Table 1: Impact of the score model $h(x, \theta)$ on detection performance (F1 Score) on MNIST.We trained the score model via sliced score matching with different numbers of training epochs 20, 50, 100, and 200 and then evaluated the impact of the score model on the change point detection performance using the F1-score.

Metric	20 Epochs	50 Epochs	100 Epochs	200 Epochs
F1 Score	0.243	0.531	0.736	0.741

Table 2: The other two real datasets come from the baseline.We use the same setting in our experiment to evaluate the five methods,the results are shown below.The Bee-Dance dataset is selected from Change-point Detection with Auxiliary Deep Generative Models (KL-CPD). Text language dataset is selected from Change Point Detection with Neural Online Density-Ratio Estimator (NODE).We use the F1-score as the evaluation metric to keep with our previous work

	Bee-Dance	Text language
BOCPDMS	0.167	F
NODE	0.372	0.571
Mstats-CPD	0.341	0.394
KL-CPD	0.401	0.532
KSD-CPD	**0.474**	**0.589**

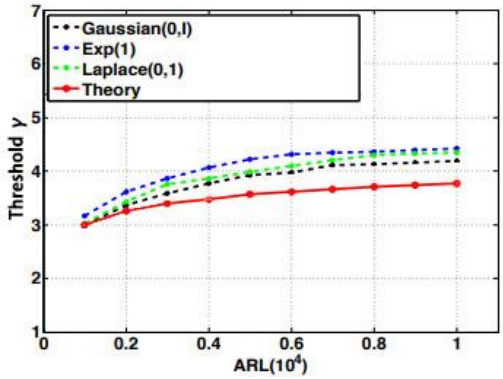

Figure 1: The simulation results are obtained from 5000 direct Monte Carlo trials.Theorem 5.1 shows that $ARL \sim O(e^{\gamma^2})$.We validate our theoretical ARL approximation through simulations on time series data (with 10,000 data points) under various null distributions,such as standard normal,exponential,and Laplace using 5,000 Monte Carlo trials.We adjust the threshold γ ,run the time series without change point and record the ARL when a false alarm is reported