Clustering-Based Cross-Sectional Regime Identification for Financial Market Forecasting

Rongbo Chen[1], Mingxuan Sun[1], Kunpeng Xu[1], Jean-Marc Patenaude[2], and Shengrui Wang[1](✉)

[1] University of Sherbrooke, Sherbrooke, QC, Canada
{rongbo.chen,mingxun.sun,kunpeng.Xu,shengrui.wang}@usherbrooke.ca
[2] Laplace Insights, Sherbrooke, QC, Canada
jeanmarc@laplaceinsights.com

Abstract. Regime switching analysis is extensively advocated in many fields to capture complex behaviors underlying an ecosystem, such as the economic or financial system. A regime can be defined as a specific group of complex patterns that share common characteristics in a specific time interval. Regime switch, caused by external and/or internal drivers, refers to the changing behaviors exhibited by a system at a specific time point. The existing regime detection methods suffer from two drawbacks: they lack the capability to identify new regimes dynamically or they ignore the cross-sectional dependencies exhibited by time series data for the forecasting. This promoted us to devise a cluster-based regime identification model that can identify cross-sectional regimes dynamically with a time-varying transition probability, and capture cross-sectional dependencies underlying financial time series for market forecasting. Our approach makes use of a nonlinear model to account for the cross-sectional regime dependencies, neglected by most existing studies, that can improve the performance of a forecasting model significantly. Experimental results on both synthetic and real-world dataset demonstrate that our model outperforms state-of-the-art forecasting models.

Keywords: Regime switch analysis · Cross-sectional regime identification · Financial market forecasting

1 Introduction

Financial markets may dynamically exhibit abrupt behavior changes. While some of these may be transitory, often the time-evolving behavior of market prices may persist over some specific time intervals [3,9]. For example, the mean, volatility, and correlation patterns in stock returns may change dramatically in a fluctuating stock market [6,20]. These sudden changes or structural break can be viewed as regime switches, some of which may be recurring and some of which may be permanent. Such switches are prevalent in the dynamic financial market. Regime switching analysis [12,13] is extensively advocated for its ability to

C. Strauss et al. (Eds.): DEXA 2022, LNCS 13427, pp. 3–16, 2022.
https://doi.org/10.1007/978-3-031-12426-6_1

capture these sudden changes or structural breaks in market behaviors hidden in financial data, making it a promising approach in financial analysis and market studies. Understanding the phenomenon of how new dynamics of prices and fundamentals persist for a certain length of time after a change helps explore financial behaviors for market forecasting.

Due to their ability to parsimoniously capture stylized behaviors of many financial series, including fat tails, persistently occurring periods of fluctuation followed by periods of low volatility, skewness, and time-varying correlations, regime switching models continue to gain in popularity [2,3,12]. In finance, a regime can be defined as a specific group of complex patterns that share the same characteristics in a specific time interval. Regime switching refers to the changing behaviors exhibited by time series transiting from one regime to another at a specific time points; such changes can be caused by external and/or internal drivers. Existing regime models are designed to predict the likelihood of a structural break resulting in a regime switch that is driven by a combination of driving variables, corresponding to pressures from within or outside the market [4,5,8]. Most, however, handle only single time series, and are not capable of dealing with multiple time series, which is more complex due to the statistically cross-sectional correlations underlying the multiple time series. Such is often the case with financial data, for example, market prices often have positive and negative correlations to one another, and stocks as broad asset classes have exhibited prolonged periods of negative correlation [11].

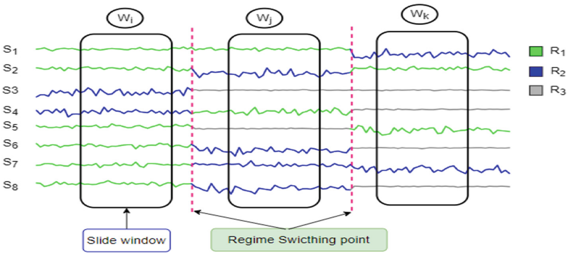

Fig. 1. An example of a multiple time series set consisting of 3 distinct regimes.

Though some regime models [14,23] have been proposed for modeling multiple time series, most of them suffer from issues such as having to specify the number of regimes and lacking the capability to identify regimes dynamically. This prevents them from achieving good performance, since regimes themselves are often thought of as approximations to underlying states that are unobserved and may arrive at any future time in the time-varying financial market, such as financial crisis. Moreover, time series may not synchronize with regime switching. For example, in Fig. 1, the two sub-sequences $\{S_3, S_4\}$ belong to regime R_2 and the others belong to regime R_1 in window W_i. However, most existing regime models [14,23] would identify the whole set of sub-sequences in window W_i (resp. W_j and W_k) as regime R_1 (resp. R_2 and R_3), and miss out on the regimes of sub-sequences $\{S_3, S_4\} \in R_2$ in window W_i, $\{S_1, S_4\} \in R_1$ and $\{S_3, S_5\} \in R_3$ in window W_j, and $\{S_1, S_7\} \in R_2$ and $\{S_2, S_5\} \in R_1$ in window W_k. Consequently, these models can not achieve satisfactory forecasting performance.

To address the issues above, we propose a financial market forecasting model utilizes clustering-based cross-sectional regime identification model. We proposed clustering-based regime identification model can identify cross-sectional regimes dynamically along with a forecasting process based on a time-varying transition probability matrix, to address the problem of specifying a fixed number of regimes and switching among in a fixed set of regimes with a static transition probability matrix. We then devise a non-linear model to capture cross-sectional regime dependencies (as presented in Fig. 1) on multiple time series for financial market forecasting. The significance of this work can be summarized as follows:

- We propose a clustering-based cross-sectional regime identification model on multiple time series, which allows the identification of multiple cross-sectional regimes dynamically, with a time-varying transition probability matrix, along with the forecasting process in the time evolving financial market, bypassing the need to specify a fixed number of regimes that switch within a fixed set of regimes with a static transition probability matrix.
- We propose a non-linear financial market forecasting model relying on a clustering-based regimes identification model, which can capture the cross-sectional dependencies among financial time series to generate forecast for the time-evolving financial market.
- We validate our model by implementing it on synthetic and real-world datasets, comprehensive experimental results, compared with state-of-the-art forecasting algorithms, demonstrate the suitability and promising performance of the proposed model.

The remainder of this paper is organized as follows. In Sect. 2, we discuss related work on financial market forecasting. Section 3 presents the proposed forecasting model in detail. Section 4 provides comprehensive experimental results on synthetic and real-world data and compares the results with other baselines. Finally, conclusions are given in Sect. 5.

2 Related Work

Financial time series forecasting is undoubtedly a hot topic for finance researchers in both academia and the finance industry due to its potential financial gain. Recent literature reports a number of methods applied to financial time series forecasting, including statistical models such as VARMA [20], TRMF [29] and the GARCH family models [2], However, most of these methods are based on linear equations, which are incapable of modeling financial data governed by complex non-linear dynamic patterns. Methods based on deep learning and graph neural networks [9,17,19,27,31] have also been proposed for financial time series forecasting, due to their capability of exploiting long-term and/or short-term dependencies and non-linear dynamic patterns underlying complex data.

Though these approaches can achieve relatively good performance, at the cost of high time complexity due to the overwhelming number of parameters, they are lacking in terms of model interpretability.

On the other hand, among the existing models employed in computational economics and econometric time series analysis, regime-switching models have proved the most preferable, due to their ability to capture non-linear patterns in the market, coupled with heightened model interpretability [12]. Boudt et al. [7] proposed a two-regime model with two state process for funding and market liquidity and TED spread. Alan et al. [1] proposed a multi-regime model to forecast the impact on volatility in global equity markets during the COVID-19 pandemic. Mahmoudi et al. [21] proposed a Markov regime-switching model for detection of structural regimes to analyze the impact of the crude oil market on the Canadian stock market. These methods require to specify the number of regimes manually. Sanquer et al. [25] proposed a hierarchical Bayesian model for automatically identifying hidden regimes. Note that all of these methods are focused on single time series with a static transition probability matrix, and can not be easily applied on multiple time series.

To model multiple time series, Hochstein et al. [14] proposed a regime switching vector autoregressive model that can deal with the changing dependency structures of multivariate time series. Matsubara [23] proposed a regime shift forecasting model on co-evolving data streams. Though it can identify regimes dynamically, it can not capture multiple regimes in one slide window as shown in Fig. 1. Tajeuna et al. [28] proposed a regime shift model for multiple time series forecasting, but it focused on regime identification on discontinuous windows and ignore continuity of time series. Overall, many authors have contributed to advances in handling cross-sectional regime identification on multiple time series, but it remains a challenging task, as many issues have not yet been addressed.

3 The Proposed Model

In this section, we give an overview of the proposed model, followed by detailed description of the cross-sectional regime identification, model description and estimation, and financial market forecasting.

3.1 Overview of the Proposed Model

This subsection gives an overview of the proposed financial market forecasting model relying on clustering-based cross-sectional regime identification, using the scenario shown in Fig. 2. We start by identifying regimes via clustering methods from the first slide window, where the optimal number of clusters in each window is determined by a silhouette score [24], and we then build a non-linear model on each of the identified regimes and obtain the regime parameters and transition probability. Finally, we make a forecast based on a non-linear regime model. Forecasting is performed on the window at the next timestamp. At this iteration step, we need to evaluate whether or not the regimes identified in the window exist in the regime database (RB) by comparing with their cluster centers. If they exist, we need to add the data to the corresponding regime and update the regime parameters; if not, we will add them to the regime database and estimate the regime parameters. The scenario of the proposed model is shown in Fig. 2.

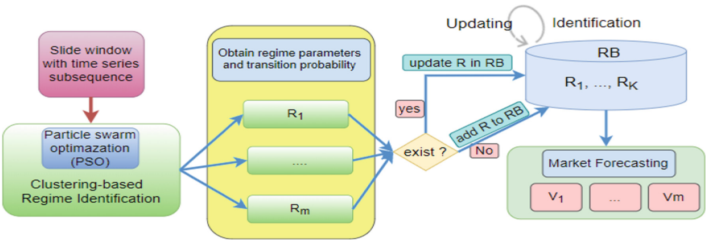
Slide window with time series subsequence
Particle swarm optimazation (PSO)
Clustering-based Regime Identification
Obtain regime parameters and transition probability
R1
....
Rm
exist ?
yes
No
update R in RB
add R to RB
Updating
Identification
RB
R1, ..., RK
Market Forecasting
V1
...
Vm

Fig. 2. Overview of the proposed model.

3.2 Cluster-Based Regime Identification

To identify the regimes in each slide window, we use a fuzzy C-Means to cluster time series to reveal the available structures within the data. Each of these structures is defined as a regime that shares some common patterns hidden in the data. To overcome the variable bias in time series data, we here used an extended FCM [18] of squared Euclidean distance to control the impact of each variable in evaluating the similarity between time series, the distance between a time series x_i with length L and a cluster center c_k is defined as follows:

$$d^2(x_i, c_k) = \sum_{j=1}^{L} \lambda_j ||x_{ij} - c_{kj}||^2 \lambda_j \geq 0, \sum_{j=1}^{L} \lambda_j = 1 \tag{1}$$

where λ_i is the importance of the i^{th} variable, and the larger the value of λ_i, the greater the importance of the i^{th} variable in the clustering process. This approach, to some extent, balances the noise underlying the data, achieving better clustering results. The coefficient $\lambda_j, (1 \leq j \leq L)$ can be estimated by Particle Swarm Optimization (PSO) algorithm, which is a tool for searching for optimal values by using a flock of particles, further details can be found in [15,30]. The objective function for the Sum of Error (SE) is defined as follows:

$$SCE = \sum_{k=1}^{K} \sum_{i=1}^{N} u_{ki}^{m} d^2(x_i, c_k) \tag{2}$$

where K is the number of clusters, $m(m > 1)$ is the fuzzification coefficient, and N is the number of time series. U and c_i are the partition matrix and center of the i^{th} cluster, respectively. By optimizing the objective function Eq. 2, the partition matrix and cluster centers can be calculated as follows:

$$v_k = \frac{\sum_{i=1}^{N} u_{ik}^{m} w_i}{\sum_{i=1}^{N} u_{ik}^{m}} \qquad U_{ki} = \frac{1}{\sum_{m=1}^{K} (\frac{||c_k - x_i||}{||c_k - x_i||})^{2/(m-1)}}, m > 1 \tag{3}$$

Thus, we can identify the regimes using the method presented above, and then build non-linear regime models based on these identified regimes, optimize the parameters of the regime models and estimate the transition probability. Finally, we make market forecast based on the regime models.

3.3 Regime Modeling and Transition Probability Estimation

The clustering approach described above only allows use to identifying regimes; that is; groups of time series which share similar patterns. For a better forecasting, we need to build an effective non-linear regime model on these clusters. Based on the work presented in [23,26], we can build a single non-linear regime model on each cluster in order to make a forecast. Thus, the non-linear regime model is defined as follows:

$$\frac{ds(t)}{dt} = \mu + \mathcal{G}g(s(t)) + \mathcal{F}f(s(t)) \tag{4}$$

$$v(t) = \epsilon_k + \mathcal{E}s(t) \tag{5}$$

where $s(t)$ is a hidden vector that evolves over time and describes the potential behaviors in the corresponding regime, and $v(t)$ is the actual observed value. $ds(t)/dt$ denotes the derivative with respect to time t. $g(\cdot)$ is a linear function, while $f(\cdot)$ is non-linear. Here, μ, $\mathcal{G}$ and $\mathcal{F}$ describe the potential activities $s(t)$, capturing linear and non-linear dynamic patterns of the regime. For parameters optimization, readers are referred to [23].

Regime transition probability describes the likelihood that the current regime stays the same or switch to another. In fact, we need to investigate whether the regimes in one window will change in the subsequent window or not. Rather than calculating static transition probabilities as elaborated in existing model, we can track the regime transition trajectory of each time series; the regime transition from regime R_i to regime R_j can be estimated as follows [28]:

$$\mathbb{Q}_1(i,j) = \begin{cases} 0 & if \sum_{i=1}^{K}\sum_{j=1}^{K}\aleph(i,j)\mathbb{N}(i,j) = 0 \\ \frac{\sum_{k=1}^{K}\aleph(i,k)\mathbb{N}(k,j)}{\sum_{i=1}^{K}\sum_{j=1}^{K}\aleph(i,j)\mathbb{N}(i,j)} & else \end{cases} \tag{6}$$

where $\aleph(i,j) = \frac{|\mathcal{N}(R_i,R_j)|}{\mathbf{N}}$ is the risk of suddenly switching from regime R_i to R_j, while $\mathbb{N}(i,j) = \frac{|\mathbf{N}_i \cap \mathbf{N}_j|}{|\mathbf{N}_i \cup \mathbf{N}_j|}$ is the probability of switching from R_i to R_j. $\mathcal{N}(R_i,R_j)$ is the number of time series appearing in the trajectory from regime R_i to R_j for the two windows. $\mathbf{N}_i$ and $\mathbf{N}_j$ are the numbers of time series present in regimes R_i and R_j, and $\mathbf{N}$ is the total number of time series. To further improve the above estimate, we also consider the difference between two cluster centers c_i and c_j in clustering-based regime identification: $\mathbb{Q}_2(i,j) = \frac{1}{|c_i - c_j|}$ for $i \neq j$ otherwise $\mathbb{Q}_2(i,j) = \frac{1}{|c_i|}$ ensuring the probability of staying at the same regime, and the effect of $s_i(t)$ and $s_j(t)$ that describes the potential behaviors in regime R_i and R_j: $\mathbb{Q}_3(i,j) = \frac{1}{|s_i - s_j|}$ for $i \neq j$ otherwise $\mathbb{Q}_3(i,j) = \frac{1}{|s_i|}$. All of these are the underlying drivers that may result in a regime switch. Thus, the transition probability $\mathbb{Q}$ of switching from regime R_i to regime R_j can be defined as follows:

$$\mathbb{Q}(i,j) = \mathbb{Q}_1(i,j)\frac{\mathbb{Q}_2(i,j)}{\sum_{i=1}^{k}\mathbb{Q}_2(i,j)}\frac{\mathbb{Q}_3(i,j)}{\sum_{i=1}^{k}\mathbb{Q}_3(i,j)}(1 \leq i,j \leq K) \tag{7}$$

Note that we can get the regime transition probability of each time series occurring within each slide window.

3.4 Financial Market Forecasting

Based on the previous section, once we build the non-linear model of each clustering-based regime, we can learn all the parameters. We make use of the fourth-order Runge-Kutta method [16] to generate l/γ ($\gamma = 3$) potential activity sets $\mathcal{S} = [s(t+\gamma), s(t+2\gamma)\cdots, s(t+l)]$ to estimate the $[\{s(t+1),\cdots, s(t+l)]$, as presented in [23], for the forecasting step. The process is defined as follows:

$$s(t+\gamma) = s(t) + \frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4) + \mathcal{O}^5 \tag{8}$$

where $ds(t)/dt = F(s(t))$, $K_1 = \gamma F(s(t))$, $K_2 = \gamma F(s(t)+\frac{1}{2}K_1)$, $K_3 = \gamma F(s(t)+\frac{1}{2}K_2)$, $K_4 = \gamma F(s(t) + K_3)$. Thus, once we obtain the potential activity set $\mathcal{S}$, the regime model Eq. 5 can be used to make l-steps-ahead-forecast $\mathcal{V}$, as follows:

$$\mathcal{V} = \epsilon_k + \mathcal{E}\mathcal{S} \tag{9}$$

The detailed framework of our forecasting model utilizing clustering-based regime identification is shown in Algorithm 1.

4 Experiments

4.1 Dataset Description

To test the performance of our model, we used one synthetic dataset and three real-world financial datasets. The synthetic dataset (SyD) consists of 400 time series of length 1350, governed by 6 distinct regime. They are generated by the regime functions RF as $RF(t) = \sum_{k=1}^{6}\alpha_k\varpi_k(t)fcn_k(t)$, where $\alpha \in \{1,0\}$ ($\sum_{k=1}^{6}\alpha_k = 1$) allows having one regime to be exhibited in a time interval. $\varpi_k(t) \in [0,1]$ is to exhibit regimes with constraint $\sum_{k=1}^{6}\varpi_k(t)$=1. The 6 regime functions are defined as follows:

$$\begin{gathered}
fcn_1(t) = \cos(\frac{2\pi t}{5}) + cos(\pi(t-3)) \qquad fcn_2(t) = \cos(\frac{2\pi t}{5}) - cos(2\pi(t-3)) \\
fcn_3(t) = \sin(\frac{2\pi t}{5} - 3) - \sin(\frac{\pi t}{6}) \qquad fcn_6(t) = \cos(\frac{3\pi t}{5}) + \sin(\frac{2\pi t}{5} - t) \\
fcn_4(t) = \tan(\frac{\pi t}{2} - 3) - \frac{1}{2}\cos(\frac{\pi(t-3)}{6}) + \cos(\pi(t-13)) \\
fcn_5(t) = \tan(\frac{\pi t}{2} - 3) * \cos(\frac{\pi(t-3)}{6}) + \cos(\pi(t-13))
\end{gathered} \tag{10}$$

Algorithm 1: Framework of the Proposed Model

```
Input: Financial time series: $X$, Slide window length: w, Threshold: $\eta$
Forecasting window: l, maximal number of regimes in a window: m
Output: Forecasting time series: F, Transition probability: $\tau P$
begin
    /* Initialization */
      – tc = w, current time point;
      – $RB$ = [], regime database;
      – F = [], forecasting time series;
      – $\tau P$ = [], transition probability;

    repeat
        /* Get slide window data $X_w$ from $X$ */
        $X_w$ = X[tc-w:tc];
        /* Identify regimes on $X_w$ by clustering-based method */
        Obtain a regime set $RS$;
        if $len(RB)$ ==0 then
            Add all the regimes in $RS$ to $RB$;
            for $R$ in $RB$ do
                /* Regime estimated value on R */
                Obtain parameters on regime R by Eq. 4 and 5;
                Generate $\mathcal{S}$ by Eq. 8 set on regime $R$;
                Obtain forecast value on regime $R$ by Eq. 9;
                Obtain $\mathcal{Q}$ transition probability by Eq. 7;
            F.append($v^l$);
            $\tau P$.append($\mathcal{Q}$);
        for $R_{RS}$ in $RS$ do
            Err =[];
            Obtain $C_{RS}$ center of regime $R_{RS}$;
            for $R_{RB}$ in $RB$ do
                Obtain $C_{RB}$ center of regime $R_{RB}$;
                Err.append($d^2(C_{RS} - C_{RB})$);
            if $min(Err) > \eta$ then
                /* Identified a new regime $R_{RS}$ */
            else
                /* Regime $R_{RS}$ already existed */
                Find the best $R_{Rm}$ of $R_{RS}$ in RB and add $R_{Rm}$ data into $R_{RS}$;
            Obtain parameters on regime $R_{RS}$ by Eq. 4 and 5;
            Generate $\mathcal{S}$ by Eq. 8 set on $R_{RS}$;
            Obtain forecast value on regime $R_{RS}$ by Eq. 9;
            Obtain $\mathcal{Q}$ transition probability by Eq. 7;
            Replace $R_{Rm}$ by $R_{RS}$;
        F.append($v^l$);
        $\tau P$.append($\mathcal{Q}$);
    until iterate for next window;
```

Three real-world datasets were selected from financial markets. The first one (Stocks) consists of 200 stocks selected from top 500 companies including AAPL, IBM, BAC, MSFT, WMT and so on. The second (Sectors) is comprised of 9 financial sector SPDR Funds: XLB, XLE, XLF, XLI, XLK, XLP, XLU, XLV, XLY. The last (ETFs) contains 18 ETF funds: EWA, EWC, EWD, EWG, EWH, EWJ, EWS, EWW, EWP, EWQ, EWM, EWL, EWI, EWN, EWO, EWK, EWU, SPY. These three datasets consist of daily frequencies (for business days only), comprising over 22 years' data, available on the yahoo finance website[1]. It is worth noting that the experimental financial data were converted into volatility based on the log-return of the close price, as described in [10].

4.2 Performance Metrics

There are many metrics for evaluating the performance of a forecasting model. Here, due to space limition, we report the results obtained on the most popular metric for evaluating forecasting performance: the root mean square error (RMSE) [22]. The performance metrics is defined as follows:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\sigma_i - \hat{\sigma_i})^2} \tag{11}$$

where T is the length of the forecasting window, σ_i and $\hat{\sigma_i}$ are the ground truth and predicted value at time t respectively. The smaller value, the better the performance of a forecasting model, meaning that the predicted value is closer to the ground truth.

4.3 Experimental Results and Discussion

In this section, we present the results of our model, and evaluate its performance against some comparable state-of-the-art methods on synthetic and real-world datasets. We tested our model using a slide window of half a year (126 business days) to forecast one month ahead (21 business days) for the real-world data, and a slide window of length 75 to forecast 15 steps ahead on the synthetic data.

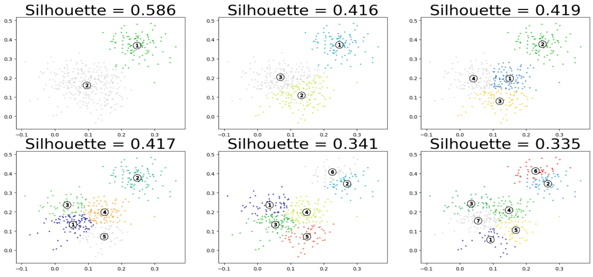
Silhouette = 0.586
Silhouette = 0.416
Silhouette = 0.419
Silhouette = 0.417
Silhouette = 0.341
Silhouette = 0.335

Fig. 3. The clustering results of the first window in stock dataset

[1] https://ca.finance.yahoo.com/.

Regime Identification Analysis. An accurate, effective regime identification model is the key to achieving good performance of our forecasting model, since the latter is built on the regime identification results. We therefore start by analyzing the regimes identified by our clustering-based regime identification model. First, we need to identify the regimes in the first slide window and then iterate to the next slide window. Note that the synthetic dataset SyD was generated with K = 6 known regimes by Eq. 10, but for our real-world datasets, we do not have ground truth information such as the number of regimes. To validate the performance of our regime identification model, we therefore test it for numbers of regimes varying from 2 to the maximal value 7 on the synthetic and real-world datasets, and make use of the silhouette score to find the optimal number of regimes in each slide window. For example, the results for the first slide window of the Stock dataset is shown Fig. 3. It is clear that all 200 time series in the window are clustered into 2 groups with the largest silhouette score (0.586), which means that there are two regimes identified in the first slide window. Look at the regimes identified in the synthetic and real-world datasets as shown in Fig. 4, it can be seen that 6 distinct regimes are identified in the synthetic data, which is corresponds to the true number of regimes generated. There are 6 distinct regimes identified both in Stock and Sector dataset. While 8 different regimes are identified in the ETFs dataset as in Fig. 4. In summary, our clustering-based regime identification model can find the regime groups accurately and effectively.

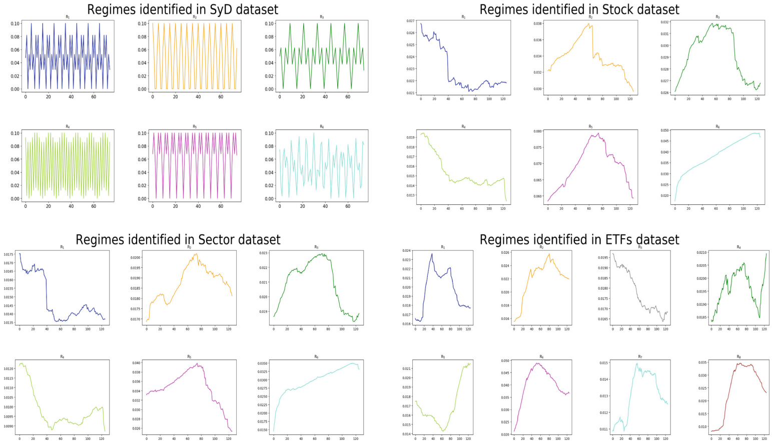

Fig. 4. The regimes identified in four datasets respectively.

Market Forecasting Analysis. To validate the performance of our model, we use a slide window of half a year with 126 business days to forecast one month with 21 days ahead for the real-world dataset, while a window of 75 steps to forecast 15 steps ahead for the synthetic dataset. For page limitation, we randomly selected two time series from the Sector and ETFs datasets, respectively, as examples to show the performance of regime identification and value forecasting of our proposed model. The results are as shown in Fig. 5.

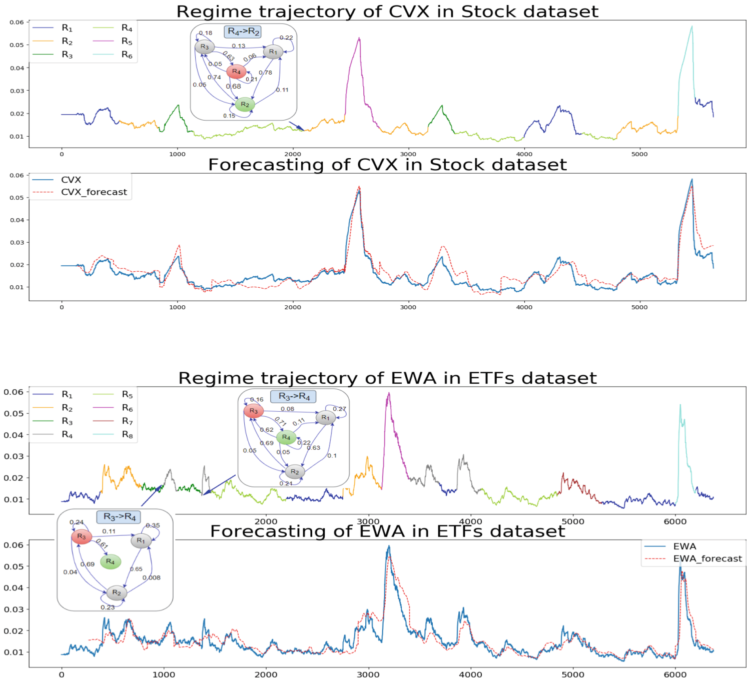

Fig. 5. Examples of regime trajectories and market forecasting, and also including three regime transition processes with the time-varying transition probabilities in trajectory.

Look at the regime trajectories in Fig. 5, it can be seen that there are six and eight regimes identified in stock and ETFs dataset respectively, and the identified regimes, marked with corresponding colors, are roughly synchronized to the regime exhibited in the real market. Moreover, we present regime switching process on trajectories of stock CVX and etf EWA. When comparing with two switches from R_3 to R_4 in trajectory of etf EWA as in Fig. 5, the probabilities of R_4 ($p_{R_3->R_4} = 0.61$) in the first switching are completely different from the second one ($p_{R_3->R_4} = 0.71$) due to our unique time-varying learning mechanism that is lacked by existing methods. Furthermore, this kind of time-varying transition probabilities can be used to explicitly explain the regime switching mechanism from the model interpretability. What the most important is that the ground truth is obviously well matched by our forecasting on stock CVX and ETF EWA as shown in Fig. 5. In summary, our model can identify regime accurately and demonstrates promising performance for the market forecasting.

Performance Analysis. To validate the forecasting quality of our model, we compared our experimental results with that generated by the baselines on the four test datasets. The baselines for the result comparison are as follows: VARMA [20] is a classical statistical model for analyzing and forecasting time series data.

TRMF [29] is a regularized matrix factorization based auto-regressive prediction model. VAR-MLP [31] is a hybrid model that combines auto-regressive and multi-layer model for time series forecasting. MAGNN [9] is a graph neural network based methods for financial time series prediction. DSTP-RNN [19] is an attention-based recurrent neural network method for multivariate time series forecasting. RegimeCast [23] is a regime shifts based forecasting model for co-evolving real-time data streams. The results are shown in Table 1; the best results are highlighted in bold and the second best are underlined.

Table 1. Comparison of RMSE on test datasets.

Models	SyD	Stocks	Sectors	ETFs
VARMA	.35303	.00349	.00316	.00328
TRMF	.26986	.00307	.00275	.00262
VAR-MLP	.23388	.00226	.00232	.00245
MAGNN	.18652	.00161	.00121	.00135
DSTP-RNN	.20421	.00172	.00086	.00237
RegimeCast	.12516	.00205	.00143	.00169
Ours	**.08072**	**.00103**	**.00041**	**.00082**

We can see that our model outperforms the baselines, earning the smallest (best) values on the performance metric, and shows a significant improvement over the second best baselines (underlined) in Table 1. The deep learning methods outperform the traditional methods (VARMA, TRMF), as the latter cannot harness the non-stationary and non-linear dependencies for the prediction. However, the deep learning based VAR-MLP can not explicitly model the cross-sectional dependencies for the prediction, putting it at disadvantage compared to the graph neural network based methods (DSTP-RNN, MAGNN). RegimeCast can capture the non-stationary and non-linear dependencies for the value prediction, but it ignores the cross-sectional regime dependencies, which is a degenerated version of our model that operates without considering multiple regimes in slide windows. In sum, our model shows promising performance on financial market forecasting.

5 Conclusion

In this paper, we have proposed a financial market forecasting model utilizing clustering-based cross-sectional regime identification. Our proposed model not only captures cross-sectional dependencies in multiple time series, but also identifies cross-sectional regimes dynamically along with the time-evolving financial markets, using a time-varying transition probability matrix. In addition, we have built a non-linear forecasting model based on a clustering-based cross-sectional regime model for financial market forecasting. Experimental results on synthetic and real-world datasets demonstrate the promising performance of our model. However, we will improve the performance of our model and test it on hourly and minutes financial time series or sensor data. Moreover, we also may apply our model into other domains, such as the energy consumption and mechanical fault diagnosis. In short, we see the significant challenges for our future work, but we are confident that the proposed method has great potential in real applications.

Acknowledgements. This work was supported by NSERC CRD - Quebec Prompt - Laplace Insights - EAM - joint program under the grants CRDPJ 537461-18 and 114-IA-Wang-DRC 2019 to S. Wang and Chinese Scholarship Council scholarship to R. Chen.

References

1. Alan, N.S., Engle, R.F., Karagozoglu, A.K.: Multi-regime forecasting model for the impact of COVID-19 pandemic on volatility in global equity markets. NYU Stern School of Business (2020)
2. Ali, G., et al.: EGARCH, GJR-GARCH, TGARCH, AVGARCH, NGARCH, IGARCH and APARCH models for pathogens at marine recreational sites. J. Stat. Econ. Methods **2**(3), 57–73 (2013)
3. Ang, A., Timmermann, A.: Regime changes and financial markets. Annu. Rev. Financ. Econ. **4**(1), 313–337 (2012)
4. Baillie, R.T., Morana, C.: Modelling long memory and structural breaks in conditional variances: an adaptive FIGARCH approach. J. Econ. Dyn. Control **33**(8), 1577–1592 (2009)
5. Banerjee, A., Urga, G.: Modelling structural breaks, long memory and stock market volatility: an overview. J. Econom. **129**(1–2), 1–34 (2005)
6. Bollerslev, T., Engle, R.F., Nelson, D.B.: Arch models. Handb. Econom. **4**, 2959–3038 (1994)
7. Boudt, K., Paulus, E.C., Rosenthal, D.W.: Funding liquidity, market liquidity and ted spread: a two-regime model. J. Empir. Financ. **43**, 143–158 (2017)
8. Charfeddine, L., Khediri, K.B.: Financial development and environmental quality in UAE: cointegration with structural breaks. Renew. Sustain. Energy Rev. **55**, 1322–1335 (2016)
9. Cheng, D., Yang, F., Xiang, S., Liu, J.: Financial time series forecasting with multi-modality graph neural network. Pattern Recogn. **121**, 108218 (2022)
10. Christensen, B.J., Prabhala, N.R.: The relation between implied and realized volatility. J. Financ. Econ. **50**(2), 125–150 (1998)
11. Faniband, M., Faniband, T.: Government bonds and stock market: volatility spillover effect. Indian J. Res. Capital Markets **8**(1–2), 61–71 (2021)
12. Hamilton, J.D.: A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica J. Econom. Soc. 357–384 (1989)
13. Hamilton, J.D.: Regime switching models. In: Durlauf, S.N., Blume, L.E. (eds.) Macroeconometrics and Time Series Analysis. TNPEC, pp. 202–209. Palgrave Macmillan UK, London (2010). https://doi.org/10.1057/9780230280830_23
14. Hochstein, A., Ahn, H.I., Leung, Y.T., Denesuk, M.: Switching vector autoregressive models with higher-order regime dynamics application to prognostics and health management. In: 2014 International Conference on Prognostics and Health Management, pp. 1–10. IEEE (2014)

15. Hu, M., Wu, T., Weir, J.D.: An adaptive particle swarm optimization with multiple adaptive methods. IEEE Trans. Evol. Comput. **17**(5), 705–720 (2012)
16. Jackson, E.A.: Perspectives of Nonlinear Dynamics: Volume 1, vol. 1. CUP Archive (1989)
17. Lai, G., Chang, W.C., Yang, Y., Liu, H.: Modeling long-and short-term temporal patterns with deep neural networks. In: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 95–104 (2018)
18. Li, J., Izakian, H., Pedrycz, W., Jamal, I.: Clustering-based anomaly detection in multivariate time series data. Appl. Soft Comput. **100**, 106919 (2021)
19. Liu, Y., Gong, C., Yang, L., Chen, Y.: DSTP-RNN: a dual-stage two-phase attention-based recurrent neural network for long-term and multivariate time series prediction. Expert Syst. Appl. **143**, 113082 (2020)
20. Lütkepohl, H.: Forecasting with VARMA models. Handb. Econ. Forecast. **1**, 287–325 (2006)
21. Mahmoudi, M., Ghaneei, H.: Detection of structural regimes and analyzing the impact of crude oil market on Canadian stock market: Markov regime-switching approach. Studies in Economics and Finance (2022)
22. Makridakis, S.: Accuracy measures: theoretical and practical concerns. Int. J. Forecast. **9**(4), 527–529 (1993)
23. Matsubara, Y., Sakurai, Y.: Regime shifts in streams: real-time forecasting of co-evolving time sequences. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1045–1054. ACM (2016)
24. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
25. Sanquer, M., Chatelain, F., El-Guedri, M., Martin, N.: A smooth transition model for multiple-regime time series. IEEE Trans. Signal Process. **61**(7), 1835–1847 (2012)
26. Scheffer, M., Carpenter, S., Foley, J.A., Folke, C., Walker, B.: Catastrophic shifts in ecosystems. Nature **413**(6856), 591 (2001)
27. Shih, S.Y., Sun, F.K., Lee, H.Y.: Temporal pattern attention for multivariate time series forecasting. Mach. Learn. **108**(8), 1421–1441 (2019)
28. Tajeuna, E.G., Bouguessa, M., Wang, S.: Modeling regime shifts in multiple time series. arXiv preprint arXiv:2109.09692 (2021)
29. Yu, H.F., Rao, N., Dhillon, I.S.: Temporal regularized matrix factorization for high-dimensional time series prediction. In: Advances in Neural Information Processing Systems 29 (2016)
30. Zhan, Z.H., Zhang, J., Li, Y., Chung, H.S.H.: Adaptive particle swarm optimization. IEEE Trans. Syst. Man Cybern. Part B (Cybern.) **39**(6), 1362–1381 (2009)
31. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing **50**, 159–175 (2003)