

# Nonlinear Modeling: Model selection

## Introduction to Statistical Modelling

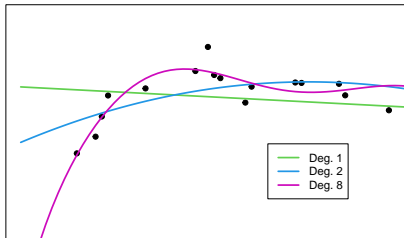
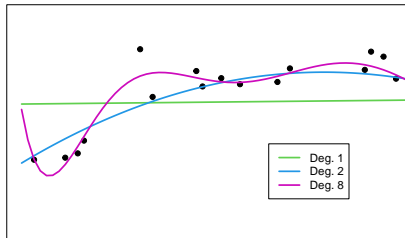
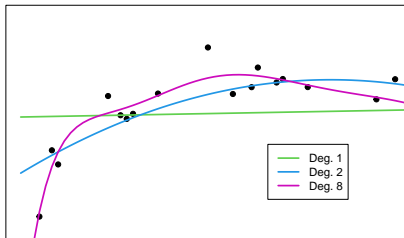
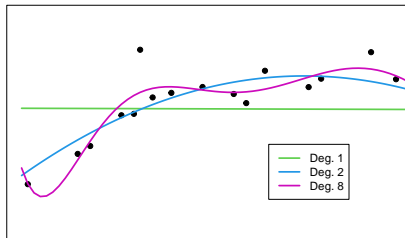
Prof. Joris Vankerschaver

# Model Selection

# Model selection

- Also called *structure characterisation*
- Problem: “perfect” model and “true” parameters are unknown.
- Goal: **Select best model structure from set of candidate models, based on experimental data**

# Which model fits the data the best?



# Two sources of error

**Bias:** *How well does the model fit the data?*

- Error due to non-modeled phenomena.
- Decreases as model gets more complex.

**Variance:** *How well does the model do on new, unseen data?*

- Decreases with more data.
- Increases as model gets more complex.

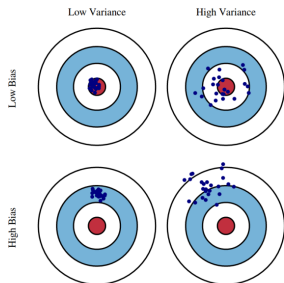


Figure adapted from <http://scott.fortmann-roe.com/docs/BiasVariance.html>

# Bias and variance are complementary

For a model  $M_D(x)$  on a dataset  $D$ , the error decomposes as

$$\text{Error}[M_D(x)] = \text{Bias}[M_D(x)]^2 + \text{Var}[M_D(x)] + \text{Noise}.$$

Goal model selection: select model with smallest total error =  
compromise between bias error and variance error

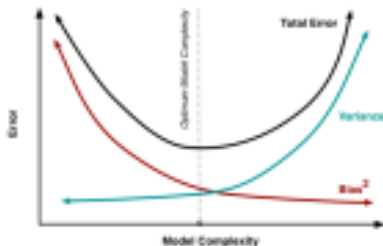
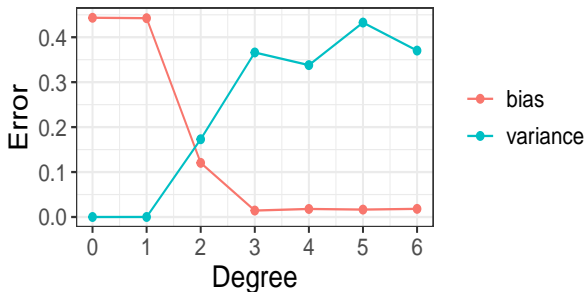


Figure adapted from <http://scott.fortmann-roe.com/docs/BiasVariance.html>

# Model selection for linear models

- Same data as before (slide 1)
- Polynomial model  $y \sim 1 + x + x^2 + \dots + x^d$



- Model of degree 2 (quadratic curve) gives best fit (not too complex, not too simple)
- Bias and variance in general **difficult to calculate**, need easier criteria.

## Case study: biodegradation test

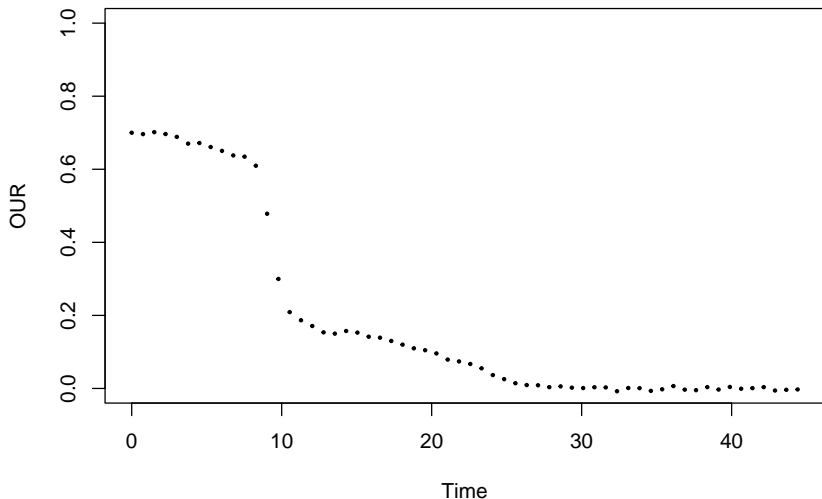
**Waste treatment:** Measure the oxygen uptake rate (OUR) during oxidation of biodegradable waste products by activated sludge.

- Shape respirogram depends on degradation kinetics and quantity added products
- Not known a priori → measure and test several models



## Case study: biodegradation data

1.5 data points per minute, acquired using dissolved oxygen (DO) sensor.



## Case study: general model

- $k$  pollutants  $S_1, \dots, S_k$ .
- Oxygen uptake rate

$$OUR = \sum_{i=1}^k (1 - Y_i) r_{S_i}$$

where  $Y_i$  is the yield, (fraction of substrate  $S_i$  that is not oxidated but transformed in biomass  $X$ ), and  $r_{S_i}$  the degradation rate of  $S_i$ .

- Candidate models differ in number of pollutants  $k$  and choice of degradation rates  $r_{S_i}$ .

## Case study: candidate models

**Model 1:** degradation of one pollutant according to first-order kinetics. Gives *exponentially* decreasing OUR-curve.

$$r_{S_1} = \frac{k_{max1}X}{Y_1}S_1$$
$$OUR = (1 - Y_1)r_{S_1}$$

## Case study: candidate models

**Model 2:** degradation of one pollutant according to *Monod kinetics*.

$$r_{S_1} = \frac{\mu_{max1} X}{Y_1} \frac{S_1}{K_{S_1} + S_1}$$
$$OUR = (1 - Y_1) r_{S_1}$$

## Case study: candidate models

**Model 3:** simultaneous degradation of two pollutants according to Monod kinetics (*double Monod*) without interaction.

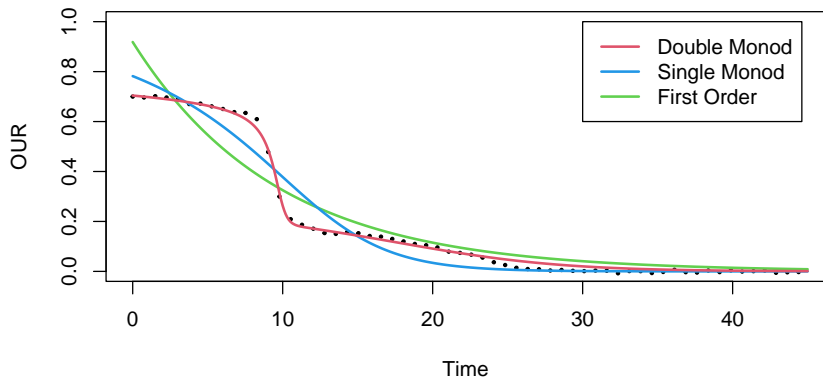
$$r_{S_1} = \frac{\mu_{max1} X}{Y_1} \frac{S_1}{K_{S_1} + S_1}$$

$$r_{S_2} = \frac{\mu_{max2} X}{Y_1} \frac{S_2}{K_{S_2} + S_2}$$

$$OUR = (1 - Y_1)r_{S_1} + (1 - Y_2)r_{S_2}$$

## Case study: parameter estimation

Dataset (dots) and best fits (calibrated candidate models based on an SSE-based objective function) of the different models



## Methods for model selection

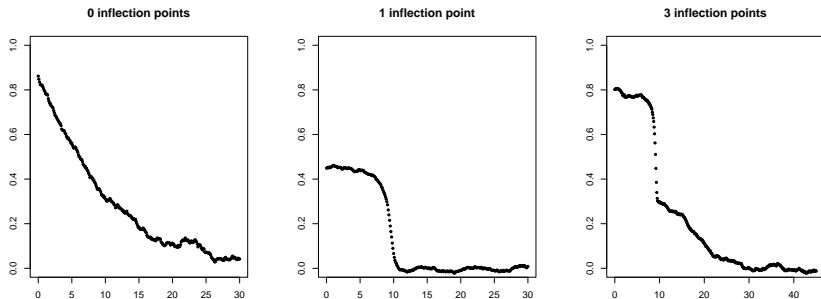
# Methods for model selection

- *A priori* model selection: before parameter estimation
  - Reduces number of parameter estimations necessary = time gain
  - Techniques not easy to determine: ad hoc methods
- *A posteriori* model selection: after parameter estimation
  - General methods available
  - Need parameter estimation for all candidate models = increase in calculation times



# A priori model selection

- Restrict set of model candidates based on properties of data that are independent of parameters.
- Biodegradation example: inflection points.



# A posteriori model selection

- Compose set of candidate models
- Collect experimental dataset(s)
- Perform parameter estimation for all models
- Rank candidate models and select best
- Methods
  - Goodness-of-fit and complexity penalization
  - Evaluation of undermodelling
  - Statistical hypothesis test
  - Residual analysis

# Goodness-of-fit and complexity penalization

Select least complex model that describes data (sufficiently) well.

Balance two terms:

- ① **Goodness of fit**, measured by sum-squared of residuals (SSR)
- ② **Complexity of the model**, as a function of number of parameters.

Many different criteria to make this concrete.

# Akaike Information Criterion (AIC)

Model complexity penalty:  $2p$ , with  $p$  number of parameters:

$$AIC = N \ln \left( \frac{SSR}{N} \right) + 2p.$$

Properties:

- Sometimes preferred when prediction accuracy is important and sample size is small
- Not necessarily consistent (will not select true model even if sample size is large)

# Bayes Information Criterion (BIC)

Model complexity penalty:  $p \ln N$

$$BIC = N \ln \left( \frac{SSR}{N} \right) + p \ln N.$$

Properties:

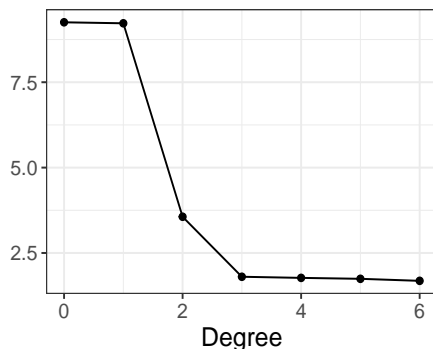
- Will select a simpler model than AIC.
- Consistent (under some conditions)

## AIC/BIC: Polynomial example

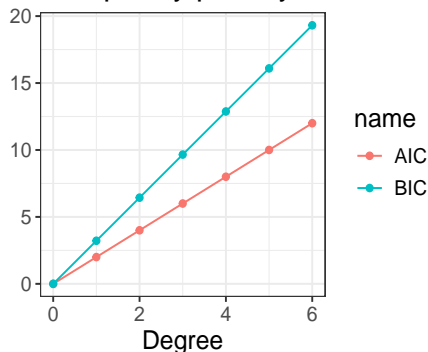
- SSR always decreases when number of parameters increases
- Penalty terms cause goodness-of-fit to increase at a certain point

Example: Select best linear model  $y \sim 1 + x + \dots + x^d$  according to AIC/BIC/... for given data.

Sum of squared residuals

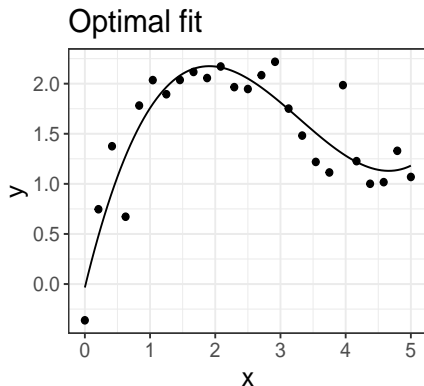
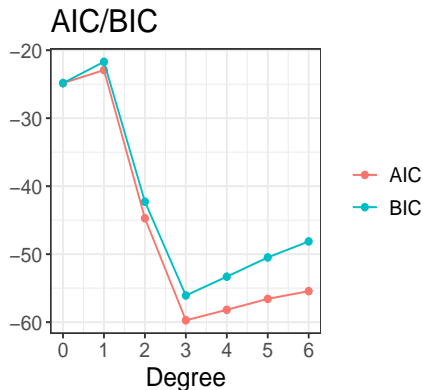


Complexity penalty



## AIC/BIC: Polynomial example

Optimal model provides a **good fit** (SSR low) and is **not too complex** (penalty low).



- Both AIC and BIC select fit of degree 3
- In general AIC and BIC don't have to agree

## AIC/BIC: Biodegradation example

Model	p	SSR	AIC	BIC
Exponential	2	0.36	-303.67	-299.48
Single Monod	3	0.16	-348.74	-342.45
Double Monod	6	0.01	-508.87	-496.30



# Statistical hypothesis test

- Choice between 2 models: simple and more complex
- Is complex model statistically speaking better?
- Verify using F-test:

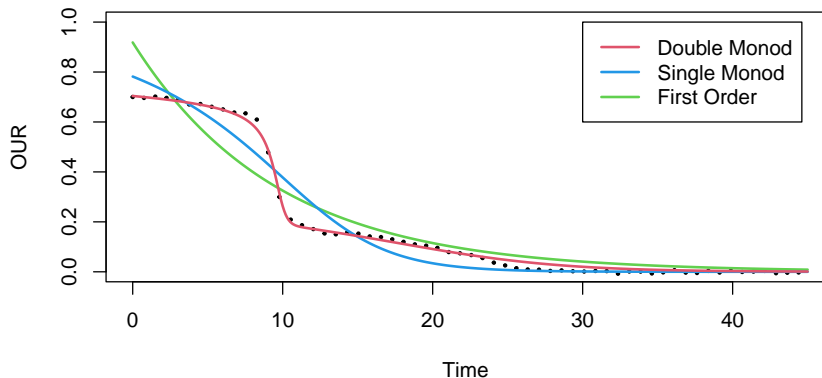
$$F = \frac{\left( \frac{SSR_{simple} - SSR_{complex}}{p_{complex} - p_{simple}} \right)}{\left( \frac{SSR_{complex}}{N - p_{complex}} \right)}$$

- Compare test criterion with tabulated  $F_{1-\alpha, p_{complex}-p_{simple}, N-p_{complex}}$  for significance level  $\alpha$
- If value larger, complex model better (and vice versa)

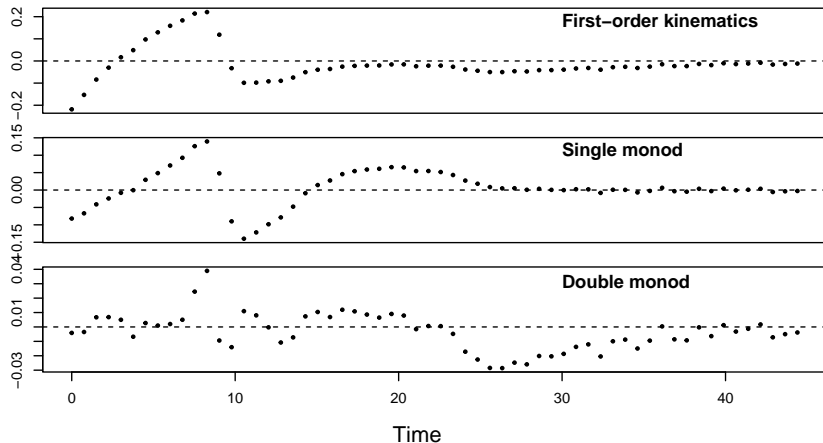
# Residual analysis

- Hypothesis: model is appropriate if properties of residuals are same as properties of measurement errors
- Two popular techniques for evaluation independence of residuals
  - Autocorrelation test (see Parameter Estimation)
  - Runs test (nonparametric test)

## Autocorrelation test: Biodegradation example



# Autocorrelation test: Residuals as a function of time



# Autocorrelation test

- Residuals show some correlation for all three models, indicating that there is some unresolved structure in the data.
- Correlations for double Monod decay much quicker than the other two models.

