Rare-Event Simulation for time series

Wei Jiang[1], Lachlan J. Gibson[1] and Dirk P. Kroese[1]

1

Abstract

This report will demonstrate a technique to estimate the probability of rare events in time series. This method uses Importance Sampling to estimate the probability through a proxy distribution, which is an approximation for a target distribution. The target distribution is represented by a fitted Gaussian Processes (GP) on the time series data and conditioned on a rare Event. And the proxy distribution is achieved by fitting Normalizing Flows on the target distribution. This report conducts two experiments to show that Normalizing Flows will approximate the target distribution quite well, and the method will be competent in Rare-Event Simulation for time series.

Keywords: Rare-Event Simulation; Time Series; Normalizing Flows; Gaussian Process; Importance Sampling

Introduction

Traditionally, time series are studied using mathematical analysis methods and numerical methods, such as the example described in (Kroese *et al.* 2013, EXAMPLE 17.4). In the era of big data, time series problems usually begin with a batch of data, and data-driven methods are more popular (Brunton and Kutz (2022)). However, the amount of data is not sufficient for rare events simulation. This is the difficulty of the Rare-Event Simulation. The common practice is to fit a model with the dataset first, and then work on this model.

This report will demonstrate a way to simulate rare events for time series. The following report contains three parts. Section Dataset and methods will introduce the time-series dataset and detail the technique of Rare-Event Simulation for time series. Section Experiments will present two experiments for time series. And Section Discussion will discuss the issues about experiments.

Dataset and methods

The time-series dataset used in this report is about the atmospheric CO_2 concentration measured monthly from the Mauna Loa observatory in Hawaii (Keeling *et al.* (2001)), as shown in Figure 1. Each point represents an observation of CO_2 concentration of each month from January 1966 to February 2019. After the year 1996, I use Gaussian Processes to predict the concentration until 2025. The blue area is one standard variance of Gaussian Distribution conditioned on time.

It is clear that the prediction captured the pattern of the data, and the uncertainty grows as time shifts away from the data. Suppose we are interested in the probability of prediction between 2022 to 2025 with CO_2 concentration above 400. This is a rare event, and the probability estimation is Rare-Events Simulation.

Following the practice mentioned in section Introduction, Rare-Event Simulation for time series can be divided into three parts:

- fitting a model A for time series. Model A should be a probabilistic model and can fit a distribution for time series conditioned on a rare event. Usually, this distribution is too complex to evaluate density or sample from it.
- fitting a model B to approximate the target distribution defined on model A. This distribution is called proxy distribution. It is used in the Importance Sample and is easy for evaluation and sampling.
- using Importance Sampling to estimate the probability for rare events.

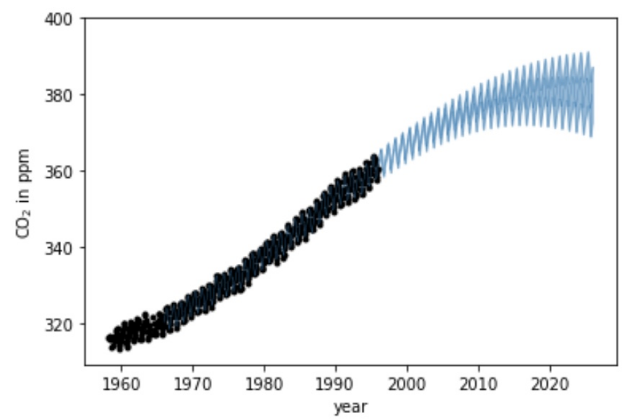

Figure 1 the atmospheric CO_2 concentration is measured monthly from the Mauna Loa observatory in Hawaii. This figure displays the data from 1966 to 1996 and the prediction afterwards.

This report will use Gaussian Processes to fit time series. The three reasons to use Gaussian Processes are:

- The Gaussian Processes is a typical method for the time series and has already been applied in many real-life applications (Cheng *et al.* (2020));
- The results of GP are easy for further analysing because the inference of GP is a conditioned Gaussian distribution;

- Many high-quality implementations of Gaussian Processes already exist, so that we can focus on the rare event simulations.

After getting a target distribution using Gaussian Processes, I will use Normalizing Flows to fit a model with a density shape that is closed to the target distribution. And finally, perform Importance Sampling to estimate the probability.

Experiments

I conducted two experiments in this report. The first experiment is used to test if the Normalizing Flows are working properly with Gaussian Processes, and the second experiment is about Rare-Event Simulation. Code is available at: https://github.com/bayesbreeze/Rare-Events/blob/main/demos/demo3.ipynb

Expriment 1

Experiment One is used to estimate the normalization constant of unnormalized distribution between the years 2022 and 2025 as shown in Figure 2. The upper figure is about the conditional distribution between the years 2022 and 2025. The blue area represents one standard deviation of the gaussian distribution conditioned on time, and the dark line in the middle is the mean. The below figure is an approximated distribution generated by fitted Normalizing flows for the above conditional distribution. I use the fitted distribution combined with the Importance Sampling to get an estimation of the integral 2.9967, which is pretty close to the correct value 3.

This experiment shows that this method is competent in time series simulation. Next, I will use Normalizing flows in Rare events simulation.

Expriment 2

Experiment Two is used to estimate the probability of CO_2 concentration level above 400 between the years 2022 and 2025. Like Expriment 1, I first use Normalizing Flows to fit the unnormalized target distribution, getting an approximated distribution as shown in Figure 3. Then, I use Importance Sampling to estimate the probability through the proxy distribution. I run 10 times of simulation, getting a probability of 0.002494, and the standard deviation is 0.000011. Later, I will try to estimate the probability using other methods, such as numerical methods, for comparison.

Discussion

Usually, it is hard to check the accuracy of simulation for rare events, because the distribution is complex, it has no closed form; and the numerical solution is also difficult because of the high dimensionality.

The training for the current method will converge to a wrong distribution when training data for each batch is small. I will try to use BatchNorm (Ioffe and Szegedy (2015)) to fix this problem. If it does not work, I will try another method: using MCMC to sample from the target distribution and then fit the model using Normalizing Flows.

Finally, I will try to figure out a way to speed up the training by using more efficient Gaussian Processes implementations.

References

Brunton SL, Kutz JN. 2022. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press.

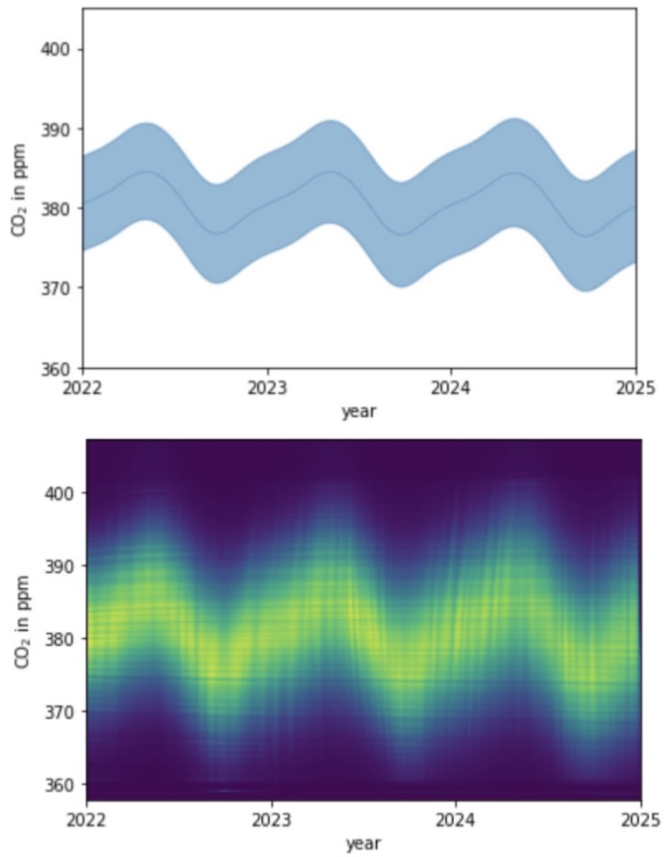

Figure 2 The distribution of CO_2 concentration between 2022 and 2025. The upper figure shows the conditional distribution predicted by GP, where the center line is the mean of the conditional distribution, and the blue area represents one standard deviation. The lower figure demonstrates the distribution fitted by Normalizing flows, and it captures the patterns of the conditional distribution above.

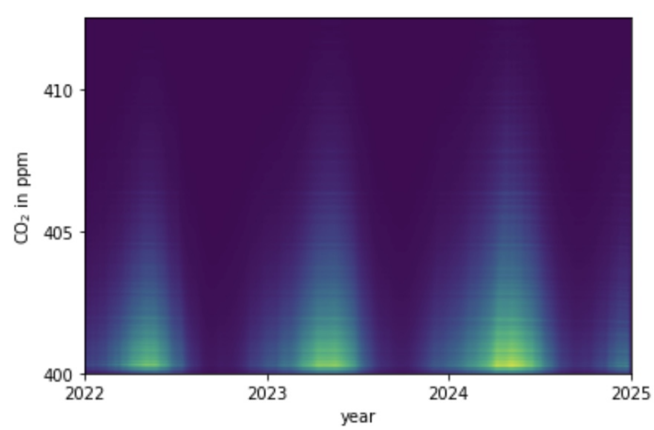

Figure 3 the distribution for rare events fitted by Normalizing Flows. The rare event defines as the CO_2 concentration above 400 between the year 2022 and the year 2025.

Cheng LF, Dumitrascu B, Darnell G, Chivers C, Draugelis M, Li K, Engelhardt BE. 2020. Sparse multi-output gaussian processes for online medical time series prediction. BMC medical informatics and decision making. 20:1–23.

Ioffe S, Szegedy C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: . pp. 448–456. PMLR.

Keeling CD, Piper SC, Bacastow RB, Wahlen M, Whorf TP, Heimann M, Meijer HA. 2001. Exchanges of atmospheric co2 and 13co2 with the terrestrial biosphere and oceans from 1978 to 2000. i. global aspects. .

Kroese DP, Taimre T, Botev ZI. 2013. *Handbook of monte carlo methods*. John Wiley & Sons.