

Benchmarking optimality of time series classification methods in distinguishing diffusions

Zehong Zhang¹, Fei Lu¹, Esther Xu Fei^{2,4}, Terry Lyons³,
Yannis Kevrekidis^{4,5}, and Tom Woolf⁶

¹Department of Mathematics, Johns Hopkins University, Baltimore, MD 21218, USA feilu@math.jhu.edu

²Department of Environmental Health and Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

³Mathematical Institute, University of Oxford, Oxford, United Kingdom

⁴Department of Applied Mathematics and Mathematics, Johns Hopkins University, Baltimore, MD 21218, USA

⁵Departments of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

⁶School of Medicine, Johns Hopkins University, Baltimore, MD 21218, USA.

Abstract

Performance benchmarking is a crucial component of time series classification (TSC) algorithm design, and a fast-growing number of datasets have been established for empirical benchmarking. However, the empirical benchmarks are costly and do not guarantee statistical optimality. This study proposes to benchmark the optimality of TSC algorithms in distinguishing diffusion processes by the likelihood ratio test (LRT). The LRT is optimal in the sense of the Neyman-Pearson lemma: it has the smallest false positive rate among classifiers with a controlled level of false negative rate. The LRT requires the likelihood ratio of the time series to be computable. The diffusion processes from stochastic differential equations provide such time series and are flexible in design for generating linear or nonlinear time series. We demonstrate the benchmarking with three scalable state-of-the-art TSC algorithms: random forest, ResNet, and ROCKET. Test results show that they can achieve LRT optimality for univariate time series and multivariate Gaussian processes. However, these model-agnostic algorithms are suboptimal in classifying nonlinear multivariate time series from high-dimensional stochastic interacting particle systems. Additionally, the LRT benchmark provides tools to analyze the dependence of classification accuracy on the time length, dimension, temporal sampling frequency, and randomness of the time series. Thus, the LRT with diffusion processes can systematically and efficiently benchmark the optimality of TSC algorithms and may guide their future improvements.

Key words Times series classification, Likelihood ratio test, Optimal benchmark, Stochastic differential equations, ResNet, ROCKET, Random forest

1 Introduction

Time series classification (TSC) is one of the central tasks in time series analysis and streaming data processing. Recent years have seen an explosion in the collection of time series data and a surge of TSC algorithms (see e.g., [1, 2, 4, 8, 13, 14, 19, 21, 22, 29, 30]). In particular, the recent reviews [2, 13, 29] have thoroughly compared dozens of TSC algorithms on hundreds of public bakeoff datasets, providing valuable understanding of the algorithms and the TSC tasks.

However, an optimality benchmark remains missing. The need for an optimality benchmark grows along with the fast-growing numbers of datasets and algorithms. Due to a lack of understanding of the complexity of the bakeoff datasets, current empirical benchmarks, which compare all methods using bakeoff datasets, have skyrocketing computational and data storage cost. Yet, even a top performer is not cleared to be optimal.

An ideal optimality benchmark would have three characteristics: (1) It has a theory-guaranteed optimal reference to provide a direct diagnosis for any TSC method. Notably, a method reaching the benchmark for a type of time series is guaranteed optimal for classifying the underlying stochastic process, and efforts can focus on improving the efficiency and scalability of the method. (2) It is flexible in design to reflect the complexity of time series data in applications, ranging from univariate to multivariate time series, from Gaussian processes to highly nonlinear non-Gaussian processes, and from small to large randomness. (3) It is computationally efficient and scalable.

We propose to benchmark the optimality of binary TSC algorithms in distinguishing diffusion processes by the likelihood ratio test (LRT). The LRT is an optimal classifier because it is uniformly most powerful by the Neyman-Pearson lemma [24]; that is, it has the lowest false positive rate among classifiers with a controlled level of false negative rate. The LRT can be computed for time series sampled from Markov processes with known distributions. Meanwhile, diffusion processes from stochastic differential equations (SDEs) provide a large variety of such Markov processes, and these processes are flexible to reflect the specific features of real-world applications [25, 26]. Additionally, the benchmarking test is computationally efficient and scalable. The LRT does not require training and has a negligible computation cost. Furthermore, the simulation of SDEs can systematically generate large datasets with different lengths, nonlinearities, and levels of randomness. Therefore, LRTs for diffusion processes provide a reference of optimality for the performance (such as the ROC curves and accuracy) of all TSC algorithms.

We demonstrate the LRT benchmarking using three state-of-the-art TSC algorithms: random forest [4], ROCKET [8], and ResNet [30], in five representative classes of diffusion processes. The five processes are Brownian motions with constant drifts, 1-dimensional nonlinear diffusions with different potentials, 1-dimensional linear and nonlinear diffusions, multivariate Ornstein-Uhlenbeck processes, and high-dimensional interacting particle systems. Test results (see, e.g., Figure 5-6) show that the three algorithms achieve optimality in the case of Brownian motions with constant drifts, and they are near optimal for the nonlinear univariate time series and multivariate Gaussian processes. However, these three model-agnostic algorithms are significantly less accurate than the model-aware LRT in the case of high-dimensional nonlinear non-gaussian processes. Thus, it may be helpful to incorporate model information in developing next-generation TSC algorithms.

Additionally, the LRT benchmarks show that the optimal accuracy of TSC depends on the time series’s length, dimension, and temporal sampling frequency. Analysis and numerical tests show that the accuracy increases with either time length or dimension, which enlarges the effective sample size. However, the classification rates are not sensitive to the frequency of the observations. Thus, in data collection, it is more helpful to collect data for a longer time rather than at a higher temporal resolution.

The rest of the paper is organized as follows. In Section 2, we cast the TSC as the learning of a function that maps a time series to a binary output so that a TSC algorithm can be viewed as a hypothesis testing method. In particular, we point out that the likelihood ratio test (LRT) is a uniformly most powerful test by the Neyman-Pearson Lemma. Additionally, we show the computation of the likelihood ratio for diffusion processes. Section 3 analytically computes the LRT for two Gaussian processes. The analysis shows the dependence of the classification accuracy on the time series’s dimension, length, and frequency. Section 4 describes three examples of nonlinear diffusion processes and specifies the data generation for benchmarking tests. These examples showcase the design of benchmarking tests. We present in Section 5 the test results of benchmarking three scalable TSC algorithms: the random forest, ResNet, and ROCKET. Finally, the Appendix briefly reviews the Girsanov theorem and hypothesis testing.

2 Time series classification and distinguishing diffusions

We recast binary time series classification as a hypothesis testing problem, so that the likelihood ratio test (LRT) provides an optimal classifier by the Neyman-Pearson Lemma. On the other hand, diffusion

Table 1: Confusion matrix of the classifier with $\{\theta_0\}$ being positive.

Truth \ Decision	Decision		Rates/ Probability of errors	
	Accept θ_0	Reject θ_0		
θ_0 (Positive)	TP	FN	TPR = $1 - \alpha_k^0$	FNR = $\alpha_k^0 = \mathbb{E}[F(\mathbf{x}, k) \theta_0]$
θ_1 (Negative)	FP	TN	FPR = $1 - \alpha_k^1$	TNR = $\alpha_k^1 = \mathbb{E}[F(\mathbf{x}, k) \theta_1]$

* FN is also called type I error and FP is called type II error. The true positive rate (TPR) is $1 - \alpha_k^0$ and the false positive rate (FPR) is $1 - \alpha_k^1$.

processes provide a large variety of time series whose LRT can be computed in a scalable fashion. Thus, we propose to benchmark the optimality of TSC classifiers by LRT in distinguishing diffusions.

2.1 TSC as a function learning problem

In the lens of statistical learning, a binary TSC algorithm learns the probabilities that the time series belongs to two classes from training data [6, 15].

Let the data be the time series (either univariate or multivariate) and their labels,

$$\mathbf{Data}: \quad \{\mathbf{x}^{(m)}, y^{(m)}\}_{m=1}^M, \quad \mathbf{x}^{(m)} \in \mathbb{R}^{d \times (L+1)}, y^{(m)} \in \{0, 1\},$$

where for each m , $\mathbf{x}^{(m)} = x_{t_0:t_L}^{(m)} = (x_{t_1}, \dots, x_{t_L})^{(m)}$ is a sample path of a stochastic process $\mathbf{X} = X_{t_0:t_L}$ with $t_0 < t_1 < \dots < t_L$ denoting time indices. Here $y^{(m)}$ has a label 1 if the time series $\mathbf{x}^{(m)}$ is in class θ_1 ; otherwise, its label is 0 if the time series is in class θ_0 . We denote the two classes by $\{\theta_0, \theta_1\}$, which will be used as parameters for the time series models.

A TSC algorithm learns a function with a parameter β from data,

$$f_\beta(\mathbf{x}) = z, \quad \mathbf{x} \in \mathbb{R}^{d(L+1)}, z \in [0, 1], \quad (2.1)$$

such that the value $f_\beta(\mathbf{x})$ approximates the probability of \mathbf{x} being in class θ_1 , i.e., $\mathbb{P}(\theta = \theta_1 | \mathbf{X} = \mathbf{x})$. This function leads to a classifier for any threshold $k \in (0, 1)$:

$$F(\mathbf{x}, k) = \mathbf{1}_{R_k}(\mathbf{x}), \quad \text{where } R_k = \{\mathbf{x} : f_\beta(\mathbf{x}) > k\}, \quad (2.2)$$

where R_k is called the *acceptance region* to classify the time series \mathbf{x} as in class θ_1 (equivalently, the *rejection region* for the class θ_0).

The confusion matrix of the binary classifier (2.2) with θ_0 as positive is shown in Table 1. For a given threshold k , we have a false negative (FN) prediction if $F(\mathbf{x}, k) = 1$ while \mathbf{x} is in class θ_0 , and we have a false positive (FP) prediction if $F(\mathbf{x}, k) = 0$ while \mathbf{x} is in class θ_1 . The definitions of true positive (TP) and true negative (TN) are similar. The false negative rates (FNR) and the true negative rates (TNR) rates are the probabilities

$$\begin{aligned} \text{FNR}(k) &= \alpha_k^0 = \mathbb{E}[F(\mathbf{x}, k) | \theta_0] = \mathbb{P}(R_k | \theta_0) \approx \frac{FN}{TP + FN}, \\ \text{TNR}(k) &= \alpha_k^1 = \mathbb{E}[F(\mathbf{x}, k) | \theta_1] = \mathbb{P}(R_k | \theta_1) \approx \frac{TN}{TN + FP}, \end{aligned} \quad (2.3)$$

where the empirical approximations are based on the number of counts.

Two popular metrics evaluating the performance of the classifier are the *Receiver operating characteristic* (ROC) curve and *accuracy*. The ROC curve is $(1 - \alpha_k^0, 1 - \alpha_k^1)_{k \in (0, 1)}$, the curve of True Positive Rate (TPR, y-axis) versus False Positive Rate (FPR, x-axis), both parametrized by the threshold k (see e.g., [10] for an introduction). The ROC curve allows the user to define the threshold and measure the

quality of a classifier by the *area under the curve* (AUC). A rule of thumb is that the larger is the AUC, the better is the classifier. The accuracy is defined by:

$$\text{Accuracy}(k) = \frac{1 - \alpha_k^0 + \alpha_k^1}{2} \approx \frac{TP + TN}{TP + TN + FP + FN},$$

where the approximate equality becomes an equality when the sample sizes in the two classes are the same. The maximal accuracy is independent of the threshold:

$$ACC_* = \max_{0 \leq k \leq 1} \text{Accuracy}(k), \quad (2.4)$$

We will use AUC and the maximal accuracy to assess the classifiers, because they are independent of a specific threshold. There are many other metrics to fit the goal of a specific field, i.e., choosing a threshold k to increase the *true positive rate* (TPR) (aka sensitivity, power, or recall) $1 - \alpha_k^0$ or to control the false positive rate (FPR) $1 - \alpha_k^1$ (aka specificity), or a balance balancing these needs [15].

Sampling errors in training and testing. Sampling errors are present in the training and the testing data, thus they affect the accuracy of the classifier. The accuracy of the function learned in a classifier can be analyzed through mathematical and statistical learning theory (see e.g., [6, 9, 15]), and non-asymptotic error bounds are available to quantify the dependence on the data size based on concentration inequalities [7, 11]. The sampling error in the testing stage, on the other hand, can be easily analyzed: the empirical approximation of the rates in (2.3) have a sampling error of order $O(\frac{1}{\sqrt{m}})$ with m being the number of test samples, as the next lemma shows (its proof is in Appendix A.2).

Lemma 2.1 (Sampling error in FNR/TNR) *For each classifier in (2.2), the sampling errors in the empirical approximations of the FNR and TNR rates in (2.3) are of order $\frac{1}{\sqrt{m}}\sigma_{k,i}$ with $\sigma_{k,i} = \sqrt{\alpha_k^i(1 - \alpha_k^i)}$ for $i = 0, 1$, where m is the number samples in the test stage. Specifically, let $\{\mathbf{x}_j\}_{j=1}^m$ be the test samples, and let $\hat{\alpha}_{k,m}^i = \frac{1}{m} \sum_{j=1}^m F(\mathbf{x}_j, k)$ conditional on θ_i . Then, $\hat{\alpha}_{k,m}^i$ converges in distribution to $\mathcal{N}(0, \sigma_{k,i}^2)$ as $m \rightarrow \infty$, and $\mathbb{P}(|\hat{\alpha}_{k,m}^i - \alpha_k^i| > \epsilon) \leq 2e^{-\frac{m\epsilon^2}{2}}$ for any $\epsilon > 0$ and $m > 0$.*

However, the learning theory does not provide empirical optimality criteria for the performance of the classifier. The likelihood ratio test in the next section fills the gap.

2.2 Hypothesis testing and the likelihood ratio test

The hypothesis testing methods construct the classifier function in a statistical inference approach (see [5, Chapter 8] and Section A.3 for a brief review). In particular, the *Neyman-Pearson lemma* provides a powerful tool for analyzing the optimality of a binary classifier: it shows that the likelihood ratio test is a *uniformly most powerful test* in the class of tests with the same level (see [5, Chapter 8] and Section A.3 for a brief review).

In hypothesis testing, we set the null hypothesis to be $H_0 : \theta = \theta_0$ and the alternative hypothesis to be $H_1 : \theta = \theta_1$, and we select a *rejection region* R_k with a threshold k to reject θ_0 . Then, the classifier rejects the null hypothesis H_0 if the time series is in the rejection region R_k . In other words, we get a false native (FN) if we mistakenly reject H_0 while the truth is θ_0 , and we get a true negative (TN) if we correctly reject H_0 when the truth is θ_1 . The false negative rate (FNR) and true negative rate (TNR) are the probabilities in (2.3). The major task in a hypothesis test is to select the rejection region R_k , particularly, to select R_k with a tunable threshold k .

The *likelihood ratio test* (LRT) is a general hypothesis testing method that is as widely applicable as maximum likelihood estimation. It determines the rejection region by statistics derived from the likelihood ratio. The commonly-used statistics is the log-likelihood ratio

$$l(\mathbf{x} \mid \theta_1, \theta_0) = \log \frac{p(\mathbf{x} \mid \theta_1)}{p(\mathbf{x} \mid \theta_0)}$$

of the time series data \mathbf{x} . From this statistics, we can define a function approximating the probability of \mathbf{x} being in class θ_1 , which is a counterpart of $f_\beta(\mathbf{x})$ in (2.1): $f(\mathbf{x}) = \frac{1}{e^{l(\mathbf{x}|\theta_1, \theta_0)} + 1}$. Then, the classifier function for LRT is $F(\mathbf{x}, k) = \mathbf{1}_{R_k}(\mathbf{x})$ with the rejection region defined by

$$R_k^{LRT} = \{\mathbf{x} : l(\mathbf{x} | \theta_1, \theta_0) > c_k\}, \quad c_k = \log \frac{k}{1-k}, \quad (2.5)$$

for each threshold $k \in (0, 1)$.

The Neyman-Pearson lemma shows that the LRT is optimal in the sense that it has the smallest false positive rate among classifiers with a controlled level of false negative rate:

Theorem 2.2 (Neyman-Pearson Lemma) *The LRT is a uniformly most powerful classifier. Specifically, let \mathbf{x} be a sample from one of two distributions with a likelihood ratio $l(\mathbf{x} | \theta_1, \theta_0)$ and assume that $\mathbb{P}(\{\mathbf{x} : l(\mathbf{x} | \theta_1, \theta_0) = k\}) = 0$. Then, the test with rejection region R_k^{LRT} defined in (2.5) is uniformly most powerful. That is, it has a false positive rate no larger than any other test with a measurable rejection region R such that $\mathbb{P}(R | \theta_0) \leq \mathbb{P}(R_k^{LRT} | \theta_0)$, i.e.,*

$$1 - \mathbb{P}(R | \theta_1) \geq 1 - \mathbb{P}(R_k^{LRT} | \theta_1), \quad \forall R \text{ s.t. } \mathbb{P}(R | \theta_0) \leq \mathbb{P}(R_k^{LRT} | \theta_0).$$

As a result, the LRT provides an ideal tool to analyze the optimality of TSC algorithms. The ROC curve of the LRT classifier provides an upper bound for the ROC curve of any TSC classifier. Similarly, the LRT classifier's accuracy provides an upper bound for other classifiers.

The LRT classifier can be readily applied to time series with a computable likelihood ratio, and there is no training stage. When the time series is sampled from a Markov process, the transition densities determine the likelihood ratio. Suppose that for each θ_i , the transition probability of the Markov process has a density function $p(x_{t+1} | x_t, \theta_i)$ for each l . Then, the probability density function of a data path $x_{t_0:t_L}$ conditional on θ_i is

$$p(x_{t_0:t_L} | \theta_i) = \prod_{l=0}^{L-1} p(x_{t_{l+1}} | x_{t_l}, \theta_i),$$

and the log-likelihood ratio of the path is

$$l(x_{t_0:t_L} | \theta_1, \theta_0) = \log \frac{p(x_{t_0:t_L} | \theta_1)}{p(x_{t_0:t_L} | \theta_0)} = \sum_{l=0}^{L-1} \log \frac{p(x_{t_{l+1}} | x_{t_l}, \theta_1)}{p(x_{t_{l+1}} | x_{t_l}, \theta_0)}. \quad (2.6)$$

However, the transition probabilities and the likelihood ratio are unavailable for most time series, except for a few simple examples such as Gaussian processes and linear models (see Section 3). In particular, to benchmark the performance of TSC algorithms, it is desirable to have nonlinear time series datasets with varying length, temporal sampling frequency, and dimension. The diffusions defined by stochastic differential equations provide a large class of such Markov processes.

2.3 Distinguishing diffusions

Diffusion processes provide a large class of time series whose likelihood ratio can be accurately computed. An Itô diffusion is defined by a stochastic differential equation

$$dX_t = b_\theta(X_t)dt + \sigma(X_t)dB_t, \quad (2.7)$$

where B_t is a standard \mathbb{R}^d -valued Brownian motion. Here for simplicity, we assume that both the diffusion coefficient $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ and the drift $b_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ with parameter θ are Lipschitz, and the diffusion satisfies the uniform elliptic condition $\sum_{1 \leq i, j \leq d} c_i c_j \sigma_{ki} \sigma_{kj}(x) \geq \gamma \sum_i c_i^2$ with $\gamma > 0$ for all x and $c_i \in \mathbb{R}$. Beyond such diffusions, we can also consider Itô processes with b and σ being general stochastic processes satisfying suitable integrability conditions [25].

The likelihood ratio of a sample path $x_{t_0:t_L}$ of the diffusion $X_{t_0:t_L}$ satisfying (2.7) can be computed by numerical approximation of the transition probabilities. In particular, when the temporal sampling frequency is high, i.e., $\max_l \{\Delta t_l = t_{l+1} - t_l\}$ is small, the Euler-Maruyama scheme

$$\Delta X_{t_l} = X_{t_{l+1}} - X_{t_l} \approx b_{\theta_l}(X_{t_l})\Delta t_l + \sigma(X_{t_l})\Delta W_l$$

yields an accurate approximation of the transition probability

$$\hat{p}(X_{t_{l+1}} | X_{t_l}, \theta_l) \propto e^{-\frac{1}{2\Delta t_l} \|\Delta X_{t_l} - b_{\theta_l}(X_{t_l})\Delta t_l\|_{\Sigma}^2},$$

where $\Sigma(x) = \sigma\sigma^\top(x) \in \mathbb{R}^{d \times d}$ and $\|z\|_{\Sigma}^2 = z^\top \Sigma^{-1}z$ for any $z \in \mathbb{R}^d$. Using it in (2.6), we obtain an approximate likelihood ratio:

$$\begin{aligned} & \hat{l}(X_{t_0:t_L} | \theta_1, \theta_0) \\ &= \sum_{l=0}^{L-1} \left([b_{\theta_1} - b_{\theta_0}](X_{t_l})^\top \Sigma(Y_s)^{-1} \Delta X_{t_l} - \frac{1}{2} [\|b_{\theta_1}\|_{\Sigma}^2 - \|b_{\theta_0}\|_{\Sigma}^2](X_{t_l}) \Delta t_l \right). \end{aligned} \quad (2.8)$$

As the temporal sampling frequency increases, i.e., $\max_l \{t_{l+1} - t_l\} \rightarrow 0$, the above likelihood ratio converges to the likelihood ratio of the continuous path $X_{[0,T]}$. The limit ratio is the Radon-Nikodym derivative between the two distributions of the path, as characterized by the Girsanov theorem (see Section A.1):

$$\begin{aligned} & l(X_{[0,T]} | \theta_1, \theta_0) \\ &= \int_0^T [b_{\theta_1} - b_{\theta_0}](Y_s)^\top \Sigma(Y_s)^{-1} dY_s - \frac{1}{2} \int_0^T [\|b_{\theta_1}\|_{\Sigma}^2 - \|b_{\theta_0}\|_{\Sigma}^2](X_t) dt. \end{aligned} \quad (2.9)$$

There are three advantages to benchmarking TSC algorithms by diffusions. First, the LRT of the diffusion processes provides the theoretical optimal rates, which can be used to detect overfitting when training TSC classifiers. Second, the diffusions provide a large variety of testing time series data, whose length, sampling frequency, dimension, and nonlinearity can vary as needed. Third, the likelihood ratio between diffusion processes can be efficiently computed by numerical approximation as in (2.8).

3 Examples with analytical likelihood ratios

The likelihood ratio can be computed analytically for Brownian motions with constant drifts and the Ornstein-Uhlenbeck (OU) processes. In particular, these two examples offer insights into how the classification accuracy depends on the temporal sampling frequency, length of paths, the randomness, and the dimension of the time series data.

3.1 Brownian motions with constant drifts

Let $(X_t, t \geq 0)$ be an \mathbb{R}^d -valued Brownian motion with a constant drift:

$$dX_t = \theta dt + \sigma dB_t, \quad \Leftrightarrow \quad X_t = X_0 + \theta t + \sigma B_t, \quad (3.1)$$

where $\theta \in \{\theta_0, \theta_1\} \subset \mathbb{R}^d$ and the process $(B_t, t \geq 0)$ is the standard Brownian motion starting at 0. Then, the exact log-likelihood ratio in (2.6) for a given sample path $X_{t_0:t_L}$ is

$$l(X_{t_0:t_L} | \theta_1, \theta_0) = \sigma^{-2} \left[(\theta_1 - \theta_0)^\top (X_{t_L} - X_{t_0}) - \frac{1}{2} (|\theta_1|^2 - |\theta_0|^2) (t_L - t_0) \right].$$

Note that $X_{t_L} - X_{t_0} = \theta(t_L - t_0) + \sigma(B_{t_L} - B_{t_0})$ for each θ . Thus, conditional on the hypotheses $\theta = \theta_0$ and $\theta = \theta_1$, the likelihood ratios have distributions

$$\begin{aligned} \text{Hypothesis } \theta = \theta_0 : & \quad l(X_{t_0:t_L} | \theta_1, \theta_0) \sim -m_l + v_l Z, \\ \text{Hypothesis } \theta = \theta_1 : & \quad l(X_{t_0:t_L} | \theta_1, \theta_0) \sim m_l + v_l Z, \end{aligned}$$

where Z is a standard Gaussian random variable and

$$m_l = \frac{1}{2}|\theta_1 - \theta_0|^2(t_L - t_0), \quad v_l = \sigma|\theta_1 - \theta_0|\sqrt{t_L - t_0}.$$

Let the rejection region be $R_k = \{X_{t_0:t_L} : l(X_{t_0:t_L} | \theta_1, \theta_0) > c_k\}$ with $c_k = \log \frac{k}{1-k}$ as defined in (2.5). Then, the false negative rate (FNR) and the true negative rate (TNR) of the LRT are

$$\begin{aligned} \text{FNR}(k) &= \alpha_k^0 = \mathbb{P}(x_{t_0:t_L} \in R_k | \theta_0) = \mathbb{P}(Z > c_k v_l^{-1} + m_l v_l^{-1}) \\ \text{TNR}(k) &= \alpha_k^1 = \mathbb{P}(x_{t_0:t_L} \in R_k | \theta_1) = \mathbb{P}(Z > c_k v_l^{-1} - m_l v_l^{-1}). \end{aligned}$$

Then, the accuracy $\frac{1}{2}(1 - \alpha_k^0 + \alpha_k^1)$ is $ACC_k = \frac{1}{2} + \frac{1}{2}\mathbb{P}(-m_l v_l^{-1} < Z - c_k v_l^{-1} < m_l v_l^{-1})$. Since Z is centered Gaussian, the threshold maximizing the accuracy is $k_* = \arg \max_{k \in (0,1)} (ACC_k) = 0$. As a result, the maximal accuracy is

$$\begin{aligned} ACC_* &= \frac{1}{2} + \frac{1}{2}\mathbb{P}(-m_l v_l^{-1} < Z < m_l v_l^{-1}) \\ &= \frac{1}{2} + \frac{1}{2}\mathbb{P}\left(-\frac{1}{2\sigma}|\theta_1 - \theta_0|\sqrt{(t_L - t_0)} < Z < \frac{1}{2\sigma}|\theta_1 - \theta_0|\sqrt{(t_L - t_0)}\right). \end{aligned}$$

The above FNR and TNR rates and the maximal accuracy depend on three factors: the path length $t_L - t_0$, the scale of the noise σ (which affects the variance of the time series), and the distance $|\theta_1 - \theta_0|$ which depends on the dimension d . As either $\sqrt{t_L - t_0}$, $|\theta_1 - \theta_0|$, or σ^{-1} increases, the maximal accuracy increases. For example, when $\theta_0 = a_0[1, \dots, 1]^\top$, and $\theta_1 = a_1[1, \dots, 1]^\top$, $|\theta_1 - \theta_0| = d^{1/2}$, and the maximal accuracy is

$$ACC_{k_*} = 1 - \mathbb{P}(|Z| \geq \frac{1}{2\sigma}|a_1 - a_0|\sqrt{d(t_L - t_0)}).$$

These rates and the maximal accuracy do not depend on the temporal sampling frequency of the time series because the likelihood ratio is exact. However, the temporal sampling frequency will affect the accuracy when the likelihood ratio is approximated numerically as in (2.8), particularly for nonlinear time series; see the numerical examples in Section 5.

3.2 Ornstein-Uhlenbeck processes

Consider two \mathbb{R}^d -valued OU processes with parameters $\theta \in \{\theta_0, \theta_1\} \subset \mathbb{R}$:

$$dX_t = \theta X_t dt + \sigma dB_t \Leftrightarrow X_{t+\Delta t} = e^{\theta \Delta t} X_t + \sigma \int_t^{t+\Delta t} e^{\theta(t+\Delta t-r)} dB_r \quad (3.2)$$

for each $t > 0$, where $(B_t, t \geq 0)$ is an \mathbb{R}^d -valued standard Brownian motion and $\sigma > 0$ is a constant. Then, conditional on X_t and θ_i , the random variable $X_{t+\Delta t}$ has a distribution $\mathcal{N}\left(X_t e^{\theta_i \Delta t}, \frac{\sigma^2}{2\theta_i} (1 - e^{2\theta_i \Delta t}) I_d\right)$, and the transition probability density of this Markov process is

$$p(x_{t+\Delta t} | x_t, \theta_i) = (2\pi\sigma_{i,\Delta t}^2)^{-d/2} \exp\left(-\frac{1}{\sigma_{i,\Delta t}^2} \|x_{t+\Delta t} - e^{2\theta_i \Delta t} x_t\|^2\right)$$

with $\sigma_{i,\Delta t}^2 = \frac{\sigma^2}{2\theta_i} (1 - e^{2\theta_i \Delta t})$. Let $X_{t_0:t_L}$ be a discrete path with $t_l = l\Delta t$ for $0 \leq l \leq L$. By the Markov property, the logarithm probability density of $X_{t_0:t_L}$ conditional on θ_i is

$$\log p(X_{t_0:t_L} | \theta_i) = C - \frac{dL}{2} \log(\sigma_{i,\Delta t}^2) - \frac{1}{2\sigma_{i,\Delta t}^2} \sum_{l=0}^{L-1} \|X_{t_{l+1}} - e^{\theta_i \Delta t} X_{t_l}\|^2,$$

where C is a constant. Thus, the log-likelihood ratio in (2.6) is

$$l(X_{t_0:t_L} | \theta_1, \theta_0) = \frac{dL}{2} \log \left(\frac{\sigma_{0,\Delta t}^2}{\sigma_{1,\Delta t}^2} \right) + \frac{1}{2} \sum_{l=0}^{L-1} \left(\frac{\|X_{t_{l+1}} - e^{\theta_0 \Delta t} X_{t_l}\|^2}{\sigma_{0,\Delta t}^2} - \frac{\|X_{t_{l+1}} - e^{\theta_1 \Delta t} X_{t_l}\|^2}{\sigma_{1,\Delta t}^2} \right).$$

Let the rejection region be $R_k = \{X_{t_0:t_L} : l(X_{t_0:t_L} | \theta_1, \theta_0) > c_k\}$. Note that conditional on θ_0 , $N_l := \frac{1}{\sigma_{0,\Delta t}} (X_{t_{l+1}} - e^{\theta_0 \Delta t} X_{t_l})$ has a distribution $\mathcal{N}(0, I_d)$ for each l , and $X_{t_{l+1}} = e^{\theta_0 \Delta t} X_{t_l} + \sigma_{0,\Delta t} N_l$. Then, with $Y_l = (e^{\theta_1 \Delta t} - e^{\theta_0 \Delta t}) X_{t_l} + \sigma_{0,\Delta t} N_l$, the false positive rate (FNR) is

$$\begin{aligned} \alpha_k^0 &= \mathbb{P}(l(X_{t_0:t_L} | \theta_1, \theta_0) > c_k | \theta_0) \\ &= \mathbb{P} \left(\sum_{l=0}^{L-1} [\|N_l\|^2 - \sigma_{1,\Delta t}^{-2} \|Y_l\|^2] > 2c_k - dL \log \left(\frac{\sigma_{0,\Delta t}^2}{\sigma_{1,\Delta t}^2} \right) \right), \end{aligned}$$

with $N_l \sim \mathcal{N}(0, I_d)$. Similarly, denoting $Y'_l = (e^{\theta_0 \Delta t} - e^{\theta_1 \Delta t}) X_{t_l} + \sigma_{1,\Delta t} N_l$, we can compute the true negative rate (TNR)

$$\alpha_k^1 = \mathbb{P} \left(\sum_{l=0}^{L-1} [\sigma_{0,\Delta t}^{-2} \|Y'_l\|^2 - \|N_l\|^2] > 2c_k - dL \log \left(\frac{\sigma_{0,\Delta t}^2}{\sigma_{1,\Delta t}^2} \right) \right).$$

The optimal threshold $k = \arg \max_k \frac{1}{2}(1 - \alpha_k^0 + \alpha_k^1)$ depends on the various factors of the time series, so is the maximal accuracy. The numerical examples in Section 5 shows that the maximal accuracy increases as either d or L increases.

4 Benchmark design: example diffusions

We demonstrate the construction of diffusions for TSC benchmarking with three representative examples. In each example, the procedure is straightforward: first, we construct pairs of diffusions through varying the drifts. Then, we generate data from these diffusions, and compute the statistics of LRT, which will be used as a reference for the performance of the state-of-the-art machine learning TSC algorithms in the next section.

4.1 Diffusions with different drifts

Nonlinear diffusions can be constructed by varying the drifts $\{b_{\theta_i}\}_{i=0,1}$:

$$dX_t = b_{\theta_i}(X_t) dt + \sigma(X_t) dB_t, \quad b_{\theta_i}(X_t) = \sum_{j=1}^J \theta_{i,j} \phi_j(X_t), \quad (4.1)$$

where $X_t \in \mathbb{R}^d$, $\theta_i = (\theta_{i,1}, \dots, \theta_{i,J}) \in \mathbb{R}^J$ are the parameters, $\{\phi_j\}$ are *pre-specified* basis functions, and $(B_t, t \geq 0)$ is the standard Brownian motion in \mathbb{R}^d . Here the diffusion coefficient $\sigma(X_t)$ the same for the two diffusions, representing either a multiplicative noise (when it depends on the state) or an additive noise (when it is a constant). To test the optimality of the TSC algorithms, we consider three pairs of nonlinear diffusions: gradient systems with different potentials, SDEs with linear and nonlinear drifts, and high-dimensional interacting particle systems with different interaction kernels.

Example 4.1 (Different potentials) Consider two gradient systems with different potentials: a double-well potential $V_{\theta_0}(x) = \frac{1}{2}(|x|^2 - 1)^2$ and a single flat well-potential $V_{\theta_1}(x) = \frac{1}{4}|x|^4$:

$$dX_t = -\nabla V_{\theta_i}(X_t) dt + dB_t.$$

Writing them in the parametric form $V_{\theta_i}(X) = \sum_{j=0}^4 \theta_{i,j} |x|^j$ with $\theta_i = (\theta_{i,1}, \dots, \theta_{i,4})$, we have $\theta_0 = \frac{1}{4}(1, 0, -2, 0, 1)$ and $\theta_1 = (0, 0, 0, 0, \frac{1}{4})$.

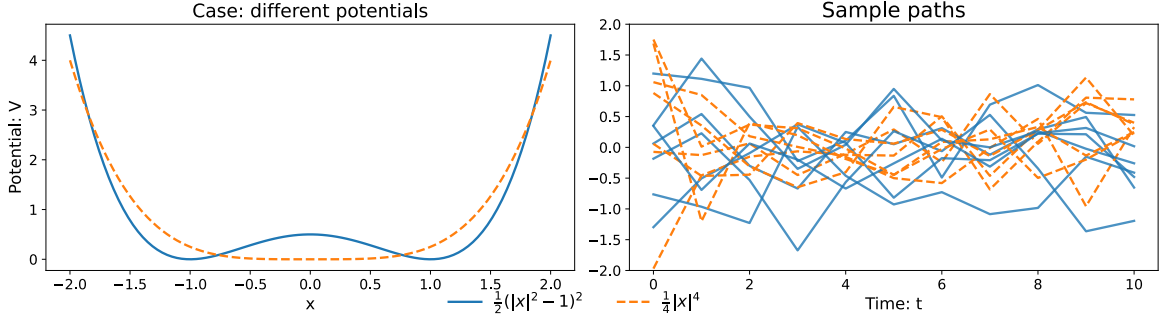


Figure 1: Different potentials in Example 4.1 and a few sample paths.

The double-well potential is a widely-used prototype model for systems with metastable states [26]. These two potentials are visually different, see Figure 1 (left). Each potential is confining and leads to an ergodic process with a stationary distribution. Thus, long sample paths that explore the full landscape of the potentials can distinguish the diffusions from the empirical densities. However, the short sample paths look similar and are difficult to distinguish, as shown in Figure 1 (right).

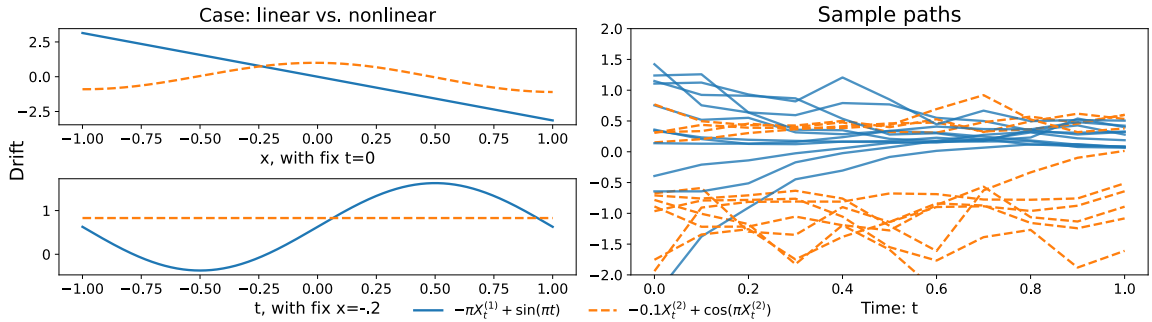


Figure 2: Linear v.s. nonlinear drifts in Example 4.2 and a few typical sample paths.

Example 4.2 (Linear v.s. nonlinear drifts) Consider two 1D Itô processes

$$dX_t = b_\theta(t, X_t)dt + X_t dB_t$$

with linear and nonlinear drifts $b_{\theta_0}(t, x) = -\pi x + \sin(\pi t)$ and $b_{\theta_1}(t, x) = -0.1x + \cos(\pi x)$, which can be written as $b_{\theta_i}(t, x) = \theta_{i,1}x + \theta_{i,2} \cos(\pi x) + \theta_{i,3} \sin(\pi t)$ with $\theta_0 = (-\pi, 0, 1)$ and $\theta_1 = (-0.1, 1, 0)$.

The two drift functions are clearly different, since $b_{\theta_0}(t, x)$ is linear in x and the other is nonlinear in x . Their sample paths are also visually different: the sample paths of b_{θ_0} are smoother than those of b_{θ_1} 's (they decay faster); see Figure 2. Thus, we expect that all TSC algorithms can achieve a high accuracy.

Example 4.3 (Interacting particles) Consider a system with N interacting agents with $X_t^j \in \mathbb{R}^d$ denoting the position or opinion the j -th agent at time t . Suppose that the agents interact with each other according to the following stochastic differential equation:

$$dX_t^j = \frac{1}{N} \sum_{i=1}^N \phi_\theta(\|X_t^j - X_t^i\|)(X_t^j - X_t^i) + \sigma dB_t^j,$$

where $\phi_\theta : \mathbb{R}^+ \rightarrow \mathbb{R}$ is the interaction kernel, $\{B_t^j, j = 1, \dots, N\}$ are independent standard Brownian motions, and $\sigma > 0$ is a scalar for the strength of the stochastic force. We will consider two types of

Table 2: Settings of the time series data in numerical tests.

Model	d	L	$t_L, \Delta t$
a) Constant drifts	1	{10, 20, 40, 80}	{1, 2, 4, 8}, 0.1
b) Different potentials	1	{20, 40, 80, 160}	{2, 4, 8, 16}, 0.1
c) OU processes	{1, 2, 4, 8}	20	2, 0.1
d) Interacting particles	{6, 12, 24, 48}	20	2, 0.1
e) Linear v.s. nonlinear	1	{5, 10, 20, 40}	1, $0.1 \times \{2, 1, 0.5, 0.25\}$
f) Interacting particles	24	{10, 20, 40, 80}	4, $0.1 \times \{4, 2, 1, 0.5\}$

* The models ‘‘Constant drifts’’ and ‘‘OU processes’’ are defined in Equations (3.1) and (3.2), and the models ‘‘Different potentials’’, ‘‘Interacting particles’’ and ‘‘Linear v.s. nonlinear’’ are defined in Examples 4.1–4.3.

interaction kernels (see Figure 3 (left))

$$\phi_{\theta_0}(r) = \begin{cases} 0.2, & r \in [0, \sqrt{2}), \\ 2, & r \in [\sqrt{2}, 2), \\ 0, & r \in [2, \infty). \end{cases} \quad \phi_{\theta_1}(r) = \begin{cases} 2, & r \in [0, \sqrt{2}), \\ 0.2, & r \in [\sqrt{2}, 2), \\ 0, & r \in [2, \infty). \end{cases}$$

This system leads to high-dimensional data, with $X_t = (X_t^1, \dots, X_t^N) \in \mathbb{R}^d$ with $d = d_1 N$. We will consider $d_1 = 2$ and $\sigma = 1$ with N varying to change the dimension of the system.

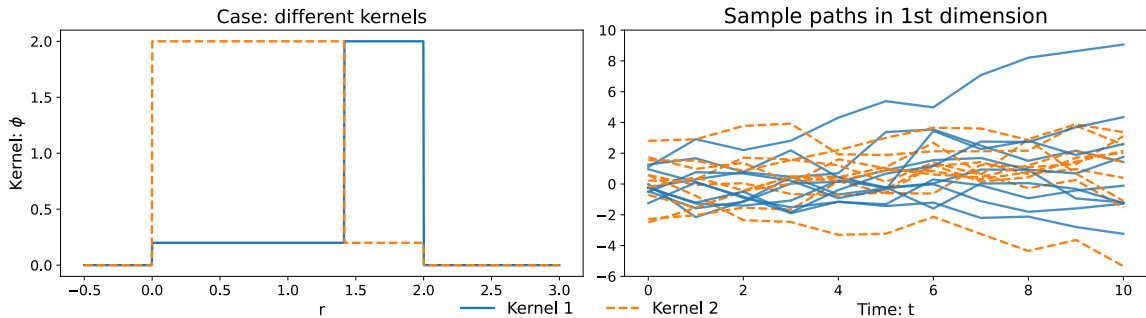


Figure 3: Interaction kernels in Example 4.3 and sample paths of the 1st dimension of an agent.

Such interacting particle systems have been increasingly studied because of their wide-range of applications in biology, engineering and social science (see e.g., [3, 17, 20, 23]). The difference between the two kernels is the strength of interaction between ‘‘far’’ and ‘‘close’’ neighbors: the kernel ϕ_{θ_1} makes the close neighbors interact stronger than those far away, whereas the kernel ϕ_{θ_0} makes the far neighbors interact stronger than those nearby. Then, the dynamics of the two systems are different, and it is shown in [23] that the more heterophilous kernel ϕ_{θ_0} enhances consensus when there is no stochastic force (i.e., the systems is deterministic). As a result, it is relatively easy to distinguish the two diffusions when the stochastic force is small. On the other hand, when the stochastic force is relatively large (e.g., $\sigma = 1$), the sample paths of the agents in the two systems are similar (Figure 3 (right)), making the classification a difficult task.

4.2 Data generation and the LRT benchmarks

The simulated diffusion processes allow us to test the dependence of classification performance on three parameters: path length in time t_L , the dimension d of the state, and temporal sampling frequency (by varying the time gap Δt). We test each of the three parameters with four values using two diffusion models, thus in total we generate 24 datasets in 6 cases with these parameters specified in Table 2.

In each dataset, the training data consists of $M = 2000$ sample trajectories $\{X_{t_0:t_L}^{(m)}\}_{m=1}^M$ of the pair of \mathbb{R}^d -valued diffusions with $\theta \in \{\theta_0, \theta_1\}$, 1000 paths for each of the pair. Here the time instances are

$t_l = l\Delta t$, and these data paths are downsampled from the solutions of the SDEs simulated by the Euler-Maruyama scheme with a fine time step $\delta = 0.01$. For example, the path with $\Delta t = 0.1$ makes an observation every 10 time steps from the fine simulated solution. The initial conditions $\{X_{t_0}^{(m)}\}_{m=1}^M$ are sampled from the standard normal distribution in \mathbb{R}^d . Each sample path is augmented with its time grid $t_0 : t_L$ with $t_0 = 0$.

For each dataset, we obtain two types of LRT benchmarks by computing the LRT in two ways: one using the fine paths and the other using the time series dataset, both compute the likelihood ratio using the Euler-Maruyama approximation in (2.8). Since there is no need of training, each classifier makes prediction directly on the whole dataset of M paths, and returns a single ROC curve, AUC and ACC_* , which will be used as references.

The LRT classifier using the fine solution is called “*LRT hidden truth*”, and it provides the optimal classification rates by the Neyman–Person lemma (see Theorem 2.2). The other LRT classifier using the training data is called “*LRT numerical*”. It does not use the hidden fine path, but it uses the diffusion model information that are not used by the TSC algorithms. It has a relatively large numerical error when the SDE is nonlinear, particularly when the observation time interval Δt is much larger than the simulation time step δ . Thus, it provides a lower baseline for the TSC algorithms. The two LRT benchmarks are the same when the time series are samples of a Gaussian process from a linear SDE, e.g., the cases of Brownian motions with constant drifts and OU processes.

4.3 Discussions on benchmark design

The LRT benchmark design has two main components: selection of the diffusion processes and generation of simulated data. In addition to the examples in Table 2, there is a large variety of diffusion processes from stochastic differential equations in the form of (4.1), such as gradient systems and stochastic Hamiltonian systems [25, 26]. The two diffusions should have the same diffusion coefficient, so that the likelihood ratio can be computed based on the Girsanov theorem.

To generate simulated data, we recommend using the Euler-Maruyama scheme so that the likelihood ratio of the fine trajectory is exact. The time series data are downsampled from the fine trajectories. It is helpful to compute two LRT benchmarks, one using the fine trajectories and the other using the downsampled data, to provide an optimality benchmark and a lower baseline benchmark. In particular, the optimality benchmark can detect the overfitting of a TSC algorithm in the training stage.

Four parameters can be tuned to adjust the theoretical classification accuracy: the time length of paths, the dimension, the temporal sampling frequency, and the strength of the driving noise (as suggested by the analysis in Section 3). The time length of paths and the dimension affect the effective sample size and hence the classification rates. The temporal sampling frequency affects the LRT baseline but it may have a limited effect on the model-agnostic TSC algorithms. At last, a large noise dims the signal from the drifts, thus lowering the accuracy of classification.

5 Benchmarking random forest, ROCKET and ResNet

5.1 Random forest, ROCKET, and ResNet

We benchmark three scalable TSC methods: random forest [4], ROCKET [8], and ResNet [30]. They have been shown to be state-of-the-art in recent review papers [2, 13, 29]. In particular, the most recent review [29] compares 11 multivariate time series classifiers that are top-performers in [2, 13], including both non-deep learning methods (including ROCKET and HIVE-COTE (Hierarchical Vote Collective Of Transformation-based Ensembles) [19]) and deep learning methods (including ResNet and InceptionTime [14]), using 26 UEA archive datasets [1]. The recommended method is ROCKET due to its high overall accuracy and remarkably fast training time.

Random Forest. The random forest (RF) is an ensemble learning technique that combines a large number of decision trees, and it is applicable to both classification and regression. The original RF described by [4] is a classifier consisting of a collection of tree-structured classifiers $\{f(\mathbf{x}, \beta_i)\}_{i=1}^{n_T}$ with

independent identically distributed parameters β_i and each tree casts a unit vote for the input \mathbf{x} to be in a class. These votes lead to a function $f_\beta(\mathbf{x}) = \frac{1}{n_T} \sum_{i=1}^{n_T} f(\mathbf{x}, \beta_i)$ approximating the probability of \mathbf{x} being in the class (i.e., the probability $\mathbb{P}(\theta = \theta_1 | \mathbf{x})$ of \mathbf{x} in the class θ_1 in our notation in Section 2.1). The classifier function with a threshold k is $F(\mathbf{x}, k)$ as in (2.2). It is user-friendly with only a few parameters easy to tune to achieve robust performance, and its performance is comparable to other classifiers such as discriminate analysis, support vector machine and neural networks [18, 28].

We use the default HalvingRandomSearchCV strategy in scikit-learn [27] to search for parameter values in the ranges listed below.

	# of trees	max depth	max features	min SS	bootstrap
RF	{10 : 100}	{3, None}	{1 : 11}	{2 : 11}	{True, False} ,

where “min SS” represents minimal samples split, and the quality of a split is measured by the Gini index. Note that number of trees is medium so as to have a comparable computational cost with other methods.

ResNet. The deep residual network (ResNet) for time series classification [30] is a network with three consecutive blocks, each comprised of three convolutional layers, followed by a global average pooling layer and a final dense layer with softmax activation function. The major characteristic is that the three consecutive blocks are connected by residual “shortcut” connections, enabling the flow of the gradient directly through them, thus reducing the vanishing gradient effect [12]. It outperforms other deep learning time series classifiers in [13], especially for univariate datasets [30].

We maintain all hyper-parameter settings from [13].

Structure	layers	activate	normalize	residue	dropout
ResNet	9+2	ReLU	batch	between blocks	none .

There are nine convolution layers in the three blocks, each with the ReLU activation function that is preceded by a batch normalization operation. The number of filters in each convolution layer is 64 in the first block; while the number is 128 for the second and third blocks. In each residual block, the kernel size (or the length of the filter) is set to 8, 5 and 3 respectively for the first, second and third convolution. The optimization settings is also similar to [13]:

Training	optimizer	rlr	epochs	batch	learning rate	weight decay
ResNet	Adam	yes	150	16	0.001	0.0 ,

where “rlr” means that the learning rate is reduced by a half if the model’s training loss has not improved for 5 consecutive epochs with a minimum learning rate set to 0.0001. Here we set the epochs to 150 to have a computational cost comparable with other methods while maintaining accuracy.

ROCKET. The ROCKET (Random Convolutional Kernel Transform) [8] is the current state-of-the-art multivariate time series classifier [29]. It uses random transformations followed by a linear classifier (ridge regression or logistic regression). In the transformation part, a large number of random convolution kernels are applied to each time series, each kernel producing a feature map. From each of these feature maps, two features are extracted: the maximal value and the proportion of positive value (ppv). Thus, each random kernel extracts two features from each time series. The linear classifier then makes classification based on these features.

We keep the default setting for ROCKET in the *sktime* repository¹, and we use the ridge regression (the parameter regularization strength α is searched by the build-in function RidgeCV). The randomness comes from the kernel’s parameters: length, weights, bias, dilation, and padding:

¹https://github.com/alan-turing-institute/sktime/blob/master/sktime/transformers/series_as_features/rocket.py.

Kernel	length	weight	dilation	padding or not	stride
ROCKET	{7, 9, 11}	$\mathcal{N}(0, 1)$	$[2^x]$	equal probability	1.

Here $x \sim \mathcal{N}(0, A)$ with $A = \log_2 \frac{l_{\text{input}} - 1}{l_{\text{kernel}} - 1}$, where l_{input} and l_{kernel} are the lengths of the time series and the kernel. The number of kernels is set to 10000, resulting in 20000 features for each time series.

5.2 ROC curves in a typical test

We compare the performance of these TSC algorithms with the LRT benchmarks in three statistics: the ROC curve in a typical test, the box-and-whisker plots of AUC (area under the ROC curve) and the optimal accuracy (denoted by ACC_*) in 40 different runs. In each run, we train the algorithms using randomly sampled 3/4 of the data paths and use the rest 1/4 of the data for prediction test. Thus, each algorithm is trained using $M_{\text{training}} = \frac{3}{4}M = 1500$ sample paths and the rates in prediction are computed using $\frac{1}{4}M = 500$ sample paths. By Lemma 2.1, each prediction rate has a standard deviation at the scale of $\frac{0.5}{\sqrt{500}} = 0.02$. Thus, two algorithms perform similarly if the difference between their rates are within the sampling error of 0.04 (in two standard deviations).

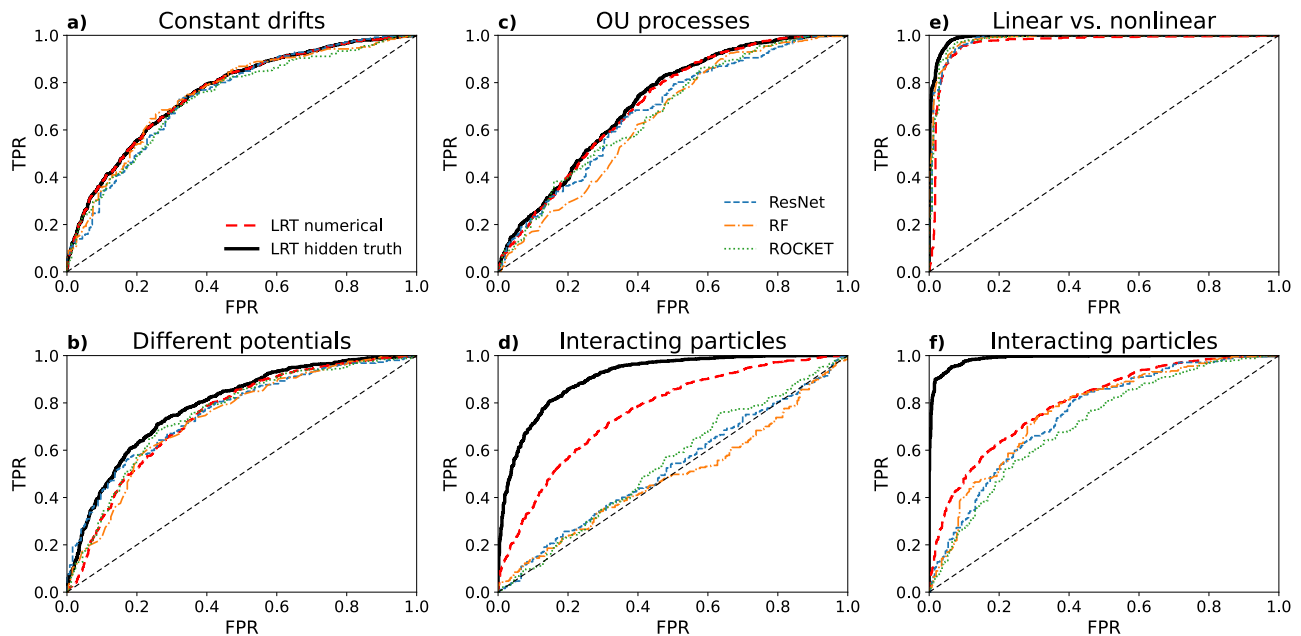


Figure 4: ROC curves in a typical test in each of the 6 cases (each using the first of the settings in Table 2). The three algorithms achieve the optimal LRT in Case (a), and they are in-between the two LRT benchmarks in Cases (b,e). They are suboptimal in comparison with the “LRT numerical” in Cases (c,f) and they have unsuccessful classification in Case (d).

Figure 4 shows the ROC curves in a typical test in each of the 6 cases in the Table 2. Each case uses its first of the four settings, e.g., the constant drifts dataset has $(t_L, d, \Delta t) = (1, 1, 0.1)$, the dataset for different potentials has $(t_L, d, \Delta t) = (20, 1, 0.1)$, and the OU processes dataset in Case (c) has $(t_L, d, \Delta t) = (2, 1, 0.1)$. The datasets for the interacting particles in Cases (d) and (f) have $(t_L, d, \Delta t) = (2, 6, 0.1)$ and $(t_L, d, \Delta t) = (4, 24, 0.4)$, respectively.

For univariate time series in the Cases (a,b,e), the three algorithms either reach or are close to the optimality benchmark by the LRT. They achieve the optimal benchmark of “LRT hidden truth” for the Brownian motion with constant drifts. They are nearly optimal with curves in-between the two LRT benchmarks in distinguishing the diffusions with different potentials and the diffusions with linear or nonlinear drifts.

For the univariate time series in Case (c) and the multivariate time series in Case (f), the three

algorithms are suboptimal as their ROC curves are below the “LRT numerical” with $\Delta t = 0.1$ and $\Delta t = 0.4$, respectively.

However, the three algorithms have unsuccessful classifications in Case (d), which is the multivariate interacting particles with $(t_L, d, \Delta t) = (2, 6, 0.1)$. Their ROC curves are around the diagonal line. In contrast, the benchmark of “LRT numerical” with $\Delta t = 0.1$ has a reasonable ROC curve and the ROC curve of “LRT hidden truth” is much higher. Thus, the data has rich information for the classification, and there is room for improvements in these algorithms. We note that the LRT makes use of the model information while the three algorithms are model agnostic. Hence, the success of the “LRT numerical” shows the importance of model information in the classification of nonlinear multivariate time series.

In particular, the contrast between the failure in Case (c) and the success in Case (f) invites further examination of the factors that affect the performance of the algorithms. Note that both Case (d) and Case (f) are for the interacting particle systems, and they are different only at $(t_L, d, \Delta t) = (2, 6, 0.1)$ and $(t_L, d, \Delta t) = (4, 24, 0.4)$. Thus, in the next section, we examine the algorithms with varying $(t_L, d, \Delta t)$. We will also examine the dependence of the classification accuracy on randomness and training sample size (Figure 8). Additionally, a single test is insufficient to draw a conclusive comparison because of the randomness in the data; hence, we run multiple tests in each setting and report the statistics of AUC and ACC in the next section to benchmark the optimality.

Also, one may notice that the random forest lags behind the other two in Case (c) and the ROCKET lags behind in Case (f), both with rate differences larger than two theoretical standard deviations (0.04). Such differences are due to the randomness in the data in this single test, the statistics from multiple tests in the next section show that no method is superior in all settings.

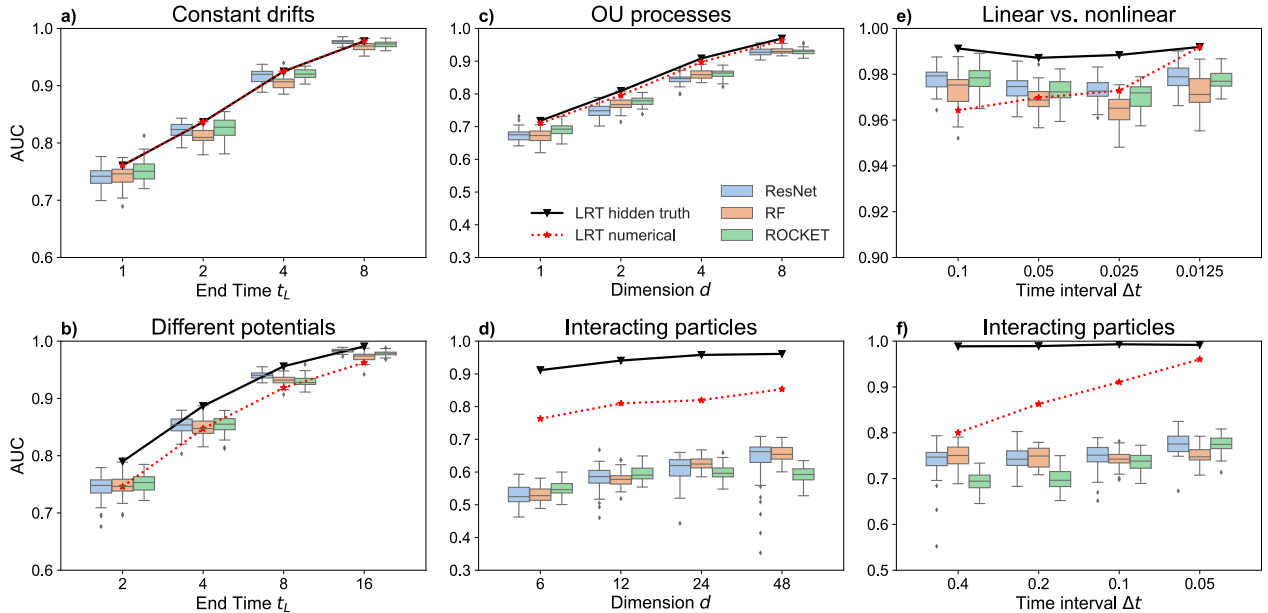


Figure 5: AUC for the 6 Cases with varying $(t_L, d, \Delta t)$ in Table 2. All three algorithms perform similarly: they reach the optimal LRT for Gaussian processes in Case (a), and they are near optimal in Cases (b,e), suboptimal in the Cases (c,f), and are unsuccessful in Case (d).

5.3 Optimality benchmarking in AUC and maximal accuracy

We benchmark the optimality of a classifier by examining the statistics of the AUC and optimal accuracy (ACC_*) in 40 independent simulations for each 4 settings of the 6 cases in Table 2. We present the box-and-whisker plots (the minimum, the maximum, the sample median, and the first and third quartiles, and the outliers) of the AUC and ACC_* , which reflects the randomness in the classifications.

Recall that the “LRT hidden truth” provides an upper bound of optimality and the “LRT numerical”

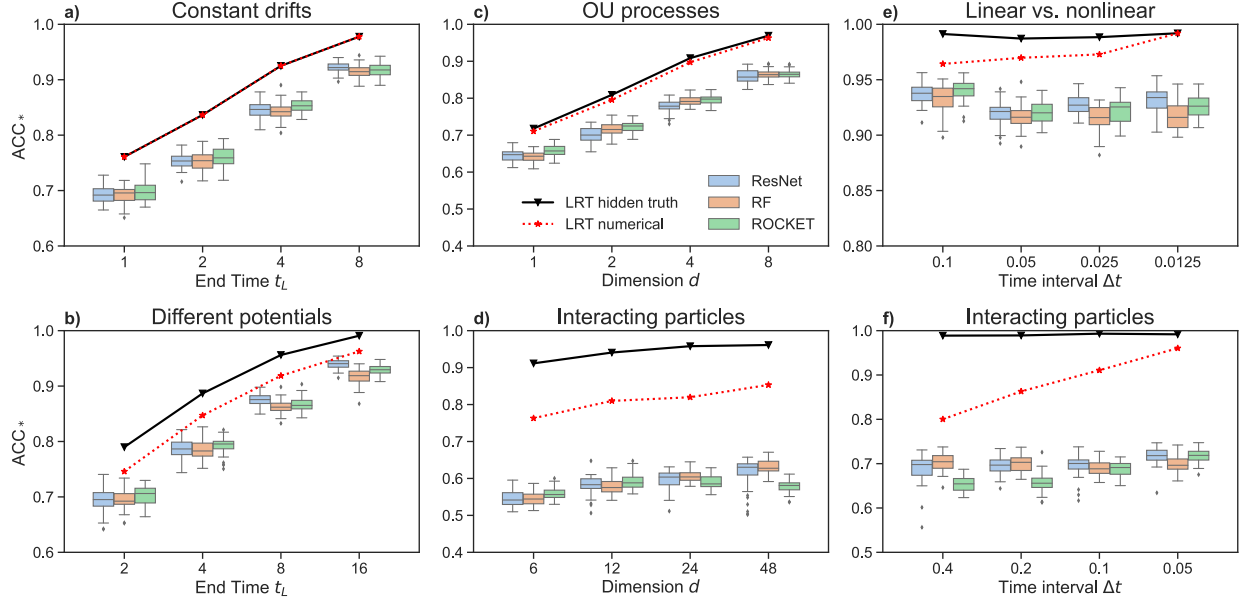


Figure 6: Maximal accuracy (ACC_*) with varying $(t_L, d, \Delta t)$ in Table 2. All three algorithms are suboptimal in comparison with the LRT benchmarks.

provides a low baseline for them. Thus, a classifier achieves the optimality for the Gaussian processes if its AUC and ACC_* concentrate around the “LRT hidden truth”. A classifier is *suboptimal* if its AUC or ACC_* is below the baseline of “LRT numerical”, particularly when the temporal sampling frequency of observation is relatively low. We say it is *near optimal* when its statistics lie in between the benchmark lines, particularly when the two lines are close.

Figure 5 shows the statistics of the AUCs in the six cases with varying path time lengths t_L , dimension d and temporal sampling frequency (through Δt). In the case of univariate time series data, the three algorithms achieve the optimality represented by the LRT hidden truth for the Gaussian process in Case (a), and they are near optimal for nonlinear time series in Cases (b,e). They are unsuccessful in all settings in Case (d), the high-dimensional interacting particle system with short sample paths, and they are suboptimal in Cases (c,f). These results agree with those from the ROC curves.

Additionally, we notice two patterns. (i) The AUC increases as the path length in time t_L or the dimension d increases, which can be clearly seen in Cases (a,b,e,d). (ii) The AUC of the three methods is not sensitive to the temporal sampling frequency of observation, because Cases (e,f) show that the AUC changes insignificantly as Δt refines. Note that the slopes of the LRT benchmarks in Case (c) are much steeper than those in Case (d). This is because the entries of the OU processes are independent, whereas the entries of the interacting particles are correlated through the interactions. Thus, the increment of AUC is due to the increased effective sample size through either d or t_L . Such patterns of AUC’s dependence on path length and sample size will be further examined in Figure 8 for the interacting particle systems.

Figure 6 shows the statistics of the maximal accuracy (ACC_*) in the six cases. It turns out that all three algorithms have smaller maximal accuracy than the benchmark of “LRT numerical” (not to mention the “LRT hidden truth”). Thus, there is a room for their improvements. On the other hand, the two patterns on the dependence of (t_L, d, Δ) are similar to those observed in AUC in Figure 5.

Figure 7 shows the statistics of the computation time in training of these tests. The computation is carried out on a node of 3.0GHz Intel Cascade Lake 6248R with 48cores, 192GB RAM 1TB NMVe local SSD. The figure shows that the random forest (RF) has a controlled computation time for all cases. The computation time of either ResNet or the ROCKET increases in the path length ($L = \frac{t_L}{\Delta t}$) as shown in Cases (a,b,e,f), and is not sensitive to the dimension d as Cases (c,d) suggests. The ResNet has the largest computation time in most cases. The LRT benchmarks are not shown here because their

computation time is negligible (since they only involve the evaluations of the likelihood ratio).

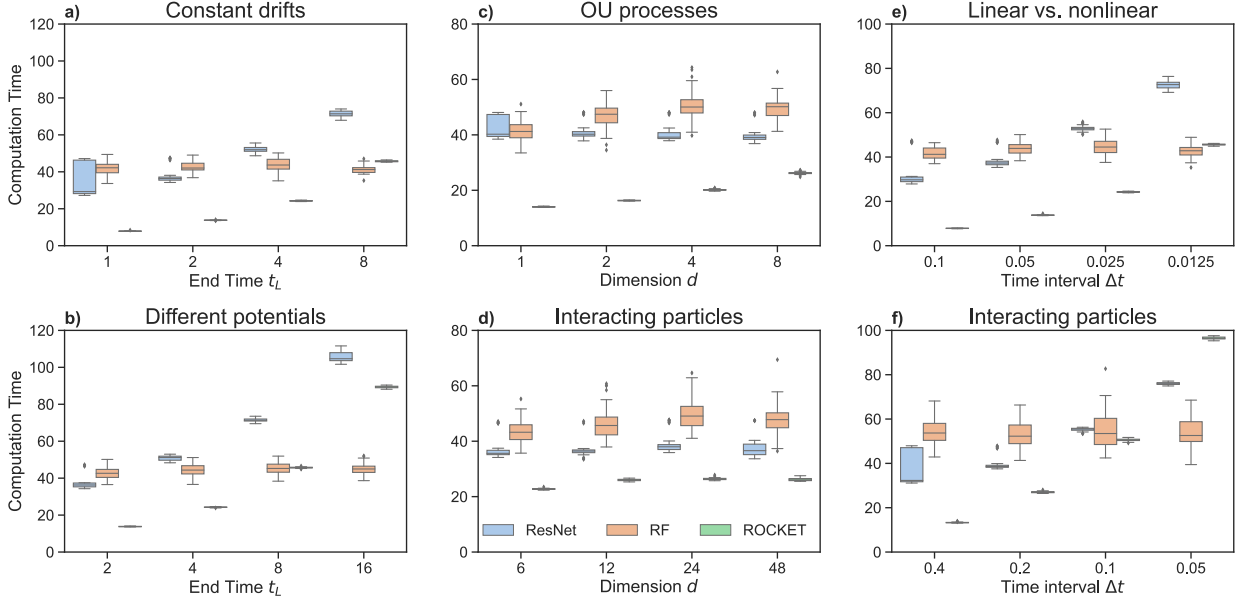


Figure 7: Computation time (in seconds) for the tests with varying $(t_L, d, \Delta t)$ in Table 2. The random forest (RF) has a controlled computation time. The ResNet and the ROCKET have computation times increasing with the path length ($L = \frac{t_L}{\Delta t}$) in Cases (a,b,e,f), and not sensitive to the dimension d in Cases (c,d). The ROCKET has the smallest computation time when the length L is not large.

Figure 8 further examines the dependence of the classification performance on the path length t_L , the randomness (in terms of σ), and the training sample size in Cases **a)**– **c)**, respectively, for the interacting particle systems. These cases show that the AUC of each method increases when either the path time length increases, or the randomness decreases, or the training sample size increases. In particular, Case **c)** shows that a growing training sample size can significantly improve the AUC of each algorithm; yet, with a training sample size of 4000, their AUCs are far below the LRT benchmarks (which do not need to be trained by taking into account the model information). Additionally, we note that the variation of each algorithm reduces as the sample size increases, indicating that the learning error decays in the sample size. The ResNet has the largest variation among the three algorithms, but its performance improves the most when the sample size increases.

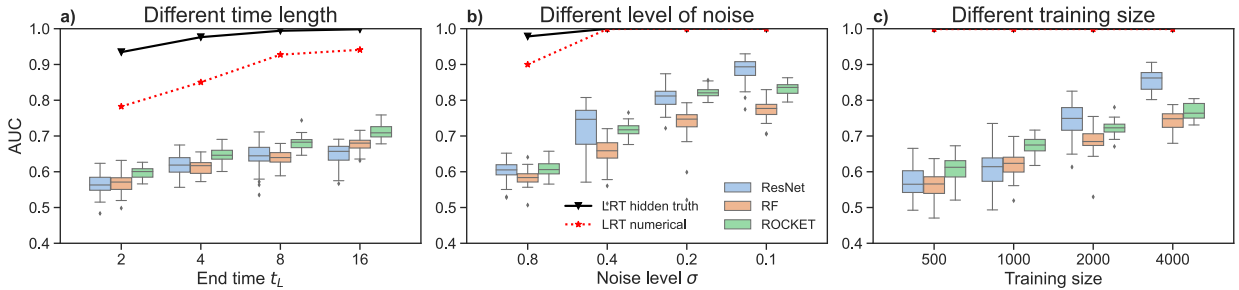


Figure 8: AUC for interacting particles in three additional settings: **a)** $t_L \in \{2, 4, 8, 16\}$, $\sigma = 1$ and $M = 2000$; **b)** $\sigma \in \{0.8, 0.4, 0.2, 0.1\}$, $t_L = 2$ and $M = 2000$; **c)** $M_{\text{training}} \in \{500, 1000, 2000, 4000\}$, $(t_L, \sigma) = (2, 0.4)$. In all cases, the test sample size is 500 and $(d, \Delta t) = (12, 0.1)$.

In summary, the LRT benchmarks show that all three algorithms can achieve the LRT optimal AUC for univariate time series and multivariate Gaussian processes. However, these model-agnostic algorithms are suboptimal in classifying nonlinear multivariate time series from high-dimensional stochastic interacting particle systems. Also, the maximal accuracy of each algorithm is below the LRT benchmark

in all cases, suggesting room for improvement. Importantly, the LRT benchmarks focuses on the

5.4 Discussion

The performance of a classifier depends on multiple factors, including the design of the classifier, the training data size, and the properties of the time series (such as its dimension, randomness, time length, and temporal sampling frequency). The LRT benchmarks help separate these factors so that we can better examine the classifier.

- The optimal classification accuracy is determined by the distribution of the underlying discrete-time stochastic process from which the time series is sampled. This distribution varies in the properties of the time series, such as its dimension, randomness, time length, and temporal sampling frequency. The optimal classification accuracy increases when the dimension or the time length increases or the randomness reduces, but it is not sensitive to the temporal sampling frequency. Thus, in data collection in practice, it is more helpful to collect data for a longer time rather than a higher sampling frequency.
- The performance of a classifier is bounded above by the optimal classification accuracy, and it is limited by its structure and the training data size. In particular, the training data size can significantly affect the classifier’s accuracy. The size needed to achieve a prescribed level of accuracy increases with the uncertainty in the distribution of the time series, as well as the structure of the classifier. A classifier with a larger complexity requires more data to train. The ResNet, which uses neural networks, improves the most from an enlarging sample size compared to the random forest and ROCKET, which use simpler designs. We would expect a bias-variance trade-off for which one can select the degree of complexity of the algorithms adaptive to data size, and we leave this as future work.
- The model-agnostic TSC algorithms do not use the model information and rely on data to learn the classifier function; thus, they require a large amount of training data. In contrast, the LRT relies on the model information and does not need to be trained. Therefore, we would expect a TSC algorithm using the model information can significantly increase the performance while reducing the training data size.

6 Conclusion

We have shown that the likelihood ratio test (LRT) distinguishing diffusion processes provides ideal optimality benchmarks for time series classification (TSC) algorithms. The benchmarking is computationally scalable and is flexible in design for generating linear or nonlinear time series to reflect the specific characteristics of real-world applications.

Numerical tests show that the three state-of-the-art TSC algorithms, random forest, ResNet, and ROCKET, can achieve the optimal benchmark for univariate time series and multivariate Gaussian processes. However, these model-agnostic methods are suboptimal compared to the model-aware LRT in classifying high-dimensional nonlinear non-Gaussian processes.

The LRT benchmarks also show that the classification accuracy increases with either the time length or the time series dimension. However, the classification accuracy is less sensitive to the frequency of the observations. Thus, in data collection, it is more helpful to collect data for a longer time rather than a higher sampling frequency.

In future work, we propose to quantitatively analyze the dependence on these factors in terms of the effective sample size, the bias-variance trade-off in the training of the algorithms, and the incorporation of model information into the algorithms.

A Appendix

A.1 Itô-diffusion and the Girsanov theorem

Theorem A.1 (Girsanov Theorem) *Let P_{θ_i} be the probability measure induced by the solution of the SDEs (4.1) for $t \in [t_0, T]$, and let P_0 be the law of the respective drift-less process. Suppose that the drifts $\{b_{\theta_i}\}$ and the diffusion $\Sigma = \sigma\sigma'$ fulfill the Novikov condition*

$$\mathbb{E}_{P_{\theta_i}} \left[\exp \left(\frac{1}{2} \int_{t_0}^T b_{\theta_i}(X_t, t)^\top \Sigma^{-1} b_{\theta_i}(X_t, t) dt \right) \right] < \infty.$$

Then, P_{θ_i} and P_0 are equivalent measures with Radon-Nikodym derivative given by

$$\frac{dP_{\theta_i}}{dP_0}(X_{[t_0, s]}) = \exp \left(- \int_{t_0}^s b_{\theta_i}^\top \Sigma^{-1} dX_t + \frac{1}{2} \int_{t_0}^s [b_{\theta_i}^\top \Sigma^{-1} b_{\theta_i}](X_t) dt \right)$$

for all $s \in [t_0, t]$ and $X_{[t_0, s]} = (X_t)_{t \in [t_0, s]}$. In particular, the likelihood ratio between P_{θ_1} and P_{θ_0} is

$$\frac{dP_{\theta_1}}{dP_{\theta_0}}(X_{[t_0, s]}) = \exp \left(- \int_{t_0}^s [b_{\theta_1} - b_{\theta_0}]^\top \Sigma^{-1} dX_t + \frac{1}{2} \int_{t_0}^s [b_{\theta_1}^\top \Sigma^{-1} b_{\theta_1} - b_{\theta_0}^\top \Sigma^{-1} b_{\theta_0}](X_t) dt \right).$$

The proof of Theorem A.1 can be found in [16, Chapter 3.5] or [25, Section 8.6].

A.2 Sampling error in the classification rates

Proof of Lemma 2.1. Fix a threshold k , the classifier defines a random variable $\xi = \xi(\mathbf{x}) = F(\mathbf{x}, k)$. Then, conditional on θ_i with $i \in \{0, 1\}$, the random variable ξ has a Bernoulli distribution that takes the value 1 with a probability α_k^i . In particular, the test samples $\{\mathbf{x}_j\}_{j=1}^m$ lead to samples $\{\xi_j\}_{j=1}^m$ of ξ , and the empirical approximations of the FNR and TNR by these samples are

$$\hat{\alpha}_{k,m}^i = \frac{1}{m} \sum_{j=1}^m \xi_j, \text{ conditional on } \theta_i, i = 0, 1.$$

Therefore, by the Central Limit Theorem, the empirical estimators converge in distribution

$$\sqrt{m}[\hat{\alpha}_{k,m}^i - \alpha_k^i] \rightarrow \mathcal{N}(0, \sigma_{\xi,i}^2), \text{ where } \sigma_{\xi,i}^2 = \alpha_k^i(1 - \alpha_k^i)$$

as $m \rightarrow \infty$ for each $i = 0, 1$. Also, the Hoeffding's inequality (see e.g., [6, 7, 11]) implies that for any $\epsilon > 0$,

$$\mathbb{P}(|\hat{\alpha}_{k,m}^i - \alpha_k^i| > \epsilon) \leq 2e^{-\frac{m\epsilon^2}{2}},$$

which provides a non-asymptotic bound for each $m > 0$. ■

A.3 Hypothesis testing and the Neyman-Pearson lemma

Here we briefly review the hypothesis testing inference method in statistics [5, Chapter 8]. Recall that a hypothesis test is a rule that specifies for which sample values the decision is made to accept a hypothesis H_0 as true, and reject the complement hypothesis H_1 . We assume that the family of distributions of the samples are parametrized by $\theta \in \Theta$, where Θ is the entire parameter space. We denote that the null alternative hypotheses by $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, respectively, where Θ_0 is a subset Θ . The binary classification is therefore a hypothesis testing with $\Theta = \{\theta_0, \theta_1\}$ and $\Theta_0 = \{\theta_0\}$.

The likelihood ratio test is as widely applicable as maximum likelihood estimation. When there are two parameters, it is defined as follows.

Definition A.2 (Likelihood Ratio Test.) Let the probability density function (or probability mass function) corresponding to θ_i be $f(x | \theta_i)$ for $i = 0, 1$. The likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is:

$$\lambda(x) = \frac{f(x | \theta_1)}{f(x | \theta_0)}.$$

A likelihood ratio test (LRT) is any test that determines the rejection region for H_0 by $\lambda(x)$.

The LRT in (2.5) determines the rejection region using the log-likelihood $l(x) = \log \lambda(x)$. The rejection region with threshold $k \in (0, 1)$ is equivalent to

$$R_k^{LRT} = \{\mathbf{x} : \frac{1}{\lambda(x) + 1} > k\} = \{\mathbf{x} : \lambda(x) > \frac{k}{1 - k}\}.$$

The reject region is selected to control the probability of falsely rejecting H_0 , i.e., false negative rate (FNR). Meanwhile, it is also desirable to control the false positive rate (FPR), e.g., reduce the possibility of false alarms.

The hypothesis tests are evaluated by the probabilities of making mistakes. A strategy to compare hypothesis tests is to control the FNR in a class and compare the FPR. The power function provides a tool to define the class.

Definition A.3 (Power function, size α test.) The power function of the hypothesis test with a rejection region R and sample x is the probability $\beta(\theta) = \mathbb{P}(x \in R | \theta)$ as a function of $\theta \in \Theta$. A test with power function β is a size α test if $\sup_{\Theta_0} \beta(\theta) = \alpha$; a test with power function β is a level α test if $\sup_{\Theta_0} \beta(\theta) \leq \alpha$.

An ideal hypothesis test would have a power function $\beta(\theta) = 0$ for all $\theta \in \Theta_0$ and $\beta(\theta) = 1$ for all $\theta \in \Theta_0^c$. Thus, a good test would have $\beta(\theta)$ close to 0 for all $\theta \in \Theta_0$ and $\beta(\theta)$ near 1 for all $\theta \in \Theta_0^c$.

Next, we define the uniformly most powerful test as the test with the smallest FPR uniformly for all $\theta \in \Theta_0^c$ in the class of tests with a controlled FNR.

Definition A.4 (Uniformly Most Powerful (UMP) Test) Let \mathcal{C} be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class \mathcal{C} , with power function $\beta(\theta)$, is a uniformly most powerful (UMP) class \mathcal{C} test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every function $\beta'(\theta)$ that is a power function of a test in class \mathcal{C} .

The Neyman-Pearson lemma shows that a LRT with a rejection region $R = \{x : \frac{f(x|\theta_1)}{f(x|\theta_0)} > c\}$ is a UMP test when $\Theta_0 = \{\theta_0\}$ and $\Theta_0^c = \{\theta_1\}$ for any $c \in (0, \infty)$ such that $\mathbb{P}(\{x : \frac{f(x|\theta_1)}{f(x|\theta_0)} = c\}) = 0$.

Theorem A.5 (Neyman-Pearson Lemma) Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the probability density function (or probability mass function) corresponding to θ_i is $f(x | \theta_i)$ for $i = 0, 1$, using a test with rejection region R that satisfies

$$\begin{cases} x \in R, & \text{if } f(x | \theta_1) > cf(x | \theta_0) \\ x \in R^c, & \text{if } f(x | \theta_1) < cf(x | \theta_0) \end{cases} \quad (\text{A.1})$$

for some $c > 0$, and

$$\alpha = P_{\theta_0}(X \in R) \quad (\text{A.2})$$

Then:

1. (Sufficiency) Any test that satisfies (A.1) and (A.2) is a UMP level α test.
2. (Necessity) If there exists a test satisfying (A.1) and (A.2) with $c > 0$, then every UMP level α test is a size α test (satisfies (A.2)) and every UMP level α test satisfies (A.1) except perhaps on a set A satisfying $P_{\theta_0}(X \in A) = P_{\theta_1}(X \in A) = 0$.

Acknowledgments F.L. and Y.K. are partially supported by the grant DE-SC0021361. The work of F.L. is partially funded by the Johns Hopkins University Catalyst Award and FA9550-20-1-0288. The computation is carried out on the clusters of the Maryland Advanced Research Computing Center. FL would like to thank Professors Geoff Webb and Xingjie Li for helpful comments on the paper.

References

- [1] A. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.
- [2] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.
- [3] N. Bell, Y. Yu, and P. J. Mucha. Particle-based simulation of granular materials. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation - SCA '05*, page 77, Los Angeles, California, 2005. ACM Press.
- [4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [5] G. Casella and R. Berger. *Statistical Inference*. Pacific Grove, CA, Thomson Learning, Australia, 2002.
- [6] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.
- [7] F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, Cambridge, 2007.
- [8] A. Dempster, F. Petitjean, and G. I. Webb. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5):1454–1495, 2020.
- [9] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, New York, 2013.
- [10] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [11] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, UK, 2015.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [13] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.
- [14] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.
- [15] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, New York, 2013.
- [16] I. Karatzas and S. E. Shreve. Brownian motion. In *Brownian Motion and Stochastic Calculus*. Springer, New York, second edition, 1998.
- [17] U. Krause. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 2000:227–236, 2000.
- [18] A. Liaw, M. Wiener, et al. Classification and regression by random forest. *R news*, 2(3):18–22, 2002.
- [19] J. Lines, S. Taylor, and A. Bagnall. Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 1041–1046. IEEE, 2016.
- [20] F. Lu, M. Maggioni, and S. Tang. Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *Foundations of Computational Mathematics*, pages 1–55, 2021.

- [21] T. Lyons. Rough paths, Signatures and the modelling of functions on streams, 2014.
- [22] J. Morrill, A. Fermanian, P. Kidger, and T. Lyons. A Generalised Signature Method for Multivariate Time Series Feature Extraction, 2021.
- [23] S. Motsch and E. Tadmor. Heterophilous Dynamics Enhances Consensus. *SIAM Rev*, 56(4):577 – 621, 2014.
- [24] J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.
- [25] B. Øksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, New York, 6th edition, 2013.
- [26] G. A. Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, New York, 2014.
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [28] P. Probst and A.-L. Boulesteix. To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, 18(1):6673–6690, 2017.
- [29] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021.
- [30] Z. Wang, W. Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.