The multiresolution criterion and nonparametric regression

Thoralf Mildenberger and Henrike Weinert
joint work with P.L. Davies and U. Gather

SFB 475
Fakultät Statistik
Technische Universität Dortmund

Workshop on current trends and challenges
in model selection and related areas
Vienna, July 2008

Technische Universität Dortmund
SFB 475: Reduction of Complexity in Multivariate Data Structures

technische universität dortmund

Outline

Nonparametric Regression

Model: $y(t_i) = f(t_i) + \varepsilon(t_i), \qquad (0 \leq t_1 < \cdots < t_N \leq 1)$

$$\varepsilon(t_1), \ldots, \varepsilon(t_N) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Goal: Find estimate $\hat{f}$ of f.

Nonparametric Regression

Model: $y(t_i) = f(t_i) + \varepsilon(t_i), \qquad (0 \leq t_1 < \cdots < t_N \leq 1)$

$$\varepsilon(t_1), \ldots, \varepsilon(t_N) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Goal: Find estimate $\hat{f}$ of f.

Problem: $\hat{f}$ usually chosen from family $(\hat{f}_h)$ indexed by smoothing parameter h (bandwidth, size of a partition, penalty etc.)

Interpretation: Often h - 'complexity' of $\hat{f}_h$.

Choosing the smoothing parameter

Risk based choice: h such that $\hat{f}_h$ minimizes risk (e.g. MSE, MISE etc.)
Risk has to be estimated from data by e.g.: Asymptotic considerations, Plug-In-Methods, Penalized Criteria, CV, Risk bounds etc.

Residual based choice: Given data, find simplest model that 'could have generated' the data, i.e. residuals 'look like noise' e.g. Taut-String Algorithm (Davies and Kovac 2001).

The Multiresolution Criterion

Given some estimate $\hat{f}$, consider residuals

$$r_i := r(t_i) := y(t_i) - \hat{f}(t_i)$$

Accept residuals as noise iff

$$\max_{I \in \mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{i \in I} r_i \right| \leq \sigma C \qquad (*)$$

$\mathcal{I}$ System of all intervals in $\{1, \ldots, N\}$

The Multiresolution Criterion

Given some estimate $\hat{f}$, consider residuals

$$r_i := r(t_i) := y(t_i) - \hat{f}(t_i)$$

Accept residuals as noise iff

$$\max_{I \in \mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{i \in I} r_i \right| \leq \sigma C \qquad (*)$$

$\mathcal{I}$ System of all intervals in $\{1, \ldots, N\}$

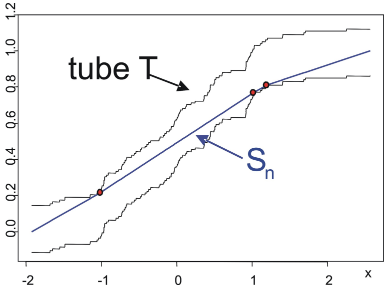
Choose estimate of smallest complexity such that $(*)$ is fulfilled.

Residual based methods

MR criterion has been combined with different measures of **complexity**:

- Number of local extrema or total variation
  (Taut-String-Algorithm, Davies and Kovac 2001)
- Number of changes between convexity and concavity
  (Davies, Kovac and Meise 2008)
- Smoothness quantified by derivatives
  (Weighted Smoothing Splines, Davies and Meise 2008)
- Number of jumps
  (Potts smoother, Boysen et al. 2008)

Taut String Method



summed process $y_n^\circ = \frac{1}{n}\sum_{t_i \le t} y(t_i)$

Tube $T\left(y_n^\circ, \frac{C}{\sqrt{n}}\right)$:

$y_n^\circ - \frac{C}{\sqrt{n}} \le g(t) \le y_n^\circ + \frac{C}{\sqrt{n}}$

String S_n: has smallest length$(S_n) = \int_0^1 \sqrt{1 + s_n^2(t)}dt$

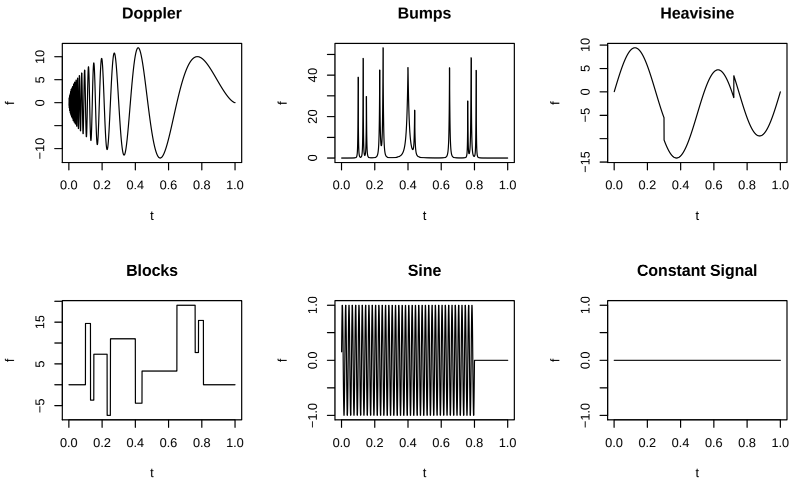
Derivative of S_n: candidate for $\hat{f}$

Check if MR criterion fulfilled, if not: local squeezing of tube

Simulation Study (Davies, Gather, Weinert, 2008)

- Wavelet-Thresholding (Donoho and Johnstone, 1994)
  $\rightarrow$ hard and soft thresholding [H,S]
- Unbalanced Haar (Fryzlewicz, 2006) [U]
- Minimum-Description-Length (Rissanen, 2000) [M]
- Adaptive weights smoothing
  (Polzehl and Spokoiny, 2003) [A]
- Local Plug-in kernel method (Herrmann, 1997) [P]
- Taut-string (Davies and Kovac, 2001) [T,V]

Simulation Study



Simulation Study

6 Test-bed functions, 4 σ-values, 5 sample sizes n
1000 simulations at each test-bed function, $\sigma-$ and $n-$level

Simulation Study

6 Test-bed functions, 4 σ-values, 5 sample sizes n
1000 simulations at each test-bed function, $\sigma-$ and $n-$level

Mean for 3 performance criteria:

L_∞-norm: $\ell(f, \hat{f}) = \max_{1 \leq i \leq n} \left| f\left(\frac{i}{n}\right) - \hat{f}\left(\frac{i}{n}\right) \right|$

L_2-norm: $\ell(f, \hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(f\left(\frac{i}{n}\right) - \hat{f}\left(\frac{i}{n}\right) \right)^2$

Simulation Study

6 Test-bed functions, 4 σ-values, 5 sample sizes n
1000 simulations at each test-bed function, $\sigma-$ and $n-$level

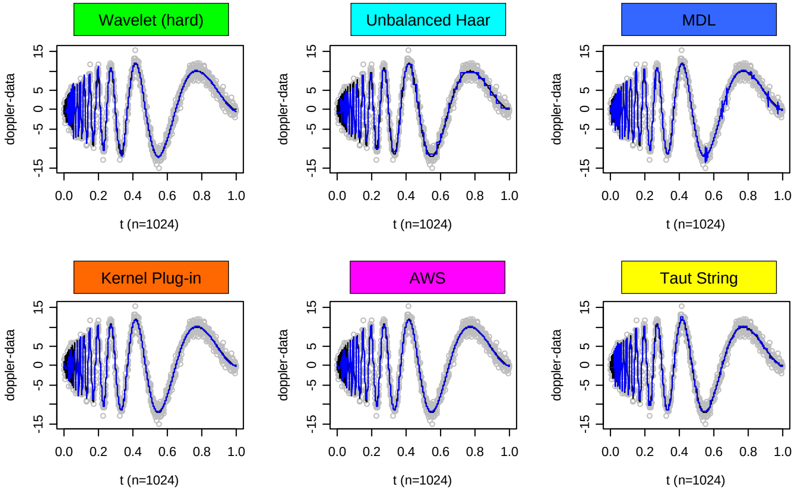
Mean for 3 performance criteria:

L_∞-norm: $\ell(f, \hat{f}) = \max_{1\leq i\leq n} \left| f\left(\frac{i}{n}\right) - \hat{f}\left(\frac{i}{n}\right) \right|$

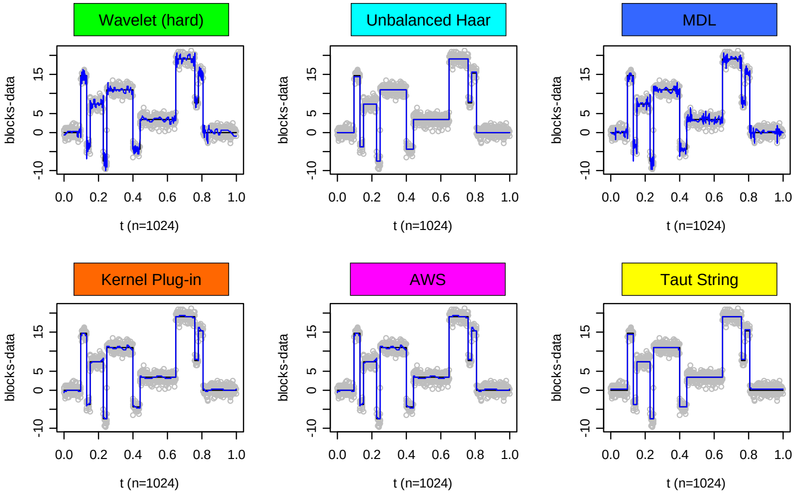
L_2-norm: $\ell(f, \hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(f\left(\frac{i}{n}\right) - \hat{f}\left(\frac{i}{n}\right) \right)^2$

Peak-identification-loss:
$\ell(f, \hat{f}) =$ number of unidentified extremes of f
+ number of superfluous extremes of $\hat{f}$
$\rightarrow$ overall error in identifying extremes of true f
with extremes of $\hat{f}$

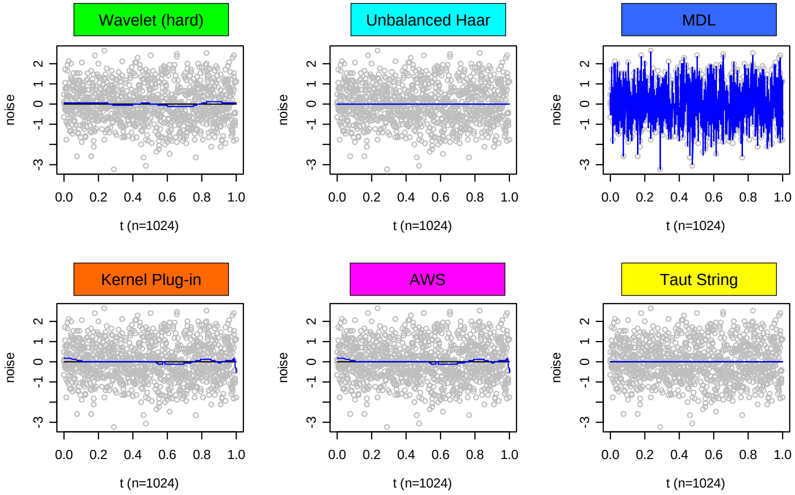
Approximations of Doppler-data



Approximations of Blocks-data



Approximations of a Constant

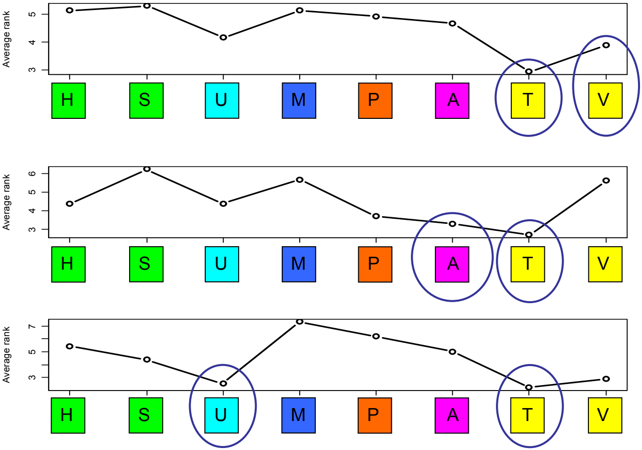


Average Ranks

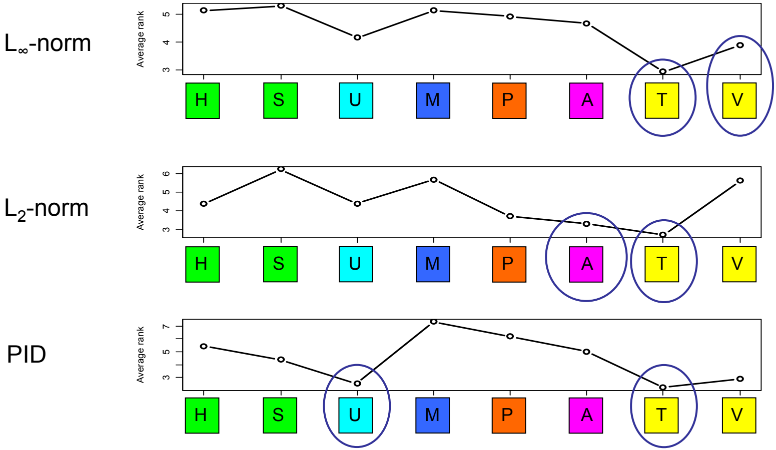
L_∞-norm

L_2-norm

PID



Average Ranks



MR-based TS algorithm performs well

MR criterion and Nadaraya-Watson kernel regression

$$r_{t,h} := \begin{cases} \frac{\sum_{i=1}^{n} K_h(t_i - t) r_i}{\sqrt{\sum_{i=1}^{n} K_h^2(t_i - t)}}, & \text{if} \quad \sqrt{\sum_{i=1}^{n} K_h^2(t_i - t)} \neq 0 \\ 0, & \text{if} \quad \sqrt{\sum_{i=1}^{n} K_h^2(t_i - t)} = 0 \end{cases}$$

for all $t \in [0, 1]$, $h > 0$, with $K_h(\cdot) := h^{-1} K(h^{-1} \cdot)$ for the uniform kernel

$$K := \mathbb{I}_{[-0.5, 0.5]}$$

MR criterion and Nadaraya-Watson kernel regression

$$r_{t,h} := \begin{cases} \frac{\sum_{i=1}^{n} K_h(t_i - t) r_i}{\sqrt{\sum_{i=1}^{n} K_h^2(t_i - t)}}, & \text{if } \sqrt{\sum_{i=1}^{n} K_h^2(t_i - t)} \neq 0 \\ 0, & \text{if } \sqrt{\sum_{i=1}^{n} K_h^2(t_i - t)} = 0 \end{cases}$$

for all $t \in [0, 1]$, $h > 0$, with $K_h(\cdot) := h^{-1} K(h^{-1} \cdot)$ for the uniform kernel

$$K := \mathbb{I}_{[-0.5, 0.5]}$$

Then:

- $r_1, \ldots, r_N \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \Longrightarrow r_{t,h} \sim \mathcal{N}(0, \sigma^2)$.
- MR criterion:

$$\sup_{t,h} |r_{t,h}| = \max_{I \in \mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{i \in I} r_i \right|$$

The Multiresolution Norm (Mildenberger 2008)

Consider:	data	$(y_1, \ldots, y_N)$
	estimate	$(\hat{f}_1, \ldots, \hat{f}_N)$
	residuals	$(r_1, \ldots, r_N)$

as vectors in $\mathbb{R}^N$ with the **multiresolution norm**

$$\|(x_1, \ldots, x_N)\|_{\mathrm{MR}} := \max_{I \in \mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{t \in I} x_t \right|$$

The Multiresolution Norm (Mildenberger 2008)

Consider:

	data	$(y_1, \dots, y_N)$
	estimate	$(\hat{f}_1, \dots, \hat{f}_N)$
	residuals	$(r_1, \dots, r_N)$

as vectors in $\mathbb{R}^N$ with the **multiresolution norm**

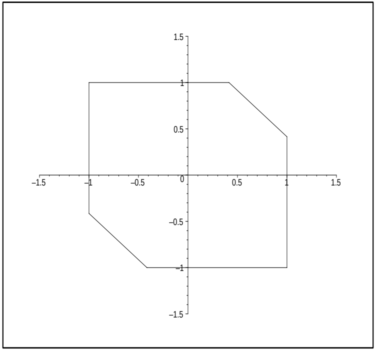
$$\|(x_1, \dots, x_N)\|_{\mathrm{MR}} := \max_{I \in \mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{t \in I} x_t \right|$$

Then: Multiresolution criterion is fulfilled

$$\Longleftrightarrow \|y - \hat{f}\|_{\mathrm{MR}} \leq \sigma C$$

i.e. $\hat{f}$ is contained in the MR-Ball of radius σC centered at y or (equivalently) residuals $r = y - \hat{f}$ lie in ball around zero

Multiresolution Norm Unit Ball in $\mathbb{R}^2$



ℓ_p-Norms

$$
\begin{aligned}
\|(x_1, \ldots, x_N)\|_p &= \left(\sum_{t=1}^{N} |x_t|^p\right)^{1/p} \qquad (1 \le p < \infty) \\
\|(x_1, \ldots, x_N)\|_\infty &= \max\{|x_1|, \ldots, |x_N|\}
\end{aligned}
$$

ℓ_p-Norms

$$\begin{aligned} \|(x_1, \ldots, x_N)\|_p &= \left(\sum_{t=1}^{N} |x_t|^p\right)^{1/p} \quad (1 \le p < \infty) \\ \|(x_1, \ldots, x_N)\|_\infty &= \max\{|x_1|, \ldots, |x_N|\} \end{aligned}$$

invariant w.r.t.:

ℓ_p-Norms

$$\|(x_1,\ldots,x_N)\|_p = \left(\sum_{t=1}^N |x_t|^p\right)^{1/p} \qquad (1 \leq p < \infty)$$
$$\|(x_1,\ldots,x_N)\|_\infty = \max\{|x_1|,\ldots,|x_N|\}$$

invariant w.r.t.:

1. Sign changes in one or several components

ℓ_p-Norms

$$\|(x_1,\ldots,x_N)\|_p = \left(\sum_{t=1}^N |x_t|^p\right)^{1/p} \quad (1 \leq p < \infty)$$

$$\|(x_1,\ldots,x_N)\|_\infty = \max\{|x_1|,\ldots,|x_N|\}$$

invariant w.r.t.:

1. Sign changes in one or several components
2. Permutation of components

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\mathrm{MR}} = \max\Big\{1, 1, 1,$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1, -1, 1)\|_{\text{MR}} = \max \Big\{ 1, 1, 1, 0/\sqrt{2}, 0/\sqrt{2},$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\mathrm{MR}} = \max\left\{1,1,1,0/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\}$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\mathrm{MR}} = \max\left\{1,1,1,0/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = 1$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\text{MR}} = \max\left\{1,1,1,0/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = 1$$

but

$$\|(1,1,-1)\|_{\text{MR}} = \max\left\{1,1,1,\right.$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\text{MR}} = \max\left\{1,1,1,0/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = 1$$

but

$$\|(1,1,-1)\|_{\text{MR}} = \max\left\{1,1,1,2/\sqrt{2},0/\sqrt{2},\right.$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\text{MR}} = \max\left\{1,1,1,0/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = 1$$

but

$$\|(1,1,-1)\|_{\text{MR}} = \max\left\{1,1,1,2/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\}$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\text{MR}} = \max\left\{1,1,1,0/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = 1$$

but

$$\|(1,1,-1)\|_{\text{MR}} = \max\left\{1,1,1,2/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = \sqrt{2}$$

Lack of Invariance

MR-norm not invariant w.r.t. these transformations:
Consider

$$\|(1,-1,1)\|_{\text{MR}} = \max\left\{1,1,1,0/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = 1$$

but

$$\|(1,1,-1)\|_{\text{MR}} = \max\left\{1,1,1,2/\sqrt{2},0/\sqrt{2},1/\sqrt{3}\right\} = \sqrt{2}$$

With $|x| := (|x_1|,\ldots,|x_N|)$, we have:

$$\|x\|_{\text{MR}} \leq \||x|\|_{\text{MR}}$$

Lack of Invariance

Furthermore:

- **Identity** and **reverse ordering** are the only permutations that do not affect the MR-norm of any $x \in \mathbb{R}^N$.
- **Identity** and **changing all signs simultaneously** are the only sign changes that do not affect the MR-norm of any $x \in \mathbb{R}^N$.

Sign Patterns

For $x \in \mathbb{R}^N$ mit $|x_1| = \cdots = |x_N| =: m > 0$:

- $\|x\|_{\text{MR}}$ attains its maximum $\iff$ all components have the same sign
- $\|x\|_{\text{MR}}$ attains its minimum $\iff$ the signs are alternating
- $\|x\|_{\text{MR}} \ \geq \ m \ \times \sqrt{\text{length of longest run}}$

Sign Patterns

For $x \in \mathbb{R}^N$ mit $|x_1| = \cdots = |x_N| =: m > 0$:

- $\|x\|_{\text{MR}}$ attains its maximum $\iff$ all components have the same sign
- $\|x\|_{\text{MR}}$ attains its minimum $\iff$ the signs are alternating
- $\|x\|_{\text{MR}} \;\geq\; m \;\times \sqrt{\text{length of longest run}}$

$\rightarrow$ Dependence of the MR-norm on sign patterns allows for residual diagnostics!

Summary

- Residual-based smoothing parameter selection performs quite well
- Multiresolution criterion corresponds to
  a ball in the *multiresolution norm*
- Detection of structure in residuals is possible because of lack of invariance properties

References 1

BOYSEN, L., KEMPE, A., MUNK, A., LIEBSCHER, V. and WITTICH, O. (2008). Consistencies and rates of convergence of jump penalized least squares estimators. The Annals of Statistics, to appear.

CHAUDHURI, P. and MARRON, J.S. (2000). Scale space view of curve estimation. Annals of Statistics 28, 408-428.

DAVIES, P.L. and KOVAC, A. (2001). Local Extremes, Runs, Strings and Multiresolution. The Annals of Statistics 29, 1-65.

DAVIES, P. L., KOVAC, A., and MEISE, M. (2008). Nonparametric regression, confidence regions and regularization. The Annals of Statistics, to appear.

DAVIES, P.L. and MEISE, M. (2008). Approximating Data with Weighted Smoothing Splines. Journal of Nonparametric Statistics 20, 207-228.

DAVIES, P. L., GATHER, U., and WEINERT, H. (2008). Nonparametric Regression as an Example of Model Choice. Communications in Statistics - Simulation and Computation 37, 274-289.

DONOHO, D.L., and JOHNSTONE, I. M. (1994).Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425-455.

References 2

DÜMBGEN, L. and SPOKOINY, V.G. (2001). Multiscale testing of qualitative hypotheses. Annals of Statistics 29, 124-152.

FRYZLEWICZ, P. (2007). Unbalanced Haar Technique for Nonparametric Function Estimation. Journal of the American Statistical Association 102, 1318-1327.

HERRMANN, E. (1997). Local Bandwidth Choice in Kernel Regression Estimation. Journal comp. graph. stat. 6, 35-54.

MILDENBERGER, T. (2008). A geometric interpretation of the multiresolution criterion in nonparametric regression. Journal of Nonparametric Statistics, to appear.

POLZEHL, J. and SPOKOINY, V. (2003). Varying coefficient regression modeling. Preprint.

RISSANEN, J. (2000). MDL-Denoising. IEEE Trans. Information Theory 42, 40-47.