Segmentation with hidden Markov models: comparison of different state path estimators

Hidden Markov models (HMMs) is a popular class of probabilistic models for modeling sequentially dependent data. HMMs are widely used for example in image analysis, bioinformatics, language processing, but also in financial mathematics.

Suppose we have a sequence of observations $X = (X_1, \ldots, X_n)$, that depend on an underlying sequence of latent variables $Z = (Z_1, \ldots, Z_n)$. We assume that the latent variables can take on K different values, meaning that the latent variable sequence can be in K different states. We can consider a sequence of electricity prices as observations, then the latent variables can model possible underlying hidden regimes of electricity prices, which could correspond for example to a low-price regime, high-price regime and a spike regime [1]. We consider a hidden Markov model

$$(X, Z) = \{(X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n)\},$$

where the hidden variable sequence Z is modeled with Markov chain.

We are interested in estimating the underlying state sequence Z given the observation sequence X, this is also called segmentation or classification problem. In the example of electricity prices we would like to estimate the hidden regimes. In practice, the most common state sequence estimators are the so-called Viterbi path and PMAP path. The Viterbi path maximizes the probability of the state path for given observation sequence:

$$\hat{z}_{Viterbi} = \arg\max_{z} P(Z = z | X = x).$$

The PMAP estimator maximizes pointwise probabilities:

$$\hat{z}_{PMAP} = \arg \max_{z=(z_1,\ldots,z_n)} \sum_{t=1}^{n} P(Z_t = z_t | X = x).$$

There exists also the whole class of so-called hybrid estimators [2], that operate between the Viterbi and PMAP path. The purpose of this Master's thesis and the tasks included are the following:

1) to become acquainted with the class of hidden Markov models and learn the main concepts connected to these models;

2) to understand the main ideas behind the Viterbi, PMAP and hybrid estimators and to characterize their behaviour with different examples;

3) to apply HMMs to a concrete data set (one can model for example electricity prices or stock market returns), perform the segmentation task with different estimators and compare the results.

The analysis can be performed with different packages in R.

[1] Apergis, N., Gozgor, G. et al. (2019). *Decoding the Australian electricity market: New evidence from three regime hidden semi-Markov model*, Energy Economics, 78, 129-142.
[2] Lember, J., Koloydenko, A. (2014). *Bridging Viterbi and posterior decoding: a generalized risk approach to hidden path inference based on hidden Markov models*, Journal of Machine Learning Research, 15, 1-58.