Statistics and Application 统计学与应用, 2021, 10(6), 1071-1082
Published Online December 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.106114

Hans汉斯

基于MS-GARCH模型的时间序列聚类

王　琳，丁孝全*

河南科技大学数学与统计学院，河南　洛阳

收稿日期：2021年11月27日；录用日期：2021年12月11日；发布日期：2021年12月27日

摘　要

聚类是时间序列数据挖掘的重要任务之一。本文基于有限混合MS-GARCH模型，提出一种时间序列聚类方法。利用贝叶斯马尔科夫链蒙特卡洛模拟方法，克服路径依赖的困难，给出了模型参数的估计。最后，选取23家中国上市公司股票数据进行实证分析，验证了所提方法的有效性。

关键词

时间序列聚类，有限混合模型，MCMC算法，MS-GARCH模型

Time Series Clustering with MS-GARCH Mixtures

Lin Wang, Xiaoquan Ding*

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang Henan

Received: Nov. 27th, 2021; accepted: Dec. 11th, 2021; published: Dec. 27th, 2021

Abstract

Clustering is one of the important tasks of time series data mining. In this paper, we propose a novel time series clustering method based on the finite mixture MS-GARCH model. By using Bayesian Markov chain Monte Carlo simulation methods to overcome the difficulty of full path dependence, we estimate the model parameters. Finally, the empirical analysis of stock data of 23 Chinese listed companies verifies the effectiveness of our proposed method.

Keywords

Time Series Clustering, Finite Mixture Model, MCMC Algorithm, MS-GARCH Model

*通讯作者。

文章引用: 王琳, 丁孝全. 基于 MS-GARCH 模型的时间序列聚类[J]. 统计学与应用, 2021, 10(6): 1071-1082.
DOI: 10.12677/sa.2021.106114

1. 引言

时间序列数据是一类非常重要的复杂数据，广泛存在于经济、金融、管理、医疗卫生、气象、水文和工程等领域。由于其高维、高冗余和随时间动态变化的特点，相比直接分析原始数据，合理利用前人建立的时间序列模型，在研究中更能容易刻画序列的变化规律。

广义自回归条件异方差(GARCH)模型是由 Bollerslev [1]在 Engle [2]研究的基础上对自回归条件异方差(ARCH)模型推广得到的。该模型以可变条件方差描述序列波动特征，条件异方差依赖于给定信息和滞后条件方差，具有更灵活的滞后结构，能更好地体现序列的记忆特征。1990 年，Lamoureux 和 Lastrages [3]指出条件方差可能存在结构转换，如果不考虑这种结构变化，可能造成波动持续性的错误估计。受其启发，Cai [4]和Hamilton，Sumsel [5]将Markov状态切换和ARCH效应结合，提出MS-ARCH模型；Gray [6]则进一步把马尔可夫状态切换引入 GARCH 模型，建立 MS-GARCH 模型。随后，许多学者对 MS-GARCH 模型做了不同的改进，并开展了深入研究[7]-[13]。

聚类是把一个数据对象的集合划分成若干簇(子集)，使簇内对象彼此相似、簇间对象不相似的过程。面对庞杂而日益增长的时间序列数据，聚类是时间序列数据挖掘的重要任务之一。近年来，涌现出许多时间序列聚类方法，大体上可以分为基于原始数据的聚类、基于数据特征的聚类和基于数据模型的聚类等三种[14] [15] [16]。

基于混合模型的聚类是一种有效、可解释、适应性强的统计分析方法[17] [18] [19]，它假定要聚类的数据来自一个有限混合的概率分布，通过计算每个数据在此概率分布下的后验概率，取后验概率最大的成分作为聚类结果。近年来，该方法也被广泛用于时间序列数据的聚类[20]-[27]。

在基于混合模型的聚类研究中，参数估计是关键工作，通常有期望最大化(EM)和Markov Chain Monte Carlo (MCMC)两种算法。EM算法基于极大似然估计，利用迭代优化估计模型的参数；MCMC算法则基于Bayes估计，利用MCMC模拟，解决了时间序列数据的高维、高冗余、路径依赖等不适用于EM算法的问题。研究表明，MCMC算法能很好地解决高维数据计算复杂的难题，拟合效果优于EM算法[12] [13] [28] [29] [30]。

本文基于有限混合 MS-GARCH 模型，利用 Bayes 理论的数据增补方法和 MCMC 模拟，探讨时间序列数据的聚类问题。文[21]利用有限混合 GARCH 模型对美国上市公司股票收益数据进行聚类，最终确定按波动率的高、中、低分为三类，但并没有考虑机制切换问题。就我们所知，未见文献把有限混合 MS-GARCH 模型用于时间序列数据的聚类分析。

2. 基于 MS-GARCH 模型的时间序列聚类

2.1. 基于有限混合模型的聚类

设数据集 $\{\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_I\}$ 是来自由 K 个成分构成的有限混合概率分布

$$f(\mathbf{y}\mid\rho,\Theta)=\sum_{k=1}^{K}\rho_k f(\mathbf{y}\mid\Theta_k) \tag{2.1}$$

的样本，其中 $\rho=(\rho_1,\rho_2,\cdots,\rho_K)$ 是混合权重，$\Theta=(\Theta_1,\Theta_2,\cdots,\Theta_K)$ 是成分参数，$f(\mathbf{y}\mid\Theta_k)$ 是第 k 个成分

的概率分布，权重 ρ 满足：

$$0 \le \rho_k \le 1 (k = 1,2,\cdots,K), \quad \sum_{k=1}^{K} \rho_k = 1.$$

基于模型的聚类方法把混合模型的成分与聚类的簇相对应，成分数就是簇数。设 z_i 代表数据 $\mathbf{y}_i$ 的簇标记，$z_i = k$ 表示 $\mathbf{y}_i$ 属于第 k 个簇。若模型的参数 ρ, Θ 已知，则由 Bayes 定理

$$\mathbb{P}\{z_i = k \mid \mathbf{y}_i\} = \frac{\rho_k f(\mathbf{y}_i \mid \Theta_k)}{\sum_{j=1}^{K} \rho_j f(\mathbf{y}_i \mid \Theta_j)}.$$

若

$$\lambda_i = \arg\min_{k \in \{1,2,\cdots,K\}} \mathbb{P}\{z_i = k \mid \mathbf{y}_i\},$$

则 $\mathbf{y}_i$ 归属于第 λ_i 个簇。

2.2. MS-GARCH 模型

称

$$\begin{cases} y_t = \sqrt{h_t} \cdot \varepsilon_t, \\ h_t = \omega_{s_t} + \sum_{i=1}^{p} \alpha_{i,s_t} y_{t-i}^2 + \sum_{j=1}^{q} \beta_{j,s_t} h_{t-j}, \end{cases}$$

为具有 Markov 状态切换的 GARCH(p, q)模型，简记为 MS-GARCH，其中 $\{s_t\}_{t \in \mathbb{N}}$ 是状态空间为 $\{1,2,\cdots,\mathcal{S}\}$、转移概率矩阵为 $\eta = (\eta_{ij})$ 的遍历 Markov 链，$\{\varepsilon_t\}$ 独立同分布，对任意 $s = 1,2,\cdots,\mathcal{S}$，$i = 1,2,\cdots,p$，$j = 1,2,\cdots,q$，有 $\omega_s > 0$，$\alpha_{i,s} \ge 0$，$\beta_{j,s} \ge 0$。

2.3. 基于 MS-GARCH 模型的时间序列聚类

设 $\{y_{i,t}, i = 1,2,\cdots,I; t = 1,2,\cdots,T\}$ 是 I 个长度为 T 的时间序列组成的观察数据，记

$$\mathbf{y}_i = (y_{i,1}, y_{i,2}, \cdots, y_{i,T}), i = 1,\cdots,I.$$

假设 $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_I\}$ 来自有限混合概率分布(2.1)，$f(\mathbf{y} \mid \Theta_k)$ 是第 k 个 MS-GARCH(p, q)模型

$$\begin{cases} y_t = \sqrt{h_t} \cdot \varepsilon_t, \\ h_t = \omega_{s_t^k}^k + \sum_{i=1}^{p} \alpha_{i,s_t^k}^k y_{t-i}^2 + \sum_{j=1}^{q} \beta_{j,s_t^k}^k h_{t-j}, \end{cases} \tag{2.2}$$

的概率密度，其中 $\{s_t^k\}_{t \in \mathbb{N}}$ 是状态空间为 $\{1,2,\cdots,\mathcal{S}\}$、转移概率矩阵为 $\eta^k = (\eta_{ij}^k)$ 的遍历 Markov 链。

为叙述方便，下面只对两状态的 MS-GARCH(1, 1)模型讨论，其它情况类似。此时，模型(2.2)的参数为 $\Theta_k = (\theta_k, \eta^k)$，其中

$$\theta_k = (\omega_1^k, \omega_2^k, \alpha_1^k, \alpha_2^k, \beta_1^k, \beta_2^k), \quad \eta^k = (\eta_{11}^k, \eta_{12}^k, \eta_{21}^k, \eta_{22}^k),$$

对 $j = 1,2$，有

$$\omega_j^k > 0, \ \alpha_j^k \ge 0, \ \beta_j^k \ge 0, \ \eta_{j1}^k \ge 0, \ \eta_{j2}^k \ge 0, \ \alpha_j^k + \beta_j^k < 1, \ \eta_{j1}^k + \eta_{j2}^k = 1.$$

当 ε_t 服从标准正态分布时，联合概率密度

$$f\left(\mathbf{y}\mid\Theta_k\right)=\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi h_t}}\exp\left(-\frac{y_t^2}{2h_t}\right), \tag{2.3}$$

其中 h_t 由(2.2)的第二式递推给出。

定义随机变量 $z_i\in\{1,2,\cdots,K\}$，$i=1,2,\cdots,I$，其概率分布

$$\mathbb{P}\left\{z_i=k\mid\rho\right\}=\rho_k.$$

记

$$z=\left(z_1,z_2,\cdots,z_I\right),\ S_T^k=\left(s_1^k,s_2^k,\cdots,s_T^k\right),\ S_{-t}^k=\left(s_1^k,s_2^k,\cdots,s_{t-1}^k,s_{t+1}^k,\cdots,s_T^k\right).$$

由于 GARCH 模型中条件方差存在路径依赖问题，参数估计的极大似然方法不再适用，我们利用 MCMC 方法对隐变量 z 和参数 ρ,Θ 进行 Gibbs 抽样。为此，需要计算 $\left(z,\rho,\Theta\right)$ 的后验分布。取先验分布 $\pi\left(z,\rho,\Theta\right)$，使得

$$\pi\left(z,\rho,\Theta\right)=\pi\left(z\mid\rho\right)\pi\left(\rho\right)\pi\left(\Theta\right), \tag{2.4}$$

其中

$$\pi\left(z\mid\rho\right)=\prod_{i=1}^{I}\rho_{z_i}=\prod_{k=1}^{K}\rho_k^{x_k},\ x_k=\#\left\{z_i=k,i=1,2,\cdots,I\right\}, \tag{2.5}$$

$$\pi\left(\Theta\right)=\prod_{k=1}^{K}\pi\left(\Theta_k\right). \tag{2.6}$$

利用 Bayes 定理，由(2.4)~(2.6)可得 $\left(z,\rho,\Theta\right)$ 的后验分布

$$p\left(z,\rho,\Theta\mid\mathbf{y}\right)\propto f\left(\mathbf{y}\mid z,\rho,\Theta\right)\pi\left(z,\rho,\Theta\right)=\pi\left(\rho\right)\prod_{k=1}^{K}\rho_k^{x_k}\pi\left(\Theta_k\right)\prod_{i=1}^{I}f\left(\mathbf{y}_i\mid\Theta_{z_i}\right). \tag{2.7}$$

下面依次讨论 z、ρ 和 Θ 的抽样。

2.3.1. 簇标记 z 的抽样

由(2.7)可得 z 的后验分布

$$p\left(z_1,z_2,\cdots,z_I\mid\rho,\Theta,\mathbf{y}\right)\propto\prod_{i=1}^{I}f\left(\mathbf{y}_i\mid\Theta_{z_i}\right)\rho_{z_i}=\prod_{i=1}^{I}p\left(z_i\middle|\rho,\Theta,\mathbf{y}_i\right). \tag{2.8}$$

于是，$z_i\mid\rho,\Theta,\mathbf{y}_i$ 服从多项分布，满足

$$P\left(z_i=k\mid\rho,\Theta,\mathbf{y}_i\right)\propto f\left(\mathbf{y}_i\mid\Theta_k\right)\rho_k,\quad k=1,2,\cdots,K. \tag{2.9}$$

2.3.2. 混合概率 ρ 的抽样

由(2.7)可得 ρ 的后验分布

$$p\left(\rho\mid z,\Theta,\mathbf{y}\right)=p\left(\rho\mid z\right)\propto\pi\left(\rho\right)\prod_{k=1}^{K}\rho_k^{x_k}. \tag{2.10}$$

取 $\pi\left(\rho\right)$ 为 Dirichlet 分布 $\mathcal{D}\left(a_1,a_2,\cdots,a_K\right)$：

$$\pi(\rho\mid a)=\frac{\prod_{k=1}^{K}\Gamma(a_k)}{\Gamma\left(\sum_{k=1}^{K}a_k\right)}\prod_{k=1}^{K}\rho_k^{a_k-1},\ a_k>0,\ \rho_i\geq 0,\ \sum_{k=1}^{K}\rho_k=1,$$

则由(2.10)可知，后验分布 $p(z\mid\rho)$ 为 Dirichlet 分布 $\mathcal{D}(a_1+x_1,a_2+x_2,\cdots,a_K+x_K)$。

2.3.3. 簇参数 Θ 的抽样

由(2.6)和(2.7)可知， Θ 的后验分布

$$p(\Theta\mid z,\rho,\mathbf{y})=p(\Theta\mid z,\mathbf{y})=\prod_{k=1}^{K}p(\Theta_k\mid\tilde{\mathbf{y}}_k)\propto\prod_{k=1}^{K}\pi(\Theta_k)\prod_{i\in I_k}f(\mathbf{y}_i\mid\Theta_k),\tag{2.11}$$

其中 $I_k=\{i\mid z_i=k\}$， $\tilde{\mathbf{y}}_k=\{\mathbf{y}_i\mid i\in I_k\}$。于是，对 Θ 的抽样可通过对每个 Θ_k 分别独立抽样来实现。取先验分布 $\pi(\Theta_k)$，使得

$$\pi(\Theta_k)=\pi(\theta_k)\pi(\eta^k).$$

下面依次讨论 S^k 、 η^k 和 θ_k 的抽样。

1) 状态标记 S_T^k 的抽样

状态变量 s_t^k 是一个隐 Markov 过程，根据 Markov 链的性质， s_t^k 的概率需要考虑 s_{t-1}^k 和 s_{t+1}^k 的影响。由

$$\begin{aligned}p(s_t^k\mid S_{-t}^k,\theta_k,\eta^k,\tilde{\mathbf{y}}_k)&=\frac{p(s_t^k,S_{-t}^k,\theta_k,\eta^k,\tilde{\mathbf{y}}_k)}{p(S_{-t}^k,\theta_k,\eta^k,\tilde{\mathbf{y}}_k)}=\frac{p(\tilde{\mathbf{y}}_k\mid S_T^k,\theta_k,\eta^k)\,p(S_T^k,\theta_k,\eta^k)}{p(\tilde{\mathbf{y}}_k\mid S_{-t}^k,\theta_k,\eta^k)\,p(S_{-t}^k,\theta_k,\eta^k)}\\&=\frac{f(\tilde{\mathbf{y}}_k\mid S_T^k,\theta_k,\eta^k)\,p(S_T^k\mid S_{-t}^k,\theta_k,\eta^k)}{f(\tilde{\mathbf{y}}_k\mid S_{-t}^k,\theta_k,\eta^k)},\end{aligned}$$

可知

$$\mathbb{P}(s_t^k=s\mid S_{-t}^k,\theta_k,\eta^k,\tilde{\mathbf{y}}_k)\propto\eta^k_{s_{t-1},s}\eta^k_{s,s_{t+1}}\prod_{i\in I(k)}\prod_{\tau=t}^{T}h_{i\tau}^{-0.5}\exp\left(-\frac{y_{i\tau}^2}{2h_{i\tau}}\right).\tag{2.12}$$

由于 s_t^k 只有两个状态，由(2.12)可知对 s_t^k 的抽样相当于从一个 Bernoulli 分布中抽样。

2) 转移概率 η^k 的抽样

给定 η^k 的先验分布 $\pi(\eta^k)$，当 η^k 独立于观察变量 $\tilde{\mathbf{y}}_k$ 和参数 θ^k 时， η^k 的后验分布

$$p(\eta^k\mid S_T^k,\theta^k,\tilde{\mathbf{y}}_k)=p(\eta^k\mid S_T^k)\propto\pi(\eta^k)\prod_{t=1}^{T}\eta^k_{s_{t-1},s_t}.\tag{2.13}$$

取 $\pi(\eta^k)=\pi(\eta_{11}^k,\eta_{12}^k)\pi(\eta_{21}^k,\eta_{22}^k)$，当 $\pi(\eta_{11}^k,\eta_{12}^k)$ 和 $\pi(\eta_{21}^k,\eta_{22}^k)$ 分别为 Dirichlet 分布 $\mathcal{D}(b_1,b_2)$ 和 $\mathcal{D}(c_1,c_2)$ 时，由(2.13)可知后验分布 $p(\eta_{11}^k,\eta_{12}^k\mid S_T^k)$ 和 $p(\eta_{21}^k,\eta_{22}^k\mid S_T^k)$ 分别为 $\mathcal{D}(b_1+n_{11},b_2+n_{12})$ 和 $\mathcal{D}(c_1+n_{21},c_2+n_{22})$，其中 n_{gl} 为从状态 g 转移到状态 l 的次数。

3) GARCH 参数 θ^k 的抽样

由于 GARCH 模型中当期条件方差受到前期条件方差影响的递推结构，各参数之间不独立，参数的后验分布没有合适的共轭先验，因此不能直接使用常规的 Gibbs 抽样。若取 θ^k 的先验分布为常数，则 θ^k 的后验分布

$$p\left(\theta^k \mid S_T^k, \eta^k, \tilde{\mathbf{y}}_k\right) \propto \prod_{i \in I(k)} \prod_{t=1}^{T} h_{it}^{-0.5} \exp\left(-\frac{y_{it}^2}{2h_{it}}\right). \tag{2.14}$$

利用 griddy-Gibbs 抽样方法[12] [28]，根据(2.14)可对 θ^k 进行抽样。

3. 实证分析

本文选取美的集团、长安汽车、格力电器、海信家电、三全食品、比亚迪、东风汽车、宇通客车、上汽集团、江淮汽车、光明乳业、海尔智家、伊利股份、长城汽车、飞科电器、中信证券等 23 家上市公司 2017 年 12 月至 2020 年 7 月的所有日收益数据进行模拟实验(数据来源：网易财经数据库)，以验证本算法的有效性。

3.1. 数据预处理

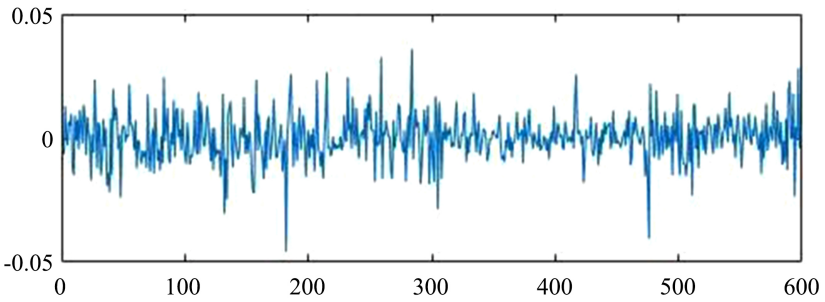
1) 由于上市公司的股权变更、违规调查等外部因素造成的停牌现象，使数据中存在少数 NULL 值。对此执行删除操作，不影响原序列的特征信息。

2) 实验中选取上市公司股票的对数收益率数据进行计算，以消除时间不连续等负面影响，

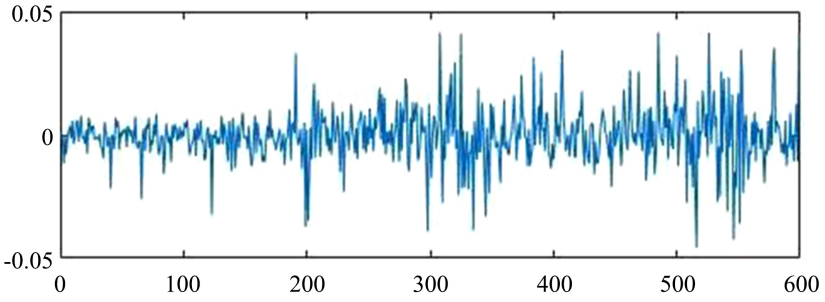
3) 本文的 MS-GARCH 模型只考虑条件方差的波动特征，取对数收益率后，将序列减去各自样本均值以消除模型中均值参数的影响，有助于简化迭代计算过程。

3.2. 数据概览

我们从中选取两家公司的收益率序列数据，画出它们的时序图，如图 1 所示。可以看出各个序列都有明显的波动聚集特征，即某些时间段有较稳定的波动，某些时间段有较剧烈的震荡。



(a) 对数收益率 $\mathbf{y}_1$



(b) 对数收益率 $\mathbf{y}_2$

Figure 1. Returns chart of logarithmic rate
图 1. 对数收益率图

3.3. 模型选择

选用两状态的 MS-GARCH(1,1)混合模型进行聚类分析。在聚类数的选择方面，参照文[21]的结论，我们把股票收益率数据按照波动率的高、中、低划分为三类，分别对应投资风险高、投资风险中等和投资风险低的三类股票。

3.4. 聚类结果

首先，我们计算得出了 23 条股票数据(包含均值、标准差等)的描述性统计，如表 1 所示。这些数据确实存在一定差异性。例如，全部数据的平均峰度为 6.3084，但范围从 3.6448 到 11.9634。

Table 1. Descriptive statistics
表 1. 描述性统计

	均值	标准差	最小值	最大值
均值	−3.8158e−05	2.3949e−04	−4.1344e−04	7.0417e−04
标准差	0.0092	0.0020	0.0046	0.0122
最小值	−0.0412	0.0081	−0.0459	−0.0154
最大值	0.377	0.0065	0.0147	0.0418
偏度	−4.4533	0.29779	−0.7997	0.5555
峰度	6.3084	1.7580	3.6448	11.9634

算法程序采用 Matlab 软件实现。抽样算法设定循环 3000 次，其中前 1000 次作为预烧阶段，保留后 2000 次作为抽样结果。

对于参数初始化，我们根据每条数据方差的差异性，对混合概率和簇类归属作了简单的初始化设置。其中，11 家公司(美的集团、格力电器、上汽集团、三元股份、海尔智家、伊利股份、长江电力、中国平安、工商银行、中国石油、建设银行)归属第一类，8 家公司(海信家电、三全食品、比亚迪、宇通客车、光明乳业、长城汽车、飞科电器、中信证券)归属第二类，4 家公司(长安汽车、贝因美、东风汽车、江淮汽车)归属第三类。初始混合概率的设置如表 2 所示。

Table 2. Initial mixed probability of groups
表 2. 初始混合概率

ρ_1	ρ_2	ρ_3
0.4783	0.3478	0.1739

23 条时间序列各自包含两个状态的参数，即共有 46 组不相同的参数。我们在程序中对这些模型参数分别设置了不同的初始值，同时它们又必须满足 GARCH 模型的平稳性条件。由于 Gibbs 抽样的收敛性质，只需要粗略估计其大致范围即可。除此之外，griddy-Gibbs 计算中设置格点数为 50 (格点的数量过大或过小都会对估计结果和运行时间产生负面影响)。最后，要注意计算过程中数值不能超过 Matlab 所允许的范围。对于 GARCH 参数初始化，我们以美的集团为例给出了详细情况，如表 3 所示。

Table 3. Initialization information of GARCH parameters
表 3. GARCH 参数初始化信息

	ω_1	ω_2	α_1	α_2	β_1	β_2
参数数值	7e−04	3.1e−03	0.2	0.3	0.56	0.52
数值范围	(8e−05, 5e−03)	(7e−05, 5.5e−03)	(0.01, 0.5)	(0.1, 0.55)	(0.2, 0.9)	(0.01, 0.9)

通过计算，我们得到了该组数据两个状态的参数后验概率密度，抽样情况良好。根据 Gibbs 抽样原理，经过迭代后的抽样值会在真实值附近波动。美的集团的 GARCH 参数的后验概率密度图，如图 2 所示。

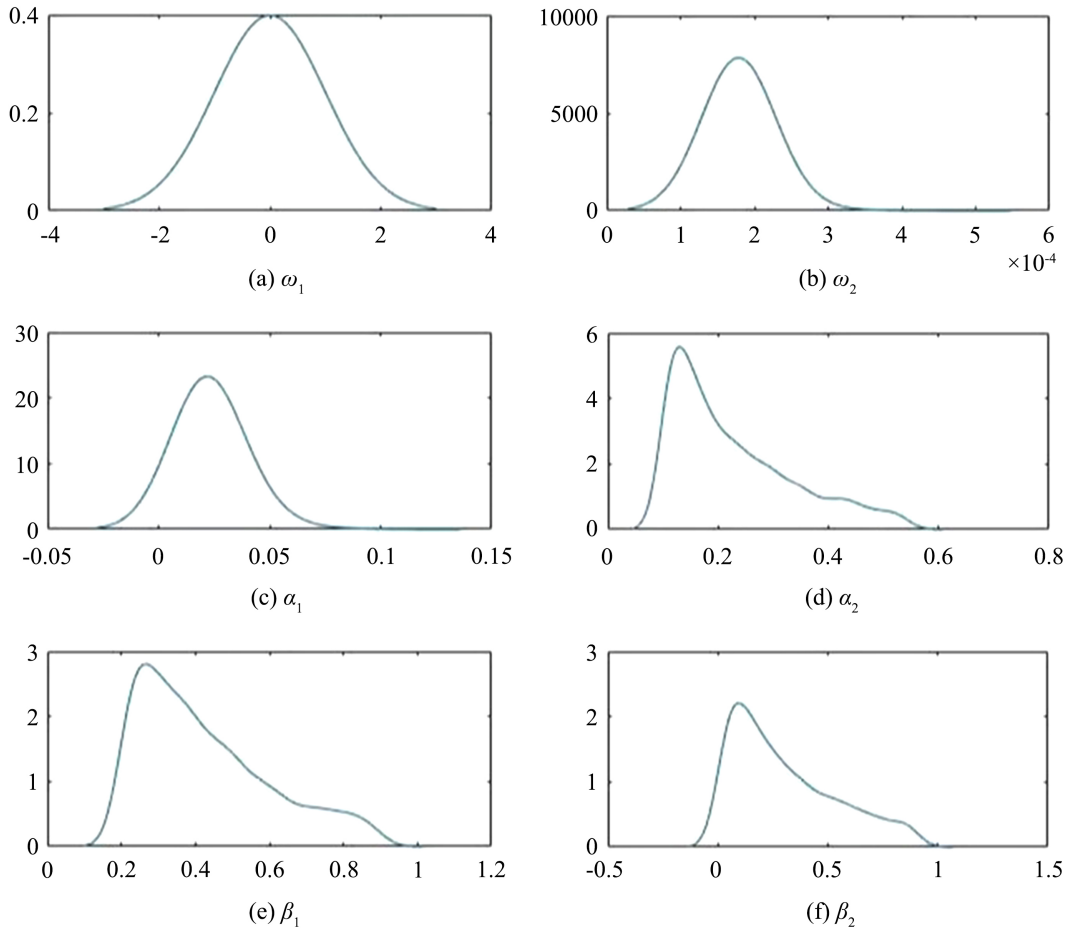


Figure 2. Posterior probability density plot of GARCH parameters
图 2. GARCH 参数后验概率密度图

由实验结果中状态转移概率后验均值与状态转移概率初始值的对比，可以看到股票处于低波动状态的持续性长于高波动状态，表示该股票大多数时间处于低幅度变动，不会频繁出现剧烈震荡。这类相对稳定的股票对于投资者来说风险偏低，若近期走势良好，值得投资。相反，若实验所得的后验概率中，高波动状态所占比重较高，那么说明股票长期处于数值较大的震荡，不利于进行投资，投资者需结合实际情况慎重分析。状态转移概率初始值与后验均值，如表 4 所示。

Table 4. Transition probability of states
表 4. 状态转移概率

转移状态	初始值	后验均值
η_{11}	0.6914	0.9644
η_{12}	0.3086	0.0356
η_{21}	0.4557	0.6124
η_{22}	0.5143	0.3876

在最终实验结果中，根据混合概率后验均值的情况可以看出在我们提供数据的基础上，三个混合概率非常接近。其中，第二类混合后验概率与初始簇类混合概率相比变化极小，也就意味着我们对第二类的初始分类状况较好。而第一类和第三类经过迭代抽样后，都产生了不同程度的变化，如表 5 所示。

Table 5. Mixed probability
表 5. 混合概率

ρ_1	ρ_2	ρ_3
0.3325	0.3348	0.3327

根据 GARCH 参数可对三个簇进行解释。三类中持久性最高的为第一组(1 状态 $\alpha_1^1+\beta_1^1$ 估计为 0.4217，2 状态 $\alpha_2^1+\beta_2^1$ 估计为 0.4621)，第一组与第三组的持久性更加接近(第三组 1 状态 $\alpha_1^3+\beta_1^3$ 估计为 0.4062，2 状态 $\alpha_2^3+\beta_2^3$ 估计为 0.4320)，即第一组代表持久性最高的股票类型，第二组代表持久性最低的股票类型，第三组代表持久性中等的股票类型。其中三组的滞后因子彼此相近(1 状态为 0.0273 左右，2 状态为 0.2160 左右)，而差异最明显的地方则是条件方差的自回归参数(第一组两状态分别为 0.3942 和 0.2462，第二组分别估计为 0.2974 和 0.2375，第三组估计分别为 0.3790 和 0.2148)。其次，三类的无条件方差具有一定差异，第一类与第二类相对接近(第一组两状态估计分别为 3.05e−05 和 1.61e−03，第二组估计分别为 1.75e−04 和 1.67e−03)，第三类最小(第三类估计分别为 4.61e−05 和 9e−04)。但由于数值较小，还不足以对各自模型的持久性产生决定性的影响。GARCH 参数的后验均值结果，如表 6 所示。

Table 6. Posterior mean of GARCH parameters
表 6. GARCH 参数的后验均值

z_i	ω_1	ω_2	α_1	α_2	β_1	β_2
1	3.05e−05	1.61e−03	0.0275	0.2159	0.3942	0.2462
2	1.75e−04	1.67e−03	0.0273	0.2159	0.2974	0.2375
3	4.61e−05	9e−04	0.0272	0.2172	0.3790	0.2148

在 23 家上市公司股票数据聚类结果中，有 5 家公司属于第一组，9 家公司属于第二组，9 家公司属于第三组。与初始化值相比较，有 9 家公司在初始化中就被正确区分，其余 14 家公司经过迭代抽样产生了类别变化，说明我们基于 Gibbs 抽样的聚类算法有明显的纠正效果。除此之外，我们还展示了各股票

数据各自模型参数 $\hat{\omega}$、$\hat{\alpha}$ 和 $\hat{\beta}$ 的后验均值，在取值上均满足 GARCH 模型的平稳性条件，如表 7 所示。

对于 Gibbs 抽样，先验分布的设定可能影响算法对于参数的估计、状态识别和聚类结果。在实验过程中，我们尝试修改各参数初始值，观察结果变化，发现初始值的改变并不会对最终结果产生太大影响，这与理论分析是一致的。

Table 7. Experimental results
表 7. 实验结果

Stock	Mean	Std	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\omega}_1$	$\hat{\omega}_2$	z_i
美的集团	2.49e−04	0.0094	1.7840e−04	1.7996e−04	0.0234	0.2271	0.4252	0.3018	3
长安汽车	−9.58e−05	0.0114	5.3200e−04	7.6242e−04	0.0272	0.2159	0.0272	0.2159	2
格力电器	9.89e−05	0.0095	6.8346e−04	8.5839e−04	0.0272	0.2158	0.2239	0.0317	2
海信家电	−2.15e−05	0.0114	4.9930e−05	8.5265e−04	0.0274	0.2158	0.3809	0.2284	2
三全食品	7.04e−04	0.0104	3.4960r−05	4.8147e−04	0.0275	0.2160	0.4135	0.3091	2
贝因美	−2.15e−04	0.0122	4.1553e−05	8.6063e−04	0.0278	0.2160	0.4528	0.2011	3
比亚迪	−3.16e−05	0.0110	4.0473e−05	6.4142e−04	0.0280	0.2160	0.4106	0.2377	3
东风汽车	−2.06e−04	0.0105	2.6773e−05	9.9240e−04	0.0283	0.2159	0.4699	0.1833	1
宇通客车	−3.89e−04	0.0097	2.6549e−05	6.8634e−04	0.0285	0.2160	0.4474	0.1996	3
上汽集团	−2.98e−04	0.0085	2.7620e−05	0.0013	0.0273	0.2160	0.3730	0.2379	3
江淮汽车	−5.14e−06	0.0117	3.1046e−05	0.0016	0.0274	0.2158	0.4570	0.2332	3
三元股份	−6.93e−05	0.0084	2.3680e−05	0.0035	0.0273	0.2160	0.4247	0.2996	2
光明乳业	−8.99e−05	0.0095	3.5175e−05	0.0017	0.0274	0.2159	0.3913	0.2753	1
海尔智家	−6.94e−05	0.0096	3.8103e−05	6.6397e−04	0.0273	0.2158	0.3634	0.2468	1
伊利股份	−3.65e−05	0.0093	2.8463e−05	0.0014	0.0274	0.2159	0.4317	0.2348	1
长城汽车	−2.22e−04	0.0103	3.1039e−05	0.0011	0.0273	0.2159	0.4305	0.2057	2
飞科电器	−3e−04	0.0091	2.6045e−05	0.0015	0.0274	0.2160	0.4204	0.2500	3
中信证券	1.87e−04	0.0098	3.4126e−05	0.0022	0.0274	0.2159	0.4142	0.2520	2
长江电力	1.26e−04	0.0046	2.0374e−05	2.2958e−04	0.0272	0.2158	0.0198	0.0308	3
中国平安	7.89e−05	0.0079	2.3178e−05	0.0011	0.0274	0.2158	0.4044	0.2409	3
工商银行	−3.03e−05	0.0055	2.4860e−05	0.0016	0.0272	0.2158	0.1328	0.2834	2
中国石油	−4.12e−04	0.0058	2.2460e−05	0.0037	0.0273	0.2160	0.2289	0.2957	2
建设银行	−2.06e−06	0.0066	2.3938e−05	0.0033	0.0273	0.2159	0.3149	0.2910	1

4. 结论

本文基于有限混合 MS-GARCH 模型，提出一种时间序列聚类方法。在新息服从标准正态分布的情况下，利用 MCMC 方法，给出了模型参数的估计，再通过计算后验概率得到聚类结果。最后，选取 23 家中国上市公司股票的收益率数据进行实证分析，验证了所提方法的有效性。本文的方法也可用于新息服从 t-分布、广义误差分布等其它分布的情况。本文的聚类数目是根据经验和理论分析事先指定的，聚类数目未知的聚类问题将在未来的工作中讨论

参考文献

[1] Bollerslev, T. (1986) Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**, 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

[2] Engle, R.F. (1982) Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation. *Econometrica*, **50**, 987-1008. https://doi.org/10.2307/1912773

[3] Lamoureux, C.G. and Lastrapes, W.D. (1990) Persistence in Variance, Structural Change, and the GARCH Model. *Journal of Business and Economic Statistics*, **8**, 225-234. https://doi.org/10.1080/07350015.1990.10509794

[4] Cai, J.A. (1994) Markov Model of Switching-Regime ARCH. *Journal of Business & Economic Statistics*, **12**, 309-316. https://doi.org/10.1080/07350015.1994.10524546

[5] Hamilton, J.D. and Susmel, R. (1994) Autoregressive Conditional Heteroskedasticity and Changes in Regime. *Journal of Econometrics*, **64**, 307-333. https://doi.org/10.1016/0304-4076(94)90067-1

[6] Gray, S. (1996) Modeling the Conditional Distribution of Interest Rates as a Regime-Switching Process. *Journal of Financial Economics*, **42**, 27-62. https://doi.org/10.1016/0304-405X(96)00875-6

[7] Klaassen, F. (2002) Improving GARCH Volatility Forecasts with Regime-Switching GARCH. *Empirical Economics*, **27**, 363-394. https://doi.org/10.1007/s001810100100

[8] Marcucci, J. (2005) Forecasting Stock Market Volatility with Regime-Switching GARCH Models. *Studies in Nonlinear Dynamics and Econometrics*, **9**, 1145. https://doi.org/10.2202/1558-3708.1145

[9] Haas, M., Mittnik S. and Paolella, M.S. (2004) A New Approach to Markov-Switching GARCH Models. *Journal of Financial Econometrics*, **2**, 493-530. https://doi.org/10.1093/jjfinec/nbh020

[10] Ané, T. and Ureche-Rangau, L. (2006) Stock Market Dynamics in a Regime-Switching Asymmetric Power GARCH Model. *International Review of Financial Analysis*, **15**, 109-129. https://doi.org/10.1016/j.irfa.2005.08.002

[11] Abramson, A. and Cohen, I. (2007) On the Stationarity of Markov-Switching GARCH Processes. *Econometric Theory*, **23**, 485-500. https://doi.org/10.1017/S0266466607070211

[12] Bauwens, L., Preminger, A. and Rombouts, J.V.K. (2010) Theory and Inference for a Markov Switching GARCH Model. *Econometrics Journal*, **13**, 218-244. https://doi.org/10.1111/j.1368-423X.2009.00307.x

[13] Henneke, J.S., Rachev, S.T., Fabozzi, F.J. and Nikolov, M. (2011) MCMC-Based Estimation of Markov Switching ARMA-GARCH Models. *Applied Economics*, **43**, 259-271. https://doi.org/10.1080/00036840802552379

[14] Liao, T.W. (2005) Clustering of Time Series Data—A Survey. *Pattern Recognition*, **38**, 1857-1874. https://doi.org/10.1016/j.patcog.2005.01.025

[15] Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y. (2015) Time-Series Clustering—A Decade Review. *Information Systems*, **53**, 16-38. https://doi.org/10.1016/j.is.2015.04.007

[16] Maharaj, E.A., D'Urso, P. and Caiado, J. (2019) Time Series Clustering and Classification. CRC Press, New York. https://doi.org/10.1201/9780429058264

[17] Fraley, C. and Raftery, A.E. (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, **97**, 611-631. https://doi.org/10.1198/016214502760047131

[18] McLachlan, G.J., Lee, S.X. and Rathnayake, S.I. (2019) Finite Mixture Models. *Annual Review of Statistics and Its Application*, **6**, 355-378. https://doi.org/10.1146/annurev-statistics-031017-100325

[19] Bouveyron, C., Celeux, G., Murphy, T.B. and Raftery, A.E. (2019) Model-Based Clustering and Classification for Data Science: with Applications in R. Cambridge University Press, New York. https://doi.org/10.1017/9781108644181

[20] Xiong, Y. and Yeung, D.Y. (2004) Time Series Clustering with ARMA Mixtures. *Pattern Recognition*, **37**, 1675-1689. https://doi.org/10.1016/j.patcog.2003.12.018

[21] Bauwens, L. and Rombouts, J.V.K. (2007) Bayesian Clustering of Many GARCH Models. *Econometric Reviews*, **26**,

365-386. https://doi.org/10.1080/07474930701220576

[22] Fröhwirth-Schnatter, S. and Kaufmann, S. (2008) Model-Based Clustering of Multiple Time Series. *Journal of Business & Economic Statistics*, **26**, 78-89. https://doi.org/10.1198/073500107000000106

[23] Samé, A., Chamroukhi, F., Govaert, G. and Aknin, P. (2011) Model-Based Clustering and Segmentation of Time Series with Changes in Regime. *Advances in Data Analysis & Classification*, **5**, 301-321. https://doi.org/10.1007/s11634-011-0096-5

[24] Frühwirth-Schnatter, S. (2011) Panel Data Analysis: A Survey on Model-Based Clustering of Time Series. *Advances in Data Analysis and Classification*, **5**, 251-280. https://doi.org/10.1007/s11634-011-0100-0

[25] Kini, B.V. and Sekhar, C.C. (2013) Bayesian Mixture of AR Models for Time Series Clustering. *Pattern Analysis and Applications*, **16**, 179-200. https://doi.org/10.1007/s10044-011-0247-5

[26] Costilla, R., Liu, I., Arnold, R. and Fernández, D. (2019) Bayesian Model-Based Clustering for Longitudinal Ordinal Data. *Computational Statistics*, **34**, 1015-1038. https://doi.org/10.1007/s00180-019-00872-4

[27] Wang, Y. and Tsay, R.S. (2019) Clustering Multiple Time Series with Structural Breaks. *Journal of Time Series Analysis*, **40**, 182-202. https://doi.org/10.1111/jtsa.12434

[28] Bauwens, L. and Lubrano, M. (1998) Bayesian Inference on GARCH Models Using the Gibbs Sampler. The Econometrics Journal, **1**, 23-46. https://doi.org/10.1111/1368-423X.11003

[29] Aielli, G.P. and Caporin, M. (2013) Fast Clustering of GARCH Processes via Gaussian Mixture Models. *Mathematics and Computers in Simulation*, **94**, 205-222. https://doi.org/10.1016/j.matcom.2012.09.015

[30] Sampietro, S. (2010) Bayesian Analysis of Mixture of Autoregressive Components with an Application to Financial Market Volatility. *Applied Stochastic Models in Business & Industry*, **22**, 225-242. https://doi.org/10.1002/asmb.613