15.2 Computer lab 2: Timeseries Modeling and Prediction

In this computer laboratory we will investigate how to analyze time series. The tasks contain

- Detrending (polynomial fitting)
- Estimation of periodical components
- AR modeling
- Prediction

The methods for time series modeling and prediction taught in the system identification course require the time series under study to be *stationary*. This means that its mean and variance are independent of time. To obtain such a time series, we usually *detrend* the data. This can for example be done by fitting a polynomial to the data and then removing this polynomial trend. The presence of deterministic *periodical components* in the data may also hamper the analysis. Such periodical components should thus be removed before modeling. There are many possible mathematical models for stationary time series: AR, MA, ARMA etc. (see Lecture 3). In this lab we will focus on describing the data as an *AR process* Once a model has been fitted to the data, this model can be used for *prediction* of the future behavior of the time series (see Lecture 3). In the case of an AR model, the optimal predictor is particularly easy to determine. This goes as follows. Suppose that the timeseries is modeled as an AR(d) model given as

$$y_t - a_1 y_{t-1} - \cdots - a_d y_{t-d} = e_t,$$

with white zero mean noise, or equivalently

$$y_t = a_1 y_{t-1} + \cdots + a_d y_{t-d} + e_t,$$

where $(a_1, \ldots, a_d)^T$ are the parameters which are estimated by (e.g.) a least squares estimator. This is also written as

$$A(q^{-1}) y_t = e_t,$$

where $A(q^{-1}) = 1 - a_1 q^{-1} - \cdots - a_d q^{-d}$. Then the best prediction of the next value y_{t+1} would be

$$\hat{y}_{t+1} = a_1 y_t + \cdots + a_d y_{t-d+1},$$

since there is no way one could have information on e_{t+1} by definition. This thus gives the optimal predictor associated to this model. Thus far, we assumed that the order d was fixed, but in the last exercise for today we also implement a method to detect the order of the system underlying the data. This issue will come back in the lecture on model selection, but for now it is enough to see the practical use.

The second step is to learn which frequencies are dominant in the timeseries. Therefore, the function `periodogram` is used. In case a few sinusoids are apparent in the timeseries, clear peaks will show up in the periodogram. It is not too difficult to detect their location, and subtract those sinusoids from the signal in order to get rid of this trend. This is implemented in `lab21a,m` using `periodogram.m` and `findpeaks.m`. This approach will be used in the next example, so do have a look at the code.

15.2.1 A stochastic Process

Lets build up the practice of analysis of timeseries gently. Consider the simple stochastic process given as

$$y_t = (1 + aq^{-1})^{-1} e_t$$

with $a = -0.8$ with $\{e_t\}$ white zero mean noise. Then the corresponding timeseries $\{y_t\}$ can be generated by the `filter` command as in the previous lab session. In the exercise sessions we saw what the corresponding covariance matrix is. Lets now see wether we see this covariance if we estimate the covariances of the data. Therefore, construct the sample covariance matrix $\hat{\mathbf{R}} \in \mathbb{R}^{d \times d}$ defined for any $i = 1, \ldots, d$ and $j = 1, \ldots, d$ as

$$\hat{\mathbf{R}}_{ij} = \frac{1}{n - |i-j|} \sum_{t=1}^{n-|i-j|} y_t y_{t+|i-j|}. \tag{15.17}$$

If you generate $n = 10$ samples from this process, compare now the sample covariance matrix $\hat{\mathbf{R}}$ to the theoretical result we derived earlier? Lets do this experiment $m = 100$ times (i.e. generate 100 times such data, and for each experiment construct $\hat{\mathbf{R}}$). Can you explain in one sentence why the average of those 100 runs tends to be the same compared to the sample covariance matrix obtained by performing the experiment that we had before with $n = 1000$?

15.2.2 Artificial Data

In this section, we will work on a “fictitious” data series having the following form:

$$y_t = \frac{1}{A(q^{-1})} e_t \tag{15.18}$$

$$z_t = y_t + p(t) + \sum_{k=1}^{K} A_k \sin(\omega_k t + \phi_k) \tag{15.19}$$

where $p(t)$ is a polynomial of order P, K is the number of sinusoids in the “periodical component” and $A(z)$ is given by (15.18). The data is generated with the MATLAB command **lab21b** (try several realizations of the data).

The objective of this exercise is to predict the future behavior of z_t given the first $n = 300$ data points. In order to solve this task we first have to make the time series stationary (by removing the polynomial trend). Then the periodical component is removed and the AR model is estimated. The predictor can then be written out as before, and can then be applied to the data series. These tasks are solved with the interactive MATLAB program `lab21c`.

- Generate the data with **lab21b**. The raw data z_t is displayed as a function of time. Also the periodogram of z_t is shown. Is the data stationary?
- Use the program **lab21c** to analyze the data. What order of the polynomial trend did you choose? What would happen if you choose a polynomial of a too high degree?
- How many sinusoids did you find in the data? What are their frequencies? Does the estimated periodical component fit the data well?

- What model order did you choose for the AR part of the data? Is the spectrum given by the estimated model close to the spectrum obtained from the data? Are there any discrepancies? If yes, how can they be explained?
- Predict the future behavior of the data. Compare the predicted data to the real time series. Did you do a good job predicting the future?

15.2.3 Real Data

Situate yourself in the happy 80's: the year is 1987, money is abundant and the real-estate market is flourishing. The politicians of a small coastal town in the far north of Sweden have plans to build a large outdoor water activity center: with tax-payer's money, of course. The first reactions from the public were negative: the weather is too cold and the season is far too short for this kind of project!

But the Mayor is not worried; he has read about global warming and is convinced that the average temperature in his town will increase rapidly, especially since he himself has bought a very large American car that pollutes just below the authorized limit.

To support his ideas, you have been given the task to look at temperature data and see if you can find evidence of warmer weather coming the Mayor's way. To your help you have the monthly temperatures measured since 1860. The data is prepared for processing by invoking the MATLAB command **lab21d**.

Write down in the report the parameters you choose. According to your predictions, what is your advice to the optimistic Mayor? Do you find any signs of warming in this part of Sweden? Will the project be a success?