Advances in Applied Mathematics 应用数学进展, 2022, 11(1), 404-417
Published Online January 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.111048

Hans汉斯

基于ARIMA-GARCH模型的股票价格预测

尹　路

云南财经大学统计与数学学院，云南　昆明

收稿日期：2021年12月26日；录用日期：2022年1月16日；发布日期：2022年1月28日

摘　要

本文引入了误差校正的思想，先利用ARIMA-GARCH模型对日收盘价进行初步预测，但是预测精度不高，通过对误差序列进行的白噪声检验，发现误差序列存在还有未被ARIMA-GARCH模型提取的信息。再利用变量间的相关关系，寻找与误差序列相关的变量。随后通过主成分分析对变量进行降维，以筛选出合适的解释变量。将解释变量与误差序列进行回归建模，对误差进行预测。最后将预测的误差值与ARIMA-GARCH模型的预测值相加，得到最终的预测值。通过将最终的预测值与ARIMA-GARCH模型的预测值相比较，观察到预测精度有了很大的提高，进而验证了引入误差校正的方法是合理的。

关键词

股价预测，ARIMA模型，GARCH模型，误差校正

Stock Price Prediction Based on ARIMA-GARCH Model

Lu Yin

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan

Received: Dec. 26th, 2021; accepted: Jan. 16th, 2022; published: Jan. 28th, 2022

Abstract

This paper introduces the idea of error correction. Firstly, ARIMA-GARCH model is used to predict the daily closing price, but the prediction accuracy is not high. Through the white noise test of the error sequence, it is found that there is information in the error sequence that has not been extracted by ARIMA-GARCH model. Then the correlation between variables is used to find the variables related to the error sequence. Then, the dimensionality of the variables is reduced by principal component analysis to screen out the appropriate explanatory variables. Regression modeling is carried out between explanatory variables and error series to predict the error. Finally, the

文章引用: 尹路. 基于 ARIMA-GARCH 模型的股票价格预测[J]. 应用数学进展, 2022, 11(1): 404-417.
DOI: 10.12677/aam.2022.111048

predicted error value is added to the predicted value of ARIMA-GARCH model to obtain the final predicted value. By comparing the final prediction value with the prediction value of ARIMA-GARCH model, it is observed that the prediction accuracy has been greatly improved, and then it is verified that the method of introducing error correction is reasonable.

Keywords

Stock Price Forecast, ARIMA Model, GARCH Model, Error Correction

1. 引言

1.1. 研究背景

股票市场在金融领域占据了越来越重要的地位。股票市场的建立和发展对促进经济改革和经济发展发挥了重要作用。因此，股票发行一直是许多学者和投资者研究的焦点，准确预测股票价格一直是研究人员感兴趣的方向。但是，股票市场的价格经常大起大落，呈现复杂的非线性和随机性，影响股票价格的因素很多，每个因素对股票价格的影响程度都不一样，这使得准确预测股票变得很困难。因此，寻找有效的预测方法是摆脱股票预测困境的关键。

1.2. 研究意义和目的

股票市场的价格预测研究不仅有重要的学术意义，而且有重要的实际意义。股票价格预测可以更好地理解和把握股票市场运行规律、股票价格波动规律及其对实体经济的影响机制和影响程度，更好地把握货币政策的传导机制。在实践中，在股票价格剧烈波动的情况下选择和实施有效的货币政策，有助于减少和消除股票市场的不稳定因素，从而进一步提高各国宏观经济的运行质量。

2. 文献综述

目前国内外对于股票价格预测的方法不尽相同，从最初的单一预测模型，如惠晓峰等人利用来自美国联邦储备局提供的统计数据(www.economagic.com)中的1994~1997年的日汇率数据建立GARCH模型。由于GARCH模型更适合于收益率的时间序列，因此作者对人民币/美元的日汇率序列进行了相应的处理，并将其转化为一个收益率序列。之后对收益性序列绘制了自相关图和偏自相关图，得出收益性序列是存在相关性的，接着又进行了ARCH检验，得出存在着明显的异方差性[1]。在得到估计模型后，对估计的结果进行有关的残差检验，来验证估计的有效性。最后利用滚动算法和递归算法分别来进行预测，选择了平均预测误差平方和的平方根(RMSE)，Theil不相等系数(U)，平均绝对误差(MAE)，平均预测误差(MFE)和平均绝对误差(MAPE)来衡理时间序列预测效果。人民币兑美元的时间序列存在GARCH效应(即异方差)，GARCH(1, 1)模型完全适用于人民币兑美元的建模。预测结果也再次证明了GARCH(1, 1)模型预测短期汇率的可行性[2]。刘国旗采用了两种GARCH修正模型：二次GARCH模型(即QGARCH模型)和Glosten、Jagannathan和Runkle即GJR模型。以及标准GARCH模型对中国股市的波动性进行了预测。首先讨论了中国股市数据的统计特征，得出了拒绝正态分布的原假设[3]。之后给出了模型的估计结果，

为了评价非线性 GARCH 模型预测金融时间序列波动性的能力，给出了收益率波动性的测量公式，得出 QGARCH 模型是最优的。杨建辉等人首先使用 EMD 方法对上证指数日收盘价数据进行分解，得到 6 个本征模态分量 IMF 和 1 个剩余分量 R [4]。然后将分解后的不同频率的数据重新组合成高频序列、低频序列和趋势序列，IMF1~IMF4 都是均值在 0 附近的高频分量，IMF5~IMF6 是均值显著偏离 0 的低频部分[5]。剩余分量代表指数的长期趋势项。为了避免预测过程中的误差累积，提高最终的数据预测精度。分别对 IMF 组合的高频项、低频项和趋势项进行了拟合和预测。利用自回归模型对低频序列和残差序列进行了模拟和预测。股票的高频序列表现出负偏态，因此拒绝原序列服从均值为 0 的正态分布的零假设，同时表现出过峰度(峰度大于 3 时称为过峰度)。J-B 统计量拒绝了零假设，即高频率的金融数据序列是非正态的。因此，选择了 GARCH 模型对其进行拟合。最后，对三个序列的预测值进行加权和重组，得到最终的预测结果。

由于模型复杂度的限制和信息冗余的影响，单个预测模型很难考虑所有的影响因素，因此预测精度无法保证。在这种情况下，人们提出了组合预测方法，即对同一对象采用两种或两种以上的预测方法进行预测。例如，张超首先将数据分为样本内数据和样本外数据。利用样本内数据进行建模，利用样本外数据检验模型的预测能力[6]。同时，为了便于建模，将股票价格数据转换为对数收益率数据。其次，建立了收益率序列的 ARMA-GARCH 模型，并对未来收益率进行了预测。将预测收益率值转换为预测股价值，计算实际值与预测值之间的误差。误差序列数据又分为样本内数据和样本外数据，样本内数据用来建立回归模型，样本外数据用来检验回归模型的预测能力。然后通过变量间的相关分析，筛选出与预测误差相关的影响因素。对筛选出的影响因素进行主成分分析，实现降维，提取具有代表性的综合指标。建立了综合指标和误差序列的回归模型[7]。利用回归模型对误差进行预测，得到预测值。最后，利用误差预测值对 ARMA-GARCH 模型预测的股价进行修正，得到最终的股价预测值。这种引入误差校正的方法(即回归模型)对误差的预测比 ARMA-GARCH 模型准确，明显减小了 ARMA-GARCH 模型预测的误差，从而在一定程度上克服了 ARMA-GARCH 模型因为考虑外部因素不足导致预测误差偏大的缺点。杨琦和曹显兵利用深圳 A 股大众公用(600635)在 2014 年 1 月 2 日至 2015 年 3 月 31 日共 301 天的收盘价数据进行分析与预测，首先，通过对数据的初步分析，建立 ARMA 拟合模型；然后，通过对残差序列的平方进行相关性检验(这里采用 Box-Ljung 检验)，认为残差的平方项存在强相关性，即序列存在条件异方差性。故通过加入 GARCH 模型消除条件异方差性。从残差的检验中作者发现，模型的残差并不服从正态分布，而且有一定的右偏，所以在建立 GARCH 模型时，将标准化残差的分布分别改为 t 分布、有偏的 t 分布、广义误差分布和有偏的广义误差分布来建立模型，并进行比较[8]。在参数通过 t 检验的条件下，选取 AIC 最小的模型。最后得到 ARMA-GARCH 拟合模型。在得到 ARMA-GARCH 模型后，用标准化残差及其平方检验拟合模型的充分性。

3. ARIMA-GARCH 模型简介

3.1. ARMA 模型

形如如下结构的模型称为自回归移动平均模型，简记为 ARMA(p, q)：

$$\begin{cases} x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \\ \phi_p \neq 0,\ \theta_q \neq 0 \\ E(\varepsilon_t) = 0,\ Var(\varepsilon_t) = \sigma_\varepsilon^2,\ E(\varepsilon_t \varepsilon_s) = 0,\ s \neq t \\ E(x_s \varepsilon_t) = 0,\ \forall s < t \end{cases}$$

当 $\phi_0 = 0$ 时，上述模型可以写为

$$\begin{cases} x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \\ \phi_p \neq 0, \theta_q \neq 0 \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) = 0, \forall s < t \end{cases}$$

该模型称为中心化 ARMA(p, q)模型。

当前序列值乘以一个延迟算子时，此时就变为前一个序列值，即

$$x_{t-1} = Bx_t$$
$$x_{t-2} = B^2 x_t$$
$$\vdots$$
$$x_{t-p} = B^p x_t$$

为简化上述结构，引入延迟算子，故 ARMA(p, q)模型简记为：

$$\Phi(B) x_t = \Theta(B) \varepsilon_t$$

其中 $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ 为 p 阶自回归系数多项式，$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$ 为 q 阶移动平均系数多项式。

3.2. 差分运算

根据 Cramer 分解定理，方差齐性非平稳序列都可以分解为：

$$x_t = \sum_{j=0}^{d} \beta_j t_j + \Psi(B) a_t$$

展开一个 d 阶差分，有

$$\nabla^d x_t = (1 - B)^d x_t = \sum_{i=0}^{d} (-1)^i C_d^i x_{t-i}$$

进一步

$$x_t = \sum_{i=1}^{d} (-1)^{i+1} C_d^i x_{t-i} + \nabla^d x_t$$

3.3. GARCH 模型

假设在已知历史数据的情况下，零均值和纯随机残差序列具有异方差性 $Var(\varepsilon_t) = h_t$，在正态分布的假设下，有 $\varepsilon_t / \sqrt{h_t} \sim N(0,1)$，那么异方差等价于残差平方的均值 $E(\varepsilon_t^2) = h_t$。具有

$$h_t = E(\varepsilon_t^2) = \omega + \sum_{j=1}^{q} \lambda_j \varepsilon_{t-j}^2$$

这种结构的模型称为 q 阶自回归条件异方差模型，简记为 ARCH(q)。

而 ARCH 模型只适用于异方差函数的短期自相关过程。在实际应用中，残差序列的异方差函数并非都是短期自相关性的。因此，根据 ARCH 模型，通过加入考虑异方差函数的 p 阶自相关，建立了 GARCH 模型。GARCH 模型能有效地拟合具有长期记忆的异方差函数。GARCH(p, q)模型的结构如下：

$$\begin{cases} x_t = f(t, x_{t-1}, x_{t-2}, \cdots) + \varepsilon_t \\ \varepsilon_t = \sqrt{h_t} e_t \\ h_t = \omega + \sum_{i=1}^{p} \eta_i h_{t-i} + \sum_{j=1}^{q} \lambda_j \varepsilon_{t-j}^2 \end{cases}$$

其中 $f\left(t,x_{t-1},x_{t-2},\cdots\right)$ 为 $\{x_t\}$ 的确定性信息拟合模型，$e_t \overset{i.i.d}{\sim} N\left(0,\sigma^2\right)$。

3.4. 参数估计

对于非中心化 ARMA(p, q)模型，有

$$x_t=\mu+\frac{\Theta_q\left(B\right)}{\Phi_p\left(B\right)}\varepsilon_t$$

其中 $\varepsilon_t \sim WN\left(0,\sigma_\varepsilon^2\right)$，$\Theta_q\left(B\right)=1-\theta_1B-\cdots-\theta_qB^q$，$\Phi\left(B\right)=1-\phi_1B-\cdots-\phi_pB^p$。

参数 μ 采用矩估计的方法来估计，即用样本均值估计总体均值。

$$\hat{\mu}=\overline{x}=\frac{\sum_{i=1}^{n}x_i}{n}$$

估计完参数 μ 后，估计参数 $\phi_1,\cdots,\phi_p,\theta_1,\cdots,\theta_q,\sigma_\varepsilon^2$，用三种方法估计 $p+q+1$ 个未知参数：矩估计法，极大似然估计法和最小二乘估计法。

1) 矩估计法

采用样本自相关系数估计总体自相关系数的方法，即

$$\begin{cases}\rho_1\left(\phi_1,\cdots,\phi_p,\theta_1,\cdots,\theta_q\right)=\hat{\rho}_1\\ \vdots\\ \rho_{p+q}\left(\phi_1,\cdots,\phi_p,\theta_1,\cdots,\theta_q\right)=\hat{\rho}_{p+q}\end{cases}$$

通过求解以上方程组，可以获得参数值 $\phi_1,\cdots,\phi_p,\theta_1,\cdots,\theta_q$ 的矩估计 $\hat{\phi}_1,\cdots,\hat{\phi}_p,\hat{\theta}_1,\cdots,\hat{\theta}_q$。再用样本方差估计总体方差 $\sigma_x^2=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n}$。在 ARMA($p$, q)模型两边同时求方差，得到 $\sigma_\varepsilon^2=\frac{1+\phi_1^2+\cdots+\phi_p^2}{1+\theta_1^2+\cdots+\theta_q^2}\sigma_x^2$，把估计值代入其中得到 σ_ε^2 的估计值。

2) 极大似然估计法

假设序列服从多元正态分布

$$x_t=\phi_1x_{t-1}+\cdots+\phi_px_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\cdots-\theta_q\varepsilon_{t-q}$$

记

$$\tilde{x}=\left(x_1,\cdots,x_n\right)$$

$$\tilde{\beta}=\left(\phi_1,\cdots,\phi_p,\theta_1,\cdots,\theta_q\right)'$$

$$\sum\nolimits_n E\left(\tilde{x}'\tilde{x}\right)=\Omega\sigma_\varepsilon^2$$

其中

$$\Omega=\begin{pmatrix}\sum_{i=0}^{\infty}G_i^2 & \cdots & \sum_{i=0}^{\infty}G_iG_{i+n-1}\\ \vdots & \ddots & \vdots\\ \sum_{i=0}^{\infty}G_iG_{i+n-1} & \cdots & \sum_{i=0}^{\infty}G_i^2\end{pmatrix}$$

那么 $\tilde{x}$ 的似然函数为：

$$L\left(\tilde{\beta};\tilde{x}\right)=p\left(x_1,x_2,\cdots,x_n;\tilde{\beta}\right)=\left(2\Pi\right)^{-\frac{n}{2}}\left(\sigma_\varepsilon^2\right)^{-\frac{n}{2}}\left|\Omega\right|^{-\frac{1}{2}}\exp\left\{-\frac{\tilde{x}'\Omega^{-1}\tilde{x}}{2\sigma_\varepsilon^2}\right\}$$

对数似然函数为：

$$L\left(\tilde{\beta};\tilde{x}\right)=-\frac{n}{2}\ln\left(2\Pi\right)-\frac{n}{2}\ln\left(\sigma_\varepsilon^2\right)-\frac{1}{2}\ln\left|\Omega\right|-\frac{1}{2\sigma_\varepsilon^2}\left|\tilde{x}'\Omega^{-1}\tilde{x}\right|$$

对未知参数求偏导数，得到似然方程组

$$\begin{cases}\dfrac{\partial}{\partial\sigma_\varepsilon^2}l\left(\tilde{\beta};\tilde{x}\right)=-\dfrac{n}{2\sigma_\varepsilon^2}+\dfrac{S\left(\tilde{\beta}\right)}{2\sigma_\varepsilon^4}=0\\ \dfrac{\partial}{\partial\tilde{\beta}}l\left(\tilde{\beta};\tilde{x}\right)=-\dfrac{1}{2}\dfrac{\partial\ln\left|\Omega\right|}{\partial\tilde{\beta}}-\dfrac{1}{\sigma_\varepsilon^2}\dfrac{\partial S\left(\tilde{\beta}\right)}{2\partial\tilde{\beta}}=0\end{cases}$$

其中 $S\left(\tilde{\beta}\right)=\tilde{x}'\Omega^{-1}\tilde{x}$。

求解似然方程组就可得到未知参数的极大似然估计值。

3.5. 模型检验

1) 模型的显著性检验

如果一个序列的信息提取是充分的，那么拟合该序列的模型是显著的。也就是说，在一个好的拟合模型中，拟合残差项不再包含任何相关信息，残差序列是一个纯随机序列。所以接下来对残差序列进行白噪声检验。

原假设 H_0： $\rho_1=\rho_2=\cdots=\rho_m=0,\ \forall m\geq 1$

备择假设 H_1：至少存在某个 $\rho_k\neq 0,\ \forall m\geq 1,\ k\leq m$

检验统计量： $LB=n\left(n+2\right)\sum_{k=1}^{m}\left(\dfrac{\hat{\rho}_k^2}{n-k}\right)\sim\chi^2\left(m\right),\ \forall m>0$

如果零假设被拒绝，则表明残差序列中仍然存在未提取的信息，拟合模型不显著。

2) 参数的显著性检验

原假设 H_0： $\beta_j=0,\ 1\leq\forall j\leq m$

备择假设 H_1： $\beta_j\neq 0,\ 1\leq\forall j\leq m$

$$E\left(\hat{\beta}\right)=E\left[\left(X'X\right)^{-1}X'\tilde{y}\right]=\left(X'X\right)^{-1}X'X\tilde{\beta}=\tilde{\beta}$$

$$Var\left(\hat{\beta}\right)=Var\left[\left(X'X\right)^{-1}X'\tilde{y}\right]=\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}\sigma_\varepsilon^2=\left(X'X\right)^{-1}\sigma_\varepsilon^2$$

对于线性拟合模型，记 $\hat{\beta}$ 为 $\tilde{\beta}$ 的最小二乘估计，有

$$\Omega=\left(X'X\right)^{-1}=\begin{pmatrix}a_{11}&\cdots&a_{1m}\\ \vdots&\ddots&\vdots\\ a_{m1}&\cdots&a_{mm}\end{pmatrix}$$

在正态分布假定下，第 j 个未知参数的最小二乘估计值 $\hat{\beta}_j$ 服从正态分布：

$$\hat{\beta}_j \sim N\left(0, a_{jj}\sigma_\varepsilon^2\right),\ 1 \le j \le m \tag{1}$$

由于 σ_ε^2 不可观测，故用最小残差平方和估计 σ_ε^2：

$$\hat{\sigma}_\varepsilon^2 = \frac{Q\left(\tilde{\beta}\right)}{n-m}$$

根据正态分布的性质，有

$$\frac{Q\left(\tilde{\beta}\right)}{\sigma_\varepsilon^2} \sim \chi^2\left(n-m\right) \tag{2}$$

由式(1)和式(2)可以构造出检验统计量 t:

$$t = \sqrt{n-m}\frac{\hat{\beta}_j}{\sqrt{a_{jj}Q\left(\tilde{\beta}\right)}} \sim t\left(n-m\right)$$

当 $|t| \ge t_{1-\alpha}\left(n-m\right)$ 时，拒绝原假设，则认为该参数显著。

4. 数据预处理

4.1. 数据来源和基本特征

本文中采用的是绿色拇指工业公司 2018 年 6 月 18 日至 2020 年 10 月 21 日的日收盘价格，其中 2018 年 6 月 18 日至 2020 年 7 月 27 日共计 531 个数据为样本内建模数据，2020 年 7 月 28 日至 2020 年 10 月 21 日共计 61 个数据为样本外数据进行预测，所有的数据均来源于 https://www.datafountain.cn/datasets/5410。

Table 1. Basic statistical characteristics of sequences
表 1. 序列的基本统计特征

序列	样本量	均值	方差	最大值	最小值	偏度	峰度
close	531	10.4746	8.8114	23.16	4.079	0.8361	1.1609

从表 1 可以看出，序列的偏度为 0.8361，大于 0，说明序列是右偏的；峰度为 1.1609，小于 3，表明该序列具有细尾特征。

4.2. 数据预处理

通过观察日收盘价序列的时序图，该序列具有明显的递增趋势，所以它是非平稳序列。由于该数据是股票数据，为了能直观感受到收益率的变化，为了使数据更平稳，不改变数据间的相关关系，同时削弱数据的异方差和共线性，本文对数据进行预处理，采取对数收益率的方式。我们将日收益率定义为日收盘价的对数的一阶差分：

$$y_t = \ln\left(x_t\right) - \ln\left(x_{t-1}\right)$$

从时序图可以看出，日收益率始终在零值附近波动，没有明显的趋势和周期，因此可以初步认为是一个平稳序列。自相关图 1 显示出除了延迟 3 阶自相关系数在 2 倍标准差范围之外，其他阶数的自相关系数在 2 倍标准差范围之内，显示出很强的短期自相关性。但是衰减到零的速度很慢，可以判断该序列

具有拖尾的特征。同时偏自相关图也显示了拖尾的性质。

通过对收益率序列进行纯随机性检验，发现延迟 6 阶的 Q 统计量的 P 值为 0.0001689 小于显著性水平 $\alpha = 0.05$，延迟 12 阶的 Q 统计量的 P 值为 0.001266 小于 α，因此，可以拒绝纯随机性的原假设，并对该序列进行统计分析。

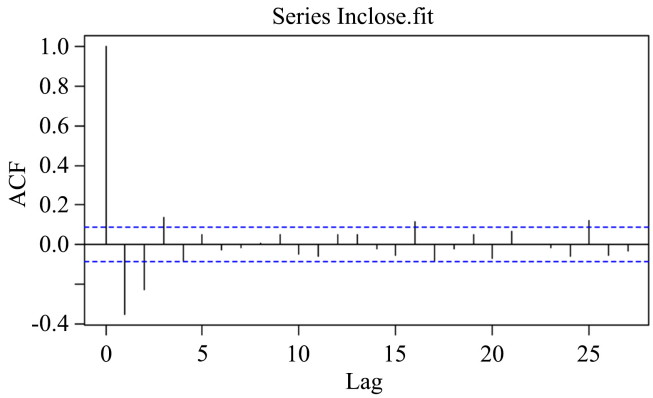

Figure 1. Autocorrelation diagram
图 1. 自相关图

5. 建立模型

5.1. 建模原理

当我们拿到一个观察值序列之后，一个完整的分析应该关注水平和波动两方面。首先提取序列的水平信息，然后分析残差序列的波动信息。所以我们首先对每日收益率序列进行 ARIMA 建模，对残差序列进行 GARCH 建模。通过用 ARIMA 模型和 GARCH 模型结合的 ARIMA-GARCH 模型进行预测。由于预测值与真实值之间存在误差，所以利用主成分分析剔除变量间的冗余信息，寻找适合回归模型的自变量，对误差序列进行回归拟合。利用该回归模型进行误差序列的预测，最后将预测值对 ARIMA-GARCH 模型的预测值进行校正。

5.2. 建模步骤

如图 2 所示，本文建模步骤划分为四个流程：① 将原始数据序列转换为收益率数据序列，针对收益率序列建立 ARIMA 模型；② 对 ARIMA 模型的残差序列建立 GARCH 模型；③ 利用 ARIMA-GARCH 模型进行预测，得到预测值 $\hat{y}$；④ 利用实际值与预测值之间的误差序列拟合回归模型并对误差序列进行预测，得到预测值 $\hat{\varepsilon}$；⑤ 将误差序列的预测值对 ARIMA-GARCH 模型得预测值校正，得到最终的预测值 $\hat{y}'$。

5.3. 建立模型

通过自相关图和偏自相关图分析，建立了 ARIMA(1, 1, 1)模型。图 3 显示了建立的 ARIMA 模型的残差检验结果。延迟 6 阶的 P 值均大于 0.05，没有充分的理由拒绝原假设。因此，残差序列是纯随机的，表明拟合模型 ARIMA(1, 1, 1)显著有效。参数的显著性检验中 P 值均小于 0.05，因此，该参数通过了显著性检验。

由于建立的模型和参数均通过了显著性检验，故所拟合模型可以用于进一步的预测。拟合的 ARIMA(1, 1, 1)模型：

$$x_t = -0.5117x_{t-1} + \varepsilon_t + 0.7406\varepsilon_{t-1}, \varepsilon_t \sim N(0, 0.002213)$$

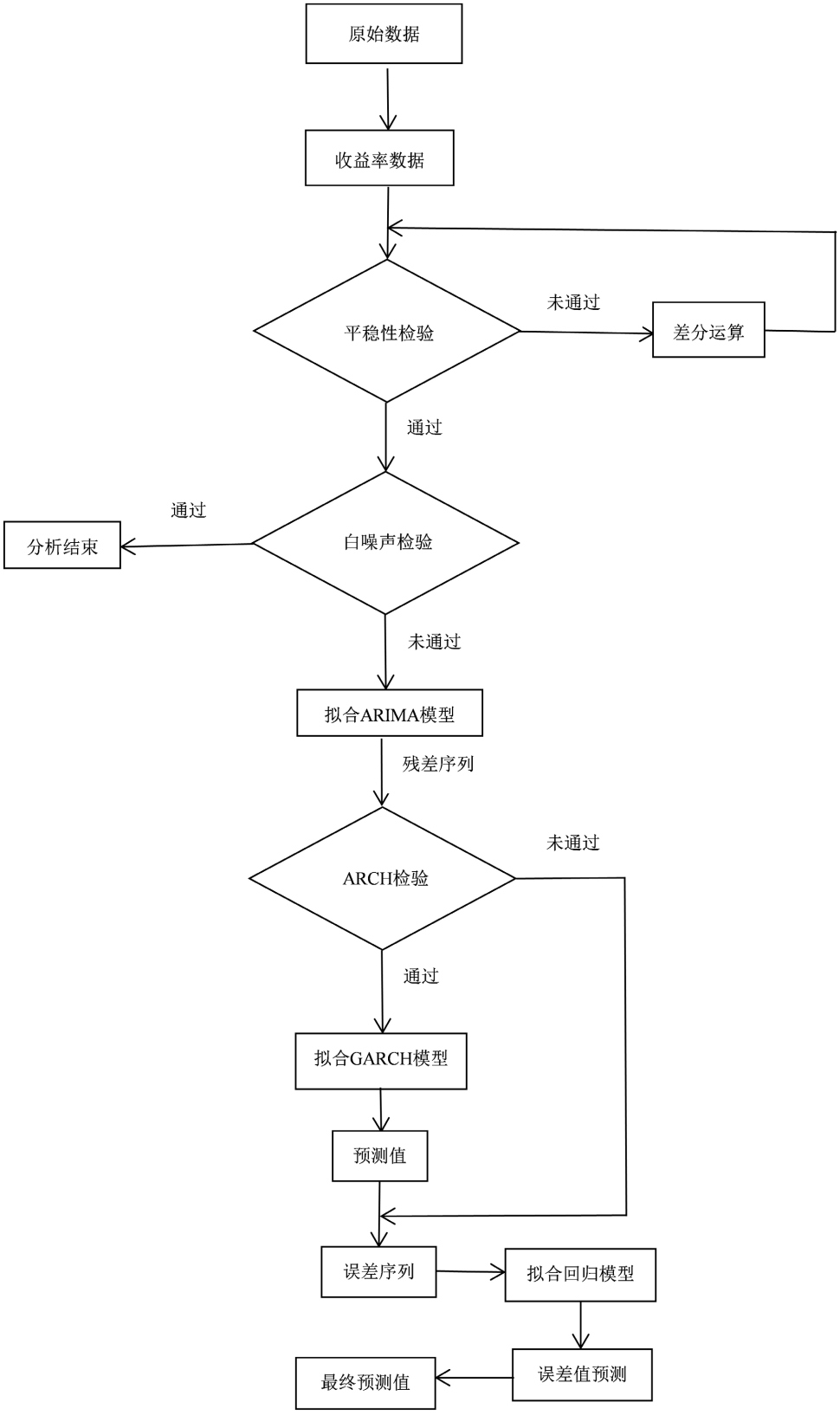

Figure 2. Flow chart
图 2. 流程图

```
        Box-Ljung test

data:  lnclose.fit$residuals
X-squared = 0.10978, df = 1, p-value = 0.7404


        Box-Ljung test

data:  lnclose.fit$residuals
X-squared = 1.1305, df = 2, p-value = 0.5682


        Box-Ljung test

data:  lnclose.fit$residuals
X-squared = 1.8125, df = 3, p-value = 0.6122


        Box-Ljung test

data:  lnclose.fit$residuals
X-squared = 1.9152, df = 4, p-value = 0.7514


        Box-Ljung test

data:  lnclose.fit$residuals
X-squared = 2.0646, df = 5, p-value = 0.8401


        Box-Ljung test

data:  lnclose.fit$residuals
X-squared = 2.2063, df = 6, p-value = 0.8998
```

Figure 3. Significance test chart of ARIMA model
图 3. ARIMA 模型的显著性检验图

通过观察 ARIMA(1, 1, 1)模型的残差序列时序图(如图 4)，发现残差序列的波动在大部分时期是平稳的，但在某些时期波动持续偏大，某些时期波动持续偏小，呈现出集群效应，有理由怀疑序列的方差是非齐性的。因此对残差序列进行 ARCH 检验，检验结果表明残差序列具有显著的方差异质性，故对残差序列进行了 GARCH(1, 1)模型拟合。

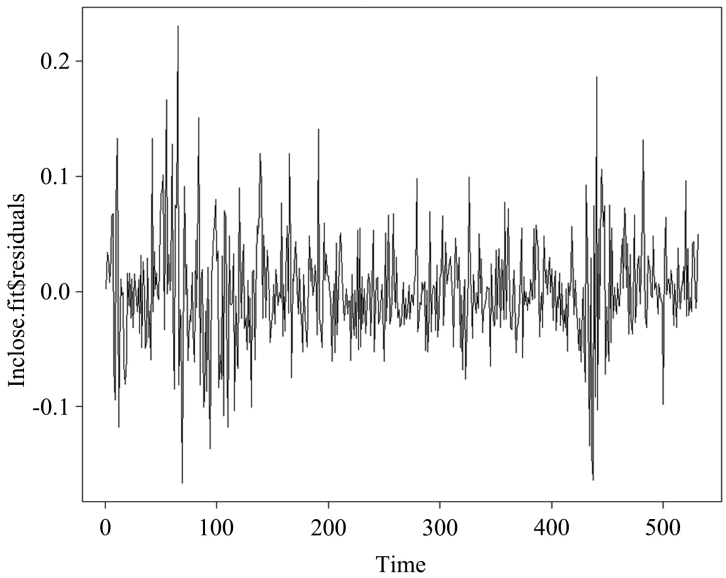

Figure 4. Residual sequence diagram
图 4. 残差时序图

GARCH(1, 1)模型拟合的结果表明模型和参数都通过了显著性检验。通过将水平模型和波动模型相结合，得到拟合的 ARIMA-GARCH 模型：

$$\begin{cases} x_t = -0.5117x_{t-1} + \varepsilon_t + 0.7406\varepsilon_{t-1} + \nu_t, \varepsilon_t \sim N(0, 0.002213) \\ \nu_t = \sqrt{h_t e_t} \\ h_t = 0.6640h_{t-1} + 0.2066\nu_{t-1}^2 \end{cases}$$

采用 ARIMA-GARCH 模型进行初步预测，预测结果 $\hat{y}$ 见下表 2。

Table 2. Predicted value of ARIMA-GARCH model
表 2. ARIMA-GARCH 模型的预测值

日期	实际值	预测值($\hat{y}$)	误差/%
2020/7/28	13.28	14.09207	−6.11500095
2020/7/29	13.635	13.94496	−2.27325859
2020/7/30	13.455	14.44043	−7.32389200
2020/7/31	13.413	14.05194	−4.76355940
2020/8/3	13.927	13.94450	−0.12567487
2020/8/4	13.88	14.08516	−1.47810007
2020/8/5	13.84	13.51243	2.36681414
2020/8/6	14.44	13.39707	7.22253479
2020/8/7	14.35	13.34368	7.01269712
2020/8/10	14.99	13.28570	11.36956049
2020/8/11	14.47	13.28375	8.19797348
2020/8/12	14.91	13.26919	11.00473582
2020/8/13	15.45	13.21299	14.47902174
2020/8/14	16.01	13.34118	16.66970602
2020/8/17	15.84	13.19656	16.68835485
2020/8/18	15.1	13.16294	12.82823706
2020/8/19	14.66	13.24223	9.67103038
2020/8/20	14.3	13.24481	7.37892418
2020/8/21	14.008	13.23708	5.50339099
2020/8/24	14.6	13.27349	9.08568656
2020/8/25	14.45	13.29719	7.97790768
2020/8/26	14.191	13.28262	6.40112505
2020/8/27	14.64	13.36046	8.74000425
2020/8/28	15.25	13.40531	12.09629869
2020/8/31	15.25	13.37560	12.29114761
2020/9/1	14.819	13.22457	10.75938074

Continued

2020/9/2	14.04	13.17579	6.15534852
2020/9/3	13.7	13.09785	4.39523671
2020/9/4	13.8	12.98293	5.92078915
2020/9/8	13.45	12.96274	3.62276513
2020/9/9	13.6	12.95624	4.73356103
2020/9/10	13.25	13.01203	1.79600402
2020/9/11	13.86	13.08871	5.56489034
2020/9/14	13.556	13.16137	2.91108513
2020/9/15	13.423	13.08937	2.48547731
2020/9/16	13.82	13.00026	5.93155791
2020/9/17	14.15	12.96799	8.35341000
2020/9/18	14.1	13.01368	7.70438689
2020/9/21	13.505	13.01740	3.61049660
2020/9/22	13.555	13.03632	3.82647190
2020/9/23	12.91	13.00128	−0.70703165
2020/9/24	12.328	12.94991	−5.04473469
2020/9/25	12.36	12.96650	−4.90694377
2020/9/28	12.67	12.95528	−2.25162182
2020/9/29	12.89	12.92789	−0.29398175
2020/9/30	12.955	12.92832	0.20590666
2020/10/1	12.93	12.95107	−0.16292963
2020/10/2	12.97	12.97858	−0.06617826
2020/10/5	13.33	13.01796	2.34090910
2020/10/6	13.308	13.00231	2.29701921
2020/10/7	13.76	13.05121	5.15111638
2020/10/8	15.107	13.04968	13.61832406
2020/10/9	15.32	13.05934	14.75629829
2020/10/12	15.84	13.13939	17.04933509
2020/10/13	15.85	13.19401	16.75700725
2020/10/14	15.658	13.19860	15.70695820
2020/10/15	15.513	13.32459	14.10691091
2020/10/16	15.88	13.28203	16.36000485
2020/10/19	15.84	13.20718	16.62132793
2020/10/20	16.209999	13.23843	18.33171958
2020/10/21	16.35	13.24379	18.99820861

从表 2 观察到预测值与真实值之间的误差比较大，并且通过对误差序列进行白噪声检验，发现误差序列不是白噪声序列，误差序列中还存在着未提取的有用信息，还可以对误差序列进行统计分析。同时考虑到 ARIMA-GARCH 模型仅仅考虑了用历史的日收盘价数据预测将来的日收盘价，没有考虑外部因素对于日收盘价的影响，建立的模型进行预测难免会使误差偏大。因此，本文通过主成分分析找出影响误差序列的因素，并对误差序列进行回归建模。

以 ARIMA-GARCH 模型 2020 年 7 月 28 日至 2020 年 10 月 12 日的拟合误差序列为样本内建模数据，把 2020 年 10 月 13 日至 10 月 21 日的误差序列作为样本外数据进行预测。取绿色拇指工业公司对应时间的 Open、High、Low 以及 AdjClose，对这 4 组数据取对数，然后利用变量间的相关性分析方法判断哪些变量与 ARIMA-GARCH 模型误差序列具有相关关系。这 4 组数据与误差序列之间的相关性，P 值均小于显著性水平 $\alpha = 0.05$，判别结果是 Open、High、Low 以及 AdjClose 与 ARIMA-GARCH 模型拟合残差都有一定的相关性。

为了降低自变量的维数，提取影响因素中包含的信息，考虑采用主成分分析法提取具有代表性的综合指标。bartlett 球形检验的结果 P 值小于显著性水平 $\alpha = 0.05$，表明可以进行主成分分析。下方的子图显示了 KMO 检验，整体的与每个变量的检验统计量 MSA 均大于 0.7，说明这 4 组数据非常适合做主成分分析。接下来针对这 4 组数据利用主成分分析法寻找误差序列的解释变量，进而对误差序列进行回归建模。

由主成分分析结果，可以看出，第一主成分所占方差贡献率为 95.22139%，然而第二主成分所占方差贡献率仅为 0.02647444%，因此，第一主成分能较好地反映 4 个一致指标的整体变化，因此本文仅提取第一主成分作为回归方程的解释变量，用 Z 表示。

$$Z = 0.494x_1 + 0.506x_2 + 0.502x_3 + 0.498x_4$$

以 Z 为解释变量，误差序列 ε_t 为被解释变量，拟合回归模型。模型结果显示参数通过了显著性检验并且 $R^2 = 0.9716$，这表明拟合的回归模型是显著的，回归方程：$\hat{\varepsilon}_t = 0.50114Z_t$。

Table 3. Analysis of error prediction results
表 3. 误差预测结果分析

日期	实际值	校正前预测值	校正后预测值	校正前误差	校正后误差
2020/10/13	15.85	13.19401	15.51296	16.75700725	2.1264554
2020/10/14	15.658	13.19860	15.53003	15.70695820	0.8173085
2020/10/15	15.513	13.32459	15.64123	14.10691091	−0.8265740
2020/10/16	15.88	13.28203	15.60944	16.36000485	1.7037535
2020/10/19	15.84	13.20718	15.52507	16.62132793	1.9882029
2020/10/20	16.209999	13.23843	15.58176	18.33171958	3.8756145
2020/10/21	16.35	13.24379	15.60184	18.99820861	4.5758738

利用回归模型给出的误差预测结果，对 ARIMA-GARCH 模型的预测误差结果进行修正，得到最终的股价预测值。从表 3 可以观察到回归模型对误差的预测明显减小了 ARIMA-GARCH 模型预测的误差，从而在一定程度上克服了 ARIMA-GARCH 模型因为考虑外因素不足导致预测误差偏大的缺点。

6. 结论

由于影响股票市场的随机因素有很多，股票价格波动大，表现出复杂的非线性、不确定性，故股票市场是一个很不平稳的动态变化过程，建立一个准确的预测模型是很困难的。仅用时间序列模型不能很好地预测股票价格。因此在对未来股票价格进行预测时，不能只关注于时间序列预测模型给出的预测值，同时要对预测的误差多加关注。因为误差序列中含有预测模型未提取的信息，充分利用误差值里所含有的重要信息，对于提高预测精度是很有帮助的。

本文引入了误差校正的思想，先利用 ARIMA-GARCH 模型对日收盘价进行初步预测，但是预测精度不高，通过对误差序列进行的白噪声检验，发现误差序列存在还有未被 ARIMA-GARCH 模型提取的信息。然后利用变量间的相关性，找出与误差序列相关的变量。接着通过主成分分析对变量进行降维，筛选出合适的解释变量。对解释变量和误差序列进行回归分析，预测误差。最后，将预测误差值与 ARIMA-GARCH 模型的预测值相加，得到最终的预测值。通过将最终的预测值与 ARIMA-GARCH 模型的预测值相比较，观察到预测精度有了很大的提高，进而验证了引入误差校正的方法是合理的。

此论文将 ARIMA-GARCH 模型与回归模型相结合对日收盘价进行了预测，提高预测精度，实现了不仅仅靠历史收盘价数据预测未来价格，还同时考虑了日开盘价、日最高价数据等对日收盘价的影响。综合模型的预测精度相较于单一的 ARIMA-GARCH 模型有所提高，可以为实际的股票预测工作提供一些参考价值。

参考文献

[1] 惠晓峰, 柳鸿生, 胡伟, 等. 基于时间序列 GARCH 模型的人民币汇率预测[J]. 金融研究, 2003(5): 99-105.

[2] 刘国旗. 非线性 GARCH 模型在中国股市波动预测中的应用研究[J]. 统计研究, 2000(1): 49-52.

[3] 杨建辉, 易慧琳. EMD 和 GARCH 模型应用于股票价格预测[J]. 河南科学, 2013(11): 2029-2034.

[4] 张超. 基于误差校正的 ARMA-GARCH 股票价格预测[J]. 南京航空航天大学学报(社会科学版), 2014, 16(3): 43-48.

[5] 杨琦, 曹显兵. 基于 ARMA-GARCH 模型的股票价格分析与预测[J]. 数学的实践与认识, 2016, 46(6): 82-88.

[6] 王燕. 时间序列分析——基于 R [M]. 北京: 中国人民大学出版社, 2015.

[7] Parzen, E. (1961) An Approach to Time Series Analysis. *The Annals of Mathematical Statistics*, **32**, 951-989. https://doi.org/10.1214/aoms/1177704840

[8] Hamilton, J.D. (2020) Time Series Analysis. Princeton University Press, Princeton.