RESEARCH ON REGIME-SWITCHING BETWEEN DIFFERENT STOCHASTIC DYNAMICAL SYSTEMS

Tianyuan Zhou

Bachelor of Science, Fudan University, 2015

A Dissertation submitted to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Statistics

University of Virginia
May 2023

Prof. Jordan Rodu, Chair
Prof. João Sedoc
Prof. Scott Doney
Prof. Taylor Brown

Research on Regime-Switching Between Different Stochastic Dynamical Systems

Tianyuan Zhou

(ABSTRACT)

Regime-Switching (RS) is an important phenomenon and modeling technique in time series analysis that the observed process follows different patterns during different time periods. To the best of our knowledge, there is no study on the RS phenomenon between different dynamical systems. In this dissertation, we finished 3 projects on this topic. In Project I, we studied the RS modeling between different stochastic dynamical systems with known parameter forms. We proposed a heteroskedasticity-based E-M algorithm to infer this model since it cannot be estimated under the likelihood framework. We also proposed a hypothesis testing procedure, named as RS testing, to test whether the RS phenomenon exists or not through testing whether the state prediction agrees with the observation or not. We demonstrated its power by comprehensive simulations and proved that VIX has the RS phenomenon. In Project II, we extended this method to the Realized Variance (RV) processes of the stock market. To calculate RV, we proposed a novel data cleaning method for the TAQ transaction-level dataset to achieve a better trade-off between the data quality and data size. We proved that the RS phenomenon is universal in the stock market's volatility. In Project III, we studied the RS phenomenon in scientific processes. The state-of-the-art forecasting method for scientific processes is time-invariant Empirical Dynamic Modeling (EDM). In this project, we proposed a time-dependent EDM framework and proposed a Periodically-Regime-Switching (PRS) model to combine

the strong periodicity and Markovian property of the latent state process. We proved our method's performance on the chlorophyll forecasting problem. At the end, we made a comprehensive discussion on these 3 projects and summarized the application scenarios.

Keywords: Regime-switching; Stochastic dynamical models; Time series analysis; Stock volatility; Chlorophyll forecasting.

Acknowledgments

First of all, I would like to express my deepest appreciation to my advisor, Prof. Jordan Rodu, for his constant support and guidance at every stage of my research. I was greatly impressed by and will benefit forever from his brilliant ideas, deep knowledge and perceptive advice.

I would like to offer my special thanks to my committee members, Prof. João Sedoc, Prof. Scott Doney and Prof. Taylor Brown, for their invaluable advice, continuous support, and patience during my research. It is impossible to complete my dissertation without their suggestions and efforts.

I'd like to acknowledge all members of the department. They offered me the wonderful opportunity to enroll into the graduate program, where I learned academic knowledge and how to become a better person.

Finally, I want to thank my parents, who always stand with me, give me unconditional love and provide me material and spiritual support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Regime-Switching (RS) is a widespread phenomenon in time series analysis which evolves across different time periods, and the fundamentals could also change. RS occurs with time series during different periods that may have different properties and a single model is insufficient. To represent this kind of time series, we should use different models for different periods, and demonstrate the state transition between different models. These kinds of models are called RS models.

RS modeling is typically done by state-space models. If there is a finite set of basic models that are present during different episodes, then the true model each time stamp is modeled by a latent state process, where at each time stamp there is a latent state controlling which basic model dominates. For example, Hidden Markov Models (HMM) (Baum and Petrie, 1966; Baum and Eagon, 1967; Baum, Petrie, et al., 1970) assume the observed time series is independent conditioned on the latent states which evolves as a Markov chain.

One important kind of time series is present in dynamical systems. Dynamical systems focus on modeling the evolution of time series by modeling the relationship between neighboring observations. For example, AutoRegressive-Moving-Average (ARMA) (Box et al., 2015) models the observations through a linear statistical model of lagged observations and noise errors. More complicated models are modeled by Stochastic Differential Equations (SDE) or Partial Differential Equations (PDE). There are

methods combining the dynamical systems and RS models, such as Markov-Switching AR (Adejumo, Albert, and Asemota, 2020).

In this dissertation, I studied the RS between different dynamical systems as this class of models handle the scenario better than a single model that is insufficient to model the time series across the whole observed time period. RS between dynamical systems assumes there exists a set of dynamical systems and a latent state process specifying the dominant system at each time stamp. These models are useful particularly for financial and scientific processes which will be further explained in this thesis. In these areas, evolution modeled by SDEs or PDEs change across different time periods because of the changing environment, such as the market factors in the financial models and the environmental variables in the water system.

Before discussing our proposed RS methods, we briefly discuss the application scenario for RS models. Theoretically RS models assume the existence of a latent state process controlling the evolution of the observations of each time period. That is to say, in the state transition view, when the underlying state transits, the observations will have a staggered jump characterized by rough and discrete. On the other hand, RS should be viewed as an approximation of a complicated system where each state or regime is modeling one dominant pattern of a time period. With the approximation view, even though the true process is not RS and transition is stable characterized by continuous and smooth, we can still use RS models and model the jumps as staggered, as stated in the famous aphorism by George Box, "All models are wrong, some are useful". If the true data is not RS but too complicated to be modeled sufficiently by basic models, we can use RS models so that during different time periods the dominant patterns can be modeled accurately.

In the financial market, my research focuses on the volatility process, including the im-

plied volatility and realized variance processes. These volatility processes are typically represented by some well-studied models, including a class of models called stochastic volatility models, such as Heston model (S. L. Heston, 1993) and 3/2 model (S. L. Heston, 1997; Platen, 1998). Different models are needed to characterize different patterns for different financial assets during different time periods. In this dissertation, we proposed the RS model between different stochastic volatility models. This class of models cannot be inferred by the Maximal Likelihood Estimation (MLE) by Expectation-Maximization (E-M) algorithm, which is the standard inference method for RS models, due to the indistinguishability of the likelihood. So we propose an algorithm that uses the Bayes factor of the heteroskedasticity test as the emission probability, which we call the heteroskedasticity-based E-M algorithm. Along with this method, we proposed a hypothesis testing procedure named as the RS test to test whether the RS phenomenon exists, by testing Kendall's rank correlation test between the emission and prediction probabilities. We prove the performance of our heteroskedasticity-based E-M algorithm and RS test on simulated datasets and real-world applications including a half-year VIX process and realized volatility processes of SPDR S&P 500 ETF Trust (SPY) and 30 Dow Jones Industrial average (DJI) component stocks during a 7-year period, and prove that the RS phenomenon is universe in volatility processes in the stock market.

In scientific process studies, we conduct research on the chlorophyll forecast, which is important for predicting toxic algal bloom. The chlorophyll process is a part of a complicated water dynamical system involving a small set of observed cross predictors and a large set of unobserved variables. These kinds of problems are usually modeled by Empirical Dynamic Modeling (EDM), which guarantees the existence of an AR type forecasting function, but it suffers from the curse-of-dimensionality problem

when the underlying dynamical system is too complicated. To solve this problem, we proposed a RS-EDM framework, which allows EDM to have a RS structure, so it helps reduce the lag we need. Further we proposed a Periodically-Regime-Switching (PRS) model which assigns a periodical property to Markov switching. The Markovian and periodical properties are contradicting but we resolved it by using a conditional Markov structure. We proved its performance on simulated datasets and real-world applications of chlorophyll forecast.

The outline of the rest of this dissertation is as follows. We will cover 3 separate projects. Chapter 2 is on Project I, in which we propose the RS model between different classes of dynamical models, and we will apply it on implied volatility processes. Chapter 3 is on Project II, in which we proposed a data cleaning method for TAQ data set, that takes a better trade-off between data quality and data size and enables us to calculate the minute-level realized volatility. We prove that the RS phenomenon is universal for stocks' realized variance. Chapter 4 is on Project III, in which we proposed the RS-EDM and PRS models, and applied it to the chlorophyll forecast. A discussion is provided at the end in Chapter 5.

Chapter 2

Project I

2.1 Summary

In this project, we first propose a Regime-Switching (RS) model that switches between different classes of dynamical systems or models. The motivation is that different dynamical models work better under different time periods, so it is natural to model it by a RS model between different dynamical systems. Then we address the inference problem and hypothesis testing problem on whether the RS phenomenon exists. In this project, we mainly discussed the RS models for implied volatility, and the generalizations to other applications are straightforward.

It is a widely spread practice to model implied volatility by RS models. However, unlike conventional RS models between the same class of models with different parameters, our research focuses on the RS models switching between different classes of models. This kind of model cannot be inferred by Maximum Likelihood Estimation (MLE) obtained by Expectation-Maximization (E-M) algorithm (Dempster, Laird, and Rubin, 1977), which is the standard inference approach for conventional RS models. So we proposed a novel inference method, heteroskedasticity-based E-M algorithm, that is an extension of E-M algorithm but relies on the Bayes factor of heteroskedasticity test as the emission probabilities. Along with model estimation, we propose a novel hypothesis testing that tests the rank correlation between the

prediction and emission probabilities to prove the predictability of the future regimes by our RS models. We name this testing procedure as the RS test and view it as a testing on the existence of RS phenomenon.

To prove the effectiveness of our method, we apply our method on both the simulated and real-world VIX data. We proved by simulations that our RS test has the power to detect the RS phenomenon for both in-sample and out-of-sample data, and also proved the RS phenomenon exists in VIX data by RS test.

The outline of the rest of this chapter is as follows. In Section 2.2 we provided a comprehensive but not exhaustive review on implied volatility modeling. In Section 2.3 we described how to infer the RS models between different classes of models and how to perform the RS test. Then we demonstrated its usage and effectiveness in Section 2.4 and provided the discussion in Section 2.5.

In this project and next project, we will use the episode to denote the time a state or basic model lasts which contains a sub-process with multiple timestamps with observations.

2.2 Background and Literature Reviews

Implied Volatility is in general calculated from option markets. Options are forms of financial derivative contracts giving the option holders the right to buy or sell a specific quantity of underlying assets at a specific price on an exercise date without assigning the obligation to the holders. The exercise date could be a fixed maturity date, which we call an European option, or before some maturity date, which we call an American option. The option assigning the holders the right to buy/sell the

underlying asset is called call/put option. A European call or put option is called a vanilla option, which is the most liquid option in the market. Besides vanilla or American options, there could be complicated contracts for the maturity structure, such as Bermudan options, or for the payout structure, such as clique options. The underlying asset could be stock, commodity, foreign exchange, or other financial instruments. Option pricing is to find the theoretical value of an option, which relies on mathematical assumptions of the price process for underlying assets, which are called the option pricing model (see Shreve et al., 2004, for more on option pricing).

A typical option pricing model for the underlying asset's price process can be written in the Stochastic Differential Equation (SDE) form $dS_t = \mu_t S_t dt + \sqrt{V_t} S_t dW_t^{(S)}$, where μ_t and V_t are the mean return and volatility parts. Different option pricing models typically differ in the volatility part V_t. The most commonly used model is Black-Scholes (B-S) model (Black and Scholes, 1973) which assumes $dS_t = \mu S_t dt + \sigma S_t dW_t$, a constant mean and volatility term, and under the B-S model, the pricing for vanilla options has closed forms. In reality, we cannot observe the volatility but can observe the option prices in the option market, so we could infer the volatility based on the vanilla option price by finding the volatility reversely from the B-S model, which is called implied volatility (see Orlando and Taglialatela, 2017, for a review).

The aggregated implied volatility of the stock market is called VIX, which is the popular name for the Chicago Board Options Exchange's Volatility Index[1]. It can be viewed as the expected implied volatility of the S&P-500 index, which is calculated based on the implied volatility from the S&P-500 index options with different maturities and strike prices. It is first introduced by Whaley (1993), is important because investors usually interpret as the fear index (Whaley, 2000), and has its own options

[1]Refer to https://www.cboe.com/tradable_products/vix/faqs for detailed calculations

(Goard and Mazur, 2013a) (see Whaley, 2009, for a detailed discussion). VIX is an important index. In Section 2.4, we will use the VIX as our application.

Options of the same underlying asset but different strike prices and maturities in general have different implied volatilities, whose relationship is called volatility surface. It contradicts the B-S model's constant volatility assumption, and there are different ways to solve this contradiction. One approach is to model the volatility surface explicitly, such as stochastic volatility inspired (Gatheral, 2004). Another approach is to use models other than B-S that allows randomness in the volatility process V_t and further model it by some SDEs, which can be further divided into two main categories, the stochastic volatility models that assumes V_t follows some SDE not explicitly dependent on the price process S_t, and the local volatility models that assumes V_t explicitly depends on S_t (Dupire et al., 1994). In this research, we mainly study stochastic volatility models.

Stochastic volatility models in general assume $V_t = \mu(V_t; \Theta)dt + \sigma(V_t; \Theta)dW_t$, where $\mu(\cdot; \Theta)$ and $\sigma(\cdot; \Theta)$ are the drift and diffusion function with model parameters Θ, which can be inferred by two categories of inference methods. One is model calibration, which is based on the option prices observed at one time stamp, and the other is model estimation, which is based on the underlying asset's historic price or volatility processes. The conceptual difference of these two categories lies in that the option market is forward-looking and the historic processes are backward-looking, but in general these two should agree with each other (Bates, 1996). Technically the model calibration is based on finding the theoretical values of options as a function of parameters, maturities and strike prices, and fitting them by the observed option prices with different maturities and strike prices through MLE or ordinary least squares to find the best parameters. The model estimation is based on the underlying asset's

downsampled historic price or volatility processes and discretized SDEs of stochastic volatility models. Since a stochastic volatility model involves a price process and a volatility process, there are different choices of which process to use. For example, we could use only the price process (Atiya and Wall, 2009), only the observed volatility process such as VIX (Goard and Mazur, 2013b), both the price and option price processes (Aït-Sahalia and Kimmel, 2007), or both the price and volatility processes (Fouque and Saporito, 2018). In this study, we will focus on the model estimation only based on the observed volatility process.

Stochastic volatility models are typically based on some basic models. Some widely used basic stochastic volatility models are as follows:

- Heston (S. L. Heston, 1993): $dV_t = \kappa(\eta - V_t)dt + \xi\sqrt{V_t}dW_t$;
- Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) (S. L. Heston and Nandi, 2000): $dV_t = \kappa(\eta - V_t)dt + \xi V_t dW_t$;
- 3/2 (S. L. Heston, 1997; Platen, 1998): $dV_t = \kappa V_t(\eta - V_t)dt + \xi V_t^{\frac{3}{2}}dW_t$.

The estimation of these models are based on Euler discretization in the following form

- Heston: $V_{t+1} \sim V_t + \kappa(\eta - V_t)dt + \xi\sqrt{V_t} \cdot \mathcal{N}(0, \Delta t)$;
- GARCH: $V_{t+1} \sim V_t + \kappa(\eta - V_t)dt + \xi V_t \cdot \mathcal{N}(0, \Delta t)$;
- 3/2: $V_{t+1} \sim V_t + \kappa V_t(\eta - V_t)dt + \xi V_t^{\frac{3}{2}} \cdot \mathcal{N}(0, \Delta t)$;

where the volatility process is discretized into $\{V_t\}_{t=1}^T$ with interval Δt. Here the GARCH model is also called the Heston-Nandi model in some literature because it is

different from the discrete time GARCH model (Bollerslev, 1986a). After discretizing, we could estimate a basic single stochastic volatility models by maximizing the likelihood function

$$\mathcal{L}(\Theta) = \prod_{t=1}^{T-1} \left[f_{\mathcal{N}(0,\sigma^2(V_t;\Theta)\Delta t)} \left((V_{t+1} - V_t) - \mu(V_t;\Theta)\Delta t \right) \right],$$

or with weight w_t for each time stamp,

$$\mathcal{L}(\Theta) = \prod_{t=1}^{T-1} \left[f_{\mathcal{N}(0,\sigma^2(V_t;\Theta)\Delta t)} \left((V_{t+1} - V_t) - \mu(V_t;\Theta)\Delta t \right) \right]^{w_{t+1}}. \tag{2.1}$$

For stochastic volatility models mentioned above, this MLE could be written as a linear regression (see Goard and Mazur, 2013b, for details), and we use this method for estimating basic models without RS.

We could extend the aforementioned basic models by assigning them additional structures. For example, we could add jumps, such as the B-S model with jump (Merton, 1976) or asymmetric jump (Kou, 2002). We could add the multi-factor property to make it satisfy the option market's term-structure (Christoffersen, S. Heston, and Jacobs, 2009). We could add two models together, such as the 4/2 model, which is the summation of the Heston model and the 3/2 model (Grasselli, 2017). Also we could add the RS structure.

In this study, we will mainly focus on the RS property. Researchers have already studied this phenomenon extensively and proposed different models with RS structures. For example, the RS B-S model inferred by Baum-Welch algorithm (Mitra and Date, 2010), the RS B-S models with a large number of hidden states (Rossi and Gallo, 2006), the RS B-S model with jumps (Costabile et al., 2014), RS local volatility model (He and Zhu, 2017), RS Heston models with application on VIX

(Goutte, Ismail, and Pham, 2017), RS Heston model with joint modeling of S&P-500 and VIX (Papanicolaou and Sircar, 2014), RS rough Heston model (Alfeus, Overbeck, and Schlögl, 2019), etc. RS models can be inferred by model estimation or model calibration. Note that all existing RS models are switching between different models within the same model class, for example, two Heston models with different parameters. In this study, we proposed the RS models between different classes of models. To the best of our knowledge, this is the first time this problem is studied.

2.3 Methodology

In this section, we first defined the mathematical model for RS. Then we showed why the standard E-M algorithm for MLE fails for our models, and how to modify it to our proposed heteroskedasticity-based E-M algorithm. Then we proposed RS testing that could test whether the RS model is predictive or not. As an ending we provided the formal framework for heteroskedasticity-based E-M algorithm.

2.3.1 Model assumption

We define a model that switches between two different regimes, which is controlled by a process following a 2-state Markov chain. RS models with more than 2 states can be easily generalized from the 2-state RS model.

We defined the basic models first. Assume we have two different stochastic dynamical models $M_1 = \mathcal{M}_1(\Theta_1)$ and $M_2 = \mathcal{M}_2(\Theta_2)$, where $\mathcal{M}_1(\cdot)$ and $\mathcal{M}_2(\cdot)$ are different

classes of models and Θ_1 and Θ_2 are associated model parameters that

$$\begin{aligned}(M_1) \quad & dV_u = \mu_1(V_u;\Theta_1)dt + \sigma_1(V_u;\Theta_1)dW_u;\\ (M_2) \quad & dV_u = \mu_2(V_u;\Theta_2)dt + \sigma_2(V_u;\Theta_2)dW_u.\end{aligned}$$

For example, we can choose $\mathcal{M}_1(\cdot)$ to be Heston model with $\Theta_1 = (\kappa_1, \eta_1, \xi_1)$ and $\mathcal{M}_2(\cdot)$ to be 3/2 model with $\Theta_2 = (\kappa_2, \eta_2, \xi_2)$, which can be written into the following form

$$\begin{aligned}(M_1) \quad & dV_u = \kappa_1(\eta_1 - V_u)dt + \xi_1\sqrt{V_u}dW_u;\\ (M_2) \quad & dV_u = \kappa_2 V_u(\eta_2 - V_u)dt + \xi_2 V_u^{\frac{3}{2}}dW_u.\end{aligned}$$

Particularly, we call this model the RS-Heston-3/2 model, and will use it extensively in the simulations and applications.

Then we could define the RS by a state-space model. We divided the whole time period into T episodes that the t-th episode is controlled by the latent state h_t. In the following, we will use t for denoting episodes and (t, m) for denoting time stamps. We assume that $\{h_t\}_{t=1}^T$ follows a 2-state Markov chain with parameters $(\pi_0, \boldsymbol{T})$, where $\pi_0 = [\pi_0(1), \pi_0(2)]^\top$ is the initial probabilities that $\mathrm{P}(h_1 = i) = \pi_0(i)$ and $\boldsymbol{T}$ is the transition matrix whose elements $\boldsymbol{T}_{ij} = \mathrm{P}(h_{t+1} = j | h_t = i)$ is the probability of transiting from state i to state j. $\{h_t\}_{t=1}^T$ are not observable, but each of them controls an episode during which the observed $\{V_u\}_{u\in t\text{-th episode}}$ follows M_s if $h_t = s$. We discretize the time into $\{V_{t,l}\}_{l=1,\cdots,L}^{t=1,\cdots,T}$ that during each episode t, we could observe $V_t = V_{t,\cdot} = \{V_{t,l}\}_{l=1}^L$. The basic models for the Heston and 3/2 models under this

discretization could be written into

$$
\begin{aligned}
(M_1) \quad & V_{t,l+1} \sim V_{t,l} + \kappa(\eta - V_{t,l})\Delta t + \xi\sqrt{V_{t,l}} \cdot \mathcal{N}(0, \Delta t); \\
(M_2) \quad & V_{t,l+1} \sim V_{t,l} + \kappa_2 V_{t,l}(\eta_2 - V_{t,l})\Delta t + \xi_2 V_{t,l}^{\frac{3}{2}} \cdot \mathcal{N}(0, \Delta t),
\end{aligned} \tag{2.2}
$$

where Δt is the time interval between $V_{t,l}$ and $V_{t,l+1}$. Other basic models could be written similarly. In summary, the RS-Heston-3/2 model can be written as

$$
\begin{aligned}
\{h_t\}_{t=1}^{T} &\sim \text{Markov Chain}(\boldsymbol{T}, \pi_0); \\
V_{t,\cdot}\big|h_t = s &\sim M_s \quad t = 1, \cdots, T,
\end{aligned}
$$

where M_1 and M_2 are defined in Eq 2.2.

2.3.2 MLE, E-M algorithm, and why it fails

Algorithm 1: E-M algorithm

Data: $\{X_t\}_{t=1}^{T}$
Result: $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \widehat{\mathcal{E}}\}$
Initialize $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \widehat{\mathcal{E}}\}$;
while *not converged* **do**
  E-step: update $\{\gamma_t(j)\}_{t=1:T}^{j=1:S}$ and $\{\xi_t(i,j)\}_{t=1:T}^{i,j=1:S}$ conditioned on $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \widehat{\mathcal{E}}\}$ by Alg 2;
  M-step: update $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \widehat{\mathcal{E}}\}$ conditioned on $\{\gamma_t(j)\}$ and $\{\xi_t(i,j)\}$ by Alg 3;
end

The standard inference method for RS models, such as HMM, is MLE obtained by the E-M algorithm. However, the likelihood-based E-M algorithm doesn't work here and we proposed a variant, which we call heteroskedasticity-based E-M algorithm. I will first briefly introduce the E-M algorithm for standard RS models and discuss

Algorithm 2: E-step of E-M algorithm

Data: $\{X_t\}_{t=1}^T$, $\{\pi_0, \boldsymbol{T}, \mathcal{E}\}$
Result: $\{\gamma_t(j)\}_{t=1:T}^{j=1:S}$, $\{\xi_t(i,j)\}_{t=1:T}^{i,j=1:S}$
Find $\{b_j(X_t)\}_{t=1:T}^{j=1:S}$;
`// Forward probabilities`
for $t \leftarrow 1$ **to** T **do**

$$\alpha_t(j) \leftarrow \begin{cases} \pi_0(j) b_j(X_1), & \text{if } t = 1 \\ \sum_{i=1}^S \alpha_{t-1}(i) \boldsymbol{T}_{ij} b_j(X_t), & \text{otherwise} \end{cases} \quad \forall j = 1, \cdots, S;$$

end
`// Backward probabilities`
for $t \leftarrow T$ **to** 1 **do**

$$\beta_t(j) \leftarrow \begin{cases} 1, & \text{if } t = T \\ \sum_{j=1}^S \boldsymbol{T}_{ij} \beta_{t+1}(j) b_j(X_{t+1}), & \text{otherwise} \end{cases} \quad \forall j = 1, \cdots, S;$$

end
`// State occupancy probabilities`
$\mathrm{P}(h_t = j) = \gamma_t(j) \leftarrow \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^S \alpha_t(i)\beta_t(i)}$ $\forall t, j$;
$\mathrm{P}(h_t = i, h_{t+1} = j) = \xi_t(i,j) \leftarrow \frac{\alpha_t(i)\boldsymbol{T}_{ij} b_j(X_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^S \alpha_t(i)\beta_t(i)}$ $\forall t, i, j$;

Algorithm 3: M-step of E-M algorithm

Data: $\{X_t\}_{t=1}^T$, $\{\gamma_t(j)\}_{t=1:T}^{j=1:S}$, $\{\xi_t(i,j)\}_{t=1:T}^{i,j=1:S}$
Result: $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \widehat{\mathcal{E}}\}$
$\widehat{\pi}_0(j) \leftarrow \gamma_1(j)$ $\forall j = 1, \cdots, S$;
$\widehat{\boldsymbol{T}}_{ij} \leftarrow \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^S \xi_t(i,k)}$ $\forall i, j = 1, \cdots, S$;
Estimate $\widehat{\mathcal{E}}_j$ based on weighted samples $\{X_t\}_{t=1}^T$ with weights $\{\gamma_t(j)\}_{t=1}^T$

why the MLE doesn't work.

A RS model contains two processes, a latent state process $\{h_t\}_t$ following a S-state time-homogeneous Markov chain and an observed process $\{X_t\}_t$. The Markov chain could be specified by 2 parameters,

- the initial probability $\pi_0 = [\pi_0(1), \cdots, \pi_0(S)]^\top$;
- the transition matrix $\boldsymbol{T} = [\boldsymbol{T}_{ij}]_{i=1:S}^{j=1:S}$,

and the relationship between the observations and the latent states is characterized by

- emission distribution $\mathcal{E} = \{\mathcal{E}_s\}_{s=1:S}$

where for each time t, the observed $X_t|h_t = s \sim \mathcal{E}_s$. If observing $X_t = x$ at time t, we could find the emission probability for each hidden state s, $b_s(x) = \mathrm{P}(X_t = x|h_t = s)$. In our study, the observed time series during an episode is a sub-process that $X_t = V_t = \{V_{t,m}\}_m$ controlled by h_t.

The standard inference framework is the MLE by the E-M algorithm, as shown in Alg 1. In short, it is an iterative method iterating over E-steps and M-steps until convergence, where the E-step is to update the belief of latent states given the current parameters, and the M-step is to update parameters conditioned on the belief of latent states. We could see that the E-M algorithm for different RS models only differs in the emission probability part while all other steps are identical.

Conceptually, the E-M algorithm relies on the distinguishable assumption, that under different hidden states, the emission probabilities for the same observation should be distinguishable. For example, if a data point X_t is generated from $\mathcal{E}_s$, then $b_s(X_t)$, the emission probability if $h_t = s$, should be much larger than $b_r(X_t)$ for $r \neq s$ with a high probability. Otherwise the belief of latent states can hardly be inferred correctly.

However, the E-M algorithm for MLE doesn't work on our problem, because the likelihood cannot distinguish different models. As a motivation example, we show two illustrative experiments. One experiment is to distinguish Heston and GARCH models, and the other is to distinguish Heston and 3/2 models. In the first example, we generate multiple data sets from both the Heston and GARCH models, and for each data set, its maximal likelihood $\mathcal{L}_{(H)}$ assuming it follows a Heston model by

Eq 2.1, $\mathcal{L}_{(G)}$ assuming it follows a GARCH model, and its likelihood ratio between Heston and GARCH $\frac{\mathcal{L}_{(H)}}{\mathcal{L}_{(G)}}$ are found. If the true model is Heston, then the likelihood ratio should be large, otherwise if the true model is GARCH, then the likelihood ratio should be small, so we expect the likelihood ratio when the true model is the Heston model is different from when the true model is the 3/2 model. The simulation settings are as follows. For both Heston and GARCH, we set $\kappa_H = \kappa_G = 5$, $\eta_H = \eta_G = 0.16$, initial values as 0.25, length as 1000 timestamps and time interval as $\Delta t = 10^{-4}$. We set the diffusion terms $\xi_H\sqrt{\eta_H} = \xi_G\eta_G = 0.25$. 1000 data sets following the Heston model and 1000 data sets following the GARCH model are generated, so we can estimate the distribution of likelihood ratios $\frac{\mathcal{L}_{(H)}}{\mathcal{L}_{(G)}}$ if the true model is the Heston model and the GARCH model. The second experiment is similar, in that we want to distinguish Heston and 3/2 models. All settings are the same except that the diffusion terms are set to be $\xi_H\sqrt{\eta_H} = \xi_{3/2}\eta_{3/2}^{\frac{3}{2}} = 0.25$.

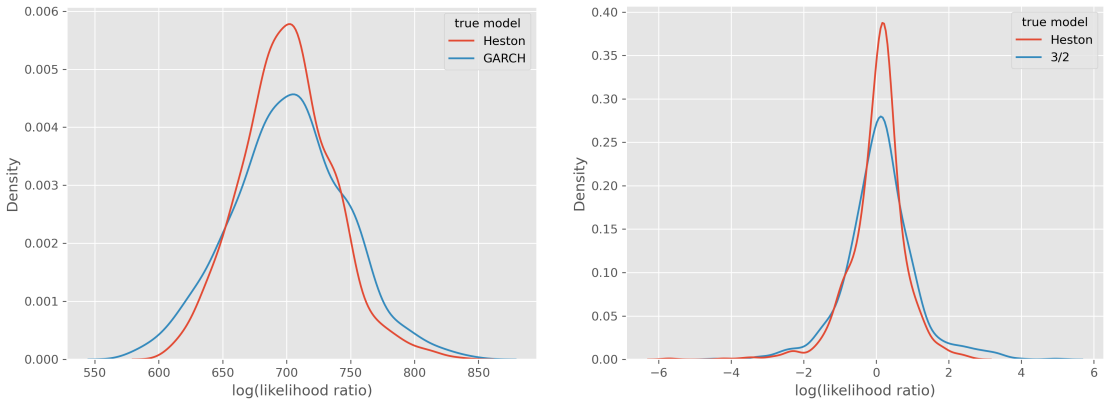

Figure 2.1: The density plot of log(likelihood ratio) between Heston and GARCH (left panel) and between Heston and 3/2 models (right panel), when the true model is Heston or GARCH / Heston or 3/2 models. We could see they are not distinguishable in both 2 experiments when Eq 2.3 is satisfied.

Fig 2.1 shows the density plot of likelihood ratios from kernel density estimation for both 2 experiments. In its left panel, we could see that whether the data is

generated from Heston and GARCH, the likelihood ratio doesn't differ in location. Similarly, in the right panel, we could see the likelihood ratio doesn't differ between the Heston and 3/2 models. That is to say, the likelihood ratios are similar and not distinguishable even if the true models are different. From more analysis, I found the following indistinguishable condition:

$$\xi_H \times \sqrt{\eta_H} \approx \xi_G \times \eta_G \approx \xi_{3/2} \times \eta_{3/2}^{\frac{3}{2}}, \tag{2.3}$$

whose interpretation is natural. The SDE of stochastic volatility models have the drift and diffusion parts. The drift part is mean-reverting which guarantees V_t to oscillate around η. So the diffusion term has standard deviation also oscillated from $\xi_H \times \sqrt{\eta_H}$ for Heston, $\xi_G \times \eta_G$ for GARCH and $\xi_{3/2} \times \eta_{3/2}^{\frac{3}{2}}$ for 3/2 model. This variance term is determining the likelihood, and if the variance of two models are similar then the likelihood will also be similar. Or from another aspect, if we have a dataset and fit different models, then different models won't differ significantly in likelihood and their estimated parameters will satisfy Eq 2.3.

2.3.3 Heteroskedasticity-based E-M algorithm

MLE doesn't work because the likelihood is not distinguishable, so we need another approach to distinguish different models. Also we need to convert it into emission probabilities to implement them into the E-M algorithm. In practice we use different stochastic volatility models not due to the likelihood, but whether the model is sufficient or not. The idea comes from the Box-Jenkins algorithm (Box et al., 2015), where we find the smallest orders for AutoRegressive-Moving-Average (ARMA) models that the residuals don't have any significant autocorrelation or partial autocorrelation.

Algorithm 4: Emission probability from heteroskedasticity regression.

Data: $\{V_t = \{V_{t,m}\}_{m=1}^{n+1}\}_{t=1}^{T}$, $\{\gamma_t(M_1), \gamma_t(M_2)\}_{t=1}^{T}$, M_1, M_2, g, $\pi(M_1)$, $\pi(M_2)$, $\pi(O)$

Result: $\{b_1(V_t), b_2(V_t)\}_{t=1}^{T}$

for $i \leftarrow 1$ **to** 2 **do**

  Estimate the model M_i by weighted data $\{V_{t,m}\}_{t=1:T}^{m=1:n+1}$ with weights $\gamma_t(M_i)$ for $V_{t,m}$ by MLE of Eq 2.1;

  for $t \leftarrow 1$ **to** T **do**

    Get residuals $\{\widehat{W}_{t,m+1}\}_{m=1}^{n}$;

    Regress $\widehat{W}_{t,m+1}^2$ against $V_{t,m}, V_{t,m}^{\frac{1}{2}}, V_{t,m}^{-\frac{1}{2}}, V_{t,m}^{-1}$ and get R_{het}^2;

    Find $BF_i^{(t)}$ from Eq 2.4;

  end

end

for $t \leftarrow 1$ **to** T **do**

  Solve $\mathrm{P}(M_1|V_t)$, $\mathrm{P}(M_2|V_t)$ and $\mathrm{P}(O|V_t)$ from

$$\begin{cases} \frac{\mathrm{P}(M_1|V_t)}{\mathrm{P}(M_2|V_t)+\mathrm{P}(O|V_t)} = \frac{1}{BF_1^{(t)}} \times \frac{\pi(M_1)}{\pi(M_2)+\pi(O)} \\ \frac{\mathrm{P}(M_2|V_t)}{\mathrm{P}(M_1|V_t)+\mathrm{P}(O|V_t)} = \frac{1}{BF_2^{(t)}} \times \frac{\pi(M_2)}{\pi(M_1)+\pi(O)} \\ \mathrm{P}(M_1|V_t) + \mathrm{P}(M_2|V_t) + \mathrm{P}(O|V_t) = 1 \end{cases};$$

  $b_1(V_t) \leftarrow \frac{\mathrm{P}(M_1|V_t)}{\mathrm{P}(M_1|V_t)+\mathrm{P}(M_2|V_t)}$;

  $b_2(V_t) \leftarrow \frac{\mathrm{P}(M_2|V_t)}{\mathrm{P}(M_1|V_t)+\mathrm{P}(M_2|V_t)}$;

end

Similarly, for AutoRegressive Conditional Heteroskedasticity / Generalized AutoRegressive Conditional Heteroskedasticity (ARCH/GARCH) (Engle, 1982; Bollerslev, 1986b), we find its smallest orders that the squared residuals don't have any significant autocorrelation or partial autocorrelation. We propose to use a similar idea to distinguish different models, that the sufficient model should not observe any heteroskedasticity and there should not be any significant correlations between squared residuals and the observed process.

To quantify heteroskedasticity, we could do a hypothesis test to test whether the squared residuals are correlated with the process or not. Ideally, when we are using the correct models, the residuals should pass this test because the residuals' variance

cannot be explained by the process. Otherwise we should reject this test because still some information of the white noise could be modeled by the process. The testing procedure is as follows. We first estimate the residuals $\{\widehat{W}_{t,m+1}\}_t^m$. Then we regress the squared residuals $\{\widehat{W}_{t,m+1}^2 = (V_{t,m+1} - \widehat{V_{t,m+1}})^2\}_m$ against the previous observations with different polynomials $\{V_{t,m}, V_{t,m}^{\frac{1}{2}}, V_{t,m}^{-\frac{1}{2}}, V_{t,m}^{-1}\}_m$. If this linear regression is significant, then the residuals could be explained by the process and there exists heteroskedasticity. Otherwise there doesn't exist heteroskedasticity and the model is sufficient. We name this regression as heteroskedasticity regression and name the hypothesis testing procedure whether heteroskedasticity regression is significant or not as heteroskedasticity test.

The E-M algorithms for different RS models with Markov switching differ only in the emission parts. To implement our heteroskedasticity test into the E-M algorithm, we need to convert it into emission probabilities. We borrow the idea of Bayes factor. Here we use the Bayes factor for practical usage and we are not to classify our method into Bayesian or frequentist. The heteroskedasticity test is based on linear regression, so we could use the Bayes factor of linear regression. There are different choices for linear regression's Bayes factors, such as the g-prior Bayes factor (Zellner, 1986) and the minimal Bayes factor (Chen, Ye, and M. Wang, 2021), and we use the widely used g-prior Bayes factor. To calculate the Bayes factor, we first find R_{het}^2, the R-squared for the heteroskedasticity regression, and the Bayes factor is a function of R_{het}^2 that

$$BF = \frac{(1+g)^{(n-p-1)/2}}{(1+g(1-R_{het}^2))^{(n-1)/2}} = \frac{\mathrm{P}(data|homoskedasticity)}{\mathrm{P}(data|heteroskedasticity)} = \frac{\mathrm{P}(data|M^{\complement})}{\mathrm{P}(data|M)}, \quad (2.4)$$

where g is a hyperparameter we need to choose to better distinguish different models, n is the sample size for regression, $p = 4$ is the number of features, and M and $M^{\complement}$ denote the condition where M is and isn't the correct model. Also this method could

be generalized to weighted samples, so that the regression and R^2_{het} are calculated based on the weighted samples.

Algorithm 5: Heteroskedasticity-based E-M algorithm

Data: $\{V_t = \{V_{t,m}\}_{m=1}^{n+1}\}_{t=1}^{T}$
Result: $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \{\widehat{\Theta_i}\}_i\}$
Initialize $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \{\widehat{\Theta_i}\}_i\}$;
while *not converged* **do**
  E-step: update $\{\gamma_t(j)\}_{t=1:T}^{j=1:S}$ and $\{\xi_t(i,j)\}_{t=1:T}^{i,j=1:S}$ conditioned on $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \{\widehat{\Theta_i}\}_i\}$ by Alg 2 with emission probabilities computed from Alg 4;
  M-step: update $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \{\widehat{\Theta_i}\}_i\}$ conditioned on $\{\gamma_t(j)\}$ and $\{\xi_t(i,j)\}$ by Alg 6;
end

Algorithm 6: M-step of heteroskedasticity-based E-M algorithm

Data: $\{V_t\}_{t=1}^{T}$, $\{\gamma_t(j)\}_{t=1:T}^{j=1:2}$, $\{\xi_t(i,j)\}_{t=1:T}^{i,j=1:2}$
Result: $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \{\widehat{\Theta_i}\}_i\}$
$\widehat{\pi}_0(j) \leftarrow \gamma_1(j)\ \forall j = 1, 2$;
$\widehat{\boldsymbol{T}}_{ij} \leftarrow \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1}\sum_{k=1}^{S} \xi_t(i,k)}\ \forall i, j = 1, 2$;
for $i \leftarrow 1$ **to** 2 **do**
  Estimate $\widehat{\Theta_i}$ by Eq 2.1 with data $\{V_{t,m}\}_t^m$ and weight $\gamma_t(i)$ for each $V_{t,m}$;
end

The above heteroskedasticity regression and Bayes factor are for testing one model. In RS, we have multiple candidate models, so we have multiple heteroskedasticity regressions and Bayes factors, and we compute the emission probabilities from them. Assume the data is switching between 2 models M_1 and M_2, for example M_1 is Heston and M_2 is 3/2. We also set a third model O to accommodate cases where the data is not generated from M_1 or M_2. For model M_1, we have BF_1. Note $M_1^{\complement}$ is M_2 or O.

We have

$$\begin{aligned} & \frac{\mathrm{P}(M_1|data)}{\mathrm{P}(M_2|data)+\mathrm{P}(O|data)} = \frac{\mathrm{P}(M_1|data)}{\mathrm{P}(M_1^{\complement}|data)} = \frac{\mathrm{P}(data|M_1)\times\pi(M_1)}{\mathrm{P}(data|M_1^{\complement})\times\pi(M_1^{\complement})} \\ = & \frac{1}{BF_1}\times\frac{\pi(M_1)}{\pi(M_1^{\complement})} = \frac{1}{BF_1}\times\frac{\pi(M_1)}{\pi(M_2)+\pi(O)}. \end{aligned}$$

If we calculate BF_1 and BF_2 from data, and determine the hyperparameter for priors $\pi(M_1)$, $\pi(M_2)$ and $\pi(O)$, when we can calculate $\frac{\mathrm{P}(M_1|data)}{\mathrm{P}(M_2|data)+\mathrm{P}(O|data)}$ and similarly $\frac{\mathrm{P}(M_2|data)}{\mathrm{P}(M_1|data)+\mathrm{P}(O|data)}$. Also we have $\mathrm{P}(M_1|data)+\mathrm{P}(M_2|data)+\mathrm{P}(O|data)=1$. From these 3 equations, we could calculate $\mathrm{P}(M_1|data)$, $\mathrm{P}(M_2|data)$ and $\mathrm{P}(O|data)$. Then we could normalize out O, and the emission probabilities for model M_1 and M_2 are

$$\begin{aligned} b_1(data) = \mathrm{P}(M_1|data, M_1\cup M_2) &= \frac{\mathrm{P}(M_1|data)}{\mathrm{P}(M_1|data)+\mathrm{P}(M_2|data)}; \\ b_2(data) = \mathrm{P}(M_2|data, M_1\cup M_2) &= \frac{\mathrm{P}(M_2|data)}{\mathrm{P}(M_1|data)+\mathrm{P}(M_2|data)}. \end{aligned}$$

These are the emission probabilities we used, and the algorithm is summarized in Alg 4. Another choice is to use $b_1(data)=\mathrm{P}(M_1|data)$ and $b_2(data)=\mathrm{P}(M_2|data)$ if we do not condition out O, but we recommend conditioning out O because it is in line with our model assumption. We name the E-M algorithm with this emission probability as heteroskedasticity-based E-M algorithm, and its pseudocode is shown in Alg 5.

2.3.4 Regime-Switching testing: prediction and emission

Besides inferring the RS model, an important question is whether the RS phenomenon exists, or how to statistically test it. This task is not trivial since we don't know the

underlying truth. In this project, we convert this task to testing whether the RS model is predictive for future regimes or not. See Section 2.5 for discussion.

We propose a hypothesis testing procedure for this specific task and we name it as RS test. Though we don't know the underlying true class labels, we could use the emission probability as an approximation of the underlying true labels. We could set a threshold on the emission probability to determine whether each state belongs to M_1 or M_2. After inferring the RS model, we can forecast the future probability of each episode belonging to each model, and similarly, we can set a threshold to determine the predicted label for each episode t. Then we could test whether these two match with each other with a high accuracy. We can marginalize out the effect of threshold if we don't want to specify it. If we marginalize out the threshold for predicted probabilities, we can use the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). If we want to marginalize out both the thresholds for emission and prediction probabilities, we can perform Kendall's τ rank correlation test between the emission and predicted probabilities on whether it is greater than 0 or not (Kendall, 1938), which is recommended.

The mathematical formulation is as follows. First we fit the model by heteroskedasticity-based E-M algorithm. For each episode t, we have the observation $V_t = \{V_{t,m}\}_{m=1}^n$, and the emission probability $b_1(V_t) = \mathrm{P}(h_t = M_1 | V_t, M_1 \cup M_2)$, and denote it as e_t. The predicted probability is the probability of observing M_1 conditioned on all information till time $t-1$ that $\mathrm{P}(h_t = M_1 | V_1, \cdots, V_{t-1}; M_1 \cup M_2)$, and denote it as p_t. It can be calculated by the standard forward algorithm. Then we could have a set of pairs of emission and prediction probabilities $\{(e_t, p_t)\}_{t=1:T}$. To test whether the prediction and emission probabilities match with each other, or whether the prediction probabilities have predictive power for emission probabilities, we use Kendall's rank

correlation coefficient τ as defined below:

$$\tau = \frac{C - D}{C + D}, \tag{2.5}$$

where C and D are the number of concordant and discordant pairs between e_t and p_t, defined that for any pair (e_t, p_t) and (e_s, p_s) that $1 \leq t < s \leq T$, if $e_t > p_t, e_s > p_s$ or $e_t < p_t, e_s < p_s$, then this pair is concordant, otherwise it is discordant. In the cases without ties, which is basically true in our cases, we have $C + D = \binom{T}{2}$ pairs that are either concordant or discordant. The hypothesis to test is

$$\begin{aligned} H_0 : &\quad p_t \text{ and } e_t \text{ are independent;} \\ H_A : &\quad p_t \text{ can predict } e_t. \end{aligned}$$

Under the null hypothesis that there is no rank-correlation between e_t and p_t, the τ defined in Eq 2.5 asymptotically follows $\mathcal{N}\left(0, \frac{2(2T+5)}{9T(T-1)}\right)$ where T is the number of pairs or episodes. If p_t can predict e_t, then the rank correlation τ should be positive, in which case we can conclude that the RS phenomenon exists. So this test can be converted to a one-sided z-test with p-value calculated by

$$p - value = 1 - F\left(\tau; \mathcal{N}\left(0, \frac{2(2T+5)}{9T(T-1)}\right)\right)$$

where $F\left(\cdot; \mathcal{N}\left(0, \frac{2(2T+5)}{9T(T-1)}\right)\right)$ is the cumulative distribution function of $\mathcal{N}\left(0, \frac{2(2T+5)}{9T(T-1)}\right)$. This is the RS test we proposed.

The RS test could be performed in-sample or out-of-sample. For in-sample hypothesis testing, the parameters $(\pi_0, \boldsymbol{T}, \mathcal{E})$ are estimated from the same dataset as for testing rank correlation. For out-of-sample hypothesis testing, these parameters are

estimated from one observed process and the hypothesis testing is performed on a different observed process. The 'in-sample' only means the parameters are estimated in the same dataset for hypothesis testing, and other than that we are not using any future information for prediction. In both two settings, the hypothesis testing procedures are the same and valid. Its pseudocode is shown in Alg 7.

Algorithm 7: RS test

Data: $\{V_t^{(train)}\}_{t=1}^{T^{(train)}}$, $\{V_t\}_{t=1}^{T}$
Result: τ, p-value.
// Training series $\{V_t^{(train)}\}_{t=1}^{T^{(train)}}$ and testing series $\{V_t\}_{t=1}^{T}$ could be the same, for in-sample testing, or different, for out-of-sample testing.
Estimate $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}, \{\widehat{\Theta_i}\}_i\}$ by Alg 5 with data $\{V_t^{(train)}\}_{t=1}^{T^{(train)}}$;
for $t \leftarrow 1$ **to** T **do**
    Calculate $e_t = b_1(V_t)$ by Alg 4;
    if $t = 1$ **then**
        for $j \leftarrow 1$ **to** 2 **do**
            $\alpha_t(j) \leftarrow \widehat{\pi}_0(j)$;
        end
    else
        for $j \leftarrow 1$ **to** 2 **do**
            $\alpha_t(j) \leftarrow \sum_{i=1}^{2} \alpha_{t-1}(i)\boldsymbol{T}_{ij}$;
        end
    $p_t \leftarrow \frac{\alpha_t(1)}{\alpha_t(1)+\alpha_t(2)}$;
    for $j \leftarrow 1$ **to** 2 **do**
        $\alpha_t(j) \leftarrow \alpha_t(j) b_j(V_t)$;
    end
end
Calculate τ by Eq 2.5 from $\{(e_t, p_t)\}_t$;
$p - value = 1 - F\left(\tau; \mathcal{N}\left(0, \frac{2(2T+5)}{9T(T-1)}\right)\right)$;

2.3.5 Choosing hyperparameters

In this algorithm, we need to pre-specify hyperparameters including g, $\pi(M_1)$, $\pi(M_2)$ and length of episode each state lasts for.

The choice of length of each episode depends on our prior knowledge. For example, if we are studying the RS phenomenon for stock market volatility, and the volatility is calculated from the market, so each episode could last one trading day as the trading hours are not consecutive overnight. In some other cases where the time process is consecutive, such as the crypto-currency market, we could break it down by hours or days. If we don't have high frequency records, we could use one week, one month or one year as an episode, which can be used for macroeconomics research. There are two requirements for each episode we should try to satisfy. First, each episode should be long enough to have sufficient data points, such as hundreds, for heteroskedasticity regression and test. Second, each episode should be short enough so it could be modeled by a single basic model. In our applications, we use minute-level data and each episode is set to be one trading day and has 390 data points each episode. If we only have hourly or daily data, we cannot let each episode be one day as we cannot have enough data points per episode.

The choice of $\pi(M_1)$, $\pi(M_2)$ and $\pi(O)$ depends on our prior belief. In general we want $\pi(O)$ to be small, such as $\pi(O) = 0.2$. For M_1 and M_2, we could allocate the prior weights according to our prior belief. If we had no prior belief, we could equally allocate the weight, such as $\pi(M_1) = \pi(M_2) = 0.4$.

The choice of g is a little bit tricky. g is for calculating the emission probabilities from heteroskedasticity tests, and we want the emission probabilities to be distinguishable. For the Bayes factor of linear regression, researchers have suggested various methods

to select g such as empirical Bayes (Liang et al., 2008). However, these methods work in the static setting that the dataset remains unchanged. In our E-M iterations, though the observed process is not changing, the weights for each episode are the state occupancy probabilities and remain changing. Thus we suggest tuning g according to the training set that makes the models as distinguishable as possible. We could choose g that makes the hypothesis testing on the training set to be significant, so that the power of this testing procedure could be high. See Section 2.5 for more on discussion on its relationship to signal-noise-ratio.

2.3.6 Formal framework and convergence criteria

The two-step heteroskedasticity test can be written in a unified formula. This framework is not limited in this problem, so we chose to write the data into $\{X_i, Y_i, Z_i\}_{i=1}^N$ form, where Y is the response, X is the features to predict Y, and Z is the part to explain conditional variance of residuals for heteroskedasticity test. Without loss of generality we assume $X \subset Z$. Then the heteroskedasticity test can be written into the following form that

$$\begin{aligned} Y &= X\beta + U \\ (Y - X\beta)^2 &= \gamma_0 + Z\gamma_1 + V \end{aligned}$$

where U is the error term for regressing Y against X and V is the error term for heteroskedasticity regression. In two-step regression, we first estimate β and then

estimate γ_0, γ_1. We could write the two-step regression into one-step that

$$\widehat{\beta} = \arg\min_{\beta} R^2_{het}(\beta; X, Y, Z) = \arg\min_{\beta} \left(\frac{\min_{\gamma_0,\gamma_1} \left[(Y - X\beta)^2 - (\gamma_0 + Z\gamma_1)\right]^2}{\min_{\gamma_0} \left[(Y - X\beta)^2 - \gamma_0\right]^2} \right).$$

Or to estimate all parameters simultaneously in a min-max form that

$$\widehat{\beta}, \widehat{\gamma_0}, \widehat{\gamma_1} = \arg \min_{\beta} \max_{\gamma_0,\gamma_1} \left(\frac{\left[(Y - X\beta)^2 - (\gamma_0 + Z\gamma_1)\right]^2}{\left[(Y - X\beta)^2 - \overline{(Y - X\beta)^2}\right]^2} \right).$$

In this framework, we are to find β that minimize the above term which is similar to the R^2 of the heteroskedasticity regression. The two-step regression is an approximate solution of the above optimization, because if β is different from its ordinary least squares estimation, then the residual term will have leakage of X and increase the R^2 of the heteroskedasticity regression.

Under this framework, we effectively minimize the R^2 for the heteroskedasticity regression. For each episode t and each model M_i, we have an R^2_{t,M_i}. We suggest to construct the convergence criteria by aggregating these $\{R^2_{t,M_i}\}^{i=1,2}_{t=1,\cdots,T}$. One recommendation is to use weighted average, that the weight is the state occupancy probability $\gamma_t(M_i)$ calculated from the E-step in Alg 2, which can be written as

$$\overline{R^2_{het}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{2} \gamma_t(M_i) R^2_{t,M_i}. \tag{2.6}$$

$\overline{R^2_{het}}$ is in general decreasing each iteration, and we could use it as the convergence criteria that when it is decreasing slowly, we stop iterations. In practice, we suggest to stop if it is decreasing less than a percentage, such as 1%. If $\overline{R^2_{het}}^{(r)}$ is the $\overline{R^2_{het}}$

after r E-M iterations, we will stop when

$$\frac{\overline{R^2_{het}}^{(r-1)} - \overline{R^2_{het}}^{(r)}}{\overline{R^2_{het}}^{(r-1)}} < 0.01$$

where 0.01 is the criteria we choose. Also we could choose other stopping criteria. For example, the most straightforward way is to choose a fixed number of iterations.

2.4 Results

In this section, we first prove the performance of our RS models and its testing procedure by simulation. Then we tested whether VIX has the RS phenomenon by RS test. We mainly focus on the RS-Heston-3/2 model.

2.4.1 Simulations

In the simulation, we want to see whether our RS test can identify the RS phenomenon. Also we want to see how our testing procedure would perform when there is no RS phenomenon. We generate data from 6 different dynamical models, one RS model and the other time-invariant models:

- RS-Heston-3/2: RS models between Heston and 3/2 with the initial probability $[0.5, 0.5]$ and the transition matrix $[[0.9, 0.1], [0.1, 0.9]]^\top$;
- Heston: Heston model without RS;
- 3/2: 3/2 model without RS;
- GARCH: GARCH model without RS;

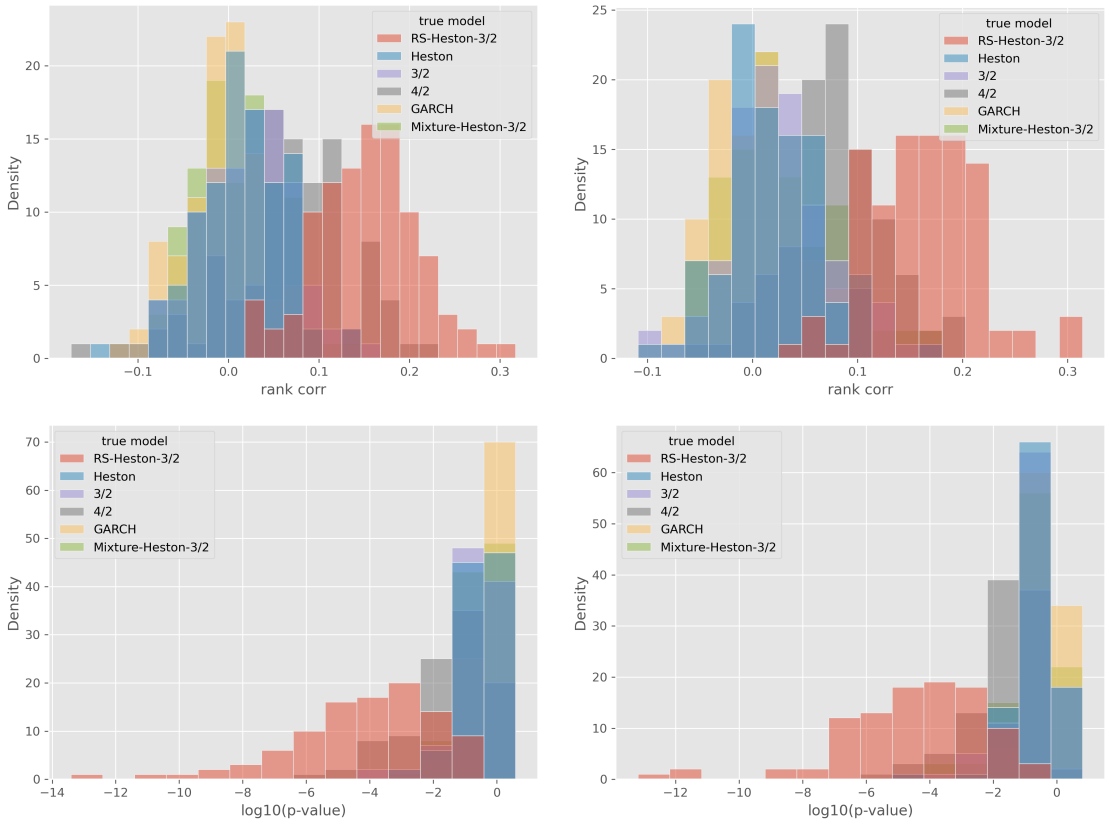

Figure 2.2: The histograms of estimated Kendall's rank correlation τ (top row) and of the p-values (bottom row, in the $\log_{10}$ scales) of repeated simulations. The left column are the in-sample results and the right column are the out-of-sample results. There are 6 kinds of true models: RS, Heston, 3/2, GARCH, 4/2 and mixture models. We could see the RS and time-invariant models' rank correlations are separated, and the hypothesis testing has a large power.

- 4/2: 4/2 model without RS, which is of the form $aV_t + \frac{b}{V_t}$ where V_t follows a Heston model;

- Mixture: a mixture model of Heston and 3/2 models, that at each episode it is equally likely to be a Heston model or a 3/2 model, and the modes of different episodes are independent.

For each model, the model parameters are $\kappa_H = \kappa_{3/2} = \kappa_G = 5$, $\eta_H = \eta_{3/2} = \eta_G = 0.25$, $\xi_H \times \sqrt{\eta_H} = \xi_{3/2} \times \eta_{3/2}^{\frac{3}{2}} = \xi_G \times \sqrt{\eta_G} = 0.25$ and initial value 0.16. The 4/2

model is $V_t = V_t^{(H)} + \frac{\eta_H}{V_t^{(H)}}$ where $V_t^{(H)}$ is generated from Heston model. For each dynamic, I generated 2 independent time series with the same model parameters, one of which is for training and the other is for out-of-sample testing. Both series contain 250 episodes, close to 252 trading days per year, and each episode lasts 400 time points, close to 390 trading minutes per trading day. We trained the RS-Heston-3/2 model on one series and performed the testing procedure on both two series to get the in-sample and out-of-sample testing results, including the testing statistics τ and p-values.

The results are shown in Fig 2.2. It contains 4 sub-figures that are the histogram of rank correlation τ's / p-values of in-sample / out-of-sample test. The in-sample and out-of-sample results are similar. We could see that if the true model is a basic model without RS, the rank correlations are centered close to 0 and in general smaller than 0.1, and the p-values are typically larger than 0.01. When the true models are RS, the rank correlations are in general larger than 0.1 and centered around 0.15, and the p-values are much smaller than 0.01, that is to say, this hypothesis test has a large power.

When the true model is time-invariant, it is still slightly biased towards positive rank correlations because even the basic model could be approximated by the RS model, though this phenomenon is not strong. This bias is different for different models. If the true model is GARCH, then the bias is nearly 0. If the true model is Heston, 3/2 model or their mixtures, the bias is slight. If the true model is 4/2 model, the bias is slightly larger though it is still much smaller than the true RS models. The reason is that the 4/2 model is the summation of a Heston model and a 3/2 model, so it has both patterns of Heston and 3/2 models, which can be modeled by RS. See Section 2.5 for more discussions.

2.4.2 Applications on VIX

We applied the in-sample RS test on VIX data[2]. The data is from 2021-07-29 to 2022-02-08 as shown in Fig 2.3. We regard each trading day as an episode, and within each episode, we use the minute-level squared VIX data as the observed volatility process. We only used the trading hour (09:30-16:00) data so each episode has 390 data points. We estimated the RS-Heston-3/2 model and we inferred the rank correlation and p-value of the in-sample RS test.

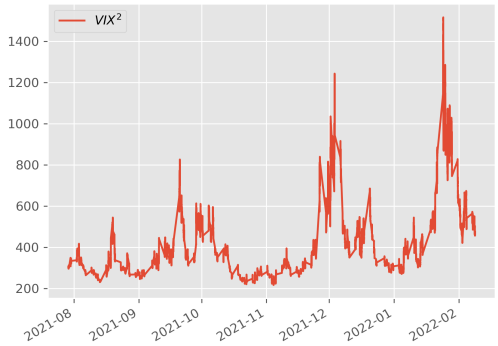

Figure 2.3: VIX data (minute-level; 2021-07-29 to 2022-02-08; squared).

The RS testing results are shown in Table 2.1. These statistics are in-sample. We could see that the rank correlation $\tau = 0.12$ is large, and the p-value is 0.018, significant. So we could conclude that the VIX has the RS phenomenon during 2021-07-29 to 2022-02-08.

Statistics (in-sample)	Values
rank correlation τ	0.12
p-value	0.018

Table 2.1: RS test (in-sample) results of VIX. From the τ and p-value, we could see it is significant and VIX has the RS phenomenon.

[2]Source: Bloomberg Finance L.P.

2.5 Discussion

First of all, we want to discuss what RS is and why the RS phenomenon exists. There are different views on why the RS phenomenon exist, for example some researchers believe it is related to business cycles (e.g. Hamilton, 1989), and some researchers think the regimes switch due to changing fundamentals (see Ang and Timmermann, 2012, for a review). In practice, we would like to view RS as a modeling approach, that RS is a simplified approximation of a complicated model. This view corresponds to the business cycles or changing fundamentals arguments but is more general. For example, we could think there is a complicated model that involves the economic fundamentals as unobserved external parameters, which is changing slowly and impact the observations largely, so we can approximate the process conditioned on different regimes of these latent fundamentals by different basic models, and RS is caused by the changing fundamentals. From this approximation view, we could also illustrate why the RS test is slightly biased. When the true model is a basic model without RS, still we can infer it by a RS model, because even for the basic models, there are some phases having the pattern of other models, or in another word, models are indistinguishable during these phases. That's why the simulation results in Section 2.4.1, the RS testing results are slightly biased. But this bias is moderate and we don't need to worry about false positiveness here. To be conservative, we could use $\tau > 0.1$ or p-value < 0.001 to decide whether it is significant or not.

The above approximation view is inherited from the usage of stochastic volatility models. For example, when the Heston model is proposed, the main advantage lies in its closed-form solution for option pricing. Later other researchers found the square-root diffusion term cannot model all processes sufficiently, so other stochastic volatility candidates are proposed, such as the 3/2 model. But whatever model is proposed, it is

not aiming at fully modeling how data is generated but to mostly capture important patterns by simple models. Our RS model is in line with this view.

Besides, we propose a trick for estimating the RS model, called the third-state trick. This trick is designed for the RS model between different dynamical systems and is a patch for E-M algorithm on separated model spaces. Conventional RS models, such as HMM, have a common model space for different states. However, our RS models have different model spaces for different states, and topologically they are separated and unconnected. So if the true model falls in between these two classes, the emission probability will be indistinguishable and there will be difficulties in inference. This kind of difficulty is only for our RS models, due to the aforementioned connectedness issue. To accommodate it, we proposed the third-state trick. This trick creates the third state outside of M_1 and M_2 for accommodating the indistinguishable cases. We need to make the following modifications:

- Add the third state M_3;
- After getting emission probabilities $b_1(V_t)$ and $b_2(V_t)$ for each episode t, we reassign the emission probabilities to 3 states $[\tilde{b}_1(V_t), \tilde{b}_2(V_t), \tilde{b}_3(V_t)]$ that $\tilde{b}_3(V_t) = \min(b_1(V_t), b_2(V_t))$, $\tilde{b}_1(V_t) = b_1(V_t) - \tilde{b}_3(V_t)/2$ and $\tilde{b}_2(V_t) = b_2(V_t) - \tilde{b}_3(V_t)/2$;
- After fitting the transition matrix $\widehat{\boldsymbol{T}}$, we reassign the transition matrix from the third state that $\widehat{\boldsymbol{T}}_{2\cdot} = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$;
- The predicted probabilities of the third state should be conditioned out.

In Section 2.3, we discussed the choice of g that we need to make the emission probabilities as distinguishable as possible. If we cannot find a suitable g, we can use this third-state trick. Also we could alter the strength of the third state by changing the

emission probability assigned to the third state by choosing an $\alpha \in [0, 1]$ and setting $\tilde{b}_3(V_t) = \alpha \min(b_1(V_t), b_2(V_t))$. If $\alpha = 1$, then it is the third-state trick described above. If $\alpha = 0$, it is equivalent to not using the third-state trick.

The heteroskedasticity test works better under a suitable signal-noise-ratio. A set of distinguishable emission probabilities appears when the heteroskedasticity tests for different models are different, that one test is insignificant and all others are significant. This requires a suitable signal-noise-ratio, because a too large noise will make all tests insignificant, and a too small noise will make all tests significant, both of which are not ideal. That's why we need to tune the hyperparameter g for emission probabilities to make them distinguishable since tuning g is effectively tuning the signal-noise-ratio in the Bayes factor.

For volatility process, we mainly study the RS-Heston-3/2 model, because the Heston and 3/2 models are inherently connected that if V_t follows a Heston model, then $\frac{1}{V_t}$ follows a 3/2 model. Also researchers have already proposed the 4/2 model, which assumes the true volatility is a summation of a Heston and 3/2 process because different models can capture different patterns. If we view the 4/2 model as a mixture of the Heston and 3/2 models, our RS-Heston-3/2 model can be viewed as a temporal mixture of these two basic models. From the simulation results in Section 2.4.1, we can see the 4/2 model has the largest τ and smallest p-value in RS test among all time-invariant models we tried, though not as significant as the true RS model. So we could think the 4/2 model is to capture the RS phenomenon, but not as powerful as the RS-Heston-3/2 model. Note that the true volatility processes in Project I and II show the universe and very significant RS phenomenon, which is more significant than the simulated 4/2 models.

The distinguishability issue of likelihood discussed in Section 2.3.2 is slightly differ-

ent from the commonly discussed identifiability issue of likelihood of mixture models, which refers to the phenomenon that we can switch the clusters' or states' labels without affecting the likelihood. For example, in HMM with Gaussian emissions, the label switching will yield an equivalent HMM model only with different state interpretation. For our RS models, if we switch the states in our RS models, we will have a very different model though the likelihood will remain similar. For example, if we switch the labels in our RS-Heston-3/2 model, then a Heston episode will become a 3/2 episode and vice versa, which will fundamentally change the model assumption and is essentially very different. That's why we name this issue as the distinguishability issue instead of the common identifiability issue of likelihood.

The RS test is testing whether the RS model has predictive power for future regimes, which is a sufficient condition of the existence of the RS phenomenon. In a special case where the RS is not predictable when the latent state process has the transition matrix $\boldsymbol{T} = [[0.5, 0.5]^\top, [0.5, 0.5]^\top]$, then still it is a RS model but the RS test will not give a significant p-value. In contrast, if the RS test is significant, we should regard the RS phenomenon as strong enough. Again we want to emphasize that we are not arguing the true process as time-varying; instead we approximate it by RS between different basic models to capture different dominant patterns during different time periods. So we should interpret the significant RS testing results as follows: the dominant patterns for each episode is predictable from the dominant patterns of its previous episode, where the dominant patterns are captured by various basic models. For example, if the RS testing for the RS-Heston-3/2 model is significant, then its interpretation is whether the Heston or 3/2 model is better to capture next episode's pattern can be determined by whether the Heston or 3/2 model is better for capturing this episode's pattern. This is the interpretation for RS testing.

We can further improve the method in several ways. One way is to better define the E-M algorithm by finding a simple specific target loss function for it. In this project, we view the E-M algorithm as a computational and heuristic method, and we don't have a meaningful and tractable loss function. Another future direction is to study our method's prediction power for the volatility level. Currently our prediction target is the state of the next episode and our hypothesis testing is based on it. We can further study how to add the predictive power for the volatility process by our RS structure. Besides, we can study the model calibration approach of our RS models. In this project, we used the model estimation approach that is based on historic observed volatility processes. The model calibration approach is based on the option prices observed at a single time stamp, where we typically first find the closed-form solution of option prices of our RS models and fit it by the option prices observed in the market at one time stamp. By this approach we can directly prove the performance of our RS models from the option market.

Chapter 3

Project II

3.1 Summary

In Project I, we proposed the RS modeling framework, including the model assumption, heteroskedasticity-based E-M algorithm and RS test. In this project, we continue on this framework and apply it on a large scale dataset. We apply on the minute-level Realized Variance (RV) for 31 assets, including SPDR S&P 500 ETF Trust (SPY) and 30 Dow Jones Industrial average (DJI) component stocks, from 2016 to 2022, and demonstrate that the RS phenomenon is universe across different assets and time periods for RV processes.

To calculate minute-level RV, we use the transaction-level high frequency stock trading data from Trade And Quote (TAQ) dataset. TAQ dataset is a rich dataset containing a lot of noisy transactions that need to be cleaned first. There is a well-established data cleaning method but we find it is not suitable for minute-level RV calculation, since it removes too many transaction records that a lot of minutes don't have any valid record so we cannot calculate RV for these minutes. To avoid this, we propose a novel data cleaning method which keeps twice as many records as the well-established rule-based method and is still highly correlated with its output in the daily RV.

To the best of our knowledge, there is no large-scale research on minute-level RV. Few researches are on the calculation or modeling of minute-level RV, and in general they only focus on specific patterns (see e.g. Mu and Zhou, 2008). In this project, we first proposed the TAQ data cleaning method for RV calculation, and proved that RS is a universal phenomenon for RV across different assets and time periods.

3.2 Background and Literature Reviews

Similar to the implied volatility studied in Project I, RV is very important in finance. Except its index[1], one could predict RV as a forecast of future volatility (Andersen and Bollerslev, 1998). Also RV is related to the profit and loss for gamma trading (Carr and Madan, 2005) and variance swap.

RV is defined on a duration of time. It could be calculated from the quadratic variation of a downsampled log-price process. For a duration of time, suppose we could observe a downsampled log-price process $\log S_{t_1}, \log S_{t_2}, \cdots, \log S_{t_N}$ where $t_1, \cdots, t_N$ is a equidistant or unequidistant partition of this duration, then the RV is

$$RV = \sum_i (\log S_{t_{i+1}} - \log S_{t_i})^2. \tag{3.1}$$

Note that $(\log S_{t_{i+1}} - \log S_{t_i})$ is the log-return from t_i to t_{i+1}. Also some researchers use $\sum_i (\frac{S_{t_{i+1}} - S_{t_i}}{S_{t_i}})^2$ to calculate RV, which will give very similar results as $\log S_{t_{i+1}} - \log S_{t_i} \approx \frac{S_{t_{i+1}} - S_{t_i}}{S_{t_i}}$ by Taylor expansion. Suppose the log-price process follows the SDE $d \log S_t = \mu_t dt + \sigma_t dW_t$, then the integrated variance of a time duration is $\int \sigma_t^2 dt$, and RV is its consistent estimation that $plim \sum_i (\log S_{t_{i+1}} - \log S_{t_i})^2 = \int_0^T \sigma_t^2 dt$

[1] Refer to https://www.cboe.com/us/indices/dashboard/gamma for details.

(Zhang, Mykland, and Aït-Sahalia, 2005). In reality, we cannot observe the exact price process, but a noisy price process from the market, which is the true price plus a market microstructure noise (Zhang, Mykland, and Aït-Sahalia, 2005; Bandi and Russell, 2008; Hansen and Lunde, 2006). In this study, we will not focus on the market microstructure noise because we are focusing on the dynamical systems of RV instead of its estimation. But we will talk about the data cleaning and downsampling later. Also we will only study RV during trading hours and will not discuss the overnight RV which is typically modeled separately (Andersen, Bollerslev, and Huang, 2011).

RV is calculated from high-frequency price data, and we use the executed stock trade data in TAQ dataset from Wharton Research Data Services[2] (Wharton Research Data Services, 2023). The TAQ high-frequency data is highly noisy, and there are different ways to clean. The most well-established way is rule-based (Barndorff-Nielsen et al., 2009; Boudt, Cornelissen, and Payseur, 2013) with the following filtering rules:

1. Delete transactions outside of trading hours;
2. Delete transactions with 0 size or price;
3. Delete corrected trades;
4. Delete transactions with sale conditions other than E (automatic execution), F (intermarket sweep order) or I (odd lot trade);
5. Delete transactions not originating from NYSE, NASDAQ or AMEX.

Another common cleaning method is data-based (Brownlees and Gallo, 2006). It filters out transactions based on the price S_t. Brownlees and Gallo (2006) suggested

[2]Refer to TAQ 3 User's Guide https://www.nyse.com/publicdocs/nyse/data/Daily_TAQ_Client_Spec_v3.0.pdf for more details

to only keep transactions satisfying the condition ($|S_t - \overline{S_t}| < 3s_t + 0.02$) where $\overline{S_t}$ and s_i are the 10%-trimmed mean and standard variance of its 40-nearest neighbor (NN) transactions' prices in the same trading day. Brownlees and Gallo (2006) also suggested to tune the number of NN transactions to use and the largest deviation from the trimmed mean.

The modeling approaches of RV could be categorized depending on whether it is based on SDEs or not. The models without SDEs are commonly used for prediction, such as Heterogeneous AutoRegressive (HAR) (Corsi, 2009), High-Frequency-Based VolatilitY (HEAVY) models (Shephard and Sheppard, 2010; Noureldin, Shephard, and Sheppard, 2012) and Rough Fractional Stochastic Volatility (RFSV) models (Gatheral, Jaisson, and Rosenbaum, 2018). This category of methods are widely used when we focus on forecasting. For example, HAR assumes the future RV is a linear combination of the daily, weekly and monthly RV. HAR has a rich class of variants, such as HAR-Q (Bollerslev, Patton, and Quaedvlieg, 2016; Clements and Preve, 2021), HAR-J (Andersen, Bollerslev, and Diebold, 2007), HAR-SJ (Patton and Sheppard, 2015) and RS-HAR (X. Wang, Shrestha, and Sun, 2019). HEAVY is a variant of GARCH that uses high-frequency RV to replace the lagged squared return in GARCH. RFSV is based on the long-term correlation assumption with the estimated Hurst parameter. However, the class without SDE cannot reveal how RV evolves. To model how RV evolves, we need to model it by SDEs, such as the Heston model (Christoffersen, Jacobs, and Mimouni, 2010). We will focus on the SDE-based approaches, especially those based on stochastic volatility models.

In this section, we slightly abuse the names of the stochastic volatility models. We use the same terminology, such as the Heston model, for both the implied volatility and realized volatility processes. Some researchers has named the corresponding stochastic

volatility models for realized volatility process. For example, in Christoffersen, Jacobs, and Mimouni (2010), the Heston-type model for RV is called the affine square root stochastic volatility model and is short as SQR, the GARCH-type model is called the model with linear rather than square root diffusion for variance and is short as VAR, and the 3/2-type model is called the stochastic volatility model whose power in the diffusion term is 3/2 and whose variance drift is nonlinear in the level of the variance and is short as 3/2N. For consistency, we will stick on the Heston, GARCH and 3/2 models for both two kinds of processes.

3.3 Methodology: Data Cleaning on TAQ Data

In this section, we propose a novel data cleaning method for TAQ dataset that takes a better trade-off between the data quality and data size. It is important not to remove too many data points for high-frequency applications because we need to calculate minute-level RV.

The motivation is to combine the rule-based and data-based cleaning. We used rule-based cleaning for preliminary filtration and used data-based cleaning for fine filtration. For rule-based preliminary filtration, we used all rules mentioned in the previous section except the last rule that filters out transactions not originating from NYSE, NASDAQ or AMEX. Instead we only filter out transactions originating from the Financial Industry Regulatory Authority which contains a lot of darkpool transactions. That is to say, we used the following rules:

1. Delete transactions outside of trading hours;

2. Delete transactions with 0 size or price;

3. Delete corrected trades;

4. Delete transactions with sale conditions other than E (automatic execution), F (intermarket sweep order) or I (odd lot trade);

5. Delete transactions originated from Financial Industry Regulatory Authority.

The remaining data are divided into two parts, (1) the transactions from NYSE, NASDAQ or AMEX, and (2) the transactions from other exchanges, and then we perform the data-based cleaning. We use the first part to calculate the trimmed mean and trimmed variance based on which we filter both two parts of data. That is to say, suppose we have a series of price S_t, then for each transaction, we keep it only when its price S_t satisfies

$$(|S_t - \overline{S_t}| < 2s_i),$$

where $\overline{S_t}$ and s_i are the 10%-trimmed mean and 10%-trimmed standard variance of the 20-NN of S_t that are originated from NYSE, NASDAQ or AMEX. The data-based cleaning method has 2 hyperparameters, the number of NN and the largest deviation from the trimmed mean. We chose to use the 20-NN and $2s_i$, denoted as $2std$, instead of 40-NN and $(3s_i+0.02)$, denoted as $3std+0.02$, as suggested by Brownlees and Gallo (2006) to remove the bias in the data, and more results on different data cleaning hyperparameters are shown in Section 3.4. We name this method as Rule aNd Data (RnD)-based cleaning method.

3.4 Results: RS on RV

In this section, we study 31 assets, SPY and the 30 DJI component stocks from 2016 to 2022. We showed the results for two parts, including the performance of our new data cleaning method and the out-of-sample RS test.

3.4.1 Data cleaning results

In this part, we want to verify our new data cleaning methods, RnD-based method, works well, that our method performs similarly to the well-established rule-based method and at the same time it keeps a much larger proportion of transactions than the rule-based method.

First, we show the performance of RnD-based methods with different hyperparameters. There are 2 hyperparameters, the number of NN and the largest deviation from the trimmed-mean. In Section 3.3, we suggest to using 20-NN and the largest deviation to be $2std$. It is tuned to be different from the original data-based method (Brownlees and Gallo, 2006) that uses 40-NN and $3std + 0.02$. To show the reason we conduct the following experiments. For each set of hyperparameters, after RnD-based data cleaning, we calculate the RV for the m-th minute of trading day t by Eq 3.1, $RV_{t,m}$, then we aggregate the RV of the whole day by $RV_t = \sum_m RV_{t,m}$. So for each day t, we have RV_t under different data cleaning methods, including the RnD-based method RV_t^{RnD} and rule-based method RV_t^r. We particularly compare the RnD-based cleaning method under different parameters, including 20/40-NN and $2std$ / $(3std + 0.02)$, with the rule-based method. For each configuration, we regress RV_t^{RnD} against RV_t^r without intercept to study their correlations, and we want RV_t^{RnD} to be close to RV_t^r with a high correlation and low bias.

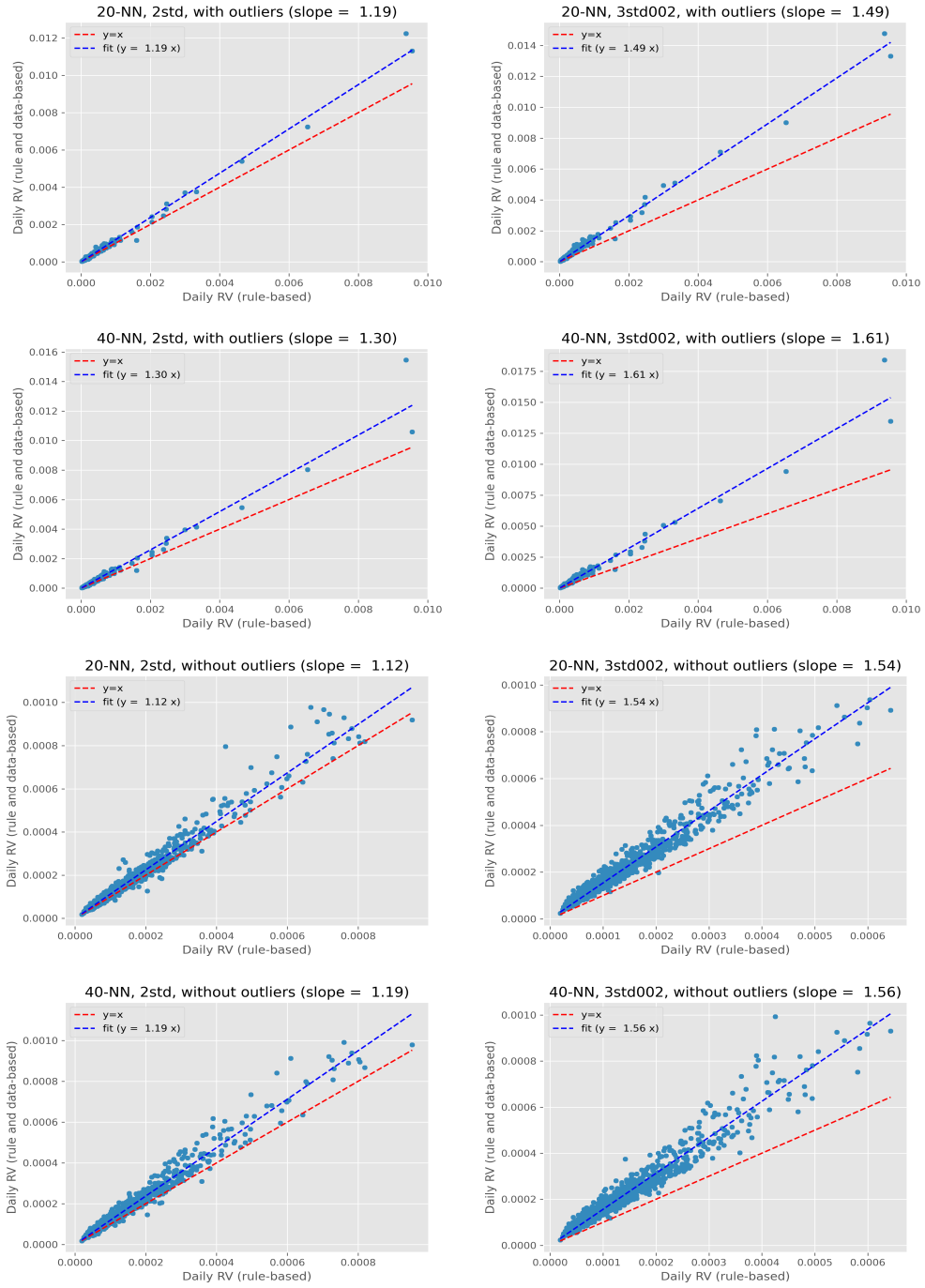

Figure 3.1: The correlation plot of daily RV for rule-based (x-axis) and RnD-based methods (y-axis). For each subfigure, the title shows the parameters of the RnD-based method, whether outliers are excluded, and the bias of the slope of regression. The scatters are each day's RV, The blue dashed line is the regression line, and the red dashed line is the unbiased $y = x$ line.

The scatter plots and regression lines are shown in Fig 3.1, and the slopes and R^2's are shown in Table 3.1. For each setting, we regressed $RV_t^{RnD} = kRV_t^r$ with or without outlier. For each regression, we mainly looked at 2 quantities. One is the R^2, that we want the R^2 to be as close to 1 as possible, which means the RV_t of the RnD-based method and of the rule-based method are highly correlated. The second quantity is the regression coefficient k that we want it to be as close to 1 as possible. If $k = 1$, then there is no bias of our RnD-based method from the well-established rule-based method. From these subfigures, we could see that no matter with or without outliers, $R^2 \approx 1$. That is to say, our new method is highly correlated with the well-established method whatever parameters are chosen. But different parameters yield different regression coefficients k, which means different levels of bias. If we choose 20-NN and $2std$ as we recommended, k is about 1.15 that the bias is about 15%. If we choose 40-NN and $3std + 0.02$ as recommended by Brownlees and Gallo (2006), we will have $k > 1.5$ that the bias is more than 50%. That's why we recommend 20-NN and $2std$.

		with outlier		without outlier	
largest deviation	NN	slope	R^2	slope	R^2
$2std$	20	1.19	0.993	1.12	0.985
$2std$	40	1.30	0.971	1.19	0.988
$3std + 0.02$	20	1.49	0.992	1.54	0.985
$3std + 0.02$	40	1.61	0.978	1.56	0.983

Table 3.1: The regression results of daily RV for RnD-based methods against the well-established rule-based method. The slope and R^2 are reported. The slope can be viewed as 1 + bias. Different parameters for RnD-based cleaning methods all have very high R^2 but different biases. The 20-NN and 2std are yielding least bias among other parameters.

Second, we could look at the proportion of data kept after filtering by the rule-based and RnD-based methods. The results are shown in Table 3.2. We could see that rule-based method keeps about 33% of records and about 20% of trading volumes. In contrast, our proposed RnD-based method keeps about 66% of records and 38% trad-

ing volumes. Basically our method keeps two times the well-established rule-based method. The trading volumes are kept less than the records, because many records with very large trading volumes are reported to the Financial Industry Regulatory Authority, which are deleted in both two methods because they are more likely to be darkpool transactions or delayed reports. Note that the calculation of RV doesn't depend on the trading volume, but the number of trading records.

3.4.2 Applications on RV

In this part, we study the out-of-sample RS test on SPY and DJI component stocks during 2016 and 2022. We clean data by the RnD-based cleaning method as proposed in Section 3.3, use the transaction level data to aggregate the RV for each minute during trading hours, and set each trading day as an episode. We break the whole sequence into two relatively independent parts, one is from 2016 to 2018 and the other is from 2019 to 2022. We view these two sequences for the same asset as two independent sequences generated from the same distribution, so we use one sequence for training and the other for the out-of-sample RS testing. We mainly focus on the RS-Heston-3/2 model.

The out-of-sample RS testing results, including the estimated out-of-sample rank correlation τ and p-value on both 2016-2018 and 2019-2022, are shown in Table 3.3. First of all, we could see that the majority of these RS tests are significant with p-values less than 0.001, and the estimated τ's are large that in general greater than 0.1. So we could conclude that the RS phenomenon is universal among different assets and different time periods in the stock market.

Also we can find that if we test on 2016-2018, most assets have significant out-of-

	Rule-based		RnD-based	
asset	% records	% volumes	% records	% volumes
SPY	23%	14%	73%	47%
AAPL	24%	16%	55%	36%
AMGN	37%	22%	68%	40%
AXP	39%	24%	71%	41%
BA	26%	18%	54%	35%
CAT	37%	23%	69%	41%
CRM	35%	21%	65%	39%
CSCO	25%	18%	65%	40%
CVX	35%	20%	68%	36%
DIS	29%	20%	57%	36%
DOW	35%	21%	69%	37%
GS	36%	20%	65%	37%
HD	34%	20%	64%	37%
HON	37%	22%	67%	38%
IBM	34%	20%	66%	36%
INTC	26%	18%	65%	40%
JNJ	33%	18%	65%	33%
JPM	32%	21%	67%	37%
KO	25%	18%	62%	33%
MCD	35%	21%	65%	38%
MMM	37%	20%	66%	34%
MRK	32%	20%	68%	35%
MSFT	30%	19%	63%	40%
NKE	35%	21%	67%	38%
PG	36%	19%	69%	33%
TRV	43%	23%	73%	39%
UNH	37%	21%	67%	36%
V	35%	19%	67%	34%
VZ	24%	18%	62%	34%
WBA	33%	21%	70%	43%
WMT	32%	22%	64%	40%
average	33%	20%	66%	38%

Table 3.2: The proportion of records and volumes kept under rule-based and RnD-based methods (20-NN and $2std$). We could see the majorities are very significant, and if we train in 2019-2022 and test in 2016-2018, RS tests are nearly all significant.

sample RS testing results with p-value less than 0.001 except HON, which has a less-than-0.01 p-value 0.0027. The reason for the universal significance could be that though 2016-2018 is a highly volatile period, it is not as volatile as 2019-2022, in the sense that its 'evolution of the changing dynamical system' is relatively stable and predictable. So it is more significant than the test in 2019-2022, during which period, the fundamentals are changing across different time periods, that the 'evolution of the changing dynamical system' is not stable. But even under this scenario, the RS test is still significant on more than half of the assets.

3.5 Discussion and Conclusion

In Project II, we first proposed a data cleaning method for TAQ data, that has a better trade-off between the data quality and data size. This method combines the well-established rule-based cleaning method and a data-based cleaning method. This new method keeps about one times more data than the well-established rule-based method, and the output RV of the new method is highly correlated with the rule-based method with only a moderate bias. This cleaning method can be used in a larger scope where the data size is an important consideration.

Second, we applied out-of-sample RS tests on stocks' RV processes and demonstrated that RS is a universal phenomenon across different assets and time periods. These results are stronger than the results for VIX in Project I which is an in-sample RS test on a single volatility process over a half-year time period. In this project, we did out-of-sample RS tests for a wide variety of assets over a 7-year period.

One future direction of this project is to study why the regime-switching phenomenon is universal among the assets we studied. The findings in this project are empirical

	Test: 2019-2022		Test: 2016-2018	
asset	τ	p-value	τ	p-value
SPY	0.046	5.7e-02	0.306	3.2e-36
AAPL	0.049	4.2e-02	0.211	3.7e-18
AMGN	0.030	2.2e-01	0.081	8.3e-04
AXP	0.166	7.4e-12	0.160	4.8e-11
BA	0.039	1.1e-01	0.155	1.9e-10
CAT	0.104	1.7e-05	0.136	2.1e-08
CRM	0.086	4.2e-04	0.194	1.3e-15
CSCO	0.183	4.8e-14	0.221	9.2e-20
CVX	0.174	7.7e-13	0.119	1.1e-06
DIS	0.103	2.0e-05	0.212	3.1e-18
DOW	0.092	3.1e-04	0.212	8.8e-11
GS	0.036	1.3e-01	0.114	2.8e-06
HD	0.079	1.2e-03	0.093	1.4e-04
HON	0.066	6.8e-03	0.073	2.7e-03
IBM	0.027	2.7e-01	0.129	1.0e-07
INTC	0.255	8.2e-26	0.273	4.1e-29
JNJ	0.107	1.2e-05	0.180	1.5e-13
JPM	0.154	2.6e-10	0.191	3.9e-15
KO	0.191	3.4e-15	0.238	1.4e-22
MCD	0.104	1.8e-05	0.105	1.7e-05
MMM	0.075	2.1e-03	0.206	2.9e-17
MRK	0.125	2.9e-07	0.175	6.2e-13
MSFT	0.132	5.7e-08	0.272	5.7e-29
NKE	0.139	1.1e-08	0.234	8.0e-22
PG	0.162	2.7e-11	0.183	5.4e-14
TRV	0.098	5.7e-05	0.101	3.0e-05
UNH	0.106	1.4e-05	0.113	3.6e-06
V	0.107	1.0e-05	0.145	2.8e-09
VZ	0.320	1.2e-39	0.214	1.6e-18
WBA	0.122	5.6e-07	0.151	5.2e-10
WMT	0.190	5.6e-15	0.165	1.3e-11
#		22 / 31		30 / 31

Table 3.3: RS test results of SPY and DJI components. The highlighted cells are $\tau > 0.1$ or p-value less than 0.001. The last row is the count of significant assets for each period with p-value less than 0.001.

without specifying the underlying economic model that governs the asset pricing. Actually we view our RS model as an approximation of a very sophisticated model that can be time-varying or time-invariant. So an important future direction is to find the economic reason behind the regime-switching phenomenon. One possible solution is to find a time-invariant SDE model that exhibits the RS phenomenon. For example, we can find a slight RS phenomenon in 4/2 models, but it is far weaker than the RS phenomenon we observed in our empirical studies. Also it is possible that the SDE model we want is too complicated for practical usage. So another solution is to provide a qualitative explanation, which could be behavioral and be out of the scope of statistics. Besides finding the reasons behind the regime-switching phenomenon, another future direction is to better calculate RV. In this study we used transaction-level data to calculate minute-level RV. However, the transactions suffer from market microstructure noise. There is a rich literature on avoiding market microstructure noise, but they are usually only applicable for daily RV calculation because they downsample the price process with a large time interval such as 5 minutes. There are two ways to solve this issue. One way is to denoise the price process in a way that allows minute-level RV calculation, but it can be too hard a problem. Another way is to prove that the RS structure is persistent. One test can be finding whether we can embed the RS structure into RV forecasting, because the predicted RV is persistent and not impacted by transient market microstructure noise. So if the RS structure helps RV forecasting, then the RS structure should be persistent and is not a pure market microstructure noise.

Chapter 4

Project III

4.1 Summary

In this project, we study the Regime-Switching (RS) phenomenon for scientific process forecasting. A RS model assumes that at different time periods the observed process follows different dynamical systems. It is an important phenomenon because the unobserved external environment could largely impact the system and change fundamentally and slowly. So we could represent it by the RS model switching between a set of simplified dynamical systems, each of which is conditioned on those unobserved external factors.

Beyond the RS phenomenon, the evolution of many systems, such as those environmental dynamical systems, is changing periodically. A period could be a day, that during the noon and night the evolution will follow different patterns. A period could also be a year, in which the summer and winter dynamical systems are different. So we need a RS model that could model this strong periodicity.

The general framework for modeling RS is by assuming a latent state process controlling the regimes, which needs to be modeled further. The most common way is to assume it follows a time-homogeneous Markov chain, so the state transition probabilities are time-invariant and memoryless, which is the RS assumption we used in

Project I and II.

The property we want to add to the latent state process is strong periodicity. One way to model the strong periodicity is by harmonic functions. Harmonic functions are widely used to represent strong periodicity, such as in periodic ARMA (Dudek, Hurd, and Wójtowicz, 2016). Harmonic functions can guarantee the patterns of extreme phases during each period, such as the summer or winter phases for a yearly trend. However, it has difficulties in modeling the intermediate phases, for example, for yearly periods, during the spring and fall we could observe the summer or winter dynamical systems because of the highly volatile external environment during these phases. These intermediate phases are oscillating randomly and we should not impose a deterministic periodicity for them. As a result, harmonic functions are not suitable because they are deterministic. We still need random components.

Our motivation is to combine the periodicity and Markov chain for modeling the latent transitions to make the latent state process simultaneously have the periodic and Markovian property. But these two properties are contradicting essentially. A time-homogeneous Markovian process in general doesn't have a strong periodicity, and a periodic process in general cannot be Markovian. In this project, we propose a Periodically-Regime-Switching (PRS) model. This model resolves the contradiction of Markov and periodically switching, by assuming the transition probabilities changing periodically. This model lies in between the spectrum from the deterministic strong periodicity to the purely random Markov chain. Note that the periodicity we discussed in this project is not the periodicity of the Markov chain, which is not a suitable model for our problems.

To show its performance, we apply our proposed method on chlorophyll forecast under the Empirical Dynamic Modeling (EDM) framework (Sugihara and May, 1990;

Sugihara, 1994). EDM framework is an important tool for forecasting a dynamical system that could be described by a set of PDEs or SDEs. It is similar to an AutoRegressive (AR) model (Box et al., 2015). We proposed the RS version of EDM that assumes there is a set of prediction functions each of which dominates different time stamps. Then we proposed PRS-EDM. In this application, we mainly work on PRS with Linear Regression (PRS-LR) model. This modeling technique that extends the RS to PRS can be straightforwardly generalized to other RS models, such as HMM.

The contribution of this project is two-folded. The first fold is to propose the PRS model that lies in between the strong periodic process and the purely random Markov process. The second fold is to propose the RS and PRS versions of the EDM framework.

4.2 Background and Literature Reviews

Algal bloom is the phenomenon that the population of algae grows rapidly in a water system. It is a global problem (Pick, 2016; Ho, Michalak, and Pahlevan, 2019; Ndlela et al., 2016) related to the climate change (Paerl and Huisman, 2008; Paerl and Huisman, 2009; Paerl and Paul, 2012; Paerl, Gardner, et al., 2016; Ho and Michalak, 2020). A toxic algal bloom is directly harmful to humans (Ciaccio et al., 2015; Chorus and Bartram, 1999), and every year, it costs the government 4 billion dollars (Kudela, Berdalet, and Urban, 2015). To mitigate its effect, researchers are trying to react earlier by finding early warnings (Pace et al., 2017; Wilkinson et al., 2018). Measuring the toxic algal bloom is costly and cannot be done automatically, so it is hard to collect a large data set. An accessible approximation for algae is the chlorophyll in the water system, which is an indicator of algal blooms (Fennel et al.,

2022) and can be measured automatically.

The accumulation and decrease of algae is a part of a complicated water dynamical system. It involves many variables and is typically represented by PDEs (Glover, Jenkins, and Doney, 2011). One example is the Nutrients-Phytoplankton-Zooplankton-Detritus (NPZD) model that is over-simplified. A slightly larger model involves different kinds of phytoplankton, zooplankton, nutrients and detritus, such as a 9-component NPZD model to model the spring algal bloom proposed by Ji et al. (2006). Modeling across different seasons needs to involve more factors such as temperature (refer to Fennel et al., 2022, for more on this topic).

To collect data involved in this complicated system, we need a lot of lab work and cannot be measured automatically, which leads to the difficulties in high-frequency data collection. To forecast a time series better, we need to take the trade-off between larger data size and better data quality. The lab experiments give us comprehensive measurements of the water quality in a low frequency, such as daily or weekly. If we have enough data of this kind, we can model and estimate the full system of PDEs (Møller, Madsen, and Carstensen, 2011), but in general it is too costly and the data is limited. Another kind of data is collected from the automatic shore stations which measure the water quality in a high frequency, such as every 15-minute, but only measures a limited part of variables. We focus on this kind of datasets and study how to better forecast based on the high-frequency measurements with limited variables.

The data set we are using is from Virginia Institute of Marine Science's automatic shore station at Willis Wharf site. It automatically measures the water quality (Ross and Snyder, 2020) including chlorophyll, along with cross predictors including water

temperature, salinity, pH, optical dissolved oxygen, turbidity and wind speed[1].

The state-of-the-art method to solve this problem is under the EDM framework (Sugihara and May, 1990; Sugihara, 1994; Dietze et al., 2018; Bailey, Doney, and Lima, 2004; Yu et al., 2021). EDM guarantees that we can predict a time series by its lagged observations if this time series can be represented by a dynamical system of PDEs. Mathematically, if we have a dynamical system involving some variables, and we can observed a subset of them $\{Y_t\}_{t=1}^T$ with a fixed time interval, then there exist a prediction function $f(\cdot)$ that $Y_{t+h} = f(Y_t, \cdots, Y_{t-k})$, where h is the forecast horizon and k is the lag to use and is mathematically proved to be finite. EDM is proposed if the data follows a deterministic dynamical system of PDEs without randomness. Later researchers have used it for systems with random noise (Cenci, Sugihara, and Saavedra, 2019), and the prediction function could be written as $\widehat{Y}_{t+h} = f(Y_t, \cdots, Y_{t-k})$. EDM framework doesn't tell how to find the prediction function $f(\cdot)$, and we can use different regression methods to find it, such as linear regression, support vector machine (Cortes and Vapnik, 1995) and random forest (Breiman, 2001). If EDM is nested with linear regression, then the model has the same mathematical form as AR(k) though their model assumptions are different. We can view the AR model as a specific case of linear regression because the inference of AR is equivalent to a linear regression if AR is fitted by ordinary least squares. A specific regression method widely used in EDM is S-map (Sugihara, 1994), which uses locally exponentially weighted linear regression for prediction.

EDM's prediction function $f(\cdot)$ is time-invariant and suffers from model complexity issues. We propose RS and PRS prediction functions $f(\cdot; t)$ to relax the time invariance assumption. Here I briefly introduced the studies on RS for linear regression or

[1]Refer to https://www.vims.edu/esl/research/water_quality/williswharf.php for the dataset.

AR. RS-AR, commonly called Markov-switching AR, is a special case of RS-Linear Regression (RS-LR), commonly called Markov-modulated linear regression. It is first introduced for studying gross national income (Hamilton, 1989) and then widely used in economics (Adejumo, Albert, and Asemota, 2020; Tuaneh, Essi, and Etuk, 2021; Deschamps, 2008) and environmental science, such as for wind prediction (Pinson and Madsen, 2012; Ailliot and Monbet, 2012). These methods don't assume any seasonality so some researchers have proposed variants including the seasonal components. For example, Aliat and Hamdi (2018) proposed to use periodic Markov-switching AR whose parameters of each state are strictly periodic; Spezia, Paroli, and Dellaportas (2004) proposed to use Markov-switching AR with additive harmonic processes with applications on air pollution prediction. To the best of our knowledge, there is no work that assumes the latent state process is changing both periodically and randomly.

4.3 Methodology

4.3.1 PRS models

We propose the PRS model for the latent state process. The motivation is to model the latent state process by a conditional Markov process, that the state transition probabilities are periodic, and conditioned on the state transition probabilities it is Markovian. This process is in between the deterministic harmonic functions and the purely random time-homogeneous Markov chain. Formally, suppose we have a state process $\{h_t\}_t$ that is transiting among S states, and we want the transition probability to have a period L. Then at time t, the conditional transition probabilities from

Algorithm 8: Algorithm to infer $\boldsymbol{T}(\cdot)$ (Form I).

Data: $\{\xi_t(i,j)\}_{t=1:(T-1)}^{i,j=1:S}$, L
Result: $\{(\lambda_{ij}^{(c)}, \lambda_{ij}^{(s)}), \lambda_{ij}^{(intercept)})\}_{i=1:S}^{j=1:S}$
for $i \leftarrow 1$ **to** S **do**
  $X, Y, W \leftarrow [], [], []$;
  for $t \leftarrow 1$ **to** $T-1$ **do**
    for $j \leftarrow 1$ **to** S **do**
      Append $[\cos(\frac{2\pi t}{L}), \sin(\frac{2\pi t}{L})]$ on X;
      Append j on Y;
      Append $\xi_t(i,j)$ on W;
    end
  end
  Fit a weighted multinomial-logistic regression with a nearly 0 regularization based on predictors X, responses Y and weight W, and $\{(\lambda_{ij}^{(c)}, \lambda_{ij}^{(s)}), \lambda_{ij}^{(intercept)})\}_{j=1}^{S}$ is the regression coefficients and intercepts.
end

$h_t = i$ to $h_{t+1} = j$ is a function of $\cos(\frac{2\pi t}{L} + \phi)$ or a function of $\cos(\frac{2\pi t}{L})$ and $\sin(\frac{2\pi t}{L})$. Mathematically, we have some function $g_{ij}(\cdot)$ that

$$\boldsymbol{T}_{ij}(t) = \mathrm{P}(h_{t+1} = j | h_t = i) = g_{ij}\left(\cos(\frac{2\pi t}{L} + \phi_{ij})\right) = g_{ij}\left(\cos(\frac{2\pi t}{L}), \sin(\frac{2\pi t}{L})\right).$$

We could use either the phase lag form or the sine-cosine form. In general they are equivalent when the sine and cosine functions are linearly added, where the phase lag form is better for scientific understanding and the sine-cosine form is better for inference.

We proposed the following parametric form of this $g_{ij}(\cdot)$:

- Form I: to use multinomial logistic regression type on harmonic terms that

$$\boldsymbol{T}_{ij}(\frac{2\pi t}{L}) = \frac{\exp\left(\lambda_{ij}^{(c)}\cos(\frac{2\pi t}{L}) + \lambda_{ij}^{(s)}\sin(\frac{2\pi t}{L}) + \lambda_{ij}^{(intercept)}\right)}{\sum_{s=1}^{S}\exp\left(\lambda_{is}^{(c)}\cos(\frac{2\pi t}{L}) + \lambda_{is}^{(s)}\sin(\frac{2\pi t}{L}) + \lambda_{is}^{(intercept)}\right)} \tag{4.1}$$

or in phase lag form

$$\boldsymbol{T}_{ij}(\frac{2\pi t}{L}) = \frac{\exp\left(\lambda_{ij}\cos(\frac{2\pi t}{L}+\phi_{ij}) + \lambda_{ij}^{(intercept)}\right)}{\sum_{s=1}^{S}\exp\left(\lambda_{is}\cos(\frac{2\pi t}{L}+\phi_{is}) + \lambda_{is}^{(intercept)}\right)};$$

- Form II: to use truncated linear type on harmonic terms that

$$\boldsymbol{T}_{ij}(\frac{2\pi t}{L}) = \frac{\left(\lambda_{ij}^{(c)}\cos(\frac{2\pi t}{L}) + \lambda_{ij}^{(s)}\sin(\frac{2\pi t}{L}) + \lambda_{ij}^{(intercept)}\right)^{+}}{\sum_{s=1}^{S}\left(\lambda_{is}^{(c)}\cos(\frac{2\pi t}{L}) + \lambda_{is}^{(s)}\sin(\frac{2\pi t}{L}) + \lambda_{is}^{(intercept)}\right)^{+}} \tag{4.2}$$

or in phase lag form

$$\boldsymbol{T}_{ij}(\frac{2\pi t}{L}) = \frac{\left(\lambda_{ij}\cos(\frac{2\pi t}{L}+\phi_{ij}) + \lambda_{ij}^{(intercept)}\right)^{+}}{\sum_{s=1}^{S}\left(\lambda_{is}\cos(\frac{2\pi t}{L}+\phi_{is}) + \lambda_{is}^{(intercept)}\right)^{+}}$$

where $(\cdot)^{+}$ is the positive part function that $(\cdot)^{+} = \max(\cdot, 0)$.

Besides these two forms, we could use any function of $\cos(\frac{2\pi t}{L})$ and $\sin(\frac{2\pi t}{L})$ as long as it is a valid transition matrix that all elements are non-negative and rows sum up 1. The proposed Form I and II are easier to infer, especially Form I since it is second-order differentiable so can be inferred by Newton-Raphson algorithm. Form II is similar to Form I because we could view $\exp(\cdot)$ as a soft approximation of $(\cdot)^{+}$. The inference of these two forms can be done by Maximum Likelihood Estimation (MLE) based on the intermediate results $\{\xi_t(i,j)\}_{t=1:(T-1)}^{i,j=1:S}$ from the E-step in E-M algorithm. The pseudocode for inferring Form I is shown in Alg 8, which converts the problem into a multinomial-logistic regression. It requires a nearly 0 regularization or set $\lambda_{ij}^{(c)} = \lambda_{ij}^{(s)} = \lambda_{ij}^{(intercept)} = 0$ for all $j = 1$ to avoid the identifiability problem. Inferring Form II is similar but not as standard as using the Newton-Raphson algorithm for

Form I, so it will be slower and more vulnerable than Form I. The advantage of Form II is that it is linear to the harmonic functions unless truncated, while Form I is linear to the harmonic terms only when it is close to 0. In practice, we recommend trying both 2 forms and finding the better one for prediction, but if you are unsure which one to use, Form I is recommended because it is standard and robust.

Any RS model can be extended to this PRS model if we believe the latent state processes are strongly-periodic and Markovian processes. In this section, we mainly focus on the PRS model with linear regression, which is the PRS version of Markov-modulated linear regression (Andronov and Spiridovska, 2019). The adoption of PRS for any other RS models, such as HMM, is straightforward. If we apply PRS on HMM, then it is similar to the HMM with exogenous input (Bengio and Frasconi, 1994) that uses $\cos(\frac{2\pi t}{L})$ and $\sin(\frac{2\pi t}{L})$ as exogenous input, and Form I parametrization by multinomial-logistic functions is widely used in this scenario (Filardo, 1994; Shirley et al., 2010).

4.3.2 RS-EDM and PRS-EDM

Algorithm 9: E-M algorithm for inferring PRS-LR model

Data: $\{(t, X_t, Y_t)\}_{t=1}^{T}$, L
Result: $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$
Initialize $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$;
while *not converged* **do**
  E-step: update $\{\gamma_t(j)\}_{t=1:T}^{j=1:S}$ and $\{\xi_t(i,j)\}_{t=1:(T-1)}^{i,j=1:S}$ conditioned on $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$ by Alg 10;
  M-step: update $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$ conditioned on $\{\gamma_t(j)\}$ and $\{\xi_t(i,j)\}$ by Alg 11;
end

EDM (Sugihara and May, 1990; Sugihara, 1994) is a powerful framework to fore-

Algorithm 10: E-step of the E-M algorithm for inferring PRS-LR model

Data: $\{(t, X_t, Y_t)\}_{t=1}^{T}$, L, $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$
Result: $\{\gamma_t(j)\}_{t=1:T}^{j=1:S}$, $\{\xi_t(i,j)\}_{t=1:(T-1)}^{i,j=1:S}$

```
// Emission probabilities
```
for $t \leftarrow 1$ **to** T **do**
  for $j \leftarrow 1$ **to** S **do**
    $b_j(t) \leftarrow pdf(Y_t - X_t\widehat{\boldsymbol{\beta}_j}; \mathcal{N}(0, \widehat{\sigma_j^2}))$;
  end
end
```
// Forward probabilities
```
for $t \leftarrow 1$ **to** T **do**

$$\alpha_t(j) \leftarrow \begin{cases} \pi_0(j) b_j(1), & \text{if } t = 1 \\ \sum_{i=1}^{S} \alpha_{t-1}(i) \boldsymbol{T}_{ij}(\frac{2\pi t}{L}) b_j(t), & \text{otherwise} \end{cases} \quad \forall j = 1, \cdots, S;$$

end
```
// Backward probabilities
```
for $t \leftarrow T$ **to** 1 **do**

$$\beta_t(j) \leftarrow \begin{cases} 1, & \text{if } t = T \\ \sum_{j=1}^{S} \boldsymbol{T}_{ij}(\frac{2\pi t}{L}) \beta_{t+1}(j) b_j(t+1), & \text{otherwise} \end{cases} \quad \forall j = 1, \cdots, S;$$

end
```
// State occupancy probabilities
```
$\mathrm{P}(h_t = j) = \gamma_t(j) \leftarrow \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{S} \alpha_t(i)\beta_t(i)}$ $\forall t, j$;
$\mathrm{P}(h_t = i, h_{t+1} = j) = \xi_t(i,j) \leftarrow \frac{\alpha_t(i)\boldsymbol{T}_{ij}(\frac{2\pi t}{L}) b_j(X_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{S} \alpha_t(i)\beta_t(i)}$ $\forall t, i, j$;

Algorithm 11: M-step of the E-M algorithm for inferring PRS-LR model

Data: $\{(t, X_t, Y_t)\}_{t=1}^{T}$, $\{\gamma_t(j)\}_{t=1:T}^{j=1:S}$, $\{\xi_t(i,j)\}_{t=1:(T-1)}^{i,j=1:S}$
Result: $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$
$\widehat{\pi}_0(j) \leftarrow \gamma_1(j)$ $\forall j = 1, \cdots, S$;
Find $\widehat{\boldsymbol{T}}_{ij}(\cdot)$ by MLE according to the dataset $\{\xi_t(i,j)\}_{i,j=1:S}^{t=1:T}$;
for $j \leftarrow 1$ **to** S **do**
  Find $\widehat{\boldsymbol{\beta}_j}$ and $\widehat{\sigma_j^2}$ by weighted least square with predictors $\{X_t\}_t$, responses $\{Y_t\}_t$ and weights $\{\gamma_t(j)\}_t$;
end

cast variables governed by a PDE-based or SDE-based dynamical system. In our application, the oscillation of the chlorophyll process is typically described by a set

Algorithm 12: Fitting and h-step-forward prediction based on PRS-LR model

Data: $\{(t, X_t, Y_t)\}_{t=1}^{T}$, L, h, X_{t+h}
Result: $\widehat{Y_{t+h}}$
`// Fitting`
Fit $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$ by Alg 9;
Find state occupancy probability $\gamma_t(j)$ for $j = 1, \cdots, S$ by Alg 10;
`// Prediction`
$\boldsymbol{T}_{(T|T+h)} \leftarrow \prod_{r=0}^{h-1} \boldsymbol{T}(\frac{2\pi(T+r)}{L})$;
for $s \leftarrow 1$ **to** S **do**
  $\gamma_{t+h}(s) \leftarrow \sum_i \left[\boldsymbol{T}_{(T|T+h)}\right]_{is} \gamma_t(i)$;
end
$\widehat{Y_{t+h}} \leftarrow \sum_s \gamma_{t+h}(s) X_{t+h} \boldsymbol{\beta}_s$;

of unknown PDEs, so we use the EDM framework to forecast it. Mathematically, suppose we have observed some equidistant processes $\{(U_t, V_t)\}_t$ governed by some PDEs or SDEs, that V_t is the target response to predict and U_t are some cross predictors. If we want to predict h-forward response V_{t+h} based on information up to t, then EDM framework guarantees that there exist a prediction function $f(\cdot)$ whose input is lagged observations $V_{(t-k):t}$ and $U_{(t-m):t}$ and whose output is the predicted $\widehat{V_{t+h}}$, i.e.

$$\widehat{V_{t+h}} = f(U_{(t-m):t}, V_{(t-k):t}). \tag{4.3}$$

If the underlying system is PDE-based and deterministic, then $\widehat{V_{t+h}} = V_{t+h}$. If we cannot observe any cross predictors, the prediction function could be simplified as $\widehat{V_{t+h}} = f(V_{(t-k):t})$. In short, EDM guarantees the existence of this AR-type prediction function with finite lags. But EDM doesn't tell us how to find this prediction function, and we can find it by any suitable machine learning methods, such as linear regression, support vector machines (Cortes and Vapnik, 1995; Drucker et al., 1996) or random forest (Breiman, 2001), and fit it from observed training processes. If we use linear regression, we can write the prediction function as

$\widehat{V_{t+h}} = \beta_0 + \sum_{i=0}^{k} \beta_i^{(U)} U_{t-i} + \sum_{j=0}^{m} \beta_j^{(V)} V_{t-j}$. We could view it as a linear regression problem that predictors are $X_t = [U_{(t-m):t}, V_{(t-k):t}]$ and the response $Y_t = V_{t+h}$. We will use the (X_t, Y_t) parametrization instead of (U_t, V_t) parameterization when possible as it is a more general framework not restricted to EDM. Particularly, a dedicated method S-map (Sugihara, 1994) is proposed along with EDM.

However, standard EDM has the model complexity issue when the underlying dynamical system is very complicated and involves a lot of variables. Though EDM doesn't require all variables in the system to be observable, it will use a large lag to accommodate it, which will make the input of the prediction function high-dimensional and suffer from the curse of dimensionality (Keogh and Mueen, 2017). This issue is not solvable under the standard EDM framework because it is too flexible a framework that the prediction function is universe and time-invariant without restrictions. So this prediction function has a high complexity and requires a large training set.

Our motivation to solve this problem is to relax the universe and time-invariant prediction function assumption. Mathematically in the standard EDM framework, the prediction function $f(\cdot)$ in Eq 4.3 is time-invariant, that the prediction function at time t $f(\cdot; t) \equiv f(\cdot)$. We relax this assumption by allowing $f(\cdot; t)$ to be time-dependent to accommodate the changing external environment. The intuition is that we believe there exist many unobserved variables that are time-dependent. There are multiple ways to assume its time-dependent structure, and in this project we propose to use the RS structure. That is to say, we simplified the prediction function conditioned on

different external environments into several categories. Mathematically, we assume

$$\begin{aligned} h_t \in \{1, \cdots, S\} \quad &\text{for} \quad t = 1, \cdots, T; \\ f(\cdot; t) = f_s(\cdot) \quad &\text{if} \quad h_t = s; \\ \widehat{Y}_t = f_s(X_t), & \end{aligned} \tag{4.4}$$

where $\{h_t\}_t$ is the latent state process that controls which prediction function is in place at each time stamp.

$\{h_t\}_t$ could be modeled in different ways. The most common choice is to model the latent state process by a Markov chain, and we name this model as the RS-EDM model. Another way to model the latent state process $\{h_t\}_t$ in Eq 4.4 is by a PRS model we proposed in Section 4.3.1. We name this method as PRS-EDM. Mathematically, it assumes

$$\begin{aligned} \{h_t\}_{t=1}^T \quad &\sim \quad PRS; \\ f(\cdot; t) = f_s(\cdot) \quad &\text{if} \quad h_t = s; \\ \widehat{Y}_t = f_s(X_t). & \end{aligned}$$

PRS-EDM guarantees the model switching has a strong periodic trend. The period could be 1 day if we want to make hour-level prediction, or could be 1 year if we want to make day-level prediction. If the period length is one year, then we could think there is a summer mode dominating the summer period and a winter mode dominating the winter period. During spring and autumn, the model is switching in between these two modes randomly. The scientific motivation behind is that many unobserved variables involved in the underlying dynamical system are changing periodically. They

have the summer mode and winter mode, making the prediction function also have these modes.

A technical detail using RS-EDM and PRS-EDM for h-step forward forecasting is that at time T, $Y_t = V_{t+h}$ is not observed if $t > T - h$. For example, if $h = 1$, then we can observe $\{(X_t, Y_t)\}_{t=1}^{T-1}$ and X_T, and we first find the state occupancy probabilities for $T - 1$, perform a one-step forward propagation, and predict $\widehat{Y_T}$. If $h \geq 2$, then we can observe a truncated training set $\{(X_t, Y_t)\}_{t=1}^{T-h}$ and the cross predictor X_T to predict, so we first find the state occupancy probabilities for $T - h$, perform a h-step forward propagation, and predict $\widehat{Y_T}$.

4.3.3 RS-LR and PRS-LR

Though RS-EDM and PRS-EDM are motivated from EDM, they are not restricted in the EDM context. In EDM, we could write the data into the $\{(X_t, Y_t)\}_t$ form where X_t is the lagged cross predictors and observations and Y_t is the h-step forward forecasting target. That is to say, EDM only tells us how to construct the predictors and responses and the following inference procedure doesn't rely on EDM. Particularly, when we use linear regression for estimating the prediction function for EDM, the EDM reduces to a linear regression, and similarly RS-EDM reduces to a Markov-modulated linear regression, which we name as RS-LR for consistent terminology in this section, and PRS-EDM reduces to PRS-LR. Note that just as linear regression is not necessarily from the EDM framework, RS-LR and PRS-LR have a wider application than EDM.

RS-LR and PRS-LR can be inferred by MLE with the E-M algorithm. The pseudocode of PRS-LR inference is shown in Alg 9. The predictions for PRS-LR are straightforward by forward algorithm, and the pseudocode of its fitting and pre-

diction procedure is shown in Alg 12. For PRS with other regression methods or distributions, the generalization is straightforward. Also the inference of RS-LR is a special case of PRS-LR. If we estimate $\boldsymbol{T}(\cdot)$ in the M-step, shown in Alg 11, by the time-invariant transition matrix that $\widehat{\boldsymbol{T}}_{ij} \leftarrow \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1}\sum_{k=1}^{S} \xi_t(i,k)}$ $\forall i, j = 1, \cdots, S$, then this PRS model reduces to RS model.

4.3.4 Feature importance

Algorithm 13: In-sample feature selection of PRS-LR model

Data: $\{(t, Z_t, X_t, Y_t)\}_{t=1}^{T}$, L
Result: p-value
Fit $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$ based on full data $\{([Z_t, X_t], Y_t)\}_{t=1}^{T}$ by Alg 9;
Find the last forward probability $\alpha_T(j)$ for $j = 1, \cdots, S$ by Alg 10;
$\mathcal{L}^{(full)} \leftarrow \sum_{j=1}^{S} \alpha_T(j)$;
Fit $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$ based on reduced data $\{(Z_t, Y_t)\}_{t=1}^{T}$ by Alg 9;
Find the last forward probability $\alpha_T(j)$ for $j = 1, \cdots, S$ by Alg 10;
$\mathcal{L}^{(reduced)} \leftarrow \sum_{j=1}^{S} \alpha_T(j)$;
P-value $\leftarrow 1 - F\left(-2[\log(\mathcal{L}^{(reduced)}) - \log(\mathcal{L}^{(full)})]; \chi^2_{\dim(X_t)\times\dim(Y_t)\times S}\right)$ where $F(\cdot; \chi^2_{df})$ is the cumulative probability function of χ^2_{df};

Algorithm 14: In-sample comparsion between PRS-LR and RS-LR models

Data: $\{(t, X_t, Y_t)\}_{t=1}^{T}$, L
Result: p-value
Fit $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$ by Alg 9;
Find the last forward probability $\alpha_T(j)$ for $j = 1, \cdots, S$ by Alg 10;
$\mathcal{L}^{(PRS)} \leftarrow \sum_{j=1}^{S} \alpha_T(j)$;
Fit $\{\widehat{\pi}_0, \widehat{\boldsymbol{T}}(\cdot), \{(\widehat{\boldsymbol{\beta}_s}, \widehat{\sigma_s^2})\}_{s=1:S}\}$ by Alg 9 that the transition matrix is time-invariant;
Find the last forward probability $\alpha_T(j)$ for $j = 1, \cdots, S$ by Alg 10;
$\mathcal{L}^{(RS)} \leftarrow \sum_{j=1}^{S} \alpha_T(j)$;
P-value $\leftarrow 1 - F\left(-2[\log(\mathcal{L}^{(RS)}) - \log(\mathcal{L}^{(PRS)})]; \chi^2_{2S(S-1)}\right)$ where $F(\cdot; \chi^2_{df})$ is the cumulative probability function of χ^2_{df};

Feature importance or feature selection problem is important. In the EDM framework, if we want to make a better prediction, which feature should be included in the model needs to be decided. Also studying the impact from each feature helps us scientifically understand the underlying dynamical system. In PRS models, we want to find whether the periodicity exists or not. Even outside the EDM framework, we still need to find which feature is important and should be included in the prediction model. We propose two approaches to solve it for RS-LR and PRS-LR.

The first approach is the Likelihood Ratio Testing (LRT) feature importance. It is based on a standard LRT to test whether a feature or a set of features has a significant impact on the model. This approach requires the RS or PRS models to be estimated by MLE and have a known degree of freedom. RS-LR and PRS-LR satisfy these requirements and the testing procedure is as follows. Suppose we have the dataset $\{(Z_t, X_t, Y_t)\}_t$, where X_t is a set of q-dimensional cross predictors that we want to test whether it has any impact or not, Z_t is a set of cross predictors conditioned on, and Y_t is a set of p-dimensional response. To perform LRT, we fit 2 nested S-state RS-LR or PRS-LR and get their maximized likelihood. One model is fitted based on the reduced dataset $\{(Z_t, Y_t)\}_t$ and we can have its maximum likelihood $\mathcal{L}^{(reduced)}$, and the other model is fitted based on the full dataset $\{([Z_t, X_t], Y_t)\}_t$ and we can have its maximum likelihood $\mathcal{L}^{(full)}$. The difference of their degrees of freedoms is pqS, so under null hypothesis that Y_t doesn't depend on X_t conditioned on Z_t, we have

$$-2[\log(\mathcal{L}^{(reduced)}) - \log(\mathcal{L}^{(full)})] \sim \chi^2_{pqS}.$$

The feature importance is the p-value of LRT that

$$LRT\ feature\ importance = 1 - F\left(-2[\log(\mathcal{L}^{(reduced)}) - \log(\mathcal{L}^{(full)})]; \chi^2_{\dim(X_t)\times\dim(Y_t)\times S}\right),$$

where $F(\cdot; \chi^2_{df})$ is the cumulative probability function of χ^2_{df}. The pseudocode is shown in Alg 13. Similarly, we can test the periodicity by comparing PRS-LR with RS-LR because they are also nested. Based on the same set of cross predictors, we could fit a RS-LR model to get its likelihood $\mathcal{L}^{(RS)}$ and a PRS-LR model to get its likelihood $\mathcal{L}^{(PRS)}$. The difference of their degrees of freedoms is $2S(S-1)$, if we use the parametric form in Eq 4.1 or Eq 4.2. Under the null hypothesis that RS-LR is the true model, we have

$$-2[\log(\mathcal{L}^{(RS)}) - \log(\mathcal{L}^{(PRS)})] \sim \chi^2_{2S(S-1)}.$$

The LRT feature importance of periodicity is

$$LRT\ feature\ importance = 1 - F\left(-2[\log(\mathcal{L}^{(RS)}) - \log(\mathcal{L}^{(PRS)})]; \chi^2_{2S(S-1)}\right).$$

Its pseudocode is shown in Alg 14.

The second approach is the predictive feature importance. The procedure is standard, that we compare the prediction performance with and without the features we want to test by time series cross-validation or any out-of-sample prediction evaluation methods (see e.g. Hyndman and Athanasopoulos, 2018). Then we compare whether the metrics, such as R^2, have been improved or not if they involve the features we want to test. With a similar idea, we can compare RS-LR and PRS-LR to see whether PRS-LR has a better prediction performance than RS-LR. That is to say, if we use R^2 as the prediction metric, then the predictive feature importance is

$$predictive\ feature\ importance = R^2_{(full)} - R^2_{(reduced)}$$

where $R^2_{(full)}$ is of the full model and $R^2_{(reduced)}$ is of the reduced model without some features or periodicity.

These two feature selection or importance approaches are standard. We want to briefly compare these two methods. The LRT feature importance is based on rigorous statistical testing whether the testing features could explain any part of the data or not, but cannot provide the extent of impact. It is not on whether these features could help prediction and is not forward-looking. The predictive feature importance is based on whether the testing features could help prediction. This method is heuristic and not rigorous statistically because the prediction performance is impacted by the data size due to overfitting. This method is forward-looking. Ideally these two methods should agree with each other if the process is stationary, but when the dynamical system is changing, the predictive feature importance might lose its power and LRT feature importance is still sensitive in general, but the in-sample method might include features that will undermine the prediction performance if these features have a time-varying impact.

4.4 Results

In this section, we test the performance of our proposed methods on simulated datasets and chlorophyll forecasting. In both simulation and real-world application settings, we prove that RS-LR and PRS-LR outperform the time-invariant models without the RS phenomenon, and that PRS-LR outperforms RS-LR when there is periodicity. For PRS-LR, we show the performance with both Form I and II for transition matrix as shown in Eq 4.1 and 4.2. We denote the PRS-LR with Form I as PRS-LR-I and with Form II as PRS-LR-II in this section.

4.4.1 Simulation Results

σ	oracle	LR	RS-LR	PRS-LR-I	PRS-LR-II
0.1	0.83	0.51	0.60	0.82	0.72
1	0.55	0.34	0.37	0.48	0.45

Table 4.1: The simulation results from PRS-LR's prediction for different σ, compared with linear regression (LR) and RS-LR. The oracle is also provided. The reported metrics are R^2, and the best two methods for each setting are highlighted.

We generate data following PRS-LR models and test the prediction performance of different methods on it. We perform 4 experiments. In the first two experiments, we test the prediction accuracy with data generated from different models, where in the first experiment all X_t's are generated independently and in the second experiment the data is generated from an AR(1) model. The second experiment is still under the EDM framework but the first experiment is not. In the third experiment we test the predictive feature importance and in the fourth experiment we test the LRT feature importance.

The first experiment's procedure is as follows. We set $\dim(X) = 5$, $\dim(Y) = 1$, $T = 550$, $L = 50$, $S = 2$. We generated 50 sequences of length 550 independently. For the sequence l, we first generate model parameters. We set the initial probability $[\frac{1}{S}, \cdots, \frac{1}{S}]^\top$, sample the the PRS transition matrix of Form I in Eq 4.1 that $\lambda_{ij}^{(l)(c)}, \lambda_{ij}^{(l)(s)} \sim Unif[10, 20]$ and $\lambda_{ij}^{(l)(intercept)} \sim Unif[0, 1]$ for $i, j = 1, \cdots, S$ independently, and sample the linear regression coefficients $\beta_j^{(l)(s)} \sim Unif[-1, 1]$ for $j = 0, \cdots, \dim(X)$ independently. Then we generate $\{h_t^{(l)}\}_{1:T}$ following the PRS model first, and conditioned on each $h_t^{(l)}$, we generate $X_t^{(l)}$ and $Y_t^{(l)}$, by sampling $X_{t,j}^{(l)} \sim \mathcal{N}(0, 1)$ independently and sampling independently $Y_t^{(l)} \sim \sum_{j=1}^{\dim(X)} \beta_j^{(l)(s)} X_{t,j}^{(l)} + \beta_0^{(l)(s)} + \mathcal{N}(0, \sigma^2)$ conditioned on $h_t^{(l)} = s$ with different signal-noise-ratio that $\sigma = 0.1, 1$. After generating a sequence, we set the last 50 time stamps as the test period,

and for each time stamp t in this test period, we predict $\widehat{Y_t^{(l)}}$ by refitting the model with all data up to time stamp $t-1$, i.e. $\{(X_u^{(l)}, Y_u^{(l)})\}_{u=1}^{t-1}$, and predict one-step forward to find $\widehat{Y_t^{(l)}}$. After predicting all these sequences, we evaluated the prediction performance by R^2 of the prediction $\{\widehat{Y_t^{(l)}}\}_t^l$ and $\{Y_t^{(l)}\}_t^l$ for all t in testing period and all these sequences $l = 1, \cdots, 50$.

We compare both PRS-LR-I and PRS-LR-II with benchmark forecasting methods linear regression and RS-LR. Also we show the oracle R^2 that if we know all underlying model parameters. The resulting R^2 are shown in Table 4.1. First, for all signal-noise-ratios, PRS-LR is pretty close to oracle and outperforms all other methods. Among other methods, RS-LR is better than LR because LR cannot model the RS phenomenon. Though RS-LR could model the RS phenomenon, it cannot model the periodic transition probabilities so it is not as competitive as PRS-LR. Within PRS-LR, we could see that PRS-LR-I is better than PRS-LR-II. In other results, PRS-LR-II occasionally outperforms PRS-LR-I but PRS-LR-I is better in the majority of cases. In most cases, the robustness of the multinomial-logistic form of PRS-LR-I is important.

R^2	persistent	LR	RS-LR	PRS-LR-I	PRS-LR-II
1-step	-0.14	0.07	0.06	0.20	0.20
3-step	-0.42	0.07	0.06	0.11	0.12
7-step	-0.99	0.07	0.06	0.11	0.11

Table 4.2: The simulation results of PRS-LR's prediction for different forecast horizons, including 1/3/7-step forwarding, for AR(1) data generation under PRS transitions, compared with persistent forecasting, linear regression (LR) and RS-LR. The reported metrics are R^2, and the best two methods for each setting are highlighted.

The second experiment's procedure is similar to the first experiment, except we generate data as an AR(1), by setting $X_t = Y_{t-1}$. Particularly, we set $\dim(X) =$

$\dim(Y) = 1$, $T = 550$, $Y_t = \begin{cases} 0.8 \times X_t + \epsilon_t & h_t = 1, \\ -0.8 \times X_t + \epsilon_t & h_t = 2, \end{cases}$ where $\epsilon_t \sim \mathcal{N}(0,1)$, $\{h_t\}_t$ follows PRS transition in Form I with $\lambda_{11}^{(c)} = 11, \lambda_{11}^{(s)} = 20, \lambda_{11}^{(intercept)} = 0.1, \lambda_{12}^{(c)} = 18, \lambda_{12}^{(s)} = 12, \lambda_{12}^{(intercept)} = 0.5, \lambda_{21}^{(c)} = 14, \lambda_{21}^{(s)} = 16, \lambda_{21}^{(intercept)} = 0.5, \lambda_{22}^{(c)} = 15, \lambda_{22}^{(s)} = 16, \lambda_{22}^{(intercept)} = 0.1$. We test the last 50 time stamp with the forecast horizon 1-step, 3-step and 7-step forward forecasting and measured their R^2. The result are shown in Table 4.2. We could see PRS-LR outperforms RS-LR, linear regression and persistent prediction in both the short-term and long-term prediction.

σ	feature removed	noise feature	PRS-LR-I	PRS-LR-II	RS-LR
1	$X_{.,1}$	Yes	-3.9e-02	-2.5e-02	-2.6e-02
1	$X_{.,2}$	Yes	-3.9e-02	-2.5e-02	-2.6e-02
1	$X_{.,3}$	No	0.25	0.28	0.27
1	$X_{.,4}$	No	0.23	0.25	0.25
1	$X_{.,5}$	No	0.21	0.23	0.23
1	periodicity	No	0.012	0.010	-
0.1	$X_{.,1}$	Yes	-1.3e-02	-1.6e-03	-2.7e-03
0.1	$X_{.,2}$	Yes	-1.3e-02	-1.6e-03	-3.6e-03
0.1	$X_{.,3}$	No	0.32	0.34	0.34
0.1	$X_{.,4}$	No	0.29	0.31	0.30
0.1	$X_{.,5}$	No	0.29	0.30	0.30
0.1	periodicity	No	0.018	0.012	-

Table 4.3: The simulation results of PRS-LR and RS-LR predictive feature importance for different σ. The feature importance is the decrease of R^2 if we remove each feature and periodicity. Important features are highlighted. We could see when removing a noisy feature, the R^2 remains nearly unchanged. If we remove a strong feature or periodicity, then the R^2 largely decreases.

The third experiment's procedure is similar to the first experiment, except we set $\beta_1^{(l)(s)}, \beta_2^{(l)(s)} \equiv 0$ and $\beta_3^{(l)(s)}, \beta_4^{(l)(s)}, \beta_5^{(l)(s)} \sim Unif[1,2]$, i.e. the predictors are 5-dimensional and the first 2 dimensions are noisy features. Then we measure the prediction performance of PRS-LR with the same experiment procedure as the first experiment, and measure the decrease of R^2 if removing each feature as the predic-

tive feature importance. The results are shown in Table 4.3. We can see that in both two signal-noise-ratio settings, the predictive feature importance is consistent with our model setting. When deleting a noise feature, the R^2 remains nearly unchanged and is sometimes even slightly positive, which means the R^2 increases if we remove a feature, because removing a noisy feature reduces the model complexity. When removing a strong feature, the prediction R^2 largely decreases. Similarly, if we remove the periodicity by using RS-LR instead of PRS-LR, the prediction R^2 also decreases.

feature removed	noise feature	PRS-LR-I	PRS-LR-II	RS-LR
$X_{\cdot,1}$	Yes	0.11	0.12	0.14
$X_{\cdot,2}$	Yes	0.29	0.29	0.12
$X_{\cdot,3}$	No	<1e-16	<1e-16	<1e-16
$X_{\cdot,4}$	No	<1e-16	<1e-16	<1e-16
$X_{\cdot,5}$	No	<1e-16	<1e-16	<1e-16
periodicity	No	6.8e-07	7.1e-07	-

Table 4.4: The simulation results of PRS-LR and RS-LR's LRT feature importance of each feature and periodicity. We could see that the p-values of noisy features are insignificant. The p-values of strong features and periodicity are significant with p-values all close to 0. Significant features are highlighted.

The fourth experiment has the same setting as the second experiment except that we generate one sequence of $T = 1000$ and $\sigma = 1.0$, and we measure the LRT feature importance on it. The results are shown in Table 4.4. We can see the p-value approach is very sensitive and specific, that if we test on a noisy feature, the p-value is insignificant, while if we test on a strong feature or the periodicity, the p-value is very significant and close to 0.

From these 4 experiments, we can see that first PRS-LR can largely improve the performance if the underlying true model is PRS-LR, whether under the EDM framework or not. Also we can see both the LRT and predictive feature importance approaches for PRS-LR can select the strong features and periodicity and rule out noisy features.

4.4.2 Application on chlorophyll forecast

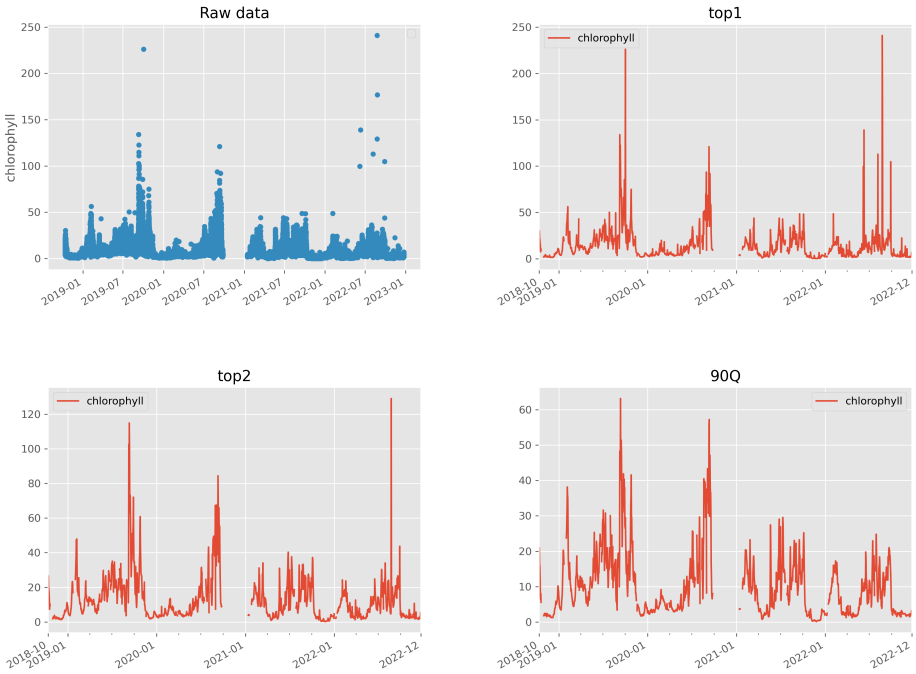

Figure 4.1: The original chlorophyll process to forecast along with different extreme value calculations (top-1, top-2, 90-Q).

In this application, our target is to forecast the chlorophyll. We will forecast the chlorophyll, especially its peak values, and evaluate the prediction accuracy of different methods, with or without cross-predictors to demonstrate the performance of PRS-LR and RS-LR. Also we will show the LRT feature importance because it is important in this scientific setting.

The dataset is high-resolution and has a record every 15 minutes from 2018 to 2022, except some gap periods without any records. The prediction target is the maximal chlorophyll level every day. Since the prediction target is day-level and every day we have approximately 100 records, we aggregated the dataset by day first. We used the maximal chlorophyll amount as the prediction target and used the following cross

predictors: max salinity (the maximal salinity), min salinity (the minimal salinity), min ODO (the minimal optical dissolved oxygen), turbidity (the average turbidity), pH (the average pH value), temperature (the average temperature) and wind (the average wind speed). After aggregation, each day has a set of cross predictors and a target observation except those gap days.

Here the maximum or minimum values for the response and cross predictors are calculated in 3 different ways: top-1 (the largest/smallest), top-2 (the second largest / second smallest) and 90-Q (the 90%-quantile / 10%-quantile) on all records each day, because the data has outliers. The choice of how to calculate valid extreme values is a scientific question as our prediction target is the heavy tails. Here we are not to address which method is better, but to show how RS-LR and PRS-LR perform under different settings by treating each of them as a separate dataset. We want to point out that whatever calculation we use, especially those methods removing values, we need solid scientific reasons. For example, if we remove the largest value, then we need a solid reason why this largest value is abnormal. Later we could see even top-1 and top-2 have different dynamical patterns and statistically speaking, top-1 has the heaviest tail and 90-Q has the least heavy tail. The data are shown in Fig 4.1, and the histogram and the zoomed-in process over 2022 are shown in Appendix A.

The forecasting target is the chlorophyll amount, which is non-negative and heavy-tailed. It is a common practice to take the logarithm transformation of the maximum chlorophyll as the prediction target. As our prediction focus is the heavy tails, we evaluate the prediction performance on the untransformed chlorophyll which is more sensitive to heavy tails than the log-transformed chlorophyll. So we transform the chlorophyll by logarithm before fitting and after predicting, and then we transform both the response and predicted values back to the original scale and evaluate the

with X	horizon	persistent	AR	S-map	PRS-LR-I	PRS-LR-II	RS-LR
top-1	1-day	0.15	0.31	0.31	0.32	0.33	0.30
	3-day	-0.19	0.23	0.24	0.25	0.25	0.25
	7-day	-0.48	0.13	0.14	0.14	0.15	0.12
top-2	1-day	0.62	0.68	0.69	0.66	0.66	0.68
	3-day	0.35	0.46	0.46	0.46	0.45	0.47
	7-day	-0.05	0.27	0.28	0.29	0.29	0.24
90-Q	1-day	0.89	0.89	0.89	0.89	0.89	0.89
	3-day	0.61	0.65	0.66	0.66	0.66	0.65
	7-day	0.21	0.33	0.34	0.20	0.30	0.36
without X	horizon	persistent	AR	S-map	PRS-LR-I	PRS-LR-II	RS-LR
top-1	1-day	0.15	0.30	0.30	0.33	0.33	0.30
	3-day	-0.19	0.21	0.22	0.23	0.23	0.22
	7-day	-0.48	0.10	0.11	0.10	0.10	0.11
top-2	1-day	0.62	0.68	0.67	0.69	0.69	0.68
	3-day	0.35	0.47	0.46	0.47	0.47	0.46
	7-day	-0.05	0.27	0.27	0.24	0.24	0.25
90-Q	1-day	0.89	0.89	0.88	0.88	0.88	0.89
	3-day	0.61	0.65	0.64	0.60	0.60	0.61
	7-day	0.21	0.19	0.20	0.10	0.13	0.19

Table 4.5: Prediction accuracy for chlorophyll (top-1, top-2 and 90-Q) of 1-day, 3-day and 7-day forward with/without cross predictors, measured by R^2. For each scenario, the optimal and sub-optimal method with highest and the second highest R^2 are highlighted. We could see PRS-LR and RS-LR outperforms time-invariant methods in most cases.

prediction accuracy by R^2. We predict under the EDM framework. We forecast 3 different horizons: 1-day forward, 3-day forward and 7-day forward. We used lag-2 for chlorophyll and lag-1 for cross predictors, because the lag-2 could model a second order PDE and the lag-1 could model a first order PDE. Mathematically, denote the cross predictors as U_t, the chlorophyll process as V_t, then we set $X_t = [U_t, \log(V_t), \log(V_{t-1})]$ and $Y_t = [\log(V_{t+h})]$. After prediction, we have the predicted $\widehat{Y}_t = \log(\widehat{V_{t+h}})$. Then we evaluate (1) the predictions performance and (2) LRT feature selection based on $\{\exp(\widehat{Y}_t)\}_t$ and $\{\exp(Y_t)\}_t$ for t in the testing period. For each test day t to predict, we assume that we could observe all pairs of observations before $t-h+1$, $\{(X_u, Y_u)\}_{u<t-h+1}$, and X_t. We train the model on all observed pairs and use h-step forward prediction. This setting is restricted in the EDM framework and the same as real-world application since there is no information leakage and all X_u and Y_u are observable at time t.

We set the testing period in the following day to accommodate the missing data during gap periods. There is no record in those gap days so no cross predictors and target observations. The missing value imputation should be in line with scientific facts and it is out of the scope of this project, so we use the following way to handle it. First, we break the processes into sub-processes by gap days, so that all sub-processes only contain consecutive days of cross predictors and target observations without missing values. We removed sub-processes shorter than 22 days. Second, we choose a warm-up length and a number of warm-up sub-processes, that guarantees the sufficient training of all methods. Note that a larger number of warm-up length means a delay in prediction if we have some gap days in practice, and a larger number of warm-up sub-processes means a smaller testing set, so we choose them as small as possible. For this dataset, we choose to use 3 sub-processes for warm-up, and for

remaining sub-processes we let its first 7 days for warm-up. The testing period is approximately from June 2019 to December 2022.

To demonstrate the performance of RS-LR and PRS-LR, we compare them with the benchmark time-invariant EDM methods including persistent, AR and S-map (Sugihara, 1994). Persistent forecasting, also called naive forecasting, is to use the latest target observation as the prediction of the future, regardless of the forecast horizon. AR uses the lagged cross-predictors and target observations and we fitted it by linear regression. S-map is similar to AR but with an exponential local kernel. Here we briefly introduced S-map. Suppose that we have a training set $\{(X_t, Y_t)\}_t$ and want to predict Y^* corresponding to a query point X^*, we assign a weight to each data point in the training set that the weight for X_t is $\exp(-\theta^{(smap)}||X_t - X^*||_2 \big/ \overline{d})$, where $||\cdot||_2$ is the Euclidean distance, $\overline{d}$ is the average of $||X_t - X^*||_2$ among all training samples and $\theta^{(smap)}$ is a hyperparameter to control the locality. If $\theta^{(smap)} = 0$, then S-map reduces to AR. We set $\theta^{(smap)} = 0.5$ in our experiment, and normalized $\{X_t\}_t$ to remove the mean and variance.

The results of the prediction accuracy are shown in Table 4.5. We can see that the prediction performance with cross predictors is better than without cross predictors. And also we can see that our methods perform well, especially for 1-day and 3-day prediction. For 7-day prediction, our method is still very competitive. Here, since for 7-day prediction, we need to do a 7-step forward propagation, so there might be some decay and averaging effect. We need to take this into consideration, but not necessarily it will give us a bad prediction. For example, in the results with cross predictors, our methods are best for 7-day prediction for top-1 and top-2. In general, our method is a competitive candidate. In Section 4.5 we provide a guideline of when to choose our RS-EDM and PRS-EDM and when to choose S-map.

		top-1			top-2			90-Q		
Horizon	Predictor	PRS-LR-I	PRS-LR-II	RS-LR	PRS-LR-I	PRS-LR-II	RS-LR	PRS-LR-I	PRS-LR-II	RS-LR
1-day	max salinity	1.00	1.00	0.67	0.66	0.97	0.07	0.08	0.04	0.04
	min salinity	0.43	0.41	0.13	0.93	0.97	0.57	0.24	0.13	0.08
	min ODO	0.0009	0.0009	0.0021	1.8e-05	1.1e-05	1.6e-05	0.19	0.22	1.00
	turbidity	2.4e-75	3.6e-75	1.8e-46	2.2e-06	3.6e-06	3.3e-04	1.00	1.00	1.00
	pH	0.70	0.70	0.76	0.60	0.56	0.68	1.00	1.00	1.00
	temperature	5.1e-06	5.3e-06	5.7e-05	2.2e-05	2.0e-05	6.3e-06	0.24	0.25	0.31
	wind	0.0047	0.0049	0.0292	0.0072	0.0074	0.0095	0.015	0.015	0.002
3-day	max salinity	1.00	1.00	1.00	0.15	0.10	0.07	3.8e-07	6.6e-06	1.3e-04
	min salinity	1.00	1.00	1.00	0.35	0.25	0.14	0.001	0.009	0.042
	min ODO	4.1e-14	5.1e-14	2.2e-08	1.3e-07	2.8e-07	3.4e-05	0.07	0.15	0.02
	turbidity	8.1e-20	5.9e-20	2.5e-15	1.2e-12	1.7e-12	5.3e-12	0.04	0.03	0.02
	pH	0.017837	0.031484	0.001226	0.33	0.28	0.24	0.59	1.00	1.00
	temperature	2.62e-12	5.38e-13	2.75e-14	1.8e-06	2.0e-06	1.5e-06	0.32	0.45	0.19
	wind	0.035246	0.043675	0.031598	0.015	0.015	0.015	0.009	0.018	0.005
7-day	max salinity	0.025	0.025	0.028	0.32	0.31	0.33	3.2e-08	1.5e-10	5.1e-07
	min salinity	0.46	0.44	0.51	0.80	0.79	0.76	0.0004	4.2e-70	0.0005
	min ODO	8.0e-06	7.8e-06	3.4e-06	1.1e-04	8.0e-05	3.9e-05	0.0026	0.0002	0.0103
	turbidity	1.2e-04	9.2e-05	1.9e-04	2.8e-05	2.9e-05	4.4e-05	0.001	1.3e-08	0.004
	pH	0.0033	0.0029	0.0031	0.011	0.009	0.008	0.06	0.0006	0.27
	temperature	4.8e-07	3.0e-06	1.1e-09	9.6e-06	5.8e-06	4.1e-07	0.08	0.19	0.09
	wind	0.23	0.24	0.24	0.41	0.43	0.44	0.22	7.5e-29	0.28

Table 4.6: LRT feature importance for chlorophyll prediction for different maximization and forecast horizons. All p-values smaller than 0.05 are highlighted. We could see PRS-LR-I, PRS-LR-II and RS-LR provide inconsistent feature selection results.

	horizon	method	p-value for PRS
top-1	1-day	PRS-LR-I	1.0
		PRS-LR-II	1.0
	3-day	PRS-LR-I	2.0e-06
		PRS-LR-II	2.6e-06
	7-day	PRS-LR-I	0.014
		PRS-LR-II	0.012
top-2	1-day	PRS-LR-I	2.5e-09
		PRS-LR-II	3.2e-09
	3-day	PRS-LR-I	0.00013
		PRS-LR-II	0.00020
	7-day	PRS-LR-I	0.008
		PRS-LR-II	0.019
90-Q	1-day	PRS-LR-I	2.8e-17
		PRS-LR-II	1.4e-16
	3-day	PRS-LR-I	7.5e-7
		PRS-LR-II	1.5e-5
	7-day	PRS-LR-I	0.00019
		PRS-LR-II	0.0014

Table 4.7: LRT p-value for periodicity for chlorophyll prediction. All p-values smaller than 0.05 are highlighted. We could see periodicity is very significant, which further supports the proposed PRS models.

The results of LRT feature importance are shown in Table 4.6. These are the LRT feature importance, i.e. p-values, given by each of PRS-LR-I, PRS-LR-II and RS-LR. We highlight significant p-values smaller than 0.05. We can see that RS-LR or PRS-LR methods are providing similar results that which feature has significant impact for the same data set and the same forecasting horizon. For example, we can see that the cross predictor temperature is very significant if we want to forecast 1-day forward chlorophyll, so it means that temperature has a strong impact on the next day's chlorophyll. Also we can see the significant features for different forecast horizons are different. For example, for top-1, the pH is insignificant for 1-day, but significant for 3-day and 7-day. The wind is significant for 1-day and 3-day, but insignificant for 7-day. That means the pH value has a long-term and delayed impact on chlorophyll, and the

wind has a short-term but transient impact. They have different response behavior in the time domain, and our LRT feature importance can tell this information and help us better understand the underlying scientific process. For the same forecast horizon, across different maximizations, the significant features are different, that top-1 and top-2 are similar but 90-Q is very different. It is indicating that different maximizations will provide different data sets following different dynamical systems. In Fig 4.1, we can see that they differ a lot in those peaks. From LRT feature importance, we can see these peaks are not necessarily incorrect records, because they have consistent patterns but different from 90-Q. One possible reason is that the 90-Q is more sensitive to to the slower evolving biological processes with timescales of days, and top-1 and top-2 might be capturing short-term and transient disturbance of physical processes such as wind, which is similar to Hawkes processes (Hawkes, 1971). The differences between these two categories could be a further direction to study. In conclusion, our model can tell which predictors are significantly impacting the time series, while S-map cannot provide this information.

Table 4.7 shows the LRT feature importance for periodicity. In general it is significant with a small p-value. That's also supporting why we need to propose this PRS modeling, because for regime-switching models, the models with periodicity are significantly different from those without periodicity. This periodicity cannot be detected and explicitly modeled by S-map or other time-invariant EDM methods.

In summary, PRS-EDM works well for chlorophyll modeling in both the prediction aspect and feature importance aspect.

4.5 Discussion

In this project, we have 2 folds of contributions. One fold is PRS models, which have the capability to represent the partially periodic partially random state transition. It can be applied on any RS models such as HMM to substitute the latent Markov chain. The other fold is RS-EDM and PRS-EDM, which extend the standard EDM by relaxing its time-invariance assumption.

The PRS model we proposed has a wide application scenario. PRS can substitute any Markovian RS process, including HMM and Markov-modulated linear regression. We could find that in a lot of RS cases, we are to use the Markov process as an approximation of potential periodic processes. For example, in the classic toy example for illustrating HMM which uses HMM to model weather (see e.g. Nguyen, 2017; Khiatani and Ghose, 2017), the weather has a strong yearly periodic phenomenon. If using HMM only, we will ignore this periodicity and model it by a latent Markov chain. In this case, we could use the PRS version of HMM instead of standard HMM. This periodicity cannot be modeled by deterministic harmonic terms, but a partially periodic and partially random process. Similarly in finance, there are PRS phenomena, such as the energy-related financial products. The PRS model has a much wider application.

The motivation behind proposing RS-EDM and PRS-EDM is from analogy. EDM suggests using the AR type prediction function. So it is natural to make an analogy between EDM and AR, and to extend EDM by combining it with the RS structure, because AR has the RS variant.

The EDM framework is widely used in scientific settings, and along with the prediction, it is important to understand which features have scientific impacts or not.

This problem is solved by the feature selection or feature importance approaches. We proposed two approaches to select features or evaluate feature importance, the LRT approach and predictive approach. Here we briefly compare these two approaches. The LRT approach has a better p-value interpretation. It is more rigorous and sensitive because it is an LRT. It doesn't suffer from the instationary changing of dynamical systems. It has a low computational cost because it only needs to fit the data once. The predictive approach is better for forward looking. It is suffering from instationary changing of dynamical systems as it is out-of-sample. It doesn't have a nice p-value interpretation and it is hard to interpret which level of R^2 decreases is large or small. But it has a nice prediction interpretation of how much we could gain for prediction if this feature is involved. This quantitative interpretation cannot be achieved by LRT feature importance which can only provide a binary decision, in-model or out-of-model, because it is essentially a p-value. One important scenario that LRT and predictive feature importance will give different results is when the underlying dynamical is non-stationary and changing and some feature has very different impact during different periods. In these cases the predictive feature importance will rule this feature out for prediction, which doesn't mean this feature is not involved in the dynamical system, while LRT will include this feature which doesn't mean it could help prediction. In short, if we consider more about prediction performance, then we should use the predictive approach; if we care more about scientific interpretation, we should use the LRT approach. The comparison between the LRT and predictive feature importance is summarized in Table 4.8.

If we want to use EDM, whether to choose time-varying or time-invariant EDM depends on the sample size and the underlying assumption we believe. If the sample size is large enough and we can endure a long computational time, then we can always

Feature importance	LRT	Predictive
Purpose	Scientific understanding	Prediction
Output	Whether any impact	How much impact
Sensitivity	Sensitive	Insensitive
To nonstationarity	Robust	Vulnerable
Computational cost	Low	High
Interpretability	Good (p-value)	Limited (for prediction)
In/out-of sample	In-sample	Out-of-sample
Theory	Rigorous theoretically	Practical and heuristic

Table 4.8: Summary of comparison between the LRT and predictive feature importance.

choose the PRS-EDM because time-invariant EDM is a special case of RS-EDM, which further is a special case of PRS-EDM. But in most scenarios using EDM, the sample size is not sufficient, since for scientific processes, the collection of data is costly and limited in frequency, time span or both. In these cases, the model complexity needs to be chosen carefully. If we believe there exists an RS phenomenon that a single prediction function cannot model the whole process, we should use RS-EDM. But note that the degrees of freedom of RS-EDM is at least S times of the time-invariant EDM, since each state of RS-EDM is a single EDM, and we need extra degrees of freedom to model the latent process. If we need to use RS-EDM, we could consider using PRS-EDM, because PRS-EDM only has $2S(S-1)$ more degrees of freedom than RS-EDM. So when the latent state process has periodicity, PRS-EDM works better than RS-EDM; even if there is no periodicity, PRS-EDM should not work too worse than RS-EDM.

We also want to compare for time-invariant EDM and RS/PRS-EDM and provide a guideline of which method to use in practice. First, if we have limited data, we should use time-invariant EDM with simple parametric form such as linear regression. If we have more data, we can choose between the time-dependent EDM and the

time-invariant EDM. S-map is the state-of-the-art method under the time-invariant EDM framework. It is a kernel regression method that can be either global or local depending on the hyperparameter we choose. If we use a global S-map, such as a linear regression, we might lose the power because the model is too simple. If we use a local S-map, we need a large data size to fit it as it is non-parametric, and the more local the prediction function, the larger training size needed. But the data size is limited in general. That's why we propose RS-EDM and PRS-EDM. Both RS-EDM and PRS-EDM allows us to use a simple model such as linear regression, because we no longer use the same time-invariant prediction function for all time stamps and we have the flexibility that different time periods could have different prediction functions. Whether to use local S-map or RS/PRS-EDM depends on which assumption we want to break. If we want to break the time-invariant assumption, we can use RS-EDM or PRS-EDM, otherwise we can use local S-map.

There are several future directions for this project. One direction is to accommodate the cases where the data has a bandwidth of periods instead of a single or several discrete periods. The PRS modeling studied in this project relies on the assumption of a small set of known periods which is from the expertise knowledge. But in reality, we might get the periods from Fourier analysis when the expertise knowledge is not available, and we will have a continuous spectrum which can hardly be modeled by the current PRS modeling. One possible solution is to involve a big set of periods and at the same time add regularization, including the penalty for the scale of impact from each period and the penalty of the roughness of these scales for neighboring periods. Another direction is for the application on chlorophyll forecast. In Section 4.4, we found that three different maximizations provide different outputs and have different patterns. So we can think the process could be a summation of a slowly

evolving biological process and a transient disturbance process. We can further study how to filter out one process and focus on the other.

Chapter 5

Discussion

In this dissertation, we studied the RS phenomenon on dynamical systems. To the best of our knowledge, there is no research on this topic. This topic has its own difficulties as we are modeling the change of how the process is changing, and this two-folded changing structure makes the inference much harder. In Project I and II, the MLE doesn't work on the second-order changing so we proposed the heteroskedasticity-based E-M algorithm to infer the model parameters, and in Project III, we used EDM to avoid this problem.

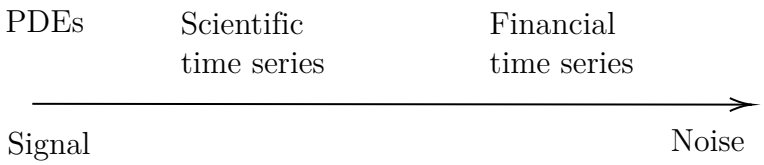

Figure 5.1: Illustrative figure of signal-noise-ratio spectrum and where financial and scientific processes are locate.

Though both financial and scientific processes are modeled by SDE, they are fundamentally different and lie on two ends of the signal-noise-ratio spectrum as illustrated in Fig 5.1. Financial process is dominated by randomness and models can only partially capture its properties, so the model in general has a simple parametric form because it is robust against noise, relatively easy to estimate, and even though the simple parametric form cannot capture all information, the bias is covered by the random noise. For example, for modeling the volatility, all the Heston model, GARCH

model and 3/2 model have a relatively simple parametric form and a considerable random component. That's why we propose the heteroskedasticity-based E-M algorithm where different simple parametric forms capture different aspects of a process. In this model, we determine whether a model is sufficient or not by heteroskedasticity test, which can be regarded as a comparison between the systematic bias and random noise. This test works best under a suitable signal-noise-ratio because if the noise is too weak, all models cannot pass the heteroskedasticity test and the emission probabilities will become indistinguishable.

In contrast, scientific processes are in general more deterministic but more complicated and involve a lot of variables. In practice it could be too complicated to infer the exact PDEs or SDEs, so its forecasting problem is solved by the end-to-end models which ignore the exact underlying system, such as EDM which does not depend on the specific parametric form of underlying PDEs or SDEs. If we assign a simple parametric form for EDM such as linear regression, it will lose the capability, unlike the low signal-noise-ratio financial process. If we assign a local non-parametric form such as the S-map, it requires a high quality data set to fit whose signal-noise-ratio should be high and data size should be large. RS-EDM and PRS-EDM provide another modeling angle, in that we use a simple but time-varying parametric form. Again, we want to emphasize that the methods under RS-EDM and PRS-EDM discussed here are not restricted to EDM. It can be extended to any regression problems with temporal information. Also the PRS model can be applied to any model involving a latent state process.

In these two ends of the signal-noise-ratio spectrum, we can see why we need RS. RS is a modeling technique which approximates a complicated system by assuming that different simple patterns will dominate different time periods, even if the underlying

true system is not dynamically changing. This approximation view largely broadens the application scenario of RS models, because we are not necessarily targeting on fully modeling the process by RS models but to capture important patterns. The famous aphorism by George Box perfectly describes the usage of RS modeling: “All models are wrong, some are useful. ”

Bibliography

Adejumo, Oluwasegun A, Seno Albert, and Omorogbe J Asemota (2020). "Markov regime-switching autoregressive model of stock market returns in Nigeria". In: *CBN Journal of Applied Statistics* 11.2, pp. 65–83.

Ailliot, Pierre and Valérie Monbet (2012). "Markov-switching autoregressive models for wind time series". In: *Environmental Modelling & Software* 30, pp. 92–101.

Aït-Sahalia, Yacine and Robert Kimmel (2007). "Maximum likelihood estimation of stochastic volatility models". In: *Journal of financial economics* 83.2, pp. 413–452.

Alfeus, Mesias, Ludger Overbeck, and Erik Schlögl (2019). "Regime switching rough Heston model". In: *Journal of Futures Markets* 39.5, pp. 538–552.

Aliat, Billel and Fayçal Hamdi (2018). "On Markov-switching periodic ARMA models". In: *Communications in Statistics-Theory and Methods* 47.2, pp. 344–364.

Andersen, Torben G and Tim Bollerslev (1998). "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts". In: *International economic review*, pp. 885–905.

Andersen, Torben G, Tim Bollerslev, and Francis X Diebold (2007). "Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility". In: *The review of economics and statistics* 89.4, pp. 701–720.

Andersen, Torben G, Tim Bollerslev, and Xin Huang (2011). "A reduced form framework for modeling volatility of speculative prices based on realized variation measures". In: *Journal of Econometrics* 160.1, pp. 176–189.

Andronov, Alexander M and Nadezda Spiridovska (2019). "Markov-modulated linear regression". In: *arXiv preprint arXiv:1901.09600*.

Ang, Andrew and Allan Timmermann (2012). "Regime changes and financial markets". In: *Annu. Rev. Financ. Econ.* 4.1, pp. 313–337.

Atiya, Amir F and Steve Wall (2009). "An analytic approximation of the likelihood function for the Heston model volatility estimation problem". In: *Quantitative Finance* 9.3, pp. 289–296.

Bailey, Barbara A, Scott C Doney, and Ivan D Lima (2004). "Quantifying the effects of dynamical noise on the predictability of a simple ecosystem model". In: *Environmetrics: The official journal of the International Environmetrics Society* 15.4, pp. 337–355.

Bandi, Federico M and Jeffrey R Russell (2008). "Microstructure noise, realized variance, and optimal sampling". In: *The Review of Economic Studies* 75.2, pp. 339–369.

Barndorff-Nielsen, Ole E et al. (2009). *Realized kernels in practice: Trades and quotes*.

Bates, David S (1996). "20 testing option pricing models". In: *Handbook of statistics* 14, pp. 567–611.

Baum, Leonard E and John Alonzo Eagon (1967). "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology". In: *Bulletin of the American Mathematical Society* 73.3, pp. 360–363.

Baum, Leonard E and Ted Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains". In: *The annals of mathematical statistics* 37.6, pp. 1554–1563.

Baum, Leonard E, Ted Petrie, et al. (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *The annals of mathematical statistics* 41.1, pp. 164–171.

Bengio, Yoshua and Paolo Frasconi (1994). "An input output HMM architecture". In: *Advances in neural information processing systems* 7.

Black, Fischer and Myron Scholes (1973). "The pricing of options and corporate liabilities". In: *Journal of political economy* 81.3, pp. 637–654.

Bollerslev, Tim (1986a). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327.

— (1986b). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327.

Bollerslev, Tim, Andrew J Patton, and Rogier Quaedvlieg (2016). "Exploiting the errors: A simple approach for improved volatility forecasting". In: *Journal of Econometrics* 192.1, pp. 1–18.

Boudt, Kris, Jonathan Cornelissen, and Scott Payseur (2013). "Highfrequency: Toolkit for the analysis of highfrequency financial data in R." In: *Recuperado em* 19, pp. 1–23.

Box, George EP et al. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45, pp. 5–32.

Brownlees, Christian T and Giampiero M Gallo (2006). "Financial econometric analysis at ultra-high frequency: Data handling concerns". In: *Computational statistics & data analysis* 51.4, pp. 2232–2245.

Carr, Peter and D Madan (2005). "FAQ's in option pricing theory". In: *Journal of Derivatives, forthcoming*.

Cenci, Simone, George Sugihara, and Serguei Saavedra (2019). "Regularized S-map for inference and forecasting with noisy ecological time series". In: *Methods in Ecology and Evolution* 10.5, pp. 650–660.

Chen, Fang, Keying Ye, and Min Wang (2021). "The minimum Bayes factor hypothesis test for correlations and partial correlations". In: *Communications in Statistics-Theory and Methods* 50.11, pp. 2467–2480.

Chorus, I and J Bartram (1999). "Toxic cyanobacteria in water. A guide to their public consequences, monitoring and management". In.

Christoffersen, Peter, Steven Heston, and Kris Jacobs (2009). "The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well". In: *Management Science* 55.12, pp. 1914–1932.

Christoffersen, Peter, Kris Jacobs, and Karim Mimouni (2010). "Volatility dynamics for the S&P500: Evidence from realized volatility, daily returns, and option prices". In: *The Review of Financial Studies* 23.8, pp. 3141–3189.

Ciaccio, Christina E et al. (2015). "Home dust microbiota is disordered in homes of low-income asthmatic children". In: *Journal of Asthma* 52.9, pp. 873–880.

Clements, Adam and Daniel PA Preve (2021). "A practical guide to harnessing the har volatility model". In: *Journal of Banking & Finance* 133, p. 106285.

Corsi, Fulvio (2009). "A simple approximate long-memory model of realized volatility". In: *Journal of Financial Econometrics* 7.2, pp. 174–196.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20, pp. 273–297.

Costabile, Massimo et al. (2014). "Option pricing under regime-switching jump–diffusion models". In: *Journal of Computational and Applied Mathematics* 256, pp. 152–167.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Deschamps, Philippe J (2008). "Comparing smooth transition and Markov switching autoregressive models of US unemployment". In: *Journal of Applied Econometrics* 23.4, pp. 435–462.

Dietze, Michael C et al. (2018). "Iterative near-term ecological forecasting: Needs, opportunities, and challenges". In: *Proceedings of the National Academy of Sciences* 115.7, pp. 1424–1432.

Drucker, Harris et al. (1996). "Support vector regression machines". In: *Advances in neural information processing systems* 9.

Dudek, Anna E, Harry Hurd, and Wioletta Wójtowicz (2016). "Periodic autoregressive moving average methods based on Fourier representation of periodic coefficients". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 8.3, pp. 130–149.

Dupire, Bruno et al. (1994). "Pricing with a smile". In: *Risk* 7.1, pp. 18–20.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the econometric society*, pp. 987–1007.

Fennel, Katja et al. (2022). "Ocean biogeochemical modelling". In: *Nature Reviews Methods Primers* 2.1, pp. 1–21.

Filardo, Andrew J (1994). "Business-cycle phases and their transitional dynamics". In: *Journal of Business & Economic Statistics* 12.3, pp. 299–308.

Fouque, J-P and Yuri F Saporito (2018). "Heston stochastic vol-of-vol model for joint calibration of VIX and S&P 500 options". In: *Quantitative Finance* 18.6, pp. 1003–1016.

Gatheral, Jim (2004). "A parsimonious arbitrage-free implied volatility parameterization with application to the valuation of volatility derivatives". In: *Presentation at Global Derivatives & Risk Management, Madrid*.

Gatheral, Jim, Thibault Jaisson, and Mathieu Rosenbaum (2018). "Volatility is rough". In: *Quantitative finance* 18.6, pp. 933–949.

Glover, David M, William J Jenkins, and Scott C Doney (2011). *Modeling methods for marine science*. Cambridge University Press.

Goard, Joanna and Mathew Mazur (2013a). "Stochastic volatility models and the pricing of VIX options". In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 23.3, pp. 439–458.

— (2013b). "Stochastic volatility models and the pricing of VIX options". In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 23.3, pp. 439–458.

Goutte, Stéphane, Amine Ismail, and Huyên Pham (2017). "Regime-switching stochastic volatility model: estimation and calibration to VIX options". In: *Applied Mathematical Finance* 24.1, pp. 38–75.

Grasselli, Martino (2017). "The 4/2 stochastic volatility model: A unified approach for the Heston and the 3/2 model". In: *Mathematical Finance* 27.4, pp. 1013–1034.

Hamilton, James D (1989). "A new approach to the economic analysis of nonstationary time series and the business cycle". In: *Econometrica: Journal of the econometric society*, pp. 357–384.

Hansen, Peter R and Asger Lunde (2006). "Realized variance and market microstructure noise". In: *Journal of Business & Economic Statistics* 24.2, pp. 127–161.

Hawkes, Alan G (1971). "Spectra of some self-exciting and mutually exciting point processes". In: *Biometrika* 58.1, pp. 83–90.

He, Xin-Jiang and Song-Ping Zhu (2017). "How should a local regime-switching model be calibrated?" In: *Journal of Economic Dynamics and Control* 78, pp. 149–163.

Heston, Steven L (1993). “A closed-form solution for options with stochastic volatility with applications to bond and currency options”. In: *The review of financial studies* 6.2, pp. 327–343.

— (1997). “A simple new formula for options with stochastic volatility”. In.

Heston, Steven L and Saikat Nandi (2000). “A closed-form GARCH option valuation model”. In: *The review of financial studies* 13.3, pp. 585–625.

Ho, Jeff C and Anna M Michalak (2020). “Exploring temperature and precipitation impacts on harmful algal blooms across continental US lakes”. In: *Limnology and Oceanography* 65.5, pp. 992–1009.

Ho, Jeff C, Anna M Michalak, and Nima Pahlevan (2019). “Widespread global increase in intense lake phytoplankton blooms since the 1980s”. In: *Nature* 574.7780, pp. 667–670.

Hyndman, Rob J and George Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts.

Ji, Rubao et al. (2006). “Spring phytoplankton bloom and associated lower trophic level food web dynamics on Georges Bank: 1-D and 2-D model studies”. In: *Deep Sea Research Part II: Topical Studies in Oceanography* 53.23-24, pp. 2656–2683.

Kendall, Maurice G (1938). “A new measure of rank correlation”. In: *Biometrika* 30.1/2, pp. 81–93.

Keogh, Eamonn J and Abdullah Mueen (2017). “Curse of dimensionality.” In: *Encyclopedia of machine learning and data mining* 2017, pp. 314–315.

Khiatani, Diksha and Udayan Ghose (2017). “Weather forecasting using hidden Markov model”. In: *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*. IEEE, pp. 220–225.

Kou, Steven G (2002). “A jump-diffusion model for option pricing”. In: *Management science* 48.8, pp. 1086–1101.

Kudela, Raphael, Elisa Berdalet, and Ed Urban (2015). “Harmful algal blooms: A scientific summary for policy makers”. In.

Liang, Feng et al. (2008). “Mixtures of g priors for Bayesian variable selection”. In: *Journal of the American Statistical Association* 103.481, pp. 410–423.

Merton, Robert C (1976). “Option pricing when underlying stock returns are discontinuous”. In: *Journal of financial economics* 3.1-2, pp. 125–144.

Mitra, Sovan and Paresh Date (2010). “Regime switching volatility calibration by the Baum–Welch method”. In: *Journal of computational and applied mathematics* 234.12, pp. 3243–3260.

Møller, Jan Kloppenborg, Henrik Madsen, and Jacob Carstensen (2011). “Parameter estimation in a simple stochastic differential equation for phytoplankton modelling”. In: *Ecological modelling* 222.11, pp. 1793–1799.

Mu, Guo-Hua and Wei-Xing Zhou (2008). “Relaxation dynamics of aftershocks after large volatility shocks in the SSEC index”. In: *Physica A: Statistical Mechanics and its Applications* 387.21, pp. 5211–5218.

Ndlela, Luyanda L et al. (2016). “An overview of cyanobacterial bloom occurrences and research in Africa over the last decade”. In: *Harmful Algae* 60, pp. 11–26.

Nguyen, Loc (2017). “Tutorial on hidden markov model”. In: *Applied and Computational Mathematics* 6.4-1, pp. 16–38.

Noureldin, Diaa, Neil Shephard, and Kevin Sheppard (2012). “Multivariate high-frequency-based volatility (HEAVY) models”. In: *Journal of Applied Econometrics* 27.6, pp. 907–933.

Orlando, Giuseppe and Giovanni Taglialatela (2017). “A review on implied volatility calculation”. In: *Journal of Computational and Applied Mathematics* 320, pp. 202–220.

Pace, Michael L et al. (2017). "Reversal of a cyanobacterial bloom in response to early warnings". In: *Proceedings of the National Academy of Sciences* 114.2, pp. 352–357.

Paerl, Hans W, Wayne S Gardner, et al. (2016). "Mitigating cyanobacterial harmful algal blooms in aquatic ecosystems impacted by climate change and anthropogenic nutrients". In: *Harmful Algae* 54, pp. 213–222.

Paerl, Hans W and Jef Huisman (2008). "Blooms like it hot". In: *Science* 320.5872, pp. 57–58.

— (2009). "Climate change: a catalyst for global expansion of harmful cyanobacterial blooms". In: *Environmental microbiology reports* 1.1, pp. 27–37.

Paerl, Hans W and Valerie J Paul (2012). "Climate change: links to global expansion of harmful cyanobacteria". In: *Water research* 46.5, pp. 1349–1363.

Papanicolaou, Andrew and Ronnie Sircar (2014). "A regime-switching Heston model for VIX and S&P 500 implied volatilities". In: *Quantitative Finance* 14.10, pp. 1811–1827.

Patton, Andrew J and Kevin Sheppard (2015). "Good volatility, bad volatility: Signed jumps and the persistence of volatility". In: *Review of Economics and Statistics* 97.3, pp. 683–697.

Pick, Frances R (2016). "Blooming algae: a Canadian perspective on the rise of toxic cyanobacteria". In: *Canadian Journal of Fisheries and Aquatic Sciences* 73.7, pp. 1149–1158.

Pinson, Pierre and Henrik Madsen (2012). "Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models". In: *Journal of forecasting* 31.4, pp. 281–313.

Platen, Eckhard (1998). *A non-linear stochastic volatility model*. Centre for Mathematics and Its Applications, Australian National University.

Ross, Paige G and Richard A Snyder (2020). "Ecological Monitoring Program at VIMS ESL-Annual Report 2018-2019". In.

Rossi, Alessandro and Giampiero M Gallo (2006). "Volatility estimation via hidden Markov models". In: *Journal of Empirical Finance* 13.2, pp. 203–230.

Shephard, Neil and Kevin Sheppard (2010). "Realising the future: forecasting with high-frequency-based volatility (HEAVY) models". In: *Journal of Applied Econometrics* 25.2, pp. 197–231.

Shirley, Kenneth E et al. (2010). "Hidden Markov Models for Alcoholism Treatment Trial Data". In: *The Annals of Applied Statistics* 4.1, p. 366.

Shreve, Steven E et al. (2004). *Stochastic calculus for finance II: Continuous-time models*. Vol. 11. Springer.

Spezia, Luigi, Roberta Paroli, and Petros Dellaportas (2004). "Periodic Markov switching autoregressive models for Bayesian analysis and forecasting of air pollution". In: *Statistical Modelling* 4.1, pp. 19–38.

Sugihara, George (1994). "Nonlinear forecasting for the classification of natural time series". In: *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* 348.1688, pp. 477–495.

Sugihara, George and Robert M May (1990). "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series". In: *Nature* 344.6268, pp. 734–741.

Tuaneh, Godwin Lebari, Isaac Didi Essi, and Ette Harrison Etuk (2021). "Markov-Switching Vector Autoregressive (MS-VAR) Modelling (Mean Adjusted): Application to Macroeconomic data". In: *Archives of Business Research* 9.10.

Wang, Xunxiao, Keshab Shrestha, and Qi Sun (2019). "Forecasting realised volatility: a Markov switching approach with time-varying transition probabilities". In: *Accounting & Finance* 59, pp. 1947–1975.

Whaley, Robert E (1993). "Derivatives on market volatility: Hedging tools long overdue". In: *The journal of Derivatives* 1.1, pp. 71–84.

— (2000). "The investor fear gauge". In: *The Journal of Portfolio Management* 26.3, pp. 12–17.

— (2009). "Understanding the VIX". In: *The Journal of Portfolio Management* 35.3, pp. 98–105.

Wharton Research Data Services (2023). *"WRDS"*. URL: wrds.wharton.upenn.edu (visited on 03/01/2023).

Wilkinson, Grace M et al. (2018). "Early warning signals precede cyanobacterial blooms in multiple whole-lake experiments". In: *Ecological Monographs* 88.2, pp. 188–203.

Yu, Peixuan et al. (2021). "Predicting coastal algal blooms with environmental factors by machine learning methods". In: *Ecological Indicators* 123, p. 107334.

Zellner, Arnold (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions". In: *Bayesian inference and decision techniques*.

Zhang, Lan, Per A Mykland, and Yacine Aït-Sahalia (2005). "A tale of two time scales: Determining integrated volatility with noisy high-frequency data". In: *Journal of the American Statistical Association* 100.472, pp. 1394–1411.

Appendices

Appendix A

More Visualizations for Project III

The histograms of the chlorophyll process are shown in Fig A.1. The zoomed-in process of 2022 is shown in Fig A.2

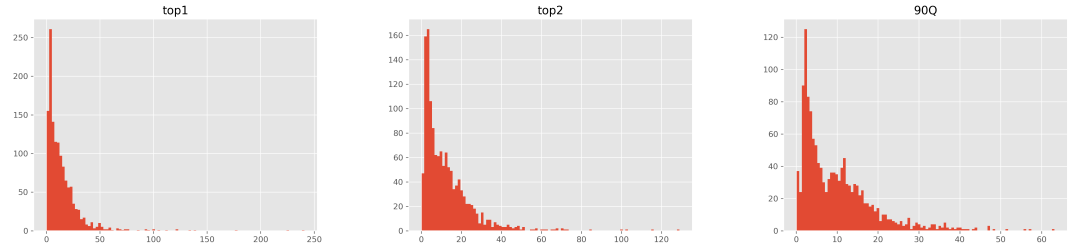

Figure A.1: Histograms of the chlorophyll process of top-1/top-2/90-Q.

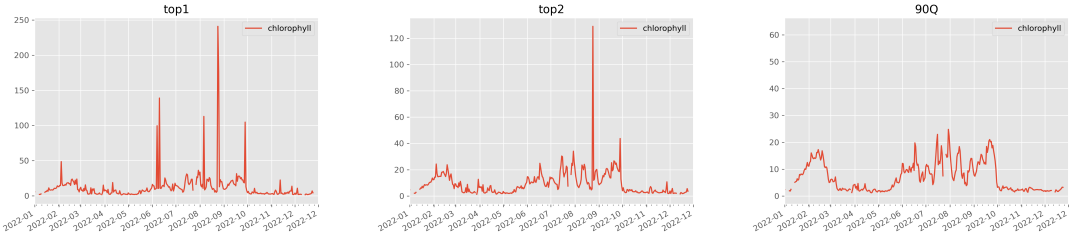

Figure A.2: The chlorophyll process during 2022 of top-1/top-2/90-Q.