# Akaike 's Information Criterion (AIC)



- Based on:
  - the maximization of the expected log-likelihood.
  - The fact that the maximum log-likelihood is a biased estimator of the expected log-likelihood, with a bias equal to the number $k$ of free parameters in the model.
- One minimizes: $\mathrm{AIC}(k) = -2l\left(\hat{\theta}_k\right) + 2k$
- AIC has been reported to consistently overestimate even the order of simple linear AR models

# The Model Selection Problem



Maximum Likelihood: $\log P(x \mid \theta)$, $x = [x_1, \ldots, x_N]$, $\theta = [\theta_1, \ldots, \theta_k]$

And if $k$ is not given? $k \longrightarrow N$ failure!

**Rissanen: All models can be regarded as codes**

$\Downarrow \quad \Downarrow \quad \Downarrow$

**The best model is the one corresponding to the shortest encoding of the data**

# How to encode the data



- Let us suppose we have a random process generating $(Y_t, Z_t)$ in R x $R^d$, and a relationship:

$$y_t = F(Z_t) + \varepsilon_t$$

with $\varepsilon_t$ i.i.d. with finite variance.

- If $P$ is a probability distribution on the data, a particular realization $x = \{(y_t, Z_t)\}$, $t = 1, \ldots, N$, can be coded with a minimum code length of

$$-\log_2 P(x) \text{ bits}$$

## How to transmit the data



- One actually transmits a two-part code.
- The first part is: $\hat{F}(X) = G(X, \theta)$
- This will allow the receiver to decode the second part, i.e. the encoded data.
- The total code length is

$$L(x, \theta) = -\log P(x \mid \theta) + L(\theta)$$

# How to encode a model

- The elements of θ are real numbers, so one as to truncate them in order to transmit them in a finite code length.

$$1011.011 \longrightarrow 1011011$$

Accuracy: $\delta = 2^{-3}$

$\Rightarrow$ **One has to know how to encode integers**

- Rissanen's approach assumes a prior distribution on the parameters.

# Universal Prior for Integers



$$L(x,\theta) = -\log P(x|\theta) + L(\theta)$$

Bayes: $$P(x|\theta) = \frac{P(x,\theta)}{P(\theta)}$$

$$\Rightarrow \quad P(x|\theta) = 2^{L(\theta)} \cdot 2^{-L(x,\theta)}$$

**Distribution!**

# Prefix Codes



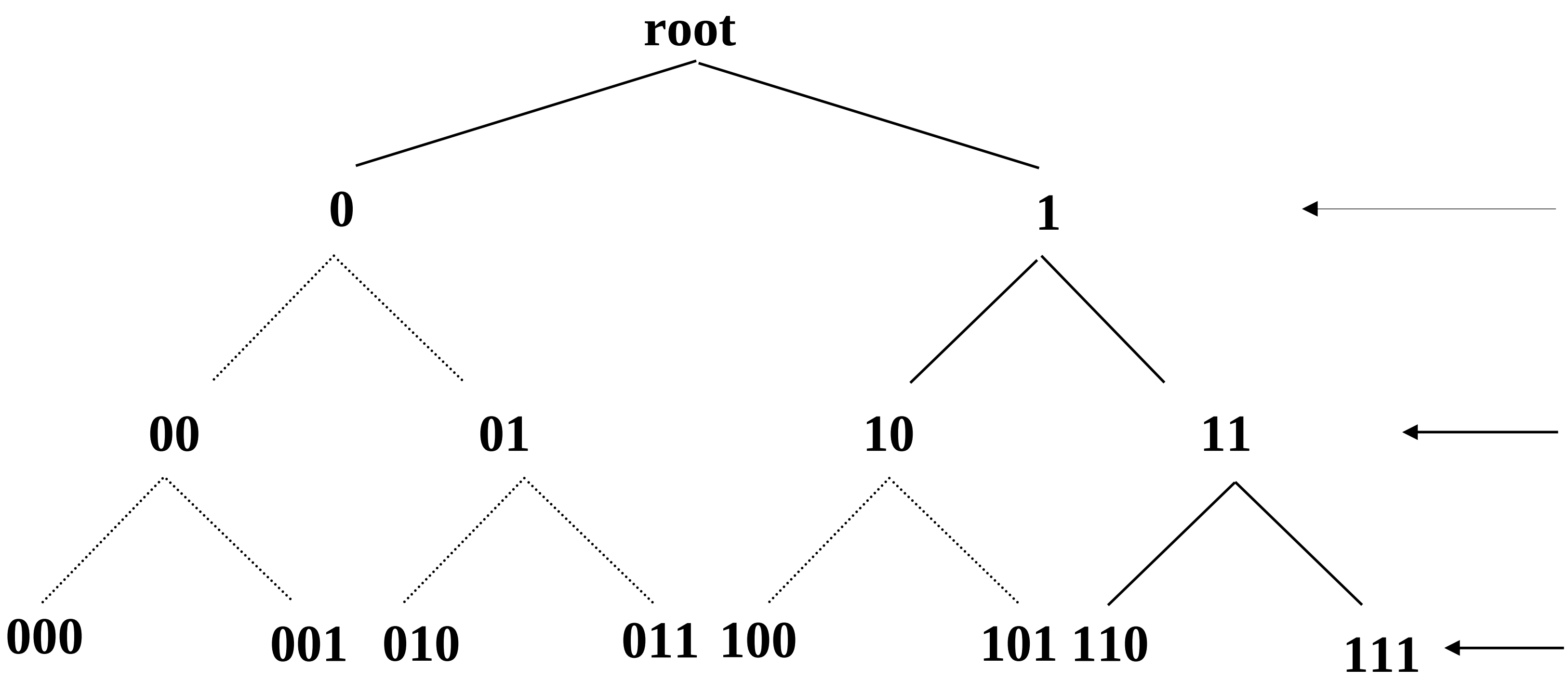
Example: $C(a) = 0 \quad C(c) = 110$

$C(b) = 10 \quad C(d) = 111$

$$1011000111010 \longrightarrow bcaadab$$

# Kraft Inequality






$$\sum 2^{-L(i)} \leq 1$$

# OPTIMAL CODES

- The average length:

$$L = -\sum_{i} p_i L(i)$$

of a prefix code is bounded below by:

$$H = -\sum_{i} p_i \log_2(p_i)$$

*H* the entropy of the code.

# Coding of Integers – first approach



- Let us define:

$$n = 9 \Rightarrow b(n) = 100 \Rightarrow \langle b \rangle = 100001$$

$$|b(n)| = \lfloor \log_2 2n \rfloor$$

- One obtains a prefix code with:

**Code word:** $\langle b(|b(n)|) \rangle b(n)$

**Code length:** $|b(n)| + 2|b(|b(n)|)|$

# Coding of Integers – first approach



**Example:** $n = 9$

$$b(9) = 1001$$

$$b(|b(9)|) = 100$$

$$\langle b(|b(9)|) \rangle = 100001$$

$$\langle b(|b(9)|) \rangle b(9) = 1000011001$$

# Coding of Integers – second approach

**Kraft Inequality does not hold for pure binary representation**

**Idea: preamble which is code string length**

**How to encode the code string length?**

**Attach preamble which is the length of the code string length ...**

| | | | | |
|---|---|---|---|---|
| ... | $l_3$ of $l_2$ | $l_2$ of $l_1$ | $l_1$ of string | string |

# Coding of Integers – second approach



**Decoding rule:** **Integer *j* announces the next length in the following *j+1* positions.**

**Example:** **10100100000|01110…**

**10 100 10000 0... → 16**

# Universal Prior for Integers



- If we define: $\log_2^* n = \log_2 n + \log_2 \log_2 n + \ldots$

- It can be proven that:

$$\sum 2^{-\log_2^*(n)} = c$$

- And with:

$$L_0(n) = \log_2^* n + \log_2 c$$

$\boxed{Q(n) = 2^{-L_0(n)}}$ is a universal prior for integers

# Prior based on computer representation



- Each parameter $\theta_j$ can be expressed with the normalized floating-point binary number:

$$0.1a_1a_2\cdots\times 2^{m_j}$$

- If it is truncated to $\bar{\theta}_j = 0.1a_1a_2\cdots a_{n_j}\times 2^{m_j}$

then the error is at most:

$$\delta_j = 2^{-n_j}$$

# Prior based on computer representation

- As a consequence, the total code length for $k$ parameters is:

$$L(\bar{\theta}) = \sum_{j=1}^{k} L_0\left(\left\lfloor 1/\delta_j \right\rfloor\right) + \sum_{j=1}^{k} L_0\left(\left\lfloor \log\left(2\max\left\{\bar{\theta}_j, 1/\bar{\theta}_j\right\}\right)\right\rfloor\right)$$

- The log log… terms vary slowly, and, most of the time, the exponent cost can be fixed at $m$ bits. Then, this expression is well approximated by:

$$\tilde{L}(\bar{\theta}) = \sum_{j=1}^{k} \log\frac{\gamma}{\delta_j} \quad \text{with } \gamma = 2^m$$

# Minimum Description Length



- The total description length is:

$$L(x,\bar{\theta}) = L(x|\bar{\theta}) + L(\bar{\theta})$$

- However, it should not be too far, and:

$$L(x,\bar{\theta}) \le L(x,\hat{\theta}) + \frac{1}{2}\delta^{\mathrm{T}} Q \delta$$

with $Q = D_{\theta\theta} L(x|\theta)\big|_{\theta=\hat{\theta}}$

# Minimum Description Length



- We obtain:

$$L(x,\bar{\theta}) \le L\left(x \middle| \hat{\theta}\right) + \frac{1}{2}\delta^{\mathrm{T}} Q \delta + k \log \gamma - \sum_{j=1}^{k} \log \delta_j$$

- Minimization over $\delta$ gives:

$$(Q\delta)_j = 1/\delta_j \quad \text{for each } j$$

- The bound on minimum description length is then:

$$\mathrm{MDL}(k) = L\left(x \middle| \hat{\theta}\right) + \left(\frac{1}{2} + \log \gamma\right) k - \sum_{j=1}^{k} \log\left(\hat{\delta}_j\right)$$

# MDL for pseudo-linear models



- These models have the general form:

$$G(z,\theta) = \sum_{i=1}^{m} \theta_i f_i(z)$$

- Typically, one assumes the errors are normally distributed, and ML estimation is simply least-squares estimation.

$$\min \| y - V\theta \|$$

$$y = [y_1, \cdots, y_N]^{\mathrm{T}} \quad \theta = [\theta_1, \cdots, \theta_m]^{\mathrm{T}} \quad V_{ij} = f_j(z_i)$$

# MDL for pseudo-linear models



- Under these assumptions:

$$MDL(k) = \frac{1}{2} N \cdot \ln \hat{\sigma}_e^2 + k\left(\frac{1}{2} + \ln \gamma\right) - \sum_{j=1}^{k} \ln \hat{\delta}_j + C$$

- For $N$ large, it is possible to show that:

$$Q = D_{\theta\theta} L\left(x \middle| \theta\right)\Big|_{\theta=\hat{\theta}} \approx \mathrm{diag}\left(\frac{1}{N}\right)$$

It corresponds to the well-known fact that the parameter estimates are asymptotically Gaussian, non-correlated, with variance $1/N$.

- This gives:

$$\hat{\delta}_j \approx \sqrt{N}$$

- In these conditions also, the term $k/2$ becomes small with respect to the other terms, and ones ends up with the simplifed MDL criterion (a factor 2 is used):

$$MDL(k) = N \cdot \ln \hat{\sigma}_e^2 + (k+1) \cdot \ln N$$

# Example: MA(5)



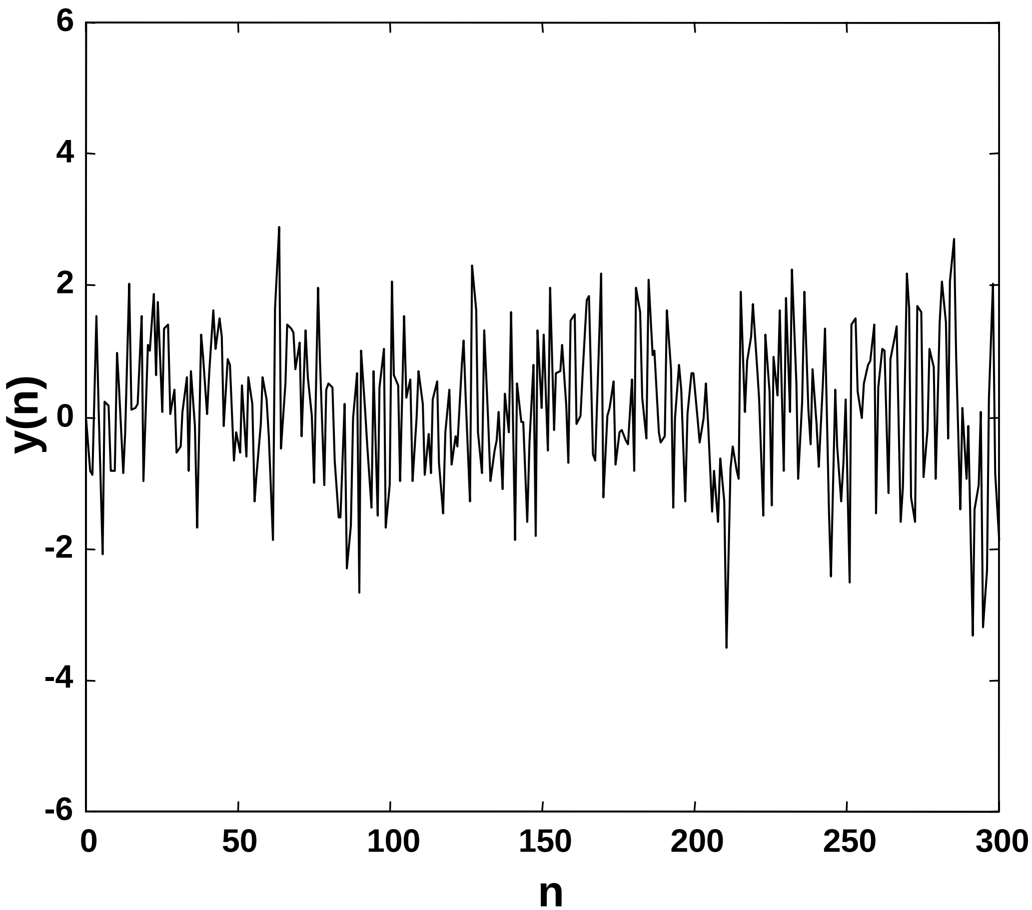
$$y(n) = \sum_{m=1}^{5} a_i x(n-i) + e(n)$$

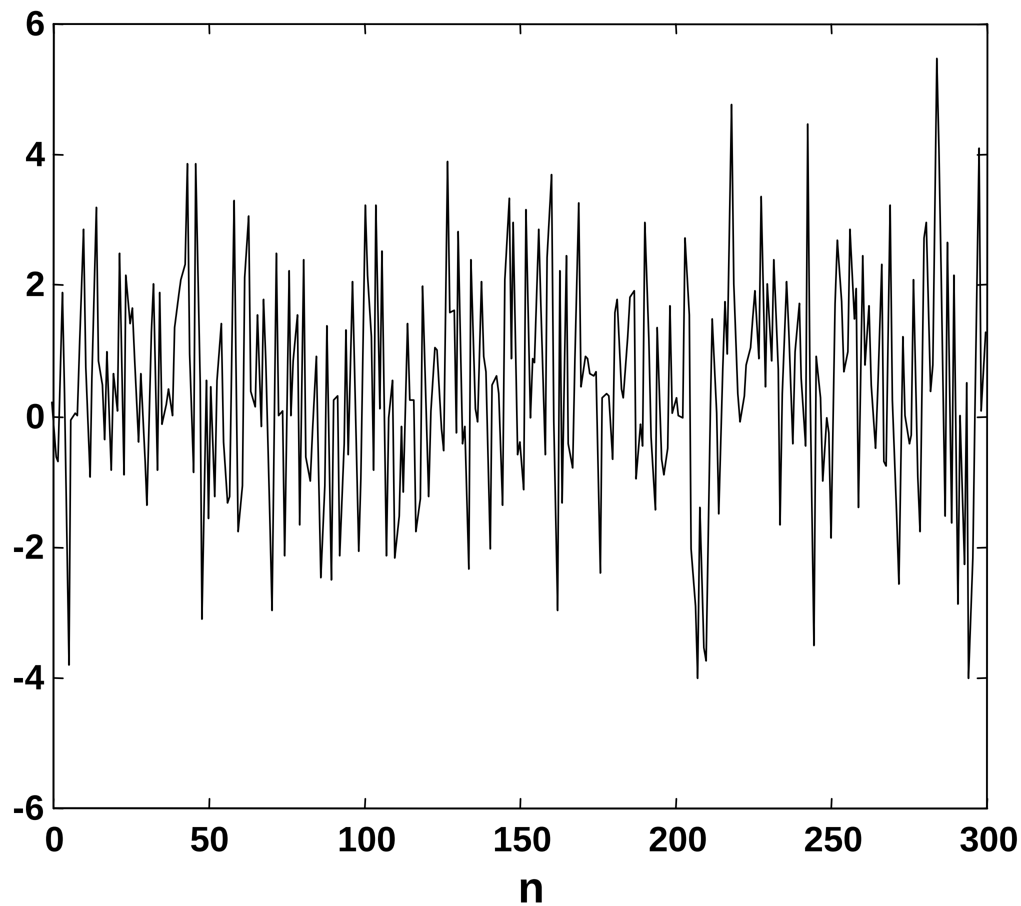
$$e(n) \sim N(0\,, \qquad )$$

# Example: MA(5)



$$\sigma_e^2 = 0.01 \qquad\qquad \sigma_e^2 = 1.0$$



# Orthogonal least squares method:

$$y(n) = \sum_{m=0}^{M} g_m w_m(n) + e(n)$$

$$\boxed{\begin{aligned} w_m(n) &= x(n-i) \\ w_m(n) &= x(n-i)\, x(n-j) \\ w_m(n) &= x(n-i)\, x(n-j)\, x(n-k) \end{aligned}}$$

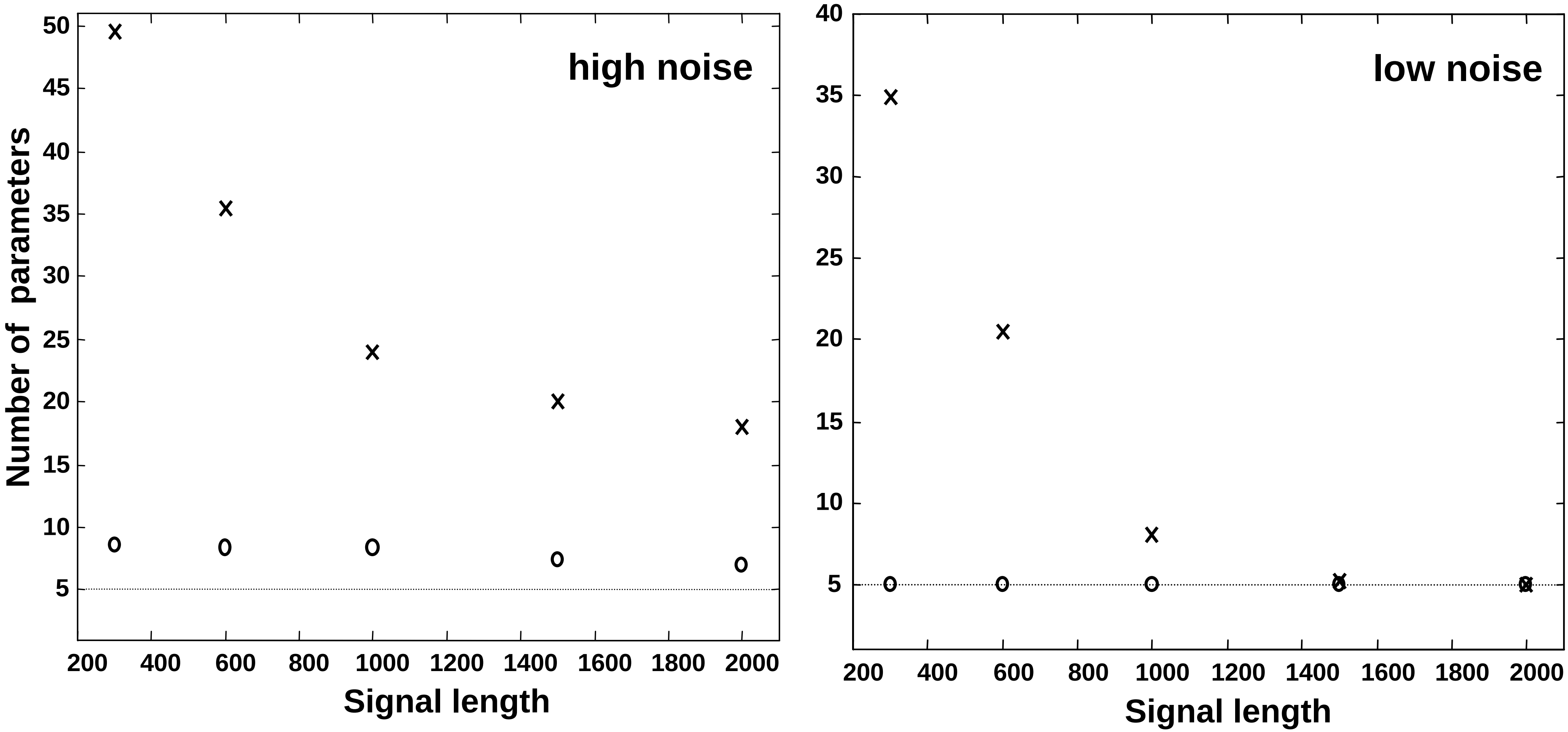
$$\hat{y}_1(n) = g_1 w_1(n)$$

$$\hat{y}_2(n) = g_1 w_1(n) + g_2 w_2(n)$$

$$\vdots$$

$$\hat{y}_M(n) = g_1 w_1(n) + g_2 w_2(n) + \ldots + g_M w_M(n)$$

# Model Selection, Polynomials of Order 3

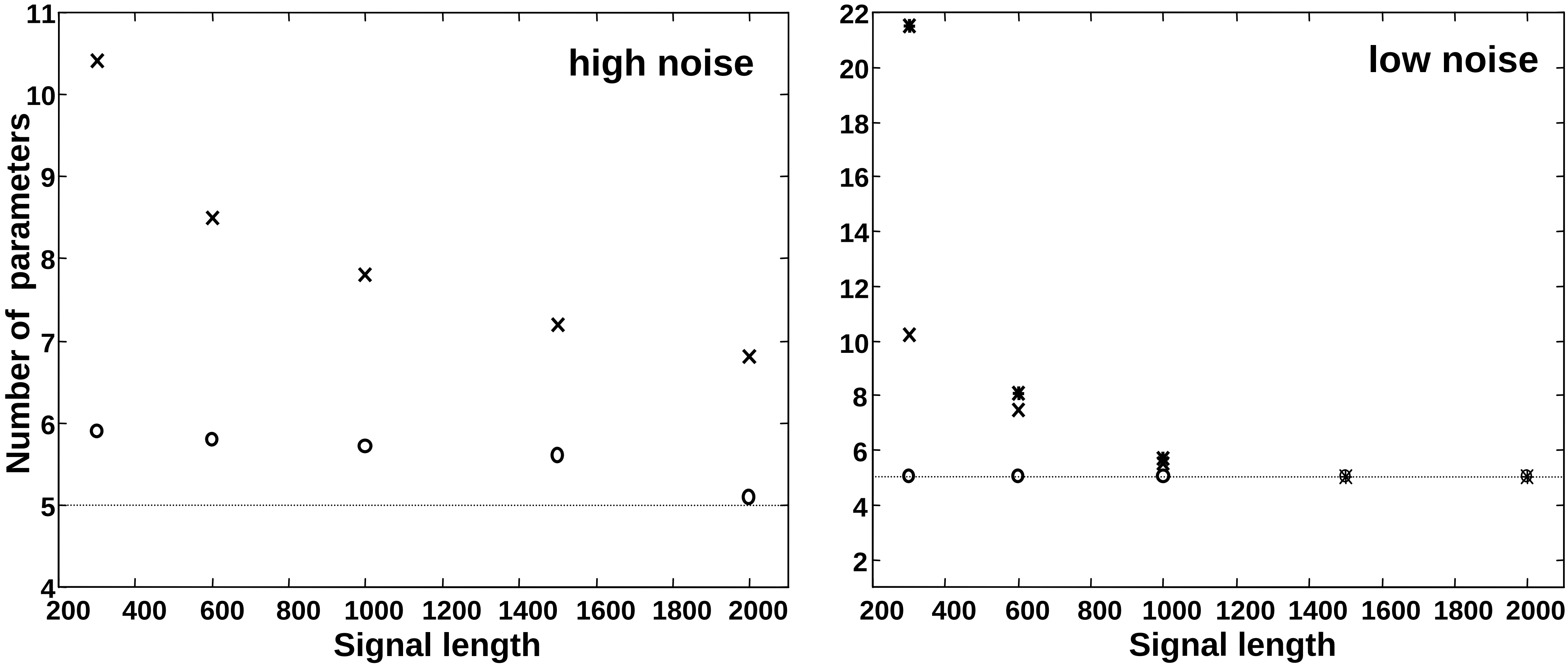



**x : MDL *without* accuracy computation**

**o : MDL *with* accuracy computation**

# Model Selection, Polynomials of Order 2





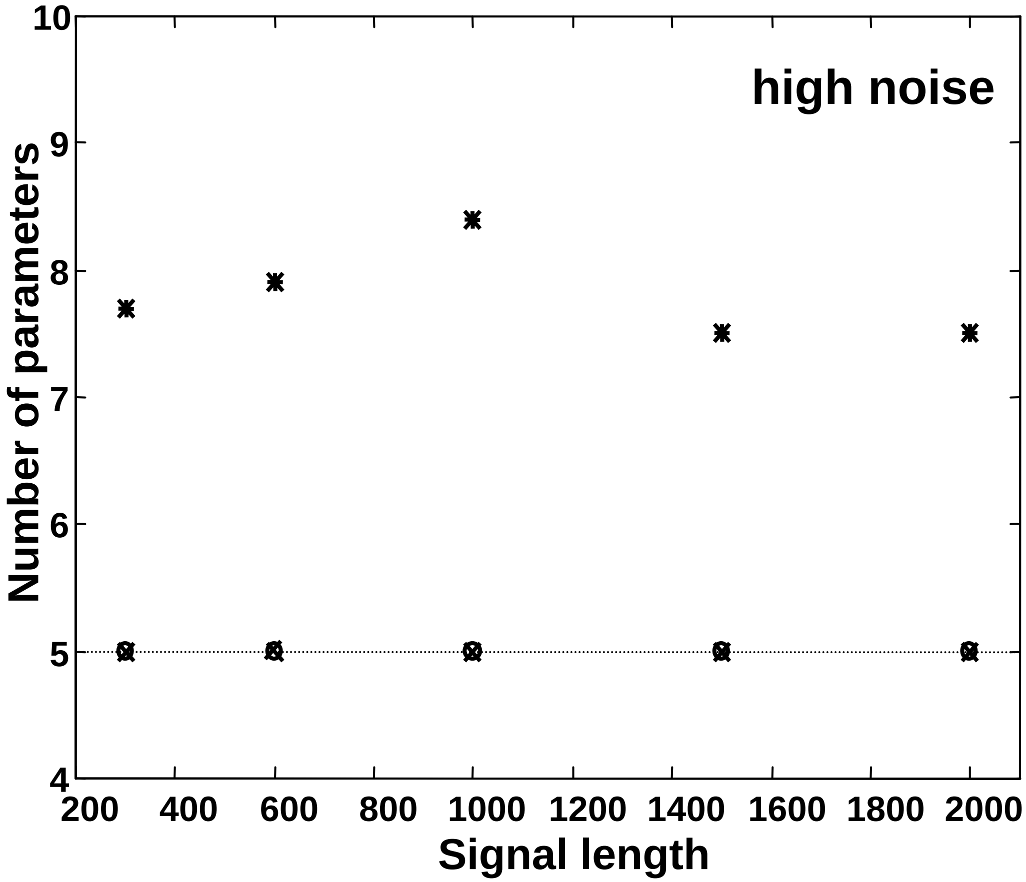

x : MDL *without* accuracy computation

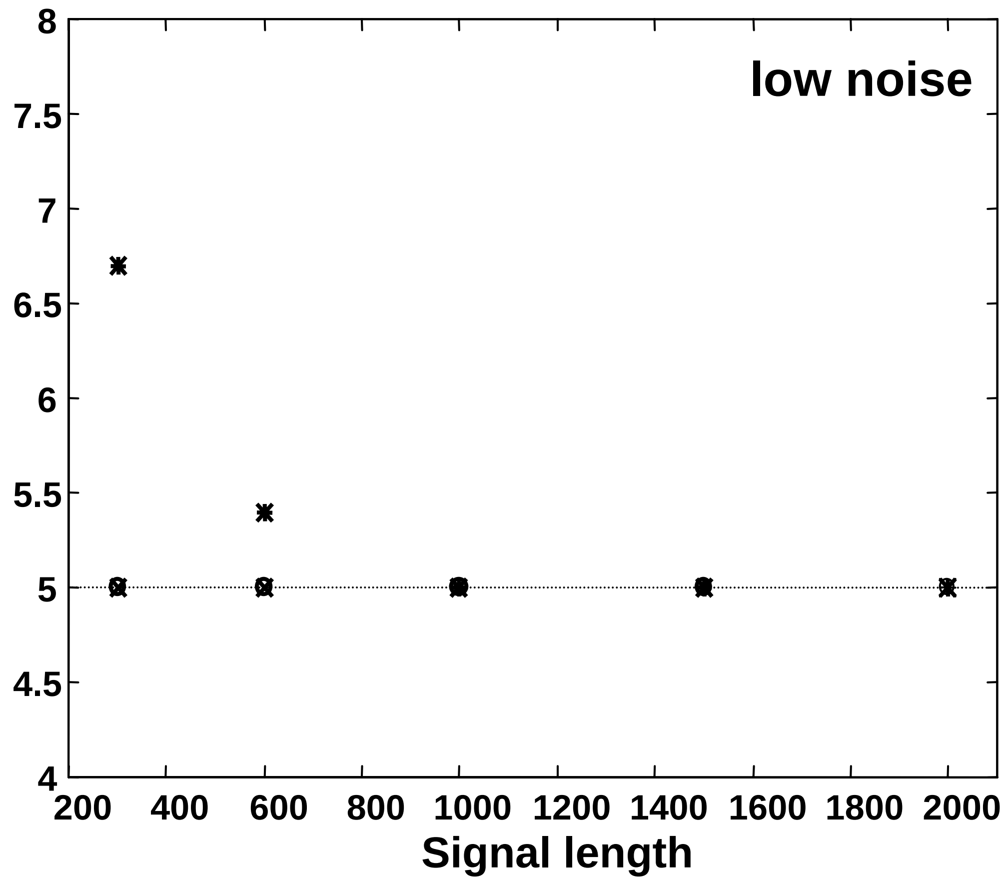
o : MDL *with* accuracy computation

* : AIC

# Model Selection, Linear Models






**x : MDL *without* accuracy computation**

**o : MDL *with* accuracy computation**

*** : AIC**

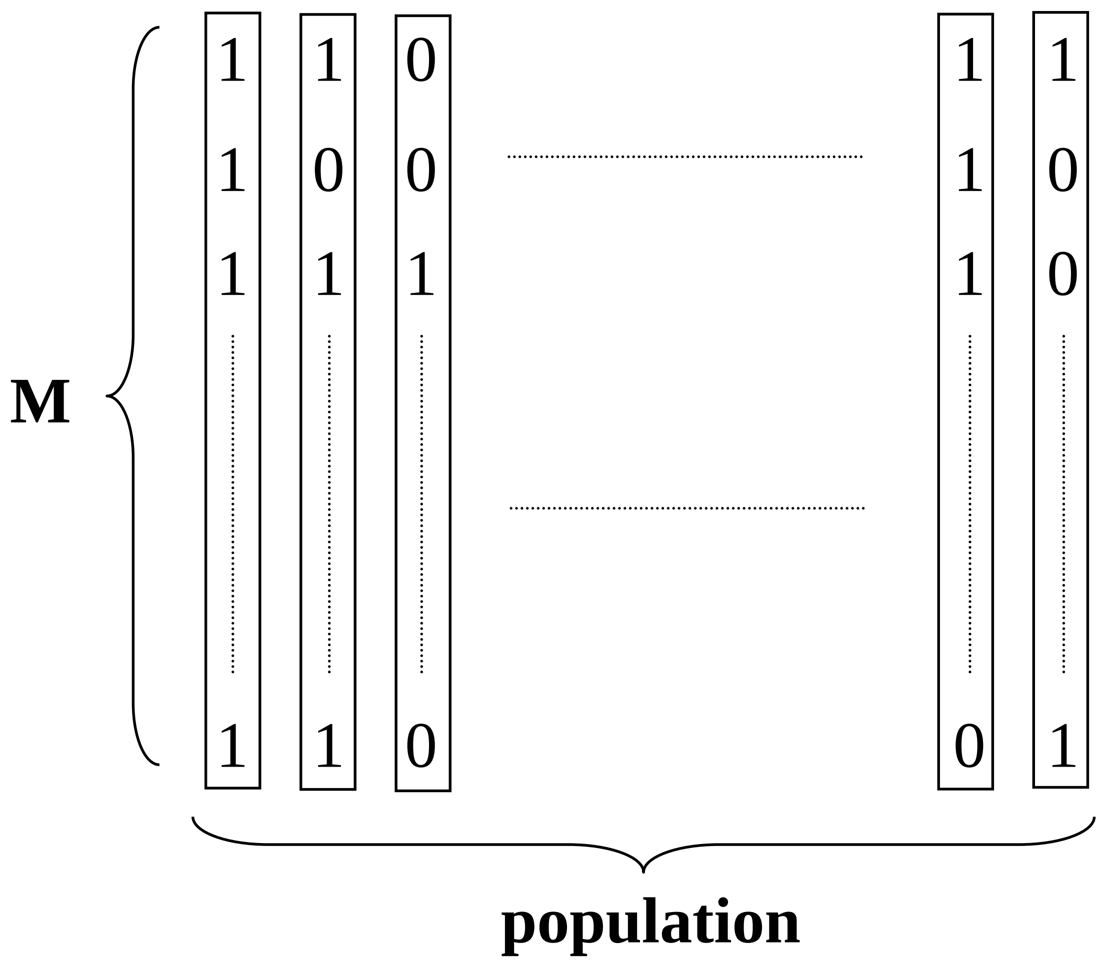
# Genetic Algorithm (GA)



- Small population
- Binary coding of regressors
- Three operator GA
  - Reproduction of the fittest
  - Mutation
  - Crossover
- Fitness function
  - Minimum description length (MDL)

# Genetic Algorithm (GA)






**1: select regressor**

**0: disregard regressor**

# Genetic Algorithm (GA)



- Mutation

10011100 ⟶ 10011000

- Crossover

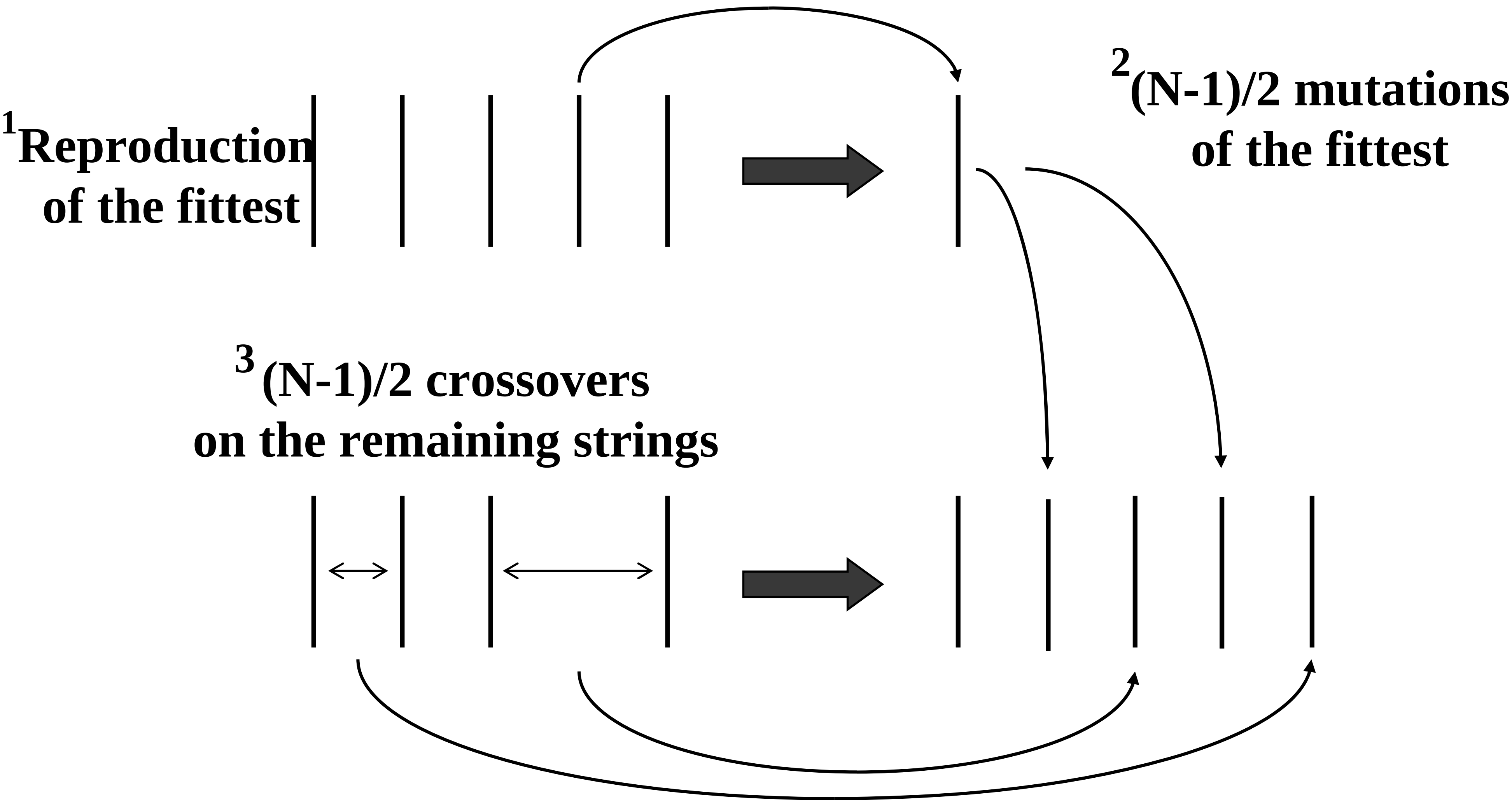
100|11100

⟶ 10001010

011|01010

# Genetic Algorithm (GA)





# Average Number of Generations



$$y(n) = a_1 y(n\text{-}4) + a_2 y(n\text{-}8) + e(n)$$

| Population | Chromosome length | | |
|---|---|---|---|
| | **10** | **20** | **30** |
| **9** | **7.67 ± 3.7** | **18.50 ± 6.7** | **31.05 ± 9.0** |
| **11** | **5.84 ± 2.3** | **13.23 ± 5.2** | **29.48 ± 8.3** |
| **13** | **6.85 ± 3.3** | **12.96 ± 4.3** | **23.80 ± 7.8** |

# Average Number of Generations and Evaluations 

$$y(n) = a_1 x(n-3) + a_2 x(n-1) x(n-3) + a_3 x(n-2) x(n-4) + a_4 x(n-3) x(n-5) + e(n)$$

| Population | Generations | Evaluations |
|---|---|---|
| **7** | **24.97 $\pm$ 8.8** | **150.82** |
| **9** | **17.88 $\pm$ 6.2** | **144.04** |
| **11** | **15.33 $\pm$ 5.1** | **154.30** |
| **13** | **14.57 $\pm$ 4.6** | **175.84** |

# References


1. K. Judd, A. Mees, ″On selecting models for nonlinear time series,″ *Physica D*, vol. 82, pp. 426-444, 1995.
2. P.D. Grünwald, *The Minimum Description Length Principle*, MIT Press, Cambridge, Mass., 2007.
3. J.-M. Vesin and R. Grueter, "Model selection using a simplex reproduction genetic algorithm," *Signal Processing*, vol. 7, no. 3, June 1999.