Studying the Effects of Training Data Distribution on Neural Network Summary Statistics of Stochastic Time Series

Domain: Bayesian Statistics, Machine Learning, Computational Biology
Project supervisor: Prashant Singh

Introduction

Artificial neural networks (ANNs) have proven to be highly effective as feature learners across diverse applications. An application of interest is learning informative features or summary statistics of stochastic time series data. Traditional summary statistics of time series include statistical moments such as mean, median, mode, etc. Such features are typically not expressive enough to capture the stochasticity of the time series. Neural network architectures such as convolutional networks are capable of extracting fine, discriminative patterns in temporal data, and thus serve as effective summary statistics of time series data [1]. Such temporal data are often encountered in different settings, including gene regulatory networks (GRNs) which forms the focus of this project. The neural network summary statistics learn a mapping from a high-dimensional space (i.e., the raw time series), to a very low dimensional space (GRN simulation model parameters).

Since we consider a stochastic GRN model setting, and since the input is fairly high dimensional (typically several hundred variables), obtaining a well-representative training set is crucial towards arriving at an accurate neural network summary statistic. Although the relationship between training data and machine learning model quality has been well-studied, the specific setting of high-dimensional stochastic time series obtained from GRN models is a challenging, special case that needs to be explored in greater detail.

Task

Within the scope of the project, the task will be to first understand the effect of varying amounts of noise on summary statistic quality, and subsequently identify means to design well-representative training sets for neural network summary statistics.

Required expertise

1. Good understanding of Python programming
2. Basic understanding of machine learning / willingness to learn about machine learning and neural networks

References:

1 - Åkesson, Mattias, Prashant Singh, Fredrik Wrede, and Andreas Hellander. "Convolutional Neural Networks as Summary Statistics for Approximate Bayesian Computation." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021). Arxiv link.