Chapter 16

Project work: Description

16.1 General Structure

The aim is to acquire hands-on experience with the tools for system identification for working with actual (real-life) data. Specifically, use of the MATLAB SI toolbox, and learning to organize different available tools in a successful application are central. We will work with real data, so there is no 'true' solution you should aim at, but I will score you based on how 'intelligent' you make use of the techniques in order to build up a sensible approach. The emphasis will be on identification of Multiple-Input, Multiple-Output (MIMO) systems. The projects will be organized in small groups (depending on the number of participants to the course), with indication of individual contribution. The work will culminate in a report containing workflow, design decisions, details of the implementation, and results of simulations. It is allowed and encouraged to work in small groups (1-5 persons) on the same task. But it is required that each individual hands in her/his own report (named).

A successful project will consist of four main steps, each of which are to be documented in the report.

1. Visualize the data, point out characterizing properties and state the solution you're after. Be creative in the economic use of time plot, histogram or frequency plots.

2. Do some simple (possibly naive) simulations: e.g. what is the best constant prediction (mean). What is the best we can do using standard techniques implemented in the `ident` tool?

3. Based on experience gathered during the previous phase, what is a proper method for identification of the system? Perform the simulations trying to get the best result possible. Most importantly, verify the result: why is this result satisfactory? How does it compare to the naive estimates obtained in (2)?

4. Summarize your contribution in an 'abstract' and 'conclusions' of your report. Which contributions to standard approaches can be claimed, and how do the models support such clams?

Those different steps (sections) should show up in the report to be handed in. Of course the different steps are intimately related in practice: for example model tuning requires one often to rethink preprocessing issues. Despite this fact, experience learns that such a linear (I-VI) presentation

helps the audience to understand better the merit and contributions of the present case study. That's why I will require the final project manuscripts to follow closely this structure. Those steps are described in some further detail for each case study in the next chapter.

16.2 Groups

I want you to work on the projects in groups of up to (and including) five persons. You can choose your mates yourself, or if more convenient you can work out a project on your own. Once you decided on a group, send me an email with who's involved, and I will file this in in Studentportalen. At the end of the day, I need a report for each of you, and each group (!) is expected to present his findings and answer questions. I expect a contribution relative to the size of the group - so if you are four then you should have worked out extra questions. If you are only one, a basic study can do. The presentations might be shared amongst the different members of the same group. Indicate in the report as well as in the representation to which part who was responsible, I want to make sure that everybody was involved to a sufficient level. Finally, it is no problem for me that different groups work on the same project, and that experiences are interchanged. But do differentiate enough between conclusions, and make sure I rate the projects as sufficiently independent.

16.3 A report

A report stresses how the present research *contributes* to the topic of interest, and will provide empirical or theoretical *evidence* for those claims. A well-manicured report describing the achieved results, motivating the design decisions and verifying the estimated models. Make sure sufficient care is given to

- Avoid Typos.
- Use of the English language: think about what you write and how you write it up.
- Be concise: reread your own text and throw out what is not needed for supporting the conclusions.
- Figures: name axes, and give units. Add a legend explaining the curves we see, and add a caption explaining what we see and what the reader should conclude from the present figure.
- MATLAB code should not be included, rather state the different steps implemented in the form of appropriate and well-defined formulas.
- A guideline would be a report of 3-4 single column, 11pt, letter pages, including 2-4 figures. Again, this depends on the size of the group you are working in.

16.3.1 A report in LaTeX

Research and development consists usually of 20% active research, and about 80% of the time goes to preparing, presenting and interpreting results. The importance of a a good presentation of the work can hardly be overestimated . A large innovation in scientific research came as such with the introduction of the TeX and LaTeX document preparation system in the early 1980s. The tool not

Figure 16.1: example caption

only freed researchers from the many practical concerns when using a typewriter, but also urged them to think proactively about the structure of the text. In fact, it was so well designed that it remains a sine qua non in technical research even now, some 30 years later.

The overall document may look as follows

```
\documentclass[11pt]{report}
% comment!
% preamble: point to all packages to be used, and give local definitions
\usepackage{graphicx}
\usepackage{amssymb}
\usepackage{epstopdf}

% define the document's overall properties
\title{Brief Article}
\author{The Author}
\date{20 November 2010}

% begin the actual document
\begin{document}
\maketitle % construct the title, author enumeration, date according to the given style
\section{My first section}
\subsection{My first subsection}
\subsubsection{My first subsubsection}

% end document
\end{document}
```

The result will be compiled by the command `latex mydoc.tex` into `mydoc.dvi`, which can be in turn translated to `ps` or `pdf` using appropriate tools. This steps are properly automated in TeX editors as `texshop` (mac) or `winedt` (windows). Those software packages come with appropriate templates which might give you good starting point.

The key idea is that every bit of text has to be placed in its proper *environment* (or box), and the LaTeX engine then takes care of how to organize those different environments into an appealing page layout. Environments come in different forms, as sections, theorems, definitions, figures, tables, title, quotations and so forth. A particularly strong feature of LaTeX is the engine to typeset formulas. It is instructive to consider a simple example

```
\begin{equation}
```

```
a_i, b^i, \hat\theta,\phi_{(n)}, \mathbb{R}, \sum_{i=1}^n q^{-i},
\end{equation}
```

which will look after typesetting as

$$a_i, b^i, \hat{\theta}, \phi_{(n)}, \mathbb{R}, \sum_{i=1}^{n} q^{-i}. \tag{16.1}$$

Another example is how to typeset figures. The following example is typeset in Fig. (17.1) using the definition in the preamble of the file, and the packages `graphicx,graphics,epic`.

```
\begin{figure}[htbp]
   \includegraphics[width=4in]{latexmatters.png}
   \caption{example caption}
   \label{fig:example}
\end{figure}
```

Lots of good manuals for how to create LaTeX documents can be found on the web, e.g. at `http://en.wikibooks.org/wiki/LaTeX` or `http://www.maths.tcd.ie/~dwilkins/LaTeXPrimer/`. But the real way to go is to give it a try yourself, and do ask me `kp@it.uu.se` if you are stuck with a question.

16.4 Presentation of the Result

Each group is assigned a slot of 5 minutes per person (that is, if you're in a group of three the presentation should take 15 min) at the end of the term in order to defend their results. A successful presentation should enumerate the claimed contributions, and provide insight to support those claims. Specifically, try to convince the audience of the following bullets:

- What are the conclusions of the effort, and how do you get there?
- How do you improve over earlier/simpler solutions?
- What is the contribution of each of the groupmembers?
- What are possible applications for your work?
- Suppose I were your manager at a company: why should I invest 1.000.000$ to implement your model?
- Suppose I were your teacher: why would I award you a top grade for your work?

After and during the presentation, I will ask some questions for each of you evaluating your insights in SI as used in the project. Here, I wont impose that you use LaTeX, you can use any presentation program you feel comfortable with.

Chapter 17

Project work: Case Studies

17.1 Identification of an Industrial Petrochemical Plant

The main application studies of system identification are found in industrial practice. The data used comes from a glass furnace plant at Philips. The data consists of 3 different measured inputs and 6 consequent output signals. The signals consist each of 1247 samples. The emphasis of this case is on the use of subspace identification and design of a subsequent controller. In order to describe a successful application study, do follow the following steps.

1. (Visualize): A first step is to visualize the data. Specifically build a scatter-plot of all combinations input-output (18 subplots in total). Look at this plot, what properties of the signal become directly apparent?

2. (Preprocess): The next step is to check wether the involved signals need preprocessing. Are means zero? Are statistics as the variance more or less time-invariant? Is there evidence for polynomial or sinusoidal trends? Can the signal reasonably be expected to follow a Gaussian process, or are they sufficiently rich? In this case a `log(Y)` transform (with basis e) would do the trick to convert the positive temperatures with large peaking values to a signal which is more or less behaving as a Gaussian process.

3. (Test): At this early stage reserve a portion of the data for testing the model you come up with at the end of the day. By putting a portion of the data aside at this early stage, you make sure that this data does not influence in any way the model building process, and that the testing of the model is completely objective.

4. (Initial): Try to build a first model using a naive approach. For example, you can convert the problem into a set of SISO estimation problems. This naive model will mainly serve to benchmark your final approach.

5. (Diagnose): Why is the aforementioned naive approach not sufficient? Or perhaps it is? Can you use insights from the naive approach in order to argue for a more involved approach? What subtleties is the naive approach missing altogether? To make this point you might want to use an intelligent plot of results, where you indicate how things go wrong.

6. (Improve): So now the stage is prepared to explain the principal strategy. In the context of this course that would involve a subspace identification strategy. Spend some time (words/slides) on which design decisions you took to get the technique to work properly.

7. (Validate): Firstly, implement a cross-validation strategy to test the identified model. Recall different methods of model selection as were described in the SISO case, and use one to validate the result. Secondly, compare the results with what you get by the naive approach: do we actually see the improvements as argued for in earlier stages?

8. (Use): An identified model is good if it serves its intended purpose well. Therefore, the ultimate verdict on the model is how well it works in practice. In this industrial context, a model is build for constructing an accurate control law. Can you derive a standard control law by pole placement using the identified model as a description of the system which is to be controlled? Why is it satisfactory? Why not?

9. (Extra): It might be clear that the above steps are only scratching the surface of the interesting things which can be done. At this stage I challenge your creativity to describe innovations you can do based on some of the elements seen in the lectures. For this project, an extra step might consists of refining the subsequent control law based on the system using a techniques of MPC. Does this control law work better in practice? To see this, assume that the identified model *equals* the system, and steer the process output to a constant output of $(1.7, 1.7, 1.7, 1.7, 1.7, 1.7)^T$.

10. (Conclude): Perhaps most importantly, what is the conclusion of your efforts thus far. Are results satisfactory, or does the problem setting pose more involved intrinsic questions? Is the identified model serving its purposes well enough? What would be a next relevant step?

17.2 Identification of an Acoustic Impulse Response

The present techniques are closely related to signal processing tools. We borrow here an application study which is commonly treated in an acoustic signal processing context, and show how tools from system identification can be used. The setting is to recover the room dynamics from sending a speech signal through a loudspeaker, after which the signal is convolved with the room dynamics, and the result is then picked up by a microphone. The question in general is how one can come up with a model for the dynamics of the room acoustics. The challenge is to come up with a method to use the derived state-space model in order to clean-up the meaningful signal. Acoustic signals are read and written in `.wav` using the MATLAB commands `wavread` and `wavwrite`. A typical feature is that acoustic model need rather high orders to capture dynamics well.

In order to describe a successful application study, do follow the following steps.

1. (Visualize): A first step is to visualize the data. Specifically build a scatter-plot of all combinations input-output (9 subplots in total). Look at this plot, what properties of the signal become directly apparent?

2. (Preprocess): The next step is to check wether the involved signals need preprocessing. Are means zero? Are statistics as the variance more or less time-invariant? Is there evidence for polynomial or sinusoidal trends? Can the signal reasonably be expected to follow a Gaussian process, or are they sufficiently rich?

3. (Test): At this early stage reserve a portion of the data for testing the model you come up with at the end of the day. By putting a portion of the data aside at this early stage, you make sure that this data does not influence in any way the model building process, and that the testing of the model is completely objective. Since in typical applications of acoustic processing data is abundant and cheap, this does not really pose a problem.

4. (Initial): Try to build a first model using a naive approach. For example, you can convert the problem into a set of SISO estimation problems. This naive model will mainly serve to benchmark your final approach.

5. (Diagnose): Why is the aforementioned naive approach not sufficient? Or perhaps it is? Can you use insights from the naive approach in order to argue for a more involved approach? What subtleties is the naive approach missing altogether? To make this point you might want to use an intelligent plot of results, where you indicate how things go wrong.

6. (Improve): So now the stage is prepared to explain the principal strategy. In the context of this course that would involve a subspace identification strategy. Spend some time (words/slides) on which design decisions you took to get the technique to work properly.

7. (Validate): Firstly, implement a cross-validation strategy to test the identified model. Recall different methods of model selection as were described in the SISO case, and use one to validate the result. Secondly, compare the results with what you get by the naive approach: do we actually see the improvements as argued for in earlier stages?

8. (Use): An identified model is good if it serves its intended purpose well. Therefore, the ultimate verdict on the model is how well it works in practice. The use of the identified model in this context is to 'de-mix' the signals picked up by the microphones in order to reconstruct the original acoustic signal.

9. (Extra): It might be clear that the above steps are only scratching the surface of the interesting things which can be done. At this stage I challenge your creativity to describe innovations you can do based on some of the elements seen in the lectures. For this project, an extra step might consists of using software to collect real signals and to report the analysis. A commonly used software packet to design such experiments, record and process the acoustic signals is the Dirac software (`http://www.bksv.com/ServiceCalibration/Support/Downloads/DIRAC/DIRAC%20Room%20Acoustics%20Software%20Evaluation%20Copy.aspx`).

10. (Conclude): Perhaps most importantly, what is the conclusion of your efforts thus far. Are results satisfactory, or does the problem setting pose more involved intrinsic questions? Is the identified model serving its purposes well enough? What would be a next relevant step?

17.3 Identification of Financial Stock Markets

In this project we will study timeseries taken from historical records of 10 main stock markets. The data comes from day-to-day index values of the stock markets summarized in Table (17.1). The start date is Dec. 1, 1999, the series runs for 1352 trading days, and the end date is in April1, 2004. This dataset contains no explicit input process, and we assume that the driving process captures the socio-economic, political and financial effects. Specifically, this implies that we will

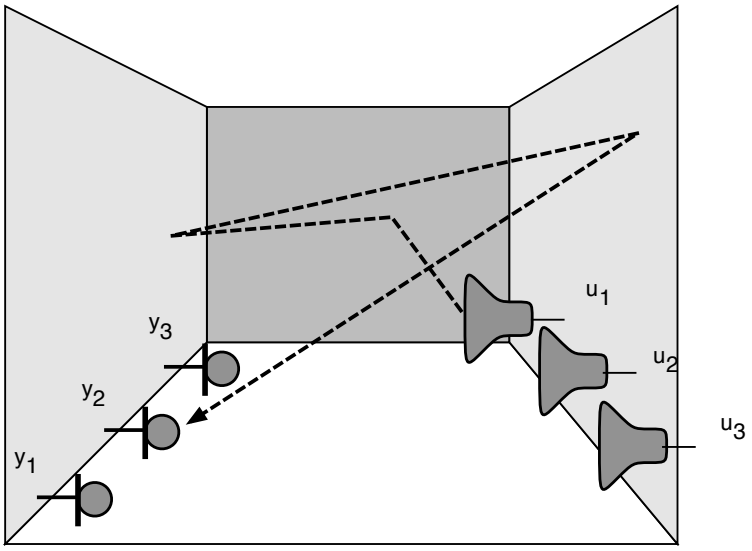

Figure 17.1: Schematic Illustration of the setup of the acoustic experiment. Here three acoustic signals are send in a room by 3 different loudspeakers (right). The room mixes those signals, and the result is picked up at different positions by three micros (left). The question is then which model of the room captures the dynamics of the mixing.

use techniques of stochastic identification to identify a state-space model which captures the present dynamics. The overall question is to see which stocks display strongly related behavior, and how many hidden states actually capture the dynamics. The specific question is to use the identified state-space system to predict the values of the stocks in a next trading day.

1. (Visualize): A first step is to visualize the data. Look at this 10 signals, what properties of the signal become directly apparent?

2. (Preprocess): The next step is to check wether the involved signals need preprocessing. Are means zero? Are statistics as the variance more or less time-invariant? Is there evidence for polynomial or sinusoidal trends? Can the signal reasonably be expected to follow a Gaussian process, or are they sufficiently rich? In this case a typical good way to process the indices into a form which resembles a Gaussian process is to consider the difference of the log of the indices. That is `y = (diff(log(x)))`. This step is essentially relating the approach to techniques as ARIMA and ARCH.

3. (Test): At this early stage reserve a portion of the data for testing the model you come up with at the end of the day. By putting a portion of the data aside at this early stage, you make sure that this data does not influence in any way the model building process, and that the testing of the model is completely objective. As data is rather scarce in this setup, think carefully which part of the data would be good for testing the model before putting it in production.

4. (Initial): Try to build a first model using a naive approach. For example, you can convert

Number	Price Index	Stock Market
1	Dow Jones Industrial Price Index	US
2	S&P 500 Composite Price Index	US
3	Nikkei 300 Price Index	JP
4	DAX 30 Performance Price Index	DE
5	CAC 40 Price Index	FR
6	Swiss Market Price Index	SH
7	Milan MIB 30 Price Index	IT
8	IBEX 35 Price Index	ES
9	BEL 20 Price Index	BE
10	FTSE 100 Price Index	UK

Table 17.1: The price indices of 10 main stock markets recorded at the end of a trading-day.

the problem into a set of SISO estimation problems. This naive model will mainly serve to benchmark your final approach.

5. (Diagnose): Why is the aforementioned naive approach not sufficient? Or perhaps it is? Can you use insights from the naive approach in order to argue for a more involved approach? What subtleties is the naive approach missing altogether? To make this point you might want to use an intelligent plot of results, where you indicate how things go wrong.

6. (Improve): So now the stage is prepared to explain the principal strategy. In the context of this course that would involve a subspace identification strategy. Spend some time (words/slides) on which design decisions you took to get the technique to work properly.

7. (Validate): Firstly, implement a cross-validation strategy to test the identified model. Recall different methods of model selection as were described in the SISO case, and use one to validate the result. Secondly, compare the results with what you get by the naive approach: do we actually see the improvements as argued for in earlier stages?

8. (Use): An identified model is good if it serves its intended purpose well. Therefore, the ultimate verdict on the model is how well it works in practice. The overall goal of such a financial application would be to predict the vale of the stocks at the end of the next trading day. It is typical that only a marginal accuracy can be obtained using such approach, but do keep in mind that even a minor gain can make a huge benefit for an investment company.

9. (Extra): It might be clear that the above steps are only scratching the surface of the interesting things which can be done. At this stage I challenge your creativity to describe innovations you can do based on some of the elements seen in the lectures. In the context of this project, it would be interesting to see how tools of nonlinear modeling can actually improve the predictions.

10. (Conclude): Perhaps most importantly, what is the conclusion of your efforts thus far. Are results satisfactory, or does the problem setting pose more involved intrinsic questions? Is the identified model serving its purposes well enough? What would be a next relevant step?

17.4 Identification of a Multimedia stream

Systems theory and identification is a perspective of looking at reality, and can as such be found anywhere. This project will study the case of a stream of video frames, which consists basically of a collection of timeseries, a single one for each pixel on the screen. While it is intuitively clear that the 'informative' part of this multivariable timeseries cannot be modeled without exploring concepts as 'meaning' and 'understanding', lots of the pixels have a value which can easily be deduced from past behavior. In other words the video can be described in terms of an informative state of a much lower dimension than the number of pixels, and the derived dynamic behavior. The overall challenge is to develop a visual feel for the dynamics which can be described by a state-space system. The specific question is to develop a simple way to 'compress' the stream of frames as a state-space system driven by the given input signals.

1. (Visualize): A first step is to visualize the data. Look at the 5 input signals, and visualize the frames. Can you work out intuitively what the inputs 'mean'? Herefore, use the MATLAB command

   ```
   >> for t=1:n, imshow(reshape(y(t,:),50,50),[0 1]); pause(.1); end
   ```

2. (Preprocess): The next step is to check wether the involved signals need preprocessing. Are means zero? Are statistics as the variance more or less time-invariant? Is there evidence for polynomial or sinusoidal trends? Can the signal reasonably be expected to follow a Gaussian process, or are they sufficiently rich? In this project, the input is indeed a signal taking values $\{0,1\}^n$, and no preprocessing is needed.

3. (Test): At this early stage reserve a portion of the data for testing the model you come up with at the end of the day. By putting a portion of the data aside at this early stage, you make sure that this data does not influence in any way the model building process, and that the testing of the model is completely objective. As data is rather scarce in this setup, think carefully which part of the data would be good for testing the model before putting it in production.

4. (Initial): Try to build a first model using a naive approach. For example, you can convert the problem into a set of SISO estimation problems. This naive model will mainly serve to benchmark your final approach.

5. (Diagnose): Why is the aforementioned naive approach not sufficient? Or perhaps it is? Can you use insights from the naive approach in order to argue for a more involved approach? What subtleties is the naive approach missing altogether? To make this point you might want to use an intelligent plot of results, where you indicate how things go wrong.

6. (Improve): So now the stage is prepared to explain the principal strategy. In the context of this course that would involve a subspace identification strategy. Spend some time (words/slides) on which design decisions you took to get the technique to work properly.

7. (Validate): Firstly, implement a cross-validation strategy to test the identified model. Recall different methods of model selection as were described in the SISO case, and use one to validate the result. Secondly, compare the results with what you get by the naive approach: do we actually see the improvements as argued for in earlier stages?

8. (Use): An identified model is good if it serves its intended purpose well. Therefore, the ultimate verdict on the model is how well it works in practice. Here the use is to 'compress' the outputs using the given inputs and the state-space model. What would be the compression rate you obtain?

9. (Extra): It might be clear that the above steps are only scratching the surface of the interesting things which can be done. At this stage I challenge your creativity to describe innovations you can do based on some of the elements seen in the lectures. In the context of this project an exciting step would be to use this approach to 'compress' real video footage. Can you come up with different data where such approach might work reasonably?

10. (Conclude): Perhaps most importantly, what is the conclusion of your efforts thus far. Are results satisfactory, or does the problem setting pose more involved intrinsic questions? Is the identified model serving its purposes well enough? What would be a next relevant step?

A main point of attention for this case study is how to convert the signal into a stream of images. The setup is then described as follows. At first, let a screen have $m = m_d \times m_w$ pixels organized in a rectangle of m_d pixels high and m_w pixels width. Let at instance t the screen be represented as the matrix $\mathbf{Y}_t$ which takes the form

$$\mathbf{Y}_t = \begin{bmatrix} y_{11,t} & \cdots & y_{1m_w,t} \\ \vdots & & \vdots \\ y_{m_d1,t} & \cdots & y_{m_dm_w,t} \end{bmatrix}. \tag{17.1}$$

It will be much more convenient in theory as well for the implementation to represent the current screen as a vector $\mathbf{y}_t \in \mathbb{R}^m$ by stacking the different columns of the matrix as

$$\mathbf{y}_t = (y_{11,t}, \ldots, y_{1m_w,t}, \ldots, y_{m_d1,t}, \ldots, y_{m_dm_w,t})^T \in \mathbb{R}^m. \tag{17.2}$$

This operation can be implemented in MATLAB using the command `reshape`.