CSIT Conference 2023, Yerevan, Armenia, September 25 - 30

Consistent Clustering of ARMA-GARCH Processes

Garik Adamyan
Yerevan State University
Yerevan, Armenia
e-mail: garik.adamyan@ysu.am

***Abstract*—This paper explores the issue of achieving asymptotically consistent clustering for time series data generated by ARMA-GARCH processes. We establish a metric in the space of invertible ARMA-GARCH processes and outline a consistent estimation method for this metric. Subsequently, we employ this metric to demonstrate the asymptotic consistency of the discussed algorithm.**

***Keywords*— time series clustering, ARMA-GARCH process, asymptotically consistency**

I. Introduction

In this paper, we consider the problem of asymptotically consistent clustering of time series datasets generated by ARMA-GARCH processes. Let $\{e_t\}$ be iid noise with $\mathbb{E}[e_t] = 0$ and $\mathbb{E}[e_t^2] = 1$ then, the ARMA(p, q)-GARCH(p', q') model is defined as follows.

Definition 1: $\{X_t\}$ random process is an ARMA(p, q)-GARCH(p', q') process if $\{X_t\}$ is weakly stationary and if for every t the following equations hold.

$$\begin{cases} \phi(B)X_t = \theta(B)\epsilon_t \\ \epsilon_t = \sigma_t e_t \\ (1-\beta(B))\sigma_t^2 = \omega + \alpha(B)\epsilon_t^2 \end{cases} \tag{1}$$

where

$$\omega > 0$$
$$\alpha_i \geq 0, i = 1, 2, ..., p$$
$$\beta_j \geq 0, i = 1, 2, ..., q$$

and polynomials $\phi(B)$, $\theta(z)$, $\beta(B)$, $\alpha(B)$ are characteristic polynomials of the corresponding lag polynomials.

Inspired by [1] and [2], we define the ground truth clusters and consistent clustering of the time series data generated by ARMA-GARCH processes as follows. 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-GARCH processes. We denote by $X^{(k)}$ the underlying ARMA-GARCH process for the cluster $\mathcal{G}_k$. The time series samples may have arbitrary lengths, and we denote the length of $\mathbf{x}_i$ time series by n_i.

Definition 2 (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 $\mathbf{x}_i$, $i = 1, 2, ..., N$ is generated by $X^{(k)}$ for some $k = 1, 2, ..., \kappa$ if and only if $i \in \mathcal{G}_k$. We call $\mathcal{G}$ a ground-truth clustering.

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 $\{1, 2, ..., N\}$.

Definition 3 (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 $\lim_{n\to\infty} P(f(\mathcal{D}, \kappa) = \mathcal{G}) = 1$

If the roots of the polynomial $\beta(z)$ are outside the unit circle then the operator $(1-\beta(B))^{-1}$ exists and we have the so-called ARCH(∞) representation of the GARCH part of the (1) process [3].

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

where

$$\psi_0 = \frac{\omega}{1 - \sum_{j=1}^q \beta_j} \tag{3}$$

and coefficients ψ_i are the coefficients of the characteristic polynomial of the $(1-\beta(B))^{-1}\alpha(B)$ which can be determined with the following recursive equations [3].

$$\psi_i = \begin{cases} \alpha_i + \sum_{j=1}^{n^*} \beta_j \psi_{i-j}, & if \quad i \leq q \\ \sum_{j=1}^{n^*} \beta_j \psi_{i-j}, & if \quad i > q \end{cases} \tag{4}$$

where $n^* = \min\{p, i-1\}$, $\beta_i = 0 \quad i > p$ and $\alpha_i = 0 \quad i > q$.

If the roots of the polynomial $\theta(z)$ are outside the unit circle then it ensures the existence of the operator $\theta(B)^{-1}$ and we can obtain an analogical representation for the ARMA part of $\{X_t\}$ (so-called AR(∞) representation).

$$\pi(B)X_t = \epsilon_t \tag{5}$$

where $\pi(B) = \theta(B)^{-1} * \phi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$. The coefficients of the sequence $\boldsymbol{\pi}_x$ are determined by the following recursive equations ([4]: p. 86):

$$\pi_j + \sum_{k=1}^{q} \theta_k \pi_{j-k} = -\phi_j, \quad j = 0, 1, ... \tag{6}$$

where $\phi_0 := -1$, $\phi_j := 0$ for $j > p$, and $\pi_j := 0$ for $j < 0$. Let us denote the space of invertible ARMA-GARCH processes by $\mathcal{U}$.

https://doi.org/10.51408/csit2023_50

II. Metric On $\mathcal{U}$

Having (2) and (5) representation of the ARMA(p, q)-GARCH(p', q') process, we define a metric on $\mathcal{U}$ as follows. Let $\{X_t\}$ and $\{Y_t\}$ be two stationary ARMA(p, q)-GARCH(p', q') processes and $\Psi_X = \{\psi_{i,X}\}_{i=0}^{\infty}$, $\pi_X = \{\pi_{i,X}\}_{i=0}^{\infty}$ and $\Psi_Y = \{\psi_{i,Y}\}_{i=0}^{\infty}$, $\pi_Y = \{\pi_{i,Y}\}_{i=0}^{\infty}$ be the corresponding sequences of $\{X_t\}$ and $\{Y_t\}$ obtained from the ARCH(∞) and AR(∞) representations. Then, for positive constants u and v (where $u + v = 1$)

$$d(X_t, Y_t) = u||\pi_X - \pi_Y||_2 + v||\psi_X - \psi_Y||_2 \tag{7}$$

It is easy to see that (7) is a well-defined metric on $\mathcal{U}$ since the coefficients π_j, ψ_j are decreasing exponentially.

We are interested in an asymptotically consistent estimator for the metric (7). To construct such an estimator, we follow the methods described in [1] for the estimation of the analogical distance defined in the space of ARMA processes. Let $\mathbf{x}_i$ be a realization of stationary ARMA(p^*,q^*)-GARCH(p'^*,q'^*) process with an unknown parameter vector $\theta^* = (\phi_1^*, \ldots, \phi_p^*, \theta_1^*, \ldots, \theta_q^*, \omega^*, \beta_1^*, , ..., \beta_p^*, \alpha_1^*, ..., \alpha_q^*)$. Then, assuming that the orders $m^* = (p^*, q^*, p'^*, q'^*)$ are known, then it is well known that the Gaussian quasi-log-likelihood estimator of the θ (denoting by $\widehat{\theta}$) is strictly asymptotically consistent [5].

$$\widehat{\theta} \xrightarrow[n\to\infty]{a.s.} \theta^* \tag{8}$$

If the orders of the $\{X_t\}$ are unknown, then the model orders and parameters need to be estimated simultaneously. This is done by model selection procedures, for example, penalizing the QML with the suitable constrained. For the given $P_{max}, Q_{max}, P'_{max}, Q'_{max}$ model orders, consider the family of model orders $\mathcal{M} = \{(p, q, p', q') : p < P_{max}, \quad q < Q_{max}, \quad p' < P'_{max}, \quad q' < Q'_{max}\}$ such that the unknown model order $m^* \in \mathcal{M}$. In [5], authors showed that for the stationary ARMA(p^*,q^*)-GARCH(p'^*,q'^*) process estimate $\widehat{\theta}$ obtained via minimizing the BIC (Bayesian Information Criterion) penalized QML, is weakly asymptotically consistent [5].

$$P(\widehat{m} = m^*) \xrightarrow[n\to\infty]{} 1, \quad \widehat{\theta}[\widehat{m}] \xrightarrow[n\to\infty]{P} \theta^* \tag{9}$$

Suppose $\{X_t^{(1)}\}, \{X_t^{(2)}\} \in \mathcal{U}$ are two invertible ARMA-GARCH processes with true parameter vectors $\theta^{(1)}$ and $\theta^{(2)}$. The $\mathbf{x}_1 = \{x_1^1, x_2^1, .., x_{n_1}^1\}$ and $\mathbf{x}_2 = \{x_1^2, x_2^2, .., x_{n_2}^2\}$ are realisation of the $\{X_t^{(1)}\}, \{X_t^{(2)}\}$ processes. Then the empirical distance of metric (7) is defined as follows:

$$\widehat{d}(\mathbf{x}_1, \mathbf{x}_2) = u||\widehat{\pi}_X - \widehat{\pi}_Y||_2 + v||\widehat{\psi}_X - \widehat{\psi}_Y||_2 \tag{10}$$

We also define the estimate between time series sample $\mathbf{x}_i$ and random process $\{X_t^{(1)}\}$ as follows:

$$\widehat{d}(\mathbf{x}_1, X^{(1)}) = u||\widehat{\pi}_X - \pi_Y||_2 + v||\widehat{\psi}_X - \psi_Y||_2| \tag{11}$$

where $\{\widehat{\pi}_{i,j}\}_{j=0}^{\infty}$, $\{\widehat{\psi}_{i,j}\}_{j=0}^{\infty}$ are given by (3), (4) and (6) calculated with estimated parameters vector $\widehat{\theta}^{(i)}$ estimated by BIC penalized QML. Having discussed the consistent estimation procedure of the ARMA-GARCH model parameters, it is easy to formulate the following propositions.

Proposition 2.1: If the orders of the stationary $\{X_t^{(1)}\}, \{X_t^{(2)}\} \in \mathcal{U}$ ARMA-GARCH process are known, then the $\widehat{d}(\mathbf{x}_1, \mathbf{x}_2)$ and $\widehat{d}(\mathbf{x}_1, X^{(1)})$ distance estimators are strictly consistent

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

Proposition 2.1 ensures the almost sure convergence but the condition of the known orders is quite impractical. The following proposition is formalized with more relaxed assumptions leading to weak convergence of the distance estimator.

Proposition 2.2: If $P_{max}, Q_{max}, P'_{max}, Q'_{max}$ are given such that the model orders of stationary $\{X_t^{(1)}\}, \{X_t^{(2)}\} \in \mathcal{U}$ ARMA-GARCH process $m_1, m_2 \in \mathcal{M}$, then the $\widehat{d}(\mathbf{x}_1, \mathbf{x}_2)$ and $\widehat{d}(\mathbf{x}_1, X^{(1)})$ distance estimators are weakly consistent

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

The provided propositions are true because the estimation of the model parameters is consistent and the defined metric and estimators are continuous functions from the estimated parameters. We also note that metric (7) and the empirical estimates (10), (11) satisfy the following triangle equations.

$$d_{PIC}\left(X^{(i)}, X^{(j)}\right) \le \widehat{d}_{PIC}\left(X^{(i)}, \mathbf{x}_i\right) + \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(j)}\right)$$
$$\widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(i)}\right) \le \widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right) + \widehat{d}_{PIC}\left(\mathbf{x}_j, X^{(i)}\right)$$
$$\widehat{d}_{PIC}\left(\mathbf{x}_i, \mathbf{x}_j\right) \le \widehat{d}_{PIC}\left(\mathbf{x}_i, X^{(i)}\right) + \widehat{d}_{PIC}\left(\mathbf{x}_j, X^{(i)}\right)$$

III. Consistent clustering of ARMA-GARCH processes

As a clustering function(algorithm) we will use Algorithm 1. described in [1]. Algorithm 1 takes the time series dataset $\mathcal{D}$, a number of clusters κ, and a number of maximal order P_{max}, Q_{max} described in Proposition 2.2. Algorithm 1 is based on the following steps 1. For each sample $\mathbf{x}_i$ estimate ARMA-GARCH models with BIC penalized QML, choosing the appropriate model from $\mathcal{M} = \{(p, q) : p < P_{max}, \quad q < Q_{max}\}$ 2. For each estimated model compute the finite truncation of the infinite sequences Ψ_i using (3) and (4). 3. Initialize cluster centers with farthest point initialization. 4. For each $\mathbf{x}_i$, assign each label to the cluster of nearest centroid. The following theorem, which can be proved using Proposition 2.1 and similarly as in [2] shows the strong consistency of Algorithm 1.

Theorem 3.1 (Strong consistency of Algorithm 1): Assume that the orders of all underlying ARMA-GARCH processes are the same and known. Then if the target number of clusters κ is known, then Algorithm 1 is strongly asymptotically consistent.

The following theorem is based on Proposition 2.2 and provides a more general framework for clustering ARMA-GARCH processes.

Theorem 3.2 (Weak Consistency of Algorithm 1): Assume that there given $P_{max}, Q_{max}, P'_{max}, Q'_{max}$ such that orders of all underlying processes are $m_i \in \mathcal{M}$, 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

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

IV. Conclustion

In this work, we define the problem of asymptotically consistent clustering for time series data generated by the ARMA-GARCH processes. We define a metric in the space of invertible ARMA-GARCH processes using AR(∞) and ARCH(∞) representations of the ARMA-GARCH processes. We define consistent estimators for this metric using BIC-penalized QMLE. Finally, we employ this metric to demonstrate the asymptotic consistency of the discussed algorithm.

References

[1] G. Adamyan, "Weakly consistent offline clustering of arma processes," *Journal of Contemporary Mathematical Analysis*, vol. 58, no. 3, pp. 183–190, 2023.

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

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

[5] J.-M. Bardet, K. Kamila, and W. Kengne, "Consistent model selection criteria and goodness-of-fit test for common time series models," *Electronic Journal of Statistics*, vol. 14(1), 2020.