# MDL THEORY AS A FOUNDATION FOR STATISTICAL MODELING

J. Rissanen
Helsinki Institute for Information Technology,
Technical Universities of Tampere and Helsinki, Finland,
and
University of London, Royal Holloway, UK

10/29/2001

# MODELING PROBLEM

**data**:

$$x^n = x_1, \ldots, x_n \quad or \quad (y^n, x^n) = (y_1, x_1), \ldots, (y_n, x_n)$$

and class of models as distributions

$$\mathcal{M}_k = \{p(x^n; \theta) : \theta \in \Omega \subseteq R^k\}, \ \ \mathcal{M} = \bigcup \mathcal{M}_k$$

**model**:

finitely describable distribution that can be fitted to data

traditional 'nonparametric' models excluded; abstractions which cannot be fitted to data

Want a model constructed in terms of given class which extracts **all** properties from data that can be expressed in terms of the class

- **NO** assumptions made about data generating mechanism; in particular, no model in the class assumed to have generated the data

*Central Problem:* How to define 'extractable properties' from 'noisy' data?

In algorithmic theory of information (Kolmogorov):

'property' of $x^n$ : set $A$ which includes $x^n$

Intuition:

- all strings in $A$ share a common property
- size $|A|$ inverse measure of amount of properties:
  - $x^n \in A$, $|A|$ large $\Leftrightarrow$ $x^n$ has few properties = restrictions
  - $x^n \in \{x^n\}$ ($|A| = 1$) $\Leftrightarrow$ $A$ captures all conceivable properties of $x^n$

*Kolmogorov-complexity* $K(x^n)$ = length of shortest program to generate $x^n$ (program = codeword)

*Kolmogorov sufficient statistics decomposition:*

$$A^* = \max\{A \ni x^n : \log|A| + K(A) \cong K(x^n)\}$$

In words: best coding (program for $A^*$) of fewest number of properties of $x^n$ together with best coding of $x^n$, given $A^*$, equals best coding of $x^n$ alone (could have $K(x^n|A)$ instead of $\log|A|$)

In general $K(x^n, A) \cong K(x^n|A) + K(A)$

$\log|A^*|$ (or better, $K(y^n|A^*)$) = code length of 'noise'

$K(A^*)$ = code length of learnable properties = 'information' in $x^n$

Want to do the same relative to model classes $\mathcal{M}_k$ (and $\mathcal{M}$):

$$\hat{L}(x^n; \mathcal{M}_k) = L(x^n|\hat{p}) + L(\hat{p})$$

(stochastic complexity = code length for noise, given best model $\hat{p}$, + information)

Traditionally:

$$\max_{\theta} p(x^n;\theta) \quad \Rightarrow \quad \hat{\theta}(x^n)$$

ML model $p(\cdot;\hat{\theta}(x^n))$

- captures both noise and learnable properties in $x^n$; cannot separate the two
- amount of information $L(\hat{\theta}(x^n))$ infinite
- $p(y;\hat{\theta}(x^n))$ is not best model to predict new data, because $\hat{\theta}(x^n)$ too 'noisy' (not much harm for large $n$; noise effect small)

Similarly

$$\max_{k} p(x^n;\hat{\theta}(x^n)) \quad \Rightarrow \quad \hat{k}(x^n) = \hat{k}$$

and $p(\cdot;\hat{\theta}^{\hat{k}}(x^n))$ not good model (well known; $\hat{k}$ tailored to data $x^n$; disastrous; only adhoc remedies)

**Summary:**

In *orthodox statistics*: accept ML estimate $\hat{\theta}(x^n)$ but reject $\hat{k}(x^n)$

Justification: None; both are parameters!

(or mean)

In *Baysian statistics*: accept Max Posterior estimates $\hat{\theta}(x^n), \hat{k}(x^n)$

Justification: faith

In *new statistics:* accept MDL estimates $\bar{\theta}(x^n), \bar{k}(x^n)$

Justification: They achieve Universal Sufficient Statistics Decomposition extracting learnable information from noisy data

$$\min_{\theta} -\log p(x^n;\theta) = -\log p(x^n;\hat{\theta}(x^n))$$

## Normalized Maximum Likelihood (NML) Model

$$\hat{p}(x^n;\mathcal{M}_k) = \frac{p(x^n;\hat{\theta}(x^n))}{C_n} \tag{1}$$

$$C_n = \int_{\hat{\theta}(y^n)\in\Omega^\circ} p(\cancel{x}^{\,y^n};\hat{\theta}(y^n))dy^n \tag{2}$$

$$= \int_{\hat{\theta}\in\Omega^\circ} h(\hat{\theta};\hat{\theta})d\hat{\theta}; \tag{3}$$

$\Omega^\circ$ interior of $\Omega$ and $h(\hat{\theta};\theta)$ density function on statistic $\hat{\theta}(x^n)$ induced by $p(y^n;\theta)$

*Fact:* $\hat{p}(x^n;\mathcal{M}_k) = \hat{q}(x^n) = \hat{g}(x^n)$ solves MinMax Problem:

$$\min_q \max_g E_g \log \frac{p(X^n;\hat{\theta}(X^n))}{q(X^n)}; \tag{4}$$

code length difference

$q$ and $g$ range over any distributions

**Proof**: The MinMax problem is equivalent with

$$\min_q \max_g D(g\|q) - D(\hat{p}\|g) + \log C_n \geq \max_g \min_q \ldots = \log C_n;$$

equality reached for $\hat{q} = \hat{g} = \hat{p}$ ; $D(g\|q)$ = KL distance

If CLT holds for $\hat{\theta}(x^n)$,

$$\log C_n = \frac{k}{2}\log\frac{n}{2\pi} + \log\int_\Omega |I(\theta)|^{1/2}d\theta + o(1) \tag{5}$$

where $I(\theta)$ is the Fisher information matrix.

Shannon : $\mathcal{M} = \{p\} \Rightarrow C_n = 1$

$$\min_q E_p \log\frac{p}{q} = \log 1 = 0$$

$\hat{q} = p$

## COMPLEXITY and INFORMATION

**Stochastic Complexity** of $x^n$, given $\mathcal{M}_k$:

$$-\log \hat{p}(x^n; \mathcal{M}_k) = -\log p(x^n; \hat{\theta}(x^n)) + \log C_n$$

Justifications:

- MinMax Problem: Best mean code length for the worst case data generating distribution; also
- For all $q(x^n)$ and all $g(x^n) = p(x^n; \theta)$, $\theta \in \Omega - \Lambda_{q,n}$,

$$E_g \log 1/q(X^n) \geq H_g(X^n) + (1-\epsilon) \log C_n,$$

where volume of $\Lambda_{q,n} \to 0$. can replace with $E_g \log 1/p(x^n; \hat{\theta}(x^n))$

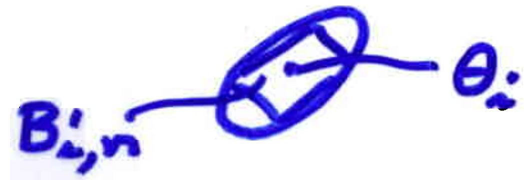
**Information** in $x^n$: $\log C_n$

Justification:

- Balasubramanian: $C_n =$ maximum number of optimally distinguishable models from $x^n$
- Universal Sufficient Statistics Decomposition (next foil)
  USSD

USSD

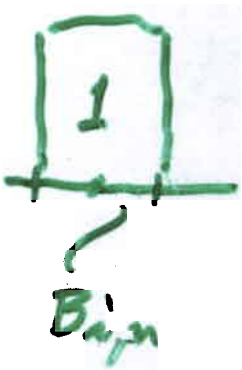
$$D_{i,n} = (\theta-\theta_i)' I(\theta_i)(\theta-\theta_i) = d$$




Partitioning $\Pi_n = \{B_{i,n}\}$ of $\Omega$ with maximal curvilinear rectangles within $D_{i,n}$

Pick $d = \bar{d}$ such that

$$\int_{\hat\theta(y^n)\in B_{i,n}} p(y^n;\hat\theta(y^n))\,dy^n = 1 = \int_{\hat\theta\in B_{i,n}} h(\hat\theta;\hat\theta)\,d\hat\theta$$




$$C_n = \int_{\hat\theta\in\Omega^0} h(\hat\theta;\hat\theta)\,d\hat\theta = \sum_{i=1}^{|\Pi_n|} 1 = |\Pi_n|$$

$$-\log \hat{P}(x^n;\mathcal{M}_k) \cong -\log p(x^n/\hat\theta_i(x^n)) + \log C_n$$

$p(x^n/\hat\theta_i(x^n)) \cong p(x^n;\hat\theta(x^n))$ for $x^n$ such that $\hat\theta(x^n)\in B_{i,n}$

$-\log p(x^n/\hat\theta_i(x^n))$ is code length for noise

$\log C_n$ is code length for optimally distinguishable models $\triangleq$ information

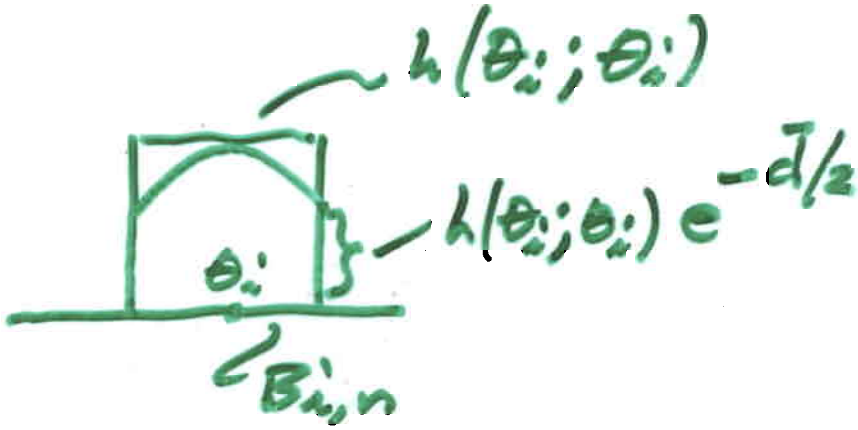
# Finite String Distinguishability

$$\mathrm{Vol}\, B_{i,n} = \left(\frac{4\bar{d}}{nk}\right)^{k/2} |I(\theta_i)|^{-1/2}$$

$$h(\theta_i;\theta_i) \cong \frac{|I(\theta_i)|^{1/2}}{(2\pi)^{k/2}} n^{k/2}$$ (peak of normal density f'n)

$$h(\theta_i;\theta_i)\, \mathrm{Vol}\, B_{i,n} \cong 1 \implies$$

$$\bar{d} = \frac{k\pi}{2}$$




## Worst case loss

$$\max_{\hat{\theta}(x^n) \in B_{i,n}} \log \frac{p(x^n;\hat{\theta}(x^n))}{p(x^n;\theta_i)} = \frac{\bar{d}}{2} - \log \frac{C_n}{\sum_j \int_{B_{j,n}} p(y^n;\theta_j)\,dy^n} < \frac{\bar{d}}{2}$$

## MDL-Principle (global ML-Principle):

Of two model classes $\mathcal{M}_k$ and $\mathcal{N}_j$ prefer former if

$$-\log \hat{p}(x^n;\mathcal{M}_k) < -\log \hat{p}(x^n;\mathcal{N}_j)$$

or equivalently

$$\hat{p}(x^n;\mathcal{M}_k) > \hat{p}(x^n;\mathcal{N}_j)$$

**Justification:** better decomposition of data into noise and the useful information by winner; smaller complexity $\Rightarrow$ shorter code length for noise (grows like $O(n)$ while information grows like $O(\log n)$) $\Rightarrow$ some of what looks like noise with the worse model class extracted as useful information by the better class

For class $\mathcal{M} = \cup_k \mathcal{M}_k$

$$\min_k\{-\log \hat{p}(x^n;\mathcal{M}_k\} \Rightarrow \hat{k}(x^n) \tag{8}$$

$$\hat{p}(x^n;\mathcal{M}) = \frac{\hat{p}(x^n;\mathcal{M}_{\hat{k}(x^n)})}{\int \hat{p}(y^n;\mathcal{M}_{\hat{k}(y^n)})dy^n} \tag{9}$$

- In modeling, to achieve the decomposition important - not to minimize this or that criterion as an estimate of mean loss function, the mean taken with respect to some imagined 'truth'
- most successful criteria are the ones that happen to be close to MDL! (justification for Bayesian techniques)
- no assumption that data be a sample from metaphysical populations

## Linear Regression

$$\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}$$
$$\left\{\begin{array}{ccc} w_{11} & \cdots & w_{1n} \\ & \cdots & \\ w_{m1} & \cdots & w_{mn} \end{array}\right\} =: W$$

$$\begin{aligned} W &= \{w_{ij}\},\ m \times n \text{ regressor matrix} \\ \gamma &= \{i_1, \ldots, i_k\},\ \text{index set, } \mathrm{k} \leq \mathrm{m} \\ W_\gamma &= \{w_{ij} : i \in \gamma\}, \Sigma_\gamma = W_\gamma W_\gamma' \end{aligned}$$

*Model Class* $\mathcal{M}_\gamma$*:*

$$\begin{aligned} x_t &= \sum_{i \in \gamma} \beta_i w_{it} + \epsilon_t,\ t = 1, \ldots, n \\ \epsilon_t &\sim N(0, \sigma^2),\ \sigma^2 = \tau \end{aligned}$$

*ML-solutions:*

$$\begin{aligned} \hat{\beta} &= \Sigma_\gamma^{-1} W_\gamma x,\ x = x^n = (x_1, \ldots, x_n)' \\ \hat{\tau} &= RSS/n = \frac{1}{n}(x'x - \hat{\beta}'\Sigma_\gamma \hat{\beta}) \end{aligned}$$

**NML-density function:** For the normal density functions

$$f(x^n; \gamma, \hat{\tau}, \hat{\beta}) = (2\pi e \hat{\tau})^{-n/2}$$

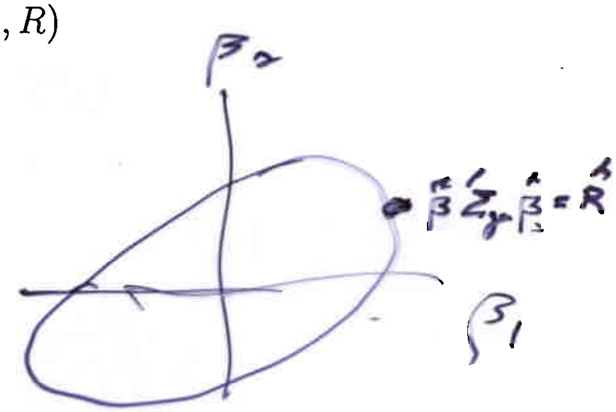
and

$$\hat{f}(x^n; \gamma, \tau_0, R) = \frac{(2\pi e \hat{\tau})^{-n/2}}{\int_{Y(\tau_0, R)} (2\pi e \hat{\tau}(z^n))^{-n/2} dz^n},$$

where

$$Y(\tau_0, R) = \{z^n : \hat{\tau}(z^n) \geq \tau_0,\ \hat{\beta}'(z^n)\Sigma_\gamma \hat{\beta}(z^n) \leq R\};$$

hyperparameters $\tau_0$ and $R$ such that $x^n \in Y(\tau_0, R)$



$\hat{\beta}$ and $\hat{\tau}$ independent and sufficient imply exact formula ($0 < k \le m$):

$$-\log \hat{f}(x^n; \gamma, \tau_0, R) = \frac{n}{2} \ln \hat{\tau} + \frac{k}{2} \ln \frac{R}{\tau_0} + F(k,n) \tag{11}$$

where

$$F(k,n) = -\ln \Gamma(\frac{n-k}{2}) - \ln \Gamma(\frac{k}{2}) + \ln \frac{4}{k^2} + \frac{n}{2} \ln(n\pi) \tag{12}$$

Problem: Optimum $\hat{\gamma}$ with $\hat{k}$ indices for (11) depends on $R$ and $\tau_0$.

**Repeat normalization** for $R$ and $\tau_0$: Optimum values $\tau_0 = \hat{\tau}$ and $R = \hat{R} = \hat{\beta}'\Sigma_\gamma \hat{\beta} \Rightarrow$

$$\begin{aligned} \hat{f}(x^n; \gamma) &= \hat{f}(x^n; \gamma, \hat{\tau}, \hat{R}) / \int_{Y(\tau_1, \tau_2, R_1, R_2)} \hat{f}(y^n; \gamma, \hat{\tau}(y^n), \hat{R}(y^n)) dy^n \\ -\ln \hat{f}(x^n; \gamma) &= \frac{n-k}{2} \ln \hat{\tau} + \frac{k}{2} \ln \hat{R} + F(k,n) - \ln \frac{2}{k} + \ln \ln \frac{\tau_2 R_2}{\tau_1 R_1}. \end{aligned} \tag{13}$$

(values of new hyperparameters irrelevant)

Extend $\hat{f}(x^n;\gamma)$ to *NML* model $\hat{f}(x^n;\Omega)$ for $\mathcal{M} = \cup_{\gamma \in 2^m} \mathcal{M}_\gamma$, ($\gamma$ over all subsets of $\{1, \dots, m\}$):

**Repeat normalization** for $\gamma$: With $\hat{\gamma} = \hat{i}_1, \dots, \hat{i}_{\hat{k}}$ maximizing $\hat{f}(x^n;\gamma) \Rightarrow$

**universal sufficient statistics** demposition:

$$-\ln \hat{f}(x^n; \Omega) = \frac{n-\hat{k}}{2} \ln \hat{\tau} + \frac{\hat{k}}{2} \ln \hat{R} - \ln \Gamma(\frac{n-\hat{k}}{2}) - \ln \Gamma(\frac{\hat{k}}{2}) + \ln \frac{1}{\hat{k}} + \mathit{Const} \tag{14}$$

- first term is code length for **noninformative** 'noise' - incompressible
- the rest define code length for **optimal** model

$$\min_{\gamma = \{i_1, \dots, i_k\}} \left\{ \frac{(n-k)}{2} \ln \hat{\tau} + \frac{k}{2} \ln(n\hat{R}) + \frac{n-k-1}{2} \ln \frac{n}{n-k} - (k+1) \ln k \right\} + \mathrm{Const.}$$

dep. on $\gamma$

## MDL Denoising Problem

Intuitively:

$$
\begin{aligned}
x_t &= \hat{x}_t + \hat{\epsilon}_t,\ t = 1, \ldots, n \\
\hat{\epsilon}_t &= \text{`noise'} \\
\hat{x}_t &= \text{`smooth' signal}
\end{aligned}
$$

Natural formalization by universal sufficient statistics:

- noise = incompressible part in data, given model class
- smooth signal = information bearing part defined by optimal model

**Model class:** linear regression with normal family for deviations

$n \times n$-matrix $W$, rows defining orthonormal basis, $WW' = I_n$

defines transform $c \leftrightarrow x$,

$$
\begin{aligned}
c &= Wx,\ x' = x^n = x_1, \ldots, x_n \\
x &= W'c,\ c' = c_1, \ldots, c_n
\end{aligned}
$$

Hence, $c'c = x'x$

Example : W defined by wavelets

For $W_\gamma = \{w_{ij} : i \in \gamma\}$, $\gamma = \{i_1, \ldots, i_k\} \in 2^n$, set of indices of nonempty subsets of $n$ basis vectors

$\hat{f}(x;\Omega) \Rightarrow$ criterion

$$\min_{\gamma \in 2^n}\{(n-k)\ln\frac{c'c - \hat{S}_\gamma}{n-k} + k\ln\frac{\hat{S}_\gamma}{k} - \ln\frac{k}{n-k}\}, \tag{16}$$

where

$$\hat{S}_\gamma = \sum_{i\in\gamma} c_i^2. \tag{17}$$

• NO arbitrarily selected parameters!

**Theorem 3** *For orthonormal regression matrices the index set $\hat{\gamma}$ that minimizes the criterion (16) is given either by the indices $\hat{\gamma} = \{(1), \ldots, (k)\}$ of the $k$ largest or the $k$ smallest $\hat{\gamma} = \{(n-k+1), \ldots, (n)\}$ coefficients in absolute value for some $k = \hat{k}$.*

- Data for denoising: $\hat{x}^n$ simpler than noise $x^n - \hat{x}^n$; hence take the largest coefficients:

$$\min_k C_{(k)}(x) = \min_k\{(n-k)\ln\frac{c'c - \hat{S}_{(k)}}{n-k} + k\ln\frac{\hat{S}_{(k)}}{k} - \ln\frac{k}{n-k}\} \tag{18}$$

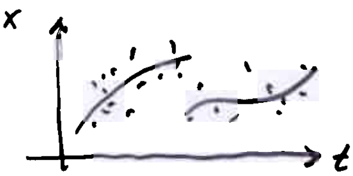
- With $\hat{c}^n$ denoting the column vector defined by the coefficients $\hat{c}_1, \ldots, \hat{c}_n$, where $\hat{c}_i = c_i$ for $i \in \{(1), \ldots, (\hat{k})\}$ and zero, otherwise,
- signal recovered is $\hat{x}^n = W\hat{c}^n$.
- threshold more intricate than Donoho-Johnstone threshold $\hat{\sigma}\sqrt{2\ln n}$.

  Notice. Donoho-Johnstone traditional risk based reasoning circular: $\hat{\sigma}$ defines noise by the threshold, and noise determines its variance! Can be resolved only by an arbitrary estimation of $\hat{\sigma}$.

**Examples with Wavelets:**

With wavelets $W$ is square $n \times n$ matrix: $c = Wx$, $x = W'c$

**Example 1:**




Data:

$x_t = f(t) + e_t$ consist of two piecewise polynomials $f(t)$ sampled at 512 points in unit interval; normal 0-mean, 0.01-variance noise $e_t$ added (G.P. Nason).

Results:

The threshold obtained with the *NML* criterion is $\lambda = 0.246$. This is between the two thresholds called VisuShrink $\lambda = 0.35$ and GlobalSure $\lambda = 0.14$, (Donoho and Johnstone); also close to $\lambda = 0.20$, obtained by Nason with very complex cross-validation procedure

**Example 2:**

Data:

128 samples from a voiced portion of speech.

Results:

The *NML* criterion retains 42 coefficients exceeding threshold $\lambda = 7.3$ in absolute value. Noise variance $\hat{\tau} = \Sigma_t (x_t - \hat{x}_t)^2 = 5.74$.

Donoho-Johnstone threshold $\lambda = \sqrt{2\hat{\tau} \ln 128} = 10.3$. Noise variance $\hat{\tau} = 10.89$.

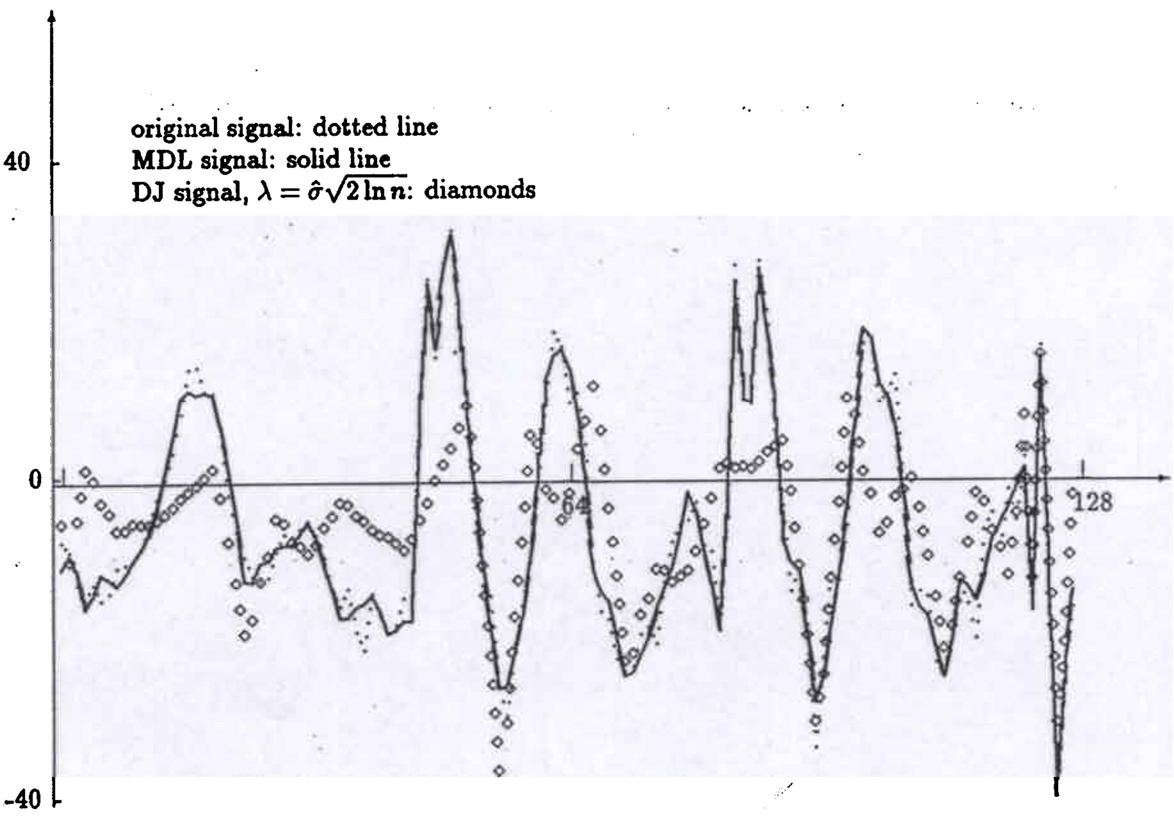



Figure 1. Speech signal smoothed with Daubechies' N=6 wavelet