Modelling using Stochastic Differential Equations

Henrik Madsen

LVI Conference of Lithuanian Mathematicians Society

hmad@dtu.dk, www.henrikmadsen.org
Outline

1 Introduction
2 Why Stochastic Differential Equations?
   • Example: Oxygen Concentration in a Lake
3 Model Building
   • The SDE based State Space Model
   • Identification, Estimation and Model Validation
   • Software
     • Example: Heat dynamics of a building
     • Example: Towards an artificial pancreas
4 More Advanced Principles
   • Hierarchical Modelling using SDEs
     • Example: Diabetes – Estimation of insulin secretion rates
   • Spatio-temporal Modeling
     • Example: Geolocation of fish
5 Summary
Various methods of advanced modelling are needed for an increasing number of complex technical, physical, chemical and biological systems.

For a model to describe the future evolution of the system, it must

1. capture the inherently non-linear behavior of the system.
2. provide means to accommodate for noise due to approximations and measurement errors.

Calls for methods that are capable of bridging the gap between physical and statistical modelling.
Introduction to SDEs

- Often systems are described by either
  1. a set of Ordinary Differential Equations (ODEs) - or
  2. a set of Stochastic Differential Equations (SDEs)

- Solutions to ODEs are deterministic functions (of time). Solutions to SDEs are stochastic processes.
Problem Scenario

Ordinary differential equation

\[ dA = -KA \, dt \]

\[ Y = A + \epsilon \]
Autocorrelated residuals!!
Problem Scenario

\[ \text{Stochastic differential equation} \]

\[ dA = -KA \, dt + \, dw \]

\[ Y = A + e \]
ODE vs SDE

- Uncorrelated residuals
- System noise
- Measurement noise
ODE vs SDE

- Uncorrelated residuals
- System noise
- Measurement noise
Grey box modelling of oxygen concentration
- A sketch of the physical system
Grey box modelling of oxygen concentration
- A white box model

Model found in the literature:

\[
\frac{dC}{dt} = \frac{K}{h\sqrt{h}} (C_m(T) - C) + P(I) - R(T)
\]

\[
P(I) = P_mE_0 \frac{I}{P_m + E_0 I} (= \beta I)
\]

\[
R(T) = R_{15} \theta^{T-15} \quad [mg/l]
\]

\[
C_m(T) = 14.54 - 0.39T + 0.01T^2 \quad [mg/l]
\]

- Simple - however, a non-linear model.
- Uncertainty of the predictions are not described (or does not depend on horizon).
Grey box modelling of oxygen concentration
- A black box model

Box-Jenkins (ARX-model):

\[ C_t - \phi C_{t-1} = \omega P_t + \omega_0 + \epsilon_t \]

- Demands equidistant data!
- Physical interpretation of the parameters is lost!
Model validation (first order model)
Grey box model of oxygen concentration

The following nonlinear SDE based state space model has been found:

***The system equation:

\[
\begin{bmatrix}
\frac{dC}{dL}
\end{bmatrix} = \begin{bmatrix}
\frac{K}{h\sqrt{h}} & -K_c \\
K_3 & -K_l
\end{bmatrix} \begin{bmatrix}
C \\
L
\end{bmatrix} dt + \begin{bmatrix}
\beta \\
0
\end{bmatrix} \begin{bmatrix}
\frac{\sqrt{C}K_b}{h} \\
\gamma
\end{bmatrix} \begin{bmatrix}
I \\
P_r
\end{bmatrix} dt \\
+ \begin{bmatrix}
\frac{K}{h\sqrt{h}} C_{m}(T) - R(T) \\
0
\end{bmatrix} dt + \begin{bmatrix}
dW_1(t) \\
dW_2(t)
\end{bmatrix}
\]

***The observation equation:

\[C_r = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} C \\ L \end{bmatrix} + e\]
The grey box modelling concept

- Combines prior physical knowledge with information in data.
- The model is not completely described by physical equations, but equations and the parameters are physically interpretable.
Why use grey box modelling?

- Prior physical knowledge can be used.
- Non-linear and non-stationary models are easily formulated.
- Missing data are easily accommodated.
- It is possible to estimate environmental variables that are not measured.
- Available physical knowledge and statistical modelling tools are combined to identify parameters of a rather complex dynamic system.
- The parameters contain information from the data that can be directly interpreted by the scientists.
- The physical expert and the statistician can collaborate in the model formulation.
The continuous-discrete time stochastic state space model
System equation (set of Itô stochastic differential eqs.)

\[ dX_t = f(X_t, U_t, \theta) \, dt + G(X_t, U_t, \theta) \, dW_t, \quad X_{t_0} = X_0 \]

**Notation**

- \( X_t \in \mathbb{R}^n \) State vector
- \( U_t \in \mathbb{R}^r \) Known input vector
- \( f \) Drift term
- \( G \) diffusion term
- \( W_t \) Wiener process of dimension, \( d \), with incremental covariance \( Q_t \)
- \( \theta \in \Theta \subseteq \mathbb{R}^p \) Unknown parameter vector
Observation equation

The observations are in discrete time, functions of state, input, and parameters, and are subject to noise:

\[ Y_{t_i} = h(X_{t_i}, U_{t_i}, \theta) + e_{t_i} \]

**Notation**

- \( Y_{t_i} \in \mathbb{R}^m \) Observation vector
- \( h \) Observation function
- \( e_{t_i} \in \mathbb{R}^m \) Gaussian white noise with covariance \( \Sigma_{t_i} \)

Observations are available at the time points \( t_i : t_1 < \ldots < t_i < \ldots < t_N \)

\( X_0, W_t, e_{t_i} \) assumed independent for all \( (t, t_i), t \neq t_i \)
Methods for Identification, Estimation and Model Validation

- **Model Identification:**
  - Identification of lag dependencies (number of states)
  - Identification of model structure

- **Parameter Estimation:**
  - (Approx.) Maximum Likelihood Methods
  - Semi-parametric MLE
  - Estimation Functions

- **Model selection:**
  - Likelihood Ratio Tests (if models are nested)
  - Information Criteria (like AIC, BIC, ...)

- **Model Validation:**
  - Test whether the estimated model describes the data.
  - Autocorrelation functions – or Lag Dependent Functions.
  - Other classical methods ...
Continuous Time Stochastic Modelling (CTSM)

- The parameter estimation is performed by using the software CTSM.
- The software has been developed at DTU Compute.
- Download from: www.ctsm.info
- The program returns the uncertainty of the parameter estimates as well.
The estimation procedure (CTSM)

CTSM is based on

- The Extended Kalman Filter
- Approximate likelihood estimation
The estimation procedure (CTSM)

CTSM is based on
- The Extended Kalman Filter
- Approximate likelihood estimation

and provides eg.
- Likelihood testing for nested models
- Calculations of smoothen state $E[X_t|Y_T]$
- Calculations of k-step predictions $E[X_t|Y_{t-k}]$. 
Flexhouse layout
RC-diagram ofte used for illustrating linear models
Model A

\[ A_w \] is the effective window area.
Model Validation for Model A

Autocorrelation function and Periodogram for the residuals.

Model is seen not to be adequate.
Model E

After some steps: (Notice that eg. the electric heating system is included)
Model Validation for Model E.

It is concluded that the model is adequate.
Model for closed loop control
Advanced Topics
Hierarchical/Population Modelling – Introduction

- Data originating from several population members/subjects
- Identical experiments
- Modelling the effect of covariates
- More data - better estimates of parameters and uncertainties.
- Software: Population Stochastic Modelling (PSM) or
- CTSM-R with some extensions - See Juhl et.al. (2015)
- Software download: www.imm.dtu.dk/psm or www.ctsm.info
Advanced Topics
Hierarchical/Population Modelling – Introduction

- Data originating from several population members/subjects
- Identical experiments
- Modelling the effect of covariates
- More data - better estimates of parameters and uncertainties.
- Software: Population Stochastic Modelling (PSM) or
- CTSM-R with some extensions - See Juhl et.al. (2015)
- Software download: www.imm.dtu.dk/psm or
  www.ctsm.info

Nonlinear mixed effects model with SDEs
Population Modelling – Stages

A population model consists of 2 stages

- 1\textsuperscript{st} stage models the process variation within a single population member/subject
- 2\textsuperscript{nd} stage models the variation in parameters between population members/subjects like:

\[
\phi_i = g(\theta, Z_i) \cdot \exp(\eta_i)
\]

\[
\eta_i \in N(0, \Omega)
\]
Population – Parameter estimation

Parameter estimation using likelihood theory

- Single member/subject - likelihood based on product of conditional densities for each time series of length $n_i$ (called $p_1$ below).
- Population likelihood is a combination of the random effects $\eta$ and the single member likelihoods

$$L(\theta | \mathcal{Y}_{Nn_i}) \propto \prod_{i=1}^{N} \int p_1(\mathcal{Y}_{in_i}|\phi_i)p_2(\eta_i|\Omega)d\eta_i$$
Data – 24H study

- 12 type 2 diabetic patients
- Three standardized meals

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>Breakfast</th>
<th>Lunch</th>
<th>Dinner</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8:00</td>
<td>12:00</td>
<td>18:00</td>
</tr>
<tr>
<td>1440</td>
<td></td>
<td></td>
<td>8:00</td>
</tr>
</tbody>
</table>

![Graph showing individual C-peptide profiles over time](image)

![Graph showing individual Insulin profiles over time](image)
ISR estimate by deconvolution

- C-peptide (the observation) is modelled with a 2-compartment model
- ISR modelled as a random walk (the third state in $x$)

$$\begin{align*}
dx &= \begin{bmatrix}
-(k_1 + k_e) & k_2 & 1 \\
k_1 & -k_2 & 0 \\
0 & 0 & 0
\end{bmatrix} x \, dt + \text{diag} \begin{bmatrix}
0 \\
0 \\
\sigma_{ISR}
\end{bmatrix} \, d\omega \\
y &= C_1 + \epsilon
\end{align*}$$
ISR estimate by deconvolution

Smoothed estimate of ISR for individual 1 and 2.

![ ISR estimate by deconvolution diagram ]
Geolocation of Fish

- Goal: Learn the structure of fish’ movements
- GPS systems do not work under water
- ’Data storage tages’ for measuring the pressure (depth under the surface)
- Data gets available at capture of the fish
Hidden Markov Model

The probability \( \phi \) for, that the fish at time \( t \) is in position \( x \), is

(System equation):

\[
\frac{\partial \phi}{\partial t} = -\nabla (u \phi - D \nabla \phi) \quad \text{(In general)}
\]

\[
\frac{\partial \phi}{\partial t} = D \left( \frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_2^2} \right) \quad \text{(Here)}
\]
Hidden Markov Model

The probability $\phi$ for, that the fish at time $t$ is in position $x$, is

(System equation):

$$\frac{\partial \phi}{\partial t} = -\nabla (u\phi - D\nabla \phi) \text{ (In general)}$$

$$\frac{\partial \phi}{\partial t} = D \left( \frac{\partial^2 \phi}{\partial x_1^2} + \frac{\partial^2 \phi}{\partial x_2^2} \right) \text{ (Here)}$$

The observations are

(Observ. equation):

$$Y_k = \text{Measured pressure/dept at time point } t_k$$
Further physical information

- Bathymetry (depths)
- Time and place for release and capture
- Information about the tide system – see the graph
Observations

Measured sequence of depths from release to capture:
Observations

Measured sequence of depths from release to capture:

Where has the fish been?
Geolocation of Fish
Summary

By using stochastic differential equations for modelling

- physical/prior knowledge and information in data are combined, i.e. we are able to bridge the gap between physical and statistical modeling.
- we obtain a description of the persistence in the differences between observations and model predictions.
- we obtain a rich framework for model identification, estimation, validation and structure modification.
- we obtain probabilistic forecasts and simulations.
Thanks to:

- Jan Kloppenborg Møller
- Lasse Engbo Christiansen
- Peder Bacher
- Anne Katrine Duun-Henriksen
- Rune Juhl
- Niels Rode Kristensen
- Martin Wæver Pedersen
- Henrik Melgaard
Some References - Generic