Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

On Consistency of Signature Using Lasso

Xin Guo; , Binnan Wang; , Ruixun Zhang; , Chaoyi Zhao

informs
https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

On Consistency of Signature Using Lasso

Xin Guo,[a] Binnan Wang,[b] Ruixun Zhang,[b,c,d,e,*] Chaoyi Zhao[f,*]

[a] Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, California 94720; [b] School of Mathematical Sciences, Peking University, Beijing 100871, China; [c] Center for Statistical Science, Peking University, Beijing 100871, China; [d] Laboratory for Mathematical Economics and Quantitative Finance, Peking University, Beijing 100871, China; [e] National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing 100871, China; [f] Sloan School of Management and Laboratory for Financial Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

*Corresponding authors

Contact: xinguo@berkeley.edu (XG); wangbinnan@stu.pku.edu.cn (BW); zhangruixun@pku.edu.cn, https://orcid.org/0000-0002-7670-8393 (RZ); cy_zhao@mit.edu, https://orcid.org/0000-0002-5921-3612 (CZ)

Received: June 23, 2024
Revised: November 27, 2024; February 26, 2025
Accepted: March 13, 2025
Published Online in Articles in Advance: May 5, 2025

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2024.1133

Abstract. Signatures are iterated path integrals of continuous and discrete-time processes, and their universal nonlinearity linearizes the problem of feature selection in time series data analysis. This paper studies the consistency of signature using Lasso regression, both theoretically and numerically. We establish conditions under which the Lasso regression is consistent both asymptotically and in finite sample. Furthermore, we show that the Lasso regression is more consistent with the Itô signature for time series and processes that are closer to the Brownian motion and with weaker interdimensional correlations, whereas it is more consistent with the Stratonovich signature for mean-reverting time series and processes. We demonstrate that signature can be applied to learn nonlinear functions and option prices with high accuracy, and the performance depends on properties of the underlying process and the choice of the signature.

Funding: R. Zhang's research was supported by the National Key Research and Development Program of China [Grant 2022YFA1007900], the National Natural Science Foundation of China [Grant 72342004 and Grant 12271013], the Fundamental Research Funds for the Central Universities (Peking University), and Yinhua Education Foundation.

Supplemental Material: All supplemental materials, including the code, data, and files required to reproduce the results were reviewed and are available at https://doi.org/10.1287/opre.2024.1133.

Keywords: signature transform • Lasso • consistency • correlation structure • machine learning • option pricing

1. Introduction

1.1. Background and Problem Statement

Originally introduced and studied in algebraic topology (Chen 1954, 1957), the signature transform, sometimes referred to as the path signature or signature, has been adopted and further developed in rough path theory (Lyons et al. 2007, Friz and Victoir 2010). The signature produced from a continuous or discrete time series is a vector of real-valued features that extracts rich and relevant information (Morrill et al. 2020a, Lyons and McLeod 2022).

Signature has proven to be an attractive and powerful tool for feature generation and pattern recognition with state-of-the-art performance in a wide range of domains in operations research, such as medical prediction (Kormilitzin et al. 2017; Moore et al. 2019; Morrill et al. 2019, 2020b, 2021; Bleistein et al. 2023; Pan et al. 2023), transportation (Gu et al. 2024), and finance (Lyons et al. 2014, 2019, Kalsi et al. 2020, Salvi et al. 2021, Akyildirim et al. 2022, Cuchiero et al. 2023, Futter et al. 2023, Lemahieu et al. 2023).[1] Comprehensive reviews of successful and potential applications of signatures in machine learning can be found in Chevyrev and Kormilitzin (2016), Lyons and McLeod (2022), and Moreno-Pino et al. (2024).

Most of the empirical success and theoretical studies of the signature are built on its striking *universal nonlinearity* property: Any continuous (linear or nonlinear) function of the time series can be approximated arbitrarily well by a linear combination of its signature (see Section 2.2). This property linearizes the problem of feature selection, and empirical studies demonstrate that the universal nonlinearity property gives the signature several advantages over neural network–based nonlinear methods (Levin et al. 2016, Lyons and McLeod 2022, Bleistein et al. 2023, Pan et al. 2023, Gu et al. 2024). First, training linear models of signature do not require the engineering of neural network architectures; second, the linear model allows for interpretability (we show an example in Section 5.2).

Despite the rapidly growing literature on the *probabilistic* characteristics of signature and its successful application in machine learning, studies on the *statistical* properties of the signature method are limited with a few exceptions such as Király and Oberhauser (2019) and Morrill et al. (2020a).[2] In particular, universal nonlinearity can be expressed under different definitions of signature, raising the question of which definition has better statistical properties for different processes and time series. To our knowledge, most empirical studies in the literature simply use a default definition regardless of the specific context and characteristics of the data. However, using an inappropriate signature definition may lead to suboptimal performance, as we demonstrate in this paper.

Given the universal nonlinearity that legitimizes the regression analysis with signature, and given the popularity of Lasso regression (Tibshirani 1996) to learn a sparse model of signature,[3] the main focus of this paper is to understand the statistical properties of different forms of signature in Lasso regression given different time series data. In particular, we study the statistical consistency in feature selection, a fundamental property for Lasso regression to achieve both explainability and good out-of-sample model performance (Zhao and Yu 2006, Bickel et al. 2009, Wainwright 2009).

1.2. Main Results and Contribution

This paper studies the consistency of Lasso regression with signature both theoretically and numerically. We compare the two most widely used definitions of signature: Itô and Stratonovich. We focus on two representative classes of Gaussian processes: multidimensional Brownian motion and Ornstein–Uhlenbeck (OU) process and their respective discrete-time counterparts, that is, random walk and autoregressive (AR) process. These data-generating processes are simple enough to allow for analytical results while being fundamental in a number of domains ranging from machine learning (Song and Ermon 2019, Ho et al. 2020) and operations management (Asmussen 2003, Zhang et al. 2018) to finance (Black and Scholes 1973, Merton 1973) and biology (Martins 1994, Hunt 2007).

Our contributions are multifold. First, we establish a probabilistic uniqueness of the universal nonlinearity given an order of truncated signature (Theorem 2), which suggests that any feature selection procedure needs to recover this unique linear combination of signature to achieve good predictive performance.

Second, to analyze the consistency of Lasso regression with signature, we explicitly derive the correlation structure of signature for the aforementioned processes. For Brownian motion, the correlation structure is shown to be block diagonal for the Itô signature (Theorem 3) and to have a special odd–even alternating structure for the Stratonovich signature (Theorem 4). In contrast, the OU process exhibits this odd–even alternating structure for either choice of the signature (Theorem 4).

Third, we establish conditions under which the Lasso regression with signature is provably consistent both asymptotically and in finite sample (Theorems 5–8), based on the classical notions of sign consistency and l_∞ consistency (Zhao and Yu 2006, Wainwright 2009).

Furthermore, numerical experiments show that the Lasso regression with the Itô signature is more consistent for time series and processes that are closer to Brownian motion and with weaker interdimensional correlations, whereas it is more consistent with the Stratonovich signature for processes with stronger mean reversion. In general, higher consistency rates yield better predictive performance.

Finally, we demonstrate that the signature can be applied to learn nonlinear functions and option prices with high accuracy. We compare stock options with interest rate options to highlight that performance depends on the properties of the underlying process. This method is interpretable because the signature allows for learning a set of Arrow–Debreu state prices that are used for transfer-learning the prices of any general financial derivatives. These results demonstrate the practical relevance of our analysis.

Overall, our study takes a small step toward understanding the statistical properties of signatures for regression analysis. It fills one of the gaps between the theory and practice of signatures in machine learning. Our findings have significant implications for various applications in operations research by guiding the selection of the appropriate signature definition to achieve better statistical properties and predictive performance. For example, our study provides a theoretical foundation for the signature-based adaptive-Lasso technique that has been recently developed and implemented by Amazon for transportation marketplace rate forecasting and financial planning (Gu et al. 2024). This simple and novel model is reported to have generated tens of millions of monetary benefits for Amazon and demonstrates strong potential for a wide range of applications, especially when compared with existing models such as Autoregressive Integrated Moving Average (ARIMA) in terms of dealing with nonstationary data and deep neural network approach in terms of interpretability and for limited and fragmented data.

1.3. Notation

Here, we define the vector and matrix norms used throughout the paper. For a vector $\mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$, we define $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$, $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$, and $\|\mathbf{x}\|_\infty = \max_{1 \le i \le n} |x_i|$; for a matrix $A \in \mathbb{R}^{m \times n}$, we define $\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^m |a_{ij}|$, $\|A\|_2 = \sqrt{\Lambda_{\max}(A^\top A)}$, and $\|A\|_\infty = \max_{1 \le i \le m} \sum_{j=1}^n |a_{ij}|$, where $\Lambda_{\max}(\cdot)$ calculates the largest eigenvalue of a matrix, while $\Lambda_{\min}(\cdot)$ represents its smallest eigenvalue.

1.4. Outline

The rest of this paper is organized as follows. Section 2 introduces the problem and key technical background.

Section 3 presents the main theoretical results, including the uniqueness of universal nonlinearity, the correlation structure of signature, and the consistency of signature using Lasso regression. Section 4 presents a simulation study to gain additional insights. Section 5 applies our results to learning nonlinear functions and option prices. Finally, Section 6 concludes.

2. Background

In this section, we present the technical background to study the consistency of feature selection with signature transform using Lasso regression.

2.1. Definition of Signature Transform

Consider a d-dimensional continuous-time stochastic process $\mathbf{X}_t = (X_t^1, X_t^2, \ldots, X_t^d)^\top \in \mathbb{R}^d$, $0 \le t \le T$ on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\ge 0}, \mathbb{P})$.[4] Its signature or signature transform is defined as follows.

Definition 1 (Signature). For $k \ge 1$ and $i_1, \ldots, i_k \in \{1, 2, \ldots, d\}$, the kth-order signature component of the process $\mathbf{X}$ with index $(i_1, \ldots, i_k)$ from time 0 to t is defined as

$$S(\mathbf{X})_t^{i_1,\ldots,i_k} = \int_{0<t_1<\ldots<t_k<t} \mathrm{d}X_{t_1}^{i_1} \cdots \mathrm{d}X_{t_k}^{i_k}, \quad 0 \le t \le T. \quad (1)$$

The 0th-order signature component of $\mathbf{X}$ from time 0 to t is defined as $S(\mathbf{X})_t^0 = 1$ for any $0 \le t \le T$. The signature of $\mathbf{X}$ is the collection of all the signature components of $\mathbf{X}$. The signature of $\mathbf{X}$ with orders truncated to K is the collection of all the signature components of $\mathbf{X}$ with orders no more than K.

The kth order signature component of $\mathbf{X}$ given by (1) is its k-fold iterated path integral along the indices $i_1, \ldots, i_k$. For a given order k, there are d^k choices of indices $(i_1, \ldots, i_k)$, and therefore the number of all kth-order signature components is d^k.

The integral in (1) can be specified using different definitions. For example, if $\mathbf{X}$ is a deterministic process, it can be defined via the Riemann/Lebesgue integral. If $\mathbf{X}$ is a multidimensional Brownian motion, it is a stochastic integral defined by either the Itô integral or the Stratonovich integral. Throughout the paper, for clarity, we write

$$S(\mathbf{X})_t^{i_1,\ldots,i_k,I} = \int_{0<t_1<\cdots<t_k<t} \mathrm{d}X_{t_1}^{i_1} \cdots \mathrm{d}X_{t_k}^{i_k} = \int_{0<s<t} S(\mathbf{X})_s^{i_1,\ldots,i_{k-1},I} \mathrm{d}X_s^{i_k}$$

when using the Itô integral, and

$$S(\mathbf{X})_t^{i_1,\ldots,i_k,S} = \int_{0<t_1<\cdots<t_k<t} \mathrm{d}X_{t_1}^{i_1} \circ \cdots \circ \mathrm{d}X_{t_k}^{i_k} = \int_{0<s<t} S(\mathbf{X})_s^{i_1,\ldots,i_{k-1},S} \circ \mathrm{d}X_s^{i_k}$$

when using the Stratonovich integral. For ease of exposition, we refer to the signature of $\mathbf{X}$ as the Itô (respectively, Stratonovich) signature if the integral is defined in the sense of the Itô (respectively, Stratonovich) integral.

2.2. Universal Nonlinearity of Signature

One of the remarkable properties of the signature is its universal nonlinearity (Levin et al. 2016, Király and Oberhauser 2019, Fermanian 2021, Lemercier et al. 2021, Lyons and McLeod 2022).[5] It is particularly relevant for feature selection in statistical and machine learning, where one needs to find or learn a (nonlinear) function f that maps the path of $\mathbf{X}$ to a target label y. Examples include learning diagnosis or signals from medical time series such as the electrocardiogram (Morrill et al. 2019, 2020b, 2021), forecasting transportation marketplace rates from the time series of supply, demand, and macroeconomic factors (Gu et al. 2024), and learning a nonlinear payoff or pricing function for financial derivatives given the time series of the underlying asset prices (Hutchinson et al. 1994, Bertsimas et al. 2001, Lyons et al. 2020).

The following theorem of Cuchiero et al. (2023) outlines the universal nonlinearity.

Theorem 1 (Universal Nonlinearity: Cuchiero et al. 2023, Theorem 2.12). *Let $\mathbf{X}_t$ be a continuous $\mathbb{R}^d$-valued semimartingale and $\mathcal{S}$ be a compact subset of paths of the time-augmented process $\tilde{\mathbf{X}}_t = (t, \mathbf{X}_t^\top)^\top$ from time 0 to T.*[6] *Assume that $f: \mathcal{S} \to \mathbb{R}$ is a real-valued continuous function. Then, for any $\varepsilon > 0$, there exists a linear functional $L: \mathbb{R}^\infty \to \mathbb{R}$ such that*

$$\sup_{s\in\mathcal{S}} |f(s) - L(\mathrm{Sig}(s))| < \varepsilon,$$

where $\mathrm{Sig}(s)$ is the signature of s.

By universal nonlinearity, any continuous function f can be approximated arbitrarily well by a linear combination of the signature of $\mathbf{X}$. This lays the foundation for learning the relationship between the time series $\mathbf{X}$ and a target label y using a linear regression.

2.3. Feature Selection with Signature Using Lasso Regression

Consider N pairs of samples, $(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_N, y_N)$, where $\mathbf{X}_n = \{\mathbf{X}_{n,t}\}_{0\le t\le T}$ is the nth path realization of $\mathbf{X}_t$ for $n = 1, 2, \ldots, N$. Given a fixed order $K \ge 1$, assume that $(\mathbf{X}_n, y_n)$ satisfies the following regression model:

$$y_n = \beta_0 + \sum_{i_1=1}^{d} \beta_{i_1} S(\mathbf{X}_n)_T^{i_1} + \sum_{i_1,i_2=1}^{d} \beta_{i_1,i_2} S(\mathbf{X}_n)_T^{i_1,i_2} + \cdots + \sum_{i_1,\ldots,i_K=1}^{d} \beta_{i_1,\ldots,i_K} S(\mathbf{X}_n)_T^{i_1,\ldots,i_K} + \varepsilon_n, \quad (2)$$

where $\{\varepsilon_n\}_{n=1}^N$ are independent and identically distributed errors following a normal distribution with zero mean and finite variance. Here the number of predictors, that is, the signature components of various orders, is $(d^{K+1}-1)/(d-1)$, including the 0th-order signature component $S(\mathbf{X})^0_T = 1$, whose coefficient is β_0.

Recall that the goal of Lasso regression is to identify a sparse set of true predictors/features among all the predictors included in Linear Regression (2). A predictor has a zero beta coefficient if it is not in the true model. We use A^*_k to represent the set of all signature components of order k with nonzero coefficients in (2), and define the set of true predictors A^* by

$$A^* = \bigcup_{k=0}^{K} A^*_k := \bigcup_{k=0}^{K}\{(i_1,\ldots,i_k) : \beta_{i_1,\ldots,i_k} \neq 0\}. \tag{3}$$

Here, we begin the union with $k = 0$ to include the 0th-order signature for notational convenience.

Given a tuning parameter $\lambda > 0$ and N samples, the Lasso estimator identifies the true predictors using

$$\begin{aligned}\hat{\boldsymbol{\beta}}^N(\lambda) = \arg\min_{\hat{\boldsymbol{\beta}}}\Bigg[\sum_{n=1}^{N}\Bigg(y_n - \hat{\beta}_0 - \sum_{i_1=1}^{d}\hat{\beta}_{i_1}\tilde{S}(\mathbf{X}_n)^{i_1}_T \\ - \sum_{i_1,i_2=1}^{d}\hat{\beta}_{i_1,i_2}\tilde{S}(\mathbf{X}_n)^{i_1,i_2}_T - \cdots \\ - \sum_{i_1,\ldots,i_K=1}^{d}\hat{\beta}_{i_1,\ldots,i_K}\tilde{S}(\mathbf{X}_n)^{i_1,\ldots,i_K}_T\Bigg)^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1\Bigg],\end{aligned} \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is the vector containing all coefficients $\hat{\beta}_{i_1,\ldots,i_k}$. Here, $\tilde{S}(\mathbf{X}_n)$ represents the standardized version of $S(\mathbf{X}_n)$ across N samples by the l_2-norm. That is, for any index $(i_1,\ldots,i_k)$,[7]

$$\tilde{S}(\mathbf{X}_n)^{i_1,\ldots,i_k}_T = \frac{S(\mathbf{X}_n)^{i_1,\ldots,i_k}_T}{\sqrt{\sum_{m=1}^{N}[S(\mathbf{X}_m)^{i_1,\ldots,i_k}_T]^2/N}}, \quad n = 1,2,\ldots,N.$$

The Lasso estimator depends on the choice of K. The universal nonlinearity demonstrates that the linear combination of *all* components of the signature of $\mathbf{X}$ can be used to approximate f. However, because of computational constraints, we must truncate the signature to a finite order K in the implementation of Lasso regression. In theory, one can exploit the signature approximation in Dupire and Tissot-Daguette (2022) and the recent results on Taylor expansions of signatures in Cuchiero et al. (2025) to develop an error-bound analysis for the choice of K. Fermanian (2022) also provides a practical approach to choose K based on the tradeoff between the approximation error and the number of coefficients in the regression model. In practice, it has also been documented that a small order K usually suffices to achieve satisfactory performances (Morrill et al. 2020a, Lyons and McLeod 2022, Gu et al. 2024). For example, Gu et al. (2024) show that $K = 3$ is sufficient for forecasting models of transportation rates in Amazon.

2.4. Consistency and the Irrepresentable Condition of Lasso Regression

Our goal is to study the consistency of feature selection with signature using the Lasso estimator in (4). Broadly speaking, consistency means that the Lasso estimator converges to the true coefficients as the number of samples increases. In this section, we introduce two widely used notions of Lasso consistency from the literature and discuss the corresponding conditions required for each notion of consistency.

Zhao and Yu (2006) propose the sign consistency for Lasso regression, which requires that the signs of all components of the Lasso estimator match those of the true coefficients as the number of samples increases without bound. Wainwright (2009) studies the consistency of Lasso regression by requiring that the l_∞ distance between the true and the estimated coefficients is bounded.

In the context of Lasso regression with signature, the sign consistency and the l_∞ consistency of Lasso are defined as follows.

Definition 2 (Sign Consistency). Lasso regression is (strongly) sign consistent if there exists λ_N, a function of sample number N, such that

$$\lim_{N\to+\infty}\mathbb{P}(\operatorname{sign}(\hat{\boldsymbol{\beta}}^N(\lambda_N)) = \operatorname{sign}(\boldsymbol{\beta})) = 1,$$

where $\hat{\boldsymbol{\beta}}^N(\cdot)$ is the Lasso estimator given by (4), $\boldsymbol{\beta}$ is a vector containing all beta coefficients of the true model (2), and the function $\operatorname{sign}(\cdot)$ maps positive entries to 1, negative entries to –1, and zero to 0.

Definition 3 (l_∞ Consistency). There exists a function of λ_N, $g(\lambda_N)$, such that the Lasso regression satisfies the l_∞ bound

$$\|\hat{\boldsymbol{\beta}}^N(\lambda_N) - \tilde{\boldsymbol{\beta}}\|_\infty \le g(\lambda_N),$$

where $\hat{\boldsymbol{\beta}}^N(\cdot)$ is the Lasso estimator given by (4), and $\tilde{\boldsymbol{\beta}}$ is a vector containing all standardized beta coefficients of the true model whose component with index $(i_1,\ldots,i_k)$ is given by

$$\tilde{\beta}_{i_1,\ldots,i_k} = \beta_{i_1,\ldots,i_k}\cdot\sqrt{\frac{1}{N}\sum_{m=1}^{N}[S(\mathbf{X}_m)^{i_1,\ldots,i_k}_T]^2}.$$

As discussed in Wainwright (2009), if the support of $\hat{\boldsymbol{\beta}}^N(\lambda_N)$ is contained within the support of $\boldsymbol{\beta}$ and the absolute values of all beta coefficients for predictors in A^* are greater than $g(\lambda_N)$, the l_∞ consistency implies the sign consistency.

To guarantee the consistency of Lasso, Zhao and Yu (2006) and Wainwright (2009) propose the following two irrepresentable conditions, respectively.

Definition 4 (Irrepresentable Condition). The feature selection in (2) satisfies irrepresentable condition I if there exists a constant $\gamma \in (0,1]$ such that

$$\text{I.} \quad \|\Delta_{A^{*c},A^*}\Delta^{-1}_{A^*,A^*}\operatorname{sign}(\boldsymbol{\beta}_{A^*})\|_\infty \le 1-\gamma,$$

and satisfies irrepresentable condition II if there exists a constant $\gamma \in (0, 1]$ such that

$$\text{II.} \quad \|\Delta_{A^{*c},A^*}\Delta^{-1}_{A^*,A^*}\|_\infty \le 1-\gamma,$$

where A^* is given by (3), A^{*c} is the complement of A^*, Δ_{A^{*c},A^*} (Δ_{A^*,A^*}) represents the correlation matrix[8] between all predictors in A^{*c} and A^* (A^* and A^*), and $\boldsymbol{\beta}_{A^*}$ represents a vector formed by beta coefficients for all predictors in A^*.

The irrepresentable conditions in Definition 4 intuitively mean that irrelevant predictors in A^{*c} cannot be adequately represented by the true predictors in A^*, implying weak collinearity between the predictors. Zhao and Yu (2006) demonstrate that the irrepresentable condition I is almost a necessary and sufficient condition for the Lasso regression to be sign consistent. Wainwright (2009) proves that the irrepresentable condition II is a sufficient condition for the l_∞ consistency of Lasso regression under specific technical assumptions. The irrepresentable condition II is slightly stronger than the irrepresentable condition I.

In the context of signature, predictors in Linear Regression (2) are correlated and have special correlation structures that differ from previous studies on Lasso (Zhao and Yu 2006, Bickel et al. 2009, Wainwright 2009). We show in the following section that in fact their correlation structures vary with the underlying process $\mathbf{X}$ and the choice of integrals in the definition of Signature (1). These different correlation structures lead to different statistical consistencies.

3. Theoretical Results

This section presents the main theoretical results. Section 3.1 shows the *uniqueness* of universal nonlinearity in a probabilistic sense. Section 3.2 characterizes the correlation structures between signature components. Section 3.3 presents the results of consistency in signature selection, both asymptotically ($N = \infty$) and for a finite sample ($N < \infty$).

As outlined in Figure 1 for our results, the statistical consistency of signature using Lasso regression depends on two factors—the underlying processes $\mathbf{X}$ (Brownian motion or OU process) and the definition of signature (Itô or Stratonovich).

3.1. Uniqueness of Universal Nonlinearity

The universal nonlinearity in Theorem 1 shows the *existence* of a linear combination of signature components to approximate any function f. We provide the following Theorem 2 to complement the universal nonlinearity, which demonstrates the *uniqueness* of this linear combination in a probabilistic sense given an order of truncated signature. To the best of our knowledge, Theorem 2 has not appeared in the literature.

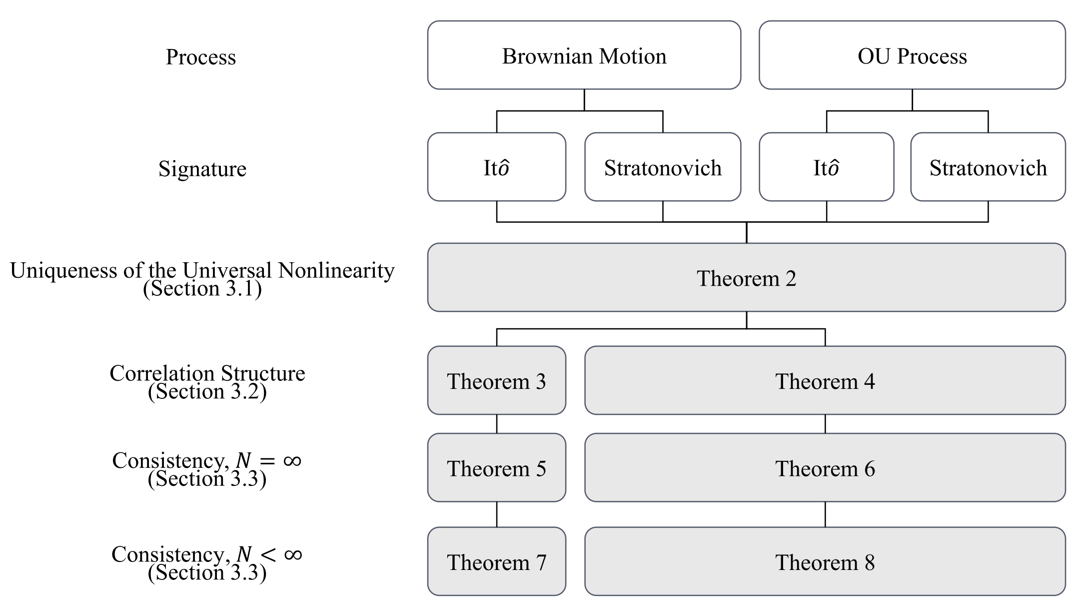
Theorem 2 (Uniqueness). *Given $K \ge 1$, let $\mathbf{S} = (S_1, S_2, \ldots, S_p)^\top$ be the vector of the signature of a stochastic process $\mathbf{X}$ with orders truncated to K, and assume $\mathbf{S}$ has a nondegenerate joint distribution. Consider two different linear combinations of signature components, $L_a = \sum_{i=1}^p a_i S_i$ and $L_b = \sum_{i=1}^p b_i S_i$, such that $a_i \ne b_i$ for at least some i. Then, there exists a constant $\theta > 0$ such that, for any $\eta \in (0, \overline{\eta})$,*

$$\mathbb{P}(|L_a - L_b| > \eta) \ge P^*_\theta(\eta) > 0. \tag{5}$$

Furthermore, if f is a function that maps $\mathbf{X}$ to a real value such that $|f(\mathbf{X}) - L_a| \le \varepsilon$ almost surely for a constant $\varepsilon < \overline{\eta}$, then for any $\eta \in (0, \overline{\eta} - \varepsilon)$,

$$\mathbb{P}(|f(\mathbf{X}) - L_b| > \eta) \ge P^*_\theta(\eta + \varepsilon) > 0. \tag{6}$$

Figure 1. Outline of Main Theoretical Results



Here

$$P^*_\theta(\eta) = \left(1 - \frac{1}{\theta}\right) \cdot \frac{\frac{\mathbb{E}\left[\left(\sum_{i=1}^p c_i S_i\right)^2 \mid \|S\|_2 \le \theta\sqrt{p\|\Sigma\|_2}\right]}{\theta\|C\|_\infty p\sqrt{\|\Sigma\|_2}} - \eta}{\theta\|C\|_\infty p\sqrt{\|\Sigma\|_2} - \eta},$$

$$\overline{\eta} = \min\left\{\frac{\mathbb{E}[(\sum_{i=1}^p c_i S_i)^2 \mid \|\boldsymbol{S}\|_2 \le \theta\sqrt{p\|\Sigma\|_2}]}{\theta\|C\|_\infty p\sqrt{\|\Sigma\|_2}}, \theta\|C\|_\infty p\sqrt{\|\Sigma\|_2}\right\},$$

with $c_i = a_i - b_i$, $C = (c_1, c_2, \ldots, c_p)^\top$, *and* $\Sigma = \mathbb{E}(\boldsymbol{S}\boldsymbol{S}^\top)$.

Theorem 2 has important implications for selecting signature components using Lasso regression. In particular, (6) shows that when a nonlinear function f is approximated by a linear combination of signature components L_a, there is always a positive probability that a different linear combination L_b has a positive gap from f, which implies that L_a is the unique linear combination to approximate f given an order of truncated signature K.[9] Therefore, given f, it is important for any feature selection procedure to recover this unique linear combination of signature components to achieve statistical consistency in feature selection.

There is a strand of literature focusing on whether signatures can uniquely determine the path of the underlying process (Hambly and Lyons 2010, Le Jan and Qian 2013, Boedihardjo et al. 2014). This literature investigates the one-to-one correspondence between $\mathbf{X}$ and its signature. This is different from Theorem 2, which characterizes the one-to-one correspondence between $f(\mathbf{X})$ and the linear combination of signature components.

3.2. Correlation Structure of Signature

Now we study the correlation structure of the four combinations of processes and signatures in Figure 1. Throughout the paper, we define $\mathbf{X} = \{\mathbf{X}_t\}_{t\ge 0}$ as a d-dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\ge 0}, \mathbb{P})$ if

$$\mathbf{X}_t = (X_t^1, X_t^2, \ldots, X_t^d)^\top = \Gamma(W_t^1, W_t^2, \ldots, W_t^d)^\top, \tag{7}$$

where $W_t^1, W_t^2, \ldots, W_t^d$ are mutually independent one-dimensional standard Brownian motions on $\mathbb{R}$, and Γ is a matrix independent of t. In particular, $\langle X_t^i, X_t^j\rangle = \rho_{ij}\sigma_i\sigma_j t$ with $\rho_{ij}\sigma_i\sigma_j = (\Gamma\Gamma^\top)_{ij}$, where $\sigma_i^2 t$ is the variance of X_t^i and $\rho_{ij} \in [-1, 1]$ is the interdimensional correlation between X_t^i and X_t^j.

We say that $\mathbf{X} = \{\mathbf{X}_t\}_{t\ge 0}$ is a d-dimensional OU process on a probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\ge 0}, \mathbb{P})$ if

$$\mathbf{X}_t = (X_t^1, X_t^2, \ldots, X_t^d)^\top = \Gamma(Y_t^1, Y_t^2, \ldots, Y_t^d)^\top, \tag{8}$$

where Γ is a $d \times d$ matrix independent of t, and $Y_t^1, Y_t^2, \ldots, Y_t^d$ are mutually independent one-dimensional OU processes on $\mathbb{R}$ driven by stochastic differential equations

$$\mathrm{d}Y_t^i = -\kappa_i Y_t^i \mathrm{d}t + \mathrm{d}W_t^i, \quad Y_0^i = 0,$$

for $i = 1, 2, \ldots, d$. Here $\kappa_i > 0$ are parameters to control the speed of mean reversion and a higher κ_i implies a stronger mean reversion. When $\kappa_i = 0$, Y_t^i reduces to a standard Brownian motion.

3.2.1. Itô Signature of Brownian Motion.

The following proposition gives the moments of the Itô signature of a d-dimensional Brownian motion.

Proposition 1. *Let* $\mathbf{X}$ *be a* d*-dimensional Brownian motion given by* (7). *For* $m, n \in \mathbb{Z}^+$ *and* $m \ne n$,

$$\mathbb{E}[S(\mathbf{X})_t^{i_1,\ldots,i_n,I}] = 0, \quad \mathbb{E}[S(\mathbf{X})_t^{i_1,\ldots,i_n,I} S(\mathbf{X})_t^{j_1,\ldots,j_n,I}] = \frac{t^n}{n!}\prod_{k=1}^n \rho_{i_k j_k}\sigma_{i_k}\sigma_{j_k},$$

$$\mathbb{E}[S(\mathbf{X})_t^{i_1,\ldots,i_n,I} S(\mathbf{X})_t^{j_1,\ldots,j_m,I}] = 0.$$

With Proposition 1, the following result explicitly characterizes the correlation structure of the Itô signature for Brownian motion.

Theorem 3. *Let* $\mathbf{X}$ *be a* d*-dimensional Brownian motion given by* (7). *If the signature is rearranged in recursive order (see Definition* B.1 *in Online Appendix* B.1*), then the correlation matrix for the Itô signature of* $\mathbf{X}$ *with orders truncated to* K *is a block diagonal matrix given by*

$$\Delta^1 = \mathrm{diag}\{\Omega_0, \Omega_1, \Omega_2, \ldots, \Omega_K\}, \tag{9}$$

where each diagonal block Ω_k *represents the correlation matrix for all kth-order signature components given by*

$$\Omega_k = \underbrace{\Omega \otimes \Omega \otimes \cdots \otimes \Omega}_{k}, \quad k = 1, 2, \ldots, K, \tag{10}$$

and $\Omega_0 = 1$. *Here* $\otimes$ *represents the Kronecker product and* Ω *is a* $d \times d$ *matrix with* ρ_{ij} *being the* (i, j)*th entry.*

Theorem 3 shows that the Itô signature components of different orders are mutually uncorrelated, leading to a block diagonal correlation structure; the correlation between signature components of the same order has a Kronecker product structure determined by the correlation ρ_{ij} of the Brownian motion.

The block diagonal structure of the correlation matrix has important statistical implications for the Itô signature. In Section 3.3, we demonstrate that the Lasso regression using signature as predictors is consistent if the correlation is weak (see Theorems 5 and 7). However, when the correlation within each block is strong, Lasso may be unstable for signature components in the same block. In such cases, one may consider using methods such as sparse principal component analysis (Zou et al. 2006, Leng and Wang 2009) or scaled Lasso (Arashi et al. 2021) to address multicollinearity and achieve a more stable Lasso estimation.

3.2.2. Stratonovich Signature of Brownian Motion and Both Signatures of OU Process.

We first provide the moments of the Stratonovich signature of Brownian motion.

Proposition 2. *Let $\mathbf{X}$ be a d-dimensional Brownian motion given by (7). For $m, n \in \mathbb{Z}^+$, we have*

$$\mathbb{E}[S(\mathbf{X})_t^{i_1,\dots,i_{2n-1},S}] = 0, \quad \mathbb{E}[S(\mathbf{X})_t^{i_1,\dots,i_{2n},S}] = \frac{1}{2^n}\frac{t^n}{n!}\prod_{k=1}^{n}\rho_{i_{2k-1}i_{2k}}\prod_{k=1}^{2n}\sigma_{i_k},$$

$$\mathbb{E}[S(\mathbf{X})_t^{i_1,\dots,i_{2n},S}S(\mathbf{X})_t^{j_1,\dots,j_{2m-1},S}] = 0,$$

and $\mathbb{E}[S(\mathbf{X})_t^{i_1,\dots,i_{2n},S}S(\mathbf{X})_t^{j_1,\dots,j_{2m},S}]$ *and* $\mathbb{E}[S(\mathbf{X})_t^{i_1,\dots,i_{2n-1},S} S(\mathbf{X})_t^{j_1,\dots,j_{2m-1},S}]$ *can be calculated using formulas provided in Proposition* B.1 *in Online Appendix* B.1.

The calculation of moments for the OU process is more complicated than those for the Brownian motion, as discussed in Online Appendix B.2. Nonetheless, the correlation matrices of both the Itô and the Stratonovich signatures of the OU process exhibit the same odd–even alternating structure as that of the Stratonovich signature of the Brownian motion, which is given below.

Theorem 4. *Consider the Stratonovich signature of a d-dimensional Brownian motion given by (7), or the Itô or the Stratonovich signature of a d-dimensional OU process given by (8). The correlation matrix for the signature with orders truncated to* 2K *has an odd–even alternating structure given by*

$$\Delta^2 = \begin{pmatrix} \Psi_{0,0} & 0 & \Psi_{0,2} & 0 & \cdots & 0 & \Psi_{0,2K} \\ 0 & \Psi_{1,1} & 0 & \Psi_{1,3} & \cdots & \Psi_{1,2K-1} & 0 \\ \Psi_{2,0} & 0 & \Psi_{2,2} & 0 & \cdots & 0 & \Psi_{2,2K} \\ 0 & \Psi_{3,1} & 0 & \Psi_{3,3} & \cdots & \Psi_{3,2K-1} & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \Psi_{2K-1,1} & 0 & \Psi_{2K-1,3} & \cdots & \Psi_{2K-1,2K-1} & 0 \\ \Psi_{2K,0} & 0 & \Psi_{2K,2} & 0 & \cdots & 0 & \Psi_{2K,2K} \end{pmatrix}, \tag{11}$$

where $\Psi_{m,n}$ is the correlation matrix between all mth and nth order signature components.[10] *In particular, if the indices of the signature components are rearranged with all odd-order signature components and all even-order signature components together, respectively, the correlation matrix has a block diagonal form given by*

$$\tilde{\Delta}^2 = \mathrm{diag}\{\Psi_{\text{odd}}, \Psi_{\text{even}}\}, \tag{12}$$

where

$$\Psi_{\text{odd}} = \begin{pmatrix} \Psi_{1,1} & \Psi_{1,3} & \cdots & \Psi_{1,2K-1} \\ \Psi_{3,1} & \Psi_{3,3} & \cdots & \Psi_{3,2K-1} \\ \vdots & \vdots & \cdots & \vdots \\ \Psi_{2K-1,1} & \Psi_{2K-1,3} & \cdots & \Psi_{2K-1,2K-1} \end{pmatrix},$$

$$\Psi_{\text{even}} = \begin{pmatrix} \Psi_{0,0} & \Psi_{0,2} & \cdots & \Psi_{0,2K} \\ \Psi_{2,0} & \Psi_{2,2} & \cdots & \Psi_{2,2K} \\ \vdots & \vdots & \cdots & \vdots \\ \Psi_{2K,0} & \Psi_{2K,2} & \cdots & \Psi_{2K,2K} \end{pmatrix}. \tag{13}$$

Theorems 3 and 4 reveal a striking difference between the four combinations of processes and signatures in Figure 1. Specifically, a Brownian motion's Itô signature components of different orders are uncorrelated, leading to a block diagonal correlation structure. In contrast, for the Stratonovich signature of Brownian motion and both signatures of the OU process, the components are uncorrelated only if they have different parity, leading to an odd–even alternating structure. This difference has significant implications for the consistency of the four combinations of processes and signatures, as will be discussed in Section 3.3.

Finally, signature-based analyses sometimes consider time augmentation in which a time dimension t is added to the original process $\mathbf{X}_t$ (Chevyrev and Kormilitzin 2016, Lyons and McLeod 2022). Online Appendix A provides the correlation structure and consistency results of the time-augmented processes.

3.3. Consistency of Signature Using Lasso Regression

This section investigates the consistency of feature selection using the four combinations of processes and signatures in Figure 1.

3.3.1. Asymptotic Results. The following theorem characterizes the conditions under which the irrepresentable condition holds for the Itô signature of Brownian motion.

Theorem 5. *For a multidimensional Brownian motion given by (7), consider its Itô signature with orders truncated to K. Both irrepresentable conditions* I *and* II *hold if and only if they hold for each Ω_k in (10). In addition, both irrepresentable conditions* I *and* II *hold if*

$$|\rho_{ij}| < \frac{1}{2q_{\max} - 1}, \tag{14}$$

where $q_{\max} = \max_{0 \le k \le K}\{\#A_k^*\}$ and A_k^* is the set of true predictors of order k defined in (3).

The sufficient condition (14) in Theorem 5 requires that different dimensions of the multidimensional Brownian motion are not strongly correlated, with a sufficient bound given by (14). Empirically, it has been documented that a small K suffices to provide a reasonable approximation in applications (Morrill et al. 2020a, Lyons and McLeod 2022). Therefore, $q_{\max}$ is typically small, which implies that the bound given by (14) is fairly easy to satisfy. In addition, Online Appendix C.1 discusses the tightness of this bound.

In fact, Zhao and Yu (2006, corollary 2) demonstrate that any Lasso regression is consistent if the absolute values of the correlations between predictors are smaller than $1/(2q-1)$, where $q = \#A^*$ is the total number of true predictors in the Lasso regression. Our sufficient Condition (14) provides a much more relaxed upper bound compared with the condition of Zhao and Yu (2006), thanks to the block diagonal correlation structure of the Itô signature for Brownian motion given by Theorem 3. This is because A_k^* is the set of true predictors in the kth block of (9), and $q_{\max}$ is the maximum number of true predictors across all these blocks. In other words, even with a large number of all true predictors ($\#A^*$) in the Lasso regression, it remains consistent as long as the number of true predictors within each block ($\#A_k^*$) is relatively small.

The following theorem characterizes the condition under which the irrepresentable condition holds for the Stratonovich signature of a Brownian motion and for both signatures of the OU process.

Theorem 6. *Consider the Stratonovich signature of a d-dimensional Brownian motion given by* (7)*, or the Itô or the Stratonovich signature of a d-dimensional OU process given by* (8)*, with orders truncated to 2K. Both irrepresentable conditions* I *and* II *hold if and only if they hold for both* Ψ_{odd} *and* Ψ_{even} *in* (13).

For these types of signatures, the irrepresentable condition may fail even when all dimensions of $\mathbf{X}$ are mutually independent, as is shown in Example 3.4 in Online Appendix B. Therefore, no sufficient conditions of the form (14) can be established. This implies that, for example, the Stratonovich signature of Brownian motion may exhibit lower consistency compared with its Itô signature, which we confirm in Section 4.

3.3.2. Finite Sample Results. Theorems 5 and 6 characterize when the irrepresentable conditions hold for the population correlation matrix of signature, which implies the sign consistency (Definition 2) of Lasso regression when $N = \infty$. In practice, however, the number of sample paths is finite, that is, $N < \infty$. Hence, the *sample correlation matrix*, denoted by $\hat{\Delta}$, may deviate from the population correlation matrix Δ. The following results demonstrate that the Lasso regression using signature maintains consistency *with high probability* in finite sample under certain conditions.[11] In addition, Online Appendix C.2 discusses the consistency of Lasso regression with general predictors in finite sample.

Theorem 7. *For a multidimensional Brownian motion given by* (7)*, consider a Lasso regression* (4) *using the Itô signature with orders truncated to K as predictors. Let* $\rho = \max_{i\neq j}\{|\rho_{ij}|\}$*,* σ *the volatility of* ε_n *in* (2)*,* $q_{\max} = \max_{0\le k\le K}\{\#A_k^*\}$*, and p the number of predictors in the Lasso regression. If* (14) *holds and the sequence of regularization parameters* $\{\lambda_N\}$ *satisfies* $\lambda_N > \frac{4\sigma(1-(q_{\max}-1)\rho)}{1-(2q_{\max}-1)\rho}\sqrt{\frac{2\ln p}{N}}$*, then the following properties hold with probability greater than*

$$P_{\min}^1 := \left(1 - \frac{8p^4\sigma_{\max}^4(\sigma_{\min}^4 + c_1)}{N\xi^2\sigma_{\min}^4}\right)(1 - 4e^{-c_2N\lambda_N^2}) \quad (15)$$

for some positive constants c_1 *and* c_2.

a. *The Lasso regression has a unique solution* $\hat{\boldsymbol{\beta}}^N(\lambda_N) \in \mathbb{R}^p$ *with its support contained within the true support, and* $\hat{\boldsymbol{\beta}}^N(\lambda_N)$ *satisfies*

$$\|\hat{\boldsymbol{\beta}}^N(\lambda_N) - \tilde{\boldsymbol{\beta}}\|_\infty \le \lambda_N\left[\frac{3-(2q_{\max}-3)\rho}{(1-(q_{\max}-1)\rho)(2+2\rho)} + 4\sigma\sqrt{\frac{2q_{\max}^{\frac{1}{2}}}{(1-(q_{\max}-1)\rho}}\right] =: h(\lambda_N).$$

b. *If in addition* $\min_{i\in A^*}|\tilde{\beta}_i| > h(\lambda_N)$*, then* $\operatorname{sign}(\hat{\boldsymbol{\beta}}^N(\lambda_N)) = \operatorname{sign}(\tilde{\boldsymbol{\beta}})$.

Here, $\xi = \min\left\{g_\Sigma^{-1}\left(\frac{(1-(2q_{\max}-1)\rho)(1-(q_{\max}-1)\rho)}{3-(2q_{\max}-3)\rho}\right), g_\Sigma^{-1}\left(\frac{1-(q_{\max}-1)\rho}{2\sqrt{pq_{\max}}}\right)\right\} > 0$ *with*

$$g_\Sigma(x) = \frac{2x\sigma_{\min}^2(p-1)\rho}{(\sigma_{\min}^2 - x)(2\sigma_{\min}^2 - x)} + \frac{(p-1)x}{\sigma_{\min}^2 - x}, \quad (16)$$

$\sigma_{\min} = \min_{1\le i\le p}\sqrt{\Sigma_{ii}}$, $\sigma_{\max} = \max_{1\le i\le p}\sqrt{\Sigma_{ii}}$*, and* Σ *the population covariance matrix of all predictors in* (2).

Part (a) of Theorem 7 demonstrates that, for a Brownian motion, when using the Itô signature as predictors, the difference between the coefficients estimated using Lasso regression and the true values can be bounded, leading to the l_∞ consistency. Part (b) shows that the sign consistency of Lasso regression holds if the magnitudes of true parameters are sufficiently large. Both results hold with a probability of at least $P_{\min}^1$. In particular, the lower bound probability (15) characterizes how likely the Lasso regression can recover the true set of signature components. This probability converges to one at a polynomial rate of N^{-1} as the number of samples increases without bound. Clearly, taking partial derivatives yields the following proposition, which illustrates how this probability varies with different parameters of the model.[12]

Proposition 3. *Holding other parameters constant, the lower bound of probability* $P_{\min}^1$ *given by* (15)

i. *Decreases with respect to* ρ*, p, and* $q_{\max}$*, which correspond to the upper bound of the interdimensional correlation of the Brownian motion, the number of predictors in the Lasso regression, and the number of true predictors, respectively; and*

ii. *Increases with respect to N, the number of sample paths.*

Proposition 3 demonstrates that the Lasso regression is (more likely to be) consistent when different dimensions of the Brownian motion are less correlated or when there are fewer predictors and true predictors in the model, or when more samples are observed. These findings align with general intuition.

The following results demonstrate the consistency of Lasso regression when using the Stratonovich signature as predictors for Brownian motion or using the Itô or Stratonovich signature as predictors for the OU process.

Theorem 8. *Consider a Lasso regression* (4) *using the Stratonovich signature as predictors for a multidimensional Brownian motion given by* (7) *or the Itô or Stratonovich signature as predictors for a multidimensional OU process given by* (8), *with orders truncated to* 2K. *Let* σ *be the volatility of* ε_n *in* (2) *and p be the number of predictors in the Lasso regression. If the irrepresentable condition* II *holds for both* Ψ_{odd} *and* Ψ_{even} *given by* (13), *and the sequence of regularization parameters* $\{\lambda_N\}$ *satisfies* $\lambda_N > \frac{4\sigma}{\gamma}\sqrt{\frac{2\ln p}{N}}$, *then the following properties hold with probability greater than*

$$P_{\min}^2 := \left(1 - \frac{8p^4\sigma_{\max}^4(\sigma_{\min}^4 + c_1)}{N\xi^2\sigma_{\min}^4}\right)(1 - 4e^{-c_2N\lambda_N^2}) \qquad (17)$$

for some positive constants c_1 *and* c_2.

a. *The Lasso regression has a unique solution* $\hat{\boldsymbol{\beta}}^N(\lambda_N) \in \mathbb{R}^p$ *with its support contained within the true support, and* $\hat{\boldsymbol{\beta}}^N(\lambda_N)$ *satisfies*

$$\|\hat{\boldsymbol{\beta}}^N(\lambda_N) - \tilde{\boldsymbol{\beta}}\|_\infty \le \lambda_N\left[\frac{\zeta(2+2\alpha\zeta+\gamma)}{2+2\alpha\zeta} + \frac{4\sigma}{\sqrt{\frac{1}{2}C_{\min}}}\right] =: h(\lambda_N); \text{ and}$$

b. *If in addition* $\min_{i\in A^*}|\tilde{\beta}_i| > h(\lambda_N)$, *then* $\text{sign}(\hat{\boldsymbol{\beta}}^N(\lambda_N)) = \text{sign}(\tilde{\boldsymbol{\beta}})$.

Here,

- $\alpha = \|\Delta_{A^{*c}A^*}\|_\infty = \max\{\|\Psi_{\text{odd},A^{*c}A^*}\|_\infty, \|\Psi_{\text{even},A^{*c}A^*}\|_\infty\}$;
- $\zeta = \|\Delta_{A^*A^*}^{-1}\|_\infty = \max\{\|\Psi_{\text{odd},A^*A^*}^{-1}\|_\infty, \|\Psi_{\text{even},A^*A^*}^{-1}\|_\infty\}$;
- $C_{\min} = \Lambda_{\min}(\Delta_{A^*A^*}) = \min\{\Lambda_{\min}(\Psi_{\text{odd},A^*A^*}), \Lambda_{\min}(\Psi_{\text{even},A^*A^*})\}$;
- $\gamma = \min\{1 - \|\Psi_{\text{odd},A^{*c}A^*}\Psi_{\text{odd},A^*A^*}^{-1}\|_\infty, 1 - \|\Psi_{\text{even},A^{*c}A^*}\Psi_{\text{even},A^*A^*}^{-1}\|_\infty\}$;
- $\xi = \min\left\{g_\Sigma^{-1}\left(\frac{\gamma}{\zeta(2+2\alpha\zeta+\gamma)}\right), g_\Sigma^{-1}\left(\frac{C_{\min}}{2\sqrt{p}}\right)\right\} > 0$;
- $g_\Sigma(\cdot)$ *is defined by* (16), $\sigma_{\min} = \min_{1\le i\le p}\sqrt{\Sigma_{ii}}$, $\sigma_{\max} = \max_{1\le i\le p}\sqrt{\Sigma_{ii}}$, *and* Σ *is the population covariance matrix of all predictors in* (2).

In comparison with the result for the Itô signature of Brownian motion (Theorem 7), Theorem 8 is mathematically more involved as a result of the more complex correlation structure (see Theorems 3 and 4). The lower bound probability (17) also converges to one at a polynomial rate of N^{-1} as the number of samples increases without bound. Clearly, taking partial derivatives yields the following proposition, which shows how the lower bound probability $P_{\min}^2$ varies with the parameters.

Proposition 4. *Holding other parameters constant, the lower bound of probability* $P_{\min}^2$ *given by* (17)

i. *Decreases with respect to* α *and p, which correspond to the upper bound for the correlation between true predictors and false predictors and the number of predictors in the Lasso regression, respectively; and*

ii. *Increases with respect to* γ *and N, which correspond to the degree of compliance with the irrepresentable condition II for the population correlation matrix and the number of samples, respectively.*

Like the result for the Itô signature of Brownian motion (Proposition 3), Proposition 4 demonstrates that the Lasso regression is (more likely to be) consistent when there are fewer predictors in the model or when more samples are observed. Furthermore, a lower correlation between predictors and greater compliance with the irrepresentable condition both improve the consistency.

4. Simulation

We use numerical simulations to illustrate our theoretical results and gain additional insights into the consistency of Lasso regression for signature transform.[13]

4.1. Consistency

Consider a two-dimensional ($d = 2$) Brownian motion with interdimensional correlation ρ.[14] Assume that there are $q = \#A^*$ true predictors in the true model (2), all of which are signature components up to order $K = 4$. We follow the steps below to perform our experiment.

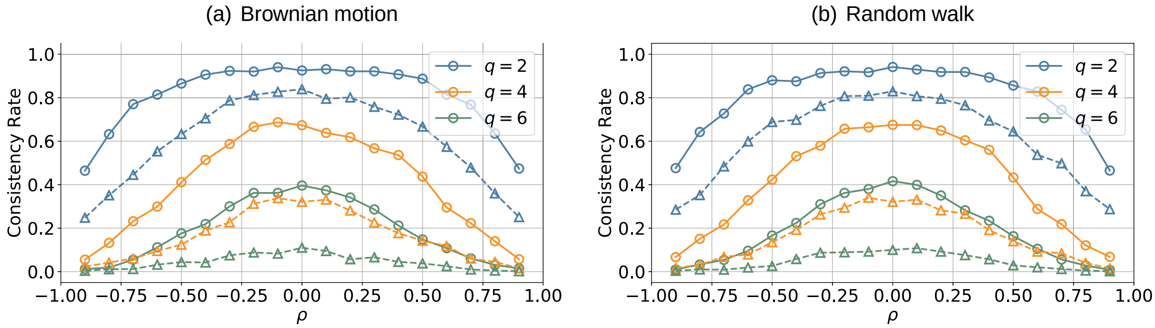
1. Randomly choose q true predictors from all $\frac{d^{K+1}-1}{d-1} = 31$ signature components;
2. Randomly set each beta coefficient of these true predictors from the standard normal distribution;
3. Generate 100 samples from this true model with error term ε_n drawn from a normal distribution with mean zero and standard deviation 0.01;
4. Run a Lasso regression given by (4) to select predictors based on these 100 samples; and
5. Check whether the Lasso regression is sign consistent according to Definition 2.

We then repeat the above procedure 1,000 times and calculate the *consistency rate*, which is defined as the proportion of consistent results among these 1,000 experiments.

Figure 2 shows the consistency rates for different values of interdimensional correlation ρ and true predictors q, with Figure 2(a) for Brownian motion and Figure 2(b) for its discrete counterpart—the random walk. First, signatures for both Brownian motion and random walk are similar. They both exhibit higher consistency rates when the absolute value of ρ is small, that is, when the interdimensional correlation of either Brownian motion or random walk is weak. Second, as the number of true predictors q increases, both consistency rates decrease. These findings are consistent with Theorem 5, Theorem 7, and Proposition 3.

Furthermore, the consistency rates for the Itô signature are consistently higher than those for the Stratonovich signature, with ρ and q fixed. This is consistent with Theorems 3 and 4—signature components of different orders are uncorrelated using the Itô signature but correlated using the Stratonovich signature.

Figure 2. (Color online) Consistency Rates for the Brownian Motion and Random Walk with Different Values of Interdimensional Correlation ρ and Different Numbers of True Predictors q



Note. Solid (dashed) lines correspond to the Itô (Stratonovich) signature.

The collinearity between the Stratonovich signature components contributes to their lower consistency for Lasso regression.

Online Appendix D.3 provides additional results for the impact of the number of dimensions d and the number of samples N.

4.2. Predictive Performance

A higher consistency rate of the Lasso regression is desirable as it is associated with better predictive performance of the model, which we confirm in this section using out-of-sample data.

To this end, we conduct additional simulations for Brownian motion and random walk, following a similar setup as in Section 4.1, with 200 samples generated from the true model for each experiment. These 200 samples are then equally divided into a training set and a test set, each containing 100 samples. Next, we run a Lasso regression on the training set and choose the tuning parameter λ using fivefold cross-validation. Finally, we calculate the out-of-sample mean squared error (OOS MSE) using the chosen λ on the test set.

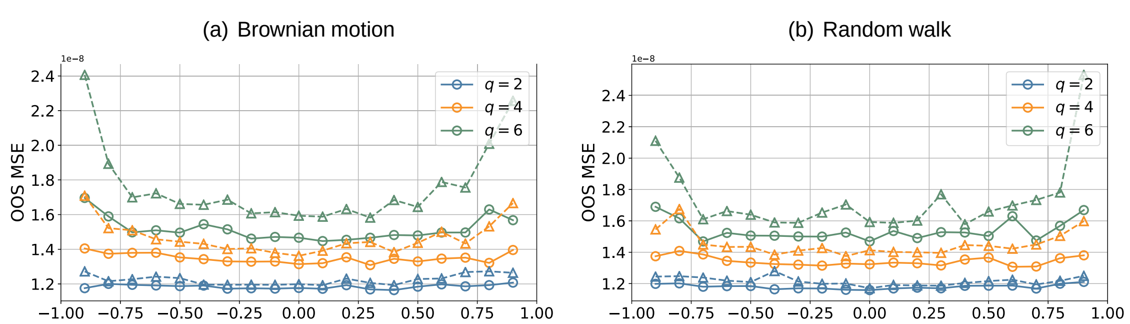
Overall, this analysis confirms that the insights derived from sign consistency extend to predictive performance metrics. In particular, Figure 3 shows the OOS MSE for different values of the interdimensional correlation ρ and different numbers of true predictors q. First, Lasso regression shows lower OOS MSE when the absolute value of ρ is small, that is, when the interdimensional correlations are weak. Second, as the number of true predictors q increases, the OOS MSE increases. Finally, the Itô signature has a lower OOS MSE compared with the Stratonovich signature with fixed ρ and q.

4.3. Impact of Mean Reversion

To study the impact of mean reversion on the consistency of Lasso regression, we run simulations for both the OU process and its discrete counterpart—the autoregressive AR(1) model with parameter ϕ. Recall that higher values of κ for the OU process and lower values of ϕ for the AR(1) model imply stronger mean reversion. We consider two-dimensional OU and AR(1) processes, with both dimensions sharing the same parameters (κ and ϕ). The interdimensional correlation matrix $\Gamma\Gamma^\top$ is randomly drawn from the Wishart(2, 2) distribution. All other setups are the same as the Brownian motion experiment in Section 4.1.

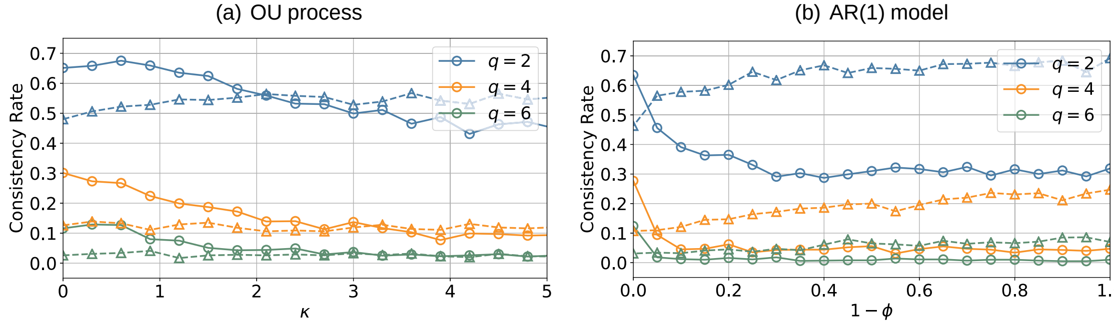
Figure 4 shows the simulation results for the consistency rates of both processes. First, the Itô signature

Figure 3. (Color online) OOS MSE for the Brownian Motion and Random Walk with Different Values of Interdimensional Correlation ρ and Different Numbers of True Predictors q



Note. Solid (dashed) lines correspond to the Itô (Stratonovich) signature.

Figure 4. (Color online) Consistency Rates for the OU Process and the AR(1) Model with Different Parameters (κ and $1-\phi$) and Different Numbers of True Predictors q



Note. Solid (dashed) lines correspond to the Itô (Stratonovich) signature.

reaches the highest consistency rate when κ and $1-\phi$ approach zero, which corresponds to a Brownian motion and a random walk. Second, when the process is sufficiently mean reverting, the Stratonovich signature has higher consistency rates than the Itô signature. Finally, Lasso regression becomes less consistent as the number of true predictors q increases, a similar observation as in the experiment for Brownian motion.

Overall, these results suggest that, for processes that are sufficiently rough or mean reverting (Gatheral et al. 2018), using Lasso regression with the Stratonovich signature will likely lead to a higher statistical consistency and better out-of-sample predictive performance compared with the Itô signature. Online Appendix B.2 provides more theoretical explanations, and Online Appendix D.4 examines the more complex ARIMA processes.

5. Applications

In this section, we use both the Itô and the Stratonovich signatures to understand the implication of their statistical properties in real applications. In particular, Section 5.1 uses the signature transform to learn option payoffs based on its universal nonlinearity, and Section 5.2 illustrates the application of the signature transform in option pricing.

5.1. Learning Option Payoffs

Option payoffs are nonlinear functions of the underlying asset. We first show that Lasso regression using signature as predictors can approximate these nonlinear functions well in terms of regression R^2. We then use the results derived in Sections 3.3 and 4 to guide the selection between the Itô and Stratonovich signatures.

5.1.1. Fitting Performance. We consider two underlying assets, X_t^1 and X_t^2, both of which following geometric Brownian motions with $X_0^1 = X_0^2 = 1$, $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = 0.2$. The correlation between the two assets is 0.6. We consider the following eight option payoff functions with time to maturity $T = 1$. In the simulation, we employ the Euler-Maruyama method for discretization and divide the time interval into 1,000 steps.

a. Call option ($d = 1$): $\max(X_T^1 - 1.2, 0)$;
b. Put option ($d = 1$): $\max(0.8 - X_T^1, 0)$;
c. Asian option ($d = 1$): $\max(\text{mean}_{0\le t\le T}(X_t^1) - 1.2, 0)$;
d. Lookback option ($d = 1$): $\max(\max_{0\le t\le T}(X_t^1) - 1.2, 0)$;
e. Rainbow option I ($d = 2$): $\max(X_T^1 - X_T^2, 0)$;
f. Rainbow option II ($d = 2$): $\max(\max(X_T^1, X_T^2) - 1.2, 0)$;
g. Rainbow option III ($d = 2$): $\max(\max_{0\le t\le T}(X_t^1) - \max_{0\le t\le T}(X_t^2), 0)$; and
h. Rainbow option IV ($d = 2$): $\max(\text{mean}_{0\le t\le T}(X_t^1) + \text{mean}_{0\le t\le T}(X_t^2) - 2.4, 0)$.

The first two are standard options most commonly used in practice; the third and fourth have payoff functions that depend on the entire path of the underlying prices; and the last four have payoff functions relying on multidimensional underlying paths, with the fifth and sixth depending only on the terminal values and the seventh and eighth depending on the entire path.

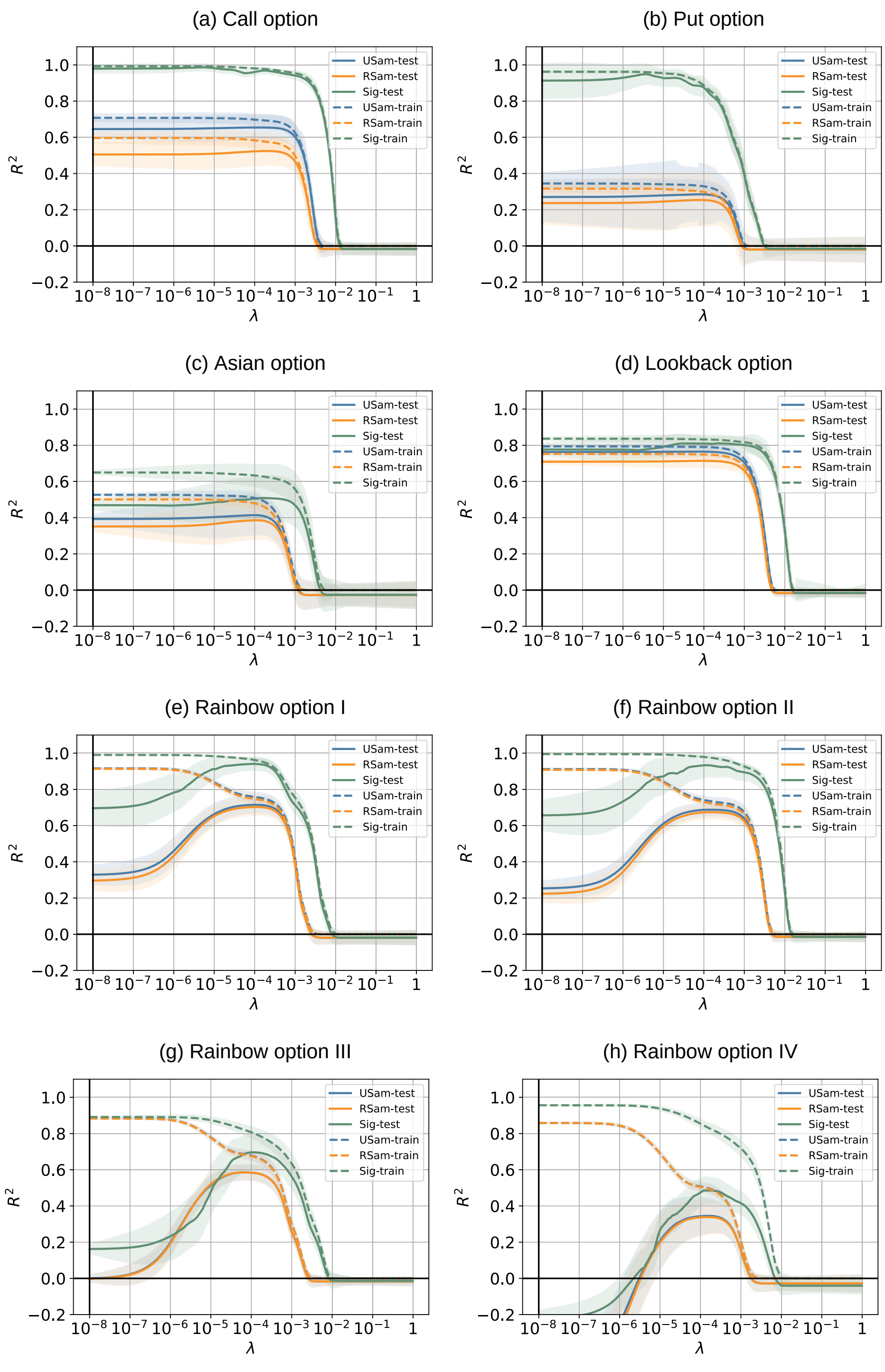
For each option payoff and for $K = 6$,[15] we perform Lasso regression using the following three different types of predictors.

1. The Stratonovich signature of the path of the underlying asset(s) with orders up to K (denoted as "Sig")
2. A set of $p\left(=\frac{d^{K+1}-1}{d-1}\right)$ randomly sampled points from the path of the underlying asset(s) (denoted as "RSam")
3. A set of $p\left(=\frac{d^{K+1}-1}{d-1}\right)$ equidistant points from the path of the underlying asset(s) (denoted as "USam")

The training set for the Lasso regression consists of 200 simulated paths, and the test set consists of 100 simulated paths. We repeat each experiment 200 times to derive confidence intervals for the estimates.

Results. Figure 5 shows the relationship between R^2 and the penalization parameter of the Lasso regression λ, when using different types of predictors. Both

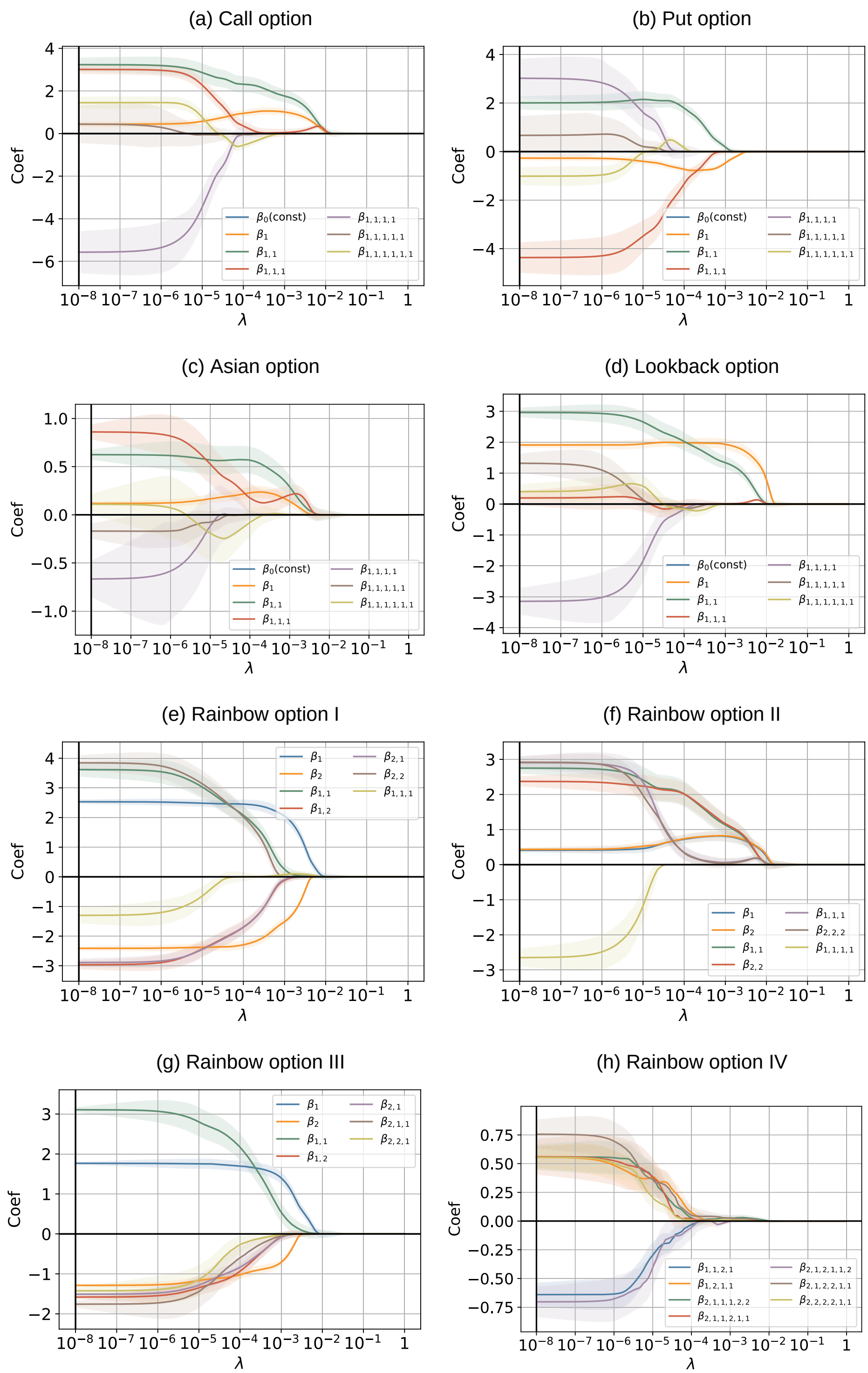
Figure 5. (Color online) In-Sample and Out-of-Sample R^2 for Learning Option Payoffs Using Different Types of Predictors



in-sample and out-of-sample R^2 for Lasso regression with signature components as predictors consistently outperform those for Lasso regression with random sampling and equidistant sampling as predictors. This demonstrates the effectiveness of the signature transform in approximating various nonlinear payoff functions, thanks to its universal nonlinearity.

Figure 6 further shows the Lasso paths as a function of the penalization parameter λ when using signature components as predictors.[16] The fairly narrow range of the 90% confidence intervals of the parameters indicates the stability of the estimated coefficients across repeated experiments, consistent with the uniqueness of the universal nonlinearity of signature (Theorem 2).

Figure 6. (Color online) Lasso Paths with Signatures as Predictors



5.1.2. Comparison Between Different Signatures. We further compare the performance of the Itô and the Stratonovich signatures in learning option payoffs. Given a discrete time series of an underlying asset $\{X_{t_j}\}_{j=1}^{1,000}$ with $t_j = j/1,000$, we consider three numerical methods to calculate the signatures, summarized in Table 1. The first two are numerical methods for computing the Itô and Stratonovich integrals, respectively.[17] The third method, called Linear, linearly interpolates the time series and then calculates the signature using Riemann/Lebesgue integrals, which is widely adopted in practice (Lyons and McLeod 2022).

Table 1. Methods for Computing Signature for a Discrete Time Series $\{\mathbf{X}_{t_j}\}$

Method	Formula
Itô	$S(\mathbf{X})^{i_1,\ldots,i_k,I}_{t_n} = \sum_{j=0}^{n-1} S(\mathbf{X})^{i_1,\ldots,i_{k-1},I}_{t_j}(X^{i_k}_{t_{j+1}} - X^{i_k}_{t_j})$
Stratonovich	$S(\mathbf{X})^{i_1,\ldots,i_k,S}_{t_n} = \sum_{j=0}^{n-1} \frac{1}{2}(S(\mathbf{X})^{i_1,\ldots,i_{k-1},S}_{t_j} + S(\mathbf{X})^{i_1,\ldots,i_{k-1},S}_{t_{j+1}})(X^{i_k}_{t_{j+1}} - X^{i_k}_{t_j})$
Linear	Linearly interpolate $\{\mathbf{X}_{t_j}\}$ and compute signature using Riemann/Lebesgue integral

We simulate two different types of processes for the underlying asset: a one-dimensional standard Brownian motion and a one-dimensional standard OU process with mean-reverting parameter $\kappa = 1$. As an example, we consider the payoff $\max(X_T, 0)$ with time to maturity $T = 1$. Similar to the settings in Section 5.1.1, the training set for the Lasso regression consists of 200 simulated paths, and the test set consists of 100 simulated paths. Each experiment is repeated 200 times, and the average out-of-sample R^2 is shown in Figure 7.[18]

Results. First, the Itô signature of the Brownian motion outperforms its Stratonovich signature in Lasso regression. Second, the Stratonovich signature of the OU process outperforms its Itô signature. These findings are consistent with our theoretical results in Section 3.2 that the Itô signature components of a Brownian motion are more uncorrelated compared with the Stratonovich signature, as well as our simulation results in Section 4.3. Third, the performance of the Linear method in Table 1 is almost identical to the results for the Stratonovich signature, suggesting that for a mean-reverting time series, it may be reasonable to use the heuristic of linearly interpolating the time series and then calculating the signature using Riemann/Lebesgue integrals as in Lyons and McLeod (2022). If the underlying time series is closer to the path of a Brownian motion, using the Itô signature may lead to improved performance compared with current heuristics.

5.2. Option Pricing

The effectiveness of the signature in learning option payoffs in the previous section suggests a new way to price and hedge options.[19] In this section, we follow the method proposed by Lyons et al. (2019) to use the signature to price stock options and interest rate options, which are two of the most important and widely traded options in the equity and fixed income markets, respectively. In addition, the dynamics of the prices of their underlying assets have different statistical properties—the former resembles a Brownian motion, whereas the latter resembles a mean-reverting process.

5.2.1. Method.

We consider a set of m options actively traded in the market with different payoffs $A_1, A_2, \ldots, A_m$, and their prices are observable (referred to as "source options"). Our goal is to determine the prices of a different set of n options with payoffs $B_1, B_2, \ldots, B_n$ (referred to as "target options"). Both sets of options share the same underlying asset with path $\mathbf{X}_t$. Therefore, their payoffs, A_i and B_j, are (different) functions of $\mathbf{X}$. For simplicity, we assume that they also share the same maturity.

To price the target options, we consider the first K signature components of the underlying asset, $S_0(\mathbf{X}), S_1(\mathbf{X}), \ldots, S_K(\mathbf{X})$. The universal nonlinearity implies that

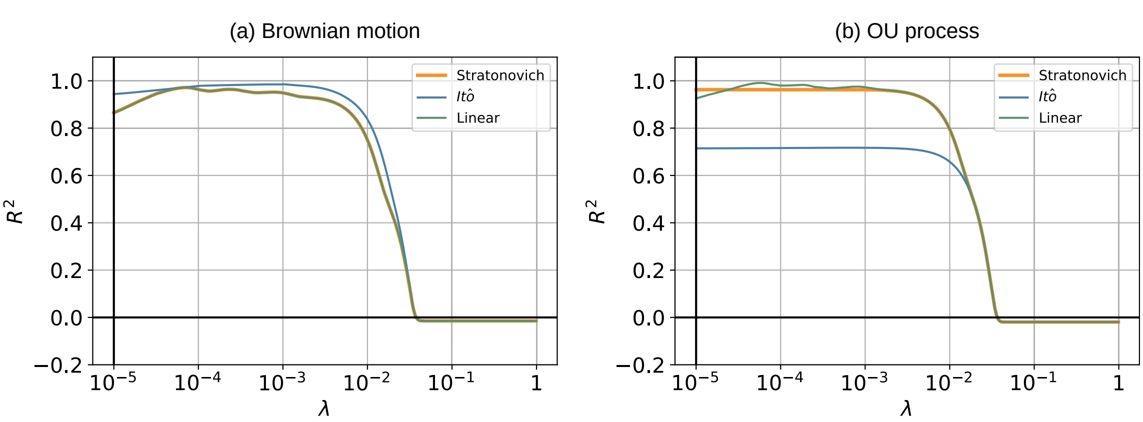
$$A_i(\mathbf{X}) \approx a_{i,0}S_0(\mathbf{X}) + a_{i,1}S_1(\mathbf{X}) + \cdots + a_{i,K}S_K(\mathbf{X}), \quad i = 1, 2, \ldots, m \tag{18}$$

and

$$B_j(\mathbf{X}) \approx b_{j,0}S_0(\mathbf{X}) + b_{j,1}S_1(\mathbf{X}) + \cdots + b_{j,K}S_K(\mathbf{X}), \quad j = 1, 2, \ldots, n. \tag{19}$$

The coefficients $a_{i,\cdot}$ and $b_{j,\cdot}$ can be estimated using Lasso regression based on data from both sets of options because their payoffs $A_i(\mathbf{X})$ and $B_j(\mathbf{X})$ are known given

Figure 7. (Color online) Out-of-Sample R^2 When Using Different Methods for Computing Signature Given in Table 1



underlying paths. Financial assets with identical payoffs must have identical prices assuming no arbitrage, thus from (18) and (19), we obtain the prices of the options

$$p(A_i) \approx a_{i,0}p(S_0) + a_{i,1}p(S_1) + \cdots + a_{i,K}p(S_K), \quad i = 1, 2, \ldots, m \tag{20}$$

and

$$p(B_j) \approx b_{j,0}p(S_0) + b_{j,1}p(S_1) + \cdots + b_{j,K}p(S_K), \quad j = 1, 2, \ldots, n, \tag{21}$$

where $p(\cdot)$ denotes the price of a derivative. Because $p(A_i)$ are observable from source options, we can estimate $p(S_0), \ldots, p(S_K)$ using (20) and then predict $p(B_j)$ for target options using (21).

We point out that this method is interpretable because the signature linearizes the problem of feature selection. In particular, $p(S_0), p(S_1), \ldots, p(S_K)$ can be understood as the prices of K latent derivatives whose payoff functions are given by the first K signature components of the underlying asset, $S_0, S_1, \ldots, S_K$. This is analogous to the Arrow–Debreu state prices of the arbitrage pricing theory of Ross (1976), and therefore we refer to these latent derivatives as "signature derivatives."

Based on this framework, we summarize the procedure for estimating $p(B_j)$ in the following steps.

i. Simulate N paths of $\mathbf{X}$: $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$;

ii. Calculate the payoffs of source options $A_i(\mathbf{X}_1)$, $A_i(\mathbf{X}_2), \ldots, A_i(\mathbf{X}_N)$ for $i = 1, 2, \ldots, m$, and target options $B_j(\mathbf{X}_1), B_j(\mathbf{X}_2), \ldots, B_j(\mathbf{X}_N)$ for $j = 1, 2, \ldots, n$;

iii. Calculate the corresponding signature $S_k(\mathbf{X}_1)$, $S_k(\mathbf{X}_2), \ldots, S_k(\mathbf{X}_N)$ for $k = 0, 1, \ldots, K$;

iv. For each i or j, estimate $a_{i,\cdot}$ and $b_{j,\cdot}$ using Lasso regression based on (18) and (19), respectively, where the predictors are $S_k(\mathbf{X}_1), S_k(\mathbf{X}_2), \ldots, S_k(\mathbf{X}_N)$ for $k = 0, 1, \ldots, K$, the dependent variables are $A_i(\mathbf{X}_1), \ldots, A_i(\mathbf{X}_N)$ or $B_j(\mathbf{X}_1), \ldots, B_j(\mathbf{X}_N)$, and the regression is performed with N samples;

v. Estimate the prices of signature derivatives $p(S_0)$, $p(S_1), \ldots, p(S_K)$ using ordinary least squares based on (20), where the predictors are $a_{i,\cdot}$, the dependent variables are $p(A_i)$, and the estimation uses m samples; and

vi. For each j, calculate the price of the target option $p(B_j)$ using (21) directly.

5.2.2. Stock Options. Following the Black–Scholes–Merton framework (Black and Scholes 1973, Merton 1973), we assume that the underlying asset $\mathbf{X}$ follows a geometric Brownian motion in the risk-neutral world with initial price 100, risk-free rate 2%, dividend yield 0%, and volatility 20%. The source options are vanilla European calls and puts with strikes at $90, 92, 94, \ldots, 110$ ($m = 22$), priced by the Black–Scholes–Merton formula. The target options are vanilla European calls and puts with strikes at $91, 93, 95, \ldots, 109$ ($n = 20$), and their true prices are determined using the Black–Scholes–Merton formula. The time to maturity for all these options is set to be 2.5 years.

The simulation is conducted as follows. We simulate N = 1,000 paths for the underlying asset, and the step size for simulating the underlying path is $1/252$ (one trading day). For each Lasso regression, signature components with orders up to four are used as predictors (K = 4), and the penalization parameter is determined using fivefold cross-validation.

The experiment is repeated 100 times, and the estimation error is computed for each experiment. The relative error is measured using the average of $|\hat{p}(B_j) - p(B_j)|/p(B_j)$ across all target options, where $\hat{p}(B_j)$ and $p(B_j)$ are the estimated price and true price of B_j, respectively.

Results. Using signature with Lasso regression provides an excellent fit for the prices of stock options. Figure 8 shows the true prices and the estimated prices of the target options for a randomly chosen experiment out of 100, showing that the estimation errors are small for both the Itô and Stratonovich signatures.

Figure 8. (Color online) Estimated Prices vs. True Prices for Stock Options

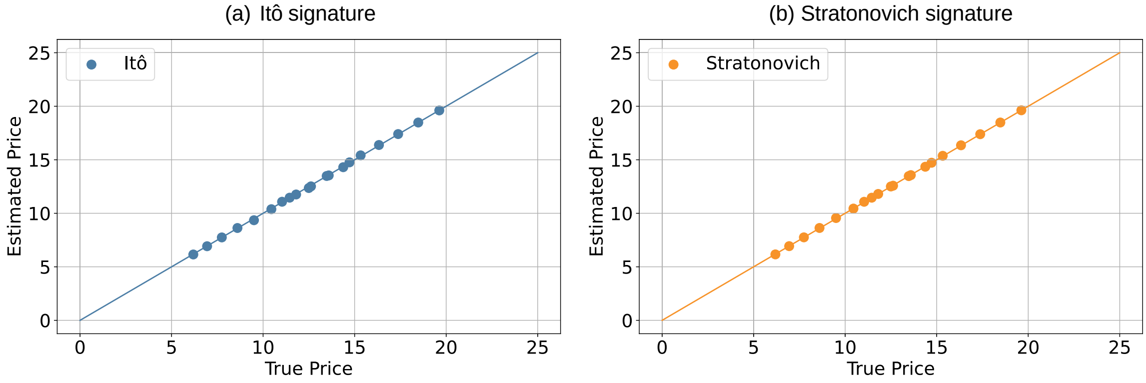
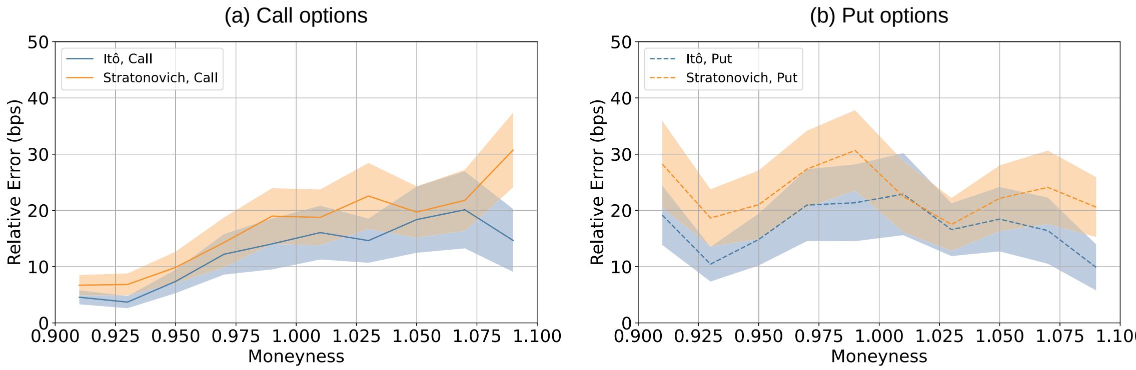


Figure 9. (Color online) Estimation Errors for Different Target Stock Options



In addition, the estimation error of stock option prices when using the Itô signature is lower compared with using the Stratonovich signature, consistent with our results in Section 4.3. Figure 9 shows the average relative errors for the Itô and the Stratonovich signatures across the 100 experiments. The x axis represents the moneyness (the strike price of the option divided by the initial asset price) of the target options, and the y axis is the average relative error. Note that the estimation error of the Itô signature is lower because the underlying price process resembles a Brownian motion.

5.2.3. Interest Rate Options. We now turn to interest rate options, whose underlying assets are commonly modeled using mean-reverting processes. In particular, consider the interest rate processes $\{r_t\}_{t\geq 0}$ by the classical Vasicek model (Vasicek 1977) in the risk-neutral world[20]:

$$\mathrm{d}r_t = \gamma(\overline{r} - r_t)\mathrm{d}t + \sigma \mathrm{d}W_t,$$

with initial rate $r_0 = 3\%$, long-term average interest rate $\overline{r} = 3\%$, mean-reverting intensity $\gamma = 0.1$, volatility $\sigma = 2\%$, and W_t a standard Brownian motion. The source options are interest rate caplets and floorlets with strikes $r_{\text{strike}} = 2.50\%, 2.60\%, \ldots, 3.40\%, 3.50\%$ ($m = 22$), and their prices $p(A_i)$ are determined using explicit formulas for caplets and floorlets under the Hull–White model (Veronesi 2010). The target options are interest rate caplets and floorlets with strikes $r_{\text{strike}} = 2.55\%, 2.65\%, \ldots, 3.35\%, 3.45\%$ ($n = 20$). The payoffs for caplets and floorlets are $\max(r(0.5,1) - r_{\text{strike}}, 0)$ and $\max(r_{\text{strike}} - r(0.5,1), 0)$, respectively, where $r(0.5,1)$ is the 0.5-year interest rate at time 0.5. Assume that each of these instruments has a notional value of \$100 and a maturity of 0.5 years. Other simulation setups are the same as in Section 5.2.2, and the experiment is repeated 100 times.

Results. Similar to the case of stock options, using signatures with Lasso regression provides an excellent fit for the prices of interest rate options. Figure 10 shows the actual and estimated prices of the target options for a randomly chosen experiment out of 100. The actual prices are determined using explicit formulas for caplets and floorlets under the Hull–White model. The prices estimated using both the Itô and Stratonovich signatures closely align with the actual prices.

Furthermore, in contrast to the case of stock options, the estimation error of interest rate option prices when using the Itô signature is higher compared with the Stratonovich signature, as shown in Figure 11. This is also consistent with our results in Section 4.3.

Overall, our results demonstrate that the Lasso regression with signature is effective in learning nonlinear

Figure 10. (Color online) Estimated Prices vs. True Prices for Interest Rate Options

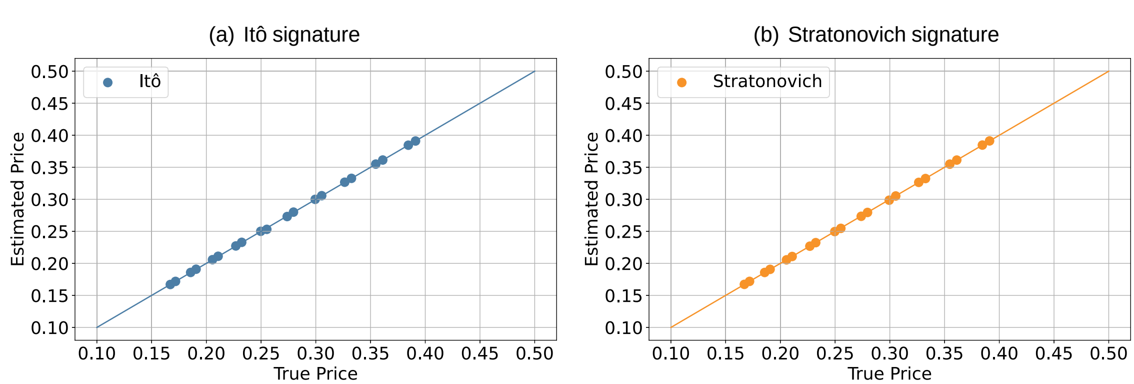
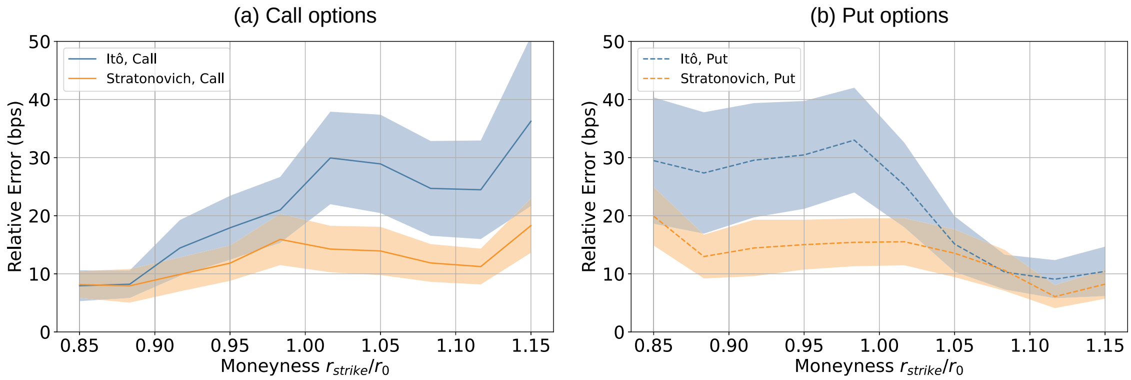


Figure 11. (Color online) Estimation Errors for Different Target Interest Rate Options



payoff functions. In addition, the statistical properties of different types of signatures suggest that the Itô signature is more appropriate if the underlying asset resembles a Brownian motion, and the Stratonovich signature is better if the underlying asset resembles a mean-reverting process.

6. Conclusion

This paper studies the statistical consistency of Lasso regression with signatures. We first establish a probabilistic uniqueness of the universal nonlinearity, which implies that any feature selection procedure needs to recover this unique linear combination of signature to achieve good predictive performance.

We find that consistency is highly dependent on the definition of the signature, the characteristics of the underlying processes, and the correlation between different dimensions of the underlying process. In particular, the Itô signature performs better when the underlying process is closer to the Brownian motion and has weaker interdimensional correlations, whereas the Stratonovich signature performs better when the process is sufficiently mean reverting.

The signature method offers an attractive interpretable framework for machine learning and pattern recognition. In fact, the first two orders of signature components correspond to the Lévy area of sample paths (Chevyrev and Kormilitzin 2016, Levin et al. 2016). In addition, the fact that the target variable can be represented as a linear function of signature components allows for interpretability with respect to the underlying features, and we offer an example in the context of option pricing (Section 5.2).

In general, these results highlight the importance of choosing the appropriate signature for different underlying data, in terms of both learning the right coefficients for interpretation and achieving predictive performance.

Our findings also call for further studies on the statistical properties of the signature before its potential in machine learning can be fully realized. First, in addition to signature, logsignature is also a widely used transform of the path of a stochastic process, which has been shown empirically to improve the training efficiency with simpler and less redundant information of the path (Morrill et al. 2020a, 2021). However, logsignature does not enjoy universal nonlinearity (Morrill et al. 2020a, Lyons and McLeod 2022), which implies that there is no theoretical basis for using a linear combination of logsignature to approximate a nonlinear function. Therefore, we choose to focus on signature in this study and defer the statistical properties of logsignature to future work.

Second, our results highlight the differences in the statistical performance of Itô and Stratonovich signatures under different probabilistic models of the underlying path. This raises a natural question: Is it possible to construct an intermediate signature transform between Itô and Stratonovich signatures? The Itô and Stratonovich integrals are defined using the left endpoints and midpoints of partition subintervals, respectively. Therefore, one may also consider a class of other stochastic integrals using other points within the subintervals (Karatzas and Shreve 1998). The location of these points may serve as a tuning parameter, which can be selected in practice using techniques such as cross-validation.[21] This approach may balance the statistical advantages of Itô and Stratonovich signatures. However, further investigation is needed to determine whether the signatures defined by these new types of integrals satisfy universal nonlinearity.

Finally, the theoretical analysis in this study focuses on the statistical consistency of the parameters of each signature component with respect to the number of sample paths, using signatures computed from continuous paths. In practice, the computation of each signature component relies on a discrete sample of the continuous path, which may introduce additional errors. The implications of this discretization on the statistical performance of signature are left for future study.

Acknowledgments

The authors thank Yacine Ait-Sahalia, Jose Blanchet, Steven Campbell, Matias Cattaneo, Nan Chen, Kay Giesecke, Paul Glasserman, Boris Hanin, Xuedong He, Jason Klusowski, Ciamac Moallemi, Hao Ni, Marcel Nutz, Markus Pelger, Shige Peng, Philip Protter, Ruodu Wang, Chao Zhang, Yufei Zhang, and seminar and conference participants at the Annual Conference of the Institute for Operations Research and the Management Sciences 2024, the International Conference of Industrial and Applied Mathematics 2023, Bernoulli-Institute of Mathematical Statistics 11th World Congress in Probability and Statistics, the First INFORMS Conference on Financial Engineering and FinTech, the 2024 Workshop on Stochastic Control, Machine Learning and Quantitative Finance at Shanghai Jiao Tong University, the 2024 Workshop on Mathematical Finance and Insurance at Peking University, Princeton University, Stanford University, Columbia University, University of Waterloo, University of Oxford, and London School of Economics and Political Science for helpful comments and discussion.

Endnotes

[1] Other examples include handwriting recognition (Yang et al. 2016b, c; Wilson-Nunn et al. 2018; Kidger et al. 2020; Ibrahim and Lyons 2022) and action recognition (Yang et al. 2016a, Li et al. 2017, Fermanian 2021, Lee et al. 2022, Yang et al. 2022, Cheng et al. 2024).

[2] Király and Oberhauser (2019) explore the statistical properties of signature in the context of kernel learning for paths, whereas Morrill et al. (2020a) summarize the statistical characteristic that signature can be regarded as an analog of the moment-generating function for time series.

[3] Examples include Lyons (2014), Chevyrev and Kormilitzin (2016), Levin et al. (2016), Moore et al. (2019), Sugiura and Hosoda (2020), Lemercier et al. (2021), Sugiura and Kouketsu (2021), Lyons and McLeod (2022), Bleistein et al. (2023), Cuchiero et al. (2023), and Lemahieu et al. (2023).

[4] In this paper, we mainly consider $\mathbf{X}_t$ as a continuous-time process for the neatness of our theoretical analysis. However, our simulations and numerical applications in Sections 4 and 5 demonstrate that our theoretical results are also applicable to discrete time series.

[5] Signature also enjoys several other nice probabilistic properties under mild conditions. First, all expected signature components of a stochastic process characterize the distribution of the process (Chevyrev and Lyons 2016, Chevyrev and Oberhauser 2022). Second, the signature of a process uniquely determines the path of the underlying process up to a tree-like equivalence (Hambly and Lyons 2010, Le Jan and Qian 2013, Boedihardjo et al. 2014).

[6] See Online Appendix A for details of the time augmentation.

[7] We perform this standardization for two reasons. First, the Lasso estimator is sensitive to the magnitudes of the predictors (Hastie et al. 2009) and the magnitudes of different orders of signature components are different (Lyons et al. 2007); therefore, standardization is necessary to ensure that the coefficients of different orders of signature are on the same scale and can be compared directly. Second, the sample covariance matrix is now equivalent to the sample correlation matrix, allowing us to focus on the correlation structure of the signature components in the subsequent analysis.

[8] In this paper, in line with Zhao and Yu (2006), all covariances and correlation coefficients are defined to be uncentered. Specifically, for random variables X and Y, we define their covariance as $\mathbb{E}[XY]$, and their correlation coefficient as $\mathbb{E}[XY]/\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$. One can easily extend our results to the centered case.

[9] This uniqueness is only for a fixed value of K. In fact, if L_a and L_b are linear combinations of signature components with orders truncated to K and $K+1$, respectively, the true function f could always be better approximated by L_b.

[10] For the Stratonovich signature of a d-dimensional Brownian motion, $\Psi_{m,n}$ is given by Proposition B.1 in Online Appendix B.

[11] This statement aligns with the convention of the literature of high-dimensional statistics; see, for example, Wainwright (2009), Ravikumar et al. (2011), and Vershynin (2018).

[12] Propositions 3 and 4 follow directly from taking partial derivatives with respect to various parameters, and we therefore omit their proofs.

[13] Online Appendix D provides technical details, computational cost, more numerical experiments, and robustness checks for the simulations conducted in this section.

[14] The choice of $d = 2$ is consistent with the simulation setup commonly used in the literature on signatures; see, for example, Chevyrev and Kormilitzin (2016).

[15] As a robustness check, we have also conducted experiments for $K \in \{3,4,5,6,7,8,9,10\}$ if $d = 1$ and $K \in \{3,4,5,6,7\}$ if $d = 2$. The results are similar to the case of $K = 6$.

[16] For rainbow options with two-dimensional underlying price processes, due to their large number of predictors in the Lasso regression, we only show the seven predictors with the largest estimated coefficients.

[17] Online Appendix D provides technical details for the schemes of numerically computing both integrals.

[18] We omit the values of in-sample R^2 as they are very close to the values of out-of-sample R^2.

[19] For example, Arribas (2018) and Lyons et al. (2019) price options using securities whose payoffs are signature. Lyons et al. (2020) use the signature transform to perform option hedging. Bayraktar et al. (2024) propose a deep learning framework to price options using signature.

[20] This is also known as a special case of the Hull–White model (Hull and White 1990); see, for example, Veronesi (2010).

[21] Specifically, given a tuning parameter $\tau \in [0,1]$, one may define the stochastic integral $\int_0^T A_t \mathrm{d}B_t$ as the limit of $\sum_{j=0}^{n-1}(\tau \cdot A_{t_j} + (1-\tau) \cdot A_{t_{j+1}})(B_{t_{j+1}} - B_{t_j})$ with $0 = t_0 < t_1 < \ldots < t_n = T$ as a partition of $[0,T]$. This reduces to the Itô integral and Stratonovich integral when $\tau = 1$ and $\tau = 0.5$, respectively (Karatzas and Shreve 1998). The tuning parameter τ can be chosen through model selection techniques such as cross-validation.

References

Akyildirim E, Gambara M, Teichmann J, Zhou S (2022) Applications of signature methods to market anomaly detection. Preprint, submitted January 7, https://arxiv.org/abs/2201.02441.

Arashi M, Asar Y, Yüzbaşı B (2021) SLASSO: A scaled LASSO for multicollinear situations. *J. Statist. Comput. Simulations* 91(15):3170–3183.

Arribas IP (2018) Derivatives pricing using signature payoffs. Preprint, submitted September 25, https://arxiv.org/abs/1809.09466.

Asmussen S (2003) *Applied Probability and Queues*, vol. 2 (Springer, Berlin).

Bayraktar E, Feng Q, Zhang Z (2024) Deep signature algorithm for multidimensional path-dependent options. *SIAM J. Financial Math.* 15(1):194–214.

Bertsimas D, Kogan L, Lo AW (2001) Hedging derivative securities and incomplete markets: An ϵ-arbitrage approach. *Oper. Res.* 49(3):372–397.

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* 37(4):1705–1732.

Black F, Scholes M (1973) The pricing of options and corporate liabilities. *J. Political Econom.* 81(3):637–654.

Bleistein L, Fermanian A, Jannot AS, Guilloux A (2023) Learning the dynamics of sparsely observed interacting systems. *Proc. 40th Internat. Conf. Machine Learn.* (PMLR, New York), 2603–2640.

Boedihardjo H, Ni H, Qian Z (2014) Uniqueness of signature for simple curves. *J. Functional Anal. (Oxford)* 267(6):1778–1806.

Chen KT (1954) Iterated integrals and exponential homomorphisms. *Proc. London Math. Soc.* 3(1):502–512.

Chen KT (1957) Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula. *Ann. Math.* 65(1):163–178.

Cheng J, Shi D, Li C, Li Y, Ni H, Jin L, Zhang X (2024) Skeleton-based gesture recognition with learnable paths and signature features. *IEEE Trans. Multimedia* 26(1):3951–3961.

Chevyrev I, Kormilitzin A (2016) A primer on the signature method in machine learning. Preprint, submitted March 11, https://arxiv.org/abs/1603.03788.

Chevyrev I, Lyons T (2016) Characteristic functions of measures on geometric rough paths. *Ann. Probability* 44(6):4049–4082.

Chevyrev I, Oberhauser H (2022) Signature moments to characterize laws of stochastic processes. *J. Machine Learn. Res.* 23(176):1–42.

Cuchiero C, Gazzani G, Svaluto-Ferro S (2023) Signature-based models: Theory and calibration. *SIAM J. Financial Math.* 14(3):910–957.

Cuchiero C, Guo X, Primavera F (2025) Funtional Itô formula and Taylor expansion of non-anticipative maps of rough paths. Preprint, submitted April 8, https://arxiv.org/abs/2504.06164.

Dupire B, Tissot-Daguette V (2022) Functional expansions. Preprint, submitted December 27, https://arxiv.org/abs/2212.13628.

Fermanian A (2021) Embedding and learning with signatures. *Comput. Statist. Data Anal. (Oxford)* 157:107148.

Fermanian A (2022) Functional linear regression with truncated signatures. *J. Multivariate Anal. (Oxford)* 192:105031.

Friz PK, Victoir NB (2010) *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*, vol. 120 (Cambridge University Press, Cambridge, UK).

Futter O, Horvath B, Wiese M (2023) Signature trading: A path-dependent extension of the mean-variance framework with exogenous signals. Preprint, submitted August 29, https://arxiv.org/abs/2308.15135.

Gatheral J, Jaisson T, Rosenbaum M (2018) Volatility is rough. *Quant. Finance* 18(6):933–949.

Gu H, Guo X, Jacobs TL, Kaminsky P, Li X (2024) Transportation marketplace rate forecast using signature transform. *Proc. 30th ACM SIGKDD Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 4997–5005.

Hambly B, Lyons T (2010) Uniqueness for the signature of a path of bounded variation and the reduced path group. *Ann. Math.* 171(1):109–167.

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2 (Springer, New York).

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. *Adv. Neural Inform. Processing Systems* 33:6840–6851.

Hull J, White A (1990) Pricing interest-rate-derivative securities. *Rev. Financial Stud.* 3(4):573–592.

Hunt G (2007) The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. *Proc. Natl. Acad. Sci. USA* 104(47):18404–18408.

Hutchinson JM, Lo AW, Poggio T (1994) A nonparametric approach to pricing and hedging derivative securities via learning networks. *J. Finance* 49(3):851–889.

Ibrahim MR, Lyons T (2022) ImageSig: A signature transform for ultra-lightweight image recognition. *Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition* (IEEE, Piscataway, NJ), 3649–3659.

Kalsi J, Lyons T, Arribas IP (2020) Optimal execution with rough path signatures. *SIAM J. Financial Math.* 11(2):470–493.

Karatzas I, Shreve SE (1998) *Brownian Motion and Stochastic Calculus*, vol. 113, 2nd ed. (Springer, New York).

Kidger P, Morrill J, Foster J, Lyons T (2020) Neural controlled differential equations for irregular time series. *Adv. Neural Inform. Processing Systems* 33:6696–6707.

Király FJ, Oberhauser H (2019) Kernels for sequentially ordered data. *J. Machine Learn. Res.* 20:1–45.

Kormilitzin A, Saunders KE, Harrison PJ, Geddes JR, Lyons T (2017) Detecting early signs of depressive and manic episodes in patients with bipolar disorder using the signature-based model. Preprint, submitted August 3, https://arxiv.org/abs/1708.01206.

Le Jan Y, Qian Z (2013) Stratonovich's signatures of Brownian motion determine Brownian sample paths. *Probability Theory Related Fields* 157:209–223.

Lee J, Jeon J, Yon Jhin S, Hyeong J, Kim J, Jo M, Seungji K, Park N (2022) LORD: Lower-dimensional embedding of log-signature in neural rough differential equations. *Proc. Internat. Conf. Learn. Representations* (OpenReview.net).

Lemahieu E, Boudt K, Wyns M (2023) Generating drawdown-realistic financial price paths using path signatures. Preprint, submitted September 8, https://arxiv.org/abs/2309.04507.

Lemercier M, Salvi C, Damoulas T, Bonilla E, Lyons T (2021) Distribution regression for sequential data. *Proc. Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 3754–3762.

Leng C, Wang H (2009) On general adaptive sparse principal component analysis. *J. Comput. Graphical Statist.* 18(1):201–215.

Levin D, Lyons T, Ni H (2016) Learning from the past, predicting the statistics for the future, learning an evolving system. Preprint, submitted March 22, https://arxiv.org/abs/1309.0260.

Li C, Zhang X, Jin L (2017) LPSNet: A novel log path signature feature based hand gesture recognition framework. *Proc. IEEE Internat. Conf. Computer Vision Workshops* (IEEE, Piscataway, NJ), 631–639.

Lyons T (2014) Rough paths, signatures and the modelling of functions on streams. Preprint, submitted May 18, https://arxiv.org/abs/1405.4537.

Lyons T, McLeod AD (2022) Signature methods in machine learning. Preprint, submitted June 29, https://arxiv.org/abs/2206.14674.

Lyons T, Caruana M, Lévy T (2007) *Differential Equations Driven by Rough Paths* (Springer, New York).

Lyons T, Nejad S, Arribas IP (2019) Numerical method for model-free pricing of exotic derivatives in discrete time using rough path signatures. *Appl. Math. Finance* 26(6):583–597.

Lyons T, Nejad S, Arribas IP (2020) Non-parametric pricing and hedging of exotic derivatives. *Appl. Math. Finance* 27(6):457–494.

Lyons T, Ni H, Oberhauser H (2014) A feature set for streams and an application to high-frequency financial tick data. *Proc. Internat. Conf. Big Data Sci. Comput.* (ACM, New York), 1–8.

Martins EP (1994) Estimating the rate of phenotypic evolution from comparative data. *Amer. Naturalist* 144(2):193–209.

Merton RC (1973) Theory of rational option pricing. *Bell J. Econom. Management Sci.* 4(1):141–183.

Moore P, Lyons T, Gallacher J, Initiative ADN (2019) Using path signatures to predict a diagnosis of Alzheimer's disease. *PLoS One* 14(9):e0222212.

Moreno-Pino F, Arroyo Á, Waldon H, Dong X, Cartea Á (2024) Rough transformers: Lightweight continuous-time sequence modelling with path signatures. *Proc. 38th Ann. Conf. Neural Inform. Processing Systems* (OpenReview,net).

Morrill J, Fermanian A, Kidger P, Lyons T (2020a) A generalised signature method for multivariate time series feature extraction. Preprint, submitted June 1, https://arxiv.org/abs/2006.00873.

Morrill J, Salvi C, Kidger P, Foster J (2021) Neural rough differential equations for long time series. *Proc. Internat. Conf. Machine Learn.* (PMLR, New York), 7829–7838.

Morrill J, Kormilitzin A, Nevado-Holgado A, Swaminathan S, Howison S, Lyons T (2019) The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. *Proc. Comput. Cardiology* (IEEE, Piscataway, NJ).

Morrill J, Kormilitzin A, Nevado-Holgado A, Swaminathan S, Howison SD, Lyons T (2020b) Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring. *Critical Care Medicine* 48(10):e976–e981.

Pan Y, Lu M, Shi Y, Zhang H (2023) A path signature approach for speech-based dementia detection. *IEEE Signal Processing Lett.* (IEEE, Piscataway, NJ), 1–5.

Ravikumar P, Wainwright MJ, Raskutti G, Yu B (2011) High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence. *Electronic J. Statist.* 5:935–980.

Ross SA (1976) The arbitrage theory of capital asset pricing. *J. Econom. Theory* 13(3):341–360.

Salvi C, Lemercier M, Liu C, Horvath B, Damoulas T, Lyons T (2021) Higher order kernel mean embeddings to capture filtrations of stochastic processes. *Adv. Neural Inform. Processing Systems* 34:16635–16647.

Song Y, Ermon S (2019) Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inform. Processing Systems*, vol. 32 (MIT Press, Cambridge, MA), 11895–11907.

Sugiura N, Hosoda S (2020) Machine learning technique using the signature method for automated quality control of argo profiles. *Earth Space Sci.* 7(9):e2019EA001019.

Sugiura N, Kouketsu S (2021) Simple el niño prediction scheme using the signature of climate time series. Preprint, submitted September 5, https://arxiv.org/abs/2109.02013.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B (Methodological)* 58(1):267–288.

Vasicek O (1977) An equilibrium characterization of the term structure. *J. Financial Econom.* 5(2):177–188.

Veronesi P (2010) *Fixed Income Securities: Valuation, Risk, and Risk Management* (John Wiley & Sons, Hoboken, NJ).

Vershynin R (2018) *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47 (Cambridge University Press, Cambridge, UK).

Wainwright MJ (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* 55(5):2183–2202.

Wilson-Nunn D, Lyons T, Papavasiliou A, Ni H (2018) A path signature approach to online Arabic handwriting recognition. *Proc. IEEE 2nd Internat. Workshop Arabic Derived Script Analysis Recognition* (IEEE, Piscataway, NJ), 135–139.

Yang W, Jin L, Liu M (2016a) DeepWriterID: An end-to-end online text-independent writer identification system. *IEEE Intelligent Systems* 31(2):45–53.

Yang W, Jin L, Ni H, Lyons T (2016b) Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network. *Proc. 23rd Internat. Conf. Pattern Recognition* (IEEE, Piscataway, NJ), 4083–4088.

Yang W, Jin L, Tao D, Xie Z, Feng Z (2016c) DropSample: A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition. *Pattern Recognition* 58:190–203.

Yang W, Lyons T, Ni H, Schmid C, Jin L (2022) Developing the path signature methodology and its application to landmark-based human action recognition. Yin G, Zariphopoulou T, eds. *Stochastic Analysis, Filtering, and Stochastic Optimization: A Commemorative Volume to Honor Mark HA Davis's Contributions* (Springer, Berlin), 431–464.

Zhang Z, Si X, Hu C, Lei Y (2018) Degradation data analysis and remaining useful life estimation: A review on Wiener-process-based methods. *Eur. J. Oper. Res.* 271(3):775–796.

Zhao P, Yu B (2006) On model selection consistency of Lasso. *J. Machine Learn. Res.* 7:2541–2563.

Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. *J. Comput. Graphical Statist.* 15(2):265–286.

Xin Guo is the Coleman Fung chair professor in financial modeling at the University of California, Berkeley, Department of Industrial Engineering and Operations Research, an affiliated faculty member of the University of California, Berkeley-University of California, San Francisco, Precision Health Program, and an Amazon Scholar. Her research interests include stochastic controls and games, mean field games, machine learning theory and applications, medical data analysis, mathematical finance and fintech, and supply chains.

Binnan Wang is a PhD student at the School of Mathematical Sciences at Peking University. His research interests include path signatures, statistical theory of machine learning, explainable artificial intelligence, and quantitative finance.

Ruixun Zhang is an assistant professor and Boya Young fellow at the School of Mathematical Sciences at Peking University (PKU), with affiliations at the PKU Center for Statistical Science, the PKU Laboratory for Mathematical Economics and Quantitative Finance, and the PKU National Engineering Laboratory for Big Data Analysis and Applications. His research interests include machine learning, sustainable investing, market microstructure, and evolutionary foundations of economic behavior and financial markets.

Chaoyi Zhao is a postdoctoral associate at the MIT Sloan School of Management and the MIT Laboratory for Financial Engineering. His research interests include investments, high-dimensional statistics, machine learning applications in finance, healthcare finance, and actuarial sciences.