth movers distance

I will replace the measure in Coifman & Wickerhauser (1992) with the Earth movers distance (EMD) of Rubner et al. (2000). This is perhaps to stretch the notion of a measure a bit. The earth movers distance is mostly used to query databases for similar images. In a separate development Ancona et al. (2002) have used wavelets in this application.

EMD is an evolution of the grey-scale measure in Peleg et al. (1989). This is best illustrated by the allusion in the method's name. Suppose that two functions, or rather their graphs are illustrated by heaps of earth and empty pits. How much work must be performed on the first function to make it look like the second, see

figure 5.8.

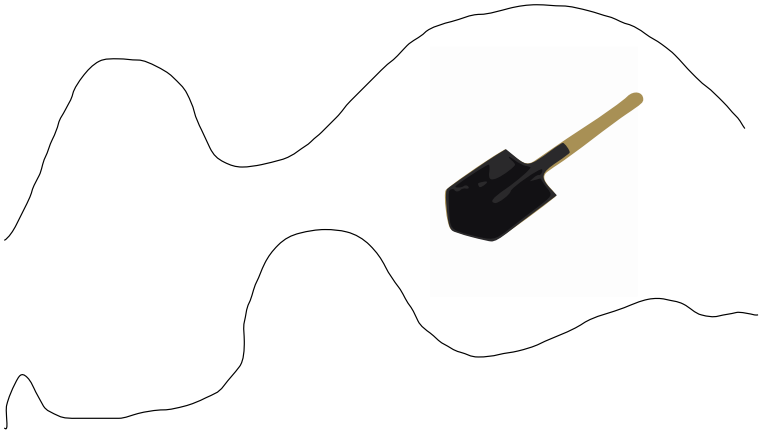

Figure 5.8: Earth movers distance: Measure the work required to fill the pits and level the heaps until resemblance between the two functions.

Rubner et al. (2000) relate this problem to the *transportation problem* of Operational Research (OR) and linear programming (LP). Let the first function be represented as discrete sources (suppliers) and the second function as sinks (consumers). With constraints to limit false moves, the minimum work required to shift the "earth" can be found if a ground distance is defined. If some earth is to be moved from coefficient $p_i$ to $q_j$, I will let the ground distance be $d_{ij} = |i - j|$.

The problem is to find a flow $\{f_{ij}\}$ to minimise:

$$\mathrm{EMD}(P, Q) = W(P, Q, F) = \sum_i \sum_j d_{ij} f_{ij} \tag{5.4.1}$$

where $P$, $Q$ and $F$ are sets of sources, sinks and flow respectively. This problem becomes a linear programming (LP) problem by the constraints. To ensure flow only in one direction: $f_{ij} \geq 0$ . The flow should be limited to what is available and what one can receive:

$$\sum_j f_{ij} \geq |P_i| \tag{5.4.2a}$$

$$\sum_i f_{ij} \geq |Q_j| \tag{5.4.2b}$$

and maximal possible flow should be carried out:

$$\sum_i \sum_j f_{ij} = \min\left(\sum_i |P_i|, \sum_j |Q_j|\right) \tag{5.4.3}$$

I will cast this transportation problem as an uncapacitated minimum cost network flow problem, and use the specialised implementation in Patrício et al. (2004).

In the Coifman-Wickerhauser algorithm I will use the EMD to measure the difference between a sample and its reconstruction with the reduced basis.

### 5.4.2 Results

The experiment is done as outlined in section 3.4. I mention that I will provoke *the curse of dimensionality* by using few samples in training. I would like to remind the reader that the classifier has only been trained with 30 (Fontainbleau), 50 (National Mall) or 90 (Pavia) samples in the results given here.

#### Method I: Above mean

In table 5.11, 5.12 and 5.13, the results for the ten best parameter combinations on the three datasets are shown. The performance is in the same league as the *above mean* method, when used with entropy as a measure. It should be noted that there is less variability among the performance of the ten best combinations. On the Fontainebleau dataset the *on mean* method with entropy still performs the best.

| | | | | 95% CI | |
|---|---|---|---|---|---|
| | level | # comp. | error | low | high |
| 1 d16 | 7 | 24 | 0.052 | 0.031 | 0.080 |
| 2 mb16 | 5 | 23 | 0.052 | 0.032 | 0.083 |
| 3 mb16 | 7 | 23 | 0.052 | 0.031 | 0.077 |
| 4 d16 | 7 | 26 | 0.053 | 0.032 | 0.083 |
| 5 mb16 | 7 | 24 | 0.053 | 0.032 | 0.081 |
| 6 bs3.1 | 6 | 28 | 0.053 | 0.033 | 0.077 |
| 7 la16 | 7 | 11 | 0.053 | 0.033 | 0.084 |
| 8 mb16 | 7 | 27 | 0.053 | 0.031 | 0.081 |
| 9 mb16 | 7 | 29 | 0.053 | 0.033 | 0.083 |
| 10 d16 | 7 | 28 | 0.053 | 0.032 | 0.082 |

Table 5.11: Method I: above mean, Pavia dataset.

#### Method II: On mean

In table 5.14, 5.15 and 5.16, the results for the ten best parameter combinations on the three datasets are shown. The performance is similar to that of the *on mean* method with entropy as the measure. In figure 5.9 on page 94 the *above mean* and *on mean* methods are compared. It might look like the EMD introduces some "noise" in the selection of few coefficients.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 la20 | 6 | 12 | 0.105 | 0.064 | 0.167 |
| 2 la16 | 6 | 11 | 0.105 | 0.063 | 0.177 |
| 3 fk22 | 8 | 8 | 0.105 | 0.065 | 0.177 |
| 4 d6 | 6 | 9 | 0.106 | 0.068 | 0.168 |
| 5 bl14 | 6 | 10 | 0.106 | 0.066 | 0.168 |
| 6 mb24 | 4 | 11 | 0.106 | 0.066 | 0.171 |
| 7 mb24 | 5 | 8 | 0.106 | 0.069 | 0.171 |
| 8 d8 | 6 | 12 | 0.107 | 0.072 | 0.164 |
| 9 la16 | 8 | 12 | 0.107 | 0.069 | 0.174 |
| 10 mb16 | 8 | 8 | 0.107 | 0.067 | 0.169 |

Table 5.12: Method I: above mean, National Mall dataset.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 la20 | 6 | 6 | 0.281 | 0.224 | 0.383 |
| 2 fk8 | 6 | 3 | 0.283 | 0.233 | 0.369 |
| 3 mb8 | 3 | 6 | 0.286 | 0.220 | 0.387 |
| 4 la16 | 6 | 6 | 0.287 | 0.227 | 0.374 |
| 5 la20 | 6 | 7 | 0.287 | 0.224 | 0.384 |
| 6 d16 | 3 | 3 | 0.288 | 0.240 | 0.366 |
| 7 d8 | 5 | 5 | 0.289 | 0.230 | 0.382 |
| 8 bl20 | 3 | 5 | 0.291 | 0.232 | 0.383 |
| 9 fk14 | 5 | 5 | 0.291 | 0.233 | 0.380 |
| 10 d16 | 5 | 3 | 0.291 | 0.236 | 0.373 |

Table 5.13: Method I: above mean, Fontainebleau dataset.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 fk8 | 5 | 24 | 0.049 | 0.029 | 0.076 |
| 2 fk8 | 6 | 25 | 0.049 | 0.029 | 0.077 |
| 3 fk8 | 6 | 27 | 0.049 | 0.030 | 0.073 |
| 4 fk8 | 5 | 22 | 0.049 | 0.029 | 0.074 |
| 5 fk8 | 6 | 26 | 0.050 | 0.029 | 0.077 |
| 6 fk8 | 4 | 30 | 0.050 | 0.031 | 0.078 |
| 7 d6 | 4 | 30 | 0.050 | 0.029 | 0.076 |
| 8 fk8 | 7 | 29 | 0.050 | 0.029 | 0.078 |
| 9 fk6 | 4 | 30 | 0.050 | 0.028 | 0.079 |
| 10 fk8 | 7 | 30 | 0.050 | 0.029 | 0.080 |

Table 5.14: Method II: on mean, Pavia dataset.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 fk22 | 3 | 9 | 0.095 | 0.062 | 0.151 |
| 2 mb16 | 3 | 8 | 0.097 | 0.063 | 0.158 |
| 3 mb16 | 3 | 7 | 0.101 | 0.065 | 0.168 |
| 4 haar | 4 | 5 | 0.101 | 0.059 | 0.173 |
| 5 d4 | 4 | 12 | 0.101 | 0.066 | 0.161 |
| 6 haar | 3 | 10 | 0.101 | 0.066 | 0.163 |
| 7 mb24 | 6 | 11 | 0.102 | 0.064 | 0.163 |
| 8 haar | 3 | 9 | 0.102 | 0.066 | 0.167 |
| 9 bl20 | 3 | 7 | 0.102 | 0.066 | 0.163 |
| 10 fk22 | 3 | 10 | 0.102 | 0.066 | 0.166 |

Table 5.15: Method II: on mean, National Mall dataset.

| | level | # comp. | error | 95% CI low | 95% CI high |
|---|---|---|---|---|---|
| 1 d16 | 7 | 5 | 0.271 | 0.211 | 0.365 |
| 2 d16 | 6 | 6 | 0.281 | 0.217 | 0.375 |
| 3 d16 | 7 | 6 | 0.282 | 0.219 | 0.383 |
| 4 mb16 | 7 | 7 | 0.291 | 0.222 | 0.390 |
| 5 mb16 | 7 | 8 | 0.292 | 0.221 | 0.388 |
| 6 bl20 | 6 | 7 | 0.292 | 0.229 | 0.388 |
| 7 fk8 | 6 | 4 | 0.293 | 0.240 | 0.378 |
| 8 mb16 | 6 | 3 | 0.294 | 0.242 | 0.392 |
| 9 bl20 | 7 | 7 | 0.294 | 0.232 | 0.390 |
| 10 fk8 | 7 | 6 | 0.295 | 0.233 | 0.391 |

Table 5.16: Method II: on mean, Fontainebleau dataset.

## 5.5 Discussion

Three methods of atomic decomposition were discussed in section 5.2. All these can be related to the Coifman-Wickerhauser algorithm either in behaviour or as initiator of optimal start values etc. In Chen (1995), *Basis pursuit* is related to the Coifman-Wickerhauser algorithm, by change of measure. Entropy is replaced by the taxicab norm and the behaviour is the same. This justifies to some degree both the *above mean* and *on mean* methods, as well as the choice of ground distance for the EMD.

### 5.5.1 Validation

In accordance with the discussion in chapter 3, a validation test set has been kept out of the analysis above. It is now brought out to validate the claims made.

In thousand repetitions the same "starved" amount of training samples are drawn from the previously used set to train the classifier with the previous selected parameters. Then the validation test sets are classified. The results can be found in table 5.17 and 5.18. Some of the EMD results are left blank as the EMD implementation used would exhaust the available memory on the extended datasets.

All but the Pavia dataset validate my claims. The reason that the validation fails on the Pavia dataset, is that the priors (see table 4.2 on page 62) differ prominently between the training and validation sets. While *simple random sampling* (SRS) ensures that the same nature is present in both the training and validation sets for the two other datasets, the expert knowledge used on the Pavia

set, comes short. This only stresses the importance of collecting ground truth (training sets) that is really typical for the nature one wants to investigate. This overfitting will be discussed further in section 7.4.

| | | | 95% CI | |
|---|---|---|---|---|
| | | error | low | high |
| Pavia | above mean | 0.67 | 0.53 | 0.78 |
| National Mall | above mean | 0.11 | 0.07 | 0.17 |
| Fontainebleau | above mean | 0.28 | 0.22 | 0.38 |
| Pavia | on mean | 0.79 | 0.61 | 0.91 |
| National Mall | on mean | 0.10 | 0.07 | 0.14 |
| Fontainebleau | on mean | 0.26 | 0.20 | 0.35 |

Table 5.17: Validation

| | | | 95% CI | |
|---|---|---|---|---|
| | | error | low | high |
| Pavia | on mean | 0.45 | 0.27 | 0.69 |
| National Mall | on mean | 0.10 | 0.07 | 0.16 |
| Fontainebleau | on mean | 0.26 | 0.21 | 0.35 |

Table 5.18: Validation EMD

Cochran's Q-test of section 3.3 was applied the classified dataof the ten best methods for each dataset. Q-statistics and p-values with confidence bands can be found in table 5.19. By all reasonable confidence levels the hypothesis that there is a difference between the methods must be rejected. This does not necessarily mean that one should not choose one method over the other.

| | | 95% CI | |
|---|---|---|---|
| | Q | low | high |
| Pavia | 9268.16 | 2058.11 | 11389.09 |
| p-value | 0.00 | 0.00 | 0.00 |
| Fontainebleau | 19325.67 | 15639.82 | 22128.34 |
| p-value | 0.00 | 0.00 | 0.00 |
| National Mall | 25640.18 | 22508.73 | 32114.78 |
| p-value | 0.00 | 0.00 | 0.00 |

Table 5.19: Cochran's Q-test

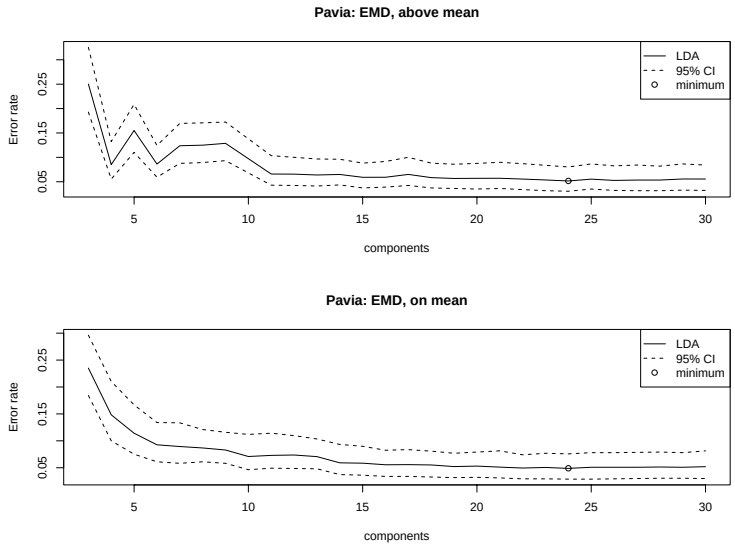


Figure 5.9: Does the EMD introduce more noise, in the selection of coefficients?

# CHAPTER 6

# Denoising

This chapter deals with denoising (i.e. removal of noise) in the wavelet domain. Wavelet coefficients are shrunk towards zero, or even set to zero, according to how much they contribute to noise. Seven methods are given.

Wavelet denoising is presumed to increase classifier performance. This is possible since wavelet denoising makes the wavelet representation more sparse. This applies even when no overt noise is present.

There are two important papers Donoho & Johnstone (1994) and Donoho et al. (1995) that lay the foundation for wavelet denoising. The monograph Jansen (2001) is also an interesting introduction to the topic.

I will first present the denoising framework. Then I give seven practical methods, and examples of their intended use. I finish this chapter by applying the methods in classification, and discuss their performance.

## 6.1 Denoising framework

This section deals with the denoising framework. Assume a function $f(\cdot)$, tainted by some noise $z(\cdot)$, so that one observes

$$y(t) = f(t) + z(t) \quad \text{or even} \quad y(t) = f(t) \times x(t) \tag{6.1.1}$$

The task of denoising is to either *recover* or *estimate* $f(t)$, given $y(t)$. I have deliberately emphasised *recover* and *estimate*, as there have been some distinctions in the literature. The "recover" camp employs methods that rely on particular heuristics in the domain under investigation. The "estimate" faction uses statistical

estimation and decision theory. It should be noted that these communities seam to converge lately. The statistical estimation path will be taken in this chapter.

At first the problem (6.1.1) seams totally unrelated to the best basis selection and atomic decomposition problems of the previous chapter. Section 2.9 of Jansen (2001) relates best basis selection (my section 5.3) to denoising in the following way. The additive noise model of (6.1.1) is given as an atomic decomposition (discrete sense)

$$\begin{aligned} \vec{y} &= \vec{f} + \vec{z} \\ \vec{y} &= \vec{a}\Phi + \vec{R} \end{aligned} \tag{6.1.2}$$

(See equation 5.2.2 on page 74 for details). The coefficients $\vec{a}$ would be found by minimising:

$$\lambda H(\vec{a}) + \tfrac{1}{2}\|\vec{R}\|_2^2 \tag{6.1.3}$$

$H(\cdot)$ is the entropy and $\lambda$ is a smoothing parameter controlling the tradeoff between noise and the Coifman-Wicherhauser best basis selection. $R$ is the residual when $\vec{a}$ is chosen. If the smoothing parameter and the last term were dropped, this would prove similar to the discourse in section 5.3.

Following discussion in Chen (1995), *basis pursuit* (section 5.2.2) fits equation 6.1.3, if the entropy $H(\cdot)$ is replaced by the $\ell^1$ norm. By adjusting its stopping criterion, *Matching pursuit* (section 5.2.3), can also be adapted to the denoising situation. It would typically have to stop some iterations earlier than usual.

### 6.1.1 Minimaxity

In statistical decision theory and in game theory the minimax method or criterion is a method that minimises the maximal loss. The loss (loss function) is the same loss that I slightly touched upon discussing the LDA classifier, see equation 3.1.4 on page 34. The minimax strategy is perhaps counterintuitive. It focuses not on winning the game, but on minimising the chance of loosing. It can be argued that this is the best strategy one can attain in complete and perfect information games. For instance in the game *noughts and crosses (tic-tac-toe)*, a draw can always be forced by this strategy.

Let the risk of an estimator $\delta$, under the loss function $L(\cdot,\cdot)$ over the parameter $\theta \in \Theta$, be the expected loss

$$R(\tilde{\delta}|y,\theta) = \mathrm{E}_\theta\Big(L(\tilde{\delta}|y,\theta)\Big) \tag{6.1.4}$$

where $y$ is the data. The estimator or decision $\tilde{\delta}$ is minimax if it satisfies

$$\sup_\theta R(\theta,\tilde{\delta}) = \inf_\delta \sup_\theta R(\theta,\delta) \tag{6.1.5}$$

The decision done here, will undeniably differ from the *maximum a posteriori (MAP)* decision done in the LDA classifier (section 3.1).

Given particular loss functions, classical statistical methods like *ordinary least squares (OLS) regression*[1] and *maximum likelihood estimation (MLE* can be seen as minimax.

The standard textbook Rice (1995) accommodates a suitable chapter (ch. 15) that explains the ideas discussed here in more detail.

### 6.1.2 Shrinkage

Having established the minimax criterion, I will leave it for a moment, while anchoring another component of the denoising framework.

In *ordinary least squares (OLS) regression*

$$y_i = \alpha + \vec{\beta} x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,1) \text{ iid} \tag{6.1.6}$$

the $\beta$ coefficients are given as

$$\hat{\vec{\beta}} = (X^t X)^{-1} X^t \vec{y} \tag{6.1.7}$$

Under *the curse of dimensionality*, the matrix inversion $(X^t X)^{-1}$ is often very ill-posed.

*Ridge regression* employs Tikhonov regularisation to this. The inversion becomes robust with the regularisation parameter $\Lambda$:

$$\hat{\beta}_{\text{ridge}} = (X^t X + \Lambda I)^{-1} X^t \vec{y} \tag{6.1.8}$$

In the Bayesian view, this is only putting a prior distribution on the coefficients, $\beta \sim N(0, \Lambda^{-1})$. The frequentist view, which I will adopt for the moment, views this as a penalised log-likelihood:

$$\min \quad \|X\vec{\beta} - \vec{y}\|^2 + \Lambda \|\vec{\beta}\|^2 \tag{6.1.9}$$

The singular value decomposition (SVD) of the *ridge regression* becomes

$$\hat{\beta}_{\text{ridge}}(\Lambda) = \sum_i \frac{\lambda_i}{\lambda_i^2 + \Lambda} (\vec{u}_i^t Y) \vec{v}_i \tag{6.1.10}$$

with eigenvalues $0 < \lambda_i \leq \lambda_{i+1} \leq \ldots$.

The penalty $\Lambda$ shrinks the coefficients towards zero. One of the goals of *ridge regression* is to reduce the influence of excessive large coefficients. Clearly as $\lambda_i^2 \to 0$, $\Lambda$ must dominate the denominator.

Hastie et al. (2001) present *ridge regression* in the context used here.

[1] at least for functions linear in the $\theta$'s.

**Different shrinkage thresholds**

The shrinkage above is called soft thresholding. It appears in Bickel (1983) and can be developed independently of my illustration with the *ridge regression*.

Given the threshold $T$, any coefficient is shrunk like:

$$\hat{\beta}^S = \text{sign}(\hat{\beta})\,(|\hat{\beta}| - T)_+ \tag{6.1.11}$$

Here

$$(\kappa)_+ = \begin{cases} \kappa & \kappa \geq 0 \\ 0 & \text{else} \end{cases} \tag{6.1.12}$$

All the coefficients are shrunk towards zero, while some are set to zero.

In contrast the hard threshold

$$\hat{\beta}^H = \hat{\beta}\; I\{|\hat{\beta}| > t\} \tag{6.1.13}$$

set all coefficients below the threshold to zero. This is the threshold I will prefer in my contest with *the curse of dimensionality*. This is more like subset selection in ANOVA (analysis of variance).

In the statistical literature the soft threshold is preferred to the hard threshold on grounds of continuity. I note that middle ground between the two thresholds is covered by *the nonnegative Garrote* of Breiman (1995).

### 6.1.3 Wavelet shrinkage: - oracles and devils

Of the noise models 6.1.1 on page 95 I will consider the additive model in the discrete form

$$y_i = f_i + z_i \tag{6.1.14}$$

where the $z_i$ is independent and identically-distributed $N(0, 1)$. This makes the theory more tractable, but it is not absolutely necessary. In chapter 4 I mention that the sensors under consideration are essentially prone only to additive noise of this kind.

Even the alternative multiplicative noise in equation 6.1.1 on page 95 can be handled. This type of noise appears in coherent sensors, like ultrasound and *Synthetic aperture radar (SAR)*. A SAR example involving wavelet denoising is given in Araújo et al. (2004).

In chapter 2 orthogonality and sparsity are shown as inherent properties of the wavelet transform. A wavelet transform of the model 6.1.14, gives a sparse representation of $f$ (few coefficients), while the orthogonality (especially the

Plancherel/Parseval relation as used in equation 2.5.4 on page 26) spreads the noise energy over all wavelet coefficients.

These properties are imperative for the success of wavelet shrinkage. The high energy coefficients contributed by $f$ can be picked out from the relative spread-out energy of the noise. This heuristic is illustrated in figure 6.1.

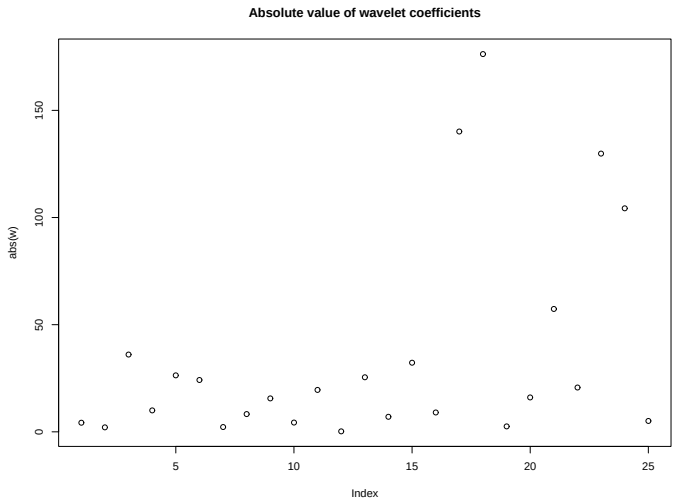


Figure 6.1: Absolute value of the wavelet decomposition of one signature from the Fontainebleau dataset.

### Oracles

Donoho & Johnstone (1994) introduce the concept of oracles into the denoising discussion. Oracles are unattainable functions with side-information that will give the best parameters for a method under consideration.

Let $T(\vec{y}, \delta)$ be the method under consideration. E.g. wavelet thresholding: The parameter $\delta$ is the threshold on the wavelet coefficients, such that:

$$\begin{aligned} T: \vec{(}y) &\rightarrow \vec{f} \\ \vec{y} \rightarrow \text{ Wavelet transform } \{\vec{y}\} &\rightarrow \text{ threshold by } \delta \rightarrow \text{ reconstruct } \rightarrow \hat{\vec{f}} \end{aligned} \tag{6.1.15}$$

The concept of oracles combined by the minimax criterion 6.1.5 on page 96 is said to give an "ideal" risk.

$$R_{\text{ideal}}(T, \vec{f}) = \inf_{\delta} R(T(\vec{y}, \delta), \vec{f}) \tag{6.1.16}$$

No oracle is available, but certain inequalities in Donoho & Johnstone (1994) can lay a bound on the risk

$$R(T, \vec{f}) \leq A \times R_{\text{ideal}}(T, \vec{f}) \tag{6.1.17}$$

The $A$ will be detailed in subsequent sections. The notion of "ideal" will now be addressed.

**Devils**

> ... Drugie dva chudesnye tvoren'ya Vlekli menya volshebnoyu krasoj:
> To byli dvuh besov izobrazhen'ya.
>
> Odin (Del'fijskij idol) lik mladoj - Byl gneven, polon gordosti uzhasnoj,
> I ves' dyshal on siloj nezemnoj. ...
>
> From the poem "V nachale zhizni shkolu pomnyu ya" (1830) by Pushkin

A partial translation found in Poggioli (1951) reads:

> Two wonderful beings fascinated me with their beauty: they were two demon's faces. One, a Delphic idol, was a youthful visage: severe, full of awful pride, he breathed the sense of an unearthly power. The other, an ideal of feminine sem-blance, passionate and deceptive, was a charming genius, false but beautiful.
>
> "In the beginning of life I remember the school" (1830) by Pushkin

The oracle has already been dealt with, now the devilish part will be remarked. By a contumelious coincidence for Oleg Besov, "bes" means devil or demon in Russian, and the "-ov" suffix is the possessive (genitive) form, hence devilish. Besides the humorous part, regularity in Besov spaces, which I deliberately have left out of chapter 2, is taxing.

The "ideal" notion introduced by the oracles is a demand for regularity. The function $f$ to be reconstructed is assumed to belong to certain function spaces, and regularity for these should be proved. The regularity can range from the familiar Lipschitz regularity[2]

$$|f(t) - p_m(t)| \leq K|t - v|^\alpha \, \forall t \in \mathbb{R} \tag{6.1.18}$$

to the more demanding Besov regularity. Besov regularity is especially important for wavelet reconstruction.

I have on purpose avoided this theory. This because in real life, no class can be assumed for the function $(f)$ we want to estimate. With this said, regularity for wide classes can be shown for wavelet reconstruction under thresholding, see Donoho et al. (1995) for details.

On page 96 I mention the convergence of the "recover" and "estimate" camps. In section 6.1.3 a heuristic (the spreading of noise energy over many coefficients) is

[2] i.e. $f(t)$ can be approximated by polynomial of degree $\alpha$

mentioned in a statistical estimation context. This can be taken even further. Breiman (2001b) was briefly mentioned in chapter 3. Breiman (2001b) discusses the differences and similarities between the statistical and data mining communities. Methods leaning towards data mining like CART and MARS (details are in Hastie et al. (2001)) are more spatial adaptive than the more strict statistical smoothing and learning methods. The impurity measure used in CART is a heuristic on the same line as the energy spreading of noise in the wavelet transform. The regularity and oracle inequalities are used in Donoho & Johnstone (1994) to justify the spatial adaptivity of wavelet shrinking.

## 6.2 Practical thresholds

This section is concerned with seven practical thresholds that will exploit the observations of the previous section. I will illustrate this on two datasets. The first is a time series of clock skew between two computers measured over a network. See (a) of figure 6.2 on page 103. The second dataset is a spectral signature from the Fontainebleau dataset, see (c) of figure 6.2 on page 103. I selected data from the Fontainebleau dataset over the two others, as I believe that this is the most noisy and will benefit the most from thresholding. The Daubechies standard 'la8' wavelet will be used in the illustrations.

I would like to remind the reader that we still are in the signal estimation/recovery framework. Classification is not considered until section 6.3.

### 6.2.1 The universal threshold

The universal threshold is an exercise in exploiting *the law of large numbers*. This law gives a bound to the ideal risk in the inequality 6.1.17 on page 99. The universal threshold is given in Donoho et al. (1995), but appears in a substitute role in DeVore & Lucier (1992).

Given

$$y_i = f_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,1) \text{ iid} \tag{6.2.1}$$

Leadbetter et al. (1983) give

$$\lim_{n\to\infty} \Pr\{\max|\varepsilon_i| \geq \sqrt{2\log n}\} = 0 \tag{6.2.2}$$

$n$ is the length of $\vec{y}$.

This means that with high probability, noise will not supersede $\sqrt{2\log n}$ as $n$ grows. The wavelet transform of $\varepsilon_i$ is by the orthogonality property of the transform also standard normal iid.

The universal threshold is

$$t_u = \sqrt{2 \log n}\, \hat{\sigma} \tag{6.2.3}$$

The universal threshold, is so called because it requires little knowledge and is easy to implement. $\sigma$ should be estimated. The standard error

$$\hat{\sigma}_{\mathrm{SE}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{6.2.4}$$

could be used with sufficient high $n$. The *median absolute deviation (MAD)*

$$\hat{\sigma}_{\mathrm{MAD}} = \Xi \, \mathrm{median}_i \left( \, |y_i - \mathrm{median}_j(y_j)| \, \right), \tag{6.2.5}$$

is more robust, and can be used with lower $n$. $\Xi$ is a "guessed" constant used to regulate overestimation. As I hope for a $\sigma = 1$ and the distribution is standard normal $\Pr\{-1 < \varepsilon_i < 1\} \approx 0.6745$ seams like a good choice.

In the *wavelet packet decomposition* I apply the universal threshold individually to each node. This fails for the standard error, but succeeds for the MAD. In figure 6.2 on the following page, most of the clock skew noise is removed, but the Fontainebleau signature seams to be smoothed too much.

I continue by applying the same threshold over all nodes (denoted as Global in the figures). Both the standard error and MAD thresholds are available, see figure 6.3 on page 104.

For the clock skew data, the MAD threshold exhibits less noise with the global MAD, however the global standard error threshold, apparently exhibits no noise. In the Fontainebleau signature, the global standard error threshold smooths too much. The global MAD threshold however, shows more structure (between index 60 and 80) than what is available in the original data.

In the original articles cited, the thresholds are employed using soft thresholding. I will as explained earlier use hard thresholding as this is closer to what I desire in a classification context. The over-smoothing and the noise "blips" seen in the figures, are not instigated by the choice between hard and soft thresholds. They are more a product of $A$ in inequality 6.1.17 on page 99 and a lack of knowledge of the underlying function class.

### 6.2.2 Visu shrink

Visu shrink (or VisuShrink) is introduced in Donoho & Johnstone (1994). It is a simpler method than the universal threshold, mainly aimed at visualisation. The threshold is simply

$$t_v = \sqrt{2 \log n} \tag{6.2.6}$$

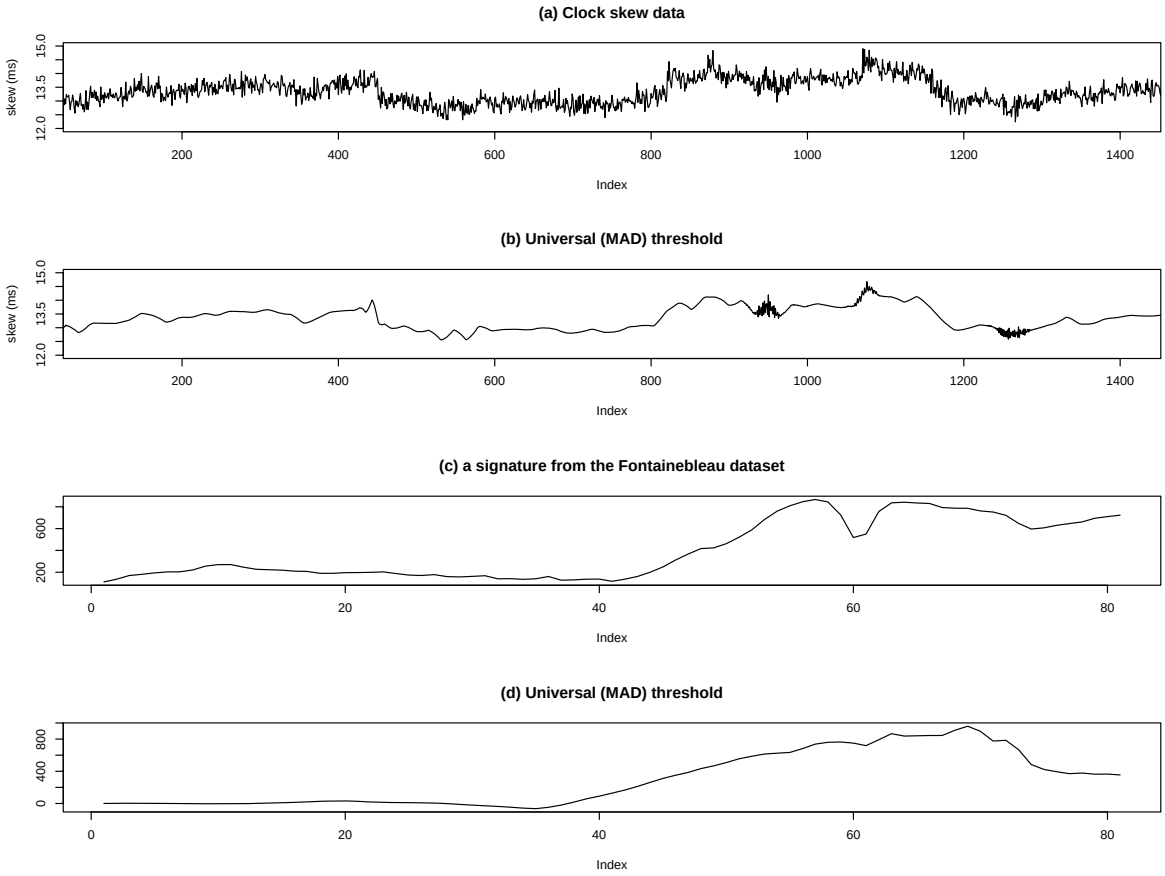


Figure 6.2: Original data and reconstruction with the universal threshold (MAD)

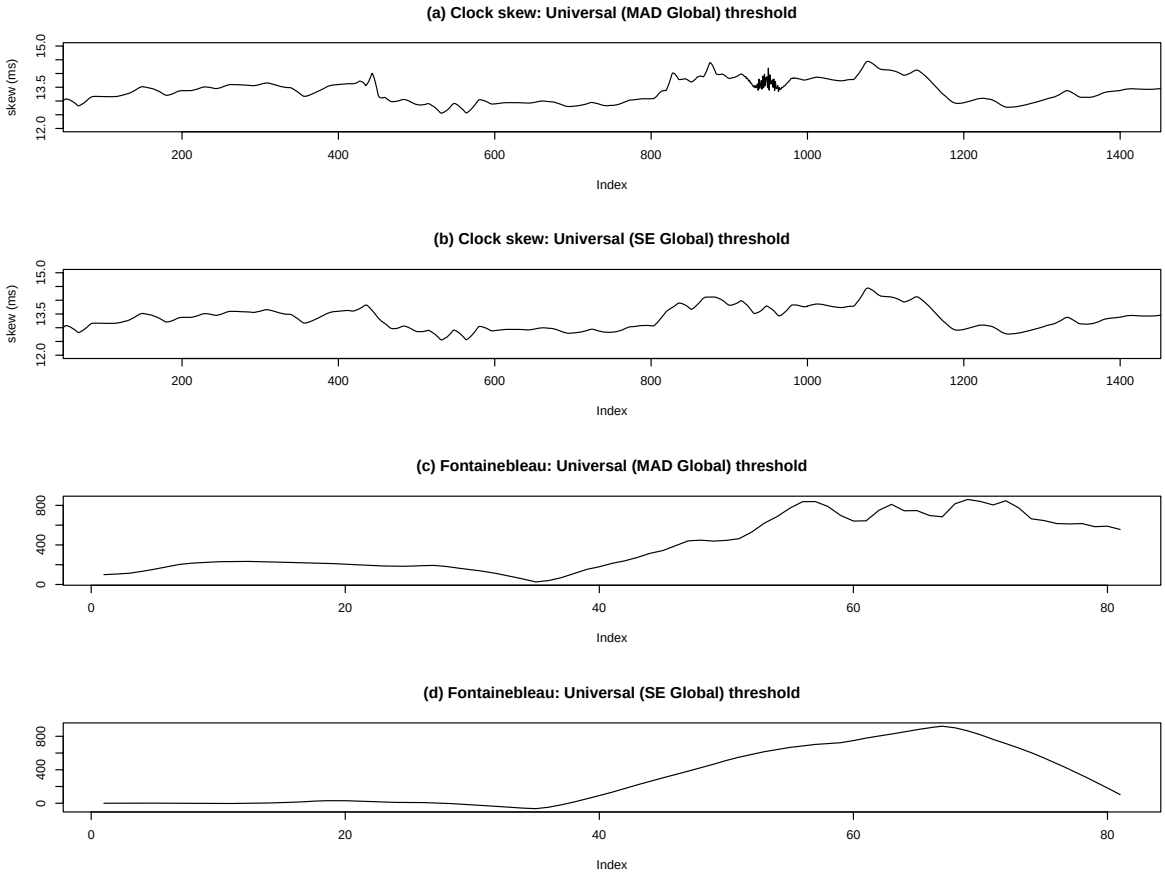


Figure 6.3: Reconstruction with the universal threshold (global MAD and SE)

In the inequality 6.1.17 on page 99, Visu shrink, has an $A$ on the same order as the universal threshold. Visu shrink avoids estimating $\sigma$ and the error that may be contributed in this process.

The Visu shrink is applied globally to all nodes in the *wavelet packet decomposition*, and can be seen in figure 6.4. The results are better than the universal threshold. The clock skew data is apparently noise free, and the Fontainebleau signature shows more structure between index 40 and 80.

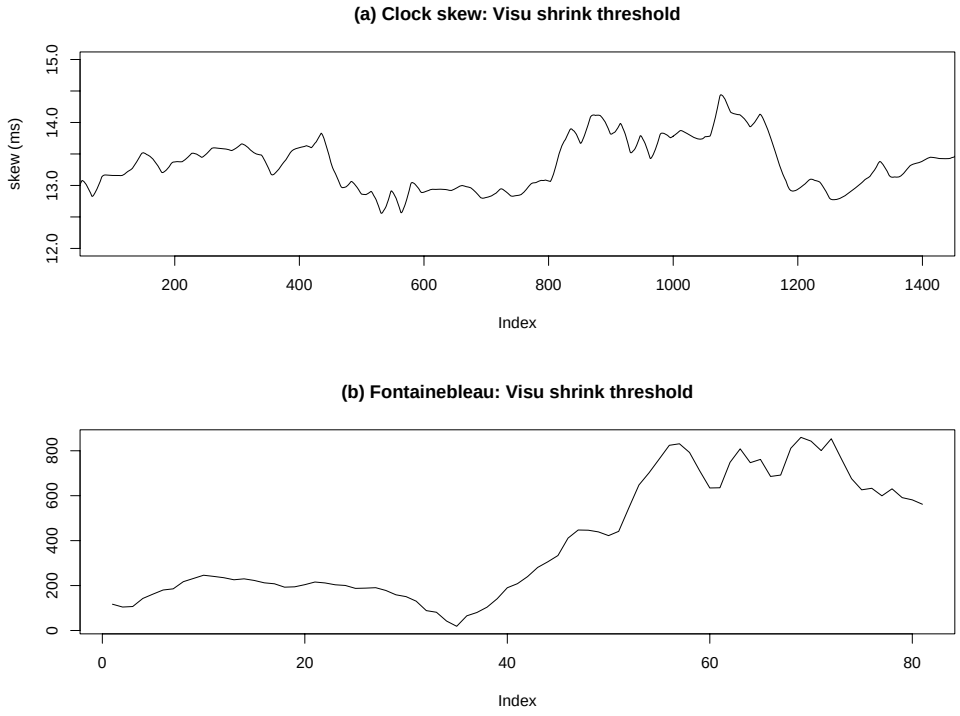


Figure 6.4: Reconstruction with the Visu shrink method. (a) Clock skew data, (b) one signature from the Fontainebleau dataset.

### 6.2.3 Risk shrink

Risk shrink (or RishShrink) introduced in Donoho & Johnstone (1994), is the only method considered in this thesis that really exploit inequality 6.1.17 on page 99.

Donoho & Johnstone (1994) contemplate how thresholding and the wavelet transform should affect coefficients that are noise free. The coefficients belonging to the function that one wishes to estimate, are some number $k < n$, where $k$ is independent of $n$. Visu shrink lets the threshold depend on $n$, this might eliminate too much of the $k$ desired coefficients.

Donoho & Johnstone (1994) assume that the function destined for estimation has a non-zero average, and that the coefficients should as a group stay away from zero. The *mean square error (MSE)* is used as the risk in inequality 6.1.17 on page 99, and by their oracle inequalities Donoho & Johnstone (1994) tabulate some thresholds, see table 6.1.

| $n \leq$ | $t_{\mathrm{RS}} = \lambda_n^*$ |
|---|---|
| 64 | 1.47 |
| 128 | 1.67 |
| 256 | 1.86 |
| 512 | 2.05 |
| 1024 | 2.23 |
| 2048 | 2.41 |
| 4096 | 2.59 |
| 8192 | 2.77 |
| 16384 | 2.95 |
| 32768 | 3.13 |
| 65536 | 3.31 |

Table 6.1: Risk shrink thresholds

The thresholds in table 6.1 are lower than the Visu shrink thresholds. Details of the inequalities and how one comes to this is given in Donoho & Johnstone (1994). Conceptually Risk shrink is only a correction to Visu shrink.

In figure 6.5 on the next page, the thresholds in table 6.1 are applied to the same data as before. The thresholds are used in the hard sense, although they are calculated for soft thresholding. Asymptotically they should be the same. The results are at least as good as those of Visu shrink (figure 6.4 on the previous page). The lower thresholds should let more of the coefficients belonging to the function one wants to estimate, survive.

### 6.2.4 James-Stein shrink

Stein's phenomenon is a paradox in estimation theory that states that when estimating more than two parameters, simultaneous estimation gives lower *mean square error (MSE)*, than when the parameters are estimated separately. Taken to the extreme: one should estimate e.g. child death rates when fitting a three tuple model from planetary orbit, although they might not be related. This can be contemplated as overfitting.

The James-Stein estimator

$$\widehat{\vec{\theta}}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\vec{y}\|^2}\right)\vec{y} \tag{6.2.7}$$

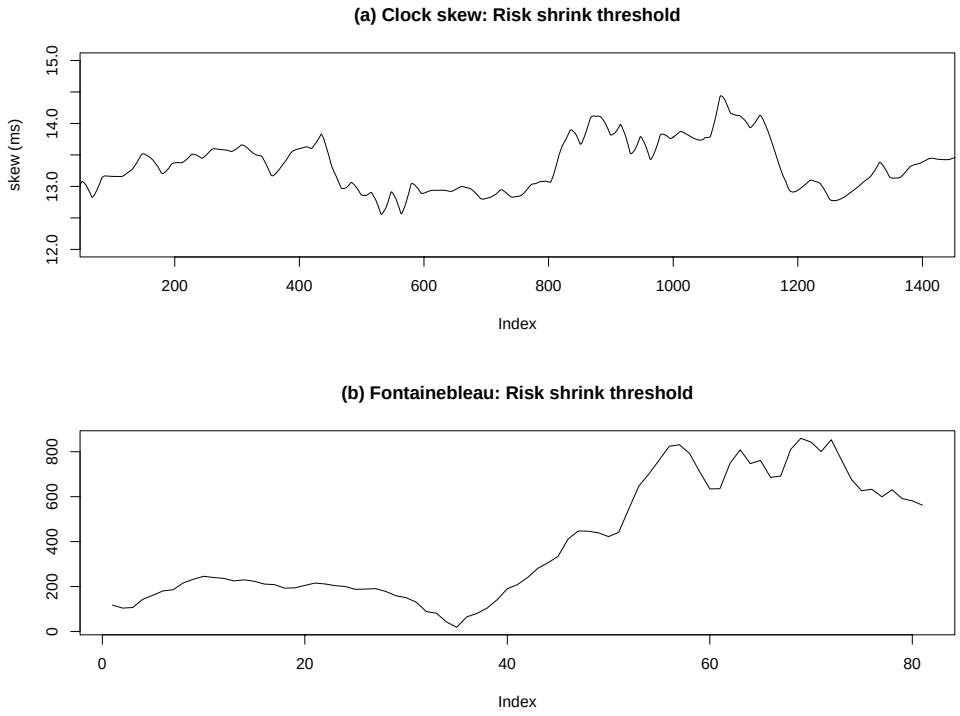


Figure 6.5: Reconstruction with the Risk shrink method. (a) Clock skew data, (b) one signature from the Fontainebleau dataset.

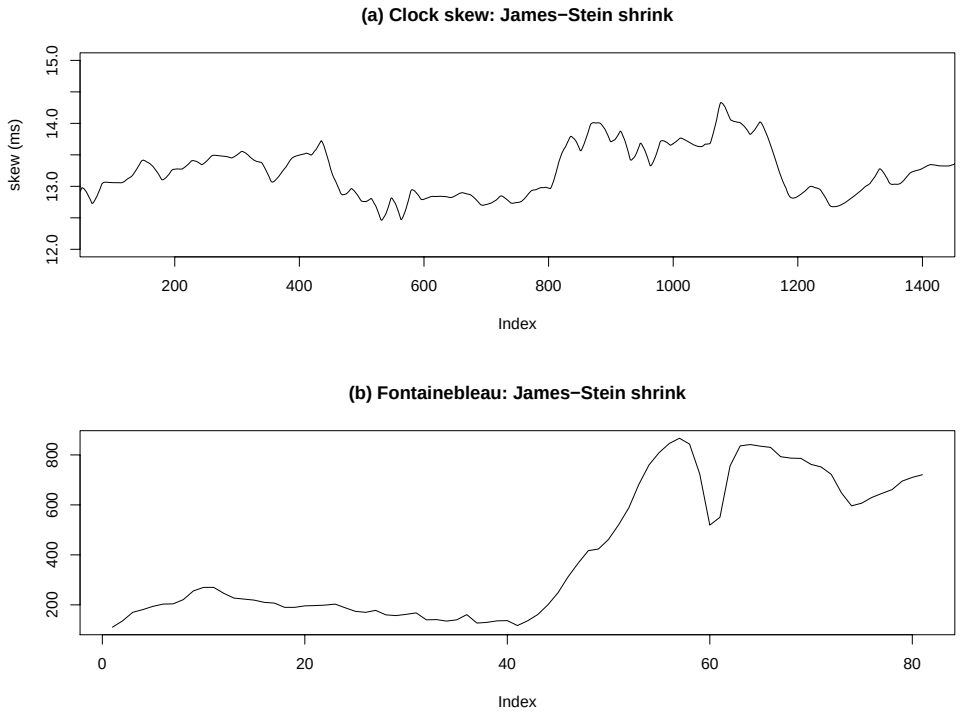


Figure 6.6: Results of shrinkage by the James-Stein shrinker. (a) Clock skew data, (b) one signature from the Fontainebleau dataset.

is an estimator that exploits this paradox. Given $m$ ($> 2$) parameters this estimator would beat e.g. *least squares estimators* in the *mean square error (MSE)* sense.

Donoho & Johnstone (1995) give the James-Stein shrink (or WavJS). This is not a shrinker in the sense of the previous given thresholding methods. Each node of the *wavelet packet decomposition* is shrunk by weighing the wavelet coefficients at a level $\vec{w}_j$ as

$$\vec{w}_j^* = \vec{w}_j \vec{s}_j \tag{6.2.8}$$

where the weights are given as

$$\vec{s}_j = \max\left[\frac{\vec{w}_j^2 - (n-2)}{\vec{w}_j^2}, 0\right] \tag{6.2.9}$$

Results of the James-Stein shrinker are give in figure 6.6. For the clock skew data, the results are similar to the Visu shrink results. For the Fontainebleau signature, less effect can be seen, it is more like it preserves the original signature.

### 6.2.5 SURE shrink

SURE shrink is given in Donoho & Johnstone (1995), but also appears to some degree in the discussion in Donoho et al. (1995) and Donoho & Johnstone (1994).

It is implicit in the inequality 6.1.17 on page 99 that we attempt to reduce risk. SURE stands for Stein's unbiased risk estimate. In Stein (1981), the same Stein as in the previous section, gives an unbiased estimate of loss (risk is the expected loss) when the estimator itself is biased. Given an estimator

$$\tilde{\delta}(\vec{x}) = \vec{x} + \underbrace{f(\vec{x})}_{\text{bias}} \tag{6.2.10}$$

the risk

$$\mathrm{E}\|\tilde{\delta}(\vec{x}) - \delta\|^2 = n + \mathrm{E}\{\|f(\vec{c})\|^2 + 2 \times \text{“some differentiation of f”}\} \tag{6.2.11}$$

is available if $f$ is “weakly” differentiable. Donoho & Johnstone (1995) make the connection to the soft threshold, and give

$$\mathrm{SURE}(t, \vec{w}) = n - 2\#\{i : |w_i| \leq t\} + \sum_{i=1}^{n} \Big( \min(|w|_i, t) \Big)^2 \tag{6.2.12}$$

This is an estimator of *the mean square error (MSE)* risk.

The SURE threshold is

$$t_{\mathrm{SURE}} = \arg \min_{0 \leq t \leq \sqrt{2 \log n}} \mathrm{SURE}(t, \vec{w}) \tag{6.2.13}$$

$\vec{w}$ are the wavelet coefficients at a given level, and the threshold is upper limited by the Visu shrink threshold.

In figure 6.7 on the following page, hard SURE thresholding was applied. The results are similar to that of Visu shrink, although the thresholds were somewhat lower.

### 6.2.6 Hybrid shrink

Donoho & Johnstone (1995) note that the SURE shrink above, works best in dense situations (i.e. few near zero coefficients), and that it may perform worse than Visu shrink in sparse situations.

Donoho & Johnstone (1995) define a measure of sparsity as

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (w_i^2 - 1) \tag{6.2.14}$$

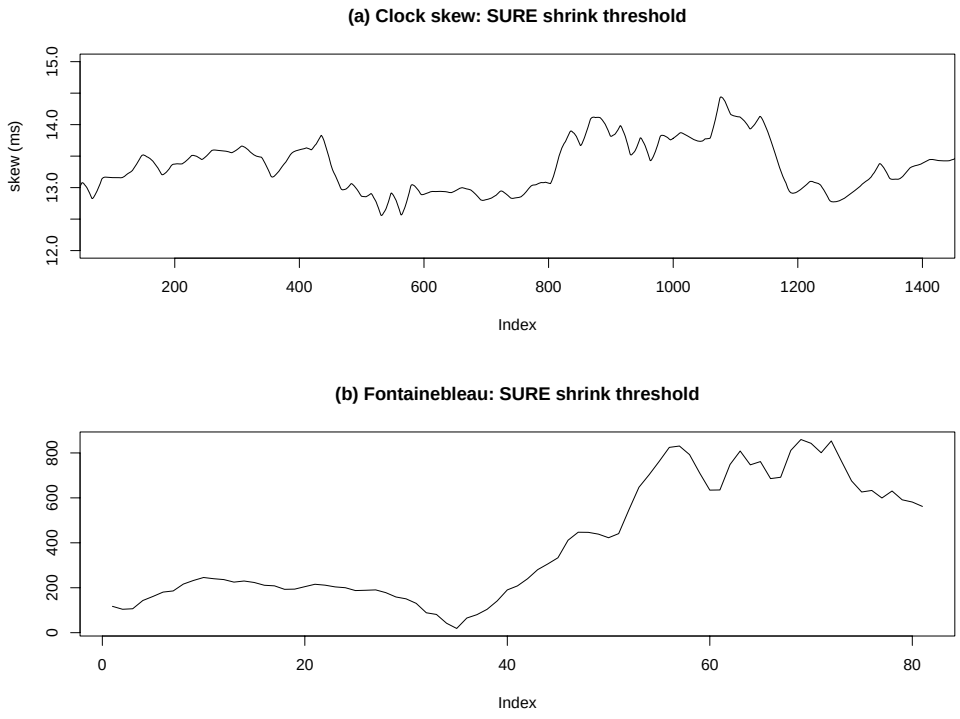


Figure 6.7: Results of SURE shrinkage. (a) Clock skew data, (b) one signature from the Fontainebleau dataset.

and decides whether $\vec{w}$ is sparse or not by the limit

$$\eta_n = \sqrt[3/2]{\log_2 n/\sqrt{n}} \tag{6.2.15}$$

If $s_n^2 \leq \eta_n$ the Visu shrink threshold is used, and else the SURE shrink threshold is used.

For my examples, this reverts to Visu shrink (figure 6.4 on page 105). I guess that my examples are too sparse for SURE shrink.

### 6.2.7 GCV shrink

The thresholds discussed, are nothing more than smoothing parameters. The SURE risk model 6.2.12 on page 109 looks very parametric, perhaps the thresholding can be handled in a more nonparametric way.

Cross-validation as discussed in section 3.2.3, can be used in parameter estimation. This is often done as *leave-one-out cross-validation (LOOCV)*, and is called *ordinary cross-validation (OCV)*. The idea is that when leaving out some samples at a time "noise" and outliers would not influence the estimate of variable unduly. To reiterate from section 3.2.3:

$$\mathrm{OCV}_\lambda = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|k_i|}\sum_{j\in k_i} L[y_j, \hat{f}_\lambda^{\setminus k_i}(\vec{x}_j)] \tag{6.2.16}$$

where $L(\cdot,\cdot)$ is a loss function, $k_i$ is the i'th set when the original set is parted in $k$-parts. $\setminus k_i$ is the original set without the i'th part. $|\cdot|$ is the cardinality operator. $\hat{f}_\lambda(\cdot)$ is the estimated function, with parameter $\lambda$. $k = n$ for LOOCV. Now select

$$\hat{\lambda} = \arg\min_\lambda \mathrm{OCV}_\lambda \tag{6.2.17}$$

the parameter $\lambda$ typically will in our context be the threshold. The estimate $\hat{\lambda}$ is more independent of the noise than earlier. Here the cross-validation minimise reconstruction error, while in section 3.2.3 the classification error was minimised.

*Ordinary cross-validation (OCV)* is not without drawbacks. The primary concern is that the same property that makes $\hat{\lambda}$ resilient to noise, also makes it underestimate noise. For many classes of $\hat{f}$, cross-validation favours high-frequency noise, while the low-frequency noise is essentially removed. This is also the case for wavelet shrinkage. In the suboptimal thresholding in (b) of figure 6.2 on page 103 it is the high-frequency noise that remains. This suggests that cross-validation also has to face this prospectus when investigating a threshold $\lambda$ at this level. A more practical deficiency of cross-validation is that for $k = n$, which is the best choice for $k$, the procedure needs to re-smooth $\hat{f}_\lambda$ $n$ times for each $\lambda$.

Craven & Wahba (1979) try to address this by introducing *generalised cross-validation* for smoothing splines. This is done for wavelet thresholding in Jansen et al. (1997), and in more detail in Jansen (2001). Jansen (2001) makes certain assumptions about what effect equation 6.2.16 on the preceding page, has on $\hat{f}$ and comes to the same formula as Craven & Wahba (1979). This is then transformed into the wavelet domain:

$$\mathrm{GCV}(\lambda) = \frac{\frac{1}{n}\|\vec{w} - \vec{w}_\lambda\|^2}{\left(\frac{n_0(\lambda)}{n}\right)^2} \tag{6.2.18}$$

where the burdensome sum in equation 6.2.16 on the previous page dematerialise. $\vec{w}_\lambda$ are the thresholded coefficients and $n_0(\lambda)$ are the number coefficients set to zero.

Discounting division by zero, (6.2.18) will be convex as the norm is convex (triangle inequality). Figure 6.8 shows that the GCV is sharply convex on a subset of the Pavia dataset. It also shows that it is sufficient smooth to let a Newton-Raphson type optimisation algorithm select

$$t_{\mathrm{GCV}} = \arg\min_{\lambda} \mathrm{GCV}(\lambda) \tag{6.2.19}$$

rather than a brute force search.

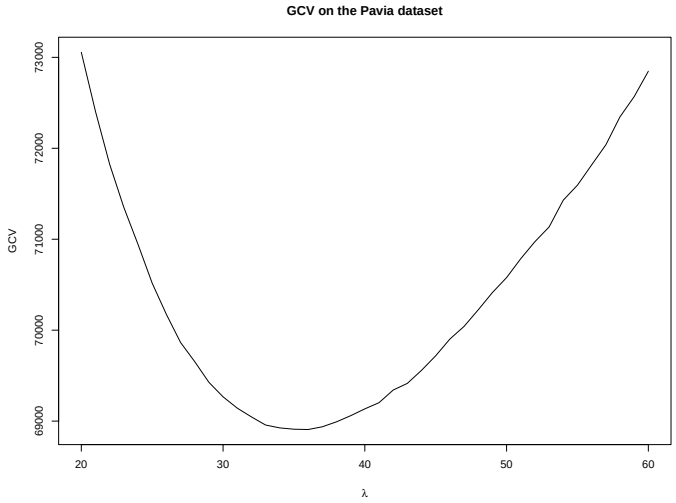


Figure 6.8: Selecting a GCV threshold

In figure 6.9 on the next page hard GCV thresholding was applied to both the clock skew data and the signature from the Fontainebleau dataset. The results are comparable with the SURE and Visu shrink results, but at a fraction of the computational cost.

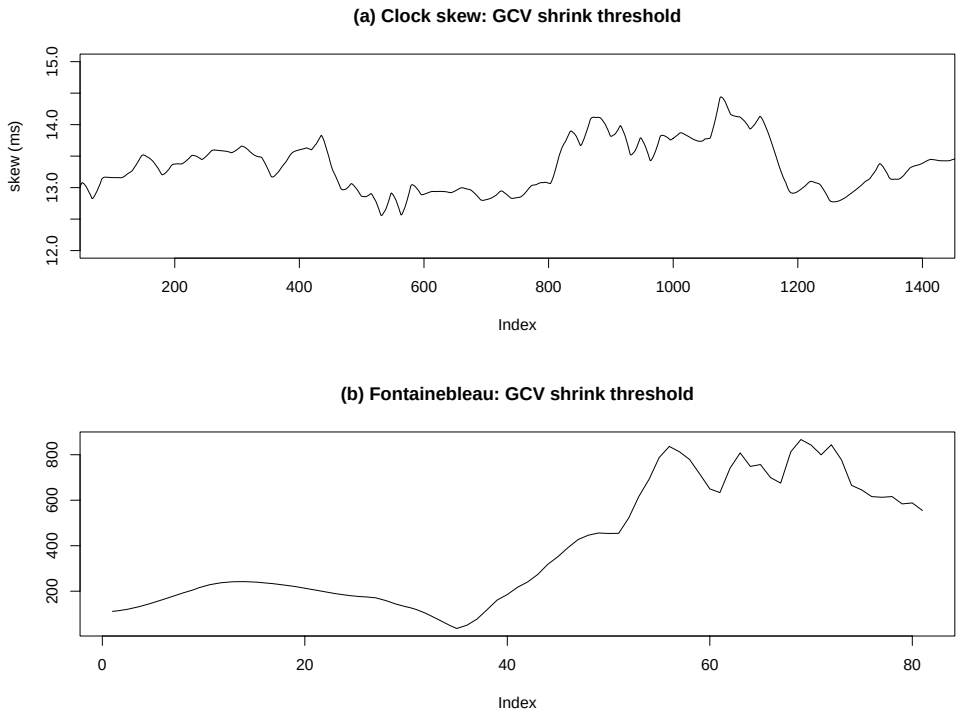


Figure 6.9: Results of GCV shrinkage. (a) Clock skew data, (b) one signature from the Fontainebleau dataset.

### 6.2.8 Discussion

The wavelet transform provides a sparse representation with high sparsity. The energy of *additive white Gaussian noise* ($+z_i$, $z_i \sim N(0,1)$ iid) is spread over many coefficients, while the energy of the function presumed to be tainted by this noise, should exist in a few high energy coefficients.

This heuristic is exploited by the seven methods detailed above. It is believed that this holds for a wider variety of noise, including slightly correlated (coloured) noise, and long tailed noise.

All the methods exhibit denoising quality on the clock skew data. Only the universal threshold with the *median absolute deviation (MAD)* estimate for $\sigma$ retains some of the presumed noise (b of figure 6.2 on page 103).

On one signature of the Fontainebleau dataset, the universal threshold smooths away too much detail. The James-Stein shrink (figure 6.6 on page 108) retains about the same detail as the original signature. The remaining five methods seem to bring out more structure in the signature. It is impossible to tell if this is the truth. Subsequent sections will shed some light on how the phenomenon affects classifier performance.

Donoho & Johnstone (1994) lay a bound on the expected loss for signal representation

$$R(T, \vec{f}) \leq A \times R_{\text{ideal}}(T, \vec{f}) \tag{6.2.20}$$

relative to the "ideal" risk on most of the discussed methods. It may be of some theoretical importance that $A$ is of a logarithmic order of $n$. I state that this inequality combined with the above mentioned heuristic, justifies wavelet thresholding as a class. The difference in performance on my (few) examples, should encourage some vigilance when selecting a shrinker. It is no apparent reason any of these methods shall perform better than the other.

Although being the simplest method, Visu shrink seems to be the best candidate on performance and computational complexity. The other methods, beside GCV shrink and the James-Stein shrink, can be seen to fall back to Visu shrink when their extended assumptions are absent.

Some of the considered methods stipulate that the soft threshold should be used. I abuse this to some degree, but most of the cited references accept that the thresholds asymptotically would be the same. Separate from this discussion, I note that Gao (1998) uses *the nonnegative Garrote*, and compares it to both the soft and hard threshold in the wavelet domain.

I have chosen to keep the overt Bayesian mindset from this limited investigation, but note that there exists an interesting "BayesShrink" in Chang et al. (2000). Chipman et al. (1997) and Barber et al. (2002) have a more thorough discussion.

## 6.3 Denoising in classification

In the discourse until now the focus has been on denoising without regard to feature extraction.

The reason for doing denoising is the hope that this makes the representation even more sparse. Section 5.3.3 describes the methods that I use to rank and select wavelet coefficients before they are used in classification. Some changes have to be done to accommodate the current situation.

There exist several ways to adapt the denoising methods of this chapter to the multivariate situation. I will not go down this path, since I have already assumed that all samples observed are independent. Methods for the multivariate situation are sometimes called ensemble denoising methods.

I will now reiterate the ranking and voting methods for feature extraction in section 5.3.3 with changes that accommodate the denoising in this chapter.

Parameter selection (which wavelet and # coefficients) is done as described in section 3.4. In figure 6.10 the conceptual system is shown. The ”reduce” step now consists of the denoising methods and the rank and vote methods described below.

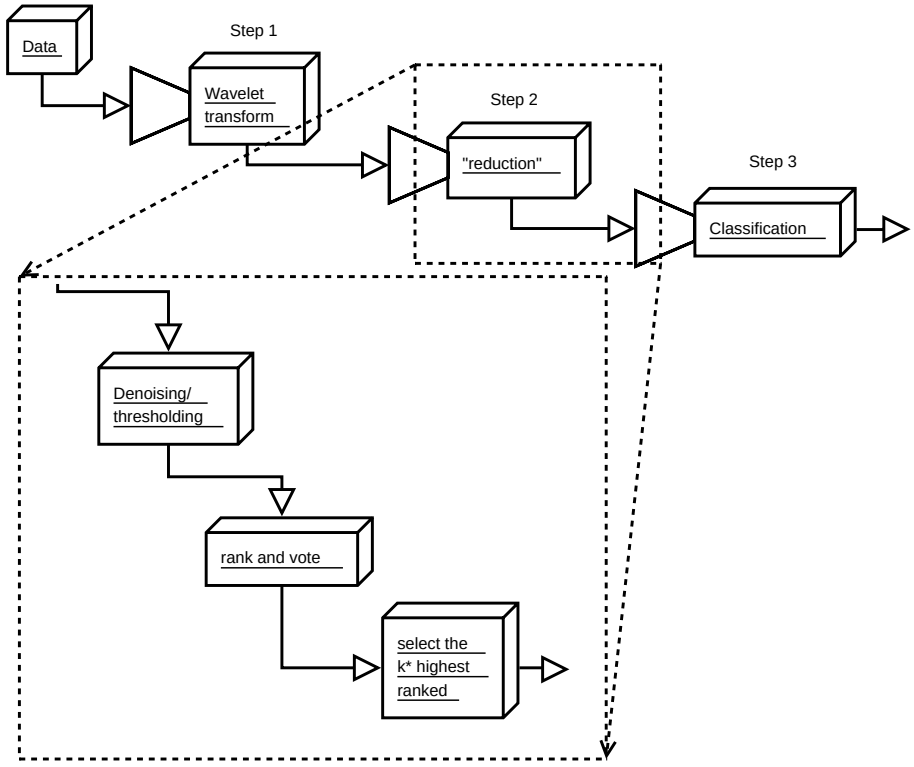


Figure 6.10: Details of the “reduce” step.

### 6.3.1 Method I: Above mean

The *wavelet packet decomposition (WPD)* gives a wavelet decomposition of the samples. The denoising methods are independently used on this wavelet decomposition of the $n$ samples. All coefficients with higher than mean magnitude (energy) are ranked by their magnitude, for each sample. Their order is a linear precedence, with a weighed vote. The $k^*$ coefficients ranked overall highest are kept. $k^*$ is selected by the method described in section 3.4.

### 6.3.2 Method II: On mean

The rationale for this method was that it would average away noise. This worked well in last chapter, but this rationale should not be present after denoising. In an experiment I will couple the Coifman-Wicherhauser algorithm of the previous chapter with the GCV denoiser.

First the wavelet transform of the mean of the data is denoised and then the Coifman-Wicherhauser algorithm is applied to this. All samples of the original data are then transformed to their wavelet form within the basis selected by the Coifman-Wicherhauser algorithm. For each sample, the coefficients are ranked after their magnitude. Their order is a linear precedence, with a weighed vote. The $k^*$ coefficients ranked overall highest are kept. $k^*$ is selected by the method described in section 3.4.

The EMD (section 5.4.1) is used instead of entropy in the Coifman-Wicherhauser algorithm. In the next sections, results derived by this method are denoted GCV EMD.

## 6.4 Results

With the ranking and voting in previous section the experiment was done as outlined in section 3.4. Unlike in the presentation of results in the previous chapter, results for the ten best parameter combinations are deferred to appendix B. In this section I will present the best parameter combination obtained for each denoising method.

In section 3.4 I mention that I will provoke *the curse of dimensionality* by using few samples in training. I would like to remind the reader that the classifier has only been trained with 30 (Fontainbleau), 50 (National Mall) or 90 (Pavia) samples in the results given here.

In table 6.2 on the next page the results on the Pavia dataset are given. The

results seem very uniform over the different denoisers. I remark that the FejerKorovkin wavelet of different orders (fk6, fk14 and fk22) features prominently on the list. This wavelet is based on the FejerKorovkin smoothing kernel, popular in neural-networks. An overview of wavelet families is given in appendix C.

| | | | | | 95% CI | |
|---|---|---|---|---|---|---|
| Denoiser | wavelet | level | # comp. | error | low | high |
| Universal MAD | fk22 | 6 | 20 | 0.057 | 0.037 | 0.080 |
| Universal SE | mb4 | 7 | 24 | 0.050 | 0.029 | 0.080 |
| Universal MAD Global | fk22 | 6 | 19 | 0.060 | 0.044 | 0.081 |
| Universal SE Global | fk14 | 7 | 13 | 0.050 | 0.030 | 0.075 |
| Visu shrink | fk22 | 7 | 12 | 0.058 | 0.044 | 0.081 |
| Risk shrink | fk22 | 6 | 17 | 0.058 | 0.041 | 0.080 |
| James-Stein | mb4 | 7 | 28 | 0.049 | 0.030 | 0.073 |
| SURE shrink | fk6 | 3 | 29 | 0.055 | 0.032 | 0.085 |
| Hybrid shrink | fk22 | 3 | 10 | 0.058 | 0.043 | 0.080 |
| GCV | d6 | 4 | 17 | 0.051 | 0.029 | 0.081 |
| GCV EMD | fk6 | 4 | 28 | 0.050 | 0.030 | 0.077 |

Table 6.2: Results for the Pavia dataset

The performance of the denoisers varies more on the Fontainebleau dataset in table 6.3 on the following page than on the Pavia dataset. Most lie at around 30% error, but both the GCV approaches are somewhat better. The Visu shrink performance is dramatically better with only 18% error.

The Visu shrink combination shown here accepts nearly twice the number of components than the other methods. The FejerKorovkin wavelet is again involved. In table B.1 on page 155 of the ten best combinations, all but two involve this wavelet. Perhaps this extended smoothing/denoising is what this dataset needs.

I am sceptical of this dramatic increase in performance, but the variance-bias tradeoff in figure 6.11 on the following page indicate no foul play. The experimental protocol that I subscribe to in chapter 3 does not allow for repeating the experiment. I suspect no integrity failings on part of the computer system, but repeat the experiment (in breach of protocol), on a separate system. The results are within three decimals of those presented here.

Result on the National Mall dataset can be found in table 6.4 on page 119. It shows some variability. Both the GCV shrinkers and one of the universal thresholds show good results.

For all three datasets the GCV shrinkers show good results (among the three best). The GCV approach as remarked in section 6.2.7 is less model dependant, and seems to thrive on this. The extra structure suggested in for instance (b) of figure 6.4 on page 105, might be at play in some of the results.

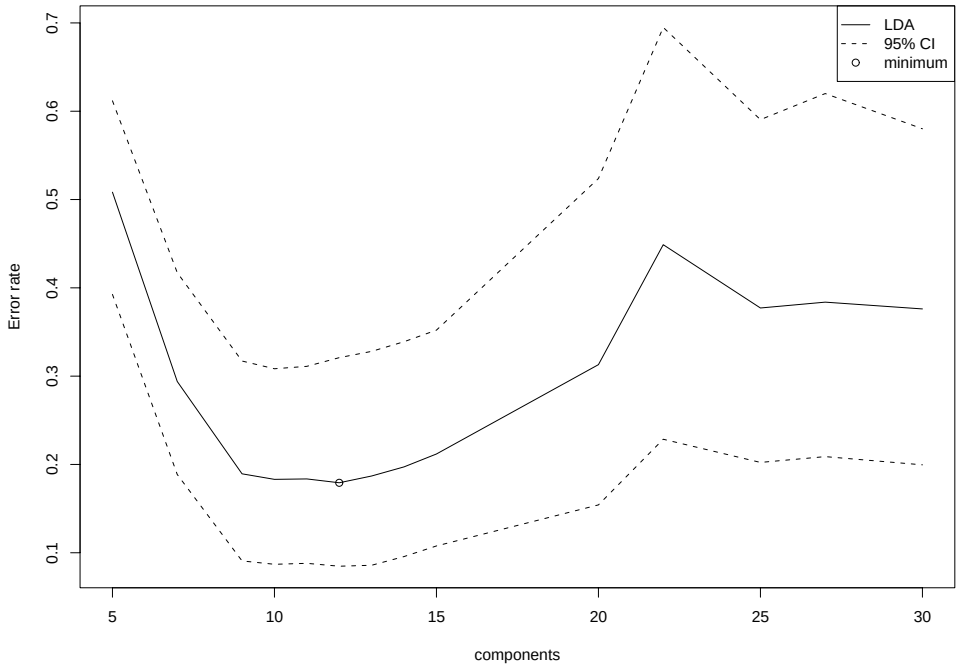


Figure 6.11: Variance-bias tradeoff for Visu shrink on the Fontainebleau dataset

| | | | | | 95% CI | |
|---|---|---|---|---|---|---|
| Denoiser | wavelet | level | # comp. | error | low | high |
| Universal MAD | haar | 7 | 6 | 0.307 | 0.250 | 0.390 |
| Universal SE | d4 | 3 | 6 | 0.315 | 0.249 | 0.409 |
| Universal MAD Global | mb8 | 3 | 3 | 0.316 | 0.270 | 0.394 |
| Universal SE Global | haar | 6 | 6 | 0.306 | 0.249 | 0.385 |
| Visu shrink | fk22 | 5 | 12 | 0.179 | 0.085 | 0.321 |
| Risk shrink | haar | 4 | 3 | 0.311 | 0.265 | 0.379 |
| James-Stein | haar | 3 | 6 | 0.308 | 0.247 | 0.407 |
| SURE shrink | w4 | 7 | 3 | 0.293 | 0.243 | 0.374 |
| Hybrid shrink | haar | 5 | 3 | 0.308 | 0.263 | 0.396 |
| GCV | d16 | 6 | 5 | 0.270 | 0.209 | 0.359 |
| GCV EMD | d16 | 7 | 5 | 0.267 | 0.210 | 0.354 |

Table 6.3: Results for the Fontainebleau dataset

| | | | | | 95% CI | |
|---|---|---|---|---|---|---|
| Denoiser | wavelet | level | # comp. | error | low | high |
| Universal MAD | mb16 | 7 | 11 | 0.116 | 0.081 | 0.176 |
| Universal SE | fk8 | 3 | 27 | 0.170 | 0.104 | 0.262 |
| Universal MAD Global | la16 | 8 | 5 | 0.149 | 0.106 | 0.211 |
| Universal SE Global | bs3.1 | 8 | 14 | 0.099 | 0.066 | 0.149 |
| Visu shrink | fk22 | 5 | 12 | 0.179 | 0.085 | 0.321 |
| Risk shrink | d6 | 4 | 9 | 0.136 | 0.096 | 0.194 |
| James-Stein | fk4 | 4 | 27 | 0.174 | 0.107 | 0.264 |
| SURE shrink | fk14 | 7 | 24 | 0.143 | 0.089 | 0.228 |
| Hybrid shrink | d6 | 3 | 9 | 0.137 | 0.096 | 0.201 |
| GCV | fk6 | 8 | 25 | 0.101 | 0.063 | 0.163 |
| GCV EMD | mb24 | 6 | 11 | 0.101 | 0.066 | 0.162 |

Table 6.4: Results for the National Mall dataset dataset

## 6.5 Discussion

Seven wavelet shrinkers or denoisers were discussed in this chapter. They all have some heuristic foundation, and show reasonable results on reconstruction of noisy functions. In the last section results from classification combined with these denoisers were shown.

Difference in performance between PCA and methods in this and the previous chapter will be discussed in chapter 7.

### 6.5.1 Validation

Following the discussion in chapter 3, a validation test set has been kept out of the analysis above. It is now brought out to validate the claims made.

In a thousand repetitions the same "starved" amount of training samples are drawn from the previously used set to train the classifier with the previous selected parameters. Then the validation test sets are classified.

The validation results in table 6.5 on the next page, show the same destitute results for the Pavia dataset as the result in section 5.5.1. On the positive side one of the universal thresholding methods, shows some resilience to the pertained difference between the training and validation sets. See section 5.5.1 for further discussion.

In table 6.6 on page 121 validation results are given for the Fontainebleau dataset. All but the Visu shrink method uphold their previous results. The striking difference from the previous Visu shrink results, only serves to stress the

| | | 95% CI | |
|---|---|---|---|
| Denoiser | error | low | high |
| Universal MAD | 0.80 | 0.59 | 0.93 |
| Universal SE | 0.81 | 0.62 | 0.94 |
| Universal MAD Global | 0.77 | 0.53 | 0.92 |
| Universal SE Global | 0.40 | 0.23 | 0.61 |
| Visu shrink | 0.82 | 0.66 | 0.93 |
| Risk shrink | 0.83 | 0.68 | 0.93 |
| James-Stein | 0.79 | 0.60 | 0.94 |
| SURE shrink | 0.66 | 0.38 | 0.87 |
| Hybrid shrink | 0.81 | 0.68 | 0.93 |
| GCV | 0.69 | 0.52 | 0.85 |
| GCV EMD | 0.66 | 0.49 | 0.83 |

Table 6.5: Validation results for the Pavia dataset

importance of proper validation. The fall in performance of 23% is not caught by the confidence bands, and could only be caught by validation.

The training and validation sets were parted by *simple random sampling (SRS)*. This ensures that the same "nature" is present in both sets. The drop in performance can not be explained by overfitting, as the "bootstrap cross-validatory" scheme (detailed in section 3.4) should undercut this. The drop in performance could then only be attributed to Visu shrink and the parameters selected for it. Visu shrink is not very complex. The FejerKorovkin wavelet selected, has a history as a smoother. However, desirable properties of the wavelet transform (orthogonality and energy preservation, see chapter 2) should avoid over smoothing by the wavelet itself.

Noise is a candidate to explain the dramatic drop in performance. Both Visu shrink and the FejerKorovkin wavelet should handle noise. There is no foul play in the variance-bias tradeoff in figure 6.11 on page 118. In the figure the expected "smile" is present, and the correct minimum is selected. The only speculative thing one can observe is that the classification algorithm does not fail until it reaches 31 parameters. With 30 samples available to train the classifier, this is less than one sample per parameter. This indicates that the wavelet combined with shrinker does the job too good.

Assume that there are minute differences in the magnitude (energy) levels of the coefficients between the training and validation sets. This will influence models with lower sample-to-parameter ratios more than others. It might be prudent to modify the way the minimum is selected. If the classification algorithm does not fail at a certain cutoff sample-to-parameter ratio, and the curve near the minimum is sufficient plane, a minimum with fewer parameters should be selected. In

figure 6.11 on page 118 nine parameters would probably be better than the twelve selected.

In table 6.6 the results for the National Mall dataset, validate the previous results.

| | | 95% CI | |
|---|---|---|---|
| Denoiser | error | low | high |
| Universal MAD | 0.30 | 0.24 | 0.40 |
| Universal SE | 0.31 | 0.24 | 0.41 |
| Universal MAD Global | 0.31 | 0.26 | 0.39 |
| Universal SE Global | 0.31 | 0.24 | 0.40 |
| Visu shrink | 0.41 | 0.34 | 0.51 |
| Risk shrink | 0.30 | 0.25 | 0.38 |
| James-Stein | 0.30 | 0.24 | 0.40 |
| SURE shrink | 0.29 | 0.24 | 0.39 |
| Hybrid shrink | 0.30 | 0.25 | 0.39 |
| GCV | 0.27 | 0.24 | 0.39 |
| GCV EMD | 0.26 | 0.20 | 0.36 |

Table 6.6: Validation results for the Fontainebleau dataset

| | | 95% CI | |
|---|---|---|---|
| Denoiser | error | low | high |
| Universal MAD | 0.12 | 0.08 | 0.18 |
| Universal SE | 0.18 | 0.11 | 0.28 |
| Universal MAD Global | 0.22 | 0.17 | 0.29 |
| Universal SE Global | 0.10 | 0.07 | 0.16 |
| Visu shrink | 0.14 | 0.10 | 0.21 |
| Risk shrink | 0.14 | 0.10 | 0.21 |
| James-Stein | 0.18 | 0.11 | 0.27 |
| SURE shrink | 0.15 | 0.10 | 0.24 |
| Hybrid shrink | 0.14 | 0.10 | 0.21 |
| GCV | 0.10 | 0.07 | 0.16 |
| GCV EMD | 0.10 | 0.07 | 0.17 |

Table 6.7: Validation results for the National Mall dataset

### 6.5.2 Difference in performance between the methods

Is there a difference in performance between the methods? In all the tables in this chapter a winner in performance can be declared. The confidence bands of most of the methods, do overlap. This signals that the declaration is not conclusive.

Cochran's Q-test of section 3.3 can answer this more decisively. In thousand repetitions a random subset was selected from the training set. For each repetition all methods are applied, and Cochran's Q-test is performed on the classified data.

Both the Q-statistic and p-values are reported with confidence bands in table 6.8. By all reasonable confidence levels the hypothesis that there is a difference between the methods must be rejected. This does not necessarily mean that one should not choose one method over the other.

| | | 95% CI | |
|---|---|---|---|
| | Q | low | high |
| Pavia | 11256.43 | 8632.41 | 13752.21 |
| p-value | 0.00 | 0.00 | 0.00 |
| Fontainebleau | 20595.90 | 14483.60 | 27726.05 |
| p-value | 0.00 | 0.00 | 0.00 |
| National Mall | 22800.54 | 18045.49 | 33494.50 |
| p-value | 0.00 | 0.00 | 0.00 |

Table 6.8: Cochran's Q-test

# CHAPTER 7

# Concluding remarks

In this thesis I have exploited the fact that the wavelet representation of hyperspectral data is sparse. My main contribution is to recognise that if one sacrifice reconstructability of the data an even sparser representation is possible. To the best of my knowledge the combination of methods in this thesis is original.

In these concluding remarks I will try to answer four questions:

- Which wavelet is the best wavelet?
- Are the wavelet methods better than PCA based methods?
- Does an alternate variance-bias strategy help in bad cases?
- What effect do more data have?

## 7.1 Which wavelet is the best wavelet?

This is perhaps the most difficult question to answer. The universe of different wavelets has no bound on its cardinality. In this thesis, I have applied 25 of the most common wavelets. An overview is given in appendix C. In section 1.3.1 I note that there exist methods to construct wavelets especially for classification. I have chosen not to take this path. Without having tested the methods in question, I fear that they might adapt too much to the data at hand. This will perhaps in the low data situation considered in this thesis, increase the generalisation error (over fitting).

In chapter 5 the 'd16', 'mb16' and 'fk22' wavelets feature prominently in the best results. In chapter 6 the same wavelets are among the best, but more focus is on particular wavelets like the 'fk22'. See comments in section 6.4.

The difference in performance between wavelets is minute. The level of decomposition and the number of coefficients selected are much more important than which wavelet one selects. If no prior information about how the data should look is available, and if no wavelet looks particularly akin to the data, one is probably better off by selecting one of the standard wavelets. In the statistical literature the 'la8' wavelet is especially used when discussing wavelets. This wavelet is related to the family of at least two of the above mentioned wavelets.

I suggest that the 'la8' wavelet, or one wavelet in the 'la', 'd' or 'mb' families is used when no prior information of the real shape of the data is available.

## 7.2 Are the wavelet methods better than PCA based methods?

In the two previous chapters Cochran's Q-test and the confidence bands make it clear that the wavelet based methods considered in this thesis can not largely be discerned from each other.

Table 4.5 on page 68 gives the performance of PCA based classification, these results are validated in table B.2 on page 155 (and shares the same validation problems on the Pavia dataset).

McNemar's test is described in section 3.3, it is ideally suited to compare classifier performance. Results are given in table 7.1 on the next page. Difference between the PCA and wavelet based methods must be accepted at all reasonable confidence levels. It should be remarked that for both the Pavia and Fontainebleau datasets the wavelet performance is on the lower confidence band of the PCA performance, while for the National Mall dataset the performance is within the confidence bounds (cf. table 4.5 on page 68).

In the table 7.1 the GCV EMD method replaces the Visu shrink method as this fails on the validation dataset. The performance on the Pavia dataset is void under all circumstances as the validation fails blatantly.

The other criterion to judge the methods on, is how the transformed data fits into the classifier model. The LDA classifier model is described in section 3.1. This model relies on normality. Methods to test for normality are described in appendix A.1. The results of these tests can be found in table 7.2 on page 126 and table 7.3 on page 127. These tests are not conclusive, but one can convince oneself of an "increase in normality" for at least the Pavia and Fontainebleau datasets. Together

| | Pavia | | Fontainebleau | | National Mall | |
|---|---|---|---|---|---|---|
| Procedure | # error | error | # error | error | # error | error |
| PCA | 142 | 0.08 | 599 | 0.12 | 1323 | 0.33 |
| Wavelet | 87 | 0.05 | 475 | 0.09 | 1043 | 0.26 |
| $\Lambda$ | | 3.57 | | 3.75 | | 5.74 |
| p | | 0.00 | | 0.00 | | 0.00 |

Table 7.1: McNemar's test, PCA vs: for the two fist datasets the Coifman-Wickerhauser algorithm with the EMD *on mean*; the last dataset GCV EMD thresholding

with the visual impression in figure 4.6 on page 60 this shows that the wavelet based methods have some merit to fulfil the model requirements of normality.

| Method | Koziol | | Mardia | | | | Royston | | |
|---|---|---|---|---|---|---|---|---|---|
| (data) | $J_n$ | p | A | df | p | B | p | H | e | p |
| PCA (P) | 0.07 | 0.50 | 362.07 | 286 | 0.00 | 952347298 | 0.00 | 7.00 | 10.08 | 0.73 |
| PCA (F) | 0.04 | 0.87 | 25.58 | 20 | 0.18 | −0 | 0.81 | 18.70 | 3.20 | 0.00 |
| PCA (NM) | 0.46 | 0.00 | 440.96 | 286 | 0.00 | 4 | 0.00 | 3.75 | 10.08 | 0.96 |
| CW92 Above (P) | 0.10 | 0.28 | 314.13 | 286 | 0.12 | 5746561 | 0.00 | 12.45 | 10.08 | 0.26 |
| CW92 Above (F) | 0.07 | 0.52 | 36.26 | 35 | 0.41 | 1550 | 0.00 | 8.08 | 4.17 | 0.10 |
| CW92 Above (NM) | 2.14 | 0.00 | 335.20 | 286 | 0.02 | 3317174 | 0.00 | 0.42 | 10.08 | 1.00 |
| CW92 On (P) | 0.19 | 0.05 | 76.76 | 286 | 1.00 | 12423810859781 93 | 0.00 | 12.18 | 10.08 | 0.28 |
| CW92 On (F) | 0.05 | 0.77 | 44.06 | 35 | 0.14 | 78259044 | 0.00 | 8.27 | 4.17 | 0.09 |
| CW92 On (NM) | 0.18 | 0.06 | 391.01 | 286 | 0.00 | 8624947068 | 0.00 | 1.27 | 10.08 | 1.00 |
| CW92/EMD Above (P) | 0.17 | 0.08 | 259.32 | 286 | 0.87 | 2670097 | 0.00 | 13.48 | 10.08 | 0.20 |
| CW92/EMD On (P) | 0.11 | 0.25 | −126.40 | 286 | 1.00 | 24310511672529 | 0.00 | 19.52 | 10.08 | 0.04 |
| CW92/EMD On (F) | 0.05 | 0.77 | 44.06 | 35 | 0.14 | 78259044 | 0.00 | 8.27 | 4.17 | 0.09 |
| CW92/EMD On (NM) | 0.36 | 0.00 | 314.92 | 165 | 0.00 | 1655 | 0.00 | 0.65 | 8.10 | 1.00 |
| Uni. MAD (P) | 1.15 | 0.00 | 4781.77 | 286 | 0.00 | 45200034 | 0.00 | 0.21 | 10.08 | 1.00 |
| Uni. MAD (F) | 0.15 | 0.12 | 41.64 | 56 | 0.92 | 9457 | 0.00 | 14.72 | 5.14 | 0.01 |
| Uni. MAD (NM) | 0.26 | 0.01 | 243.64 | 286 | 0.97 | 75021697 | 0.00 | 0.75 | 10.08 | 1.00 |
| Uni. SD (P) | 0.09 | 0.35 | 225.42 | 286 | 1.00 | 47287 | 0.00 | 19.34 | 10.08 | 0.04 |
| Uni. SD (F) | 0.11 | 0.22 | 42.83 | 56 | 0.90 | 11746 | 0.00 | 18.84 | 5.14 | 0.00 |
| Uni. SD (NM) | 0.04 | 0.84 | 279.35 | 286 | 0.60 | 448867 | 0.00 | 0.09 | 10.08 | 1.00 |
| Uni. MAD/G (P) | 1.08 | 0.00 | 6577.01 | 286 | 0.00 | 8599871 | 0.00 | 0.48 | 10.08 | 1.00 |
| Uni. MAD/G (F) | 0.88 | 0.00 | 1.89 | 10 | 1.00 | 26 | 0.00 | 113.16 | 2.25 | 0.00 |
| Uni. MAD/G (NM) | 0.75 | 0.00 | 47.90 | 35 | 0.07 | 9558997 | 0.00 | 1.24 | 4.17 | 0.89 |
| Uni. SD/G (P) | 0.23 | 0.02 | 353.95 | 286 | 0.00 | 205509 | 0.00 | 21.00 | 10.08 | 0.02 |
| Uni. SD/G (F) | 0.15 | 0.12 | 41.64 | 56 | 0.92 | 9457 | 0.00 | 14.72 | 5.14 | 0.01 |
| Uni. SD/G (NM) | 0.41 | 0.00 | 501.03 | 286 | 0.00 | 64965 | 0.00 | 0.27 | 10.08 | 1.00 |

Table 7.2: Normality tests

| Method | Koziol | | Mardia | | | | | Royston | | |
|---|---|---|---|---|---|---|---|---|---|---|
| (data) | $J_n$ | p | A | df | p | B | p | H | e | p |
| Visu (P) | 0.09 | 0.36 | 340.83 | 286 | 0.01 | 21920939 | 0.00 | 23.86 | 10.08 | 0.01 |
| Visu (F) | 0.04 | 0.87 | 25.58 | 20 | 0.18 | −0 | 0.81 | 18.70 | 3.20 | 0.00 |
| Visu (NM) | 0.33 | 0.00 | 195.77 | 120 | 0.00 | 12524149420 | 0.00 | 0.88 | 7.11 | 1.00 |
| Risk (P) | 0.14 | 0.15 | 374.46 | 286 | 0.00 | 351782212 | 0.00 | 27.94 | 10.08 | 0.00 |
| Risk (F) | 0.06 | 0.62 | 10.36 | 10 | 0.41 | 12495440 | 0.00 | 71.40 | 2.25 | 0.00 |
| Risk (NM) | 0.38 | 0.00 | 236.97 | 165 | 0.00 | 1201810894 0 | 0.00 | 0.71 | 8.10 | 1.00 |
| JS (P) | 0.09 | 0.35 | 225.42 | 286 | 1.00 | 47287 | 0.00 | 19.34 | 10.08 | 0.04 |
| JS (F) | 0.15 | 0.12 | 41.64 | 56 | 0.92 | 9457 | 0.00 | 14.72 | 5.14 | 0.01 |
| JS (NM) | 0.05 | 0.76 | 268.66 | 286 | 0.76 | 782112 | 0.00 | 0.09 | 10.08 | 1.00 |
| SURE (P) | 0.08 | 0.39 | 247.24 | 286 | 0.95 | 153307880273 | 0.00 | 6.67 | 10.08 | 0.76 |
| SURE (F) | 0.04 | 0.87 | 10.86 | 10 | 0.37 | 309 | 0.00 | 20.20 | 2.25 | 0.00 |
| SURE (NM) | 0.03 | 0.90 | 279.47 | 286 | 0.60 | 1906950 | 0.00 | 1.00 | 10.08 | 1.00 |
| Hybrid (P) | 0.35 | 0.00 | 217.36 | 220 | 0.54 | 1039360884297 | 0.00 | 21.54 | 9.09 | 0.01 |
| Hybrid (F) | 0.24 | 0.02 | 29.00 | 10 | 0.00 | 138362 | 0.00 | 73.14 | 2.25 | 0.00 |
| Hybrid (NM) | 0.38 | 0.00 | 236.97 | 165 | 0.00 | 1201810894 0 | 0.00 | 0.71 | 8.10 | 1.00 |
| GCV (P) | 0.11 | 0.24 | 278.21 | 286 | 0.62 | 21310 | 0.00 | 10.49 | 10.08 | 0.41 |
| GCV/EMD (P) | 0.09 | 0.32 | −5081.41 | 286 | 1.00 | 342335635291658 | 0.00 | 16.38 | 10.08 | 0.09 |
| GCV/EMD (F) | 0.05 | 0.77 | 44.06 | 35 | 0.14 | 78259044 | 0.00 | 8.27 | 4.17 | 0.09 |
| GCV/EMD (NM) | 0.48 | 0.00 | 248.75 | 286 | 0.95 | 108915106938 | 0.00 | 0.80 | 10.08 | 1.00 |

Table 7.3: Normality tests

## 7.3 Does an alternate variance-bias strategy help in bad cases?

In section 6.5.1 the validation of the Visu shrink results on the Fontainebleau dataset fails. In the same section I suggest another rule to select the "best" point in the variance-bias tradeoff. This alternate rule would select a minimum with fewer parameters if the curve near the minimum is sufficient plane. In figure 6.11 on page 118 nine parameters would be selected rather than the twelve of the ordinary rule.

Table 7.4 shows that the alternate rule does not help much.

| | error | validation error |
|---|---|---|
| Old rule | 0.17 | 0.41 |
| New rule | 0.18 | 0.38 |

Table 7.4: Old and new rule

## 7.4 What effect do more data have?

In table 7.5 the classification performance at different data levels is illustrated. Apparently at three times away from the presumed *curse of dimensionality* border all datasets perform better. The extra training data do not help the validation on the Pavia dataset.

| | 1x | 2x | 3x |
|---|---|---|---|
| Pavia | 0.05 | 0.04 | 0.04 |
| validation | 0.67 | 0.66 | 0.67 |
| Fontainebleau | 0.30 | 0.26 | 0.25 |
| validation | 0.28 | 0.24 | 0.23 |
| National Mall | 0.10 | 0.08 | 0.08 |
| validation | 0.11 | 0.09 | 0.08 |

Table 7.5: Coifman-Wickerhauser algorithm with the above mean method of section 5.3.3 for different data amounts. 1x is respectively 90, 30 and 50 samples in the different datasets.

## 7.5 Future research

Wavelets are one set of atoms that can be used in atomic decomposition and "wavelet like" denoising. A host of "wavelet like" atoms like Ridglets (Candès & Donoho (1999)) and Beamlets (Donoho & Huo (2001)) is available and might adapt better to the data.

### 7.5.1 Generalisation to non-spectral data

In the introduction I indicate that more and more data become available. A discipline where this is apparent is genetics. With microarray technology the expression of several thousand genes can be simultaneously measured on the same samples.

In table 7.6 the Coifman-Wickerhauser algorithm with the above mean method of section 5.3.3 was applied to the cancer classification data of Golub et al. (1999). The error measuring methods are described in section A.2.1. Without much effort the results are near those of the original article. With more research the wavelet methods presented in this thesis can be adapted to this type of data.

| | $O$ | $OS$ | $D$ | $DS$ | $U_1$ | $U_2$ | holdout error |
|---|---|---|---|---|---|---|---|
| CW above mean | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.38 | 0.32 |

Table 7.6: The Coifman-Wickerhauser algorithm with the above mean method on the cancer data of Golub et al. (1999)

### 7.5.2 Combination of PCA and wavelets

It is not given that the wavelet methods presented here give the best features for classification. It can be hypothesised that more standard feature extraction methods can further improve on the representation of the data given by my methods.

Recently Aminghafari et al. (2006) suggest applying a modification to the *principal components analysis (PCA)* to wavelet thresholded data. In table 7.7 on the following page simple PCA was through cross-validation applied to one of my methods. Slight improvements can be seen.

| | | 95% CI | |
|---|---|---|---|
| | error | low | high |
| Original | 0.10 | 0.06 | 0.16 |
| Original+PCA | 0.08 | 0.06 | 0.14 |

Table 7.7: Test set error: PCA improvement on the universal threshold with global standard error, National Mall dataset.

## 7.6 Acknowledgement

The computations in this thesis were done in the R statistical system (R Development Core Team (2004)). The Condor batch system (Litzkow et al. (1988)) running on both the Mathematical and CS departments of the University of Oslo was used.

I would like to thank my supervisor *Anne Solberg* for providing comments and suggesting that I use the *earth movers distance (EMD)*.

# BIBLIOGRAPHY


Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**, 716–723.

Aminghafari, M., Cheze, N. & Poggi, J.-M. (2006). Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis* **50**, 2381–2398.

Ancona, M., Cazzola, W., Raffo, P. & Corvi, M. (2002). Image Database Retrieval Using Wavelet Packets Compressed Data. In *Proceedings of the Sixth SIMAI National Conference*. Chia Laguna, Italy.

Anderson, T. (1951). Classification by multivariate analysis. *Psychometrika* **16**, 31–50.

Antoniadis, A., ed. (1995). *Wavelets and Statistics*, chap. Wavelab and reproducible research (J. Buckheit and D. L. Donoho). New York: Springer-Verlag, pp. 55–83.

Araújo, R. T. S., de Medeiros, F. N. S., Costa, R. C. S., Marques, R. C. P., Moreira, R. B. & Silva, J. L. (2004). Locating oil spill in sar images using wavelets and region growing. In *IEA/AIE'2004: Proceedings of the 17th international conference on Innovations in applied artificial intelligence*. Springer Springer Verlag Inc.

Barber, S., Nason, G. P. & Silverman, B. W. (2002). Posterior probability intervals for wavelet thresholding. *Journal of the Royal Statistical Society series B - methodology* **64**, 189–205.

Bartle, R. G. (1995). *The elements of integration and Lebesgue measure*. New York: Wiley. "A Wiley-Interscience publication.".

Basedow, R. W., Carmer, D. C. & Anderson, M. E. (1995). Hydice system: implementation and performance. In *Imaging Spectrometry (Proc. SPIE Int. Soc. Opt. Eng.)*, M. R. Descour, J. M. Mooney, D. L. Perry & L. R. Illing, eds., vol. 2480. Orlando, FL, USA: SPIE.

Bellman, R. (1961). *Adaptive control processes : a guided tour*. Princeton, N.J.: Princeton University Press.

Benedetto, J. J. (1997). *Harmonic analysis and applications*. Boca Raton: CRC Press.

Bickel, P. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. Recent advances in statistics, Pap. in Honor of H. Chernoff, 511-528 (1983).

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

Breiman, L. (2001a). Random forests. *Machine Learning* **V45**, 5–32. 10.1023/A:1010933404324.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science* **16**, 199–215.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

Brown, M. & Costen, N. P. (2005). Exploratory basis pursuit classification. *Pattern Recognition Letters* **26**, 1907–1915.

Bruce, L. M., Koger, C. H. & Li, J. (2002). Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. *Geoscience and Remote Sensing, IEEE Transactions on* **40**, 2331–2338.

Calderón, A. P. (1963). Intermediate spaces and interpolation. *Stud. Math., Ser. spec. No.* **1**, 31–34.

Calderón, A. P. (1964). Intermediate spaces and interpolation, the complex method. *Studia Math.* **24**, 113–190.

Candès, E. J. & Donoho, D. L. (1999). Ridgelets: a key to higher-dimensional intermittency? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **357**, 2495–2509. Doi:10.1098/rsta.1999.0444.

Capobianco, E. (2004). Effective decorrelation and space dimensionality reduction of multiscaling volatility. *Physica A: Statistical Mechanics and its Applications* **340**, 340–346.

Chang, S. G., Yu, B. & Vetterli, M. (2000). Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing* **9**, 1532–1546.

Chen, S. S. (1995). *Basis Pursuit*. Ph.D. thesis, Department of Statistics, Stanford University.

Chen, S. S. B., Donoho, D. L. & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM REVIEW* **43**, 129–159.

Chipman, H. A., Kolaczyk, E. D. & McCullogh, R. E. (1997). Adaptive bayesian wavelet shrinkage. *Journal of the American Statistical Association* **92**, 1413–1421.

Christie, M. (2004). Data collection and the ozone hole: Too much of a good thing? In *Proceedings of the International Commission on History of Meteorology*.

Clark, R., Swayze, G., Gallagher, A., King, T. & Calvin, W. (1993). The u.s. geological survey, digital spectral library: version 1: 0.2 to 3.0 microns. Tech. rep., U.S. geological survey, Reston, Va. Open File Report 93-592, http://speclab.cr.usgs.gov/spectral.lib04/spectral-lib04.html.

Cochran, W. (1950). The comparison of percentages in matched samples. *Biometrika* **37**, 256–266.

Coifman, R. & Weiss, G. (1977). Extensions of hardy spaces and their use in analysis. *Bulletin of The American Mathematical Society* **83**, 569–645.

Coifman, R. R. & Wickerhauser, M. V. (1992). Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on* **38**, 713–718.

Cox, D. & Small, N. (1978). Testing multivariate normality. *Biometrika* **65**, 263–272.

Cox, D. R. (2001). Biometrika: The first 100 years. *Biometrika* **88**, 3–11.

Cramer, H. (1945). *Mathematical methods of statistics*. Uppsala: Hugo Gebers Førlag.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 337-403.

Currin, C., Mitchell, T., Morris, M. & Ylvisaker, D. (1991). Bayesian prediction of deterninistic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association* **86**, 953–963.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics* **41**, 909–996.

Daubechies, I. (1992). Ten lectures on wavelets. Philadelphia: Society for Industrial and Applied Mathematics.

Daubechies, I. & Sweldens, W. (1998). Factoring wavelet transforms into lifting steps. *Journal of Fourier Analysis and Applications* **V4**, 247–269. 10.1007/BF02476026.

Davidson, K. R. & Donsig, A. P. (2002). *Real analysis with real applications.* Upper Saddle River, NJ: Prentice Hall.

Dempster, A. & Weisberg, H. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society series B - Methodoligical* **30**, 205–&.

DeVore, R. A. & Lucier, B. J. (1992). Fast wavelet techniques for near-optimal image processing. In *Military Communications Conference, 1992. MILCOM '92, Conference Record. 'Communications - Fusing Command, Control and Intelligence'., IEEE.*

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

Donoho, D. L. & Huo, X. (2001). Beamlets and multiscale image processing. Tech. rep., Department of Statistics, Stanford University, Stanford, Ca.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. Royal Stat. Soc. Ser. B* **57**, 301–369.

dos Anjos, A., Ellis, N., Haller, J., Landon, M., Spiwoks, R., Wengler, T., Wiedenmann, W. & Zobernig, H. (2006). Configuration of the atlas trigger. *Nuclear Science, IEEE Transactions on* **53**, 990–994.

Dwight, J. S. (1839). *Specimens of foreign standard literature ; 3*, chap. Select minor poems : translated from the German of Goethe and Schiller. Boston : Hilliard, Gray, and company. Half title page : Specimens of foreign standard literature / edited by George Ripley.

Edwards, A. L. (1948). Note on the correction for continuity in testing the significance of the difference between correlated proportions. *Psychometrika* **13**, 185–187. 10.1007/BF02289261.

Efron, B. (1979). 1977 Rietz lecture - Bootstrap methods - another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule - improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association* **92**, 548–560.

Farman, J., Gardiner, B. & Shanklin, J. (1985). Large losses of total ozone in Antarctica reveal seasonal CLOX/NOX interaction. *Nature* **315**, 207–210.

Fazel-Rezai, R. & Ramanna, S. (2005). Brain signals: Feature extraction and classification using rough set methods. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference, RSFDGrC 2005, Regina, Canada, August 31 - September 3, 2005, Proceedings, Part II*, D. Slezak, J. Yao, J. F. Peters, W. Ziarko & X. Hu, eds., vol. 3642 of *Lecture Notes in Computer Science*. Springer.

Fefferman, C. & Stein, E. (1972). Hp spaces of several variables. *Acta Mathematica* **129**, 137–193. 10.1007/BF02392215.

Feichtinger, H. & Grochenig, K. (1988). A unified approach to atomic decompositions via integrable group-representations. *LECTURE NOTES IN MATHEMATICS* **1302**, 52–73.

Field, D. J. (1999). Wavelets, vision and the statistics of natural scenes. *Philosophical transactions of The Royal Society of London series A-Mathenatical Physical and Enginerring sciences* **357**, 2527–2542.

Fisher, R. A. (1936a). Has Mendel's work been rediscovered? *Annals of Science* **1**, 115–137.

Fisher, R. A. (1936b). The use of multiple measurments in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Flandrin, P. (1992). Wavelet analysis and synthesis of fractional brownian motion. *Information Theory, IEEE Transactions on* **38**, 910–917.

Franklin, P. (1928). A set of continuous orthogonal functions. *Mathematische Annalen* **100**, 522–529. 10.1007/BF01448860.

Friedman, J. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association* **82**, 249–266.

Gabor, D. (1946). Theory of communication. *Journal of the Institution of Electrical Engineers* **93**, 429–457.

Gabor, D. (1947). Acoustical quanta and the theory of hearing. *Nature* **159**, 591–594.

Gamba, P. (2004). A collection of data for urban area characterization. In *Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004 IEEE International*, vol. 1.

Gao, H. Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *Journal of Computational and Graphical statistics* **7**, 469–488.

Geary, R. (1947). Testing for normality. *Biometrika* **34**, 209–242.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: Wiley.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Goodman, I. R. (1997). *Mathematics of data fusion*. Dordrecht: Kluwer.

Goupillaud, P., Grossmann, A. & Morlet, J. (1984). Cycle-octave and related transforms in seismic signal analysis. *Geoexploration* **23**, 85–102.

Grossmann, A. & Morlet, J. (1984). Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis* **15**, 723–736.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen* , 331–371 Transelation in Heil & Walnut (2006).

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Healey, G. & Slater, D. (1999). Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions. *Geoscience and Remote Sensing, IEEE Transactions on* **37**, 2706–2717.

Heath, T. L. (1921). *A history of Greek mathematics*. Oxford: Clarendon Press. Vol. 2.

Hecht, E. (1975). *Schaum's outline of theory and problems of optics*. New York: McGraw-Hill.

Heil, C. & Walnut, D. (2006). *Fundamental papers in wavelet theory*. Princeton, N.J.: Princeton University Press.

Henze, N. & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods* **19**, 3595–3617.

Hubbard, B. B. (1996). *The world according to wavelets: the story of a mathematical technique in the making*. Wellesley, Mass.: A.K. Peters.

Huber, P. (1985). Projection pursuit. *Annals of Statistics* **13**, 435–475.

Hyvärinen, A., Hoyer, P. & Oja, E. (2001). *Image Denoising by Sparse Code Shrinkage*. New York: IEEE Press. In Intelligent signal processing, Haykin, S. (ed.).

Jansen, M. (2001). *Noise reduction by wavelet thresholding*. New York: Springer.

Jansen, M., Malfait, M. & Bultheel, A. (1997). Generalized cross validation for wavelet thresholding. *Signal Processing* **56**, 33–44.

Jimenez, L. O. & Landgrebe, D. A. (1998). Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *Systems, Man and Cybernetics, Part C, IEEE Transactions on* **28**, 39–54.

Johnstone, I. M. & Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society series B - Methodoligical* **59**, 319–351.

Kaewpijit, S., Le Moigne, J. & El-Ghazawi, T. (2003). Automatic reduction of hyperspectral imagery using wavelet spectral analysis. *Geoscience and Remote Sensing, IEEE Transactions on* **41**, 863–871.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **23**, 187–200. 10.1007/BF02289233.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering* **82**, 35–45.

Koziol, J. (1982). A class of invariant procedures for assessing multivariate normality. *Biometrika* **69**, 423–427.

Kreutz-Delgado, K. & Rao, B. D. (1998). Measures and algorithms for best basis selection. In *Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on*, vol. 3.

Kullback, S. & Leibler, R. (1951). On information and sufficiency. *Annals of Mathemetical statistics* **22**, 79–86.

Lachenbruch, P. (1968). On expected probabilities of misclassification in discriminant analysis necessary sample size and a relation with multiple correlation coefficient. *Biometrics* **24**, 823–&.

Lachenbruch, P. & Mickey, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1–11.

Landgrebe, D. A. (2003). *Signal theory methods in multispectral remote sensing*. Hoboken, N.J.: Wiley-Interscience. With CD-ROM.

Leadbetter, M. R., Lindgren, G. & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. New York: Springer. Har bibliografi.

Lee, C. & Landgrebe, D. A. (1993). Feature extraction based on decision boundaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**, 388–400.

Litzkow, M., Livny, M. & Mutka, M. (1988). Condor - a hunter of idle workstations. In *Proceedings of the 8th International Conference of Distributed Computing Systems*.

Mallat, S. (1989). A theory for multiresolution signal decomposition - the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

Mallat, S. (1999). *A wavelet tour of signal processing*. San Diego, Calif.: Academic Press.

Mallat, S. G. & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]* **41**, 3397–3415.

Mallet, Y., Coomans, D., Kautsky, J. & De Vel, O. (1997). Classification using adaptive wavelets for feature extraction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **19**, 1058–1066.

Mallows, C. (1973). Some comments on Cp. *Technometrics* **15**, 661–675.

Mallows, C. (2006). Tukey's paper after 40 years. *Technometrics* **48**, 319–325.

Mardia, K. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–&.

Mardia, K. (1974). Applications of some measures of multivaritae skwenwss and kurtosis in testing normality and robustness studies. *Sankhya-The Indian Journal of Statistics series B* **36**, 115–128.

Mardia, K. V., Kent, J. & Bibby, J. (1979). *Multivariate analysis*. London: Academic Press.

Mckay, M. D., Beckman, R. J. & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **V12**, 153–157. 10.1007/BF02295996.

Mecklin, C. & Mundfrom, D. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation* **75**, 93–107. Doi:10.1080/0094965042000193233.

Mecklin, C. J. & Mundfrom, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review* **72**, 123–138.

Meyer, Y. (1993). *Wavelets : algorithms & applications*. Philadelphia: Society for Industrial and Applied Mathematics.

Morrison, D. F. (1976). *Multivariate statistical methods*. New York: McGraw-Hill. Bibliografi: s. 346-361.

Oakley, J. & O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89**, 769–784.

Okamoto, M. (1963). An asymptotic-expansion for distribution of linear discriminant function. *Annals of Mathematucal Statistics* **34**, 1286–&.

Page, J. (1985). Error-rate estimation in discriminant-analysis. *Technometrics* **27**, 189–198.

Patrício, J., Portugal, L., Resende, M., Veiga, G. & Júdice, J. (2004). Fortran subroutines for network flow optimization using an interior point algorithm. Tech. Rep. TD-5X2SLN, AT&T Labs Research.

Peleg, S., Werman, M. & Rom, H. (1989). A unified approach to the change of resolution: space and grey-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-11(7), July, 1989* , 739–742.

Percival, D. B. (2000). *Wavelet methods for time series analysis*. Cambridge: Cambridge University Press.

Pesquet, J. C., Krim, H., Leporini, D. & Hamman, E. (1996a). Bayesian approach to best basis selection. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 5.

Pesquet, J. C., Krim, H., Leporini, D. & Hamman, E. (1996b). Bayesian approach to best basis selection. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 5.

Poggioli, R. (1951). Realism in russia. *Comparative Literature* **3**, 253–267.

Potter, R. (1945). Visible patterns of sound. *Science* **102**, 463–470.

Proakis, J. G. & Manolakis, D. G. (2007). *Digital signal processing*. Upper Saddle River, N.J.: Pearson Prentice Hall.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, C. (1948). Tests of significance in multivariate analysis. *Biometrika* **35**, 58–79.

Reichhardt, T. (2003). Photos stop as landsat 7 defies engineers. *Nature* **423**, 907–907. 10.1038/423907a.

Rice, J. A. (1995). *Mathematical statistics and data analysis*. Belmont, Calif.: Duxbury Press. 1 diskett (9 cm) i lomme.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Royston, J. (1982). An extension of shapiro and wilk-w test for normality to large samples. *Applied Statistics-Journal of the Royal Statistical Society series c* **31**, 115–124.

Royston, J. (1983). Some techniques for assessing multivariate normality based on the shapiro-wilk-w. *Applied Statistics-Journal of the Royal Statistical Society series c* **32**, 121–133.

Royston, P. (1992). Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing* **V2**, 117–119. 10.1007/BF01891203.

Rubner, Y., Tomasi, C. & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision* **40**, 99–121.

Schauder, J. (1927). Bemerkungen zu meiner arbeit â??zur theorie stetiger abbildungen in funktionalrã?umenâ?? *Mathematische Zeitschrift* **26**, 417–431. 10.1007/BF01475462.

Schowengerdt, R. A. (1997). *Remote sensing : models and methods for image processing*. San Diego: Academic Press. 1. utg. med tittel: Techniques for image processing and classification in remote sensing.

Schwarz, G. (1978). Estimating dimension of a model. *Annals of Statistics* **6**, 461–464.

Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. New York: Wiley.

Shannon, C. (1948). A mathematical theory of communication. *Bell system technical journal* **27**, 379–423.

Shapiro, S. & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–&.

Slepian, D. (1976). On bandwidth. *Proceedings of the IEEE* **64**, 292–300.

Song, S.-p. & Que, P.-w. (2006). Wavelet based noise suppression technique and its application to ultrasonic flaw detection. *Ultrasonics* **44**, 188–193.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology* **100**, 441–471. Special Centennial Issue.

Srivastava, M. & Hui, T. (1987). On assessing multivariate normality based on shapiro-wilk w-statistic. *Statistics & Probability letters* **5**, 15–18.

Stein, C. (1981). Estimation of the mean of a multivariate normal-distribution. *Annals of Statistics* **9**, 1135–1151.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society series C-methodilogical* **36**, 111–147.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64**, 29–35.

Strang, G. & Nguyen, T. (1996). *Wavelets and filter banks*. Wellesley, MA: Wellesley-Cambridge Press.

Tian, J., Baraniuk, R. G., Wells, R. O. J., Tan, D. M. & Wu, H. R. (2000). Wavelet folding and decorrelation across the scale. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 1.

Tukey, J. (1962). Future of data-analysis. *Annals of Mathematical statistics* **33**, 1–&.

Vermote, E. F., Tanre, D., Deuze, J. L., Herman, M. & Morcette, J. J. (1997). Second simulation of the satellite signal in the solar spectrum, 6s: an overview. *Geoscience and Remote Sensing, IEEE Transactions on* **35**, 675–686.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. & Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association* **99**, 659–672.

# APPENDIX A

# Assesment of normality and misclassification

## A.1 Assessing normality

Geary (1947) gives a fairly good overview of testing for univariate normality. It is clear how one should test for univariate normality. Mecklin & Mundfrom (2004) give an overview of the current state in testing for multivariate normality. The multivariate situation is not so clear about how one should proceed. Mecklin & Mundfrom (2004) note that there exists over fifty different procedures and that neither of them sticks out as the best method.

Cox & Small (1978) is an earlier overview article. Section 5.4 of Gnanadesikan (1977) is a great exposition of nearly all underlying consideration for the tests mentioned in Mecklin & Mundfrom (2004).

Marginal normality does not imply joint (multivariate) normality, however departure from joint normality is often reflected in departure from marginal normality.

Under this observation, standard goodness-of-fit tests like the Pearson's chi-square and the Kolmogorov-Smirnov type tests might be used.

Other popular tests are based on testing for multivariate skewness and kurtosis. There is also some leniency towards tests based on QQ- and transformation plots.

As Mecklin & Mundfrom (2004) note no one test is by itself good in all situations.

I will employ a battery of three tests, which I find attractive. Neither of these is true multivariate tests. The only reasonable such test, given in Henze & Zirkler (1990), is hard to implement.

No assessment of statistical power is given; Mecklin & Mundfrom (2005) provides some limited results.

### A.1.1 Multivariate Shapiro-Wilk W

In the univariate case Shapiro & Wilk (1965) provides a test statistic:

$$W = \frac{\left(\sum a_i X_{(i)}\right)^2}{\sum (X_i - \bar{X})^2} \tag{A.1.1}$$

$X_{(\cdot)}$ are the ordered observations, and

$$a^t = \frac{m^t V^{-1}}{\sqrt{m^t V^{-1} V^{-1} m}} \tag{A.1.2}$$

The order expectation $m_i$ and order covariance $V_{ij}$ are not exactly known for all sample sizes and dimensions.

The whole idea of the Shapiro-Wilk W revolves around measuring the linearity in a normal probability plot (QQ-plot). Equation A.1.1 was derived through techniques based on the Gauss-Markov theorem (BLUE see Mardia et al. (1979)).

I will employ approximations to the order statistics $m_i$ and $V_{ij}$ found in Royston (1982):

$$\tilde{m}_i = \Phi^{-1}\left(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}\right) \qquad \text{Where } \Phi^{-1} \text{ is the inverse normal cdf.} \tag{A.1.3}$$

$$\phi = \begin{cases} \frac{\tilde{\vec{m}}^t \tilde{\vec{m}} - 2\tilde{m}_n^2}{1 - 2\tilde{a}_n^2} & n \leq 5 \\ \frac{\tilde{\vec{m}}^t \tilde{\vec{m}} - 2\tilde{m}_n^2 - 2\tilde{m}_{n-1}^2}{1 - 2\tilde{a}_n^2 - 2\tilde{a}_{n-1}^2} & n > 5 \end{cases} \tag{A.1.4}$$

Thus besides the end points ($\tilde{a}_1$, $\tilde{a}_n$ or $\tilde{a}_2$, $\tilde{a}_{n-1}$)

$$\tilde{a}_i = \phi^{-\frac{1}{2}} \tilde{m}_i \tag{A.1.5}$$

for the end points

$$\tilde{a}_1 = \tilde{a}_n = C_n + 0.221157x - 0.147981x^2 - 2.071190x^3 + 4.434685x^4 - 2.706056x^5 \tag{A.1.6}$$

$$\tilde{a}_2 = \tilde{a}_{n-1} = C_{n-1} + 0.042981x - 0.293762x^2 - 1.752461x^3 + 5.682633x^4 - 3.582663x^5 \tag{A.1.7}$$

where $x = n^{-\frac{1}{2}}$ and

$$C_n = (\tilde{\vec{m}}^t \tilde{\vec{m}})^{-\frac{1}{2}} \tilde{m}_n \tag{A.1.8}$$

The exact null distribution of $W$ is only exact known for $n = 3$. In Royston (1982) this is alleviated by the usual standard normalising transform and an approximation of $W$:

$$z = \frac{(1 - W)^\lambda - \mu}{\sigma} \tag{A.1.9}$$

where Royston estimates $\lambda$, $\mu$ and $\sigma$ through simulation for different sample sizes. All on the form

$$\sum_i C_i (\log n - d)^i \tag{A.1.10}$$

$C_i$ and $d$ is tabulated in table A.1.

| | $\lambda$ | | $\mu$ | | $\sigma$ | |
|---|---|---|---|---|---|---|
| | $n \leq 20$ | $n > 20$ | $n \leq 20$ | $n > 20$ | $n \leq 20$ | $n > 20$ |
| $C_0$ | 0.1188980 | 0.4803850 | −0.3754200 | −1.9148700 | −3.1580500 | −3.7353800 |
| $C_1$ | 0.1334140 | 0.3188280 | −0.4921450 | −1.3788800 | 0.7293990 | −1.0158070 |
| $C_2$ | 0.3279070 | 0.0000000 | −1.1243320 | −0.0418321 | 3.0185500 | −0.3318850 |
| $C_3$ | | −0.0241665 | −0.1994220 | 0.1066339 | 1.5587760 | 0.1773538 |
| $C_4$ | | 0.0087970 | | −0.0351367 | | −0.0163878 |
| $C_5$ | | 0.0029896 | | −0.0150461 | | −0.0321502 |
| $C_6$ | | | | | | 0.0038526 |
| d | 3 | 5 | 3 | 5 | 3 | 5 |

Table A.1: Coefficients for equation A.1.10

Critical values of $z$ can now be found in any table of critical values for the standard normal distribution.

**Multivariate**

Until now the test described has been a univariate test. A multivariate extension is given in Royston (1983).

This extension is based on the observation that given multivariate observations

$$X = (x)_{ij} \qquad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, m \end{matrix} \tag{A.1.11}$$

the $w_j = W(\vec{x}_{.,j})$ (as in A.1.1 on page 143) are quite uncorrelated, while $\vec{x}_{.,j}$ and $\vec{x}_{.,k}$ might show moderate or high correlation.

Royston (1983) goes on from where Royston (1982) left. I will however continue from equation A.1.9 on the previous page, which uses the more accurate $W$ of Royston (1992).

$$z_j = \frac{(1 - W(\vec{x_{.,j}}))^\lambda - \mu}{\sigma} \tag{A.1.12}$$

Let

$$k_j = \left( \Phi^{-1}[\tfrac{1}{2}\Phi(-z_j)] \right)^2 \tag{A.1.13}$$

The statistic

$$G = \frac{1}{m} \sum_{j=1}^{m} k_j \tag{A.1.14}$$

will under the null hypothesis be

$$G_0 \sim \frac{1}{m} \chi^2_m \tag{A.1.15}$$

This combination of univariate $w$-statistics is not without critique. See for instance Srivastava & Hui (1987), which proposes a principal component approach instead.

The assertion A.1.15 is only valid if

$$\operatorname{corr}(\vec{x}_{.,j}, \vec{x}_{.,k}) = 0 \qquad \forall_{j \neq k}\, j, k \tag{A.1.16}$$

Between nil and perfect correlation, $G$ will assume a $\frac{1}{e}\chi^2_e$ distribution with an *equivalent degrees* of freedom. By noting

$$\operatorname{Var}\left( \frac{1}{e} \chi^2_e \right) = \frac{2e}{e^2} \tag{A.1.17}$$

and noting that $k_i$ is the square of a standard normal distributed variable,

therefore $\mathrm{Var}(k_i) = 2$

$$\begin{aligned}
\mathrm{Var}(G) &= \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} k_i\right) \\
&= \frac{1}{m^2}\,\mathrm{Var}\left(\sum_{i=1}^{m} k_i\right) \\
&= \frac{1}{m^2}\left(\sum_{i=1}^{m}\mathrm{Var}(k_i) + 2\sum_{i<l}^{m}\mathrm{Cov}(k_i,k_l)\right) \\
&= \frac{2m + 2\sum_{i<l}^{m}\mathrm{Cov}(k_i,k_l)}{m^2}
\end{aligned} \tag{A.1.18}$$

$e$ can be found:

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{e}\chi_e^2\right) &= \mathrm{Var}(G) \\
\frac{2e}{e^2} &= \frac{2m + 2\sum_{i<l}^{m}\mathrm{Cov}(k_i,k_l)}{m^2} \\
\frac{e}{2} &= \frac{m^2}{2m + 2\sum_{i<l}^{m}\mathrm{Cov}(k_i,k_l)} \\
e &= \frac{2m^2}{2(m + \sum_{i<l}^{m}\mathrm{Cov}(k_i,k_l))} \\
e &= \frac{m^2}{m + \sum_{i<l}^{m}\mathrm{Cov}(k_i,k_l)} \\
e &= \frac{m}{1 + \frac{1}{m}\sum_{i<l}^{m}\mathrm{Cov}(k_i,k_l)}
\end{aligned} \tag{A.1.19}$$

Until now $\mathrm{Cov}(k_i, k_l)$ has been a theoretical quantity. Let

$$\hat{e} = \frac{m}{1 + \frac{1}{m}\sum_{i<l}^{m}\hat{c}_{il}} \tag{A.1.20}$$

Through simulation Royston relates the sample correlation matrix $R = (r)_{il}$ to $\hat{c}_{il}$.

$$\hat{c}_{il} = \begin{cases} 1 & i = l \\ g(r_{jl}) & i \neq l \\ 0 & r_{jl} = 0 \end{cases} \tag{A.1.21}$$

where

$$\begin{gathered}
g(\rho, n) = \rho^{\lambda}\left(1 - \frac{\mu}{v(n)}(1-\rho)^{\mu}\right) \\
\mu \text{ and } \lambda \text{ as in table A.1 on page 144} \\
v(n) = 0.21364 + 0.015124\log^2(n) - 0.0018034\log^3(n)
\end{gathered} \tag{A.1.22}$$

The final statistic will be:

$$\begin{aligned} H &= \frac{\hat{e}}{m} \sum_{j=0}^{m} k_j \\ &\sim \chi^2_{\hat{e}} \end{aligned} \tag{A.1.23}$$

### A.1.2 Koziol's Cramér-von Mises type test

This statistic is based on the empirical distribution and derived through some neat results in empirical process theory. These results will not be brought into play here. Interested readers should refer to the references contained in Koziol (1982).

Generally the Cramér-von Mises statistic is

$$W_n^2 = n \int_0^1 \{F_n(X) - X\}^2 \, dx \tag{A.1.24}$$

where $F_n(X)$ is the empirical distribution function

$$F_n(X) = \frac{1}{n} \sum_{i=1}^{n} I(X_i < X) \tag{A.1.25}$$

$I(\cdot)$ is the indicator function.

The statistic A.1.24 has an alternative form

$$W_n^2 = \sum_{i=1}^{n} \left( X_i - \frac{i - \frac{1}{2}}{n} \right)^2 + \frac{1}{12n} \tag{A.1.26}$$

which is the one used in practice. Critical values for $W_n^2$ are tabulated. It should be noted that Cramér-von Mises type statistics have higher statistical power than both Pearson's chi-square and the Kolmogorov-Smirnov type statistics.

Koziol (1982) uses the squared Mahalanobis distance

$$Y_i = (X_i - \bar{X})^t S^{-1} (X_i - \bar{X}) \tag{A.1.27}$$

which is chi-square distributed with k-degrees of freedom, if $X_i$ is k-dimensional and normal distributed. $\bar{X}$ and $S$ are the usual sample estimates for the mean and covariance.

Let $Z_i = F_k(Y_i)$, where $F_k(\cdot)$ is the cumulative function of the chi-square distribution.

Now let the Cramér-von Mises type statistic be:

$$J_n = \frac{1}{12n} + \sum_{i=1}^{n} \left( Z_i - \frac{i - \frac{1}{2}}{n} \right)^2 \tag{A.1.28}$$

$Z_{(\cdot)}$ is the ordered $Z_\cdot$. Asymptotic critical values of $J_n$ are known.

I will use critical values found by drawing $n$ $k$-dimensional multivariate normal deviates, one million times, evaluating $J_n$ and fitting natural cubic splines for the reverse look-up of the distribution of $J_n$. For an example see figure A.1.

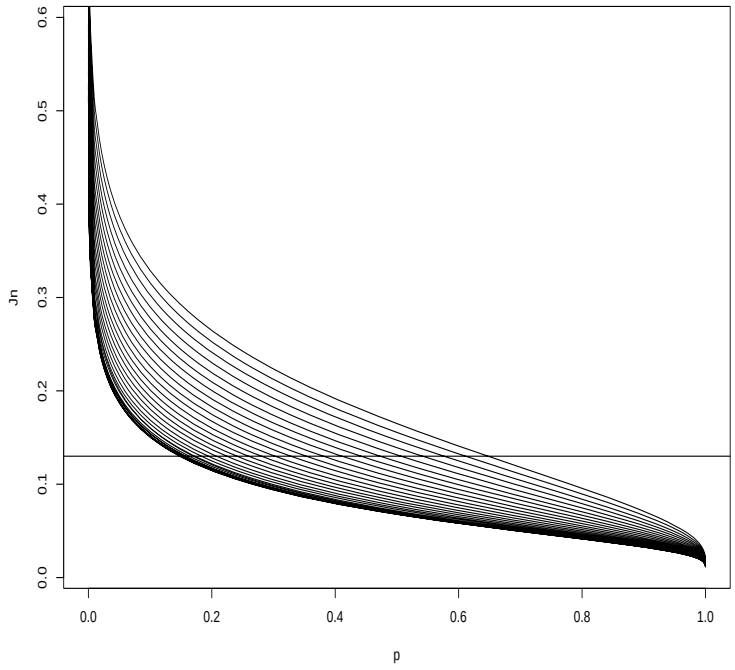


Figure A.1: Example of reverse look up for $J_n = 0.13$ for different $k = 1, \ldots, 30$ $n = 90$, $p$-values ranging from less than 0.2 for $k = 3$ to slightly less than 0.7 for $k = 30$

### A.1.3 Tests based on multivariate skewness and kurtosis

Skewness and kurtosis have often been used as loose measures of normality as they are easy to visualise. A few examples are given in figure A.2 on the next page.

Besides being the third standardised central moment, skewness is a measure of asymmetry. Kurtosis as the fourth standardised central moment is a measure of the sharpness of peaks and elongateness of the tails.

$$\begin{aligned} \mu_i &= \mathrm{E}([X-\mu]^n) \quad & \sigma &= \sqrt{\mathrm{E}[(X-\mathrm{E}\ X)]^2} \\ \gamma_3 &= \frac{\mu_3}{\sigma^3} & \gamma_4 &= \frac{\mu_4}{\sigma^4} \end{aligned} \tag{A.1.29}$$

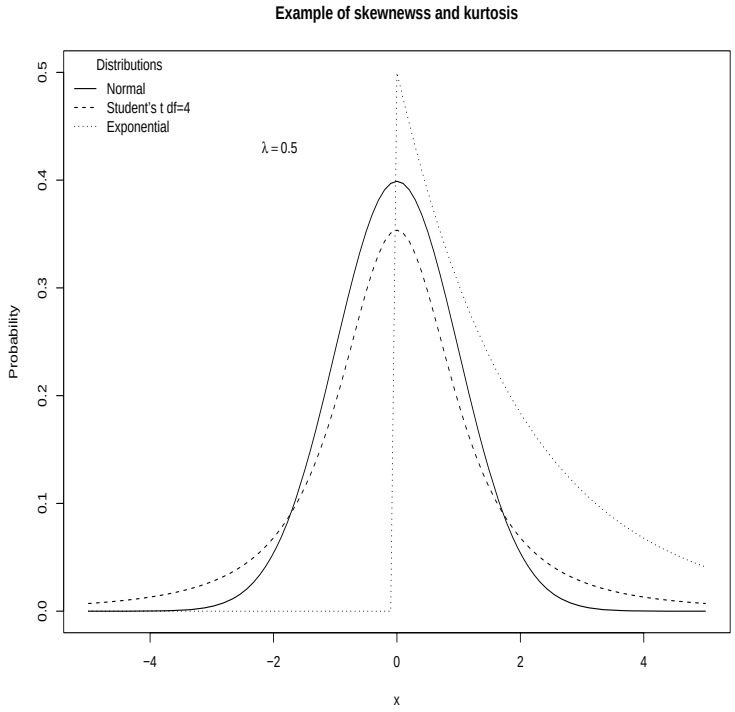


Figure A.2: Univariate example of skewness and kurtosis: The normal distribution has neither, Student's t-distribution has kurtosis, and the exponential distribution has both.

For univariate observations Gnanadesikan (1977) notes the use of the following sample coefficients for skewness and kurtosis

$$\sqrt{b}_1 = \frac{\sqrt{n}\sum_i (x_i - \bar{x})^3}{[\sum_i (x_i - \bar{x})^2]^{3/2}} \tag{A.1.30}$$

$$b_2 = \frac{n\sum_i (x_i - \bar{x})^4}{[\sum (x_i - \bar{x})^2]^2} \tag{A.1.31}$$

for which tables of critical values exist.

Mardia (1970) and Mardia (1974) takes this further with multivariate extensions.

The sample multivariate skewness coefficient:

$$b_{1,p} = \sum_{r,s,t}^{p} \sum_{r',s',t'}^{p} s_{rr'} s_{ss'} s_{tt'} M_{rst} M_{r's't'} \tag{A.1.32}$$

with $p = \dim(X_\cdot)$, $S^{-1} = (s)_{ij}$ and

$$M_{rst} = \frac{1}{n}\sum_{i=1}^{n} (x_{ri} - \bar{x}_r)(x_{si} - \bar{x}_s)(x_{ti} - \bar{x}_t) \tag{A.1.33}$$

The sample multivariate kurtosis coefficient is the sample mean of the squared Mahalanobis distance:

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}[(\vec{x}_{.i} - \bar{X})^t S^{-1}(\vec{x}_{.i} - \bar{X})]^2 \tag{A.1.34}$$

Mardia gives two statistics based on the above coefficients

$$A = \frac{b_{1,p}}{6} \qquad \sim\chi^2_d \tag{A.1.35}$$

$$B = \frac{b_{2,p} - p(p+2)}{\sqrt{\frac{8p(p+2)}{n}}} \qquad \sim \mathrm{N}(0,1) \tag{A.1.36}$$

where $d$, the degrees of freedom for the $A$ statistic are

$$d = \frac{p(p+1)(p+2)}{6} \tag{A.1.37}$$

**An example**

| Model | Koziol | | Mardia | | | | | Royston | | |
|---|---|---|---|---|---|---|---|---|---|---|
| complexity | $J_n$ | p | A | df | p | B | p | H | e | p |
| low | 0.12 | 0.20 | 3.86 | 10 | 0.95 | −1.79 | 0.07 | 46.10 | 2.25 | 0.00 |
| high | 0.04 | 0.82 | 27.82 | 35 | 0.80 | −1.69 | 0.09 | 4.96 | 4.17 | 0.31 |

Table A.2: An example of normality tests. Notice how only the Shapiro-Wilk type test of Royston discerns between the two complexities.

## A.2 Tailor-made tests for classifiers

Until now I have looked into assessing the underlying assumptions of the LDA classifier.

This is in itself important, however the probability of misclassification is of more practical importance.

Lachenbruch & Mickey (1968) and Page (1985) investigate several methods for the estimation of error rates, which is the realisation of the probability of misclassification.

Historically the classifier employed is the $W$ classifier rule of Anderson (1951). Assuming two classes distributed as $\mathrm{N}(\mu_i, \Sigma)$ $i \in \{1, 2\}$ :

$$W = \vec{X}_{\text{New}}^T S^{-1}(\bar{\vec{X}}_1 - \bar{\vec{X}}_2) - \tfrac{1}{2}(\bar{\vec{X}}_1 + \bar{\vec{X}}_2)^T S^{-1}(\bar{\vec{X}}_1 - \bar{\vec{X}}_2) \tag{A.2.1}$$

$\bar{\vec{X}}_i$ is the class mean, and $S$ is the pooled sample covariance. $\vec{X}_{\text{New}}$ represents new observations.

A sample $\vec{X}_{\text{New}}$ is classified as

$$\begin{cases} \text{Class 1} & W > c \\ \text{Class 2} & W < c \end{cases} \tag{A.2.2}$$

where $c$ is decided by the prior probabilities and the loss function used. Lachenbruch (1968) uses

$$c = \log\left(\frac{q}{1-q}\right) \tag{A.2.3}$$

where $q$ is the prior probability of belonging to class 1.

This classifier is more economical to employ than the LDA classifier, and is somewhat looser in its assumptions.

Both LDA and Anderson's rule assume multivariate normal distributions for the class populations, and are close relatives. The distribution of $W$ is intangible.

In its basic formulation Anderson's considers only two classes. Extensions to multiple classes exist for both $W$, and the forthcoming methods of estimating the probability of misclassification.

### A.2.1 Estimating the probability of misclassification

Lachenbruch & Mickey (1968) note that some methods that I mention in section 3.2 already existed in 1968, but that neither of them was satisfactory for small sample sizes compared with the number of covariates (parameters).

Even today, with refinement of the above mentioned non-parametric techniques, the parametric methods introduced shortly, will have applications. A good example that will benefit by these methods, is the cancer classification experiment found in Golub et al. (1999).

Given the true class means ($\mu_1$, $\mu_2$) and the true covariance ($\Sigma$) the probability of

misclassification in Anderson's rule can be found:

$$\begin{aligned} P_1 &= \Pr[W(\vec{X}) < c | \vec{X} \in \{\text{Class 1}\}] \\ &= \Pr\left( \frac{\vec{X}^T a - \vec{\mu}_1^T a}{\sqrt{a^T \Sigma a}} \leq \frac{\frac{1}{2}(\bar{\vec{X}}_1 + \bar{\vec{X}}_2)^T a - \vec{\mu}_1^T a}{\sqrt{a^T \Sigma a}} \right) \\ &= \Phi\left( \frac{\frac{1}{2}(\bar{\vec{X}}_1 + \bar{\vec{X}}_2)^T a - \vec{\mu}_1^T a}{\sqrt{a^T \Sigma a}} \right) \end{aligned} \tag{A.2.4}$$

where $a = S^{-1}(\bar{\vec{X}}_1 - \bar{\vec{X}}_2)$, see section 6.4 of Morrison (1976) for a fuller account. Likewise:

$$P_2 = \Phi\left( \frac{\vec{\mu}_2^T a - \frac{1}{2}(\bar{\vec{X}}_1 + \bar{\vec{X}}_2)^T a}{\sqrt{a^T \Sigma a}} \right) \tag{A.2.5}$$

**The U-method**

The approach in equation A.2.4 is theoretical and depends on generally unknown terms ($\mu_1$, $\mu_2$ and $\Sigma$).

Lachenbruch & Mickey (1968) give the U-method, and get past this. Let

$$\begin{aligned} \mathrm{E}(W_1) &= \vec{\mu}_1 S^{-1}(\bar{\vec{X}}_1 - \bar{\vec{X}}_2) - \tfrac{1}{2}(\bar{\vec{X}}_1 + \bar{\vec{X}}_2)^T S^{-1}(\bar{\vec{X}}_1 - \bar{\vec{X}}_2) \\ &= [\vec{\mu}_1 - \tfrac{1}{2}(\bar{\vec{X}}_1 + \bar{\vec{X}}_2)]^T S^{-1}(\bar{\vec{X}}_1 - \bar{\vec{X}}_2) \end{aligned} \tag{A.2.6}$$

and

$$\mathrm{Var}(W_1) = (\bar{\vec{X}}_1 - \bar{\vec{X}}_2)^T S^{-1} \Sigma S^{-1} (\bar{\vec{X}}_1 - \bar{\vec{X}}_2) \tag{A.2.7}$$

Noting that

$$(-1)\frac{\mathrm{E}(W_1)}{\sqrt{\mathrm{Var}(W_1)}} \sim \mathrm{N}(0,1) \tag{A.2.8}$$

one might set:

$$P_1 = \Phi\left( (-1)\frac{\mathrm{E}(W_1)}{\sqrt{\mathrm{Var}(W_1)}} \right) \tag{A.2.9}$$

For the estimate $\hat{P}_1$, one needs:

$$\bar{W}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} W(\vec{x}_i) \qquad \vec{x}_i \in \{\text{Class 1}\} \tag{A.2.10}$$

$$S(W)_1 = \frac{\sum_{i=1}^{n_1} \left( W(\vec{x}_i) - \bar{W}_1 \right)^2}{n_1 - 1} \tag{A.2.11}$$

The estimate of the error rate is now:

$$\hat{P}_1 = \Phi\left( \frac{-\bar{W}_1}{\sqrt{S(W)_1}} \right) \qquad \hat{P}_2 = \Phi\left( \frac{\bar{W}_2}{\sqrt{S(W)_2}} \right) \tag{A.2.12}$$

For details see Lachenbruch & Mickey (1968) and section 6.4 of Morrison (1976).

**The D-method**
With similar arguments as for the U-method:

$$\begin{aligned} P_1 &= \Pr[W(\vec{X}) < c | \vec{X} \in \{\text{Class 1}\}] \\ &= \Pr[(\vec{\mu}_1 - \vec{\mu}_2)^T \Sigma^{-1} X \ \leq \ \tfrac{1}{2}(\vec{\mu}_1 - \vec{\mu}_2)^T \Sigma^{-1} (\vec{\mu}_1 + \vec{\mu}_2)] \\ &= \Phi\left( -\tfrac{1}{2} \sqrt{(\vec{\mu}_1 - \vec{\mu}_2)^T \Sigma^{-1} (\vec{\mu}_1 - \vec{\mu}_2)} \right) \end{aligned} \tag{A.2.13}$$

The sample variant is:

$$\hat{P}_1 = \Phi\left( -\tfrac{1}{2} \underbrace{\sqrt{(\bar{\vec{x}}_1 - \bar{\vec{x}}_2)^T \Sigma^{-1} (\bar{\vec{x}}_1 - \bar{\vec{x}}_2)}}_{=D} \right) = \hat{P}_2 \tag{A.2.14}$$

where $D$ is the squared Mahalanobis distance. For details see Lachenbruch & Mickey (1968) and section 6.4 of Morrison (1976).

**The O-method**
In a way not too easy to follow, Okamoto (1963) gives an asymptotic expansion of equation A.2.4 on the previous page and ends up with

$$\begin{aligned} \hat{P}_1 &= \Phi\left(-\frac{D}{2}\right) + \frac{a_1}{n_1} + \frac{a_2}{n_2} + \frac{a_3}{n_1 + n_2} \\ &\quad + \frac{b_{11}}{n_1^2} + \frac{b_{22}}{n_2^2} + \frac{b_{12}}{n_1 n_2} + \frac{b_{13}}{n_1^2 (n_1 + n_2)} + \frac{b_{23}}{n_2^2 (n_1 + n_2)} + \frac{b_{33}}{(n_1 + n_2)^2} \\ \hat{P}_2 &= \Phi\left(-\frac{D}{2}\right) + \frac{a_2}{n_1} + \frac{a_1}{n_2} + \frac{a_3}{n_1 + n_2} \\ &\quad + \frac{b_{22}}{n_1^2} + \frac{b_{11}}{n_2^2} + \frac{b_{12}}{n_1 n_2} + \frac{b_{23}}{n_1^2 (n_1 + n_2)} + \frac{b_{13}}{n_2^2 (n_1 + n_2)} + \frac{b_{33}}{(n_1 + n_2)^2} \end{aligned} \tag{A.2.15}$$

where $D$ is found in equation A.2.14 and the coefficients $a_i$ and $b_{ij}$ are tabulated for various input values.

**Shrunken methods**

Lachenbruch & Mickey (1968) also argue for shrunken versions of the D- and O-methods. D in equation A.2.14 on the previous page is replaced by

$$\mathrm{DS} = \frac{(n_1 + n_2 - p - 3)D}{n_1 + n_2 - 2} \tag{A.2.16}$$

where $p$ is the number of covariates.

The shrunken D-method DS is:

$$\hat{P}_1 = \Phi(-\tfrac{1}{2}\,\mathrm{DS}) = \hat{P}_2 \tag{A.2.17}$$

Similarly the shrunken O-method OS:

$$\hat{P}_1 = \hat{P}_2 = \Phi(-\tfrac{1}{2}\,\mathrm{DS}) + \ldots \tag{A.2.18}$$

**An example**

| Classifier complexity | $O$ | $OS$ | $D$ | $DS$ | $U_1$ | $U_2$ | holdout error |
|---|---|---|---|---|---|---|---|
| low | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.38 | 0.32 |
| high | n/a | n/a | 0.34 | 0.55 | 1.00 | 0.00 | 0.32 |

Table A.3: An example of tailor-made tests, the holdout error is described in section 3.2.2 on page 37.

In table A.3 an example of tailor-made tests on the data of Golub et al. (1999) is given. Notice how Okamoto's method fails in both cases. This is due to data starvation. The Okamoto method is asymptotic. It is not clear at which data level this asymptotics kick in. Among the other methods, considering the prior distribution, the U method is the only to show reasonable consistent behaviour.

Some of this behaviour can be attributed to *the curse of dimensionality*. The methods that fail, depend on inverting an estimate of the covariance matrix. Matrix inversion is very prone to numerical instabilities and at this data level the estimate of the covariance matrix is poor. Actually the estimate of the covariance matrix is not positive definite (nor non-singular) which can be attributed to sampling error (*the curse of dimensionality*).

This can be worked around either by increasing the data level, using back-solving inversion techniques, or regularising the covariance estimate (akin to ridge-regression and partial least squares).

# APPENDIX B

# Additional results

| | level | # comp. | error | low | high |
|---|---|---|---|---|---|
| 1 fk22 | 5 | 12 | 0.179 | 0.085 | 0.321 |
| 2 fk22 | 4 | 12 | 0.180 | 0.081 | 0.322 |
| 3 d16 | 4 | 9 | 0.180 | 0.093 | 0.312 |
| 4 fk22 | 3 | 10 | 0.180 | 0.082 | 0.310 |
| 5 fk22 | 6 | 10 | 0.180 | 0.086 | 0.310 |
| 6 fk22 | 7 | 10 | 0.181 | 0.085 | 0.320 |
| 7 fk22 | 4 | 10 | 0.181 | 0.088 | 0.308 |
| 8 fk22 | 3 | 11 | 0.182 | 0.088 | 0.322 |
| 9 fk22 | 4 | 13 | 0.182 | 0.087 | 0.321 |
| 10 d16 | 7 | 9 | 0.182 | 0.093 | 0.321 |

Table B.1: Visu shrink: Fontainebleau dataset

| | | error | |
|---|---|---|---|
| Dataset | # components | test | validation |
| Pavia | 12 | 0.08 | 0.86 |
| National Mall | 23 | 0.12 | 0.01 |
| Fontainebleau | 4 | 0.33 | 0.31 |

Table B.2: PCA: Test- vs validationset

# APPENDIX C

# Overview of wavelet families

| Family | Order |
|---|---|
| Haar | 2 |
| Daubechies (d) | 4, 6, 8, 10, 12, 14, 16, 20 |
| Least asymmetric (Daubechies) (la) | 8, 10, 12, 14, 18, 20 |
| Minimum bandwidth (mb) | 4, 8, 16, 24 |
| FejerKorovkin (fk) | 4, 6, 8, 14, 22 |
| Coiflet (Daubechies) (c) | 6, 12, 18, 24,30 |
| Battle-Lemarie (bl) | 14, 18, 20 |
| Biorthogonal (bs3.1) | 3.1 |
| W4 | |

Table C.1: Wavelets used

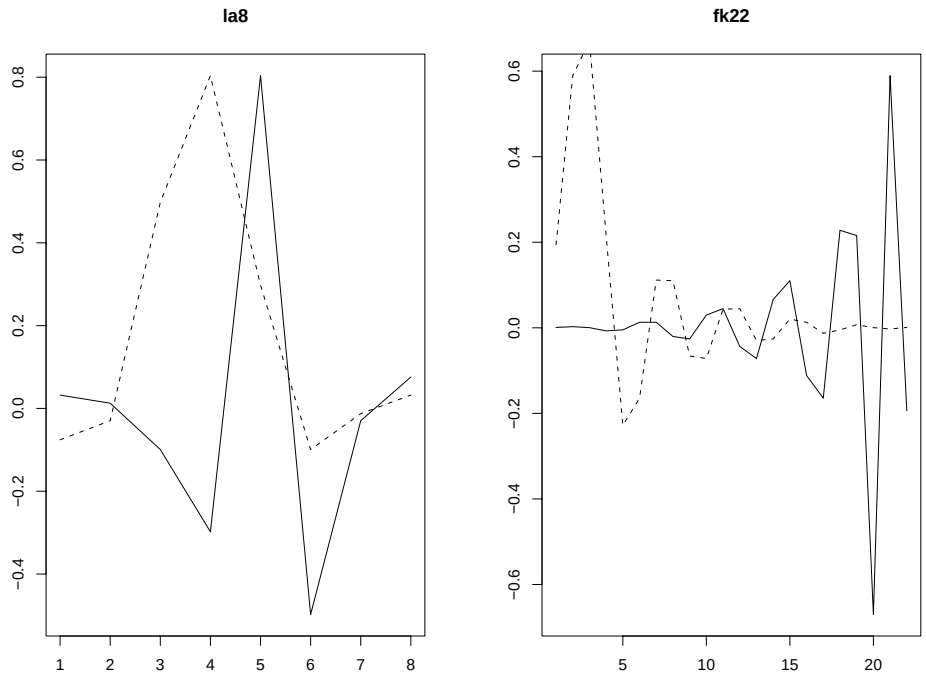


Figure C.1: The mother (solid line) and father (dashed line) wavelets of 'la8' and 'fk22'