

时间序列平稳性分析的自动机制研究

管河山, 邹清明^①, 邓灵斌^②

(南华大学 经济管理学院, 湖南 衡阳 421001)

[摘 要] 平稳性分析是建立时间序列自回归滑动平均模型的一个预处理过程,已有的分析方法研究颇多,但尚缺乏一个自动判定平稳性的机制。文章在分析自相关函数的基础上,引入非线性转换理论进行改进,并采用聚类方法建立时间序列平稳性分析的自动机制,为大批量时间序列的平稳性自动判定提供了一条新途径。文中采用了一组模拟数据和一组金融数据来进行实证分析,实验表明,平稳性分析的自动机制能得到较好的结果。

[关键词] 时间序列; 平稳性; 自动机制

[中图分类号] TP311 [文献标识码] A [文章编号] 1673-0755(2010)03-0041-03

一 平稳性分析

平稳性分析是建立时间序列自回归滑动平均模型 (autoregressive moving average, 简称为 ARMA) 的一个必要预处理过程,数据非平稳时,大样本下的统计推断基础——“一致性”要求将受到破坏。现实中收集到的时间序列数据大都是非平稳的,需要借助差分方法进行平稳化处理。

现有的时间序列平稳性分析方法较多,比如,时间序列的散点图、自相关函数 (Autocorrelation Function, 简称为 ACF)^[1 2]、偏自相关函数 (Partial Autocorrelation Function)^[1 3 4]、逆自相关函数 (Inverse Autocorrelation Function)^[1 3 4]、交叉相关矩阵序列 (Cross - Correlation Matrix)^[5]、单位根检验和协整检验等方法为平稳性判定提供了有利的支持。尽管目前已有一些时间序列平稳性分析的方法,但尚缺乏一个自动判定平稳性的机制。

本文将采用 ACF 进行分析,给定时间序列样本 $X = (X_1, X_2, \dots, X_n)$, 可根据公式 (1) 计算 ACF 系数

$$\rho_{x,k} = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x}_i)(x_{i+k} - \bar{x}_{i+k})}{\sqrt{\sum_{i=1}^{n-k} (x_i - \bar{x}_i)^2 \sum_{i=1}^{n-k} (x_{i+k} - \bar{x}_{i+k})^2}} \quad (1)$$

其中 $(k=0, 1, 2, \dots, n-1)$, $\bar{x}_i = \frac{1}{n-k} \sum_{p=1}^{n-k} x_p$, $\bar{x}_{i+k} = \frac{1}{n-k}$

$\sum_{i=1}^{n-k} x_{i+k}$ 。计算不同滞后阶数的 ACF 值,记为序列 $\rho_x = \{\rho_{x,0}, \rho_{x,1}, \dots, \rho_{x,n-1}\}$, 显然 $\rho_{x,0} = 1$ 。可以根据序列 ρ_x 的收敛状况判定时间序列 X 的平稳性,如果序列 ρ_x 快速收敛,则时间序列 X 为平稳的,否则为非平稳的。根据 ρ_x 收敛状况判定时间序列的平稳性,只能得到近似结果,因而判定准确

率仍然是个关键问题。我们对 ACF 进行改进,采用非线性转换函数改良 ACF 的性能,建立一种新的平稳性分析方法,称之为自相关函数的非线性趋势特征 (Nonlinear Trend of Autocorrelation Function, 简称为 NTACF)。

对 $\rho_x = \{\rho_{x,0}, \rho_{x,1}, \dots, \rho_{x,n-1}\}$ 进行非线性变换,具体的非线性变换方式如下:

$$\rho_{x,k} = \sum_{i=0}^k |\rho_{x,i}|^2 - \ln(\sum_{i=0}^k |\rho_{x,i}|^2) \quad (2)$$

其中 $(k=1, 2, \dots, n-1)$ 。非线性变换后得到序列 $\rho_x = \{\rho_{x,0}, \rho_{x,1}, \dots, \rho_{x,n-1}\}$, 容易验证,序列 ρ_x 收敛可推出序列 ρ_x 收敛^[6], 因此可根据 ρ_x 的收敛状况来判定时间序列 X 的平稳性。如果 ρ_x 快速收敛,则时间序列 X 是平稳的,反之是非平稳的。这部分的研究我们已经取得了一些成果^[6 7]。本文将继承、拓展我们已有的研究成果,并建立平稳性分析的自动机制。

二 平稳性分析的自动机制

针对大批量时间序列 ARMA 建模的研究需要建立一个自动机制,避免复杂的人工操作,提高建模的效率。我们采用聚类方法实现时间序列的平稳性分类,并建立平稳性分析的自动机制,具体的平稳性自动判定机制包括两大部分:基于 NTACF 的平稳性分析方法和基于聚类方法的自动判定功能。机制可以根据以下算法来实现。

算法: 平稳性自动判定

Step1: 给定时间序列样本集合 $X = \{X_1, X_2, \dots, X_q\}$, 样本个数为 q 个。先随机模拟两组时间序列样本,其中一组是平稳时间序列样本集合 $Y = \{Y_1, Y_2, \dots, Y_p\}$, 另一组是非

[收稿日期] 2010-04-20

[作者简介] 管河山 (1981-), 男, 湖南祁东人, 南华大学经济管理学院讲师, 博士。

①南华大学副教授。②南华大学副教授。

平稳时间序列样本 $Z = \{Z_1, Z_2, \dots, Z_p\}$, 其样本个数都为 p 个。

Step2 将样本集合 Y 和 Z 合并到 X 中, 得到一个新的样本集合 $W = \{W_1, W_2, \dots, W_q, W_{1+q}, \dots, W_{2p+q}\}$ 。

Step3 采用聚类方法 (k -means 算法, 参数 $k = 2$) 将集合 W 中的样本分成两类, 分别记为集合 $W^{(1)}$ 和 $W^{(2)}$, 并进一步判定时间序列样本 W_i 的平稳性, 判定方式如下

$$W_i = \begin{cases} \text{平稳} & \text{if } w_Y^{(1)} > w_Z^{(1)} \\ \text{非平稳} & \text{if } w_Y^{(2)} > w_Z^{(2)} \\ \text{失效} & \text{else} \end{cases} \quad (3)$$

平稳序列集合 Y 在第一类和第二类中的样本个数分别为 $w_Y^{(1)}$ 和 $w_Y^{(2)}$, 非平稳序列集合 Z 在第一类和第二类中的样本个数分别为 $w_Z^{(1)}$ 和 $w_Z^{(2)}$ 。

Step4 如果出现判定失效的情形, 则可改变模拟序列个数 (参数 p) 来分析。

本文模拟一个平稳时间序列集合 $Y = \{Y_1, Y_2, \dots, Y_p\}$ 和一个非平稳时间序列集合 $Z = \{Z_1, Z_2, \dots, Z_p\}$ 用于分析。参数 p 的取值可以根据需要来选择。统计理论上分析, 参数 p 取值越大, 遍历所有可能的平稳序列和非平稳序列, 则得到的模拟样本集合 Y 和 Z 的代表性较强, 分析得到的效果较好。但是考虑到计算的简化问题, 则应该寻找最少量的模拟序列样本进行分析, 即参数取值为 $p = 1$ 。但为了避免模拟模型、模拟样本的片面性及样本分类后出现等分情况, 应坚持对模拟模型个数、单个模拟模型的模拟样本数取奇数原则。本文取最小奇数 3, 此时模拟样本总个数为 $p = 3 \times 3 = 9$ 个样本。当判定失效时, 则可以将参数 p 的取值进行逐次递增, 递增过程中仍坚持“奇数原则”, 然后进行分析。根据以上算法, 可以实现大批量时间序列的平稳性自动判定, 为进一步进行 ARMA 建模奠定了基础。

表 1 模拟时间序列集合

模型	平稳时间序列集合 Y
1	AR(1), 其中 $\varphi_1 = 0.9$
2	ARMA(1, 1), 其中 $\varphi_1 = 0.95, \varphi_2 = 0.1$
3	MA(1), 其中 $\theta_1 = -0.9$
模型	非平稳时间序列集合 Z
1	ARMA(1, 1, 0), 其中 $\varphi_1 = -0.1$
2	ARMA(0, 1, 1), 其中 $\theta_1 = 0.1$
3	ARMA(1, 1, 1), 其中 $\varphi_1 = 0.1, \theta_1 = -0.1$

本文进行时间序列模拟时, 具体的模拟方式如表 1。模拟的时间序列维度分别为 100、200、500、1000 和 5000。文中将采用这个模拟数据集进行实证分析。

三 实证分析

本文例举了 2 个数据集进行实证分析, 从实践角度检验 NTACF 方法的有效性, 即验证 NTACF 方法对 ACF 方法的改进力度, 及其自动机制的有效性。

(一) 模拟数据

本文参考了文献 [1] 中的实验方法。生成 6 组平稳时间序列 $[(a) - (f)]$ 和 6 组非平稳时间序列 $[(g) - (l)]$, 每组时间序列各有 100 个时间序列样本, 共计 1200 个序列样本, 构成模拟时间序列样本集合 X 。共进行 5 次模拟实验, 模拟的时间序列维度大小分别为 100、200、500、1000 和 5000。

将事先模拟得到的序列样本集合 Y 和 Z 合并到模拟数据集 X 中, 得到 1218 个序列样本。然后采用 ACF 方法和 NTACF 方法分别进行分析, 采用本文的自动机制判定序列的平稳性, 将所有时间序列样本自动分成平稳时间序列和非平稳时间序列, 根据公式 (3) 容易判定出每个时间序列的平稳性特征, 这验证了自动机制在实践中的可行性。

进一步分析平稳性判定的结果, 该模拟数据得到的 1200 个样本, 其样本的平稳性分析结果都是已知的, 而采用本文的平稳性分析自动机制进行分析, 也可以得到 1200 个样本的平稳性分析结果, 对这两个平稳性分析结果进行对比研究, 容易得到平稳性判定的准确率。本文通过统计分析正确的样本个数, 得到的结果如表 2 所示。显然, NTACF 方法得到的准确率较高, NTACF 方法对时间序列平稳性的判定更加精确。

表 2 准确率分布, 总样本数为 1200

维度 \ 方法	ACF	NTACF
n= 100	996	1002
n= 200	1004	1105
n= 500	993	1128
n= 1000	997	1132
n= 5000	1000	1106

(二) 金融股票数据

收集中国股市中 100 支股票的历史数据, 时间段为 2005.5.7-2006.2.20 (时间序列维度取值为 200), 取股票每日的历史价格数据进行分析。采用 ACF 方法和 NTACF 方法分别进行分析, 然后采用本文的自动机制判定序列的平稳性, 将所有样本自动分成平稳时间序列和非平稳时间序列, 并识别出每个时间序列样本的平稳性特点。

进一步分析平稳性判定的结果, 分析两种方法的平稳性判定结果, 共有 8 个股票数据的判定结果不一致, 其中 ACF 方法判定为非平稳, NTACF 方法判定为平稳。

对 ACF 方法判定的结果, 采用一阶差分方法来处理这些数据, 然后进行时间序列 AR 建模, 对 NTACF 方法判定的结果, 直接进行时间序列 AR 建模。求解模型的误差, 得到的结果如表 3。从表 3 中分析, NTACF 方法得到的建模结果中有 5 个误差取值相对小, 而 ACF 方法得到的建模结果中有 3 个误差取值相对小, 因此总体来说, NTACF 判定结果要比 ACF 判定结果更加有效。

表 3 模型的误差

股票代码	Acf方法	ntacf方法
SS600100	0.090134	0.093214
SS600200	0.116801	0.112598
SS600656	0.008695	0.008415
SS600841	0.059409	0.056147
SS600742	0.004825	0.004637
SS600138	0.037609	0.038725
SS600266	0.040159	0.040028
SS600728	0.007109	0.008333

四 结论

本文在 ACF方法的基础上, 引入非线性转换理论改良 ACF的性能, 并采用聚类方法建立时间序列平稳性分析的自动机制。大量实验表明, 该方法为大批量时间序列的平稳性自动判定提供了一个新的途径, 大大提高了平稳性判定的效率。完成时间序列的平稳性分析后, 还需要对时间序列差分的自动机制进行研究, 以便进一步完成时间序列的 ARMA 建模的工作, 因此时间序列差分的自动机制将是下一步的研究重点。

完成平稳性分析自动机制和差分自动机制的研究后, 借助 ARMA 建模的相关参数估计理论便可以进行大批量时间

序列的自动建模工作, 这将值得进一步深入研讨。

[参考文献]

[1] Jorge C, Nuno C, Daniel P. A periodogram – based metric for time series classification[R]. Computational Statistics & Data Analysis 50, 2006, 2668– 2684.

[2] Galeano, P., Peña, D. Multivariate analysis in vector time series[J]. Resenhas 4, 2000, 383 – 404.

[3] Cleveland, W.S. The inverse autocorrelations of a time series and their applications[J]. Technometrics 14, 1972, 277 – 293.

[4] Chatfield, C. Inverse autocorrelations[J]. J. Roy. Statist. Soc. Ser. A 142, 1979, 363 – 377.

[5] Guan HS, Jiang QS, Hong ZL, A New Metric for Classification of Multivariate Time Series[Z]. The 4th International Conference on Fuzzy Systems and Knowledge Discovery, Hainan, China, 2007, 453– 457.

[6] 管河山, 姜青山, Shengrui Wang, 郑宇泉. 基于 ACF 非线性趋势特征的时间序列聚类 [J]. 计算机研究与发展, 2007, 44(suppl): 111– 116.

[7] 管河山. 金融多元时间序列挖掘方法研究与应用 [D]. 厦门大学博士学位论文, 2008.

Research on Automatic Mechanism for Time Series Stationarity Analysis

GUAN He-shan, ZOU Qing-ming, DENG Ling-bin
(University of South China, Hengyang 421001, China)

Abstract Stationarity analysis has been widely studied, which is a pre-process of time series ARMA modeling, but there is no automatic mechanism for stationarity analysis. In this paper, we better autocorrelation function with the non-linear translation theory. The automatic mechanism for stationarity analysis is also proposed using time series clustering. We utilize two datasets to experiment in this paper: one financial data and one synthetic. Empirical evidence has strongly suggested that our method is shown to yield useful and get robust result.

Key words time series, stationarity, automatic mechanism