ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Գարիկ Լևոնի Ադամյան

Մոդելների վրա հիմված ժամանակային շարքերի կլաստերիզացիա

Ը.00.08 - «Տնտեսության մաթեմատիկական մոդելավորում» մասնագիտությամբ ֆիզիկամաթեմատիկական գիտությունների թեկնածուի գիտական աստիճանի հայցման համար

ՍԵՂՄԱԳԻՐ

ԵՐԵՎԱՆ 2023

YEREVAN STATE UNIVERSITY

Garik Adamyan

Model-based time series clustering

SYNOPSIS

of the thesis for the degree of candidate of physical and mathematical sciences in the specialty Ը.00.08 - "Mathematical modeling of economy"

YEREVAN 2023

Ատենախոսության թեման հաստատվել է Երևանի պետական համալսարանի մաթեմատիկայի և մեխանիկայի ֆակուլտետի խորհրդի կողմից:

Գիտական ղեկավար՝ տնտեսագիտության դոկտոր
Ռ.Ա. Գևորգյան

Պաշտոնական ընդդիմախոսներ՝ ֆիզ. մաթ. գիտ. դոկտոր
Վ.Կ. Օհանյան
ֆիզ. մաթ. գիտ. թեկնածու
Վ. Գ. Բարդախչյան

Առաջատար կազմակերպություն՝ Հայաստանի Ազգային
Պոլիտեխնիկական Համալսարան

Պաշտպանությունը կկայանա 2023թ. դեկտեմբերի 27-ին, ժ. 15 : 00-ին Երևանի պետական համալսարանում գործող ԲՈԿ-ի 050 "Մաթեմատիկա" մասնագիտական խորհրդի նիստում (0025, Երևան, Ալեք Մանուկյան 1):

Ատենախոսությանը կարելի է ծանոթանալ ԵՊՀ գրադարանում: Սեղմագիրն առաքվել է 2023 թ.-ի նոյեմբերի 20-ին:

Մասնագիտական խորհրդի գիտական քարտուղար՝ Կ.Լ. Ավետիսյան

The topic of the thesis was approved at Yerevan State University.

Scientific advisor doctor of economic sciences
R.A. Gevorgyan

Official opponents doctor of phys.-math. sciences
V.K. Ohanyan
Ph.D. of phys.-math. sciences
V.G. Bardakchyan

Leading organization National Polytechnic University of Armenia

The defense will be held on December 27, 2023 at 15 : 00 at a meeting of the specialized council of mathematics 050 operating at Yerevan State University (0025, 1 Alek Manukyan St, Yerevan).

The thesis can be found at the YSU library. The synopsis was sent on November 20th, 2023.

Scientific secretary of the specialized council K. L. Avetisyan

General Characteristics Of The Work

Actuality of the subject. The clustering problem is an unsupervised learning problem to group similar observations. Time series clustering, in turn, is a family of clustering methods that study realizations of random processes as samples.

This topic has deep historical roots, tracing its origins back to some of the earliest statistical studies in the middle of the 20th century. The rapid growth of computational power and the increasing abundance of time series data, starting in the late 20th century, has led to a renaissance in this field, transforming it from a niche topic into an area of broad and pervasive importance.

Due to their unsupervised nature, time series clustering algorithms have a broad range of uses in numerous fields. In finance, for instance, clustering techniques are used to group stocks with similar price movements, helping portfolio managers diversify their portfolios and hedge against risk. In medicine, time series clustering allows for the identification of common patterns in patient data, leading to more accurate diagnoses and more effective treatment plans. The field of climate science also benefits from this technique, as it allows for the classification of weather patterns, aiding in the prediction and understanding of climate change. These and many other applications demonstrate the practical relevance and broad impact of time series clustering. [1]

Purpose and goals of the thesis. Although the literature on time series clustering is extensive, it is limited by algorithms and methods with strong theoretical evidence. To fill this gap, several approaches based on theoretical results of random processes have recently been proposed to study the asymptotic properties of time series clustering algorithms. A time series clustering algorithm is asymptotically consistent if it can recover the ground truth clusters, for large enough samples. The goal of the thesis is to examine the consistent clustering conditions and methods for the time series datasets generated by model-based procedures. This includes an examination of the datasets generated by well-known models ARMA, GARCH, ARMA-GARCH, and ARIMA. In the thesis, we also conduct extensive experiments to show the practical applicability of the methods discussed.

The object of research. The object of this research is a time series dataset generated by some of the common time series models such as ARMA, GARCH, ARMA-GARCH, ARIMA models. We examine the several metrics defined in the space of discussed models, clustering algorithms, and clustering evaluation measures. In the application section, we examine the foreign exchange (FX) market.

The methods of research. The methods of research include both theoretical and practical methods. We define the asymptotically consistent clustering problem for common time series models and provide a generic framework for clustering time series datasets. The asymptotic consistency of a described algorithm is proved by using asymptotic properties of defined metrics and their empirical estimates. The numerical experiments and applications are implemented with the Python programming language.

Scientific novelty. The theoretical novelty of the thesis is listed below.

- We define the theoretical dissimilarity measures for ARMA-GARCH models. We define empirical dissimilarity measures for the common time series models, including ARMA, GARCH, ARMA-GARCH, and ARIMA models, and show their asymptotic consistency.

- We define the asymptotically consistent clustering problem for time series data generated by the above-mentioned models. For clustering time series data generated by model-based procedures, we examine two problem setups. In the first scenario, when the orders of the underlying models are known, we show the strong asymptotic consistency of the described clustering algorithm. In the second problem setup, we assume that the underlying processes are unknown, and only the upper limits of the orders of the models are known. In this problem setup, we prove the weak consistency of the clustering algorithm.

- We applied the above-mentioned methods for the dynamic clustering of the FX market. With an empirical approach, we showed that the proposed methods are applicable to the problem of clustering of the FX market and analyzing the dynamic structure of the market, and resulting clusters reflect several important characteristics specific to the FX market structure.

Practical significance. The practical novelty of the thesis is listed below.

- To show the practical applicability of the discussed methods, in this thesis, we conduct several experiments. We evaluate several model-free algorithms for clustering time series datasets generated by GARCH processes. Several experiments show that model-free, algorithms generally speaking do not show the desired asymptotically consistency properties. In contrast to model-free algorithms, we also evaluated the methods proposed in this thesis, which show strong performance of clustering time series datasets generated by the ARMA, GARCH, and ARMA-GARCH processes.

- We consider the dynamic clustering of the foreign exchange market. The resulting clusters incorporate several important properties of the FX market. 1. Currencies with fixed exchange rates mostly appear in the same cluster. 2. The resulting clusters reflect the relationships of currencies circulating in the same geographical region. 3. Clusters reflect economic associations between countries. 4. Clusters reflect the industrial connections between currencies (countries).

- Having the results of the dynamic clustering of the foreign exchange market, we describe a market stability analysis method by comparing the clustering results for each consecutive period of time. We showed that the dynamic comparison of the FX market also can serve as a useful tool to analyze the effect of major economic events on the market structure.

The approbation of obtained results. The results of the thesis were reported

- in the scientific seminars held in the Department of Mathematical Modeling in Economics of the Faculty of Economics and Management of Yerevan State University,

- in the 14th International Conference on Computer Science and Information Technologies CSIT 2023 September 25 - 30, 2023, Yerevan, Armenia. `https://csit.am/2023/proceedings/ITCT/ITCT_1.pdf`

- in the internal scientific seminars of the international artificial intelligence company TurinTech.ai

Publications. The thesis is based on results published in 3 scientific articles.

- G. L. Adamyan, “Comparison of model-free algorithms for clustering GARCH processes,” Mathematical Problems of Computer Science, vol. 58, pp. 32–41, 2022.

- G. Adamyan, “Weakly consistent offline clustering of ARMA processes,” Journal of Contemporary Mathematical Analysis, vol. 58, no. 3, pp. 183–190, 2023

- G. Adamyan, “Model-based clustering of foreign exchange market”, PROCEEDINGS OF ENGINEERING ACADEMY OF ARMENIA (PEAA), vol. 20, no. 1, pp. 24-33, 2023

The structure and the content of the thesis. The thesis consists of an introduction, four chapters, a summary, and a bibliography. The number of references is 62. The thesis consists of 86 pages with the Appendix section and 77 pages without it.

The Content Of The Thesis

In **chapter 1**, we reviewed basic time series concepts, common time series models, and consistent model estimation conditions. The definitions, and stationarity conditions of ARMA, GARCH, and ARMA-GARCH models are presented.

Chapter 2 is dedicated to the review of existing approaches of time series clustering, clustering accuracy evaluation metrics and applications of time series clustering.

In **chapter 3**, we provide the main theoretical results of the thesis. We studied the problem of consistent clustering of time series datasets generated by the most common models including ARMA, GARCH, ARMA-GARCH, and ARIMA models. We start, by defining a consistent clustering framework and consistent clustering algorithm for clustering ARMA processes. The framework of asymptotic consistent clustering algorithms for ergodic and stationary processes in online and offline problem setups was first introduced in [2]. We are given a time series dataset with N samples $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$. We assume that each $\mathbf{x}_i$ is generated from one of the κ unknown ARMA process with unknown forecasting function $\mathcal{F}^{(k)}, k = 1, 2, ..., \kappa$, where $\kappa < N$. Note that time series samples may have arbitrary lengths, and we denote the length of $\mathbf{x}_i$ time series by n_i.

Definition 3.1.1 (Ground-truth $\mathcal{G}$)**.** *Let $\mathcal{G} = \mathcal{G}_1, ..., \mathcal{G}_k$ be a partitioning of the set $\{1, 2, ..., N\}$ into κ disjoint subsets $\mathcal{G}_k$, $\mathcal{G}_k \neq \emptyset$, $k = 1, 2, ..., \kappa$, such that the forecasting function of the process that generates $\mathbf{x}_i$, $i = 1, 2, ..., N$ is $\mathcal{F}^{(k)}$ for some $k = 1, 2, ..., \kappa$ if and only if $i \in \mathcal{G}_k$. We call $\mathcal{G}$ the ground-truth clustering.*

We denote by $X^{(k)}$ the underlying ARMA process for the cluster $\mathcal{G}_k$. The domain of the clustering function f is the finite set of samples $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ and a parameter κ (the number of target clusters) and the range is a set of partitions $f(\mathcal{D}, \kappa) := \{C_1, ..., C_\kappa\}$ of the index set $\{1, 2, ..., N\}$. The following definitions represent the rigid formulation of the asymptotically consistent clustering.

Definition 3.1.2 (Consistency: offline settings)**.** *A clustering function f is consistent for a set of sequences $\mathcal{D}$ if $f(\mathcal{D}, \kappa) = \mathcal{G}$. Moreover, denoting by $n = min\{n_1, ..., n_N\}$, f is called strongly asymptotically consistent in the offline sense if with probability 1 $P(\exists n' \forall n > n' f(\mathcal{D}, \kappa) = \mathcal{G}) = 1$. We call it weakly asymptotically consistent if* $\lim_{n\to\infty} P(f(\mathcal{D}, \kappa) = \mathcal{G}) = 1$

To construct an asymptotically consistent algorithm, we start by defining a metric on ARMA processes. Let us denote by $\mathcal{L}$ the class of invertible ARMA models. The invertibility assumption ensures that X_t can be represented in terms of its past values according to the $AR(\infty)$ formulation.

$$\pi(B)X_t = \epsilon_t \tag{3.1.1}$$

where $\pi(B) = \theta(B)^{-1} * \phi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$. The coefficients of sequence $\boldsymbol{\pi}_x$ are determined by the following recursive equations ([3]: p. 86):

$$\pi_j + \sum_{k=1}^{q} \theta_k \pi_{j-k} = -\phi_j, \quad j = 0, 1, ... \tag{3.1.2}$$

where $\phi_0 := -1, \phi_j := 0$ for $j > p$, and $\pi_j := 0$ for $j < 0$. Having (3.1.1), we note that given initial values and known orders, any process $\{X_t\} \in \mathcal{L}$ is fully characterized by the sequence $\boldsymbol{\pi}_x$. Defined sequence also completely specifies the forecasting function $\mathcal{F}_t = \mathbb{E}[X_t|X_{t-1}, X_{t-2}, ...] = \pi_1 X_{t-1} + \pi_2 X_{t-2} + \ldots + \epsilon_t$ of the processes $\{X_t\}$ [4].

Recalling the (3.1.1) representation of the invertible ARMA process, Piccolo in work [5] introduced metric on $\mathcal{L}$ as a measure of structural diversity between stochastic processes $X^{(1)}, X^{(2)} \in \mathcal{L}$. The metric function d_{PIC} on $\mathcal{L}$ is defined as

$$d_{PIC}(X^{(1)}, X^{(2)}) = \left\{ \sum_{j=0}^{\infty} (\pi_{1,j} - \pi_{2,j})^2 \right\}^{1/2} \tag{3.1.3}$$

where $\{\pi_{1,j}\}_{j=0}^{\infty}$ and $\{\pi_{2,j}\}_{j=0}^{\infty}$ is the $\boldsymbol{\pi}$ sequences for the $X^{(1)}$ and $X^{(2)}$ processes respectively. The d_{PIC} distance is well defined for all $X \in \mathcal{L}$ and can be computed even for processes with arbitrary orders and parameters. As for given ARMA process X the sequence $\{\pi_{x,j}\}_{j=0}^{\infty}$ fully characterizes the forecasting function $\mathcal{F}$, therefore, the defined distance between two ARMA processes, with given orders, is zero if, for the provided same set of initial values, the corresponding models produce the same forecasts [4]. Having this fact, if the $\mathbf{x}_i$ and $\mathbf{x}_j$ are two realizations of the two invertible ARMA processes $X^{(i)}$ and $X^{(j)}$, then if $i, j \in \mathcal{G}_k$ for some $k \in 1, ..., \kappa$, then given the same initial values the processes in the same cluster, will produce the same forecast, since the corresponding distance between processes $d_{PIC}(X^{(i)}, X^{(j)}) = 0$.

We aim to demonstrate certain properties of the d_{PIC} measure that will be useful for subsequent results. Let us denote by $\beta = (\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ the parameters vector of the process $\{X_t\} \in \mathcal{L}$, and by $\mathcal{B}^0 = \{\beta \in \mathbb{R}^{p+q} : \theta(z) \textit{ is invertible}\}$, then the following proposition holds.

Proposition 3.1.1. *The $\pi_j = h(\beta), j = 1, \ldots, \infty$ is a continuous function on $\mathcal{B}^0$.*

Let $X^{(1)}, X^{(2)} \in \mathcal{L}$ be two invertible ARMA processes, with (p_1,q_1), β^1, $\boldsymbol{\pi}_1 = \{\pi_{1,j}\}_{j=0}^{\infty}$ and (p_2,q_2), β^2, $\boldsymbol{\pi}_2 = \{\pi_{2,j}\}_{j=0}^{\infty}$ orders, parameter vectors and associated $\boldsymbol{\pi}$ coefficients respectively. Denoting by $\mathcal{B}^i = \{\beta \in \mathbb{R}^{p_i+q_i} : \textit{roots of the } \theta^i(z) \textit{ are distinct}\}$ for i = 1,2 we can formalize the following important proposition.

Proposition 3.1.2. *(Continuity of d_{PIC}) $d_{PIC}(\cdot, \cdot)$ is continuous as a function of the vectors β^1, β^2 on $\mathcal{B}^1 \times \mathcal{B}^2$.*

In addition to the properties listed, we can show that d_{PIC} has a weakly consistent estimator $\widehat{d}_{PIC}$. Let us consider samples $\mathbf{x}_1 = \{x_1^1, x_2^1, .., x_{n_1}^1\}$ and $\mathbf{x}_2 = \{x_1^2, x_2^2, .., x_{n_2}^2\}$ generated from the $X^{(1)}$ and $X^{(2)}$ ARMA processes. Then the QMLE for the estimating ARMA(p,q) processes has the following form:

$$\widehat{L}_n(\theta) = -\frac{1}{2} \sum_{t=1}^{n} \frac{(x_t - \sum_{i=1}^{t} \pi_j x_{t-j})^2}{\sigma^2} + log\sigma^2 \tag{3.1.10}$$

We define the estimator of d_{PIC} the Euclidean distance between sequences $\boldsymbol{\pi}_i$ $(i = 1, 2)$ of the estimated parameters with QMLE and samples $\mathbf{x}_1, \mathbf{x}_2$. Let us denote the empirical estimates of d_{PIC} as follows

$$\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2) = \left\{ \sum_{j=1}^{\infty} (\widehat{\pi}_{1,j} - \widehat{\pi}_{2,j})^2 \right\}^{1/2}$$

$$\widehat{d}_{PIC}(\mathbf{x}_1, X^{(1)}) = \left\{ \sum_{j=1}^{\infty} (\widehat{\pi}_{1,j} - \pi_{1,j})^2 \right\}^{1/2} \tag{3.1.11}$$

where $\{\widehat{\pi}_{i,j}\}_{j=1}^{\tau}$ are given by (3.1.2) and parameters vectors $\widehat{\beta^i}$ estimated by QMLE (3.1.10).

We proved the consistency of these estimators for two problem configurations, firstly, for the case where orders of the $\{X_t^{(1)}\}, \{X_t^{(2)}\} \in \mathcal{L}$ ARMA process are known and for the case where exact process orders are unknown but are given some constants positive P_{max}, Q_{max} such that the orders of $\{X_t^{(1)}\}, \{X_t^{(2)}\} \in \mathcal{L}$ ARMA process the $p_1, p_2 < P_{max}$ and $q_1, q_2 < Q_{max}$.

Proposition 3.1.3. *If the orders of the $\{X_t^{(1)}\}, \{X_t^{(2)}\} \in \mathcal{L}$ ARMA process are known, then under stationarity condition the $\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2)$ and $\widehat{d}_{PIC}(\mathbf{x}_1, X^{(1)})$ distance estimators are strongly consistent*

$$\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow[n \to \infty]{a.s.} d_{PIC}(X^{(1)}, X^{(2)})$$
$$\widehat{d}_{PIC}(\mathbf{x}_1, X^{(2)}) \xrightarrow[n_1 \to \infty]{a.s.} d_{PIC}(X^{(1)}, X^{(2)})$$

This approach is intuitive and ensures almost sure consistency but it limits us to applying the estimated distance to a clustering problem defined earlier since it is impractical to assume that the orders of all underlying processes are known. The next proposition provides a more generic framework for estimating Autoregressiv distance.

Proposition 3.1.4. *If there are given $P_{max}, Q_{max} \in \mathbb{N}^+$ such that the orders of $\{X_t^{(1)}\}, \{X_t^{(2)}\} \in \mathcal{L}$ ARMA process the $p_1, p_2 < P_{max}$ and $q_1, q_2 < Q_{max}$, then under stationarity condition the $\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2)$ and $\widehat{d}_{PIC}(\mathbf{x}_1, X^{(1)})$ distance estimators are weakly consistent*

$$\widehat{d}_{PIC}(\mathbf{x}_1, \mathbf{x}_2) \xrightarrow[n \to \infty]{P} d_{PIC}(X^{(1)}, X^{(2)})$$
$$\widehat{d}_{PIC}(\mathbf{x}_1, X^{(2)}) \xrightarrow[n_1 \to \infty]{P} d_{PIC}(X^{(1)}, X^{(2)})$$

It is a noteworthy observation that for any $X^{(i)}, X^{(j)} \in \mathcal{L}$ and $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}$ the distance d_{PIC} and their empirical estimate $\widehat{d}_{PIC}$ satisfy the triangle equations.

$$
\begin{aligned}
&d_{PIC}\left(X^{(i)},X^{(j)}\right)\leq \widehat{d}_{PIC}\left(X^{(i)},\mathbf{x}_i\right)+\widehat{d}_{PIC}\left(\mathbf{x}_i,X^{(j)}\right)\\
&\widehat{d}_{PIC}\left(\mathbf{x}_i,X^{(i)}\right)\leq \widehat{d}_{PIC}\left(\mathbf{x}_i,\mathbf{x}_j\right)+\widehat{d}_{PIC}\left(\mathbf{x}_j,X^{(i)}\right)\\
&\widehat{d}_{PIC}\left(\mathbf{x}_i,\mathbf{x}_j\right)\leq \widehat{d}_{PIC}\left(\mathbf{x}_i,X^{(i)}\right)+\widehat{d}_{PIC}\left(\mathbf{x}_j,X^{(i)}\right)
\end{aligned}
\tag{3.1.12}
$$

Algorithm 1 Clustering ARMA models

Require: $\mathcal{D}$, κ, (P_{max}, Q_{max})
Estimate models and model parameters
for $i = 1..N$ **do**
    $\widehat{m}^i, \widehat{\beta}^i \leftarrow$ *estimate model parameters*
end for
Initialize κ-farthest points as cluster-centres:
$c_1 \leftarrow 1$
$C_1 \leftarrow \{c_1\}$
for $k = 2..\kappa$ **do**
    $c_k \leftarrow \underset{i=1.N}{\operatorname{argmax}} \min_{j=1..k-1} \widehat{d}\left(\mathbf{x}_i, \mathbf{x}_{c_j}\right)$
    $C_k \leftarrow \{c_k\}$
end for
Assign the remaining points to closest centres:
for $i = 1..N$ **do**
    $k \leftarrow \operatorname{argmin}_{j \in \bigcup_{k=1}^{\kappa} C_k} \widehat{d}\left(\mathbf{x}_i, \mathbf{x}_j\right)$
    $C_k \leftarrow C_k \cup \{i\}$
end for
OUTPUT: clusters $C_1, C_2, ..., C_k$

We want to mention that notations in Algorithm 1 are provided for the general case of model-based clustering since the model estimation and the distance estimators can be different for different problem configurations. If the underlying model orders are known then the model parameters are estimated with QML (3.1.10) and for the model selection the BIC penalized QMLE. We start by formulating the theorem of the strong consistency of the Algorithm 1.

Theorem 3.1.1 (Strong consistency of Algorithm 1)**.** *Assuming that the orders of all underlying ARMA processes are the same and known. Then if the target number of clusters κ is known, then Algorithm 1 is strongly asymptotically consistent.*

The proof of the theorem is the same as Theorem 11 in the [2] since the defined distance estimators are strongly consistent (Proposition 3.1.3) and satisfy the triangle inequalities (3.1.12). The following theorem is based on Proposition 3.1.4 and provides a more general framework for clustering ARMA processes.

Theorem 3.1.2. *(Weak consistency of Algorithm 1) Assuming that there exists (P_{max}, Q_{max}) such that orders of all underlying ARMA processes are less than P_{max} and Q_{max}, and the target number of clusters κ are known, then Algorithm 1 is weakly asymptotically consistent. Moreover, for the given $\eta \in (0,1)$ there exists n, such that if $n_{\min} = \min_{i\in 1..N} n_i > n$, then*

$$P(f((\mathcal{D},\kappa)) = \mathcal{G}) \geq (1-(N-\kappa)(4-4\eta))(4\eta-3)^{\kappa-1}$$

In the **section 3.2**, we discussed consistent clustering of the time series dataset generated by the invertible GARCH(p,q). If the operator $(1-\beta(B))^{-1}$ exists, then we have the so-called ARCH(∞) representation of the GARCH(p,q) process ([6], [7]).

$$\sigma_t^2 = \psi_0 + \sum_{i=1}^{\infty} \psi_i \epsilon_{t-i}^2 \tag{3.2.1}$$

where

$$\psi_0 = \frac{\omega}{1-\sum_{j=1}^{q} \beta_j} \tag{3.2.2}$$

and coefficients ψ_i are the coefficients of the characteristic polynomial of the $(1-\beta(B))^{-1}\alpha(B)$ and can be determined with the following recursive equations [6].

$$\psi_i = \alpha_i + \sum_{j=1}^{n^*} \beta_j \psi_{i-j} \tag{3.2.3}$$

where $n^* = \min\{p, i-1\}$, $\beta_i = 0 \quad i > p$ and $\alpha_i = 0 \quad i > q$. Having (3.2.1) representation of the GARCH(p,q) process we define a metric on $\mathcal{U}$ as follows. Let $\{X_t\}$ and $\{Y_t\}$ are two stationary, invertible GARCH processes and $\Psi_X = \{\psi_{i,X}\}_{i=0}^{\infty}$ and $\Psi_Y = \{\psi_{i,Y}\}_{i=0}^{\infty}$ are the corresponding sequences of $\{X_t\}$ and $\{Y_t\}$ obtained from the equations (3.2.2) and (3.2.3). Then

$$d(X_t, Y_t) = \left\{ \sum_{j=0}^{\infty} (\psi_{j,X} - \psi_{j,Y})^2 \right\}^{1/2} \tag{3.2.4}$$

Under the same definitions of asymptotically consistent clustering, the empirical estimate of the metric (3.2.4) all the results discussed in the previous sections can be established also for the time series dataset generated by GARCH(p,q) processes since the exponential decrease of the coefficients ψ_i and the consistent model selection for GARCH(p,q) processes discussed in Proposition 1.5.2 in the thesis.

Another interesting extension of the presented results is a case of a clustering time series dataset generated by ARMA(p, q) models with GARCH(p', q') errors. This problem is discussed in the **section 3.3**. We denote the class of ARMA(p, q)-GARCH(p', q') processes with invertible ARMA and GARCH components as $\mathcal{LU}$. As previously explained for the process $X_t \in \mathcal{LU}$, we can derive two infinite sequences from the $AR(\infty)$ and $ARCH(\infty)$ representations, which fully characterize the model X_t. For given $X_t \in \mathcal{LU}$, and positive constants u and v (where $u+v=1$), we define the norm of the processes X_t in $\mathcal{LU}$ as follows.

$$||X_t||_{u,v} = u||\boldsymbol{\pi}||_2 + v||\boldsymbol{\psi}||_2 \tag{3.3.1}$$

And the distance between two processes $\{X_t\}, \{Y_t\} \in \mathcal{LU}$ is defined using the $||\cdot||_{u,v}$ norm.

$$d(X_t, Y_t) = u \left\{ \sum_{j=0}^{\infty} (\pi_{j,X} - \pi_{j,Y})^2 \right\}^{1/2} + v \left\{ \sum_{j=0}^{\infty} (\psi_{j,X} - \psi_{j,Y})^2 \right\}^{1/2} \tag{3.3.2}$$

Since the consistent estimation of the QMLE and BIC penalized QMLE of the ARMA-GARCH processes and the continuity of the metric (3.3.2) from the parameters of the considerable models it is easy to see that all the previously discussed results are true for the time series datasets generated by the ARMA-GARCH models.

In the **section 3.5** we discuss the theoretical similarities and differences between the proposed methods with the existing model-based approaches. The key points are summarized here.

- We established conditions for an asymptotically consistent clustering algorithm for the common time series models.
- Discussed method does not require conditional independence of samples.
- We propose clustering GARCH processes based on their ARCH(∞) representation, avoiding assumptions on the invertibility of certain polynomials.
- The consistency of the provided algorithm is obtained with the QMLE, which will keep it consistent even if the Gaussian assumption is not perfectly met, though it may be less efficient than MLE under certain conditions.

Section 3.6 is dedicated to some important considerations that need to be taken into account for the practical implementation of the discussed methods. Here we want to note that although the Algorithm 1 is asymptotically consistent, other algorithms can also be implemented. This is achieved by pre-computing the distance matrix $C = [\widehat{d}(\mathbf{x}_i, \mathbf{x}_j)]_{i,j}$ for all sample pairs $(\mathbf{x}_i, \mathbf{x}_j)$, totaling $N(N-1)/2$ distances, and then employing clustering algorithms specifically designed to operate on distance matrices. A commonly selected option is the K-Medoids algorithm [8].

In **chapter 4**, we demonstrated the results of several numerical experiments and practical applications of the methods discussed. We start with an experimental comparison of model-free algorithms for clustering time series datasets generated by GARCH processes. Motivated by [9], for comparison we choose well-known partition-based time series clustering models: K-Means, K-Means with dynamic time warping and DTW barycenter averaging, K-Shape and Kernel K-Means models. Furthermore, we can find open-source implementations of these algorithms [10].

To evaluate non-parametric models, we simulate random datasets with different setups. In the first experiment, we measure the ability of the models to cluster different numbers of clusters. In Table 4.1.1, we present the results of the first experiment evaluated with the AMI metric. We can see that the KM-DTW model outperforms other models.

In the second experiment, we measure the asymptotic consistency of the discussed models. We generate datasets with 5 clusters and 100 samples in each cluster. We

κ	KM-E	KM-DTW	k-Shape	KKM-GAK
2	0.003+-0.001	0.325+-0.403	0.004+-0.009	0.003+-0.002
4	0.004+-0.001	0.463+-0.129	0.02+-0.007	0.002+-0.001
6	0.018+-0.016	0.578+-0.151	0.043+-0.021	0.001+-0.0005
8	0.006+-0.003	0.498+-0.077	0.005+-0.011	0.001+-0.0005
10	0.005+-0.01	0.624+-0.03	0.062+-0.022	0.0001+-0.00005

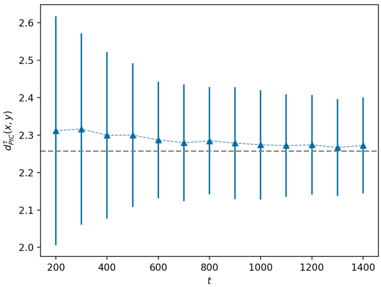
Table 4.1.1: AMI score for different κ

set $T = 1000$ and consider 5 intervals on the time axis. We train and evaluate models in the first interval and consequently add a number of samples. From the second experiment, we observed that the KM-DTW model outperforms other models, but we do not observe significant asymptotic patterns.

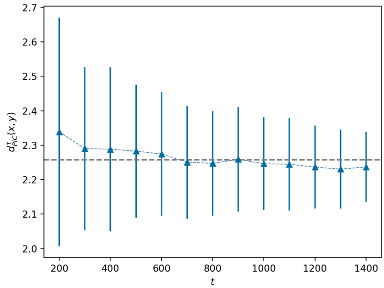
The **section 4.2** is focused on the evaluation of proposed methods with the already discussed KM-DTW model. We start this section by providing experimental results for the asymptotic behavior of the distance estimates. Conducted experiments show that the proposed estimates show desired asymptotic properties. To test the convergence we conduct a two-way Wilcoxon signed-rank test with the null hypothesis as the estimated and the real values are different. The rates of minimal time sample sizes for the convergence for the given examples are also highlighted in the thesis. The results for the experiment with the d_{PIC} and its estimator are presented in Figure 4.2.1. The vertical dotted line presents the true value of $d^{\tau}_{PIC}(X^{(1)}, X^{(2)})$.

Taking into account results from the previous section from the model-free algorithms we will consider only the KM-DTW model as it outperforms all other considered methods. From the mentioned above we will evaluate the Algorithm 1, K-Medoids, and KM-DTW methods on randomly randomly generated models. To generate the presented models we fix the $P_{max} = Q_{max} = 3$, and generate 5 (number of clusters κ) different model orders and parameters. In each step, we ensure that the generated parameters satisfy stationarity and invertibility conditions.

To evaluate the asymptotically consistent properties of the clustering algorithms, we generally follow the same expanding window approach described earlier. For each



(a) Estimation of $d^{\tau}_{PIC}(X^{(1)}, X^{(2)})$ with known model orders.



(b) Estimation of $d^{\tau}_{PIC}(X^{(1)}, X^{(2)})$ with unknown model orders.

Figure 4.2.1: Estimation of $d^{\tau}_{PIC}(X^{(1)}, X^{(2)})$.

underlying model, we generate 50 realizations with 1000 samples each. We set up, the 200-step size and in each step fit all 3 models and evaluate clustering results with an Adjusted Mutual Information score. The described process is repeated 10 times for averaging purposes, and the evaluation results are presented in Figure 4.2.3.

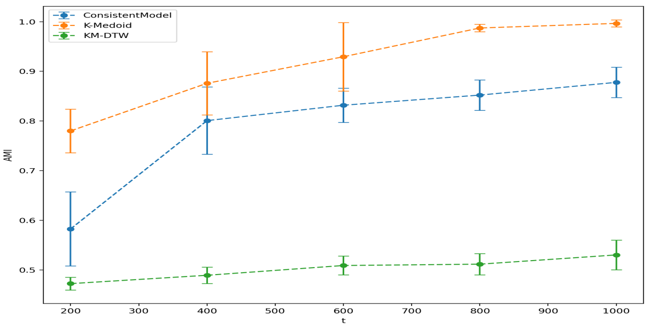


Figure 4.2.3: Asymptotic properties of Algorithm 1.

From Figure 4.2.3 it can be observed that Algorithm 1 and K-Medoids models are outperforming the KM-DTW model in every time window. The clearly increasing AMI score for both models are indicators of the asymptotic consistent properties of the models.

To ensure the robustness of the general comparison procedure, we did not fix the specific model structures in the data generation procedure, therefor underlying model parameters and the realizations are selected randomly. We fix $\delta > 1$, $P_{max} = Q_{max} = 3$, and the number of samples in each realization to 1000. We also range the number of clusters κ from 2 to 6, to show the dependence of algorithms on the number of clusters. In each step, we generate random processes according to the δ, invertibility, and stationarity constraints, and for each process, we generate 100 time series each having 1000 samples. We divide the data set into training and testing datasets with a 0.25 test ratio and evaluate the results on the test dataset with the AMI metric. The results of this experiment are provided in Table 4.2.1. For general evaluation, we generate GARCH(p, q) processes with $P_{max} = Q_{max} = 3$, $\delta > 1$, and $n_{min} = 2000$ since we previously observed that the distance between GARCH processes is converging with higher values than in the case of ARMA processes. Even more, we observed that for the smaller values ($n_{min} < 500$) the estimated GARCH processes were nearly nonstationary, and $\psi_0 = \frac{\omega}{1-\sum_{j=1}^{q}\beta_j}$ vanishes.

A similar experiment is conducted for general evaluation of the provided methods to cluster time series datasets generated by ARMA-GARCH processes. For the case of ARMA-GARCH processes, we limit with $\kappa = 4$ for the reason mentioned earlier. As can see in Table 4.2.1 Algorithm 1 inherits the same issues from the case of GARCH processes and we do not observe significant improvements of AMI metric from the baseline model KM-DTW. In this case also K-Medoids model outperforms the rest of the approaches.

As mentioned earlier, the consistency of the discussed algorithms is proven to have the fact that the consistent distance estimators are providing the conditions that the time series dataset is satisfying so-called strict separability conditions. To measure the effect of the separability assumption, the following experiment is conducted. We fix $\kappa = 2$ and for each model ARMA, GARCH, and ARMA-GARCH we generate underlying models that have a predefined distance δ measured with the corresponding distances. Then the δ is increased during the experiment. The results of the experiment are shown in Table 4.2.2. From Table 4.2.2 that we can observe that algorithms are dependent on the separability condition and for the lower level of

Process	κ	Algorithm 1	K-Medoid	KM-DTW
ARMA	2	1.0 ± 0.0	1.0 ± 0.0	0.48 ± 0.45
	4	0.9 ± 0.093	0.99 ± 0.02	0.51 ± 0.035
	6	0.85 ± 0.11	0.93 ± 0.054	0.53 ± 0.1
GARCH	2	0.95 ± 0.06	0.92 ± 0.11	0.8 ± 0.4
	4	0.74 ± 0.06	0.96 ± 0.05	0.51 ± 0.21
	6	-	-	-
ARMA-GARCH	2	1.0 ± 0	1.0 ± 0	0.4 ± 0.49
	4	0.66 ± 0.35	1.0 ± 0	0.66 ± 0.9
	6	-	-	-

Table 4.2.1: AMI score for clustering processes with different κ values.

separability the clustering results are poor. Despite this fact, the clustering results are reliable from $\delta > 0.4$.

In the **section 4.3** we show the practical applicability of the proposed methods, for clustering and analyzing the structure of the foreign exchange market. This section includes a detailed discussion about the considered dataset, methodology, and obtained results. We download the daily exchange rates from 2002-01-01 to 2023-01-01 via APILayer[1] API, which includes the 44 top-traded currencies. The list of currencies included in the analysis is available in Table 4.4. To analyze the dynamic structure of the FX market, we divide the time period from 2002-01-01 to 2023-01-01 into equal time periods and perform the model estimation and clustering procedure in each interval. The selection of a long time period will incorporate more information in time series but, having the fact that currencies generally speaking have dynamic structure and regime changes, it is possible to skip important changes in market structure. On the other hand, the smaller time periods can affect model estimation resulting in a poorly estimated distance matrix. Having this consideration we examined clustering in different time periods and chose the clustering period of 6 months (totaling 42 clustering periods) as a good trade-off between consistent model estimation and the possibility to examine the dynamic structure of the FX market.

The analysis of the dynamic behavior of the FX market is done, by clustering

[1] `https://apilayer.com/marketplace/exchangerates_data-api`

Process	δ range	Algorithm 1	K-Medoid
ARMA	$0 < \delta < 0.2$	0.49 ± 0.37	0.67 ± 0.21
	$0.2 < \delta < 0.4$	0.93 ± 0.13	0.93 ± 0.09
	$0.4 < \delta < 0.6$	0.91 ± 0.17	1 ± 0
	$0.6 < \delta < 0.8$	0.82 ± 0.25	0.85 ± 0.25
	$0.8 < \delta < 1$	1 ± 0	1 ± 0
GARCH	$0 < \delta < 0.2$	0.23 ± 0.24	0.27 ± 0.18
	$0.2 < \delta < 0.4$	0.77 ± 0.19	0.88 ± 0.12
	$0.4 < \delta < 0.6$	0.49 ± 0.33	0.82 ± 0.22
	$0.6 < \delta < 0.8$	0.81 ± 0.13	0.83 ± 0.09
	$0.8 < \delta < 1$	0.8 ± 0.18	0.93 ± 0.09
ARMA-GARCH	$0 < \delta < 0.2$	0.64 ± 0.44	0.7 ± 0.46
	$0.2 < \delta < 0.4$	0.78 ± 0.36	0.87 ± 0.26
	$0.4 < \delta < 0.6$	0.59 ± 0.48	0.99 ± 0.04
	$0.6 < \delta < 0.8$	0.5 ± 0.5	0.93 ± 0.11
	$0.8 < \delta < 1$	0.8 ± 0.4	0.96 ± 0.09

Table 4.2.2: AMI score for clustering processes with different δ values.

different time periods and comparing clustering results. In each time interval, we estimate the underlying processes and compute the distance matrix D. Due to the dynamic structure of the FX market, it is natural to assume that, the number of clusters can change over time, and we need to choose the number of clusters in each time interval. As a method for selecting a number of clusters, we choose the Silhouette [11] method, which does not make strong assumptions about the data-generating process and composes both interpretability and balance between cohesion and separation. Silhouette score is between 1 and -1, where higher scores indicate better-defined clusters. In each time interval and for all clustering methods we choose the number of clusters, restricting the maximum number of clusters to 10. To investigate the FX market dynamics, we compare the clustering results in each time period with is previous and next clusters. In our method, we compare clusters with the Adjusted Mutual Information score [12]. Figure 4.3.1 represents the maximum Silhouette score for the clustering models and the number of estimated clusters. We can see that the calculated median Silhouette score is slightly higher for the K-Medoids model and the ConsistentModel estimate a higher number of clusters for the same periods.

In Table 6.1 we present clustering results for the K-Medoids model for all time periods. Despite the fact that at first glance it is difficult to find an obvious interpretation for the resulting clusters, we should note that they reflect a number of important features that characterize the market. The key findings are listed below

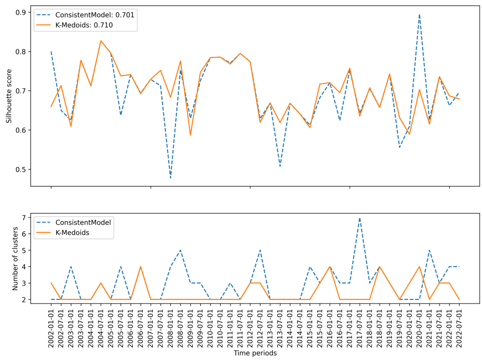


Figure 4.3.1: Silhouette score for different time periods and models. In legend showed the median values of Silhouette scores.

1. Fixed exchange currencies mostly appear in the same cluster. Well-known currencies from the Middle East are considered as peng currencies with USD. The AED and USD are clustered in the same cluster in 69% and USD and BHD are clustered together in 88% of clustering periods.

2. The resulting clusters reflect the relationships of currencies circulating in the same geographical region. For example, EUR, CHF, and GBP are appearing in the same clusterin in 71%, and PKR, INR, and IRR in 81% of clustering periods.

3. Clusters reflect economic associations between countries. For example, Armenia and the Russian Republic are members of the Commonwealth of Independent States and Eurasian Economic Union, the RUB and AMD currencies have appeared in the same cluster in 95% of the clustering period.

4. Clusters reflect the industrial connections between currencies (countries). For example, the examined oil-based currencies RUB and COP are clustered in the same

cluster in 92% of clustering periods.

To illustrate FX market dynamics, for each model, we compare clustering for each time period t with the next $t+1$ time period clustering results. Figure 4.3.3 presents the comparison results, done by external validity measures Adjusted Mutual Information for both models.

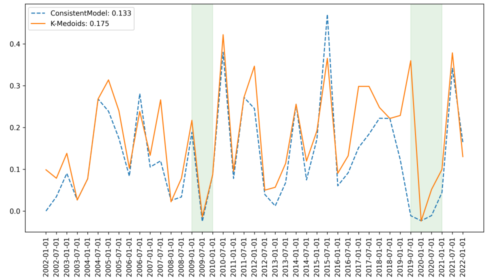


Figure 4.3.3: AMI for different time periods

The higher values can indicate stable market periods, and lower values indicate dramatic changes in market structure. The mean values of comparison metrics can indicate that clusters of FX market generated from the described methodology have dynamic structure. Although the two models show different behaviors in general, we can notice that they strongly agree with the sharp changes in the market. Since the maximum values of AMI for both models over 2002-2023 are smaller than 0.47, we can say that in each 6-month period, clusters of both models are changed by at least 0.47, compared with AMI metrics.

From Figure 4.3.3, we can also observe the periods for which the consecutive clusters are highly dissimilar, which can indicate the high changes in the market structure. The first major drop in comparison metric can be noticed from 2008-01-01 to 2009-07-01. This period coincides with the Global Financial Crisis. The second major dropdown in AMI is observed from 2019-07-01 to 2021-07-01, which coincides with the global economic crisis caused by the COVID-19 pandemic.

The estimated distance matrices can also act as a useful method to analyze dynamic relationships between sets of currencies. As an example, we examine the dynamic behavior of the RUB currency. In Figure 4.3.4, we present the estimated

Figure 4.3.4: Estimated distances between RUB and USD, and RUB and COP.

distance in the all-time period between RUB, USD, and COP, where the COP currency is chosen as a representative of oil-based currencies. First of all, let us note that the RUB currency, on average, is closer to the COP currency, which has a natural explanation considering the dependence of the RUB foreign currency on oil prices. The second notable observation is that since 2022-07-01, the RUB currency shows an anomalous behavior, significantly departing not only from the USD but also from the COP currency observed in the same oil-based sector. This effect can be interpreted as an impact of the Russian-Ukrainian war on the RUB currency.

Bibliography

[1] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering – a decade review," *Information Systems*, vol. 53, p. 16–38, 2015.

[2] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux., "Consistent algorithms for clustering time series," *Journal of Machine Learning Research*, vol. 17(3), p. 1–32, 2016.

[3] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting.* 2002.

[4] M. Corduas and D. Piccolo, "Time series clustering and classification by the autoregressive metric.," *Computational Statistics Data Analysis*, vol. 52(4), p. 1860–1872, 2008.

[5] D. Piccolo, "A distance measure for classifying arima models.," *Journal of Time Series Analysis*, vol. 11(2), p. 153–164, 1990.

[6] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, p. 307–327, 1986.

[7] C. Francq and J.-M. Zakoian, *GARCH models: structure, statistical inference and financial applications.* John Wiley & Sons, 2019.

[8] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, p. 3336–3341, 2009.

[9] A. Javed, B. S. Lee, and D. M. Rizzo, "A benchmark study on time series clustering," *Machine Learning with Applications*, vol. 1, p. 100001, 2020.

[10] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods, "Tslearn, a machine learning toolkit for time series data," *Journal of Machine Learning Research*, vol. 21, no. 118, pp. 1–6, 2020.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, p. 53–65, 1987.

[12] S. Romano, N. the Vinh, J. C. Bailey, and K. M. Verspoor, "Adjusting for chance clustering comparison measures," *Journal of Machine Learning (JMLR) 17(1)*, pp. 4635–4666, 2016.

Ամփոփում

Ատենախոսության նպատակն է ուսումնասիրել մոդելների վրա հիմված ժամանակային շարքերի տվյալների կլաստերիզացիայի ասիմպտոտիկորեն կայուն ալգորիթմները և դրանց կիրառությունները: Իրենց չվերահսկվող բնույթից ելնելով ժամանակային շարքերի կլաստերիզացիայի ալգորիթմներն ունեն լայն կիրառություններ բազմաթիվ ոլորտներում:

Գլուխ 1-ը նվիրված է թեզում ուսումնասիրվող հիմնական ժամանակային շարքերի մոդելների և դրանց հետ առնչվող հասկացությունների սահմանումներին: Այս գլխում նաև քննարկվում են ժամանակային շարքերի պարամետրերի գնահատականների ասիմպտոտիկորեն կայուն մեթոդները:

Գլուխ 2-ում քննարկվում են ժամանակային շարքերի կլաստերիզացիայի առկա ալգորիթմները, դրանց ճշգրտության գնահատականների առկա մեթոդները և այդ ալգորիթմների մի շարք կիրառություններ:

Գլուխ 3-ը նվիրված է ատենախոսության հիմնական տեսական արդյունքներին: Մասնավորապես, **բաժին 3.1 -ում** սահմանված է ARMA պրոցեսներով գեներացված ժամանակային շարքերի հավաքածուների ասիմպտոտիկորեն կայուն կլաստերիզացիայի խնդիրը: Ցույց է տրվել հակադարձելի ARMA(p,q) պրոցեսների վրա սահմանված d_{PIC} մետրիկայի ասիմպտոտիկորեն կայուն գնահատակների գոյությունը՝ քննարկելով ARMA(p,q) պրոցեսների կարգերի հայտնի կամ անհայտ լինելու դեպքերը: Ունենալով վերը քննարկված գնահատակները, ցույց է տրված Ալգորիթմ 1.-ի ասիմպտոտիկորեն կայունությունը ARMA պրոցեսներով գեներացված ժամանակային շարքերի կլաստերիզացիայի խնդրի երկու կոնֆիգուրացիաների համար: Առաջին դեպքում, երբ տվյալների գեներացման հիմնական պրոցեսների կարգերը հայտնի են և նույնն են, ցույց է տրվել Ալգորիթմ 1-ի ուժեղ կայունությունը՝ ենթադրելով, որ իրական կլաստերների թիվը հայտնի է: Երկրորդ դեպքում, մենք ենթադրում ենք, որ տվյալների գեներացման հիմնական պրոցեսների կարգերը անհայտ են, և հայտնի են միայն դրանց վերին սահմանները: Ունենալով d_{PIC}-ի էմպիրիկ գնահատման թույլ կայունությունը՝ մենք ապացուցում ենք Ալգորիթմ 1-ի թույլ ասիմպտոտիկ կայունությունը, ենթադրելով, որ իրական կլաստերների թիվը հայտնի է: **Բաժին 3.2-ից 3.4-ում** մենք ցույց ենք տալիս, թե ինչպես քննարկված մեթոդները կարող են ընդհանրացվել պատահական պրոցեսների ավելի մեծ դասերի վրա, ինչպիսիք են GARCH(p,q), ARMA(p,q)-GARCH(p',q') և ARIMA(p,d,q):

Գլուխ 4-ում ներկայացված են թեզում առաջարկվող մեթոդների թվային համեմատությունները և կիրառությունները: Թվային մեթոդներով ցույց է տրվել, որ թեզում առաջարկվող մեթոդները գերազանցում են համեմատվող ոչ պարամետրիկ մեթոդներին քննարկված ժամակային շարքերի կլաստերիզացիայի և ասիմպտոտիկորեն կայուն կլաստերիզացիայի խնդիրներում: Աշխատանքը նաև ներառում է ստացված տեսական արդյունքների մի շարք կիրառություններ, մասնավորապես՝ արտարժույթի շուկայի կլաստերիզացիայի և ստացված կլաստերների դինամիկայի ուսումնասիրություն: Ստացված կլաստերները ընդգծում են արտարժույթի շուկայի մի քանի հիմնական հատկություններ.

1. Ֆիքսված փոխարժեքով արժույթները սովորաբար հայտնվում են նույն կլաստերում:
2. Կլաստերները արտացոլում են նույն աշխարհագրական տարածքում գործող արժույթների փոխհարաբերությունները:
3. Կլաստերները ցույց են տալիս երկրների միջև տնտեսական կապերը:
4. Կլաստերները ցույց են տալիս արժույթների միջև արդյունաբերական կապերը:

Ստացված կլաստերների հիման վրա առաջարկվել է արտարժույթի շուկայի կայունության վերլուծության մեթոդ, որը հիմնված է իրար հաջորդող ժամանակաշրջանների համար ստացված կլաստերների համեմատության վրա:

Резюме

В данной диссертации мы исследуем проблему асимптотически согласованной кластеризации наборов данных временных рядов, сгенерированных на основе моделей.

Глава 1 посвящена определениям основных моделей временных рядов и связанных с ними понятий, изучаемых в диссертации. В этой главе также обсуждаются асимптотически устойчивые методы оценки параметров временных рядов.

В **главе 2** рассматриваются существующие алгоритмы кластеризации временных рядов, методы оценки точности кластеризации, и приложения этих алгоритмов.

В **главе 3** Мы исследуем меры несходства, определенные на пространствах обратимых ARMA процессов. Мы рассматриваем эмпирические оценки обсуждаемых метрик и демонстрируем их асимптотическую состоятельность. Используя эти оценки, мы исследуем Алгоритм 1 и показываем его асимптотическую согласованность для двух конфигураций задачи. В первом сценарии, когда порядки основных процессов ARMA одинаковы и известны, мы показываем строгую согласованность Алгоритма 1, предполагая, что целевое количество кластеров известно. Во второй постановке задачи мы предполагаем, что основные процессы неизвестны и известны только верхние пределы порядков моделей. Имея слабую состоятельность эмпирической оценки для неизвестных моделей, мы доказываем слабо асимптотически согласованность Алгоритма 1 для второй конфигурации задачи, предполагая, что заданное число кластеров известно. Мы показываем, как обсуждаемые методы можно распространить на другие, более крупные классы случайных процессов, такие как GARCH(p, q), ARMA(p, q)-GARCH(p', q') и ARIMA(p, d, q). Мы также анализируем теоретические и практические вопросы обсуждаемой структуры и даем практические и теоретические предложения.

Чтобы оценить предлагаемие методы кластеризации, мы провели несколько экспериментов, чтобы показать как точность кластеризации, так и асимптотическую согласованность обсуждаемых методов. Эти и другие вопросы касаемые практических проблем предлагаемых методов обсуждаются в **главе 4.** Предложенные методы превосходят непараметрические методы во всех рассмотренных экспериментах. В качестве приложения мы применили предлагаемие методы кластеризации к реальному набору данных для кластеризации и анализа структуры валютного рынка. Полученные кластеры подчеркивают несколько ключевых атрибутов валютного рынка:

1. Валюты с фиксированным обменным курсом обычно появляются в одном кластере.
2. Кластеры отражают взаимоотношения валют, действующих в одном географическом регионе.
3. Кластеры обозначают экономические связи между странами.
4. Кластеры показывают промышленные связи между валютами.

Анализ устойчивости конфигурации валютного рынка осуществляется путем сравнения кластеров для каждого последующего временного промежутка. Мы продемонстрировали, что динамический обзор валютного рынка также может служить полезным инструментом для анализа влияния значительных экономических событий на структуру рынка.