Chapter 1

Time Series

1.1 Goal

We are looking to model insurance data using machine learning methodologies. At the its core our models use time series models such as ARIMA to model observables or hidden variables which control the distribution of observables. An ARIMA time series A can be considered as a transformation $A(N)$ on a vector N of i.i.d. noise samples which are usually normally distributed.

We would like to apply cross validation techniques to select the hyperparameters of the ARIMA time series and other model options.

1.2 ARIMA Cross-Validation

In the usual setting time series cross-validation is done by fitting to a time contiguous training data set and evaluating fit on a future time contiguous testing data set (see here and references within). This limits the number of cross-validation folds that can be generated from limited data (as is the case in insurance).

In principle, time series model cross-validation can be done for arbitrary training and testing data sets. For example, we can ask given future samples (the training data set) what is the probability of the past samples being as they are (the testing data set), or alternatively given a training data set excluding some middle contiguous time section of data what is the probability of observing the excluded section.

1.2.1 Task I

Implement the following in PyTorch and submit your work in a public Git repository

- Create a PyTorch module describing an ARIMA(0,1,1) time series.
- Generate a random 20 sample long ARIMA(0,1,1) time series with drift.
- We define the model fit process as finding the maximum likelihood parameters that fit the training data set.
- Fit the model parameters to a training data set comprised of the first 14 samples of the time series generated in the second item.
- Calculated the probability of observing a testing data set comprised of last 6 samples given the model fitted in the previous step.

1.2.2 Task II

Describe how you would do the task in 1.2.1 given a training data set comprised of first and last 7 samples and a testing data set comprised of the remaining samples.