Data Driven Modelling
for Complex Systems

Dr Hua-Liang (Leon) Wei

Senior Lecturer in System Identification and Data Analytics

Head of Dynamic Modelling, Data Mining & Decision Making (3DM) Lab

Complex Systems & Signal Processing Research Group

Department of Automatic Control & System Engineering

University of Sheffield

Sheffield, UK

Data Driven Modelling

Similar Terminologies:

- Black-box Modelling
- Data-based Modelling
- Learning from Data
- System Identification

…, etc.

Outline of the Talk

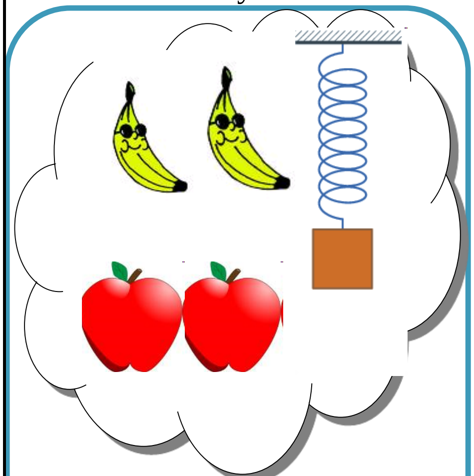
1) A quick selective review of system models
   – Model types and generalised linear models
2) Three classes of models
   – White, grey and black-box models
3) Data-driven modelling
   – Letting data speak and learning from data
4) Examples
   – Using data-driven modelling techniques

PART 1

A Quick Selective Review of System Models

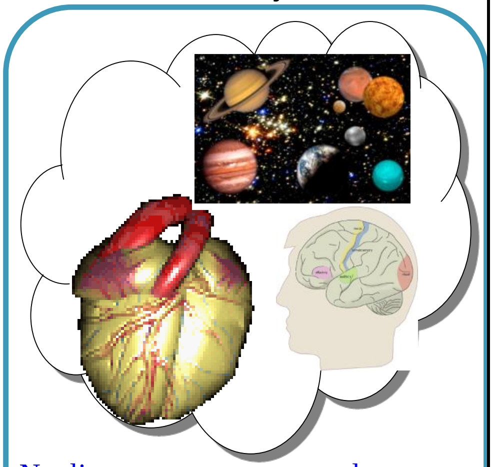
Linear vs Nonlinear

Linear Systems



Linear systems are simple systems whose structure and behavior are regular and easy to understand.

Nonlinear Systems



Nonlinear systems are complex systems whose structure and behavior are normally difficult to analyse.

Linear vs Nonlinear Models

Linear Models

- Continuous-time model
- Continuous-time *state space*
- Continuous-time *transfer function*

- Discrete-time model
- Discrete-time *state space*
- Discrete-time *transfer function* (e.g. Z-transfer function)

Nonlinear Models

- Continuous-time NL model
- Continuous-time NL *state space*
- Continuous-time *transfer function* for a nonlinear system? (*No, not really*!)

- Discrete-time NL model
- Discrete-time NL state space
- Discrete-time transfer function for nonlinear systems? (*No, not really*!)

Parametric & Nonparametric Models

- **Parametric models**
  - ODE's (lumped parameter model)
  - PDE's (distributed parameter model)
  - AR(X) (AutoRegressive with eXogenous inputs)
  - ARMA (AautoregRessive Moving Average)
  - NAR(X) (Nonlinear Autoregressive with eXogenous inputs)
  - NARMA(X) (Nonlinear Autoregressive Moving Average with eXogenous inputs)
- **Non-parametric approaches**
  - Correlation analysis (auto- and cross-correlation)
  - FFT and Spectral analysis
  - Time-frequency analysis
  - Wavelet transform
  - PCA, ICA, …
  - Bayesian, Gaussian, kernel methods, … etc.

Linear-In-The-Parameters vs Linear-In-The-Variables Models

Let *x, z* be independent ('input') variables, *y* be the dependent variable, and *e* the noise signal; also let *a*, *b*, *c* be the model parameters. Then,

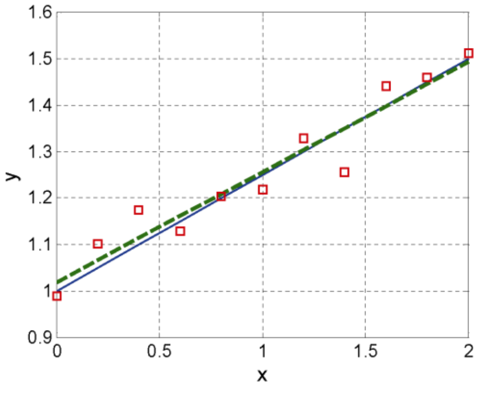
- $y = a + bx$ (linear in both)
- $y = a + bx + cx^2$ (linear-in-the-parameters)
- $y = a + x^b$ (linear-in-the-variables)
- $y = a + \ln(x^b/z^c)$ (linear in neither)

Regression Model (1)
– The Simplest Case

The simplest case of linear regression is used for line fitting

$y = ax + b$ (1)

The unbiased estimates of *a* and *b* can be calculated using given data points (x_k, y_k) (k=1,2, …, N).



Regression Model (2)
– The Multiple Regression

Multiple linear regression model has a form below

$$y = f(x_1, x_2, \cdots, x_n) + e = \sum_{i=0}^{n} a_i x_i + e$$

where x_i is the *i*th 'input' (independent) variable, a_i is the associated model parameter.

- Given observational data points (x_k, y_k) (k=1,2, …, N), the regression model becomes

$$y_k = f(x_{1,k}, x_{2,k}, \cdots, x_{n,k}) + e_k = \sum_{i=0}^{n} a_i x_{i,k} + e_k$$

Regression Model (3)
– The Vector and Matrix Form

$$X = \begin{bmatrix} 1, x_{1,1}, x_{2,1}, \cdots, x_{m,1} \\ 1, x_{1,2}, x_{2,2}, \cdots, x_{m,2} \\ \cdots \quad \cdots \quad \cdots \\ 1, x_{1,N}, x_{2,N}, \cdots, x_{m,N} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix}$$

$$\hat{\mathbf{a}} = \begin{bmatrix} a_0 \\ a_1 \\ \cdots \\ a_m \end{bmatrix} = (X^T X)^{-1}(X^T \mathbf{y})$$

$$\begin{aligned} y_1 &= a_0 + a_1 x_{1,1} + a_2 x_{2,1} + \cdots + a_m x_{m,1} + e_1 \\ y_2 &= a_0 + a_1 x_{1,2} + a_2 x_{2,2} + \cdots + a_m x_{m,2} + e_2 \\ &\cdots \qquad \cdots \qquad \cdots \\ y_N &= a_0 + a_1 x_{1,N} + a_2 x_{2,N} + \cdots + a_m x_{m,N} + e_N \end{aligned}$$

$$\begin{cases} \mathbf{y} = X\mathbf{a} + \mathbf{e} \\ (X^T X)\mathbf{a} \approx X^T \mathbf{y} \\ \hat{\mathbf{a}} = (X^T X)^{-1} X^T \mathbf{y} \end{cases}$$

Normal equation

Least squares estimator

Regression Model (4)
– The General Form

- In general sense, regression model can be expressed in a linear-in-the-parameters form

$$\begin{aligned} y &= a_0\phi_0(x_1,\cdots,x_n) + a_1\phi_1(x_1,\cdots,x_n) + \cdots + a_m\phi_m(x_1,\cdots,x_n) + e \\ &= a_0\phi_0(\mathbf{x}) + a_1\phi_1(\mathbf{x}) + \cdots + a_m\phi_m(\mathbf{x}) + e \\ &= \sum_{i=0}^{m} a_i\phi_i(\mathbf{x}) + e \end{aligned}$$

where $\varphi_i(\mathbf{x})$ are called model regressors or model terms formed by the n model variables through some linear or nonlinear manners.

Regression Model (5)
– The General Form: Examples

- Consider the model

$$y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \varepsilon$$

$\varphi_0(x_1, x_2)=1$ (for β_0), $\varphi_1(x_1, x_2)=\log x_1$ (for β_1), $\varphi_2(x_1, x_2)=\log x_2$ (for β_2)

- Consider a nonlinear IO system described by the model

$$\begin{aligned} y(k) &= a_1 y(k-2) && \leftarrow \varphi_1(\mathrm{x}(k)) \\ &+a_2 y(k-1)u(k-1) && \leftarrow \varphi_2(\mathrm{x}(k)) \\ &+a_3 u^2(k-2) && \leftarrow \varphi_3(\mathrm{x}(k)) \\ &+a_4 y^3(k-1) && \leftarrow \varphi_4(\mathrm{x}(k)) \\ &+a_5 y(k-2)u^2(k-2) && \leftarrow \varphi_5(\mathrm{x}(k)) \\ &+e(k) \end{aligned}$$

$x_1(k)= y(k\text{-}1)$

$x_2(k)= y(k\text{-}2)$

$x_3(k)= u(k\text{-}1)$

$x_4(k)= u(k\text{-}2)$

$\mathrm{x}(k)=[x_1(k), x_2(k), x_3(k), x_4(k)]^T$

Regression Model (6)
– The General Form: Parameter Estimation

- Define the design matrix

$$\Phi = \begin{bmatrix} \phi_0(\mathbf{x}(1)) & \phi_1(\mathbf{x}(1)) & \cdots & \phi_m(\mathbf{x}(1)) \\ \phi_0(\mathbf{x}(2)) & \phi_1(\mathbf{x}(2)) & \cdots & \phi_m(\mathbf{x}(2)) \\ \vdots & \vdots & & \vdots \\ \phi_0(\mathbf{x}(N)) & \phi_1(\mathbf{x}(N)) & \cdots & \phi_m(\mathbf{x}(N)) \end{bmatrix}$$

- Parameters can be estimated by least squares estimator

$$\hat{\theta} = (\Phi^T \Phi)^{-1}(\Phi^T \mathrm{y})$$

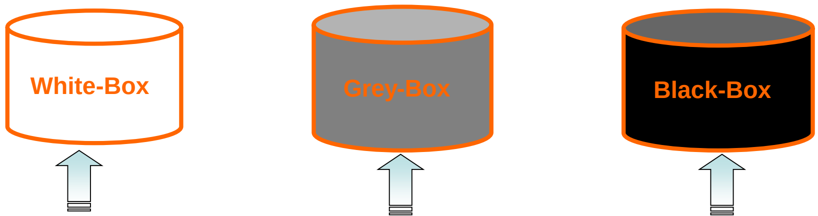
Regression Model (7)
– Least Squares Method Is **BLUE**

- Least squares (LS) is **BLUE**
  - **B**est (Minimum Variance)
  - **L**inear
  - **U**nbiased
  - **E**stimator
- LS provides a unique solution if the design matrix is full rank in column.
- LS is equivalent to maximum-likelihood if noise is Gaussian.

PART 2

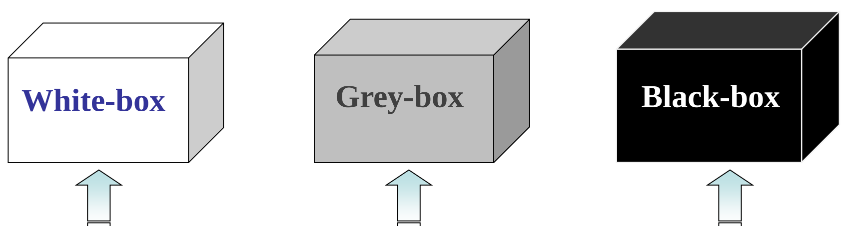
Three Levels of Modelling for Complex Systems

Three Classes of Boxes (Systems): White, Grey and Black-Boxes



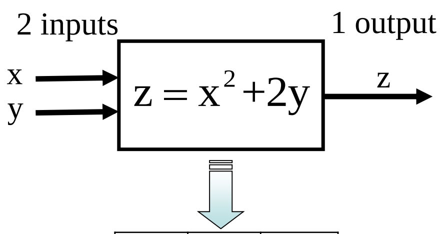
All necessary information is completely known	System information is partly known (the remaining part is unknown)	No a priori information of the system is available

White-, Grey- and Black-Box Modelling Approaches

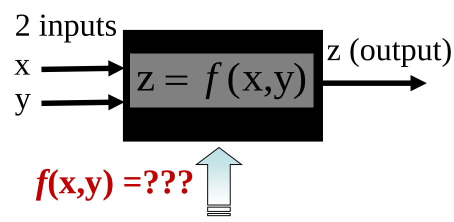


First principles e.g. physical, mechanical, and chemical laws	First principles & data driven modelling techniques	Data driven modelling (let data speak for themselves)

Two Trivial Examples



x	y	z
0	1	2
0	2	4
1	1	3
1	2	5
2	1	6
2	2	8



x	y	z=f(x,y)
0	1	-3
0	2	-6
1	1	-2
1	2	-5
2	1	1
2	2	-2

White-Box Modelling: An Example - Pendulum

Using Newton's second law of motion, along the tangential direction, we have

$$mL\ddot{\theta} = -mg\sin(\theta) - kL\dot{\theta}$$

Taking the state variables as

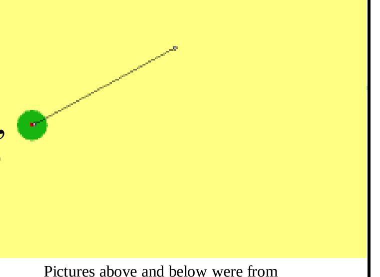
$$x_1 = \theta, x_2 = \dot{\theta}$$

we have the system state model

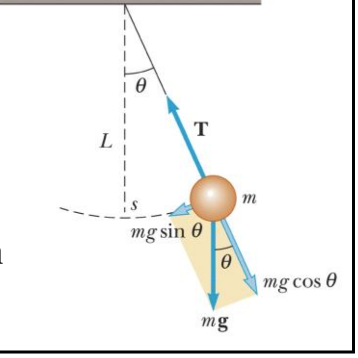
$$\dot{x}_1 = f_1(x_1, x_2) = x_2$$

$$\dot{x}_2 = f_2(x_1, x_2) = -\frac{g}{L}\sin(x_1) - \frac{k}{m}x_2$$

Using this model, we can do simulation and analysis etc.

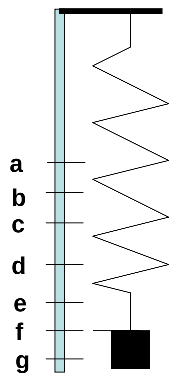


Pictures above and below were from Richard Seto's lecture on Physics 2000.

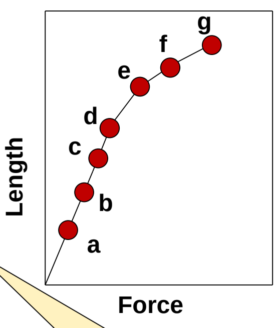


Grey-Box Modelling
A Simple Example - Spring

Experimental data



Record	Force (N)	Length (in)
1, a	1.1	1.5
2, b	1.9	2.1
3, c	3.2	2.5
4, d	4.4	3.3
5, e	5.9	4.1
6, f	7.4	4.6
7, g	9.2	5.0

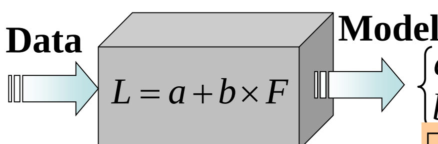


When F = 5, L= ?

Assume that the change in length of the spring is proportional to the force applied (Hooke's law),i.e., Length = $a + b \times$ Force. With the known model structure, we can estimate the unknown model parameters a and b using some standard algorithm.

Grey-Box Modelling
A Simple Example – Parameter Estimation

F	L
1.1	1.5
1.9	2.1
3.2	2.5
4.4	3.3
5.9	4.1
7.4	4.6
9.2	5.0



$$\begin{cases} a = ? \\ b = ? \end{cases}$$

$$\begin{cases} L_1 = a \times 1 + b \times F_1 + e_1 \\ L_2 = a \times 1 + b \times F_2 + e_2 \\ \quad \vdots \qquad\qquad \vdots \\ L_7 = a \times 1 + b \times F_7 + e_7 \end{cases}$$

$$\begin{bmatrix} 1.5 \\ 2.1 \\ 2.5 \\ 3.3 \\ 4.1 \\ 4.6 \\ 5.0 \end{bmatrix} = \begin{bmatrix} 1 & 1.1 \\ 1 & 1.9 \\ 1 & 3.2 \\ 1 & 4.4 \\ 1 & 5.9 \\ 1 & 7.4 \\ 1 & 9.2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{bmatrix}$$

$$L = \mathrm{A} \times \theta + \mathbf{e}$$

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (A^T A)^{-1}(A^T L) = \begin{bmatrix} 1.20 \\ 0.44 \end{bmatrix}$$

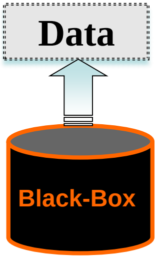
The suggested linear model L=1.2+0.44×F can only approximately represent the data, that is why there is an error term **e** in each of the above equations.

Black-Box Modelling

- **White-Box:** **EVERYTHING** of the system is known (including model structure, parameters, relevant governing laws and rules etc).
- **Black-Box:** **NOTHING** (or little) of the TRUE system model is known or can be known.

What is available are recoded **data** of the system behaviour of interest.

The objective is to learn a model or a set of models from data.



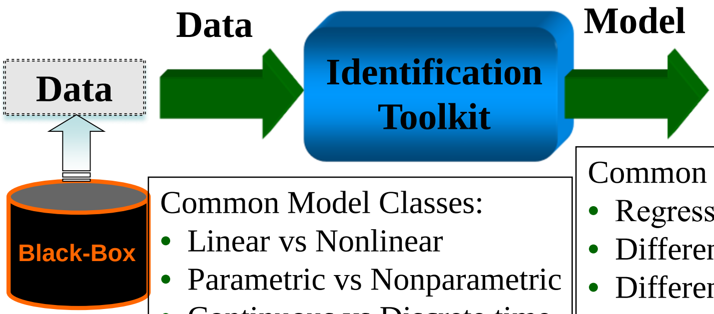
PART 3

Uncovering A Black-Box

Letting Data Speak
and
Learning from Data

Data Driven Modelling

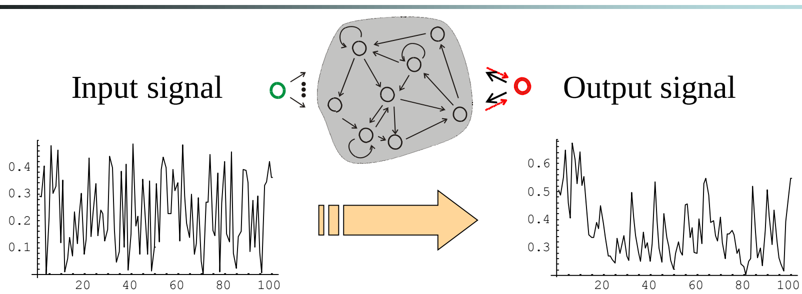
- **Black-Box Modelling:** To model black-box systems, what we can do is to let data speak and learn from data.
- **System Identification:** A science of generating models from data with no (or very limited) a priori knowledge of the inherent system dynamics.



Common Model Classes:
- Linear vs Nonlinear
- Parametric vs Nonparametric
- Continuous vs Discrete time
- Lumped vs Distributed

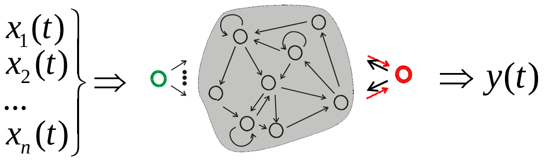
Common Model Types:
- Regression, …
- Differential eqns, …
- Difference eqns, …
- Artificial neural nets, …
- Statistical models, …

Input-Output Models (1)



- **Assumption:**
  a) System input and output signals are measurable.
  b) The true system model structure is **NOT** known.
- **Main Objective and Task:**
  To generate a model that well represents the relationship and reveals the dynamics between the input and output.

Input-Output Models (2)



System identification aims to find a model that represents the relationship between system input and output such that

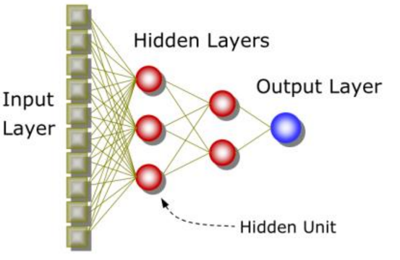
$$y(t) = f(x_1(t), x_2(t), \cdots, x_n(t)) + e(t)$$

$e(t)$ is noise or error that inevitably exists in any data-driven modelling

- $f(\bullet)$ can be any arbitrary function but often is chosen to be those that are easily interpreted, or that have good properties and performance.
- $x_1(t)$, $x_2(t)$, …, $x_n(t)$ are called input variables (also called independent variables, explanatory variables, or simply predictors).
- $y(t)$ is the output variable (also known as dependent variable).

Input-Output Models (3) - ANNs

- **Artificial Neural Networks (ANNs):**
  ANNs are among the most popular approaches to data driven modelling.
- Appropriately trained ANNs can have a very good generalization property (ie prediction capability)



- However, ANN models are *opaque* and cannot be written down, and are difficult to interpret.
- ANN models may not be applicable in many real application scenarios.

I/O Models (4) – Linear Regression

- Linear regression models are of the form

$$y(t) = f(x_1(t), x_2(t), \cdots, x_n(t)) + e(t)$$
$$= a_0 + a_1 x_1(t) + a_2 x_2(t) + \ldots + a_n x_n(t) + e(t)$$

- Note that in many data-based modelling tasks, the parameters $a_0, a_1, a_2, \ldots, a_n$ are unknown and need to be estimated from observed data.
- The estimates of these parameters can often be obtained by means of a least squares (LS) method.

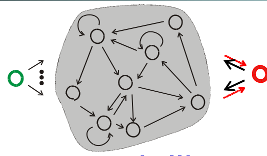
I/O Models (5): Generalized Linear Regression

- Consider a simple case with only 3 explanatory variables: $x_1(t)$, $x_2(t)$, $x_3(t)$. A generalized linear model for such a case can be chosen as:

$$y = f(x_1, x_2, x_3) + e$$
$$= a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$$ ← All linear model terms
$$+ a_4 x_1 x_1 + a_5 x_1 x_2 + a_6 x_1 x_3$$
$$+ a_7 x_2 x_2 + a_8 x_2 x_3 + a_9 x_3 x_3$$ ← All cross product model terms of degree 2
$$+ a_{10} x_1 x_1 x_1 + a_{11} x_1 x_1 x_2 + \ldots$$ ← All cross product model terms of degree 3
$$+ \ldots$$
$$+ \ldots$$ ← All model terms of higher nonlinear degrees
$$+ e$$ ← e is modelling error

I/O Models (4) – Linear Difference Eqns

$u(t)$ Input signal



$y(t)$ Output signal

- In dynamic system modelling, the explanatory variables $x_1(t), x_2(t), \ldots, x_n(t)$ are defined as the lagged system input and output variables; in this case,

$$
\begin{aligned}
y(t) &= f[x_1(t), x_2(t), \cdots, x_n(t)] + e(t) \\
&= a_0 + a_1 y(t-1) + a_2 y(t-2) + \cdots + a_p y(t-p) \\
&\quad + b_1 u(t-1) + b_2 u(t-2) + \cdots + b_q u(t-q) \\
&\quad + e(t)
\end{aligned}
\qquad
\begin{cases}
x_1(t) = y(t-1) \\
\cdots \\
x_p(t) = y(t-p) \\
x_{p+1}(t) = u(t-1) \\
\cdots \\
x_{p+q}(t) = u(t-q)
\end{cases}
$$

- This is usually referred to as an AutoRegressive with eXogenous (ARX) model, where p and q are called model orders.

I/O Models (5) – Nonlinear Difference Eqns

- Consider a simple nonlinear difference equation:

$$y(t) = f[y(t-1), y(t-2), u(t-1)] + e(t)$$

- We can use polynomials to approximate the nonlinear function $f[\bullet]$ as below:

$$
\begin{aligned}
y(t) &= a_0 + a_1 y(t-1) + a_2 y(t-2) + a_3 u(t-1) \\
&\quad + a_4 y(t-1) y(t-1) + a_5 y(t-1) y(t-2) + a_6 y(t-1) u(t-1) \\
&\quad + a_7 y(t-2) y(t-2) + a_8 y(t-2) u(t-1) + a_9 u(t-1) u(t-1) \\
&\quad + \ldots \\
&\quad + \ldots \\
&\quad + e(t)
\end{aligned}
$$

All linear model terms

All cross product model terms of degree 2

All model terms of higher nonlinear degrees

- This is usually referred to as a Nonlinear AutoRegressive with eXogenous input model (NARX).

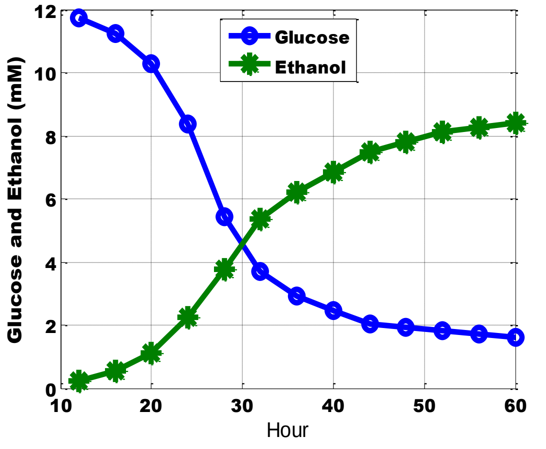
PART 4

Application Examples
of
Data-Driven Modelling

Example 1: Bioethanol Production (1)

Experimental Data

Time (hr)	Glucose (mM)	Ethanol (mM)
12	11.71731	0.212928
16	11.2481	0.558031
20	10.26968	1.132366
24	8.363881	2.251393
28	5.422382	3.756482
32	3.691161	5.358566
36	2.910126	6.206967
40	2.471344	6.847968
44	2.032576	7.485361
48	1.930023	7.812776
52	1.827471	8.140191
56	1.716091	8.273835
60	1.604711	8.407479

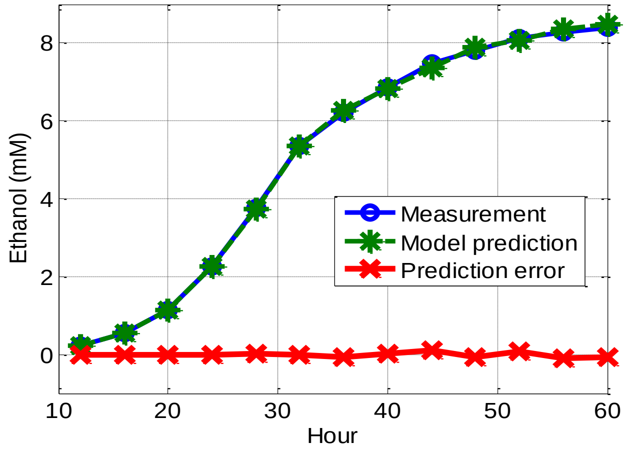


We have applied the NARX modelling approach to the biomedical data, aiming to reveal the relationship between ethanol product and glucose, without using any a priori.

Example 1: Bioethanol Production (2)

The identified NARX Model is

$$\text{Ethanol}(t) = 0.951356 \times \text{Ethanol}(t-1) + 0.403146 \times \text{Glucose}(t-2) - 0.041709 \times \text{Glucose}(t) \times \text{Glucose}(t-3) + e(t)$$

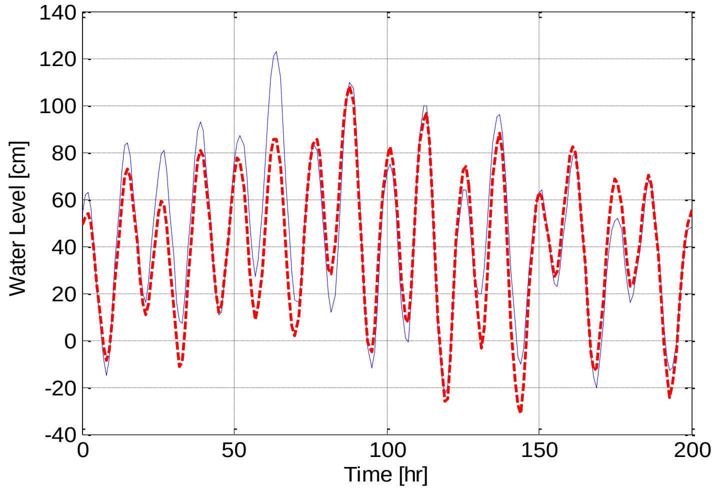


This is a simple model that can perfectly link the system output (ethonal) to the input (glucose).

Example 2: **High Tide Forecast at the Venice Lagoon (1)**

- A multiscale Cardinal B-spline NARMAX model was employed
- For a case study, the hourly water level data for year 1992 were used for model estimation
- The identified model was used to predict water levels of year 1993

Example 3: (cont.)
High Tide Forecast at the Venice Lagoon (2)



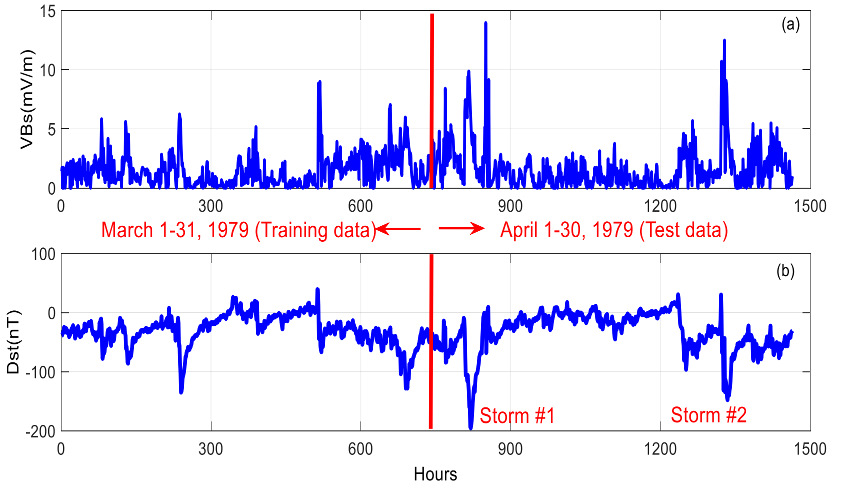
- 24 hours ahead prediction of water level at the Venice Lagoon in 1993
- Thin solid line: measurements; thick dashed: model prediction

Example 3: Dst Index Prediction (1)

- Output signal
  $y(t) = Dst(t)$ (disturbance storm time, [nT]) index
- Input variable
  $u(t) = VBs(t)$ (solar wind rectified electric field [mv/m])
- Training data
  Hourly Dst index and VBs data, March 1-31, 1979
  No. of samples = 744
- Test data
  Hourly Dst index and VBs data, April 1-30, 1979
  No. of samples = 720
- Identified model

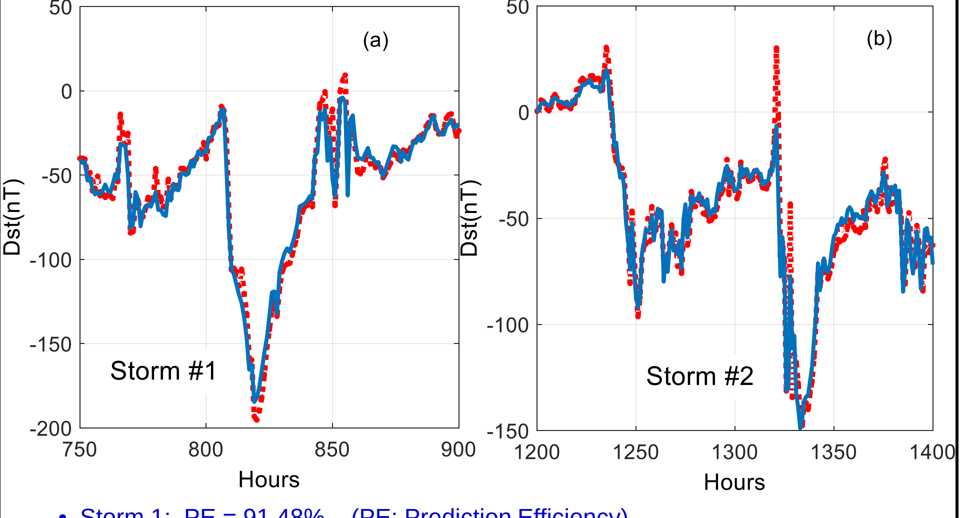
$$y(t) = 0.02486 + 0.98368y(t-1) - 0.92130y^3(t-1)u(t-1)$$
$$+ 0.51936y(t-1)y^2(t-3)u(t-2) - 1.25977y(t-1)u^2(t-1)u(t-2)$$

Example 4: Dst Index Prediction (2)



- Training data: Hourly Dst index and VBs data, March 1-31, 1979 (~ 744)
- Test data: Hourly Dst index and VBs data, April 1-30, 1979 (~ 720)

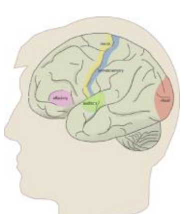
Example 4: Dst Index Prediction (3)

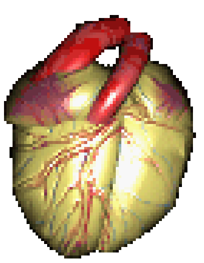


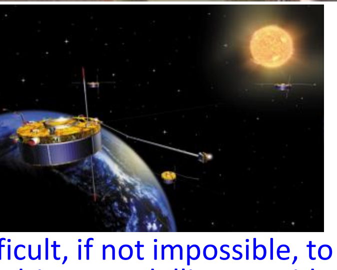
- Storm 1: PE = 91.48% (PE: Prediction Efficiency)
- Storm 2: PE = 92.17%
- All test data: PE = 93.34%

The Understanding of Complex Systems Needs Data-Driven Modelling

System Identification Has Many Applications for the Analysis of Complex Systems







In practice, theoretical models are very difficult, if not impossible, to obtain (as they need first principles) . Data-driven modelling provides a complementary but powerful tool for understanding complex systems.

Tips for Data Driven Modelling
and
Concluding Remarks

Tips for Data Driven Modelling

- Always try the simplest possible models first (e.g. linear).
- If a simpler model works, then forget complex models.
- Keep in mind the main purpose of your modelling task.
- Transparent, parsimonious and easily interpretable models (e.g. regression models) are desirable if you are aiming to reveal dependency and interaction relationships between different explanatory variables/factors.
- If the modelling task is merely focused on prediction or classification, then either parametric models (e.g. linear regressons) or complicated opaque models (such as ANN models) can be an option.

Concluding Remarks

- System identification and data driven modelling, as powerful state-of-the-art approaches, have been widely applied to various areas of science and engineering.
- No particular modelling methods are always the '**best**' and/or '**universal**' for all applications, and therefore it is not possible or appropriate to claim that one approach is always better than all the others for solving all problems.

Key References

1. Wei, H.L., Billings, S.A., & Liu, J. (2004) 'Term and variable selection for nonlinear system identification', *International Journal of Control*, 77, 86-110.
2. Wei, H.L., Billings, S.A., & Balikhin, M.A. (2004) 'Prediction of the Dst index using multiresolution wavelet models' *Journal of Geophysical Research*, 109, A07212.
3. Wei, H.L., & Billings, S.A. (2006) 'Long term prediction of nonlinear time series using multiresolution models', *International Journal of Control*, 79, 569-580.
4. Wei, H.L., & Billings, S.A. (2006) 'An efficient nonlinear cardinal B-spline model for high tide forecasts at the Venice Lagoon', *Nonlinear Processes in Geophysics*,13,577-584.
5. Wei, H.L., Billings, S.A., & Balikhin, M.A.(2006) 'Wavelet based nonparametric NARX models for nonlinear input-output system identification', *International Journal of Systems Science*, 37(15), 1089-1096.
6. Wei, H.L, Zhu, D., Billings, S.A., & Balikhin, M.A. (2007) 'Forecasting the geomagnetic activity of the Dst index using multiscale radial basis function networks', *Advances in Space Research*, 40, 1863-1870.
7. Billings, S.A., Wei, H.L., & Balikhin, M.A. (2007) 'Generalised multiscale radial basis function networks', *Neural Networks*, 20(10), 1081-1094.
8. Billings, S.A., & Wei, H.L.(2008) 'An adaptive search algorithm for model subset selection and nonlinear system identification', *International Journal of Control*, 81,714-724.
9. Wei, H.L., & Billings, S.A. (2008) 'Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information', *International Journal of Modelling, Identification and Control*, 3, 341-356.
10. Billings, S.A. (2013) *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*: John Wiley & Sons.
11. Ayala Solares, J. R., Wei, H.L., Boynton, R. J., Walker, S. N., & Billings, S. A. (2016) 'Modelling and Prediction of Global Magnetic Disturbance in Near-Earth Space: a Case Study for Kp Index using NARX Models', *Space Weather*, 14(10), 899-916.

Acknowledgement

We gratefully acknowledge that part of this work was supported by:

- EC Horizon 2020 Research and Innovation Action Framework Programme (Grant No 637302 and grant title "PROGRESS").
- Engineering and Physical Sciences Research Council (EPSRC) (Grant No EP/I011056/1)
- EPSRC Platform Grant (Grant No EP/H00453X/1)

European Commission

EPSRC Engineering and Physical Sciences Research Council

The University Of Sheffield.

Thank You !
Any Questions?