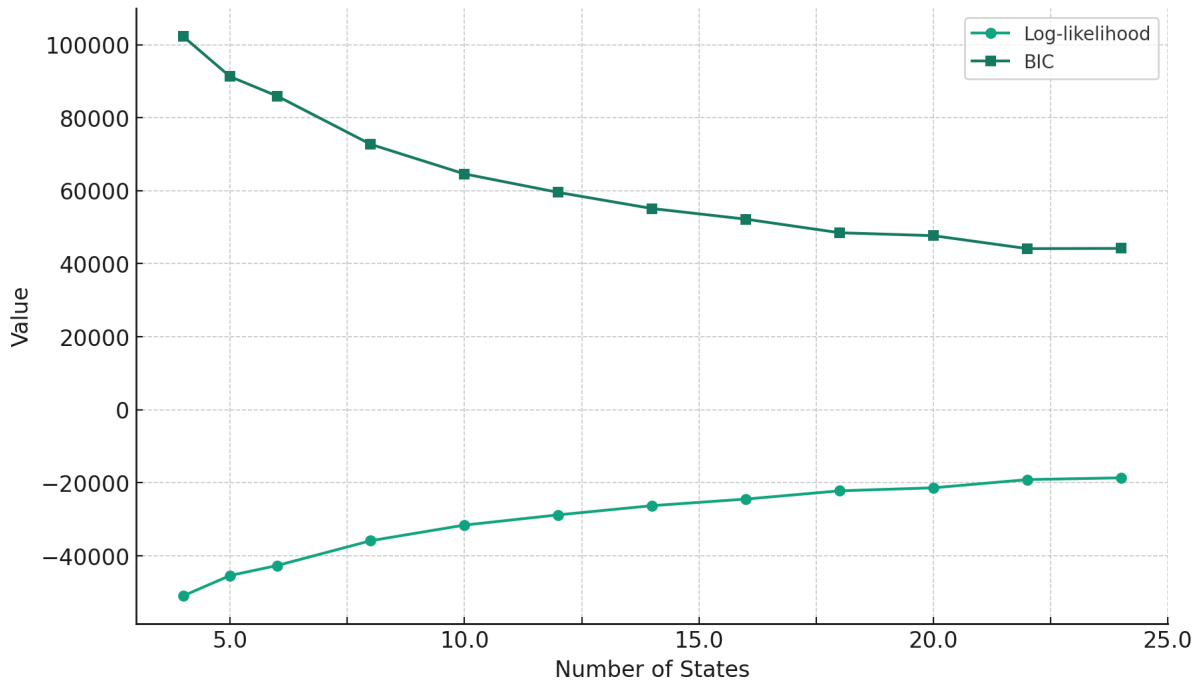
TERM PROJECT STEP BY STEP

- Divide your data into **train** and **test** subsets.
- Clean extreme values and take care of N/A values (interpolation).
- **Scale** the train data.
- Find the best selection of features using **PCA** on the train data (2 or 3 features).
- Pick a time window.

- Train/build a **good model** with the right selection of features and the right number of states (using BIC value and log-likelihood).

  NOTE: Do not forget to use ntimes when you are training models and make sure all models are **converged**.

  NOTE: Although a positive log-likelihood doesn’t necessarily mean you've done something wrong, aim to achieve a negative log-likelihood (perhaps by using the discretization technique).



- Use the specification of the fit model and create a similar model with the same specification. Then, feed the **scaled test data** to this new model in order to calculate log-likelihood of the test data on the trained model.

- **Compare the normalized train and test log-likelihood** to make sure your model is not overfit nor underfit.
  NOTE: If the size/length of the train data/observation is different than the size of the test data you must normalize both train and test loglikelihoods in order to compare them. This means, divide each log-likelihood by the length of its data.

- Divide the test data into chunks, each containing a few weeks. Feed these chunks into your model to get the normalized log-likelihood, and then identify the maximum deviation from the normalized train log-likelihood.