INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0623

A MULTISCALE APPROACH FOR STATISTICAL CHARACTERIZATION OF FUNCTIONAL IMAGES

A. ANTONIADIS, J. BIGOT and R. von SACHS

http://www.stat.ucl.ac.be

A multiscale approach for statistical characterization of functional images

Anestis Antoniadis,
Laboratoire IMAG-LMC, University Joseph Fourier,
BP 53, 38041 Grenoble Cedex 9, France.

Jéremie Bigot,
Department of Probabilty and Statistics,
University Paul Sabatier,
31062 Toulouse Cedex 9, France.

and

Rainer von Sachs,
Institute of statistics, Université catholique de Louvain,
Voie du Roman Pays, 20,
B-1348 Louvain-la-Neuve, Belgium.

October 31, 2006

Authors' Footnote:

Anestis Antoniadis is Professor at Laboratoire IMAG-LMC, University Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France. email: `Anestis.Antoniadis@imag.fr`

Jérémie Bigot is Associate Professor in Department of Probabilty and Statistics, University Paul Sabatier, Toulouse, France. email: `Jeremie.Bigot@math.ups-tlse.fr`

Rainer von Sachs is Professor, Institute of statistics, Université catholique de Louvain, Belgium. email: `rvs@stat.ucl.ac.be`

Abstract

In this paper we use an approach of spatial multiscales for an improved characterization of functional pixel intensities of images. Examples are numerous such as temporal dependence of brain response intensities measured by fMRI or frequency dependence of NMR spectra measured at each pixel. The overall goal is to improve the misclassification rate in clustering (unsupervised learning) of the functional image content into a finite but unknown number of classes. Hereby we adopt a non-parametric point of view to reduce the functional dimensionality of the observed pixel intensities, modelled to be of a very general functional form, by a combination of "aggregation" and truncation techniques. Clustering is applied via an EM-algorithm for estimating a Gaussian mixture model in the domain of the discrete wavelet transform of the pixel intensity curves.

We show improvements of our multiscale method, based on complexity-penalised likelihood estimation for *Recursive Dyadic Partitioning* of the image, over existing *monoscale* approaches, by simulated and real data examples, and we give some theoretical treatment of the resulting misclassification rate in the simplified set-up of the "horizon" model of two classes.

KEYWORDS: Mixture model; Recursive dyadic partition; Multiresolution trees; Aggregation; Wavelets.

1. INTRODUCTION

In this paper we use an approach of spatial multiscales for an improved characterization of functional pixel intensities of images. Examples are numerous such as temporal dependence of brain response intensities measured by fMRI or frequency dependence of NMR spectra measured at each pixel (voxel). Another example is satellite remote sensoring images of landscapes. The overall goal is to improve the misclassification rate in clustering of the functional image content into a finite but unknown number of classes. That is we place ourselves into the context of an *unsupervised* learning approach. Hereby we adopt a nonparametric point of view to reduce the functional dimensionality of the observed pixel intensities. Note that we model the pixel intensities to be of a very general functional form and use wavelet thresholding techniques to be able to treat a large scale of functions of possibly very low regularity. We combine statistical aggegration of nonlinear projection estimators and truncation tests on the significance of the resulting coefficients with respect to their importance of discriminating between different class memberships. Our approach is thus opposed to commonly used parametric feature extraction based on a priori knowledge on the nature of the functional response.

Our point of reference for comparisons are *monoscale* statistical models which are typically used to clean the map of noise and to help extract structure in the underlying measurements: they work at a pixel level resolution and result in a low degree of aggregation of the information underlying the data, since the appropriate choice of scale usually varies with spatial location. One prominent example for a functional pixel-scale approach based on wavelet methods is the work by Whitcher *et al.* (2005) which partially motivated our own work. The authors examine MRI time series experiments (e.g., brain responses to pharmacological stimuli) that produce statistical data in 2 (or 3) spatial dimensions which evolve over time. Their nonparametric approach consists in grouping pixels (or voxels) with similar time courses based on the Discrete Wavelet Transform representation of each time course thus giving a localised pixel information over time and spatial (mono-) scale. Note that such an approach is inherently different from using cross-correlation in the pixel domain. On top of this, discarding scales in the DWT that are associated with noise, leads to a smoothed (denoised) reconstruction. To produce a finite number of pixel clusters some clustering is applied to the remaining wavelet coefficients. The choice of the number of cluster classes is done using a recent proposal by Sugar and James (2003). Evaluation of this wavelet-based cluster analysis is done using two representative classes.

Monoscale approaches are on the basis of a variety of methods that have been proposed, including those based on maximum likelihood, decision trees, nearest neighbors, and neural networks. Consider, for instance, the example displayed in Figure 1 in which one might characterize the classes in a given rectangular region. In Figure 1(a), the image consists in individual pixels that are labeled according to underlying circular regions of different intensity (different classes), while Figure 1(b) displays the same image with additive noise. The purpose of a statistical classification method is to infer for each pixel in the noisy image the label of its class. In looking at Figure 1(c), a monoscale, i.e. pixel-scale, approach leads to a segmentation restricted to a spatial resolution of the original pixels, with many false positives. A drawback of the monoscale approach, as already noted by Bouman and Shapiro (1994), is that it uses uniform pixel sizes across the image and therefore does not take into account local spatial variation, i.e. the characteristic shape and size of patches for various classes.

On the contrary, the method developed in this paper is guided by an adaptive choice of spatial scale at a given location. The main idea is that each potential class has its own spatial resolution within a scene. We use a multiscale method, based on *recursive dyadic partitioning* (RDP) of the image to adaptively choose the locally best scale for clustering. RDP is a by now widely used method in image processing (see e.g. Donoho 1997 or Kolaczyk *et al.* 2005). Starting our functional description of the pixel intensities on the finest resolution scale of a dyadic tree of square quads we furnish a candidate model through a statistical likelihood, assuming statistical independence

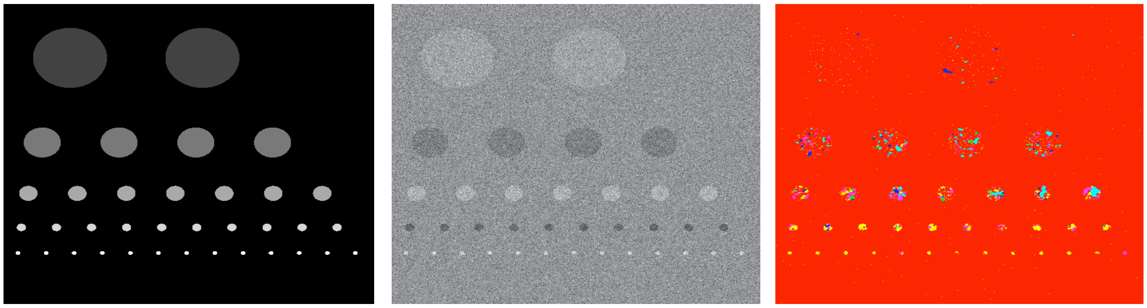


Figure 1.1: Example: Monoscale classification of pixels distributed in 5 classes of different intensities.

between signal measurements among pixels. We then use a method of penalization to encourage aggregation of pixels where useful: similar regions of pixels likely to belong to the same cluster class will be combined to a larger quad. A suitable penalty prevents from ending up with a tree with too many splits, i.e. too many unnecessary small regions. Our RDP-based approach on combining adjacent spatial information replaces correlation-model based approaches such as "Hidden Markov" models which are frequent in the literature (see, e.g., Malfait and Roose 1997, or Choi and Baraniuk 2001).

Our approach is a functional one, we recall that the pixel intensities are actually curves which are modelled as signal plus noise. As there are many, and as we need, for the subsequent clustering step, to appropriately combine the information in all of these curves to a common one, we choose a procedure of dimension reduction which allows combination with modern denoising techniques. This leads us to an "aggregation" technique (Bunea *et al.* 2006) of nonlinear wavelet threshold estimators for each of the intensity curves. As a consequence of the not sufficiently reduced dimensionality after aggregation we have to subsequently apply Neyman truncation tests (Fan 1996) on the significance of the resulting thresholded coefficients. Here "significance" is to be understood with respect to their ability of discriminating between different class memberships. To cope with the multiplicity of the test we use ideas based on the False Discovery Rate (FDR) to control the level of the test. Using wavelet threshold estimators allows to combine three attractive features: first, wavelets provide orthogonal bases they are ideal for dimension reduction by means of sparse representations; in addition they allow for powerful denoising. Finally, wavelets allow to easily include not only heteroscedasticity but also serial temporal correlation into the model we are going to use, without losing independence of empirical wavelet coefficients over locations within scales, i.e. our time-discretized curves can actually be serially correlated and hence be also seen as time series realisations.

We base our clustering on the application of an EM-algorithm for estimating a Gaussian mixture model in the domain of the discrete wavelet transform of the pixel intensity curves. For this clustering step we use now that we have one common "reduced" dimensionality for all the curves, in order to encode the location of the resulting wavelet coefficients determined in the preceeding dimension reduction step. In order to include the choice of the number of cluster classes from the data we use the more refined EM-algorithm of Law *et al.* (2004). With this we also benefit from an additional sparsification via selection of the most important "features", i.e. those coefficients that carry the most important signal information for discrimination.

The rest of the paper is organised as follows. In the following section we describe in detail our multiscale set-up including our model for both the pixel intensity curves and the Gaussian mixture model in the wavelet domain used for the clustering step. A "nut-shell" description of our complete algorithm is completed by some more details on the EM-Algorithm used for feature selection and clustering, including the choice of the number of clusters. Section 3 gives methodological and theoretical backing-up of the properties of the aggregation estimator and the additional dimension reduction step by Neyman truncation tests combined with FDR. Concering the theoretical treatment of the encountered problems, we apply the "horizon" model of Donoho (1999), already implicitly being used by Korostolev and Tsybakov (1993) in their seminal work. Paralleling a similar derivation to be found in Kolaczyk *et al.* (2005), we show that our method, from an asymptotic point of view, is able to correctly identify pixels belonging to two different cluster classes. In Section 4 we investigate the performance of our algorithm by simulated and real data examples, including a comparison with a monoscale scheme (see above). We conclude with a discussion section, where we also sketch possible extensions left for future work. All proofs are deferred to an appendix section.

2. MULTIRESOLUTION TREE-STRUCTURED SPATIAL FRAMEWORK AND ADOPTED METHODOLOGY.

In this section we describe our spatial multiscale approach using recursive dyadic partitioning and our model of functional pixel intensities. This includes the Gaussian mixture model we use in the wavelet domain for our clustering step. We then describe step by step our algorithm composed of a dimension reduction, a clustering and a spatial aggregation step. The section is completed by some more details on the EM-Algorithm used for feature selection and clustering, including the choice of the number of clusters.

We begin by preparing the set-up of spatial multiscales. Consider the notion of consecutive spatial scales through the use of windows of different sizes, starting from the original pixels and then moving up to windows of sizes 2 x 2, 4 x 4, 8 x 8, and so on. In fact, a commonly used multiscale data structure in image processing that precedes wavelets is the quad-tree (see also Donoho, 1997). At the coarsest scale the entire image is treated as a single "pixel", and at the next finer scale, the image can be split into four square quads. Each of these quads can be further split into four quads, and so on, until the finest scale is reached. This *recursive dyadic partitioning* (RDP) results into a hierarchical presentation of the image. For an image with $2^J \times 2^J$ pixels, a maximal $J+1$ layer of representations is produced. If one of these layers is used in its entirety, the representation is *monoscale* (i.e. with uniform quad sizes). However, the idea is to exploit the nesting of these layers to allow the quad size to vary across the image.

Denote the class of all potential models as the set of all RDPs, i.e. the collection of all possible prunings of the complete tree. We want to quantify the goodness-of-fit of a candidate model through a statistical likelihood, assuming statistical independence between signal measurements among pixels. We will then use a method of penalization to encourage aggregation of pixels where useful; see Section 2.2 below.

2.1 Formal model description, including a motivating example

Consider a finite spatial region of an image, say the unit square $[0,1]^2$. We have a $N = r \times r$ discretization of this region consisting of pixels $I_i,\ i = 1, \ldots, N$. For each pixel we have a time history over times $t \in \{t_1, t_2, \ldots, t_n\}$, $n = 2^{J_n}$ for some integer J_n. Focus now on just one pixel, say I, and denote by $x_I(t)$ its intensity and by $\mathbf{x}_I = (x_I(t_1), \ldots, x_I(t_n))^T$ the time history (i.e., the discretized sample path of the measurements for pixel I). Assuming that the pixel belongs to one among L possible classes, we write

$$x_I(t) \ = \ f^\ell(t) + \varepsilon^\ell(t) \ , \qquad t = t_1, \ldots, t_n,\ n = 2^{J_n} \ ,$$

where f^ℓ is the underlying mean intensity for each pixel in class ℓ, $1 \leq \ell \leq L$, and $\varepsilon^\ell(t)$ is a zero-mean noise. To be fairly general, we allow for weak serial dependence in the noise process, and a convenient condition to match with our model in the wavelet domain below is to assume that the autocovariances of the noise are absolutely summable.

The symbol $c(i)$ is used to denote the class that is assigned to pixel i, and each of those $c(i)$ takes on some values within the set of pure classes $\{c_1, \ldots, c_L\}$ where naturally $L < N$. For simplicity in the sequel we label the pure classes with $\{1, \ldots, L\}$.

Having introduced this notation, a rewriting of the above model for the intensity $x_i(t)$ of the $i-$th pixel which is more convenient for the future, is

$$x_i(t) \ = \ f^{c(i)}(t) + \varepsilon^{c(i)}(t) \ , \qquad t = t_1, \ldots, t_n, \ n = 2^{J_n} \ ,$$

or, yet,

$$x_i(t) \ |\{c(i) = \ell\} \ = \ f^{\ell}(t) + \varepsilon_i(t) \ , \qquad t = t_1, \ldots, t_n, \ n = 2^J \ , \ 1 \leq \ell \leq L \ . \tag{1}$$

Starting from this model (1) in the functional domain, we apply to each of these N discretized curves, say $\mathbf{x}_i, \ i = 1, \ldots, N$, a discrete orthogonal wavelet transformation over time t, resulting into the following coefficient model for pixel i on scale $0 \leq j \leq J-1 \ , \ k = 0, \ldots 2^j - 1$:

$$w^i_{jk} \ |\{c(i) = \ell\} \ = \ \theta^{\ell}_{jk} + \varepsilon^i_{jk} \ , \ 1 \leq \ell \leq L < N \ , \tag{2}$$

where we add now a parametric assumption for the model densities:

$$\varepsilon^i_{jk} \ |\{c(i) = \ell\} \ \sim \ \mathcal{N}(0, \sigma^2_{j,\ell}) \ .$$

Note that the variances $\sigma^2_{j,\ell}$ of the empirical wavelet coefficients of pixel i, given the class label ℓ, are modelled to be scale but not location dependent. This is in accordance with our model in the time domain which allows for serially correlated errors $\varepsilon_i(t)$ with absolutely summable autocovariances (see also Johnstone and Silverman 1997).

The interpretation of this time-dependent intensity model is that the intensity functions per pixel can vary over time, e.g. due to, for medical images, a change of concentration of the blood flow or the decay of the reaction to the stimulus. Another example where the maximum number of classes is actually known and which will deserve a special attention in our numerical examples section, comes from magnetic resonance spectroscopy with the purpose of diagnosis of brain tumors. Since this example particularly fits our framework, we now describe it with more details. When a brain tumor has been diagnosed on a patient, the next very important step is to identify the type of tumor and possibly the tumor grade. The tumor type is dependent on the type of cell the tumor originates from. Frequently one encounters three different types of tumor (oligodendroglioma, astrocytoma and meningioma). The first two tumor types arise from the brains supportive tissue, and are collectively called gliomas. The tumor grade indicates the level of tumor malignancy. Tumors are graded on their growth rate, vascularity (blood supply), presence of a necrotic center, invasive potential (border distinctness) and similarity to normal cells. As an alternative to invasive brain biopsy for diagnosis, magnetic resonance spectroscopic imaging (MRSI) has become one of the most important non-invasive diagnostic aids in clinical decision making, mostly because of the good visibility of soft tissue structures in order to assess location and size of the tumor; see e.g Meyerand *et al.* (1999). In brain tumor diagnosis, the voxel normally includes the tumorous area. Each voxel contains an MR spectrum, that provides metabolic information about the volume it is measured from. Our approach consists in using the appropriate statistical tools for analyzing non-invasively obtained MR spectroscopy data. Spectra data of brain cells from healthy tissue (e.g. gray matter, white matter), or unhealthy tissue (e.g. tumor, necrotic tissue), labeled according to the type and grade of tumor, for which consensus about the histopathology was reached, are stored as specific spectral patterns in MR spectra, and these are the reference pure classes that may be used as the known maximum number $L_{\max}$ of classes in our approach.

As an approximation of reality the content (e.g. tissue content in our MRSI example) within any pixel may be described by one element of the set of pre-specified labels $\{1, \dots, L\}$. We want to create an infered visual description of the $\{c(i), i = 1, \dots, N\}$ from the "data" $\mathbf{x} = \{\mathbf{x}_i, i = 1, \dots, N\}$, where each vector $\mathbf{x}_i$ is the vector of length n made by the observation of the signal $x_i(t)$ over the discretized time grid $t = t_1, \dots, t_n$. Without loss of generality, take N to be a power of 2. We will take the $\mathbf{x}_i$ to be conditionally independent draws from a finite set of component densities, i.e. $\mathbf{x}_i \sim g(\mathbf{x}, i)$ given the known collection $\{1, \dots, L\}$ of true class labels. We will assume that the subregions at the finest level are homogeneous in one of the "pure" classes $\{1, \dots, L\}$.

The above intensity modelling, for each pixel $i = 1, \dots, N$, will be done on several spatial levels of the following *multi-scale approach* in two spatial image dimensions. Essentially, we will work on the finest observed resolution level, i.e. the pixel level, to assign a class label to each pixel, whereas in a subsequent step we will use a complexity-penalised maximum likelihood approach to find on coarser spatial scales regions of similar class membership. In order to describe this hierarchy of scale levels, we introduce now the notion of recursive dyadic partitioning (RDP).

2.2 The multiscale approach using recursive dyadic partitioning and complexity-penalised maximum likelihood

Definition of recursive dyadic partitioning The following description is motivated from the one in Kolaczyk *et al.* (2005). A *recursive dyadic partitioning* (RDP) $\mathcal{P}$ is any partition of $[0, 1]^2$ that may be produced by

(1.) $\mathcal{P}_0 = [0, 1]^2$

(2.) if $\mathcal{P}$ is RDP composed of m squares $R_1, \ldots, R_m$, then any partition obtained by splitting one of the R_j into four subsquares of equal size is also a RDP. The finest partition at the pixel size is a C-RDP ("complete RDP") and denoted by $\mathcal{P}_N^*$.

The set of all RDPs $\mathcal{P} \subset \mathcal{P}_N^*$ is in one-to-one correspondance with the set of all quad-trees. The model class we are going to explore is defined as:
Let $\mathcal{P}_N^*$ be a C-RDP of $[0,1]^2$. A model M in the set $\mathcal{M}$ of all models is a tuple $(\mathcal{P}, C(\mathcal{P}))$ with $\mathcal{P} << \mathcal{P}_N^*$ ("$<<$"meaning "coarser") and $C(\mathcal{P}) = C(R)_{R\in\mathcal{P}}$, along with densities $g(x(\cdot)|c)$ for $c \in \{1, \ldots, L\}$ (i.e. $c(i)$ a label for coloring the square R_i associated to this label) such that the $\mathbf{x}_j$, $j \in R_i$ are sampled independently according to

$$f_M(\mathbf{x}|c(i)) = g(\mathbf{x}|c(i)) \ ,$$

where $c(i) = c(R_i)$ is the label assigned to pixel I_i through its association with R_i. In other words, for each region R of an RDP $\mathcal{P}$ the model specifies that $\mathbf{x}$ for all pixels $I_i \in R$ are sampled independently from $g(\mathbf{x}|c(R))$. Implicit in the definition we assume that the $g(\mathbf{x}|\ell)$, $\ell = 1, \ldots, L$ are known (they will be efficiently estimated in a previous step).

A Gaussian mixture model in the wavelet domain More specifically we will use a Gaussian mixture model in the wavelet coefficient domain of $x(\cdot)$, with $\mathbf{w}_i$ denoting the set of discrete wavelet coefficients of $\mathbf{x}_i$, i.e.

$$\mathbf{w}_i \sim \sum_{\ell=1}^{L} \pi_\ell \ g_\ell(\theta_\ell, \Sigma_\ell; \mathbf{w}) \ , \ i = 1, \ldots, N \ , \tag{3}$$

where the parameters to be estimated are the prior mixture probabilities π_ℓ, the means θ_ℓ and variances Σ_ℓ of the Gaussian densities $g_\ell(\cdot,\cdot)$ of class ℓ. We recall that following our model (2) of independent wavelet coefficients each variance matrix Σ_ℓ is diagonal (but not necessarily a multiple of the identity). Note that $\mathbf{w}_i$ is a vector-valued compactified notation for the vector of the empirical wavelet coefficients at scales j and locations k of the $i-$the pixel intensity representation, with means θ_ℓ which also depend on scale and location and a diagonal matrix Σ_ℓ with non-identical variances depending on scale j.

In order to come up with meaningful estimators, the number N of pixels has to be considerably larger than L, and hence, for the rest of the paper, a convenient bound on the maximum number $L_{\max}$ of classes is to assume that $L_{\max} \le C_L \cdot \log(N)$ (which does however not mean that we allow L to depend on N).

With the above description, the estimated likelihood for coefficient vector $\mathbf{w}_i$ writes as

$$\hat{f}_M(\mathbf{w}_i|c_i) = \sum_{\ell=1}^{L} \hat{\pi}_\ell(R_i)\ \hat{g}_\ell(\mathbf{w}_i)\ , \tag{4}$$

noting that $\hat{g}_\ell(\mathbf{w}_i)\ := g(\hat{\theta}_\ell^{(i)}, \hat{\Sigma}_\ell^{(i)}; \mathbf{w}_i)$ denotes the Gaussian likelihood of coefficient $\mathbf{w}_i$ with estimated mean $\theta_\ell^{(i)}$ and estimated variance $\hat{\Sigma}_\ell^{(i)}$. In words we have a model with regions R_i and mixtures with mixture probabilities $\pi_\ell(R_i)$ for each of the classes $1 \le \ell \le L$, assigned to all the pixels in the region R_i.

Complexity-penalised maximum likelihood We will identify the best model in $\mathcal{M}$ by maximising a complexity-penalised likelihood function

$$\widehat{M}\ = \arg\max_{M\in\mathcal{M}} \{l(\mathbf{x}|M) - 2\ \mathrm{pen}(M)\}\ ,$$

where the choice of the appropriate likelihood $l(\mathbf{x}|M)$ is motivated by analogy to Kolaczyk *et al.* (2005). The choice of the appropriate penalty rather follows general ideas of model selection following the by now well-known ideas of Birgé and Massart (1998) (which are somehow on the base of Kolaczyk's penalty, too). More specifically, as one possibility to choose the penalty in practice we propose an adaptation of the approach of Lavielle (2005) which we detail in Section 3.3 below.

As explained above, $l(\mathbf{x}|M)$ will be calculated in the wavelet coefficient domain (where the empirical coefficients are independent over pixels i) by the product of the estimated likelihood over all pixels in those regions R that correspond to model M, where for pixel i the above Gaussian mixture model is used, in formulae

$$l(\mathbf{x}(\cdot)|M) = l(\mathbf{w}|M) = \prod_{\{i:\ I_i\in R\sim M\}} \hat{f}_M(\mathbf{w}_i|c_i)\ .$$

Optimisation in this penalised-complexity problem can be done using a standard bottom-up tree pruning algorithm in which optimal submodels from coarser spatial resolutions are compared in a recursive fashion. Beginning at the finest spatial resolution (i.e. $\mathcal{P}_N^*$) select the class c for each pixel I_i with the largest log-likelihood. Next for each quad of four pixels compare the complexity-penalized likelihood of two submodels:
(i) the union of the four most-likely single pixel models with its penalty for a single spatial split (i.e. the quad into four pixels)
(ii) the single model for the set of four pixels that is most likely among all allowable $c \in \{1, \ldots, L\}$ with its smaller penalty.

Continuing this in a recursive fashion leads to model $\widehat{M}$. For later reference we note that such a model using Gaussian mixtures is also called "mixlet"-model in the literature (Kolaczyk *et al.* 2005).

Note that this RDP-approach, though using the assumption of independency over pixels, allows to recombine pixels in the same label class and hence is an alternative to more classical approaches which work with correlation models over neighboring pixels. In order to respond to the possible criticism of being restricted with RDP to spatially only dyadically representable regions we mention that Kolaczyk *et al.* (2005) have a translation-invariant (TI-) version of this 2-d RDP which is able to approximate more general regions of spatial homogeneity. We will use such a TI-version in our numerical examples in Section 4.

2.3 Overall procedure

Having set-up our model in the coefficient domain, several questions of how to estimate its parameters arise. Foremost we have to control the complexity of our estimation problem: it is (even numerically) impossible to estimate for $j = 0, \ldots, J-1$, all occuring parameters (e.g. by conditional maximum likelihood estimation - see Section 2.4 below). Hence, a *dimension reduction* step is necessary: adopting the approach of curves that do not behave "too differently" over time, we will apply an "aggregation" approach (Bunea *et al.* 2006) to find the optimal "set" of positions of empirical wavelet coefficients to keep common to all non-linear wavelet estimators. In practice, in this first step of dimension-reduction, we actually just take the union of all those wavelet coefficients that have survived a near-optimal thresholding on the basis of the individual curves. Embedding this into the framework of the statistical approach of "aggregation" is just used to prove that this union remains to be near-optimal when applied to the individual curves. After aggregation we have to sparsify additionally by some truncation procedure which will reduce further the number of wavelet coefficients to incorporate into the parametric estimation problem. Find in Section 3.1 below more on aggregation and optimizing the aggregation risk, as well as the details of the truncation test procedure in Section 3.2.

Our proposed algorithm can now be summarized as follows.

1. Reduction of complexity/dimension reduction:

   (a) For each pixel $1 \leq i \leq N$ smooth its observed pixel intensity $\mathbf{x}_i$ by a non-linear wavelet threshold estimator. In principle any estimator which just has to have the near-optimal L_2 rate of convergence for estimating $f_{c_i}(t)$ is fine enough. For reason of near-optimal denoising our preferred choice is hard-thresholding. Call this estimator $\hat{f}_i(t)$.

(b) Aggregation step: Construct an aggregated estimator $\hat{f}_\lambda(t) = \sum_{i=1}^N \lambda_i \; \hat{f}_i(t)$, and choose the optimal $\lambda = (\lambda_1, \ldots, \lambda_N)$ as in Bunea *et al.* (2006), by ℓ_1-penalized least squares in the functional domain. Details are to be found in Section 3.1 below but basically this amounts to taking the union of all the wavelet coefficients having survived thresholding of the first step.

(c) Final dimension reduction step by truncation of this aggregated estimator: Apply a Neyman truncation thresholding based test (Fan, 1996) on the wavelet coefficients of this estimator $\hat{f}_\lambda(t)$ to "sparsify" it additionally (for this control the level of the test by FDR). Again, the details are to be found below, in Section 3.2. Collect the position of the surviving wavelet coefficients. Construct new "dimension-reduced" estimators $\hat{x}_i(t)$ by taking the empirical wavelet coefficients of $x_i(t)$ at these positions. The reduced dimensionality of this estimator will be refered to in the sequel as K^*.

2. Once this set of coefficients of dimensionality K^* (sufficiently small) is determined for each curve i, we apply the EM algorithm of Law *et al.* (2004) to estimate all the parameters of model (2) and simultaneously determine the appropriate number of clusters L represented in the observed data. Here we work with the Gaussian mixture model (3), and run EM on this (see the detailed description in Section 2.4 below). To get back to the "pure class" model (2) for assigning a label to each pixel we apply an additional step to identify from the output of EM the class with the highest estimated prior probablity. This is actually our clustering step.

3. For the final spatial segmentation phase we use the penalized RDP approach to pass from mono- to multiscale. We recall that the above modelling, estimation and clustering are first done on the finest of the RDP spatial resolutions, but that in the complexity-penalized maximum likelihood algorithm we use again EM to calculate the mixing probabilities for the accordingly enlarged regions. See also the detailed description of RDP in Section 3.3, including a justification of using the penalty approach of Lavielle (2005) or - for a simplified algorithm in practice - using the "heuristics of the slope" (Gey and Lebarbier 2003). We emphasize that to display our results in the end, we use again a majority vote to assign a coloring to each region.

2.4 Details on the EM algorithm used for clustering

We recall that we are using the Gaussian "mixture (likelihood) model" of equation (3)

$$\mathbf{w}_i \sim \sum_{\ell=1}^{L} \pi_\ell \; g(\theta_\ell, \Sigma_\ell; \mathbf{w}) \; , \; i = 1, \ldots, N \; ,$$

where the parameters to be estimated are the prior mixture probabilities π_ℓ, the means θ_ℓ and variances Σ_ℓ of the Gaussian densities $g_\ell(\mathbf{w}_i) := g(\theta_\ell, \Sigma_\ell; \mathbf{w}_i)$ of class ℓ.

As input vectors for an algorithm to estimate the mixture model parameters, we emphasize that we cannot use the full high-dimensional wavelet coefficient vectors $\mathbf{w}_i$ but need to work with the output of our dimension reduction steps. These steps, explained in more detail in Sections 3.1 and 3.2, give us a dimensionality, which we will call K^*, over i, of the resulting wavelet coefficients to keep for the $i-$th curve. With slight abuse of notation we continue to call this wavelet coefficient vector $\mathbf{w}_i$.

Now, instead of the "classical" version of the iterative EM-algorithm to estimate the parameters of the above mixture model we are going to use the more refined EM-algorithm of Law *et al.* (2004) which allows us to also determine the number L of actual clusters. For the reader's convenience however we first describe what the standard EM-algorithm would look like here:
Initialise by $\pi_\ell^{(0)},\ \theta_\ell^{(0)},\ \Sigma_\ell^{(0)}$.
Iterate over $p \geq 1$ (until numerically convergence):
E-Step:

$$\pi_\ell^{(p+1)} \ = \ \frac{1}{N}\sum_{i=1}^{N} \nu_\ell^{(p)}(i) \ , \quad \text{where} \quad \nu_\ell^{(p)}(i) = \frac{\pi_\ell^{(p)}\ g_\ell^{(p)}(w_i)}{\sum_{\ell=1}^{L} \pi_\ell^{(p)}\ g_\ell^{(p)}(\mathbf{w}_i)} \ ,$$

with $g_\ell^{(p)}(\mathbf{w}_i) := g(\theta_\ell^{(p)}, \Sigma_\ell^{(p)}; \mathbf{w}_i)$.
M-Step:

$$\theta_\ell^{(p+1)} \ = \ \frac{\sum_{i=1}^{N} \nu_\ell^{(p)}(i)\ w_i}{\sum_{i=1}^{N} \nu_\ell^{(p)}(i)} \ , \quad \text{and} \quad \Sigma_\ell^{(p+1)} \ = \ \frac{\sum_{i=1}^{N} \nu_\ell^{(p)}(i)\ (\mathbf{w}_i - \theta_\ell^{(p+1)})'(\mathbf{w}_i - \theta_\ell^{(p+1)})}{\sum_{i=1}^{N} \nu_\ell^{(p)}(i)} \ .$$

Once the cluster densities and the mixture priors are estimated one can get back to the " pure class" model by assigning to each i a class $c(i)$ via exploration of the vector of "converged mixture probabilities" for index i denoted by $(\nu_1(i), \ldots, \nu_L(i))$: Let $\ell_i^* = \arg\max_{1 \leq \ell \leq L} \nu_\ell(i)$. Then we assign the pure class ℓ_i^* to the pixel of index i.

The variant of this EM-algorithm that we are using is the one of Law *et al.* (2004). The idea of this algorithm is to include not only the estimation of the actual number L of classes but also the selection of the relevant "features" that is, the elements of the vector of wavelet coefficient $\mathbf{w}_i$ which carry the most important information for "feature extraction" (i.e. clustering). Roughly speaking a feature is kept if its probability density gives sufficiently high likelihood for carrying a label that belongs to one of the L classes. If the density is somehow the "same" independently of any specific class label then the feature is supposed to be not informative. This idea is realized by introducing, for fixed i, probability weights $\rho_q,\ q = 1, \ldots, Q$ which measure the probability that the $q-$th component of the vector $\mathbf{w}_i$ should be included, i.e. that the feature represented by this component is relevant. Note that the number of components of each of the input vectors $\mathbf{w}_i$ is the same Q, which, by the output of our dimension reduction step by combined aggregation and testing equals K^* (the reduced dimensionality after temporal aggregation and testing). The probability weights ρ_q have to be estimated from the data and for the final purpose to be set to binary 0/1 (exclusion/inclusion) by obvious rounding. (This is because in the subsequent complexity-penalised RDP approach we are not going to continue to work with a probability distribution on the feature vector.)

More specifically, two important modifications have to be done to our likelihood function to take this feature selection into account. For one, we need to let the parameter $(\theta_\ell,\ \Sigma_\ell)$ of the (Gaussian) densities $g_\ell(\mathbf{w}_i)$ now depend on the index $q = 1, \ldots, Q_i\ (= Q)$ of the feature label, i.e. this density $g_{\ell,q}(\mathbf{w}_i) = g(\theta_{\ell,q}, \Sigma_{\ell,q}; \mathbf{w}_i^{(q)})$ describes now the pdf of the $q-$th feature $\mathbf{w}_i^{(q)}$ in the $\ell-$th class. As the features, i.e. the components of the wavelet vector $\mathbf{w}_i$ are independent in our set-up, we have that $g_\ell(\mathbf{w}_i) = \prod_{q=1}^{Q_i} g_{\ell,q}(\mathbf{w}_i)$.

A second modification comes from the idea that the $q-$th feature component is irrelevant if its distribution is independent of the class label ℓ, i.e. if it follows a common density, called $h_q(\mathbf{w}_i^{(q)}) := h(\lambda_q, \Gamma_q; \mathbf{w}_i^{(q)})$. For simplicity, and as this leads to reasonable results in practice, it is suggested to use again a Gaussian density for $h(\lambda_q, \Gamma_q; \cdot)$. With this, the final likelihood becomes

$$\mathbf{w}_i \sim \sum_{\ell=1}^{L} \pi_\ell \prod_{q=1}^{Q_i} g_{\ell,q}(\mathbf{w}) \ = \sum_{\ell=1}^{L} \pi_\ell \prod_{q=1}^{Q_i} \Big(\rho_q\ g_{\ell,q}(\mathbf{w}^{(q)})\ +\ (1-\rho_q)\ h_q(\mathbf{w}^{(q)}) \Big),\ i = 1, \ldots, N\ ,$$

where now the set of parameters to be estimated by a more sophisticated variant of the EM-algorithm is $(\pi_\ell, \theta_{\ell,q}, \Sigma_{\ell,q}, \rho_q, \lambda_q, \Gamma_q)$. We recall that both matrices $\Sigma_{\ell,q}$ and Γ_q continue to be diagonal (but not necessarily a multiple of the identity matrix).

In this modified EM-algorithm the idea is now to work with two “hidden” (i.e. latent) random variables, one set of N variables $Z = \{z_1, \ldots, z_N\}$, describing the N missing labels, with each $z_i = (z_{i1}, \ldots, z_{iL})$, where $z_{i\ell}$ is a binary variable saying whether pixel i belongs to the hidden class ℓ. Another vector of binary variables $\Phi = (\phi_1, \ldots, \phi_Q)$ is used to describe the relevance of feature q, i.e. $\rho_q = P(\phi_q = 1),\ q = 1, \ldots, Q$. This set-up then allows to run a standard EM-algorithm, for details we refer to Law *et al.* (2004, Section 3.2.1).

It remains to explain how the number L of cluster classes is selected, including the question of a good general initialisation of this EM algorithm which is essential to reach a good local optimum. For this Law *et al.* (2004), Section 3.3, use an approach based on the MML (Minimum Message Length) criterion which is closely related to MDL (Minimum Description Length) by adding to the negative loglikelihood a series of terms which "penalise" against the number of model parameters. Interestingly enough, as long as a minimum number of components L_{min} is given by the user, optimisation of this "penalised" likelihood can be directly encorporated into the M-Step of the standard EM-algorithm by only a slight modification. Since this model selection algorithm determines the number L of components, it can be initialized with a large value L_{max} of L (the known maximum number of possible "pure" classes), thus alleviating the need for a good initialisation. Hence, the following componentwise version of EM can be used (taken from Figure 6 of Law *et al.* 2004):

- Initialisation: Set the parameters of a large number of mixture components randomly and set the common (Gaussian) distribution h_q to cover all data. Choose $\rho_q = 0.5$ for all features q.

- Loop over descending $L = L_{max}, \ldots, L_{min}$:

- Inner loop, for each L, over EM until a local minimum is reached (if π_ℓ becomes zero, the $\ell-$th component is pruned, and if ρ_q becomes 1, the density $h_q(\cdot)$ is pruned, whereas if ρ_q becomes 0, the densities $g_{\ell,q}(\cdot)$ are pruned for all $\ell = 1, \ldots, L$). For given L, record the current model parameters and its message length. Then, in order to descend L, remove the component with the smallest weight.

- End of outer loop over L: Return to the model parameters that yield the smallest message length. Postprocess the estimators for the probability weights $\pi = (\pi_1, \ldots, \pi_L)$ for the $i-$th pixel to assign to it a pure class label (as described above) and also the estimators for $\rho = (\rho_1, \ldots, \rho_Q)$ to a vector of binary 0/1 weights for its feature components (i.e. the selected components of the vector $\mathbf{w}_i$) by rounding the probabilities ρ_q to zero and one.

Note that with this, in general the dimension of the feature component vector, which is equal to $\sum_q 1_{\{\phi_q=1\}}$, will be smaller than $Q = K^*$.

3. STATISTICAL ASPECTS OF DIMENSION-REDUCTION AND OF PENALISED MAXIMUM LIKELIHOOD

In this section we are going to give more details on justifying our dimension reduction approach combining the framework of aggregation estimators and Neyman truncation tests in the wavelet coefficient domain. Further we elaborate how to choose an appropriate penalty such that our penalised maximum likelihood approach is likely to select a "good" RDP. We end by furnishing some theoretical results on asymptotically correct classification by use of the "horizon" model. We note that in order to show consistency of our estimation algorithm is consistent, we adopt the following asymptotic approach. We first show near-optimality of the aggregation step by letting, for a given number of pixels N, the temporal dimension n tend to infinity. Later, in order to achieve consistent estimators of each true "class intensity" θ^{ℓ}_{jk} by the complexity-penalised RDP approach, we let N tend to infinity.

3.1 **Dimension-reduction by aggregation**

We recall our model (1) in the functional domain, with the observed intensity for pixel i with $1 \le i \le N$,

$$x_i(t) \;|\{c(i)=\ell\} \;=\; f^{\ell}(t) + \varepsilon_i(t) \;, \qquad t = t_1, \dots, t_n, \; n = 2^J \;, \; 1 \le \ell \le L \;,$$

In our present situation, the parametric form of each deterministic mean function $f^{\ell}(\cdot)$ is unspecified. We therefore seek a parsimonious representation of a huge number of possible representations. In order to perform efficient denoising of each time path of pixel intensities we will use a wavelet domain representation exploiting to a full extent the sparseness and de-correlation properties of the DWT. Indeed, it is reasonable hereafter to assume that each mean pixel intensity class curve belongs to a ball of the Besov space $B^s_{p,q}[0,1]$. It is then known, that for regular enough wavelet bases, functions in such spaces admit a sparse representation in the sense that only few large coefficients dominate the representation since sparsity is measured by $(\sum |\theta_{j,k}|^p)^{1/p} \le R$ (see, e.g. Donoho and Johnstone, 1998) with $0 < p \le 1$. Indeed, since the mean intensity class functions are within the Besov balls, tail sums of coefficients become small after some point and the error in zeroing out the coefficients from this point is perfectly controlled.

Applying a discrete wavelet transform to (1) we obtain a representation of the wavelet coefficients of $x_i(t)$ which is, we recall from (2),

$$\mathbf{w^i} \;|\{c(i)=\ell\} \;= \boldsymbol{\theta}^{\ell} + \boldsymbol{\varepsilon^i} \;, \; 1 \le \ell \le L < N \;,$$

where

$$\varepsilon^i_{jk} \;|\{c(i)=\ell\} \;\sim\; \mathcal{N}(0, \sigma^2_{j,\ell}) \;.$$

To proceed to our clustering one could apply a standard cluster-EM procedure to the empirical wavelet coefficients, but such a procedure would not take advantage of the "sparse" situation of the wavelet coefficients that is likely present and would result in a very unstable clustering. Our purpose therefore is to proceed to some kind of dimension reduction without sacrificing the discriminative characterisation of the classes. The idea is as follows.

For sparse signal nonlinear wavelet thresholding procedures lead to quasi-minimax denoising. One may see wavelet soft thresholded estimation as a procedure that, for each pixel intensity curve i $(i = 1, \ldots, N)$, selects a subset m_i of $\mathcal{J} = \{(j,k), j = 0, \ldots, J-1; k = 0, \ldots, 2^{j-1}\}$ and for each index (j,k) within m_i estimates the corresponding coefficient by $\hat{\theta}^{(i)}_{j,k} = \eta_{t_{\ell,j}}(w^{(i)}_{j,k})$ setting all other coefficients to 0. In the previous expressions η_t denotes the scale dependent thresholding procedure and $t_{\ell,j}$ the universal threshold $\sigma_{j,\ell}\sqrt{2\log n}/\sqrt{n}$.

Denote by $\bar{m} = \cup_{i=1}^{N} m_i \subset \mathcal{J}$ the set of indices of all coefficients retained from all the N curves by wavelet denoising. By the minimax properties of the wavelet denoising procedure these carry all the relevant information necessary to discriminate the L mean intensity curves. Note that thresholding the coefficients indexed by $\bar{m}$ for each of the observed curves will obviously give the m_i originally selected coefficients for the corresponding ith curve, since all other coefficients with indices in $\bar{m}\backslash m_i$ are taken to be below the threshold $t_{\ell,j}$ and considered as noise. Since the total number of classes is finite, with probability tending to 1 as $n \to \infty$, thresholding any new curve from a given class, will select the same coefficients (noise-free reconstruction property of wavelet thresholding). Therefore, adding to the estimator of a new curve which comes from the same class as the $i-$th, a set of $d_i := \#\bar{m}\backslash m_i$ coefficients, will affect the asymptotic MSE properties by an amount of the order $\max d_i \log n/n$. A rough upperbound of $\max d_i$ is N. Therefore the resulting denoising procedure, while not as efficient, still produces consistent estimators as long as $N \log n/n$ tends to zero.

What we would have liked is to show that $\max d_i = C(B^s_{p,q}, L)$ is not depending on N or n, but this is not obvious, since wavelet thresholding is not truly a model selection procedure. The use of $\#\bar{m}$ coefficients produces consistent estimates but their rates can be far from the minimax rates that individual thresholding produces. To address this issue we will therefore justify our choice by showing that our estimates based on the total $\#\bar{m}$ coefficients are not far from a linear aggregated estimator. This is the purpose of what follows.

For each pixel i we estimate the underlying $f^{c_i}(\cdot)$ by soft (or hard) thresholding or even (level-dependent) SURE thresholding. We focus the discussion on the version of the threshold estimator expressed by its sequence of wavelet coefficients $\hat{\theta}_i$. Under our assumption that each of the unknown curves $f^\ell(\cdot)$ belongs to a Besov ball $B^s_{p,q}(M)$, for $1 \leq p,q \leq \infty;\ M < \infty;\ s > (1/p - 1/2)_+$, we have that

$$\sup_{\theta^{c_i} \in B^\alpha_{p,q}(M)} \|\hat{\theta}_i - \theta^{c_i}\|_2^2 \ \leq \ C\ \log(n)\ n^{-\frac{2s}{2s+1}}\ (1 + o(1))\ .$$

Therefore, for each i, the smooth estimate obtained by wavelet thresholding the corresponding curve is an optimal estimate of the underlying mean intensity curve $f^{c(i)}(\cdot)$. We would like to construct a new estimate as a linear combination (in the wavelet domain) of all these estimates, that mimics, without even knowing from which class the observations are drawn, the behavior of the best (in L^2 risk) among the individual estimates. Note that such an estimate is necessary based on the $\bar{m}$ coefficients described before.

For $\lambda = (\lambda_1, \ldots, \lambda_N)\ \in\ \mathbf{R}^N$, we aggregate the above estimators (in the wavelet domain) by defining

$$\hat{f}_\lambda(t)\ =\ W^T\ \left(\sum_{i=1}^N \lambda_i\ \hat{\theta}_i\right) = \sum_{i=1}^N \lambda_i \hat{f}_i(t),$$

where, due to the linearity of the wavelet (back) transform, aggregating the coefficients amounts to directly combining linearly the estimators in the time domain.

The purpose of aggregation of M individual estimators is therefore to construct a single adaptive estimator (denoiser) that shares their advantages on terms of global L_2 risk. The combined procedure will pay a price and the main purpose of optimal aggregation is to obtain an upper bound for this price that is asymptotically small. In what follows we denote M_ℓ the number of curves within the M that are drawn from the $\ell-$th class and we assume, without any loss of generality that M is large enough in order that the ratio M_ℓ/M can be considered as equal to p_ℓ (the weight of the $\ell-$th class within the entire population). We will further assume that $p_\ell > 0$ for $\ell = 1, \ldots, L$.

In summary, the performance of an aggregate $\tilde{f}$, say, is evaluated by the target

$$\frac{1}{M}\sum_{\ell=1}^L M_\ell \mathbb{E}_{f^\ell}\|\tilde{f} - f^\ell\|^2\ \leq\ \frac{1}{M}\sum_{\ell=1}^L M_\ell \inf_{\lambda \in \mathbf{R}^M} \mathbb{E}_{f^\ell}\|\hat{f}_\lambda - f^\ell\|^2\ +\ \Delta_{n,M}\ ,$$

where $\inf_{\lambda \in \mathbf{R}^M} \frac{M_\ell}{M}\mathbb{E}_{f^\ell}\|\hat{f}_\lambda - f^\ell\|^2$ represents the best approximation of the unknown mean intensity f^ℓ of class ℓ (weighted by its representation in the image) by the linear combination of the M estimators and $\Delta_{n,M}$ is the price to be paid for aggregation (a constant to be found which has to be independent of f^ℓ) and which should tend to 0 as n tends to ∞.

Following the standard literature on the construction of aggregation estimators we will use a part of the original sample which could be half of the sample (e.g. the even pixels) say $\mathcal{D}_{1,N}$, called the training sample, to construct the estimators $\hat{f}_i(t_k))$ and the other part , say $\mathcal{D}_{2,N}$ (e.g. the intensity curves $x_i(\cdot)$ based on the odd pixels), called the learning sample, to construct the aggregated estimator. The cardinal of $D_{2,N}$ will be denoted M hereafter. Following Nemirovski (2000) and as long as the training sample is large enough to cover all classes, the estimators $\hat{f}_i$ based on $\mathcal{D}_{1,N}$ will be considered as fixed function during the aggregation step, and we will therefore focus our aggregation on learning, letting M going to ∞. As in Bunea *et al.* (2006), we will make the assumption that all estimators $\hat{f}_i$, as well as the mean intensity functions f^ℓ are uniformly bounded by an unknown finite bound U. This assumption is reasonable for f^ℓ but, although it seems more restrictive on the estimates $\hat{f}_i$, it is not really since one may replace the individual estimates $\hat{f}_i$ by an appropriate truncation as in Kohler (2003) without affecting their optimal properties. We will assume further that, for M large enough, almost surely the matrix $\Psi_M = \left(1/n \sum_{k=1}^{n} \hat{f}_i(t_k)\ \hat{f}_j(t_k)\right)_{1\le i,j \le M}$ is positive definite for any given $n \geq 1$. Let ξ_{min} be the smallest eigenvalue of Ψ_M. It is easy to see that under the above assumptions it follows that almost surely

$$0 < \xi_{\text{min}} \leq 1/n \sum_{k=1}^{n} \hat{f}_i^2(t_k) \leq U^2, \quad i \in \mathcal{D}_{2,N}.$$

Let

$$\hat{S}(\lambda) \ = \ \frac{1}{nM} \sum_{i \in \mathcal{D}_{2,N}} \sum_{k=1}^{n} (x_i(t_k) - \hat{f}_\lambda(t_k))^2 \ ,$$

and

$$\text{pen}(\lambda) \ = \frac{1}{M} \sum_{j \in \mathcal{D}_{2,N}} \tau_{n,j}\ |\lambda_j| \ , \ \text{ with } \ \tau_{n,j} = 2\sqrt{2}\ \sigma \|\hat{f}_j\|_n \ (\frac{2\log(M) + \log(n)}{n})^{1/2} \ ,$$

where $\sigma^2 = \sup_i \sigma_i^2$ with σ_i^2 the integrated variance of the i-th estimated curve $\hat{f}_i$. Note that such a variance is of the form $\sum_{j=0}^{J-1} \sigma_{j,c(i)}^2$ where $\sigma_{j,c(i)}^2$ is the variance of the wavelet estimator of the mean curve $f^{c(i)}$ at scale j. In practice σ^2 is unknown but can be consistently be estimated by the $\max_i \hat{\sigma}_i^2$ where $\hat{\sigma}_i^2$ is the MAD level dependent robust estimator based on each curve x_i.

Then the aggregate estimator is the one that minimizes $\hat{S}(\lambda) \ + \ \text{pen}(\lambda)$ over the set

$$\Lambda_{M,T,2} \ = \ \{\lambda \in \mathbf{R}^M : \sum_{j=1}^{M} \lambda_j^2 \ < \ T^2\} \ ,$$

with $T > 0$ such that $T^2\ \xi_{\text{min}} > 2U^2$ and $T \leq (\log(\max\{M, n\}))^{1/4}$.

We can state the following proposition:

Proposition 3.1. *Under the assumptions made in this section, let* $\xi_{min}^{-1/2}\sqrt{2}U \leq T \leq (\log(\max\{M, n\}))^{1/4}$. *Then there exists a constant* $C = C(T, U, \sigma^2, \xi_{min})$ *such that for all* $\eta > 0$ *and for all integers* $n \geq 1$ *and* $M \geq 2$,

$$\frac{1}{M} \sum_{\ell=1}^{L} M_\ell \mathbb{E}_{f^\ell} \|\tilde{f} - f^\ell\|_n^2 \ \leq (1+\eta) \frac{1}{M} \sum_{\ell=1}^{L} M_\ell \inf_{\lambda \in \mathbf{R}^M} \mathbb{E}_{f^\ell} \|\hat{f}_\lambda - f^\ell\|^2 \ + C(1+\eta+\eta^{-1}) \frac{\log(\max(M,n))}{nM}.$$

Proof: see Appendix

The above proposition shows that our aggregate is nearly-optimal in the sense that the price of adaptation through aggregation $\Delta_{n,N}$ is of order $\log(\max\{M,n\})/(Mn)$. The result will follow from adapting to our case general results on aggregation from Bunea *et al.* (2006).

To perform denoising we have applied thresholding procedures to each intensity curve one at a time and we have shown, by means of aggregation that the use of the union of wavelet coefficients selected from individual data curves to construct a "combined" representative set of coefficients for approximating the original data curves is optimal (in terms of denoising) for multiple data curves. However, whenever N is large, the aggregated wavelet denoiser, while maintaining sparsity, may keep too many coefficients for each wavelet processed intensity curve. In practice, therefore, the EM algorithm used for clustering may perform poorly and the use of such a large dimension will also degrade considerably the RDP fitting process. It is therefore important to proceed to a dimension reduction, before applying a simultaneous feature selection and clustering using our mixtures models.

3.2 Final dimension reduction step by a Neyman truncation test

We propose hereafter a dimension reduction based on an appropriately defined multiple adaptive Neyman truncation test based procedure for testing that, for each wavelet coefficient within the representative union of coefficients, its expectation among the N intensity curves remains constant against the assumption that its expected behaviour is different among curves. Since the maximum number of possible classes is assumed to be small with respect to the total number of pixels, the procedure is inspired by one-way ANOVA procedure where the levels of the nominal factor are the class labels, which in our case are sparse but unfortunately unknown. Our final selection procedure rests then on an implicit hypothesis that the active coefficients allowing good discrimination are also sparse.

To begin with, let $k_{\max}$ be the cardinality of the set $\bar{m}$ of the indices of the wavelet coefficients retained by aggregation. When a wavelet-position k is fixed, denote by w_k^i, $i = 1, \ldots, N$, the collection of coefficients from all curves at this position and let $d_k^i = w_k^{i+1} - w_k^i$, $i = 1, N-1$ be the vector of first order differences of these coefficients. It is easy to see that d_k^i can be considered as a random sample from a Gaussian distribution $\mathcal{N}(\mu_k^i, \tau^2)$, where $\mu_k^i = \mathbb{E}(w_k^{i+1} - w_k^i)$. Note that the variance τ^2 is closely related to the variance $\sigma_{j,\ell}^2$ of the wavelet coefficients w_{jk}^i of model (2).) If the kth coefficient, even if it is important for representing the curves, has not any discriminative power for separating the classes, then the vector $\boldsymbol{\mu}_k$ will be identically equal to zero. Our testing procedure to retain the coefficient is then inspired from the adaptive Neyman testing procedure for testing

$$H_{0,k}: \ \boldsymbol{\mu}_k = 0 \text{ versus } H_{1,k}: \ \text{at least one component of } \boldsymbol{\mu}_k \text{ is not zero.}$$

Since the number of possible classes is small with respect to the total number of curves, most of the components of $\boldsymbol{\mu}_k$ will be 0, and a simple procedure to test the eventual presence of non-zero components is to use the thresholding based Neyman test of Fan (1996). For our purpose we use the hard thresholding statistic, T_k^H, defined by

$$T_k^H = \sigma_H^{-1}(T_k^{H*} - \nu_k^H),$$

where

$$T_k^{H*} = \tau^{-2} \sum_{i=1}^{N} (d_k^i)^2 I(|d_k^i| \geq \tau \delta_H),$$

and

$$\begin{aligned} \nu_k^H &= \sqrt{2/\pi} a_N^{-1} \delta_H (1 + \delta_H^{-2}) \\ \sigma_H^2 &= \sqrt{2/\pi} a_N^{-1} \delta_H^3 (1 + 3\delta_H^{-2}) \end{aligned}$$

The threshold value is given by $\delta_H = \sqrt{2 \log N a_N}$ and a_N is given by

$$a_N = \min\left(4 \left(\max_i \frac{|d_k^i|}{\tau} \right)^{-4}, \ \log^{-2} N \right).$$

By Theorem 2.3. of Fan (1996), under the null hypothesis, we have $T_k^H \rightarrow \mathcal{N}(0,1)$ and, for a significance level α we reject the null hypothesis (we keep the selected coefficient) if

$$T_k^H > \Phi^{-1}(1-\alpha),$$

where Φ is the standard cumulative distribution function.

To select the appropriate wavelet coefficients for discrimination we will use the False Discovery Rate (FDR) procedure which has been developed in the context of multiple hypotheses testing by Benjamini and Hochberg (1995). Given the set of our $k_{\max}$ hypotheses out of which an unknown number k_0 are true, the FDR method identifies the hypotheses to be rejected, while keeping the expected value of the ratio of the number of false rejections to the total number of rejections below q, a user specified control value. In addition, this technique can handle problems in which $k_{\max}$ is very large at a very low computational cost.

The FDR procedure, as suggested by Benjamini and Hochberg (1995) is as follows: First find test statistics $T_1^H, \ldots, T_{k_{\max}}^H$ based on the Neyman's procedure with corresponding p-values $\pi_1, \ldots, \pi_{k_{\max}}$. Then, for any user-specified value $q \in (0,1)$, perform the following steps:

- Order the p-values $\pi_{(1)} \leq \cdots \leq \pi_{(k_{\max})}$.
- Compute $\hat{k} = \max\{k : \pi_{(1)} \leq \frac{k}{k_{\max}} q / \sum_{k=1}^{k_{max}} k^{-1}\}$.
- Reject all $H_{0,(j)}$, $1 \leq j \leq \hat{k}$ where $\boldsymbol{\mu}_{(j)} = 0$ is the null hypothesis corresponding to the ordered p-value $\pi_{(j)}$. If no such $\hat{k}$ exists, do not reject any hypothesis.
- Estimate the set of I_0 of wavelet coefficients to keep by the set $\hat{I}$ of indices corresponding to the first $\hat{k}$ ordered p-values.

The above procedure is completely justified by Proposition 2.1 of Bunea *et al.* (2006). We will therefore take the above chosen $\hat{k}$ as the choice for our best dimension K^* for this purpose of dimension reduction.

3.3 Choosing the RDP penalty

Our approach for producing a segmentation of an image is based on extracting the information in a particular mixlet model that is optimal in some well-defined sense with respect to the measurements x. In order to evaluate the goodness of fit of each candidate model, we have used a complexity-penalized maximum likelihood criterion, balancing the fidelity to the data measured by the log-likelihood with the parsimony of the model measured by a penalty function $p(\mathcal{M})$.

Inspired by the recent work of Lavielle (2005) on change-point problems, the purpose of this section is to propose an adaptive procedure for choosing and calibrating the penalty function on the number of subregions in the RDP fitting process. In our framework any RDP $\mathcal{P}$ of size $m = |\mathcal{P}|$ may be described through its corresponding quad-tree, pruned from that corresponding to $\mathcal{P}_N^*$. Furthermore, there are a total of $b(L)$ free parameters associated with the mixture model in each region $R \in \mathcal{P}$, but these remain constant for a given categorization of the pixels. The form of $p(\mathcal{M})$ depends only on m and increases with m. In the following we suggest to use the simplest penalty function $p(\mathcal{M}) = \beta\ m$. Such a choice finds its justification in Birgé and Massart (1998) and is similar in spirit to the penalties used by Kolaczyk *et al.* (2005) or Castro *et al.* (2004). Moreover, it is justified, when, as described in Kolaczyk *et al.* (2005), the RDP fitting process is seen as a natural extension of the 'horizon model' by Donoho (1999) - see also Section 3.4 below - i.e. when the regions defined by the partition P associated to the true model are separated by Hölder smooth boundaries.

Let $m_{\max}$ be an upper bound on the size of the RDP $\mathcal{P}$, and for any $1 \leq m \leq m_{\max}$, let $\mathcal{T}_m$ be the set of all RDP's of size m. By definition the best mixlet model of size m is the one which maximizes the loglikelihood $\ell(\mathbf{x}(\cdot)|m)$ over $\mathcal{T}_m$. For a given β and for a penalty proportional to the size m, the best fitted model is the model of size $\hat{m}(\beta)$. Under the 'horizon model' and if β is a function of N that goes to 0 at an appropriate rate as N goes to infinity, the estimated number of regions $\hat{m}(\beta)$ converges in probability to the true number of regions. However, we are interested in a choice of β in a non-asymptotic sense.

The way that $\hat{m}(\beta)$ varies with the regularization parameter β is given in the following proposition.

Proposition 3.2. *For almost all* $\mathbf{x}$*, there exists a sequence* $m_1 = 1 < m_2 < \cdots < m_{max}$ *and a sequence* $\beta_0 = \infty > \beta_1 > \cdots$ *with*

$$\beta_i = \frac{-\ell(\mathbf{x}(\cdot)|m_i) + \ell(\mathbf{x}(\cdot)|m_{i+1})}{m_{i+1} - m_i}, \quad i \geq 1,$$

such that $\hat{m}(\beta) = m_i$*, for all* $\beta \in (\beta_i, \beta_{i-1})$*.*

Note that by Proposition 3.2 the subset $\{(m_i, -\ell(\mathbf{x}(\cdot)|m_i), i \geq 1\}$ is just the convex hull of the set $\{(k, -\ell(\mathbf{x}(\cdot)|k), k \geq 1\}$. This also suggests the following graphical method for selecting the parameter β and the corresponding dimension m: examine how the negative log-likelihood $-\ell(\mathbf{x}(\cdot)|k)$ decreases when k increases, and select the dimension k for which the negative log-likelihood ceases to decrease significantly. The above procedure is very similar to the nonlinear L-curve regularization method used for determining a proper regularization parameter in penalized nonlinear least squares problems (see Gulliksson and Wadin 1998). In our context, the L-curve is defined as the curve $(-\ell(\mathbf{x}(\cdot)|m(\beta)); m(\beta))_{\beta \geq 0}$ and defines a strictly decreasing convex function with a derivative with respect to $m(\beta)$ equal to $-\beta$. The L-curve has usually a distinct corner, defined as the point where the L-curve has its greatest curvature and corresponding in our case to the point β where the negative loglikelihood ceases to decrease.

One option to find the "optimal" β and the corresponding "optimal" dimension $\hat{m}(\beta)$ is based on the idea of the "heuristic of the slope".

The contrast (i.e. the log-likelihood) associated to a RDP is the sum of two terms: the first term represents some approximation error within the associated clustering, i.e. a bias term, and a second term which represents the complexity of the model and can therefore be interpreted as a variance term. The idea of the heuristic of the slope is that when a model is high dimensional, the associated bias is close to zero, and so the contrast of the RDP is essentially an estimation of the variance of the model which is directly related to the size m of the RDP. (We recall that the choice of the penalty is in order to balance the increase of the variance with increasing m.) Hence, for large m, the negative log-likelihood should become a linear function of m. The choice of the dimension $\hat{m}$ beyond which the negative log-likelihood becomes linear is left to the user. The choice of $\hat{m}$ can be based on visual inspection of the L-curve $(-\ell, m(\beta))$, and is obtained by selecting the dimension m for which the negative likelihood ceases to decrease significantly and becomes linear. To choose $\hat{m}$, one can also plot the curve $(\beta, m(\beta))$ and search for the parameter β associated to the first significant jump in this curve. Once we have chosen the appropriate dimension $\hat{m}$, the basic principle is to fit a linear regression of $-\ell$ with respect to m for $m \geq \hat{m}$. If we denote by $\hat{\alpha}$ the estimated regression coefficient, then as suggested by Birgé and Massart (2001), an appropriate estimator for β is given by $\hat{\beta} = \kappa\hat{\alpha}$ where κ is a constant close to 2 (in practice we shall take $\kappa = 2$).

3.4 A theoretical result on correct clustering: the "horizon model" framework

To show that our RDP approach has some theoretical optimal property we use a "horizon model" (Donoho, 1999), a two label model with an arbitrary Lipschitz-frontier between the two regions. Note that these techniques have already been suggested by Korostelev and Tsybakov (1993) on their "model of boundary fragments" to prove near-optimal minimax results of classification. We simply refer to what has been derived in the similar RDP-context of Kolaczyk *et al.* (2005), Section 3.3., and give, for the reader's convenience, a short summary here on their results. Note that Castro *et al.* (2004) essentially use the same techniques to prove their results, however using an explicit basis of wedgelets to approximate their piecewise constant $d-$dimensional functions ($d \geq 2$) with $d-1$-dimensional Hölder-2 smooth boundaries (for example, twice continuously differentiable curves).

In the sequel we follow the description from Kolaczyk *et al.* (2005), Section 3.3. A horizon model consists of a partitioning of the domain $[0,1]^2$ into two compact regions, say G and $\overline{G}$, which are separated by a smooth, i.e. Lipschitz-continuous, boundary δG, and of a density function f with two distinct parts, on G and $\overline{G}$, which defines the underlying structure in the image.

More specifically, assume that $G \in \mathcal{G} = \{(t_1, t_2) \in [0,1]^2 : \ 0 \leq t_1 \leq 1, \ 0 \leq t_2 \leq h(t_1), \ h \text{ Lipschitz on } [0,1]\}$. Further, let $\pi : \ [0,1]^2 \to [0,1]^L$ be a function of L components on the unit square with each component constant on each of the regions G and $\overline{G}$, with at least one component differing on G and $\overline{G}$, and with $\sum_{\ell=1}^{L} \pi_\ell(t) = 1 \ \forall \ t \in [0,1]^2$. Then we use the horizon model density

$$f(\mathbf{w}) = \sum_{\ell=1}^{L} \pi_\ell(I_i) \ g_\ell(\mathbf{w}) \ ,$$

which, as we are using "the pure class per pixel" modelling, does not say anything else than assigning to each pixel of a given region the class label with its "highest likelihood". (Note that this is slightly different from the modelling in Kolaczyk *et al.* (2005) and actually a bit simpler.)

Let finally $\mathcal{F}$ denote the collection of all these horizon model densities, and observe that under this class, sampling on the two subregions G and $\overline{G}$ can be done with respect to densities f with any combination of the L-component densities g_ℓ.

As in Kolaczyk *et al.* (2005), one can measure the quality of the model $\widehat{M}$ selected by our complexity-penalized likelihood approach by the following risk

$$R(f, \hat{f}) = N^{-1} \ \mathbb{E} H^2(f, \hat{f}) \ ,$$

where $H^2(f, g) = \int (\sqrt{f} - \sqrt{g})^2$ denotes the squared Hellinger distance between densities f and g. We have the following analogue to Kolaczyk*et al.* (2005), Theorem 1, the proof of which follows the lines of the cited paper. We recall that in our set-up the total number $L_{G,\overline{G}}$ of true component densities $g(\mathbf{w})$ corresponding to G and $\overline{G}$ under f is given by our number L of classes, with $L = O\ (\log(N))$.

Proposition 3.3. *Suppose that* $\mathbf{w} \sim f$, *for* $f \in \mathcal{F}$. *Let* f_M *be the density corresponding to our model (4), and consider the model* $\widehat{M}$ *selected by our complexity-penalized likelihood approach where optimization is over a collection* M_N^* *produced through the discretization (over time) of the mixing weights to an accuracy of* $N^{-1/2}$. *Assume finally, in order to be able to discriminate between two classes, that* $\max_{\ell,\ell'} \sup_{\mathbf{w}} \dfrac{g_\ell(\mathbf{w})}{g_{\ell'}(\mathbf{w})} \leq B < \infty$. *Then*

$$R(f, f_{\widehat{M}}) \ \leq \ L \ \cdot \ O\ (\frac{\log N}{N})^{1/2}$$

for each $f \in \mathcal{F}$, *and*

$$\sup_{f \in \mathcal{F}} R(f, f_{\widehat{M}}) = O\ (\frac{\log^3 N}{N})^{1/2} \ .$$

Note first that the per density risk is of the same order as usual adaptive multiscale estimators in the standard horizon model (e.g., Donoho 1999). Note further that the additional factor of order $\log(N)$ of the optimal order over the entire class $\mathcal{F}$ can be considered as the price paid for not knowing the number of component densities g used by f on G and $\overline{G}$.

4. NUMERICAL EXAMPLES

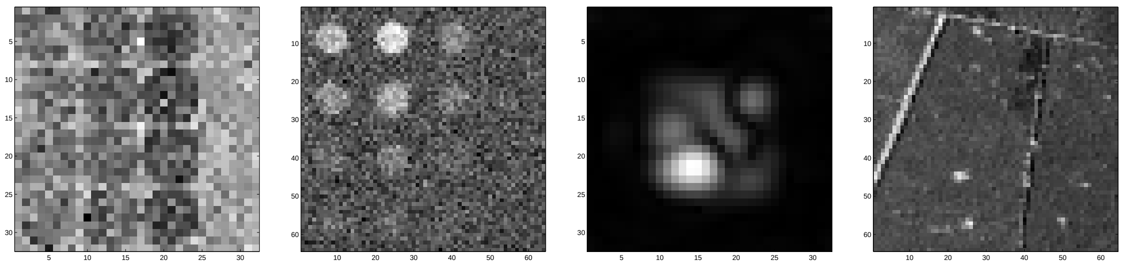
In this section we present some results, both on simulated and on real data, that we obtain by combining dimension reduction (i.e. statistical aggegration followed by the Neyman testing procedure with FDR control of the type-1 error, as described in Section 3.2), the variant of the EM-algorithm by Law *et al.* (2004) and our complexity-penalized RDP approach including an automatic choice of the regularizing parameter β based on the heuristic of the slope (as desribed in Section 3.3). We consider the following four examples:

- the "Circles" image: a $N \times N$ simulated image consisting of 12 circles that correspond to 12 active regions in 3 columns of 4 rows. The 4th column corresponds to an inactive region. The size parameters are $N = 64$ and $n = 256$.
- the "Whitcher" image: simulated data of Whitcher *et al.* (2005) - see description below -, the size parameters are $N = 64$ and $n = 128$.
- the MRSI image (see our description of this example in Section 2.1): the size parameters are $N = 32$ and $n = 4096$.
- the ONERA image: a multiband satellite image of remote sensoring measurements in various spectral bands of an area which contains roads, forests, vegetation, lakes and fields. The size parameters are $N = 64$ and $n = 128$ (frequencies).

For the reader's convenience we give a short description of the "Whitcher" image. Similarly to the "Circles" image, in 12 active regions in 3 columns of 4 rows (with 4th column corresponding to an inactive region), the simulated signals were generated using the difference of the exponentials function $S(t) = a + b(e^{-t/T_o} - e^{-t/T_i})$, with different values of T_i and T_o. The maximum signal amplitudes were normalized relative to the standard deviation of the additive Gaussian white noise such that the contrast-to-noise ratio in each row was 6,4,2, and 1. The spatial form of the active regions was a circular region of diameter 8 pixels, containing maximum intensity at each active region. This was surrounded by a Gaussian taper to zero within a square of 12 pixels, this providing a locally-varying contrast-to-noise ratio.

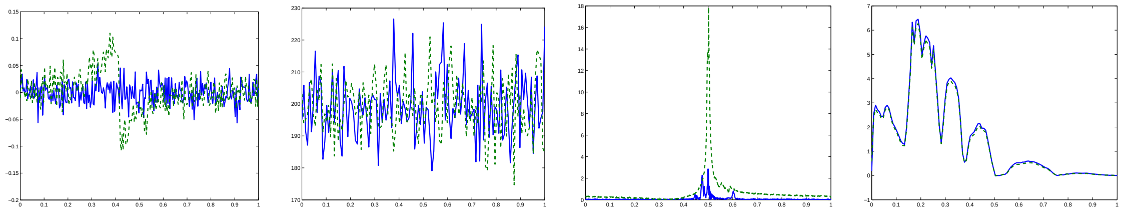
For each of this four examples, we display a typical temporal cut in Figure 4.2 and we also plot in Figure 4.3 each time two noisy curves corresponding to two pixels to give an idea of the signals that we have to cluster.

For each image, we have first applied wavelet hard tresholding using a universal threshold and a MAD estimator on the finest scale to estimate the variance of the wavelet coefficients. Then we applied the dimension reduction step by truncation as described in Section 3.2 on the union of the wavelet coefficients of all pixel intensity curves that survived this first hard thresholding. Note that in practice using the union amounts virtually to the aggregation estimator of Section 3.1 which has its justification only for the theoretical treatment of near-optimality of the latter one. In Figure 4.4, we plot the first 100 statistics $T_1^H, \ldots, T_{k\max}^H$ based on the Neyman's procedure (sorted in absolute value and decreasing order).



(a) (b) (c) (d)

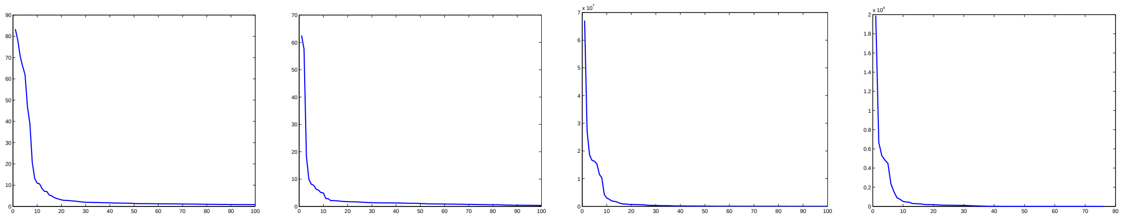
Figure 4.2: A temporal cut of the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image



(a) (b) (c) (d)

Figure 4.3: Noisy curves associated to some pixels of the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image

As explained by Bunea *et al.* (2006), the first very large values of these sorted statistics correspond to the null hypotheses to be rejected and thus to the most discriminative wavelet coefficients to separate the classes. In the FDR procedure, one has to choose a user-specified value $q \in (0, 1)$ which controls the ratio of the expected number of false rejections to the total number of rejections. However, in Buena *et al.* (2006) it is shown that the FDR procedure is consistent if we choose $q = q_p \to 0$ as $p \to +\infty$ where p is the number of observations used to compute the statistics

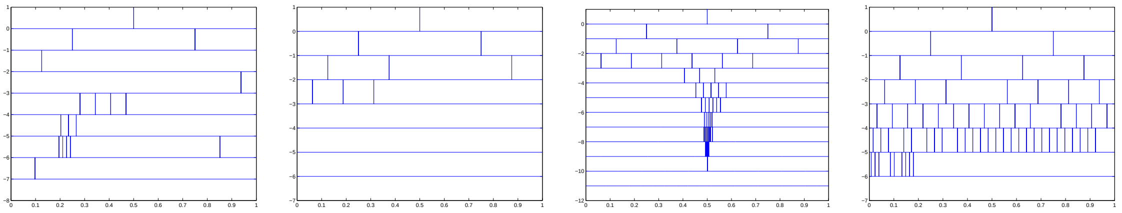


(a) (b) (c) (d)

Figure 4.4: Sorted test statistics $T_1^H, \dots, T_{k\text{max}}^H$ in absolute value and decreasing order for the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image

$T_1^H, \dots, T_{k\max}^H$. The rate at which the parameter q_p converges to zero depends on the rate at which the expected value μ_i (this is the notation in Bunea *et al.* 2006) of T_i^H converges to zero for the indices i which correspond to the index set I_0 of non-zero components of the parameters $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{k\max})$ (see the assumption **T** page 5 and Theorem 2.2 in Bunea *et al.* 2006). It is recommended to choose $q_p = p^{-\alpha}$ for some $\alpha > 0$, but no data-dependent choice for this parameter is proposed. For the four images, we chose $q = 0.05$ but this is obviously not an optimal tuning for this parameter. As we can see in Figure 4.4, the amplitude of these statistics are very different for the four images. Actually, the magnitude of the largest T_i^H's is proportional to the signal-to-noise (SNR) ratio of the curves that we have to cluster. For the MRSI image, the SNR is very high (see Figure 4.3c) and the amplitude of the largest test statistics is therefore extremely high ($\geq 10^{-7}$). For the Whitcher image, the SNR is very low (see Figure 4.3b) and the amplitude of the largest test statistics is therefore extremely low (≈ 60) as compared to the values of the test statistics in Figure 4.3c.

If we apply the FDR procedure with $q = 0.05$ to each of the four images, the number $\hat{k}$ of selected wavelet coefficients are 18 for the Circles image, 10 for the Whitcher image, 2629 for the MRSI image and 66 for the ONERA image. For the MRSI image, the selected $\hat{k}$ is obviously too large. Hence, based on visual inspection of the decay of the test statistics in Figure 4.4c, we prefer to choose $\hat{k} = 50$. The selected wavelet coefficients for each image are displayed in Figure 4.5.

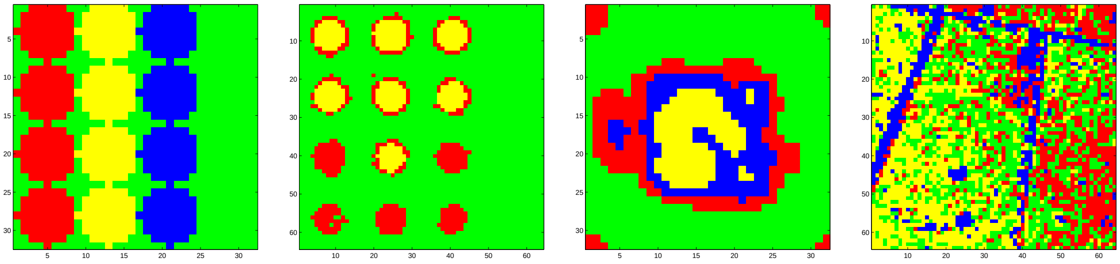


(a) (b) (c) (d)

Figure 4.5: Selected wavelet coefficients by the FDR procedure for the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image. Note that for the MRSI Image we have arbitrarily chosen the first 50 smallest p-values.

After applying the EM algorithm of Law *et al.* (2004), we obtain the following estimation $\hat{L}$ for the number of classes: 4 for the Circles image, 3 for the Whitcher image, 4 for the MRSI image and 6 for the ONERA image. For each image, the corresponding clustering at the pixel scale is given in Figure 4.6.

Note that the EM algorithm of Law *et al.* (2004) is initialised randomly which may give different results for the estimated number of classes and the final pixel scale clustering (especially for the



(a)

(b)

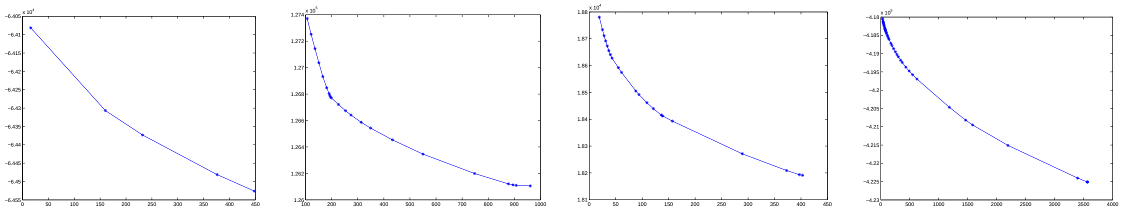
(c)

(d)

Figure 4.6: Clustering at the pixel scale for the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image.

Whitcher image). Even for the same estimated number of classes, the clustering at the pixel scale may be very different, and hence in practice it is highly recommended to re-launch the algorithm a couple of times. Note that this problem is shared by any EM algorithm and not specific to our set-up here.

Finally we apply the automatic method for choosing the parameter β for the RDP as described earlier in Section 3.3. For each image, we display the L-curve $(-\ell, m(\beta))$ in Figure 4.7 and the curve $(\beta, m(\beta))$ in Figure 4.8.



(a)

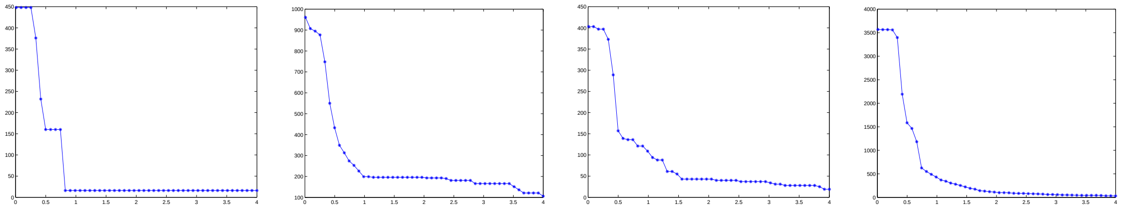
(b)

(c)

(d)

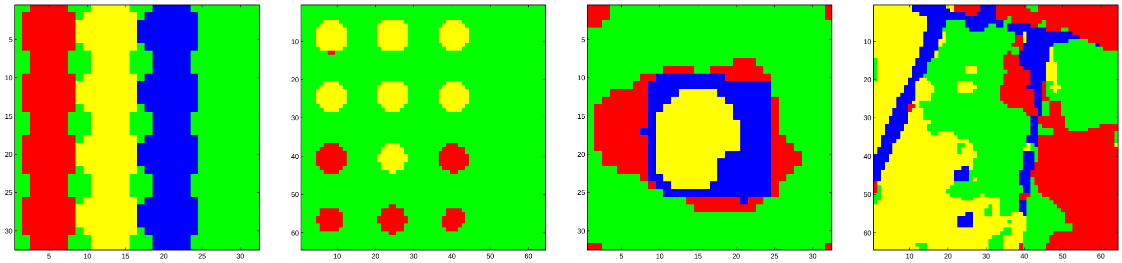
Figure 4.7: L-curve for the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image.

The selected dimension $\hat{m}$ beyond which the L-curve is considered as a linear function is 150 for the Circles image, 193 for the Whitcher image, 130 for the MRSI image and 700 for the ONERA image, and based on the heuristic of the slope we obtain the following estimation $\hat{\beta}$: 1.5222 for the Circles image, 1.8870 for the Whitcher image, $\hat{\beta} = 1.6908$ for the MRSI image, $\hat{\beta} = 2.1146$ for the ONERA image. A TI-version of the RDP with the above estimated $\hat{\beta}$'s for each image is given in Figure 4.9.



(a) (b) (c) (d)

Figure 4.8: Curve $(m(\beta), \beta)$ for the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image.



(a) (b) (c) (d)

Figure 4.9: TI RDP for the (a) Circles image, (b) Whitcher image, (c) MRSI Image, (d) ONERA image.

In summary, we observe that for the Circles image (a) our method of clustering already gives a perfect classification even without the final RDP step (which however does not take anything away from this result). Examples (b) and (d) do greatly benefit from the final RDP step (for example (d) it is even highly necessary), and we believe that the MRSI image (c) also gains clarity. Remember that the result at the pixel level gives many pixels that are spatially dispersed which for this example would not make much sense because we are trying to recover a specific segmentation of the tissues where the disease (cancer) has spread gradually (regions with a different severity).

5. CONCLUSION

We are perfectly aware of the fact that our approach is not the only possible one when trying to cluster images with a functional pixel intensity. However, we believe that it is a very appropriate *non-parametric* possibility which addresses simultaneously the problem of high dimensionality (spatial times functional dimension) and the drawbacks of working with a pixel-scale approach. To mention comparisons we have tested a variety of algorithms which turned out to be very sensitive to the high dimensionality of the problem, e.g. the work by Sugar and James (2003) for finding the number of clusters and the information criteria methods by Biernacki *et al.* (2000). Similarly, applying an approach of modelling spatial correlation (between pixels), as in kriging, would suffer from the too high spatial dimensionality (a possibility here would be to use reversible MCMC like Vannucci *et al.* (2005) which is however very cost-consuming).

Wavelets seem to offer clear advantages in addressing the functionality aspect of the problem, but as our various numerical examples have shown, there is obviously a necessity to use *multi-* over monoscale approaches such as in Whitcher *et al.* (2005).

For the spatial multiscale modelling, in this paper we have used an approach based on RDP motivated from the work by Kolaczyk *et al.* (2005) but one could also use other spatial models, e.g. Hidden Markov modelling. Comparing our method with those of a multiscale hidden Markov approach; see, e.g., Malfait and Roose (1997) or Choi and Baraniuk (2001), we note that both mentioned papers deal with *supervised* classification/segmentation. However, it would be interesting to replace the complexity-penalised RDP-step of our approach by using in particular the wavelet-domain hidden Markov tree, to see how the latter one would perform *in combination with our non-supervised* clustering step (after our non-parametric dimension reduction step).

Other comparisons could include the approach of Berlinet *et al.* (2005) which is however again on classification, i.e. a supervised method. Moreover, we found out that a criterion based on ordering the energy in the domain of the squared coefficients - as in Principal Components Analysis - would not necessarily have a good discriminating power and would in general not work in practice. Cf. also the example in Figure 3 of the paper by Law *et al.* (2004).

ACKNOWLEDGEMENTS

Jérémie Bigot and Rainer von Sachs would like to thank Anestis Antoniadis for his continuing hospitality during several visits at UJF in Grenoble. Anestis Antoniadis would like to thank Rainer von Sachs for excellent hospitality while visiting the Institute of Statistics, UCL to carry out this work. Fruitful discussions with Eric Kolaczyk, Hernando Ombao and Brandon Whitcher are gratefully acknowledged. An earlier version of this work has been presented at the 2006 Graybill Conference on “Multiscale Methods and Statistics” at the CSU, Fort Collins. Financial support form the IAP research network Nr. P5/24 of the Belgian government (Belgian Science Policy) is gratefully acknowledged.

APPENDIX

Proof of Proposition 3.1 The proof will follow from the following inequality. Assume first that T is such that $\xi_{\min}^{-1/2}\sqrt{2}U \leq T$ and let $\hat{\lambda}$ the optimal vector of weights that minimizes $\hat{S}(\lambda) + \text{pen}(\lambda)$ over $\Lambda_{M,T,2}$. For convenience of notation let $\tilde{f} = \hat{f}_{\hat{\lambda}}$ and $f_{\lambda} = \hat{f}_{\lambda}$. Then, for all $a > 1$ and for all integers $n \geq 1$ and $M \geq 2$ we have

$$\frac{1}{M}\sum_{\ell=1}^{L} M_\ell \mathbb{E}_{f^\ell}\|\tilde{f} - f^\ell\|_n^2 \quad \leq \quad \inf_{\lambda \in R^M}\left\{\frac{a+1}{a-1}\sum_{\ell=1}^{L}\frac{M_\ell}{M}\mathbb{E}_{f^\ell}\|f_\lambda - f^\ell\|_n^2 + \frac{16a^2}{a-1}\left(\frac{\sigma^2 U^2}{M\xi_{\min}}\right)\frac{2\log M + \log n}{n}\right\}$$
$$+\frac{(T+M^{-1/2})^2U^2}{n\sqrt{\pi(2\log M + \log n)}}. \tag{5}$$

To prove the above inequality, note first that by definition of $\tilde{f}$ we have

$$\hat{S}(\hat{\lambda}) + \frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}\ |\hat{\lambda}_j| \leq \hat{S}(\lambda) + \frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}\ |\lambda_j|$$

for all $\lambda \in \Lambda_{M,T,2}$, which can be rewritten as

$$\sum_{\ell=1}^{L}\frac{M_\ell}{M}\|\tilde{f} - f^\ell\|_n^2 + \frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}\ |\hat{\lambda}_j|$$

$$\leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\|\hat{f}_\lambda - f^\ell\|_n^2 + \frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}\ |\lambda_j| + \frac{2}{M}\sum_{j\in\mathcal{D}_{2,N}}\langle x_j - f^{c(j)}, \tilde{f} - f_\lambda\rangle_n\ .$$

Define the random variables

$$V_j = \frac{1}{n}\sum_{k=1}^{n}\hat{f}_j(t_k)(x_j(t_k) - f^{c(j)}(t_k)), \quad j \in \mathcal{D}_{2,N}\ ,$$

and let A be the event

$$A = \cap_{j\in\mathcal{D}_{2,N}}\{2|V_j| \leq \tau_{n,j}\}.$$

Since the estimates $\hat{f}_j$ are all computed on the training sample $\mathcal{D}_{1,N}$ which is independent of the learning one $\mathcal{D}_{2,N}$, and since the noise is Gaussian, we first have

$$\sqrt{n}V_j \sim N(0, \sigma_j^2\|\hat{f}_j\|_n^2), \quad j \in \mathcal{D}_{2,N}.$$

By the standard tail bound for the $N(0,1)$ distribution, we obtain

$$\begin{aligned} P(A^c) \ &\leq\ \sum_{j\in\mathcal{D}_{2,N}} P\{\sqrt{n}|V_j| \geq \sqrt{n}\tau_{n,j}/2\} \\ &\leq\ \sum_{j\in\mathcal{D}_{2,N}}\frac{4}{\sqrt{2\pi}}\frac{\sigma_j\|\hat{f}_j\|_n}{\sqrt{n}\tau_{n,j}}\exp\left(-\frac{n\tau_{n,j}^2}{8\sigma_j^2\|\hat{f}_j\|_n^2}\right) \leq \sum_{j\in\mathcal{D}_{2,N}}\frac{4}{\sqrt{2\pi}}\frac{\sigma\|\hat{f}_j\|_n}{\sqrt{n}\tau_{n,j}}\exp\left(-\frac{n\tau_{n,j}^2}{8\sigma^2\|\hat{f}_j\|_n^2}\right) \\ &\leq\ \frac{1}{Mn\sqrt{\pi(2\log M + \log n)}}. \end{aligned}$$

On the set A, we now have

$$\frac{2}{M}\sum_{j\in\mathcal{D}_{2,N}}\langle x_j - f^{c(j)}, \tilde{f} - f_\lambda\rangle_n = \frac{2}{M}\sum_{j\in\mathcal{D}_{2,N}}V_j(\hat{\lambda}_j - \lambda_j) \leq \frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}|\hat{\lambda}_j - \lambda_j|\tau_{n,j},$$

and therefore, still on A,

$$\sum_{\ell=1}^{L}\frac{M_\ell}{M}\|\tilde f-f^\ell\|_n^2 \leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\|f_\lambda-f^\ell\|_n^2+\frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}|\hat\lambda_j-\lambda_j|+\frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}|\lambda_j|-\frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}|\hat\lambda_j|.$$

By the triangle inequality we therefore have

$$\sum_{\ell=1}^{L}\frac{M_\ell}{M}\|\tilde f-f^\ell\|_n^2 \leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\|f_\lambda-f^\ell\|_n^2+2\frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}|\hat\lambda_j-\lambda_j|.$$

Since for all $j\in\mathcal{D}_{2,N}$, we have $\|\hat f_j\|_n^2\geq\xi_{\text{min}}>0$, it follows that

$$\xi_{\text{min}}^{-1}\|\tilde f-f_\lambda\|_n^2\geq\sum_{j\in\mathcal{D}_{2,N}}|\hat\lambda_j-\lambda_j|^2.$$

Using the fact that $\sum_\ell M_\ell/M=1$ and combining the above with the Cauchy-Schwarz inequality and triangle inequalities we find that on the set A

$$\begin{aligned}
\sum_{\ell=1}^{L}\frac{M_\ell}{M}\|\tilde f-f^\ell\|_n^2 &\leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\|f_\lambda-f^\ell\|_n^2+2\frac{1}{M}\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}|\hat\lambda_j-\lambda_j|\\
&\leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\|f_\lambda-f^\ell\|_n^2+\frac{1}{M}2\xi_{\text{min}}^{-1/2}\sqrt{\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}^2}\|\tilde f-f_\lambda\|_n\\
&\leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\|f_\lambda-f^\ell\|_n^2+2\frac{1}{M}\xi_{\text{min}}^{-1/2}\sqrt{\sum_{j\in\mathcal{D}_{2,N}}\tau_{n,j}^2}\sum_{\ell=1}^{L}\frac{M_\ell}{M}(\|\tilde f-f^\ell\|_n+\|f_\lambda-f^\ell\|_n)\\
&\leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\|f_\lambda-f^\ell\|_n^2+2\xi_{\text{min}}^{-1/2}\tau_nM^{-1/2}\sum_{\ell=1}^{L}\frac{M_\ell}{M}(\|\tilde f-f^\ell\|_n+\|f_\lambda-f^\ell\|_n)\\
&\leq \sum_{\ell=1}^{L}\frac{M_\ell}{M}\left(\|f_\lambda-f^\ell\|_n^2+2\xi_{\text{min}}^{-1/2}\tau_nM^{-1/2}(\|\tilde f-f^\ell\|_n+\|f_\lambda-f^\ell\|_n)\right),
\end{aligned}$$

where

$$\tau_n=2\sqrt{2}U\sigma\sqrt{\frac{2\log M+\log n}{n}}.$$

Taking each term on the left hand side and each term within the parenthesis in the above inequality, the resulting inequality is of the form $v^2\leq c^2+vb+cb$ with $v=\|\tilde f-f^\ell\|_n$, $b=2\xi_{\text{min}}^{-1/2}\tau_nM^{-1/2}$ and $c=\|f_\lambda-f^\ell\|_n$. We then obtain, using an argument similar to the one of Bunea *et al.* (2006),

$$\sum_{\ell=1}^{L}\mathbb{E}\left\{\frac{M_\ell}{M}\|\tilde f-f^\ell\|_n^2I_A\right\}\leq\inf_{\lambda\in\Lambda_{M,T,2}}\left\{\frac{a+1}{a-1}\sum_{\ell=1}^{L}\mathbb{E}\left\{\frac{M_\ell}{M}\|f_\lambda-f^\ell\|_n^2\right\}+\frac{2a^2}{(a-1)\xi_{\text{min}}M}\tau_n^2\right\},\quad\forall\, a>1.$$

Now, using the definition of $\Lambda_{M,T,2}$ and the Cauchy-Schwarz inequality we have

$$\sum_{\ell=1}^{L}\frac{M_\ell}{M}\|\tilde f-f^\ell\|_\infty\leq U(\sum|\hat\lambda_j|+1)\leq(\sqrt{M}T+1)U,$$

which gives

$$\begin{aligned}\sum_{\ell=1}^{L}\mathbb{E}\left\{\frac{M_\ell}{M}\|\tilde f - f^\ell\|_n^2\right\} &\le \sum_{\ell=1}^{L}\mathbb{E}\left\{\frac{M_\ell}{M}\|\tilde f - f^\ell\|_n^2 I_A\right\} + (\sqrt{M}T+1)^2U^2P(A^c)\\ &\le \inf_{\lambda\in\Lambda_{M,T,2}}\left\{\frac{a+1}{a-1}\sum_{\ell=1}^{L}\mathbb{E}\left\{\frac{M_\ell}{M}\|f_\lambda - f^\ell\|_n^2\right\} + \frac{2a^2\tau_n^2}{(a-1)\xi_{\min}M}\right\}\\ &\quad + \frac{(T+M^{-1/2})^2U^2}{n\sqrt{\pi(2\log M+\log n)}},\quad \forall\, a>1.\end{aligned}$$

The above inequality is in fact valid not only within the set $\Lambda_{M,T,2}$ but on the entire $\mathbb{R}^M$, because the value of the whole expression under the infimum in the inequality for $\lambda = 0$ is strictly smaller than the value of the same expression for any $\lambda \notin \Lambda_{M,T,2}$. The requested inequality (5) is then proved by substituting the value of τ_n^2.

To end the proof of the proposition, note first that by our assumptions one has trivially that $T > M^{-1/2}$. This implies that the last summand in (5) is $O(1/n)$. The proof is then completed by using an argument similar to the one used by Bunea *et al.* (2006) in the proof of their Corollary 3.4.

Proof of Proposition 3.2 The proof follows the lines of Lavielle (2005). For any $m \ge 1$, and any $\mathbf{x}$, let $\hat m(\beta) = m$. Then we have

$$\begin{aligned}-\ell(\mathbf{x}(\cdot)|m_i) + \beta\ m &< \min_{k>m}(-\ell(\mathbf{x}(\cdot)|k) + \beta\ k)\\ -\ell(\mathbf{x}(\cdot)|m_i) + \beta\ m &< \min_{k<m}(-\ell(\mathbf{x}(\cdot)|k) + \beta\ k)\end{aligned}$$

Thus β satisfies

$$\max_{k>m}\frac{-\ell(\mathbf{x}(\cdot)|m) + \ell(\mathbf{x}(\cdot)|k)}{k-m} < \beta < \min_{k<m}\frac{-\ell(\mathbf{x}(\cdot)|k) + \ell(\mathbf{x}(\cdot)|m)}{m-k},$$

which concludes the proof.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995), Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Hypothesis Testing, *J. R. Statist. Soc., B*, **57**, 289–300.

Berlinet, A., Biau, G., and Louvière, L. (2005), Functional classification with wavelets, Technical report, Department of Statistics, University of Montpellier, France.

Biernacki, C. , Celeux, G. & Govaert , G. (2000), Assessing a Mixture Model for Clustering with the IntegratedCompleted Likelihood, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22** (7), 719–725.

Birgé, L., & Massart, P. (1998), Minimum contrast estimator on sieves: Exponential bounds and rates of convergence, *Bernoulli*, **4**, 329–375.

Birgé, L., & Massart, P. (2001), A generalized Cp criterion for Gaussian model selection. Technical report 647, Universités de Paris 6 et Paris 7, France.

Bouman, C. & Shapiro, M. (1994), A multiscale random field model for Bayesian image segmentation, *IEEE Trans. Image Processing*, **3** (2), 162–177.

Bunea, F. ,Tsybakov , A., & Wegkamp, M. (2006), Aggregation for Gaussian regression, *Annals of Statistics*, (in press).

Bunea, F., Wegkamp, M. & Auguste, A. (2006), Consistent Variable Selection in High Dimensional Regression via Multiple Testing, *Journal of Statistical Planning and Inference* (in press).

Castro, R. , Willett , R. & Nowak, R. (2004), Coarse-to-Fine Manifold Learning, in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Montreal, Canada.

Choi, H. & Baraniuk, R. G. (2001), Multiscale image segmentation using wavelet-domain hidden Markov models, *IEEE Trans. Image Processing*, **10** (9), 1309–1321.

Donoho, D. L. (1997), Dyadic Cart and Ortho-Bases: a connection, *Annals of Statistics*, **25**, 1870–1911.

Donono, D. L. (1999), Wedgelets: Nearly-minimax estimation of edges, *Annals of Statistics*, **27**, 859–897.

Donoho, D. L. & Johnstone, I.M. (1998), Minimax estimation via wavelet shrinkage, *Annals of Statistics*, **26**, 879–921.

Fan, J. (1996), Test of significance based on wavelet thresholding and Neyman's truncation, *J. Am. Statist. Assoc.*, **91**, 674–688.

Gey, S. & Lebarbier, E. (2003), Using CART to detect multiple changepoints in the mean, Technical report, University of Paris 5, Paris, France.

Gulliksson, M. & Wadin, P.-A. (1998), Analyzing the nonlinear L-curve, Technical Report, Department of Computing Science, University of Umea, Sweden.

Johnstone, I. M., & Silverman, B. (1997), Wavelet threshold estimators for data with correlated noise, *J. R. Statist. Soc., B*, **59**, 319–351.

Kohler, M. (2003), Nonlinear orthogonal series estimates for random design regression, *Journal of*

Statistical Planning and Inference, **115**, 491–520.

Kolaczyk, E., Ju, J., & Gopal, S. (2005), Multiscale multigranular statistical image segmentation, *J. Am. Statist. Assoc.*, **472**, 1358–1369.

Korostelev, A.P. & Tsybakov, A. (1993). *Minimax Theory of Image Reconstruction*, Springer-Verlag, New York.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem, *Signal Processing*, **85** (8), 1501–1510.

Law, M., Figueiredo, M., & Jain, A. (2004). Simultaneous Feature Selection and Clustering Using Mixture Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26** (9),1154–1166.

Malfait, M. and Roose, D. (1997). Wavelet-based image de- noising using a Markov random field a priori model, *IEEE Trans. Image Processing*, **6**, 549–565.

Meyerand, M. E, Pipas, J. M, Mamourian, A., Tosteson, T. D. & Dunn, J. F. (1999). Classification of biopsy-confirmed brain tumors using single-voxel MR spectroscopy, *American Journal Neuroradiology*, **20**,117–123.

Nemirovski, A. (2000). Topics in Non-parametric Statistics, In: *Ecole dEté de Probabilités de Saint-Flour XXVIII - 1998*, Lecture Notes in Mathematics, vol. **1738**, 85–277, Springer, New York.

Sugar, C. and James, G. (2003). Finding the Number of Clusters in a Data Set : An Information Theoretic Approach, *J. Am. Statist. Assoc.*, **98**, 750–763.

Vannucci, M., Sha, N. & Brown, P. J. (2005). NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection, *Chemometrics and Intelligent Laboratory Systems*, **77**, 139–148.

Whitcher, B., Schwarz, A. J., Barjat,H., Smart, S. C., Grundy, R. I. & James, M. F. (2005). Wavelet-Based Cluster Analysis: Data-Driven Grouping of Voxel Time-Courses with Application to Perfusion-Weighted and Pharmacological MRI of the Rat Brain, *NeuroImage*, **24** (2), 281–295.