02433 – Written exercise 3

This is the third and final out of three written exercises. The written report must be handed in by the week after course week 13 via the assignment facility on campusnet. The written report is mandatory and will be part of the final oral examination.

Background

This exercise considers a real data set consisting measurements of the wind power production (WPP) at the Klim wind farm. This wind farm is located along the North-Western coast of Jutland in Denmark, in the area of Klim Fjordholme. The wind farm has 35 wind turbines of 600kW each spread over various rows, for a total capacity of 21MW. A complete description of the site is available via this www-link.

Figure 1: Klim wind farm.

The WPP data have the following structure:

```
...
200201151615   14.67708   7597
200201151620   14.68056   7872
200201151625   14.68403   7216
...
```

where

1. column: time stamp indicating the time of the measurement. Ex: 200201151615 for the first measurement recorded on the 15 January 2002 at 16:15.
2. column: other form of time stamp indicating the time of the year. Ex: 14.67708 for the 15 January 2002 at 16:15, so 67.7% of the 14th day of the year.
3. column: wind power production measurements (in kW).

The data have a 5-minute temporal resolution with 10,000 data points from the 15 January 2002 at 16:15 and onwards. A power measurement for a given point in time corresponds to the average WPP for the whole wind farm over the previous 5-minute time interval. These power measurements then take values between 0 and 21.000kW, which is the installed capacity.

Suggested analysis approach

WPP data often show time varying dynamics. That is, the fluctuations in the wind speed cause changes in WPP magnitude over time. An obvious example is that the fluctuations in WPP are not present at times when the power reaches the maximum (or minimum) capacity. Changes in WPP fluctuations may however also occur when WPP is a away from its bounds. This is due to the fact that the meteorological conditions change, which leads to variations in the wind speed. To capture these switches between multiple dynamics in the data it may be advantageous to apply a Hidden Markov Model to obtain a dependent mixture of different time series models.

In general, the analysis of the wind power data should focus on fitting different time series models to a subset of the data (e.g. the first 8000 data points) and then assessing the models' forecasting ability. The analysis should start out with simple models (such as random walk and ARIMA models), and then be extended to HMMs (with two, three and perhaps four states) and increasingly advanced models in each state (e.g. AR with different number of lags, or even more complex models).

Here are some additional topics that such an analysis could address

- Evaluate the performance of the EM algorithm versus direct optimization of the likelihood in terms of convergence, sensitivity to starting values, and computing time for a selected HMM (e.g. three-state).
- Interpret the estimated values of the process error variances in the different regimes (i.e. the variance of the state-dependent density). They should reflect the size of the fluctuations they pertain to. What do these variances mean?
- Use model selection and model checking to benchmark alternative models against each other.
- Evaluate the forecasting ability of the fitted models on e.g. the remaining 2000 data points in the dataset. This can be done by calculating the root mean square error (R) of the conditional expectation of the one-step-ahead observation $\mu_{t+1} = E(X_{t+1}|\mathbf{X}^{(t)} = \mathbf{x}^{(t)})$, i.e.

  $$R = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1}(X_{t+1} - \mu_{t+1})^2}. \tag{1}$$

  Experiment with and discuss other ways to summarise and evaluate the forecasting performance of a model.
- Since the data are bounded by the maximum and minimum production capacity it can be difficult to select an appropriate state-dependent density because most densities will extend outside the bounds. Still, Gaussian densities can provide good results even though they do not adhere to the bounds. Discuss possible alternatives to the Gaussian density. Could transforming the data be an option to consider?

Exercise

Analyse the WPP dataset where you focus on parameter estimation and forecasting. You can get inspiration to the analysis from the approach suggested above, however taking your own approach to analysis is also fine. The dataset can be downloaded from the course website. Summarise and discuss your findings in a written report. Please include any computer code in the appendix.

End of exercise