Is System Identification Just Machine Learning?

Keith Worden
Dynamics Research Group
Department of Mechanical Engineering
University of Sheffield

Uncertainty

- Engineering dynamics has largely assumed throughout its history that *deterministic* models are appropriate for system modelling and prediction.
- Recent (and not so recent) developments suggest otherwise.
- For example, the modelling of biomechanical systems faces the problem that the mechanical properties of tissue vary considerably from individual to individual and even within a single individual.
- Because of uncertainty, probabilistic reasoning is becoming much more common in the analysis of dynamical problems.
- Many of the lessons learned recently have come from *machine learning*.
Met Office got it wrong over ban on flights

By Caroline Gammell
David Millward
and Bruno Waterfield

THE Met Office was blamed last night for triggering the 
“unnecessary” six-day closure of British airspace that has 
cost airlines, passengers and the economy more than 
£1.5 billion.

The government agency was accused of using a scientific model based on “probability” rather than fact to forecast the spread of the volcanic ash cloud, despite thousands of spaces available on cross-Channel ferries and the Eurostar. Two carriers, HMS Ark Royal and HMS Ocean, and the landing ship HMS Albion are being sent to rescue passengers from £100,000

£1m flotilla Warships used as ferries

The Navy ships can carry only 4,000 passengers between them and run only two or three return journeys per day. HMS Ark Royal will be on ferry duties returning from Afghanistan.

Admiral Boyce, the former chief of the defence staff, have relied on a single source of scientific evidence before imposing the widespread ban. He suggested that the no-fly zone should have been restricted to a 20 to 30-mile limit around the volcano. “The science behind the model we are running at the moment is based on certain assumptions where we do not have clear scientific evidence,” he said.

“We don’t even know what density the cloud should be in order to affect jet engines. We have a model that runs on mathematical projections. It is probability rather than actual things happening.”

Mr Fluote said the commission had to intervene to allow airlines to make test flights in order to check the VAAC data.
Worden’s First Law of Uncertainty Management

Whenever possible, work with *facts*
Uncertainty II

- In some areas, uncertainty has been (at least partially), accommodated in theory and practice for a long time.
- *System identification* is a good example.
- To identify a parametric model from measured data, one has to allow for the fact that noise may be present in any measurements, in order that the identified parameters for the model are meaningful.
- In general, the inclusion of *noise models* in linear and nonlinear approaches has often been considered sufficient.
- The main objective of noise models has been to ensure that there is no systematic bias in estimated parameters.
Probabilistic Analysis

- Probabilistic reasoning that now underlies system identification (SI) in structural dynamics, is often *hidden*.
- Many least-squares estimators used for SI are maximum-likelihood estimators under given assumptions. SI user will often implement algorithms in linear algebra and treat the resulting crisp parameters as constituting 'the model'.
- Even if covariance matrix is found, usually only used to provide confidence intervals or 'error bars' on the parameters.
- Predictions will still be made using the crisp parameters.
- Such approaches are powerful, but do not fully accommodate the fact that the data may be consistent with a number of different parametric models.
A more robust approach to parameter estimation, and also model selection, can be formulated on the basis of Bayesian principles.

Among the potential advantages offered by a Bayesian formulation are:

- The estimation procedure will return parameter distributions rather than parameters.
- Predictions can be made by integrating over all possible models consistent with the data weighted by their probabilities.
- Evidence for a given model structure can be computed, leading to a principled means of model selection.
White/Black Box Models

- Useful to divide predictive models into two classes: white and black-box models.
- A *white-box* model is one where the equations of motion have been derived from the underlying physics of the problem and the model parameters have direct physical meanings. E.g. finite element models.
- A *black-box* model is formed by taking a class of models with some universal approximation property and learning the parameters from data; in such a model, like a neural network, the parameters will not generally be physical.
- SI or learning from data, is essential to a black-box approach; for the white-box model, parameters may come from data or from physical laws.
Bayesian Inference for White/Black Box Models

- Recent developments in SI and machine learning give Bayesian approaches for estimation of parameters in white and black-box models.
- Methods for black-box models arguably emerged first e.g. Bayesian learning algorithms for Multi-Layer Perceptron (MLP) neural networks.
- Not suggesting here that Bayesian methods are new to structural dynamics - consider 20 years of work by Jim Beck and colleagues; argument is that they have not been fully exploited. Bayesian view offers advantages mentioned earlier.
- Recently, Bayesian ID methods for differential equation models have emerged in the context of systems biology (Girolami etc.). Work is also concentrated on nonlinear state-space models.
System Identification

- Problem of SI is simply stated: given measured data from a structure, how does one infer the equations of motion which 'generated' the data.
- Although the problem can be stated simply, it is not at all easy to solve.
- Inverse problem of the second kind and can be extremely ill-posed even if the underlying equations are assumed to be linear in the parameters of interest.
- 'Solution' may not be unique.
- If equations of motion are not linear in the parameters, difficulties multiply.
SI and Uncertainty

- This issue is there because measurements or data from a system will almost always be contaminated by random noise.
- Assume data $D = \{(x_i, y_i), i = 1, \ldots, N\}$ of sampled inputs $x_i$ and outputs $y_i$.
- If there is no noise, then identification algorithm, will give deterministic estimate of system parameters $\underline{w}$,

$$\underline{w} = id(D)$$

where the function $id$ represents the algorithm acting on the data $D$.
- If noise $\epsilon(t)$ is present, $\underline{w}$ will become a random variable conditioned on $D$. In this context one no longer wishes an estimate of $\underline{w}$, but to specify ones belief in its value.
SI and Uncertainty II

- Noise is assumed Gaussian with (unknown) variance $\sigma$ ($\sigma$ will be subsumed into $w$, since it is to be inferred).
- In probabilistic terms, instead of deterministic $id$, one now has,

$$w \sim p(w|D, M)$$

where $M$ represents the choice of model.
- Question of bias is interesting in Bayesian context.
- In the presence of noise, the most we can learn from any data is the probability density function of the parameters; in fact, in probabilistic context, *this is everything.*
Predictive Models

- Usual objective of SI is to provide predictive model i.e. a mathematical model which can estimate or predict system outputs if a different system input is given.

- In the probabilistic context, best one can do is find a predictive distribution.

- Suppose a new input sequence $\mathbf{x}^*$ is applied, one wishes to find the density for the predicted outputs,

$$\underline{y}^* \sim p(\underline{y}^* | \mathbf{x}^*, \mathbf{w}, D, M)$$

- Mean of this distribution would give 'best' estimates for predictions; covariance allows one to establish confidence intervals.

- Note the presence of $\mathbf{w}$. In practice, one might use the mean or the mode of the parameter distribution - a point estimate.
Bayesian Predictive Models

- Bayesian prediction requires one to marginalise over parameter estimates, i.e.

\[ p(y^*|x^*, D, \mathcal{M}) = \int p(y^*|x^*, w, \mathcal{M}) p(w|D, \mathcal{M}) \, dw \]

- This is a very powerful idea: allowing for a fixed model structure, one makes predictions using an entire set of parameters consistent with the training data; each point in parameter space weighted according to likelihood given data.

- In practice, there are problems in implementing full Bayesian approach.
Model Evidence I

- Bayesian approach can give evidence for competing model forms.
- Suppose true model structure is one of a finite number \( \{M_i, i = 1, \ldots, M\} \)
- Can imagine computing the probability of observing the data \( P(D|M_i) \) and selecting the model with highest probability.
- Even more in Bayesian spirit, one could marginalise over all possible model structures e.g. for prediction,

\[
p(y^*|x^*, D) = \sum_{i=1}^{M} p(y^*|x^*, M_k, D) P(M_k|D)
\]

- Posterior over models \( P(M_i|D) \) is difficult to compute.
Model Evidence II

- If one appeals to Bayes theorem in the form,

\[
P(M_i|D) = \frac{p(D|M_i)P(M_i)}{p(D)}
\]

and assumes equal priors on models, one arrives at the Bayes factor,

\[
B_{ij} = \frac{P(M_i|D)}{P(M_j|D)} = \frac{p(D|M_i)}{p(D|M_j)}
\]

which weights the evidence for two models in terms of marginal likelihoods of the data given the models.

- Unfortunately, marginal likelihoods are intractable integrals.
So, is SI Just Machine Learning?

▶ It looks like it is, so far.
▶ Problems we have raised relate to difficulties in numerical calculations; are all the ideas we need in place?
▶ I’m going to argue no, and the first argument will be based on going back to the idea of *uncertainty*.
▶ We need to address the issue that there are two main types of uncertainty.
Types of Uncertainty

It is useful to distinguish between two types of uncertainty:

**Aleatory Uncertainty**: is essentially *randomness*. Examples are measurement noise superimposed on data or the behaviour of truly stochastic systems (i.e. Brownian motion). This is uncertainty which cannot be removed – *irreducible* uncertainty – and is what we have talked about up until now; machine learning methods are good at accommodating it.

**Epistemic Uncertainty**: is essentially *ignorance*. It commonly arises because all of the underlying causes (physics) of a problem are not known. This type of uncertainty can be removed by designing experiments to learn the missing physics - it is *reducible*. I would argue that machine learning is not good at dealing with this.
Grey Box Models

- Dealing with Epistemic Uncertainty leads us naturally to the idea of a _Grey Box_ model.
- A grey box model is one for which some of the underlying physics is specified i.e. it has a white-box component; we can attempt to reduce any model error then by adding a nonparametric component and learning its behaviour from data.
- This leads to the idea of two types of grey-box models:
  - We will say that a grey-box model is of **type A** if the nonparametric component is a true black-box model.
  - We will say that a grey-box model is of **type B** if the nonparametric component is motivated by physics rather than the possession of a universal approximation property.
- Type B models are arguably the result of physics and creativity – we will look at two examples.
Friction Models

- Friction is *dynamically* the resistance to motion produced by interfacial contacts between two bodies in relative motion.
- The phenomenon has a microstructural origin, and is the subject of the discipline of *tribology*.
- The most simplistic physical representation is via the *Coulomb* model; in the context of an SDOF oscillator, one has,

\[
m\ddot{y} + F(\dot{y}) + ky = x(t)
\]

\[
F(\dot{y}) = F_c\text{sgn}(\dot{y})
\]

- The model is very limited, but is very convenient for SI.
The Dahl Model

- ... was introduced in 1968 as a means of representing hysteresis loops in dynamic response.
- It has the form,

\[ m\ddot{y} + \sigma_0 z + ky = x(t) \]

\[ \dot{z} = \dot{y} \left( 1 - \text{sgn}(\dot{y}) \frac{\sigma_0 z}{F_c} \right) \left| 1 - \text{sgn}(\dot{y}) \frac{\sigma_0 z}{F_c} \right|^\delta_D \]

where the \( z \) is a state variable interpreted as the elastic deformation of surface asperities of adjacent bodies and \( \sigma_0 \) represents a sort of average asperity stiffness.

- SI problem is harder now, but better representation.
Physically motivated:

Graphical representation of variable state $z$: (a) Dahl model, (b) LuGre model \(^1\).

\(^1\)Piatkowski (T.) 2014 *Mechanism and Machine Theory* **73** pp.91-100.
Better Still: The Lugre Model

- Accounts better for static/dynamic friction

\[
\begin{align*}
m\ddot{y} + \sigma_0 z + \sigma_1 \dot{z} + \sigma_2 \dot{y} + k y &= x(t) \\
\dot{z} &= \dot{y} - \frac{\sigma_0 |\dot{y}|}{s(\dot{y})} z \\
s(\dot{y}) &= F_C + (F_S - F_C) \exp \left\{ - \left( \frac{\dot{y}}{v_s} \right)^{\delta_{vs}} \right\}
\end{align*}
\]

where \( F_S \) and \( F_C \) are the static and Coulomb friction coefficients respectively, \( s(\dot{y}) \) is the St librarian curve and \( v_s \) is the St libraniem velocity.
Distinguishes between static and dynamic friction:

LuGre model predictions.\(^2\)

Can discover Striebeck curve now as part of the learning problem.

\(^2\)Piatkowski (T.) 2014 *Mechanism and Machine Theory* 73 pp.91-100.
Hysteresis Models

- One of the most commonly used hysteresis models is the *Bouc-Wen* model.
- Hysteresis produced by addition of an unmeasured state (like the friction models, but lacking the direct interpretation as a friction force).
- It has the form,

\[
\begin{align*}
    m\ddot{y} + c\dot{y} + ky + z &= x(t) \\
    \dot{z} &= A\dot{y} - \beta|z|z^{n-1} - \gamma|z|^n
\end{align*}
\]
- It allows a versatile representation of a family of hysteresis loops.
Sadly, it isn’t versatile enough.

Consider pinched hysteresis in wooden structures etc.

Illustration of the nailed sheathing connection and pinching hysteresis curve.\

\(^3\)Judd (J.P) 2005 *PhD Thesis, Brigham Young University* Analytical modeling of wood frame shear walls and diaphragms.
Bouc-Wen-Baber-Noori Model

- Includes pinching effect and strength and stiffness degradation.

\[
\dot{z} = \frac{h(z)}{\eta(\epsilon)} \dot{y} \left\{ A(\epsilon) - \nu(\epsilon)[\beta \text{sgn}(\dot{y})|z|^{n-1}z + \gamma|z|^n] \right\}
\]

where \( \eta(\epsilon), \nu(\epsilon) \) and \( h(z) \) are parameters associated with the strength, stiffness and pinching, degradation effects; \( \eta(\epsilon), \nu(\epsilon) \) and \( A(\epsilon) \) are increasing functions of the absorbed hysteretic energy \( \epsilon \),

\[
\eta(\epsilon) = \eta_0 + \delta_\eta \epsilon(t)
\]

\[
\nu(\epsilon) = \nu_0 + \delta_\nu \epsilon(t)
\]

\[
A(\epsilon) = A_0 + \delta_A \epsilon(t)
\]
The pinching function $h(z)$ is specified as,

$$h(z) = 1 - \zeta_1(\epsilon) \exp \left( -\frac{(z \text{sgn}(\dot{y}) - qz_u)^2}{\zeta_2(\epsilon)^2} \right)$$

where,

$$\zeta_1(\epsilon) = (1 - \exp(p\epsilon))\zeta$$

$$\zeta_2(\epsilon) = (\psi_0 + \delta\psi\epsilon)(\lambda + \zeta_1(\epsilon))$$

and $z_u$ is the ultimate value of $z$, specified by,

$$z_u^n = \frac{1}{\nu(\beta + \gamma)}$$
Once we have the model (and some data) machine learning can take over, but getting the model is another matter.

One can argue about whether extraction of a model of this complexity is SI, or whether it is fundamental physics. I’d argue it is SI - it is not intended as an exploration of basic physics, but as a means of providing a predictive model.

Last example of model error/model design comes from Benchmark 3 - the cascaded tanks.
Cascaded Tanks - Model Provided: Normal Operation

When \( x_1(t) < 10, \ x_2(t) < 10, \)

\( \Delta \text{top tank level} = \text{loss through top tank outlet} + \text{gain from pump} \)
\( \Delta \text{bot tank level} = \text{gain from top tank out} + \text{loss through bot tank out} \)

\[
\dot{x}_1(t) = -k_1 \sqrt{x_1(t)} + k_4 u(t) + w_1(t) \quad (1)
\]
\[
\dot{x}_2(t) = k_2 \sqrt{x_1(t)} - k_3 \sqrt{x_2(t)} + w_2(t) \quad (2)
\]
\[
y(t) = x_2(t) + e(t) \quad (3)
\]

Conservation of mass implies \( k_2 = k_1. \)
Model Incorporating Overflow

\[
\dot{x}_1(t) = -k_1 \sqrt{x_1(t)} + k_4 u(t) + w_1(t) \quad (4)
\]

\[
\dot{x}_2(t) = \begin{cases} 
    k_1 \sqrt{x_1(t)} - k_3 \sqrt{x_2(t)} + w_2(t), & \text{no top overflow} \\
    k_1 \sqrt{x_1(t)} - k_3 \sqrt{x_2(t)} + k_5 u(t) + w_3(t), & \text{top overflow}
\end{cases} \quad (5)
\]

- The new parameter \(k_4\) represents the proportional of input which, on average, overflows directly into the lower tank when the top tank is overflowing.
- It should be noted also that if the top tank is full and inflow exceeds outflow, then \(\dot{x}_1(t) = 0\) and similarly for the bottom tank.
- 6 free parameters to identify, \(\{x_1^{(0)}, x_2^{(0)}, k_1, k_3, k_4, k_5\}\).
Genetic algorithm parameter search result

A genetic search was carried out using the simulation model and root mean squared error as the cost function.

- Initial levels: Top 4.55 cm, Bottom 5.26 cm
- $x_1^{(0)} = 4.55$ cm, top tank, initial level
- $x_2^{(0)} = 5.26$ cm, bottom tank, initial level
- $k_1 = 0.049$, flow rate from top tank outlet (enforced $k_1 = k_2$)
- $k_3 = 0.048$, flow rate from bottom tank outlet
- $k_4 = 0.040$, input rate
- $k_5 = 1.00$, proportion of overflow entering bottom tank.

$$\frac{k_4}{k_3} = 0.84$$, therefore, $y = 0.84u^2$ at steady state.
Genetic Algorithm Result - Training

![Graphs showing input, predicted top tank level, and lower tank level with error metrics](image-url)

Keith Worden Dynamics Research Group Department of Mechanical Engineering University of Sheffield

Is System Identification Just Machine Learning?
Genetic Algorithm Result - Validation

- Input
- Predicted top tank level
- Lower tank level

\[ eRMSt = 0.512 \]
Augmented Model: Normal Operation

To account for shape dependent pressure loss, due to flow through the bottom of the tubes, we hypothesised a correctional term proportional to the square of the velocity, which is in turn proportional to the height of the fluid.

When $x_1(t) < 10, \quad x_2(t) < 10,$

$$
\dot{x}_1(t) = -k_1 \sqrt{x_1(t)} + k_6 x_1(t) + k_4 u(t) + w_1(t) \tag{6}$$
$$
\dot{x}_2(t) = k_2 \sqrt{x_1(t)} - k_7 x_1(t) - k_3 \sqrt{x_2(t)} + k_8 x_2(t) + w_2(t) \tag{7}$$
$$
y(t) = x_2(t) + e(t) \tag{8}
$$

Conservation of mass implies $k_2 = k_1$ and $k_7 = k_6$.

Thus there are 8 free parameters to identify,

$$
\{x_1^{(0)}, x_2^{(0)}, k_1, k_3, k_4, k_5, k_6, k_8\}.
$$
GA Search Result - Extended Model

A genetic search was carried out using the extended simulation model and root mean squared error as the cost function.

- $x_1^{(0)} = 6.55$ cm, top tank, initial level
- $x_2^{(0)} = 5.19$ cm, bottom tank, initial level
- $k_1 = 0.046$, flow rate from top tank outlet (enforced $k_1 = k_2$)
- $k_3 = 0.066$, flow rate from bottom tank outlet
- $k_4 = 0.043$, input rate
- $k_5 = 0.84$, proportion of overflow entering bottom tank
- $k_6 = 0.0003$, flow correction from top tank
- $k_8 = 0.0060$, flow correction from bottom tank.
Genetic Algorithm Result, Extended Model - Training
Genetic Algorithm Result, Extended Model - Validation

![Graph showing pump voltage, predicted top tank level, and actual vs. predicted lower tank level over time. The lower graph includes a legend indicating actual lower tank level in red and predicted lower tank level in blue. The eRMSE is shown as 0.297.](image)
Conclusions I

- Bayesian viewpoint on nonlinear SI offers many advantages over point parameter estimation, even when schemes allow estimates of parameter confidences.
- This insight has come from machine learning work, along with very powerful parameter estimation and model structure detection techniques.
- Machine learning isn’t everything though. System identification needs physical insight and expertise in order to overcome problem of model form uncertainty. This is just as true for grey-box models as white-box models.
- Although it isn’t discussed here, the problems of developing an optimal test or data collection strategy is still not completely possible using automated analysis (this will be discussed in the context of our Benchmark 1 results).

Keith Worden  Dynamics Research Group  Department of Mechanical Engineering  University of Sheffield

Is System Identification Just Machine Learning?