Model selection using wavelet decomposition and applications*

Anestis Antoniadis[1]
Laboratoire LMC - IMAG
Université Joseph Fourier
BP 53, 38042 Grenoble Cedex, France

Irène Gijbels[2]
Institut de Statistique
Université Catholique de Louvain
B-1348 Louvain-la-Neuve, Belgium

Gérard Grégoire[1]
Laboratoire LMC - IMAG
Université Joseph Fourier
BP 53, 38042 Grenoble Cedex, France

Abstract

In this paper we discuss how to use wavelet decomposition to select a regression model. The methodology relies on a minimum description length criterion which is used to determine the number of non-zero coefficients in the vector of wavelet coefficients. Consistency properties of the selection rule are established and simulation studies reveal information on the distribution of the minimum description length selector. We then apply the developed model selection rule to a few testing problems, among which testing for pure white noise. The power of this test is investigated via simulation studies. Furthermore, the selection criterion is applied to testing for no effect in nonparametric regression and for martingale structure in time series.

1 Introduction

1.1 Outline

One of the most common questions arising in statistics is to select an appropriate model for a given data set. There is a variety of possible situations, depending on the particular

* Part of this work was done while the second author was visiting the LMC - IMAG Laboratory in Grenoble. The LMC - IMAG Laboratory is grateful acknowledged for its hospitality and support.

[1] Supported by CNRS-INRIA, Program 6, IDOPT Project.

[2] Partially supported by 'Programme d'Action de Recherche Concertée', N^o 93-98 – 164.

AMS 1991 subject classification. Primary: 62G07. Secondary: 62G10.

Key words and phrases: consistency, hypothesis testing, minimum description length criterion, model selection, simulations, wavelet decomposition.

phenomenon under study for which the data were collected. We give a few examples. Often first interest goes in determining whether or not a covariate (or a set of covariates in the multi-dimensional case) affects a response variable in a regression setting. Related problems are the selection of the order of an autoregressive model, the determination of a particular martingale structure in time series and the comparison of regression functions.

In this paper we first discuss a particular methodology for estimating a regression function nonparametrically. The basic ingredients for the methodology are a wavelet decomposition of the unknown function and a selection of the non-zero coefficients in this decomposition. Fan (1995) also relies on wavelet techniques to develop testing procedures. In that paper wavelet-based goodness-of-fit tests are compared with the classical nonparametric tests, such as for example the Kolmogorov-Smirnov test and the Cramer-von Mises test, under a variety of alternatives. These comparisons revealed that wavelet-based tests perform in general at least as well as the classical tests, but are more powerful for detecting sharp aberrations and subtle changes in distribution functions. This reflects a first reason for using wavelet-based methods. Moreover for a large variety of functions, wavelet decompositions typically only have a few non-zero coefficients. Hence, such decompositions are very well suited for model selection or testing problems involving these functions.

As a selection criterion we use the Minimum Description Length (MDL) criterion which was proposed by Rissanen (1983, 1984) and recently discussed extensively in Saito (1994). The selection methodology developed for regression estimation can then be applied to several other situations, such as the testing situations mentioned above. To illustrate this we discuss a few specific domains of applications.

The paper is organised as follows. In the next section we describe briefly the idea of wavelet decomposition and its application in estimating a regression function. In Section 2 we discuss the Minimum Description Length criterion and point out how it relates to other commonly-used selection criteria. Some asymptotic properties of the MDL-selector are provided in Section 3. It is quite difficult to determine the distribution of the MDL-selector. To gain some insight in this problem some simulation studies are carried through in Section 4. In that section we also show in detail how to apply the proposed methodology in some applications. We close with a discussion and some open questions in section 5. Detailed proofs are relegated to an Appendix.

1.2 Regression model and wavelet decomposition

We consider the following fixed design model

$$Y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, N \tag{1.1}$$

with $t_i = (i-1)/N$, and where ε_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. So, the Y_i are noisy measurements of the mean regression function f taken at equidistant fixed design points t_i (without loss of generality within $[0, 1]$).

There are several techniques available for estimating the unknown function f based on the observations $(t_1, Y_1), \cdots, (t_N, Y_N)$. See the books by Eubank (1988), Müller (1988), Härdle (1990), Wahba (1990), Green and Silverman (1994) among others. In this paper we use wavelet techniques. For an introduction to these techniques and its applications see for example Daubechies (1992) and Meyer (1993).

We consider the simple model (1.1) only for convenience. The restriction of working with equidistant fixed design points t_i can be dropped, and one can work with the situation were the design points t_i are either nonrandom and unequally spaced or random and satisfying some regularity conditions (see e.g. Neumann and Spokoiny (1993) and Antoniadis, Grégoire and McKeague (1994)). The assumption on the random errors ε_i can also be relaxed: it suffices to require that they are centered, independent and not necessarily identically distributed.

We assume that the sample size is a power of two, i.e. $N = 2^n$. Here again this restriction is assumed for convenience of presentation, and can be relaxed. See for example McGill and Taswell (1993) and Herley (1995). In order to develop our theory we will specify some fixed class $\mathcal{F}$ of functions to which f is supposed to belong. We shall assume the existence of wavelet decompositions of f and the description of $\mathcal{F}$ by wavelet coefficients. Our interest is rather in the compression of the wavelet decomposition. Recall that a *wavelet decomposition* of a function f defined on $\mathbb{R}$ is an expression of the form

$$f = \sum_{j \in \mathbb{Z}} \sum_{\ell \in \mathbb{Z}} a_{j,\ell} \psi_{j,\ell}, \tag{1.2}$$

where the coefficients $a_{j,\ell} := a_{j,\ell}(f)$ depend on f, and the functions $\psi_{j,\ell}$ are defined as $\psi_{j,\ell}(x) := 2^{j/2}\psi(2^j(x - \ell/2^j))$, the dyadic dilates (by 2^j) and translates (by $\ell/2^j$) of a single function ψ called a *wavelet*. Higher values of j correspond to higher frequency scales or higher resolution features of f.

Of course, not every choice of ψ allows one to decompose general functions f as in (1.2). Some examples of functions ψ that *can* be used in (1.2) are the orthogonal wavelets of Meyer (1990) and Daubechies (1992).

For notational brevity, we will sometimes index the $(j, \ell)^{th}$ term of (1.2) by a single I. Then, denoting by $\mathcal{D}$ the set of all relevant indexes, (1.2) can be rewritten as

$$f = \sum_{I \in \mathcal{D}} a_I \psi_I. \tag{1.3}$$

Our list of candidate models for f in (1.1) will consist of functions in $\mathcal{F}$ for which the infinite sum (1.2) is replaced by a finite sum in an efficient manner. We can justify this assumption as follows. For a given positive integer K let Σ_K denote the non-linear manifold consisting of all functions $S = \sum_{I \in \Lambda} a_I \psi_I$ where the set Λ consists of at most K indexes (from arbitrary levels j). We approximate functions f by elements from Σ_K. The error of approximating f in a quasi-normed space $\mathcal{X}$ is

$$\sigma_K(f)_{\mathcal{X}} := \inf_{S \in \Sigma_K} \|f - S\|_{\mathcal{X}}. \tag{1.4}$$

It is now known that with certain conditions on ψ we have, for each $0 < p < \infty$ and for a certain range of α, that

$$\sum_{K=1}^{\infty} [K^\alpha \sigma_K(f)_{L_p(\mathbb{R})}]^\tau \frac{1}{K} < \infty \quad \Longleftrightarrow \quad f \in \mathcal{B}^\alpha \tag{1.5}$$

where $\mathcal{B}^\alpha := \mathcal{B}^\alpha_\tau(L_\tau)$, $\tau := (\alpha + 1/p)^{-1}$ are the Besov spaces. There is in fact no simple characterization of functions with approximation order $\sigma_K(f)_{L_p(\mathbb{R})}$ precisely $O(K^{-\alpha})$, however (1.5) is a close substitute. In one direction, it does follow from (1.5) that each $f \in \mathcal{B}^\alpha$ can be approximated with order of approximation $\sigma_K(f)_{L_p(\mathbb{R})} = O(K^{-\alpha})$:

$$\sigma_K(f)_{L_p(\mathbb{R})} \leq CK^{-\alpha}||f||_{\mathcal{B}^\alpha}. \tag{1.6}$$

Conversely, we know that any $f \in L_p$ with this approximation order is in $\mathcal{B}^{\alpha-\epsilon}$ for all $\epsilon > 0$.

The proof of (1.6) is provided in Devore, Jawert and Popov (1988) and implicitly gives a numerical algorithm for determining "good" approximations $S \in \Sigma_K$. Actually, there are two algorithms. The simplest of these chooses the K largest in absolute value coefficients in the decomposition (1.2). Note however that this choice results in constants in the bound of the error of approximation (1.6) that depend on p.

According to the above considerations, the assumption that the unknown function f in (1.1) belongs to $\Sigma_K \subset \mathcal{B}^\alpha$ is not very restrictive in practice, if K is chosen large enough. We further assume that the elements of the basis have been ordered, such that the coefficients in the decomposition have a natural order: the largest coefficients are connected with the first basis functions.

All of our results are obtained under the assumption that such a finite dimensional model generates the data under discussion. There is no doubt that such an assumption can be weakened, but the derivations of similar results are expected to be much more involved. More discussion on finite-dimensional models vs. infinite-dimensional models can be found in Section 5.

With this background we are now ready to describe the associated model selection procedure. We start with a large model (K large) which is assumed to be correct, but possibly redundant. We then subsequently eliminate coefficients which appear to be superfluous. The decomposition (1.2) of the function f involves the unknown coefficients $a_{j,\ell}$. In order to estimate f one needs to get access to such unknown coefficients. An approach to this issue is provided in Antoniadis (1994), who shows that certain coefficients in the decomposition can be approximated by $f(\ell/2^n)/\sqrt{N}$, when using regular enough coiflets as orthogonal compactly supported wavelets. He also determines the error of this approximation under certain smoothness conditions on the underlying function space. The advantage of this approach is that the sampled values $f(\ell/2^n)/\sqrt{N}$ can be assessed via the observed values $Y_{\ell+1}$, leading to an estimator for f. See Antoniadis (1994) for more details. Here we follow this approach and will, according to the approximation with $f(\ell/2^n)/\sqrt{N}$, work with the data $\beta_{N,\ell} = Y_{\ell+1}/\sqrt{N}$. By performing the pre-conditionned, interval-adapted, pyramid wavelet transform of Cohen, Daubechies and Vial (1993) to the data $\beta_{N,\ell}$, one obtains the output vector $\boldsymbol{d}$, containing the empirical wavelet coefficients. This leads to the noised wavelet decomposition of f:

$$\boldsymbol{d} = \boldsymbol{\alpha}^{(K)} + \boldsymbol{\varepsilon}, \tag{1.7}$$

where $\boldsymbol{\alpha}^{(K)} \in I\!R^N$ denotes the vector of expansion coefficients of f with only its first K coefficients non-zero (according to the assumptions on f), and where ϵ_ℓ, the components of $\boldsymbol{\varepsilon}$, are i.i.d. $\mathcal{N}(0, \sigma^2/N)$. Thus the original formulation of the nonparametric estimation problem (1.1) is equivalent with the one stated in (1.7).

The model in (1.7) actually reveals a whole class of models, namely all models with a vector $\boldsymbol{\alpha}^{(K)} \in I\!R^N$ for which only the first K out of N coefficients are possibly non-zero. Hence, this includes all such N-dimensional vectors with only k $(0 \leq k \leq K)$ non-zero coefficients located somewhere at the first K positions. Solving problem (1.7), or equivalently problem (1.1), can be done by selecting an appropriate model from the above class of models and for this model estimate the vector of coefficients. Denote by $\boldsymbol{\alpha}^{(k)}$ such a vector of coefficients, having only k non-zero coefficients at the first K positions. Problem (1.7) then consists of selecting an integer k, $0 \leq k \leq K$, and estimating the associated vector of coefficients $\boldsymbol{\alpha}^{(k)}$. In Section 2 we explain how to select k in a data-driven manner, as well as how to estimate $\boldsymbol{\alpha}^{(k)}$.

2 Minimum Description Length criterion

There are several techniques which may be used to select an appropriate model from a class of available models to fit a given data set. Most of these selection rules are obtained by estimating risk functions. In particular, Mallows' (1973) C_p, Akaike's (1970, 1973) information criteria AIC and FPE, Allen's (1971) PSS and Schwarz's (1978) BIC have been recommended for use in model selection. See Nishii (1984) for a brief review of these criteria. Numerical studies comparing the performance of the different selection information criteria are given in Sugiura (1979) and Hashimoto *et al.* (1981).

The Minimum Description Length (MDL) criterion was proposed by Rissanen (1983, 1984). The criterion is based on the fact that the *best model* within the class of all models is the one with the shortest description of the data and the model itself. The description length is defined as the sum of the codelength needed for encoding the candidate model, and the codelength for encoding the data using this particular model. A nice discussion of the minimum description length criterion and its motivation from information-theoretic considerations can be found in Saito (1994). That paper also provides ample of examples where the minimum description length principle can be used. This minimum complexity or minimum description length principle provides another class of versatile criteria, having properties different than those of AIC and FPE, and particulary useful for model selection in general. Saito (1994) also describes how the minimum description length principle can be applied in nonparametric regression, but no consistency issues are considered in his work. We next explain this selection rule for our class of models (1.7).

Consider a fixed model k out of the class of models (1.7). Then, an estimator for the vector of coefficients $\boldsymbol{\alpha}^{(k)}$ is given by

$$\widehat{\boldsymbol{\alpha}}^{(k)} := \mathrm{argmin}_{\boldsymbol{\alpha}^{(k)} \in \mathbb{R}^N} \|\boldsymbol{d} - \boldsymbol{\alpha}^{(k)}\|^2. \tag{2.1}$$

Clearly, this minimum is achieved by taking the largest k coefficients in absolute value of the observed vector $\boldsymbol{d}$. So, we can write

$$\widehat{\boldsymbol{\alpha}}^{(k)} = \boldsymbol{\Theta}^{(k)}\boldsymbol{d}, \tag{2.2}$$

where $\boldsymbol{\Theta}^{(k)}$ denotes the operation which keeps the largest k coefficients in absolute value intact and sets all other coefficients equal to zero.

The MDL selection algorithm of Saito selects the model for which the Approximate

Minimum Description Length (AMDL)

$$\text{AMDL}(k) := 1.5k \log_2 N + 0.5N \log_2 \|(I - \boldsymbol{\Theta}^{(k)})\boldsymbol{d}\|^2$$

is minimized. Here I is a $N \times N$ matrix corresponding to the N-dimensional identity operator. The specific choice of the constants 1.5 and 0.5 in the above criterion is fully discussed in Saito (1994). The first term represents the codelength for encoding a k-dimensional model, while the second term is the codelength for encoding the data using this particular model under a Gaussian assumption for the errors. Hence, the selected model is the model indexed by $\widehat{k}$ for which

$$\text{AMDL}(\widehat{k}) = \min_{0 \le k \le K} \left(1.5k \log_2 N + 0.5N \log_2 \|(I - \boldsymbol{\Theta}^{(k)})\boldsymbol{d}\|^2\right). \tag{2.3}$$

We refer to $\widehat{k}$ as the AMDL-selector. Note that selection of k not only involves selection of the k largest coefficients in magnitude, but also of their positions, somewhere among the K first positions, in a N-dimensional vector.

This selection criterion is closely related to the BIC information criterion and more specifically to its generalization GIC, used particularly in linear models and models involving time series. To see this note that the term $\log_2 \|(I - \boldsymbol{\Theta}^{(k)})\boldsymbol{d}\|^2$ represents the difference between the observed data vector $\boldsymbol{d}$ and the fitted vector $\widehat{\boldsymbol{\alpha}}^{(k)} = \boldsymbol{\Theta}^{(k)}\boldsymbol{d}$, i.e. is the residual from the considered regression model indexed by k. The GIC information criterion would select the model indexed by k for which

$$\text{GIC}(k) = N \log \text{RSS}(k) + a_N k,$$

with $\text{RRS}(k)$ the residual sum of squares after fitting model k, is minimized. The sequence $a_N > 0$ is such that $\lim_{N\to\infty} a_N = +\infty$ and $\lim_{N\to\infty} N^{-1} a_N = 0$. See also Nishii (1984). Taking $a_N = 3 \log N$ we see the correspondence between the AMDL selection rule and the above GIC selection rule.

Interestingly, as pointed out by a referee, the AMDL criterion is also closely connected to the denoising algorithm of Donoho, Johnstone, Kerkyacharian and Picard (1995) by hard thresholding of wavelet coefficients. To see this consider the finite differences of

AMDL(k) with respect to k. For k large enough, it is not difficult to see that these differences behave like

$$\mathrm{AMDL}(k+1) - \mathrm{AMDL}(k) \sim -\frac{N}{2} \frac{d_{(k)}^2}{\|(I - \boldsymbol{\Theta}^{(k+1)})\boldsymbol{d}\|^2} + \frac{3}{2} \log_2 N.$$

This means, that for any k less than (resp. greater than) $\hat{k}$, the AMDL functional decreases (resp. increases), where the optimal value $\hat{k}$ is the largest $k \leq K$ such that

$$|d_{(k)}| > \sqrt{\frac{\|(I - \boldsymbol{\Theta}^{(k+1)})\boldsymbol{d}\|^2}{N}} \sqrt{3 \log_2 N} = \hat{\sigma}\sqrt{3 \log_2 N}.$$

In the next two sections we investigate the performance of the above AMDL selection rule. We first of all evaluate the quadratic risk and show that it tends to zero as N tends to infinity. In Section 4 we first study the distribution of the AMDL-selector and then apply the selection criterion in various testing problems.

3 Consistency results for the AMDL-selector

Recall that there are two important issues in the selection of a particular model indexed by k. There is first of all the k number of largest coefficients in absolute value among the first K coefficients, and secondly there is their positions among these first K positions. For a fixed k we have a whole class of models $\underline{j} := \{j_1, \cdots, j_k\}$ with j_i, $i = 1, \cdots, k$, referring to for example the position of the i^{th} largest coefficient in magnitude. For a fixed k we denote the class of possible models by J_k. Hence, the estimated vector $\widehat{\boldsymbol{\alpha}}^{(k)}$ defined in (2.1) can be rewritten. Indeed, note that

$$\begin{aligned}
\min_{\boldsymbol{\alpha}^{(k)} \in I\!R^N} \|\boldsymbol{d} - \boldsymbol{\alpha}^{(k)}\|^2 &= \min_{\underline{j} = \{j_1, \cdots, j_k\} \in J_k} \min_{\boldsymbol{\alpha}(j_1, \cdots, j_k) \in I\!R^N} \left(\sum_{\ell=1}^{k} (d_{j_\ell} - \alpha_{j_\ell})^2 + \sum_{s \notin \{j_1, \cdots, j_k\}} d_s^2 \right) \\
&:= \min_{\underline{j} = \{j_1, \cdots, j_k\} \in J_k} S(\underline{j}), \qquad (3.1)
\end{aligned}$$

where we used the notation $\boldsymbol{\alpha}(j_1, \cdots, j_k)$ to indicate that we have an N-dimensional vector $\boldsymbol{\alpha}$ with only non-zero coefficients on the positions $j_1, \cdots, j_k$. So the second minimization, i.e. the one inside $S(\underline{j})$, is a minimization over all N-dimensional vectors having only non-zero coefficients on the positions $j_1, \cdots, j_k$. Hence, the estimated vector can be written as

$$\widehat{\boldsymbol{\alpha}}^{(k)} = \mathrm{argmin}_{\underline{j} \in J_k} S(\underline{j}). \tag{3.2}$$

Suppose that the model indexed by k_0 and with positions $\underline{j_0} = \{j_{01}, \cdots, j_{0k_0}\}$ is the true unknown model. Then $\boldsymbol{\alpha}^{(K)}$ is a N-dimensional vector with only k_0 non-zero coefficients located at the positions $j_{01}, \cdots, j_{0k_0}$ among the first K positions, and $\boldsymbol{\alpha}^{(K)} = \boldsymbol{\alpha}^{(k_0)}$. We assess the goodness of the AMDL selection criterion in terms of the quadratic risk

$$R_N = E\|\boldsymbol{\alpha}^{(K)} - \widehat{\boldsymbol{\alpha}}^{(\widehat{k})}\|^2, \tag{3.3}$$

which expresses the difference of the estimated vector of coefficients and the true vector in quadratic mean.

This risk R_N can be rewritten as follows. Denote by $\widehat{\underline{j}} \in J_{\widehat{k}}$ the positions selected by the AMDL-criterion. Then,

$$\begin{aligned} R_N &= \sum_{k=0}^{K} E\left(\|\boldsymbol{\alpha}^{(K)} - \widehat{\boldsymbol{\alpha}}^{(k)}\|^2 \cdot I_{\{\widehat{k}=k\}}\right) \\ &= \sum_{k=0}^{K} \sum_{\underline{j} \in J_k} E\left(\|\boldsymbol{\alpha}^{(K)} - \widehat{\boldsymbol{\alpha}}_{\underline{j}}^{(k)}\|^2 \cdot I_{\{\widehat{\underline{j}}=\underline{j}\}}\right), \end{aligned} \tag{3.4}$$

where $\widehat{\boldsymbol{\alpha}}_{\underline{j}}^{(k)}$ denotes the estimated vector obtained from the data vector $\boldsymbol{d}$ by taking in $\boldsymbol{d}$ only the coefficients on positions $j_1, \cdots, j_k$ and putting all other coefficients equal to zero. A "marginal" risk function is popping up in the above expression, namely

$$R_N(\underline{j}) := E\left(\|\boldsymbol{\alpha}^{(K)} - \widehat{\boldsymbol{\alpha}}_{\underline{j}}^{(k)}\|^2 \cdot I_{\{\widehat{\underline{j}}=\underline{j}\}}\right), \tag{3.5}$$

which assumes k as well as the positions $\underline{j} = \{j_1, \cdots, j_k\}$ to be fixed.

We will evaluate the risk function R_N via the risk function $R_N(\underline{j})$. It is easier to study the latter risk function since it involves the estimated vector $\widehat{\boldsymbol{\alpha}}_{\underline{j}}^{(k)}$ which can be written as

$$\widehat{\boldsymbol{\alpha}}_{\underline{j}}^{(k)} := M_{\underline{j}}^{(k)} \boldsymbol{d}, \tag{3.6}$$

where now $M_{\underline{j}}^{(k)}$ is a linear operator. More precisely $M_{\underline{j}}^{(k)}$ is a $N \times N$ matrix with non-zero elements (one's) only on the diagonal on the positions $j_1, \cdots, j_k$. In contrast, the risk function R_N involves the estimated vector $\widehat{\boldsymbol{\alpha}}^{(k)} = \boldsymbol{\Theta}^{(k)}\boldsymbol{d}$ as defined in (2.2) in which the non-linear operator $\boldsymbol{\Theta}^{(k)}$ appears.

Shibata (1976) used similar measures of goodness-of-fit in his study on the selection of the order of an autoregressive model by AIC. Our study of the risk functions R_N and its marginals $R_N(\underline{j})$ is in the same spirit as the investigation of the risks in Nishii (1984) in a situation of selecting variables in a linear multiple regression problem.

We will now establish the asymptotic properties of the risk functions. Some extra notation is needed. Consider the set J^2 of all models including the true model $\underline{j_0}$, with number of non-zero coefficients k_0, i.e.

$$J^2 = \bigcup_{k=k_0}^{K} \{\underline{j} \in J_k : \underline{j} \supseteq \underline{j_0}\}. \tag{3.7}$$

Similarly, let J^1 denote the set of all possible models which do not include $\underline{j_0}$ and are different from $\underline{j_0}$, i.e.

$$J^1 = \bigcup_{k=0}^{K} \{\underline{j} \in J_k : \underline{j} \not\supseteq \underline{j_0}\}. \tag{3.8}$$

Clearly, the set of all models is the union of J^1 and J^2.

Convergence of the risk function R_N is obtained in a few steps. In a first step we study the convergence of the marginal risk $R_N(\underline{j})$. In the formulation of the result we refer to the following condition.

Condition 1: $\forall \underline{j} \in J^1 : \quad N \lim_{N\to\infty} \mathrm{pr}(\widehat{\underline{j}} = \underline{j}) = 0.$

Condition 2: $\forall \underline{j} \in J^2 - \{\underline{j}_0\} : \quad \lim_{N\to\infty} \mathrm{pr}(\widehat{\underline{j}} = \underline{j}) = 0.$

Using the linearity of the operator $M_{\underline{j}}^{(k)}$ and model (1.7) we can write

$$\|\boldsymbol{\alpha}^{(K)} - \widehat{\boldsymbol{\alpha}}_{\underline{j}}^{(k)}\|^2 = \|\boldsymbol{\alpha}^{(K)} - M_{\underline{j}}^{(k)}\boldsymbol{d}\|^2 = \|(I - M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)}\|^2 + \|M_{\underline{j}}^{(k)}\boldsymbol{\varepsilon}\|^2,$$

and therefore

$$\begin{aligned} R_N(\underline{j}) &= E\left(\|(I - M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)}\|^2 \cdot I_{\{\widehat{\underline{j}}=\underline{j}\}}\right) + E\left(\|M_{\underline{j}}^{(k)}\boldsymbol{\varepsilon}\|^2 \cdot I_{\{\widehat{\underline{j}}=\underline{j}\}}\right) \\ &:= I_1 + I_2. \end{aligned} \tag{3.9}$$

A straightforward implication of the above calculations is given by the following Theorem:

Theorem 1.

If the selection criterion satisfies Condition 1, then $N \lim_{N\to\infty} R_N(\underline{j}) = 0$, *for all* $\underline{j} \in J_1$. *If the selection criterion satisfies Condition 2, then* $N \lim_{N\to\infty} R_N(\underline{j}) = 0$, *for all* $\underline{j} \in J_2 - \{\underline{j}_0\}$. *If it satisfies both conditions then* $N \lim_{N\to\infty} R_N(\underline{j}_0) = k_0\sigma^2$.

The previous theorem reveals that the risk function $R_N(\underline{j})$ converges to zero, for all $\underline{j}$, if the selection criterion satisfies Conditions 1 and 2. That these conditions are indeed satisfied by our AMDL-criterion is then demonstrated in Theorem 2.

Theorem 2.

The AMDL-selector satisfies Conditions 1 and 2.

An immediate consequence of Theorems 1 and 2, is the evaluation of the risk R_N. Indeed,

$$\lim_{N\to\infty} R_N = \sum_{k=0}^{K} \sum_{\underline{j}\in J_k} \lim_{N\to\infty} R_N(\underline{j}) = 0,$$

since the number of terms involved in the double sum is finite and independent of N. The result is stated in the next corollary.

Corollary 1.

The risk R_N *of the AMDL selection criterion tends to zero if* N *tends to infinity, i.e.*

$$\lim_{N\to\infty} R_N = 0.$$

Moreover, the relative risk $R_N/R_N(\underline{j}_0) \to 1$ *as* N *tends to infinity, where* $\underline{j}_0$ *is the true model.*

So far we showed that the AMDL selection criterion is consistent and asymptotically efficient. A next step would be to determine the distribution of the AMDL-selector $\widehat{k}$, exactly or approximately. But this is in general a difficult task. Attempts to establish the asymptotic distribution of the AMDL-selector using arguments similar to these in Shibata (1976) and Eubank and Hart (1992) did not work out. Simulation studies in Section 4 will shed some light on the distribution. We also show that it is possible to calculate the exact distribution of the AMDL-selector under the null hypothesis when testing for pure white noise.

4 Applications

The results established in Section 3 can be applied in various testing situations. In this section we illustrate how to do this for some particular applications.

4.1 Testing for pure white noise

A first testing problem is to test whether the observations $\boldsymbol{d}$ in model (1.7) represent pure white noise. In other words, we would like to test the null hypothesis

$$H_0 : \quad \boldsymbol{\alpha}^{(K)} = \mathbf{0}, \tag{4.1}$$

or equivalently

$$H_0 : k_0 = 0. \tag{4.2}$$

We will consider a test statistic based on $\widehat{k}$, the minimizer of

$$1.5kc_\alpha \log_2 N + 0.5N \log_2 \|(I - \boldsymbol{\Theta}^{(k)})\boldsymbol{d}\|^2, \tag{4.3}$$

where $c_\alpha > 0$ will depend on the size of the test. If H_0 is true, i.e. $\boldsymbol{\alpha}^{(K)} = \mathbf{0}$, then the above quantity is likely to be minimized at zero. Therefore, the null hypothesis H_0 is rejected at level α if and only if

$$\widehat{k} \geq 1. \tag{4.4}$$

See also Eubank and Hart (1992) for similar ideas. The constant c_α is such that

$$\mathrm{pr}_{H_0}(\widehat{k} = 0) \approx 1 - \alpha.$$

But in order to determine c_α we need to know the distribution of the selector $\widehat{k}$ under the null hypothesis. In the Appendix we show that the probability $\mathrm{pr}_{H_0}(\widehat{k} = 0)$ is continuous and monotone in c_α, for any fixed k. Hence, for each $0 < \alpha < 1$, there is a unique c_α satisfying the above equation. Although the exact calculation of the distribution of $\widehat{k}$ under H_0 seems possible (see the Appendix), it is analytically untractable. Therefore we approximate the critical values of our test via simulations. The main steps of this simulation study read as follows:

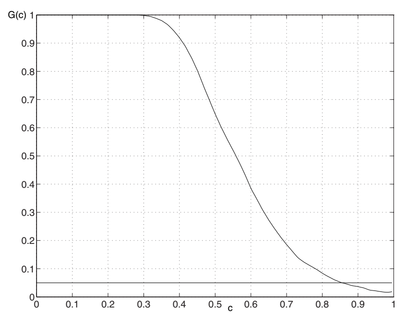
Step 1. For a fixed sample size N, simulate $\varepsilon_1, \cdots, \varepsilon_N$ and, for a given $c_\alpha > 0$ find the AMDL-selector $\widehat{k}$.

Step 2. For the same c_α, repeat Step 1 B times (e.g. $B = 1000$), leading to $\widehat{k}_i, i = 1, \cdots, B$. Estimate $\mathrm{pr}_{H_0}(\widehat{k} = 0)$ by

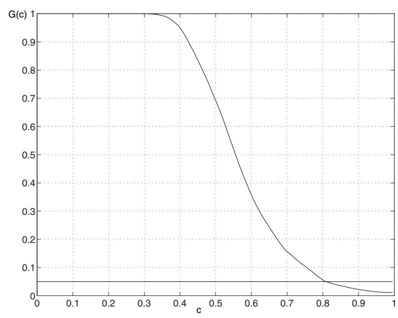
$$\frac{1}{B}\sum_{i=1}^{B} I\{\widehat{k}_i = 0\}.$$

Step 3. Repeat Steps 1 and 2 for a grid of possible values for c_α and select c_α for which $B^{-1}\sum_{i=1}^{B} I\{\widehat{k}_i = 0\} \approx 1 - \alpha$.

The Figure below displays the plot of $G(c) = \mathrm{pr}_{H_0}(\widehat{k} \geq 1)$ as a function of c, for a grid of 60 values over $B = 1000$ simulations, for sample sizes 128 and 256, and smoothed by a local polynomial fit of order 1 with variable bandwidth (see Fan and Gijbels (1995)).



Sample size $N = 128$



Sample size $N = 256$

Table 1 below summarizes the results and reports the empirical levels of our simulation for sample sizes 128 and 256.

Table 1. Empirical levels c_α

	$N = 128$	$N = 256$
$\alpha = 0.1$	0.771	0.751
$\alpha = 0.05$	0.857	0.809

4.2 Testing for no effect in nonparametric regression

When studying the relationships between the response variable and a predictor variable, all of the inference is based on the assumption that the response variable actually depends on the predictor variable. Testing for no effect is the same as checking this assumption.

There are a number of tests available for the no effect hypothesis or, more generally, for testing goodness-of-fit of a linear model. For example Cox and Koh (1989) and Buckley (1991) applied a bayesian approach to construct such test statistics. Eubank and Spiegelman (1990), Müller (1992), Eubank and Hart (1993) and Chen (1994) use various nonparametric smoothing techniques to find an estimator of the regression function and construct a test statistic for no effect.

Testing for no effect is equivalent to testing that the regression function is a constant. Hence, we rewrite model (1.1) as

$$y_i = \beta + g(t_i) + \epsilon_i, \tag{4.5}$$

where β is an unknown constant and g is an unknown function that can be assumed, without loss of generality, to satisfy $\int_0^1 g(t)dt = 0$. The null hypothesis H_0 of no effect is equivalent to H_0: $g = 0$. Since, the wavelet ψ has at least one vanishing moment, the constants are within the kernel of the wavelet transform and therefore, if $\boldsymbol{\alpha}^{(K)}$ denotes the wavelet coefficients of $\beta + g$, $\boldsymbol{\alpha}^{(K)} = \mathbf{0}$, when H_0 is true. Therefore, here again, the null hypothesis H_0 will be rejected at level α if and only if $\widehat{k} \geq 1$.

To ascertain how well a test based on $\widehat{k}$ might work with fixed alternatives we conducted a small scale simulation experiment. Normal errors were used to generate random samples of sizes 128 and 256 from model (4.5). Without loss of generality we took $\beta = 0$. The error variance was equal to 1. For the function g in (4.5) we chose $g(t) = \rho\cos(2\pi\nu t)$. Since $\sigma = 1$, the value used for ρ can be regarded as a signal to noise ratio with smaller values indicating increasing difficulty in discriminating H_0. We considered the specific choices of a signal to noise ratio snr = 0.5, 1 and 2 in our power study. In order to investigate the sensitivity of the test for various alternatives we consider $\nu = 1$, 3 and 9, so that the alternatives will be of higher frequency as ν increases.

Critical values for our test with nominal level 0.05 were all obtained from the approximation used in the previous section. Table 2 contains the empirical powers of our test against the alternative $g(t) = \rho\cos(2\pi\nu t)$. Each of these entries correspond to 500 samples of sizes 128 and 256. Standard errors for non-unit entries in Table 2 are ranged from 0.0185 to 0.008.

Table 2. Proportion of rejections in 500 samples.

		snr=0.5	snr=1	snr=2
$\nu = 1$	$N = 128$	0.148	0.410	0.990
	$N = 256$	0.190	0.684	1.000
$\nu = 3$	$N = 128$	0.085	0.625	1.000
	$N = 256$	0.140	0.648	1.000
$\nu = 9$	$N = 128$	0.125	0.558	1.000
	$N = 256$	0.158	0.762	1.000

Examination of Table 2 shows that the test performs very well for the signal to noise ratio snr=2. Even for a small value of this ratio, say snr=1, the test still performs reasonably well. Looking at the rows for $\nu = 3$ or $\nu = 9$, we find that there is no evidence that the power decreases with ν, showing that our test remains sensitive to higher frequency alternatives. Clearly, the power of the test increases with the sample size N as indicated by the asymptotic analysis in Section 3. Note that the entries of the column with snr=0.5 are only slightly above the nominal level of the test. This is not surprising since the signal to noise ratio is very low, and hence the power of the test will be close to the nominal level 0.05.

To further study the performance of our test, we compared the power of our test with Fan's hard and soft thresholding tests for his examples 1 and 2. As in Fan's paper (1995), for each example we generated a multivariate normal $d \sim N(\theta, 1)$ of dimension 256 and computed (by 400 simulations) the power of our test at a sequence of alternatives indexed by k:

$$H \,:\, \theta_{k+j} = \eta_j, forj = 1, \ldots, \ell_0; \quad \theta_j = 0 \quad \text{otherwise}$$

For example 1, the AMDL test outperforms both tests, with a power equal to 1, whatever the value of k is. For his example 2, the power of our test remains practically constant at a level 0.82, outperforming the soft thresholding test. Fan's hard thresholding test performs a little better with a power around 0.9. These results confirm the robusteness of our procedure against the locations of large coefficients.

The previous results extend directly to the nonparametric comparison of two regression functions. More precisely, suppose that data y_1, y_2, …, y_N are observed according to the model $y_i = f(t_i) + \epsilon_i$ where f is a smooth function and ϵ_1, ϵ_2, …, ϵ_N are i.i.d. $N(0, \sigma^2)$, and consider a second set of data z_1, …, z_N that obey the model $z_i = g(t_i) + \eta_i$ where g is also smooth and the η_i's are i.i.d. $N(0, \tau^2)$. It is obvious then that the test for no effect previously developped can be used for comparing the two curves, that is for testing the null hypothesis H_0: $f = g$ by considering the differences $y_i - z_i$.

4.3 Testing for martingale structure in times series

Our previous results provide also a method for testing whether a time series can be described as a martingale, and more generally for testing a specific pattern of dependence in a time series.

Anderson (1993) proposed the sample standardized spectral distribution for testing goodness-of-fit hypotheses about correlations and asserted that the asymptotic distribution of the difference between the sample and process standardized spectral distribution multiplied by the sample size converges weakly to a Gaussian process under several alternative conditions.

Consider a stationary stochastic process $\{y_t\}$, t = …, -1, 0, 1, …, with $E(y_t) = 0$, autocovariance function $\sigma(h)$, autocorrelation function $\rho_h = \sigma(h)/\sigma(0)$ and spectral density $f(\lambda)$. Inference will be based on a sample $y_1, \cdots, y_T$. The process and sample standardized spectral distribution functions of $\{y_t\}$, as defined by Anderson (1993), are respectively

$$F(\lambda) = \frac{1}{\pi}\left(\lambda + 2\sum_{h=1}^{\infty} \rho_h \frac{\sin \lambda h}{h}\right),$$

and

$$\hat{F}_T(\lambda) = \frac{1}{\pi}\left(\lambda + 2\sum_{h=1}^{T-1} r_h \frac{\sin \lambda h}{h}\right),$$

where

$$r_h = r_{-h} = \frac{\sum_{t=1}^{T-h} y_t y_{t+h}}{\sum_{t=1}^{T} y_t^2}, \quad h = 0, 1, \ldots$$

denotes the sample autocovariance sequence. By assuming that $\{y_t\}$ is a linear process

$$y_t = \sum_{s=0}^{\infty} \gamma_s u_{t-s}$$

where $\sum \sqrt{s}\gamma_s^2 < \infty$, and u_t are square integrable martingale differences, Anderson (1993) shows that,

$$\hat{Y}_T(u) = \frac{\sqrt{T}}{2\sqrt{\pi G(\pi)}} \left[\hat{F}_T\left(G^{-1}[G(\pi)u]\right) - F\left(G^{-1}[G(\pi)u]\right)\right] \tag{4.6}$$

where

$$G(\lambda) = 2\int_0^{\lambda} f^2(\nu)d\nu,$$

as a stochastic process on $[0, 1]$, converges weakly, as $T \to \infty$, to a centered Gaussian process $Z(u)$ with covariance function

$$\{\min(u, v) - uv + q(u)q(v)\}, \quad 0 \leq u, v \leq 1,$$

where

$$q(u) = u - F\left(G^{-1}[G(\pi)u]\right). \tag{4.7}$$

Note that $q(u) = 0$ when there is lack of correlation.

Consider testing the null hypothesis

$$H_0 : \; f(\lambda) = f_0(\lambda)$$

where $f_0(\lambda)$ is completely specified. It is easy to see that for any N, when H_0 is true, the random vector Y_N, whose components are given by $Y_j = \int_0^1 \hat{Y}_T(u) f_j(u)du$, where $f_j(u) = \sqrt{2}\sin(j\pi u)$, converges to a Gaussian random vector, with mean 0 and covariance

$$\Gamma_N = \Lambda_N + \boldsymbol{\alpha}_N \boldsymbol{\alpha}_N^T$$

where Λ_N is the diagonal matrix with $1/(\pi j)^2$ as the jth diagonal element and $\boldsymbol{\alpha}_N$ is the vector with $\alpha_j = \int_0^1 q(u) f_j(u)du$ as the jth component, where $q(u)$ is as in (4.7) based on $f_0(\lambda)$.

On this basis, testing the null hypothesis H_0, is equivalent to test that the random sequence $\Gamma_N^{-1/2} Y_N$ is a pure Gaussian white noise. This now can be tested exactly by the same way as in the first subsection.

5 Discussion

The results presented seem to suggest that if testing is part of the objective of a regression analysis, then model selection carried out using the AMDL criterion has some desirable properties. As pointed out by the referees and the associate editor, there is a qualification: in deriving these results we have assumed that the model generating our data is (i) fixed through the asymptotics; (ii) finite-dimensional; (iii) belongs to the class of models being examined. However, these assumptions are at least in general agreement with a number of analyses in the literature (see e.g. Shibata (1981), Venter and Steel (1992), Speed and Yu (1993). Moreover, as pointed out in section 2, the class of functions tested by our criterion, allows to approximate quite efficiently alternatives composed by complicated functions with inhomogeneous smoothness.

Shibata's (1983) results in the context of linear least squares regression with growth restrictions on the dimension of the candidate models and assuming that the true function is not contained in any of the candidate models, suggest that by using other model selection criteria such as FPE or AIC, consistency results similar to ours may be obtained to situations where one or more of (i), (ii) and (iii) are dropped. Under such settings, it is known that the BIC criterion is not mean asymptotically efficient in the sense of Shibata, and the AMDL criterion (whose behaviour is similar to that of BIC), does not seem to be the appropriate criterion to use. However, when the true model is finite, it is known that both AIC and FPE criteria are not consistent (they have a positive probability to select models that properly include the true model), while BIC and MDL are consistent, which explains the good performance of the AMDL criterion.

As noted by a referee, Saito's work for nonparametric denoising incorporates a best basis search with selection of the wavelet packet coefficients by the MDL criterion, thus allowing competition between various kind of approximating bases. While Saito's approach is appropriate for denoising, its use for testing purposes may be a problem, since the task of searching through a number of bases will lead to inevitable mistakes - bases which apparently look good, but only because of noise fluctuations. It is true that the power of our tests, based on the AMDL criterion for a fixed wavelet basis, is tied to the extend to which the wavelet basis compresses the signal to be tested and this imposes some limitation on the kind of alternatives which can adequately be tested. Note, however, that for many functions having local features such as bumps and discontinuities, whatever regular enough wavelet is used, there are only few large values of wavelet coefficients (see Donoho, Johnstone, Kerkyacharian and Picard (1992) or Fan (1995)).

6 Appendix: Proofs.

Most of the arguments given below are straightforward. We have tried to be explicit wherever possible in order to keep this paper self-contained.

Proof of Theorem 2.

Recall that the AMDL-selector consists of the model indexed by $\widehat{k}$ and with positions $\widehat{\underline{j}}$, and is obtained by minimizing the Approximate Minimum Description Length quantity

$$1.5k\log_2 N + 0.5N\log_2 \|(I-\mathbf{\Theta}^{(k)})\boldsymbol{d}\|^2 = 1.5k\log_2 N + 0.5N\log_2 \|(I-M_{\underline{j}}^{(k)})\boldsymbol{d}\|^2$$

over k and $\underline{j}=\{j_1,\cdots,j_k\}$. Hence, for $\underline{j}\in J^1$, with associated index k, we get

$$\begin{aligned}\mathrm{pr}(\widehat{\underline{j}}=\underline{j}) &\le \mathrm{pr}\Big\{1.5k\log_2 N + 0.5N\log_2 \|(I-M_{\underline{j}}^{(k)})\boldsymbol{d}\|^2 \\ &\qquad \le 1.5k_0\log_2 N + 0.5N\log_2 \|(I-M_{\underline{j_0}}^{(k_0)})\boldsymbol{d}\|^2\Big\} \\ &= \mathrm{pr}\Big\{\|(I-M_{\underline{j}}^{(k)})\boldsymbol{d}\|^2 - \|(I-M_{\underline{j_0}}^{(k_0)})\boldsymbol{d}\|^2 \le \Big(N^{3(k_0-k)/N}-1\Big)\|(I-M_{\underline{j_0}}^{(k_0)})\boldsymbol{d}\|^2\Big\}.\end{aligned} \tag{6.1}$$

Now using model (1.7) we can write

$$\|(I-M_{\underline{j}}^{(k)})\boldsymbol{d}\|^2 = \|(I-M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)}\|^2 + 2\boldsymbol{\varepsilon}^T(I-M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)} + \|(I-M_{\underline{j}}^{(k)})\boldsymbol{\varepsilon}\|^2,$$

where ^{T} stands for the transposition of a vector. For $\|(I - M_{\underline{j_0}}^{(k_0)})\boldsymbol{d}\|^2$ we have a similar expression. With these decompositions, and using that $(I - M_{\underline{j_0}}^{(k_0)})\boldsymbol{\alpha}^{(K)} = 0$, we obtain:

$$\begin{aligned}
\mathrm{pr}(\widehat{j} = \underline{j}) &\leq \mathrm{pr}\Big\{\|(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)}\|^2 + 2\boldsymbol{\varepsilon}^T(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)} + \boldsymbol{\varepsilon}^T(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\boldsymbol{\varepsilon} \\
&\qquad \leq \Big(N^{3(k_0-k)/N} - 1\Big)\|(I - M_{\underline{j_0}}^{(k_0)})\boldsymbol{\varepsilon}\|^2\Big\} \\
&:= \mathrm{pr}\left\{c + X_N + Y_N \leq Z_N\right\} \\
&= \mathrm{pr}\left\{\frac{\sqrt{c}\sqrt{N}}{2\sigma} + \Big(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\Big)^{-1}X_N + \Big(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\Big)^{-1}Y_N \leq \Big(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\Big)^{-1}Z_N\right\},
\end{aligned} \tag{6.2}$$

where we introduced the notations

$$\begin{aligned}
c &= \|(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)}\|^2 \\
X_N &= 2\boldsymbol{\varepsilon}^T(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\boldsymbol{\alpha}^{(K)} \\
Y_N &= \boldsymbol{\varepsilon}^T(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\boldsymbol{\varepsilon} \\
Z_N &= \Big(N^{3(k_0-k)/N} - 1\Big)\|(I - M_{\underline{j_0}}^{(k_0)})\boldsymbol{\varepsilon}\|^2.
\end{aligned}$$

Important is to note that $c > 0$ since $\underline{j} \in J^1$. Hence the division by the term $2\sqrt{c}\frac{\sigma}{\sqrt{N}}$ in (6.2) is allowed. The expectation of the random variable X_N is zero and its variance equals $4\frac{\sigma^2}{N}c$, and hence the central limit theorem gives

$$\Big(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\Big)^{-1}X_N \xrightarrow{\mathrm{D}} \mathcal{N}(0,1), \tag{6.3}$$

and consequently $\Big(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\Big)^{-1}X_N = O_P(1)$.

For the third term in the probability bound in (6.2) we can write

$$\Big(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\Big)^{-1}Y_N = \Big(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\Big)^{-1}\frac{\sigma^2}{N}\widetilde{\boldsymbol{\varepsilon}}^T(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\widetilde{\boldsymbol{\varepsilon}}, \tag{6.4}$$

with $\widetilde{\boldsymbol{\varepsilon}} = (\sigma/\sqrt{N})^{-1}\boldsymbol{\varepsilon}$ a vector of i.i.d. $\mathcal{N}(0,1)$ random variables. Since $\widetilde{\boldsymbol{\varepsilon}}^T(M_{\underline{j_0}}^{(k_0)} - M_{\underline{j}}^{(k)})\widetilde{\boldsymbol{\varepsilon}}$ is a finite sum of squared standard normal distributed random variables we find that the term in (6.4) tends to zero in probability.

The term involving Z_N in (6.2) can be rewritten as

$$\Big(2\sqrt{c_N}\frac{\sigma}{\sqrt{N}}\Big)^{-1}\Big(N^{3(k_0-k)/N} - 1\Big)\frac{\sigma^2}{N}\widetilde{\boldsymbol{\varepsilon}}^T(I - M_{\underline{j_0}}^{(k_0)})\widetilde{\boldsymbol{\varepsilon}}. \tag{6.5}$$

The factor $\left(N^{3(k_0-k)/N}-1\right)$ tends to zero at the rate $(\log N)/N$ if N tends to infinity. Hence, since $N^{-1}\tilde{\boldsymbol{\varepsilon}}^T(I-M_{\underline{j}_0}^{(k_0)})\tilde{\boldsymbol{\varepsilon}}$ is bounded in probability, the term in (6.5) tends to zero in probability.

The inequality (6.2) is dominated by

$$\mathrm{pr}\left\{\left(2\sqrt{c}\frac{\sigma}{\sqrt{N}}\right)^{-1}X_N \leq -\frac{\sqrt{c}\sqrt{N}}{2\sigma}+2N^{1/4}\right\} \quad +\mathrm{pr}\left\{-2\left(\sqrt{c}\frac{\sigma}{\sqrt{N}}\right)^{-1}Y_N > N^{1/4}\right\}$$
$$+\mathrm{pr}\left\{2\left(\sqrt{c}\frac{\sigma}{\sqrt{N}}\right)^{-1}Z_N > N^{1/4}\right\} \quad (6.6)$$

By (6.3), and the tails of the normal distribution one obtains:

$$\mathrm{pr}\left\{2\left(\sqrt{c}\frac{\sigma}{\sqrt{N}}\right)^{-1}X_N \leq -\frac{\sqrt{c}\sqrt{N}}{2\sigma}+2N^{1/4}\right\} = \mathcal{O}(e^{-c^2N/2}).$$

By the relationship between a distribution function and its moment generating function, relations (6.4) and (6.5) give:

$$\mathrm{pr}\left\{-2\left(\sqrt{c}\frac{\sigma}{\sqrt{N}}\right)^{-1}Y_N > N^{1/4}\right\} \leq \mathcal{O}(e^{-N^{1/4}}),$$

and

$$\mathrm{pr}\left\{2\left(\sqrt{c}\frac{\sigma}{\sqrt{N}}\right)^{-1}Z_N > N^{1/4}\right\} \leq \mathcal{O}(e^{-N^{1/4}}).$$

Finally, since $c > 0$, combination of the above inequalities shows that condition 1 is satisfied.

Suppose now that $\underline{j} \in J^2 - \{\underline{j}_0\}$. We now have:

$$\mathrm{pr}(\widehat{\jmath} = \underline{j}) \leq \mathrm{pr}\left\{\boldsymbol{\varepsilon}^T(M_{\underline{j}_0}^{(k_0)} - M_{\underline{j}}^{(k)})\boldsymbol{\varepsilon} \leq \left(N^{3(k_0-k)/N}-1\right)\|(I-M_{\underline{j}_0}^{(k_0)})\boldsymbol{\varepsilon}\|^2\right\}$$

Condition 2 follows by an inequality on a chi-squared distribution (see Shibata (1981)).

□

Distribution of $\hat{k}$.

Denote by $d_{(1)} \leq \cdots \leq d_{(N)}$ the ordered observed data. Then, when $\alpha^{(K)} = 0$, we have, for any integer $0 \leq k \leq K$,

$$\begin{aligned}\mathrm{pr}(\widehat{k} = k) &= \mathrm{pr}\left\{N^{3kc_\alpha/N}\sum_{s=1}^{N-k} d_{(s)}^2 \leq N^{3\ell c_\alpha/N}\sum_{s=1}^{N-\ell} d_{(s)}^2, \forall k < \ell \leq K\right\}\\ &= \mathrm{pr}_{H_0}\left\{N^{3kc_\alpha/N}S_{N-k} \leq N^{3\ell c_\alpha/N}S_{N-\ell}, \forall k < \ell \leq K\right\},\end{aligned}$$

where $S_\ell = \sum_{s=1}^{\ell} X_{(s)}^2$ is a partial sum of the order statistics of $\chi^2(1)$ distributed random variables $X_1, \cdots, X_N$. The above probability equals

$$\mathrm{pr}\left\{\bigcap_{\ell=k+1}^{K}\left\{\frac{S_{N-k}}{S_{N-\ell}} \le N^{3(\ell-k)c_\alpha/N}\right\}\right\} = \mathrm{pr}\left\{\bigcap_{\ell=k+1}^{K}\left\{\prod_{s=k}^{\ell-1}\frac{S_{N-s}}{S_{N-s-1}} \le N^{3(\ell-k)c_\alpha/N}\right\}\right\}$$

which is continuous and monotone in c_α, for any fixed k.

Acknowledgements. We would like to thank the referees and tha associate editor for constructive criticism which led to substantial improvement of the original paper.

REFERENCES

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, **22**, 203 – 217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 267 – 281. Eds. B.N. Petrov and F. Czáki. Akademiai Kiadó, Budapest.

Allen, D. (1971). The prediction sum of squares as a criterion for selecting prediction variables. *Technical Report* No. 23, Dept. of Statistics, Univ. of Kentucky.

Anderson, T. W. (1993). Goodness of fit tests for spectral distributions. *Ann. Statist.*, **21**, 830 – 847.

Antoniadis, A. (1994). Smoothing noisy data with coiflets. *Statistica Sinica*, **4**, 651 – 678.

Antoniadis A., Grégoire G. and McKeague, I. (1994). Wavelet Methods for Curve Estimation. *J. Amer. Stat. Assoc.*, **89**, 1340 – 1353.

Buckley, M. J. (1991). Detecting a smooth signal: optimality of cusum based procedures, *Biometrika*, **78**, 253 – 262.

Chen, J. C. (1994). Testing for no effect in nonparametric regression via spline smoothing techniques, *Ann. Inst. Statist. Math*, **46**, 251 – 265.

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms, *Applied and Comp. Harmonic Analysis*, **1**, 54 – 81.

Cox, D. D. and Koh, E. (1989). A smoothing spline based test of model adequacy in polynomial regression, *Ann. Inst. Statist. Math*, **41**, 383 – 400.

Daubechies, I. (1992). *Ten lectures on wavelets*. SIAM, Philadelphia.

Devore, R., Jawerth, B. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Amer. Math. Soc.*, **305**, 397 – 414.

Donoho, D. L., Johnstone, I. M., Kerkyacharian G. and Picard, D. (1995). Wavelet Shrinkage: Asymptopia. *J. Royal Statist. Soc.*, Ser B, **57**, 301 – 369.

Eubank, R.L. (1988). *Spline smoothing and nonparametric regression.* Marcel Dekker, New York.

Eubank, R.L. and Hart, J.D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.*, **20**, 1412 – 1425.

Eubank, R.L. and Hart, J.D. (1993). Commonality of cusum, Von Neumann and smoothing based goodness-of-fit tests, *Biometrika*, **80**, 89 – 98.

Eubank, R.L. and Spiegelman, C. H. (1990). Testing the goodness-of-fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Assoc.*, **85**, 387 – 392.

Fan J. (1995). Test of significance based on wavelet thresholding and Neyman's truncation. *Manuscript*.

Fan, J. and Gijbels, I. (1995). Data-driven Bandwidth selection in local polynomial fitting : variable bandwidth and spatial adaptation. *J. R. Statist. Soc. B*, **57**, 371 – 394.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric regression and generalized linear models.* Chapman and Hall, London.

Härdle, W. (1990). *Applied nonparametric regression.* Cambridge University Press, New York.

Hashimoto, A., Honda, M. Inque, T. and Taguri, M. (1981). Selection of regression models by several information criterions. *Rep. Statist. Appl. Res.*, *JUSE*, **28**, 57 – 72.

Herley, C. (1995). Boundary Filters for Finite-Length Signals and Time-Varying Filter Banks. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, to appear.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661 – 675.

McGill, K. and Taswell, C. (1993). Wavelet Transforms algorithms for Finite-Duration Signals. In *Proceedings of the International Conference on Wavelets and Applications, Toulouse, France*, Eds Y. Meyer and Roques S., Editions Frontières, 221 – 224.

Meyer, Y. (1993). *Wavelets: Algorithms & Applications*. Translated and revised by R. D. Ryan. SIAM, Philadelphia.

Müller, H.-G. (1988). *Nonparametric regression analysis for longitudinal data*. Lecture Notes in Statistics, **46**. Springer-Verlag, Berlin.

Müller, H.-G. (1992). Goodness-of-fit diagnostics for regression models. *Scand. J. Statist.*, **19**, 157 – 172.

Neumann, M.H. and Spokoiny, V.G. (1993). On the efficiency of wavelet estimators under arbitrary error distributions. *Discussion Paper No. 4*, Hümboldt Universität zu Berlin.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758 – 765.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, **11**, 416 – 431.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, **30**, 629 – 636.

Saito, N. (1994). Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion. In *Wavelets in Geophysics*, Eds. E. Foufoula-Georgiou and P. Kumar. Academic Press, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461 – 464.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117 – 126.

Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrika*, **68**, 45 – 54.

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statist. Math*, **35**, 415 – 423.

Speed, T. P. and Bin Yu. (1993). Model Selection and Prediction: Normal Regression. *Ann. Inst. Statist. Math*, **45**, 35 – 54.

Sugiura, N. (1978). Numerical experiment on information criteria (in Japanese). Reported in the 11th Symposium on Statistical Data Analysis at Osaka University, organized by M. Okamoto, Osaka Univ.

Venter, J. H. and S. J. Steel. (1992). Some Contributions to Selection and Estimation in the Normal Linear Model. *Ann. Inst. Statist. Math*, **44**, 281 – 297.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia.