Towards a best practice of documenting and publishing system identification algorithms

“Some thoughts on the use of real data and reproducibility”

Thomas Schön
Division of Systems and Control
Department of Information Technology
Uppsala University.

Email: thomas.schon@it.uu.se,
www: user.it.uu.se/~thosc112
The nature of system identification

System identification is the art and science of building mathematical models of dynamic systems from observed input-output data.

System identification is both a theoretical and an experimental science.

It is an engineering subject and it should be very natural to work with real data.

The use of real data in publications should be natural.
The use of real data

A lot of nice initiatives do exist! Benchmark sessions (this afternoon) and software demonstration sessions.

Should strive for making this a natural part in papers, rather than just isolated events.

Running dedicated events along this line is of course a great way to get more data sets available.

Another initiative from the Brussels group (open now)

homepages.vub.ac.be/~mschouke/benchmark2016.html
The use of real data

Scope and Objectives

The objective of this workshop is to advance the current knowledge in nonlinear system identification by encouraging the exchange of ideas and the establishment of formal collaborations between the systems and control, mechanical and machine learning communities.

These three communities have developed over the years various and numerous nonlinear modeling approaches driven by the different backgrounds, constraints and end-uses. Moreover, they generally focus on different aspects of the modeling problem as they face different limiting factors in terms of model quality and identification cost. This is why we believe that, by promoting interaction, significant benefit can be mutually gained.

This workshop will be structured around three benchmark systems featuring state-of-the-art challenges in nonlinear system identification, namely dynamic nonlinearity, process noise, and short data record. They are a Bouc-Wen hysteretic system, a Wiener-Hammerstein system with process noise, and a cascaded tanks setup, respectively.

Solicited contributions should describe solutions to one or several of these benchmark problems. In particular, comparative overviews of methods would be particularly appreciated.

Participant Registration and Deadlines

Researchers wishing to participate in the workshop are invited to signal their interest as soon as possible via email. They will be kept informed of various benchmark developments and updates, in particular the Wiener-Hammerstein measurement opportunities.

Registration for the workshop is to be completed by January 29, 2016. Presenting participants are also requested to submit a 1-page abstract summarizing the main aspects of their contribution by March 25, 2016. Participation involves no registration fee.
**Reproducibility**

**Definition:** Reproducibility is the ability of an entire experiment or study to be duplicated, either by the same researcher or by someone else working independently. Reproducibility is one of the main principles of the scientific method.  
(Source: Wikipedia)

Who is responsible for this? (publishers, or editors, or individual researchers, or research groups, or, ...)

Martin Vetterli at EPFL: lcav.epfl.ch/martin.vetterli
lcav.epfl.ch/reproducible_research
Reproducibility – example of one researcher

Prof. Martin Vetterli

Research Areas
Mathematical Signal Processing
Signal Processing for Communications
Signal Processing Applications
Theory of wavelets and their applications
Communication systems and multimedia
Sensor networks

The work in the laboratory follows the Reproducible Research philosophy, thus all papers, source code and data is made available.
Why produce reproducible research?

1. Other researchers can start from the current state of the art, instead of spending time trying to figure out what was exactly done in a certain paper.
2. It will help you to foster “better coding habits” and to reproduce figures, etc.
3. Makes it easier to compare a new method to existing methods.
4. It increases the impact of our research.
5. ...

Example (Lennart’s toolbox)

By making the algorithms of system identification widely available via his toolbox, a lot of the system identification research has become reproducible.

This has opened huge markets for system identification.
Example of initiatives in neighbouring fields

**Robotics:** The International Journal of Robotics Research (IJRR) has 1) Data papers category and 2) Multimedia extensions (movies, experimental data, code, etc.)

**Machine Learning:** The use of real data is second nature.

**Statistics:** Dedicated journals for publishing software. *Journal of statistical software* (IF: 3.8). Publishes peer-reviewed articles about statistical software, together with the source code.

"Statistical software is the key link between statistical methods and their application in practice."

**Information retrieval:** Reproducible research track at the European conference on information retrieval (ECIR)

ecir2016.dei.unipd.it/call_for_papers.html
What can SYSID do – one thought

One thought to (possible) get the discussion going.

Introduce a “solid paper” category. Besides the standard requirements, also require the following:

1. Show results using real data.
2. Data and code published online.
3. ...

Should SYSID require/encourage data and code to be published for all papers?
A few references to get started

