Complexity, Information, and Noise:
Denoising signals by the MDL principle

Teemu Roos

Complex Systems Computation Group
Department of Computer Science
University of Helsinki

Denoising

- Removing noise from signals
- Not obvious how to formalize.
- Given for $1 \le t \le n$

unobserved

observed

$$z(t) = x(t) + y(t)$$

observation = signal + noise

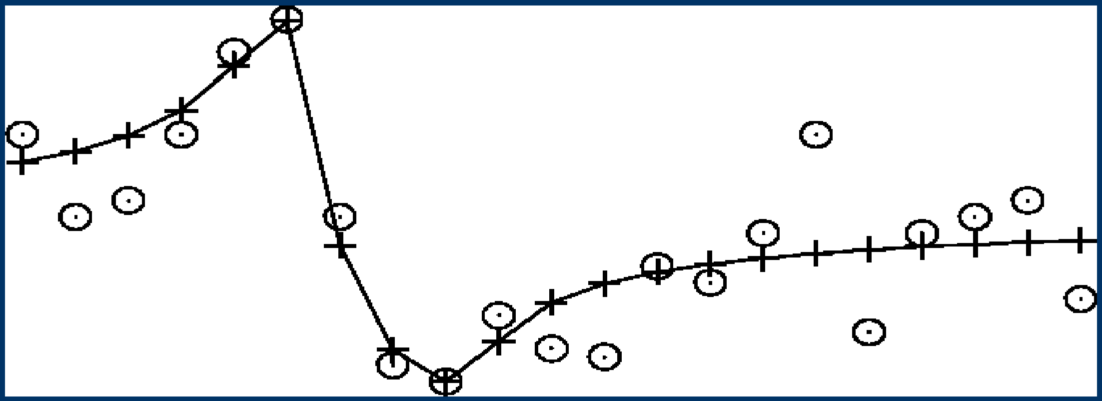
is it possible to recover $x(t)$?

Approaches: I. Statistical Estimation

- Signal = Parameter
- Denoising = Estimation

$$z(t) = \theta(t) + \varepsilon(t)$$

- Assume: $\varepsilon(t)$ i.i.d. Gaussian
- Estimate mean of multivariate Gaussian density
- Regression



Approaches: I. Statistical Estimation *(contd.)*

- Signal = Parameter
- Denoising = Estimation

$$z(t) = \theta(t) + \varepsilon(t)$$

- Only one observation of $\theta(t)$!
- Stein's paradox: $z(t)$ is not an admissible estimator of $\theta(t)$ (Stein, 1956)
- <u>Minimax approach:</u>
  - θ smooth (e.g. bounded Sobolev norm)
  - minimize estimator's worst-case risk

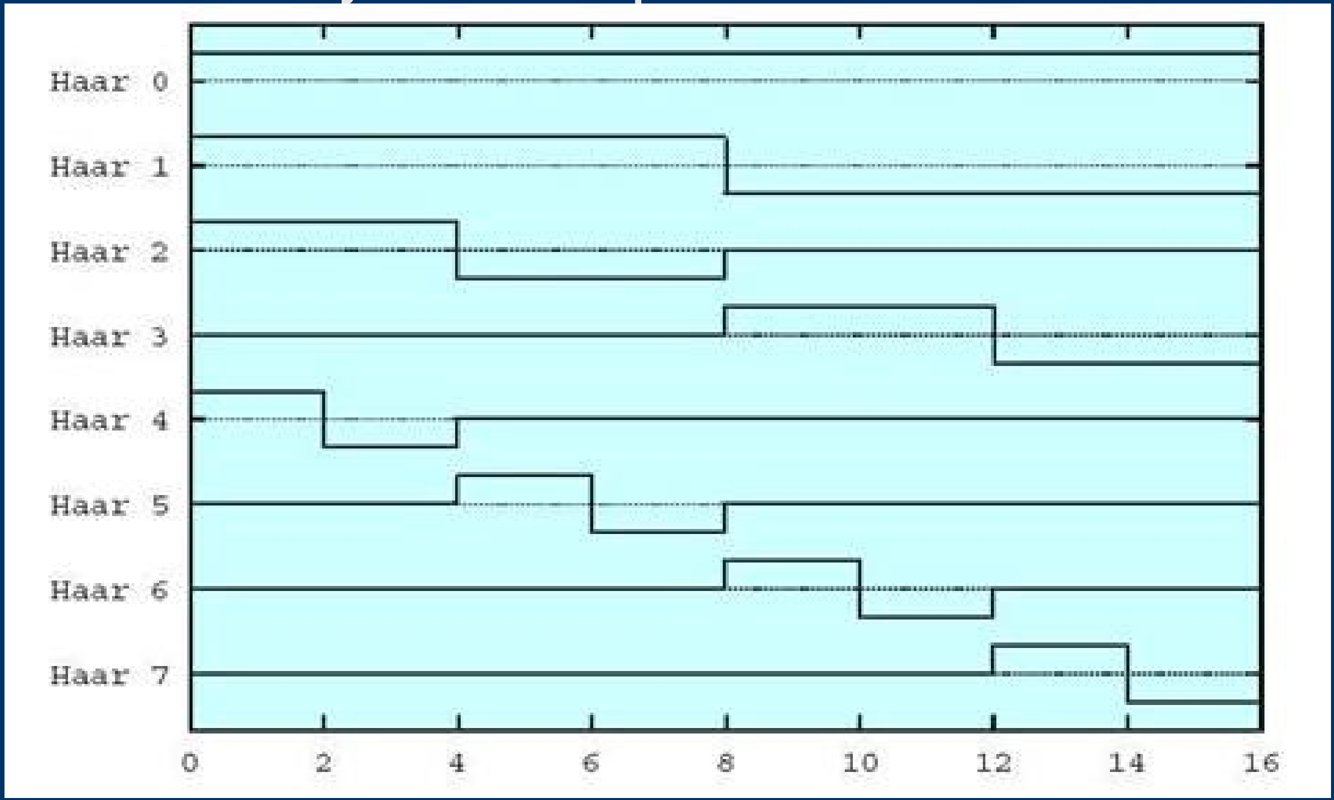
Approaches: II. Bayes

- Prior distribution for $\theta(t)$
  - restricted: find estimator of $\theta(t)$ with small *expected* risk
  - full Bayesian: find posterior of $\theta(t)$
- Good performance
- Full Bayes computationally demanding

Wavelets

- Wavelet transformations
  - both time and frequency resolution
- Similar to Fourier which only has frequencies
- *Example:*

  Haar basis



Wavelet Denoising

- Wavelet transforms concentrate the 'energy' of most natural signals
  - Many coefficients near zero, few very large ones
  - Does *not* happen with noise
- Idea: Identify coefficients that are negligible in the original signal and set them to zero
- Should retain most of the signal
- Removes a lot of noise

MDL

- Minimum Description Length (MDL) principle: minimize

$$L(\text{model}) + L(x \mid \text{model})$$

MDL

- Minimum Description Length (MDL) principle:

  minimize

$$L(\text{model}) + L(x \mid \text{model})$$

code-length of model

MDL

- Minimum Description Length (MDL) principle: minimize

$$L(\text{model}) + L(x \mid \text{model})$$

code-length of model

code-length of data *given* model

- 'Model' gives the regular features in the data
- 'Regular' = 'compressible'

Foundation of MDL

- Kolmogorov complexity K(x)
  - length of the shortest program to output x
- Kolmogorov sufficient statistic
  - finite set S such that $\underbrace{K(S)}_{\text{complexity of S}} + \underbrace{\log |S|}_{\text{log of size}} \le \underbrace{K(x)}_{\text{complexity of x}} + O(1)$

Foundation of MDL

- Kolmogorov complexity K(x)
  - length of the shortest program to output x
- Kolmogorov sufficient statistic
  - finite set S such that $K(S) + \log |S| \leq K(x) + O(1)$

complexity of S

log of size

complexity of x

$x \in S$

Foundation of MDL

- Kolmogorov complexity K(x)
  - length of the shortest program to output x
- Kolmogorov sufficient statistic
  - finite set S such that $K(S) + \log |S| \le K(x) + O(1)$

$$\underbrace{\underbrace{K(S)}_{\text{complexity of S}} + \underbrace{\log |S|}_{\text{log of size}}}_{\text{minimum code-length of x using S as a model}} \le \underbrace{K(x)}_{\text{complexity of x}} + O(1)$$

- log |S| is the best achievable code-length of x *given* S (for almost all $x \in S$)

Foundation of MDL

- Kolmogorov complexity $K(x)$
  - length of the shortest program to output x
- Kolmogorov sufficient statistic
  - finite set S such that $K(S) + \log |S| \leq K(x) + O(1)$

let's drop these

complexity of S

log of size

minimum code-length of x using S as a model

complexity of x

- log |S| is the best achievable code-length of x *given* S (for almost all $x \in S$)

Foundation of MDL

- Kolmogorov complexity K(x)
  - length of the shortest program to output x
- Kolmogorov sufficient statistic
  - finite set S such that $K(S) + \log |S| \le K(x)$

complexity of S

log of size

complexity of x

minimum code-length
of x using S as a model

- log |S| is the best achievable code-length of x *given* S (for almost all $x \in S$)

Foundation of MDL

- Kolmogorov minimal sufficient statistic (KMSS)
  - the least complex Kolmogorov sufficient statistic
- Includes all regular features of the object *but not more*
- *Example:* given a random string r = 0 1 1 0 0 0 1 0 1 1...

  duplicate all bits: x = 00 11 11 00 00 00 11 00 11 11...

Foundation of MDL

- Kolmogorov minimal sufficient statistic (KMSS)
  - the least complex Kolmogorov sufficient statistic
- Includes all regular features of the object *but not more*
- *Example:* given a random string r = 0 1 1 0 0 0 1 0 1 1...

  duplicate all bits: x = 00 11 11 00 00 00 11 00 11 11...
  - $S_1 = \{x\}$ is a sufficient statistic because

$$K(S_1) + \log |S_1| = K(x) + 0 \leq K(x)$$

Foundation of MDL

- Kolmogorov minimal sufficient statistic (KMSS)
  - the least complex Kolmogorov sufficient statistic
- Includes all regular features of the object *but not more*
- *Example:* given a random string r = 0 1 1 0 0 0 1 0 1 1...

  duplicate all bits: x = 00 11 11 00 00 00 11 00 11 11...
  - $S_1 = \{x\}$ is a sufficient statistic because

    $$K(S_1) + \log |S_1| = K(x) + 0 \leq K(x)$$
  - $S_2 = \{\text{duplicated strings}\}$ is the KMSS because

    $$K(S_2) + \log |S_2| = \log(2^{n/2}) = n/2 = K(x)$$

Ideal MDL vs Practical MDL

- Ideal MDL: Choose the KMSS model[1]
- Problems with Kolmogorov complexity
  - not computable
  - depends on the universal language (Turing machine)
- Practical MDL
  - Kolmogorov complexity replaced by computable codes
  - probabilistic models: $L(x) = -\log p(x)$
- Same idea: Identify regular features in data

[1] under certain conditions

MDL & Denoising

- Identify regular features in data
- Naturally applicable to denoising

$$K(x) = K(S) + \log |S|$$

$$\text{code-length} = L(\text{model}) + L(x \mid \text{model})$$

$$\text{complexity} = \text{information} + \text{noise}$$

- Not always clear how to interpret model as a (denoised) signal
- Rissanen (2000): subset selection in wavelet regression

MDL Denoising I

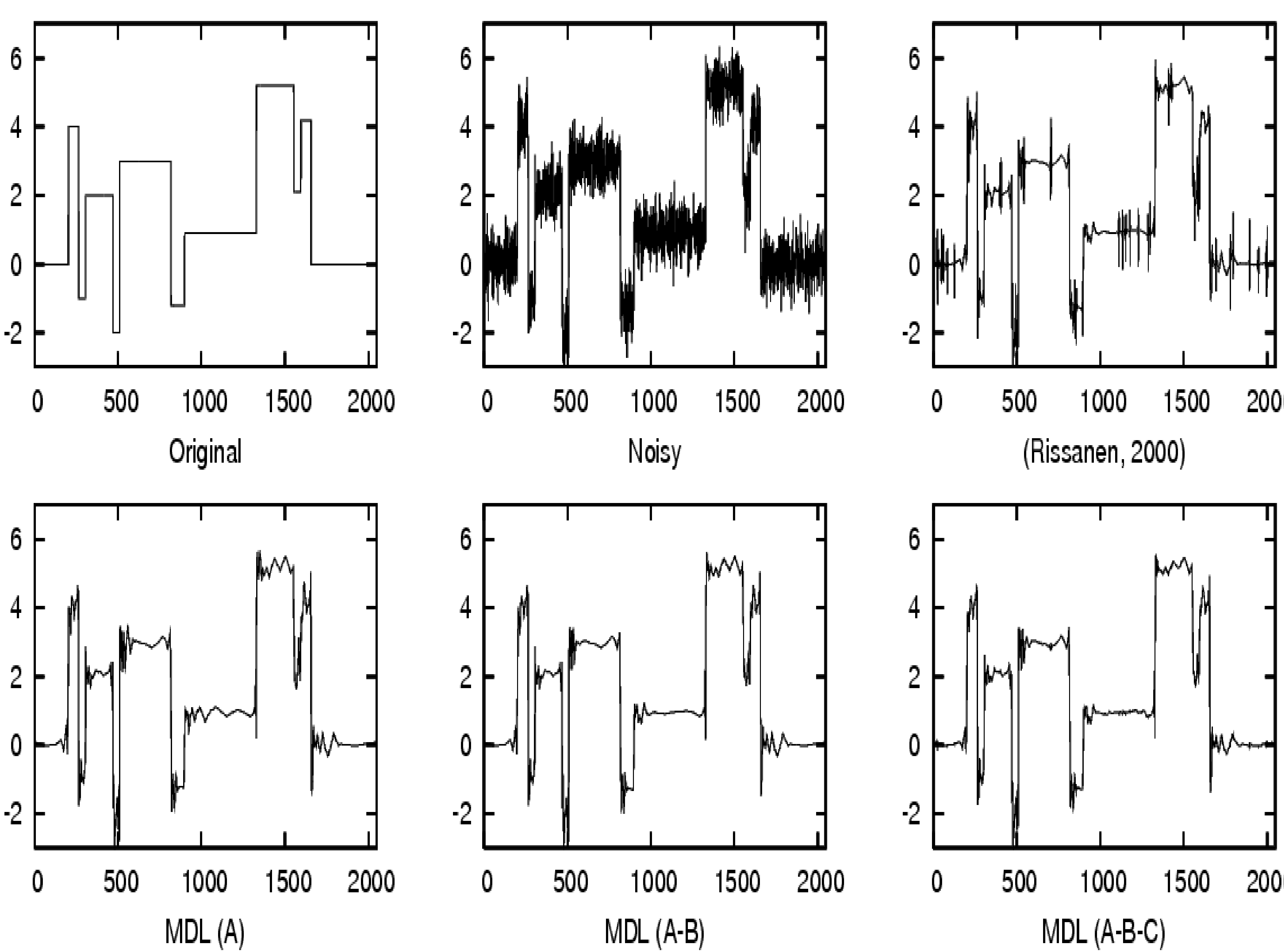
- Select a subset of wavelet basis functions
  - MDL gives a criterion: minimize

    $$L(\text{subset}) + L(x \mid \text{subset})$$

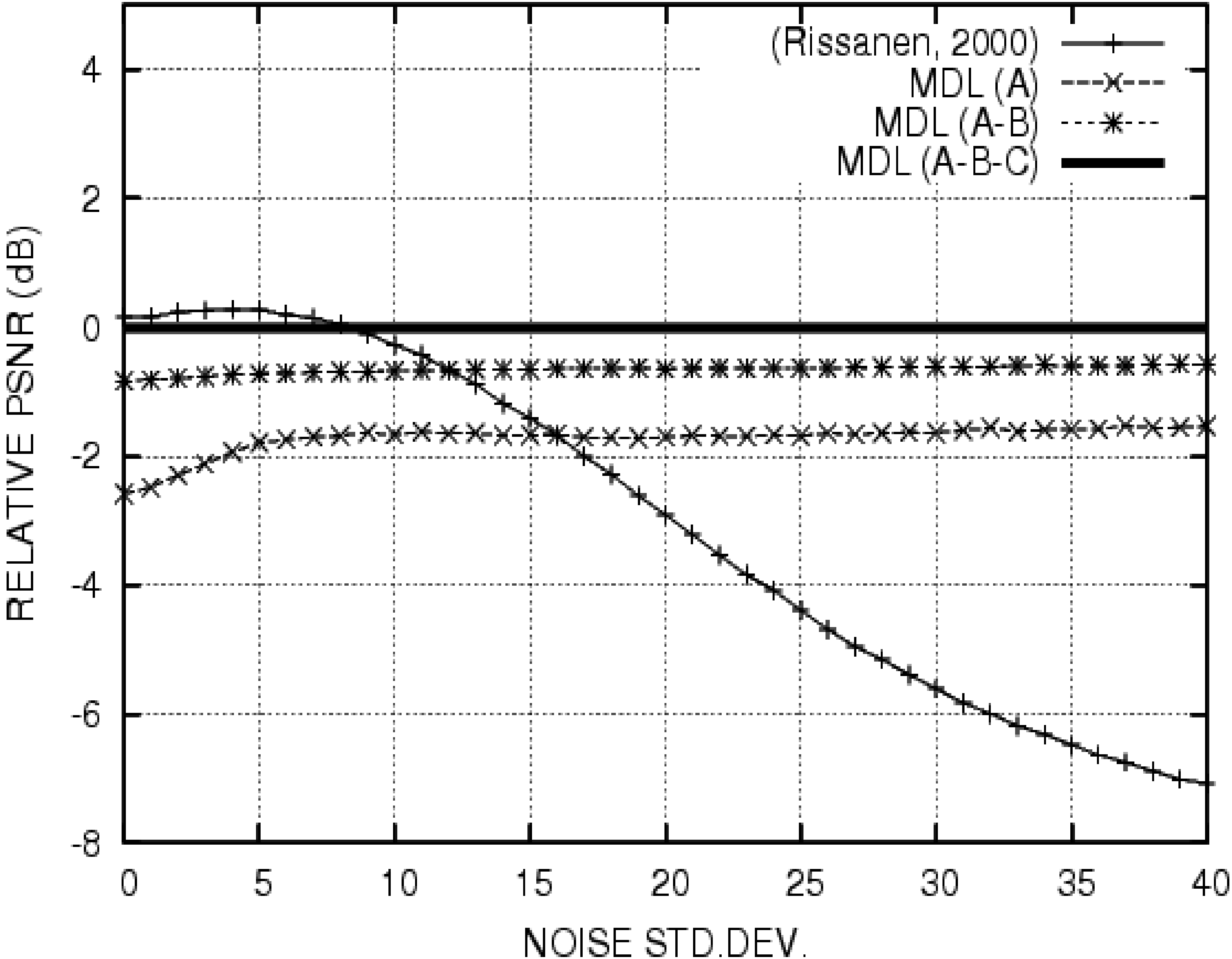
- How to evaluate $L(x \mid \text{subset})$?
- Rissanen (2000): Renormalized NML
- Difficulties in interpreting
- Works well in some cases, fails in others

MDL Denoising II

- Roos, Myllymäki, Tirri (2005): Standard NML with same behavior as renormalized NML
- Roos, Myllymäki, Rissanen (submitted):
  - encoding of L(subset) — huge space: cannot be ignored
  - model (some) coefficient dependencies / subband adaptation
  - predictive mixture codes
- Earlier problems explained by omission of L(subset)
- Improved performance


6
4
2
0
-2
0
500
1000
1500
2000
Original
Noisy
(Rissanen, 2000)
MDL (A)
MDL (A-B)
MDL (A-B-C)


Original
Noisy
PSNR=22.1
MDL
PSNR=24.3
MDL (A)
PSNR=23.9
MDL (A-B)
PSNR=24.9
MDL (A-B-C)
PSNR=25.7



(Rissanen, 2000)
MDL (A)
MDL (A-B)
MDL (A-B-C)
4
2
0
-2
-4
-6
-8
RELATIVE PSNR (dB)
0
5
10
15
20
25
30
35
40
NOISE STD.DEV.