Kharazmi University

Mathematical Researches

Homepage: https://mmr.khu.ac.ir/

Print ISSN: 2588-2546
Online ISSN: 2588-2554

Identification of outliers in time series with VARMA models using genetic algorithm

Masoud Yarmohammadi[1] ✉, **Ahad Rahimpoor**[2]

1. Department of Statistics, Payame Noor University, Tehran, Iran. ✉E-mail: masyar@pnu.ac.ir

2. Department of Statistics, Payame Noor University, Tehran, Iran. E-mail: rahimpoorA@gmail.com

Article Info

Article type:
Research Article

Article history:
Received:
20 June 2020
Received in revised form:
26 January 2021
Accepted:
31 January 2021
Published online:
31 December 2022

Keywords:
Multivariate time series,
Outlier,
Detection,
Genetic algorithms.

ABSTRACT

Introduction

Time series data may be contaminated with different types of outliers. i.e. additive outliers, innovation outliers, level shift and temporary changes. Of course, the effects of these outliers in univariate and multivariate time series are different and detection of outliers in multivariate data are more complicated. Multivariate time series often, modeled using vector autoregressive moving average (VARMA) model and presence of outliers can violate the stationarity assumptions and may lead to wrong modeling, biased estimation and inaccurate prediction. Thus, detection and properly dealing with these data, especially in relation to modeling and parameter estimation of VARMA model is necessary. By detection of outliers, their effects can be removed or robustified and the adjusted data could be prepared for further analysis.

The multivariate detection methods, suggested by Baragona and Battaglia (2007) and Cucina, et al. (2014), could not detect the innovational outliers and temporary changes. In this paper, after introducing the VARMA models, the principles of genetic algorithm and its application in detecting outliers are discussed. Then, the Tsay, Pena and Pankratz (TPP) method (Tsay, et al.(2000)) and the genetic algorithm (GA) for detecting different types of outliers in multivariate time series are described. Also, the efficiency of GA and TPP detection methods is discussed using simulation studies and real data. At the end, a conclusion is presented.

Material and methods

A genetic algorithm makes uses of techniques inspired from evolutionary biology such as selection, mutation, inheritance and recombination to solve a problem. The most commonly employed method in genetic algorithms is to create a group of individuals randomly from a

given population. The individuals thus formed are evaluated with the help of the evaluation function provided by the programmer. Individuals are then provided with a score which indirectly highlights the fitness to the given situation. The best two individuals are then used to create one or more offspring, after which random mutations are done on the offspring. Depending on the needs of the application, the procedure continues until an acceptable solution is derived or until a certain number of generations have passed.

For minimization of a function, GA operates by first generating, at random or optionally, several minimal solutions to the function, where this set of solutions called the initial population and each solution as a chromosome. Then, using reproductive operators, we combine chromosomes and make a jump into them. If the function of newly produced chromosomes is lower than the previous chromosomes, these chromosomes can be added to the initial population or replaced with chromosomes with less function in this population. This process is repeated until convergence occurs or the end number of iteration obtained.

Furthermore, we introduce another method for detecting outliers, i.e. the Tsay Pena and Pankratz (TPP) method. TPP uses some test statistics based on outlier's size and vector autoregressive (VAR) parameters. This method detects outliers in three stages. In stage I, it detects one by one outliers and remove their effects. Iteration continues until no outlier found. In stage II, for detected outlier in stage I, the estimation of outliers effects is obtained simultaneously. Then, outliers with insignificant effects are removed. The VAR parameters re-estimated based on modified series of this stage. In stage III, we repeated stage I and II with new VAR parameters estimation.

In each iteration of TPP, an outlier is detected and removed from the series. Then the parameter estimation is obtained from the modified series and the next outlier detection is continued using these estimates. This may lead to biased estimates and wrong detection of the next outlier point. In other words, in the TPP method, one detected outlier hides another outlier (masking), or one detected outlier reveals the usual observation as an outlier (swamping). This method often undetects the type of outliers. But in each iteration of GA, a random pattern of outliers (for testing) is first generated and a temporary modified series is obtained by removing effect of this pattern from series. Then the estimation of the parameters obtained and the detection of this pattern is tested. This method reduces the effect of the previously identified outliers on the full pattern of them. In fact, if the random pattern of all outliers is correctly generated, almost effect of them will be eliminated in the modified series. Therefore, using this temporary modified series, the GA obtains more accurate estimates and detects outliers more accurately.

Results and discussion

The simulation results confirm the validity of the GA method and the percentage of correct outlier detection in this method is higher than the TPP method. GA, of course, needs more time to calculate. Also, although the VAR model is used in both detection methods, the percentage of correct outlier detection in the VARMA model data is similar to the VAR model.

The Gas-furnace data set, called Series J by Box and Jenkins (1994), contains sequentially recorded measurements of two variables (gas rate and CO_2) were analyzed and modeled. It was determined that GA and TPP methods detect similar outliers. Fitting the VAR (6) model to these data shows that the variance of input gas error in modified data of GA is reduced by 17% as compared with TPP and the variance of carbon dioxide error in the modified data of GA is reduced by 43% respectively.

Conclusion

The following conclusions were obtained from this research:

- In each iteration of the TPP outlier detection method, an outlier is detected and the effect of this point is adjusted. Then the estimation of the parameters is obtained from the modified series and the next point detection continues using these estimates. This may lead to bias of estimates and misdiagnosis of the next point. In other words, using the TPP method, one detected outlier hides another outlier.
- In the proposed method based on the genetic algorithm in each iteration, first a random design is generated from all outliers and the modified series is obtained from the same design. This reduces the effect of previously detected outliers on revealing the complete design of the outliers. In fact, if the random design of all outliers are generated correctly, the effect of them will be eliminated in the modified series. Thus, with this modified series, GA obtains more accurate estimates and detects outliers more correctly.
- The simulation results confirm the accuracy of the GA method and the percentage of correct detection of outlier in this method is higher than the TPP approach. Of course, GA requires more time for calculations.

How to cite: Yarmohammadi, M., Rahimpoor, A. (2022). Identification of outliers in time series with VARMA models using genetic algorithm. *Mathematical Researches*, 8 (4), 256-283.

Publisher: Kharazmi University

Mathematical Researches
Print ISSN: 2588-2546
Online ISSN: 2588-2554
Kharazmi University
Homepage: https://mmr.khu.ac.ir/

شناسایی نقاط دورافتاده در سری‌های زمانی دارای مدل VARMA با استفاده از الگوریتم ژنتیک

مسعود یارمحمدی[۱]✉، احد رحیم‌پور[۲]

۱. نویسندهٔ مسئول، گروه آمار، دانشگاه پیام نور، تهران، ایران. رایانامه: masyar@pnu.ac.ir
۲. گروه آمار، دانشگاه پیام نور، تهران، ایران. رایانامه: rahimpoorA@gmail.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۳۹۹/۰۴/۰۹
تاریخ بازنگری: ۱۳۹۹/۱۱/۰۷
تاریخ پذیرش: ۱۳۹۹/۱۱/۱۲
تاریخ انتشار: ۱۴۰۱/۱۰/۱۰

واژه‌های کلیدی:
سری زمانی چند متغیره،
نقطه دورافتاده،
آشکارسازی،
الگوریتم ژنتیک.

چکیده

در تحلیل سری‌های زمانی چند متغیره، نقاط دورافتاده می‌توانند منجر به شناسایی غلط مدل، برآورد اریب پارامترها و پیش‌بینی‌های ضعیف شوند. لذا آشکارسازی این نقاط بسیار مهم بوده و مورد توجه می‌باشد. در این تحقیق، روش آشکارسازی جدیدی بر اساس الگوریتم ژنتیک در سری‌های زمانی دارای مدل VARMA استفاده می‌شود. این روش علاوه بر پیدا کردن مکان نقاط دورافتاده، شناسایی نوع دورافتادگی این نقاط نیز انجام می‌شود. سپس به معرفی روش تسای، پناه و پانکراتز (TPP) پرداخته و با مطالعات شبیه‌سازی نشان می‌دهیم که درصد آشکارسازی صحیح نقاط دورافتاده در الگوریتم ژنتیک نسبت به روش TPP بیشتر است. همچنین داده‌های مربوط به گاز-کوره بررسی و مدل‌بندی شده و مشخص شد که روش‌های الگوریتم ژنتیک و TPP، نقاط دورافتاده مشابهی را آشکار می‌سازند.

استناد: یارمحمدی، مسعود؛ رحیم‌پور، احد؛ (۱۴۰۱). شناسایی نقاط دورافتاده در سری‌های زمانی دارای مدل VARMA با استفاده از الگوریتم ژنتیک. *پژوهش‌های ریاضی*، ۸ (۴)، ۲۸۳-۲۵۶.

ناشر: دانشگاه خوارزمی

مقدمه

استفاده از مدل‌های میانگین متحرک اتورگرسیو چند متغیره[1] (VARMA) برای تحلیل داده‌های سری زمانی چند متغیره بسیار متداول است. نظیر حالت یک متغیره، سری‌های زمانی چند متغیره نیز ممکن است دارای یک یا چند مشاهدهٔ دورافتاده[2] بوده که از سایر مشاهدات مجزا شده یا از طرح کلی داده‌ها منحرف شده‌اند. وجود نقاط دورافتاده در این نوع داده‌ها، ناقض فروض مانایی بوده و معمولاً منجر به شناسایی غلط[3] مدل، اریبی برآورد پارامترها و پیش‌بینی‌های ضعیف می‌شود. بنابراین، آشکارسازی[4] این نقاط و نحوهٔ برخورد صحیح با آنها خصوصاً در رابطه با مدل‌بندی و برآورد پارامترهای مدل VARMA ضروری به نظر می‌رسد.

با آشکارسازی این نقاط، اثر آنها را می‌توان از سری زمانی حذف کرده و داده‌های تعدیل شده را به دست آورد. بعد از تعدیل داده‌ها، برآوردهایی از مدل VARMA به دست می‌آید که انتظار می‌رود کمترین تأثیرپذیری را از نقاط دورافتاده داشته باشند. از طرف دیگر، گاهی اوقات آشکارسازی نقاط دورافتاده برای پیدا کردن یک رویداد تأثیرگذار خارجی در طی زمان مهم است. به عنوان مثال با پیدا کردن نقاط دورافتاده در داده‌های میزان آب رودخانه می‌توان زمان‌های وقوع سیل را به دست آورد [۶].

مشاهدات چند متغیره سری زمانی ممکن است با انواع مختلف نقاط دورافتاده آلوده شده باشند. این نقاط در انواع جمع‌پذیر، نوساز، تغییر سطح و تغییر موقت بوده که به ترتیب به معنای جهش یک نقطه سری، جهش در خطای مدل یک نقطه، جهش همه نقاط سری بعد از یک نقطه و جهش در یک نقطه همراه با کاهش نمایی بعد از آن می‌باشند. البته تأثیر گونه‌های مختلف نقاط دورافتاده در حالت چند متغیره و یک متغیره متفاوت بوده و این مشاهدات باید به روشی چند متغیره بررسی شوند. اما اغلب روش‌های آشکارسازی چند متغیره مانند باراگونا و باتاگلیا [۱] و کاسیانا و همکاران [۶] نقاط دورافتاده نوساز و تغییر موقت را بررسی نمی‌کنند. لذا در این تحقیق، الگوریتم ژنتیک کاسیانا و همکاران [۶] برای شناسایی انواع مختلف نقاط دورافتاده مورد تعمیم قرار می‌گیرد. البته، بهتر بودن این روش نسبت به روش آشکارسازی چند متغیره تسای و همکاران [۹] برای انواع مختلف این نقاط نشان داده می‌شود.

در ادامه این مقاله نخست مدل‌های VARMA در حالت کلی و هنگامی که به داده‌های دورافتاده آلوده می‌شوند، معرفی می‌شوند. سپس کلیات الگوریتم ژنتیک و کاربرد آن در آشکارسازی نقاط دورافتاده مورد بحث و بررسی قرار می‌گیرد. در ادامه، روش تسای، پنا و پانکراتز (TPP) و الگوریتم ژنتیک[5] (GA) برای آشکارسازی (شناسایی[6]) انواع نقاط دورافتاده در سری‌های زمانی چند متغیره بیان می‌شود. همچنین میزان کارایی روش‌های شناسایی GA و TPP با استفاده از مطالعات شبیه سازی و داده‌های واقعی مورد بحث و بررسی قرار می گیرد. در پایان نتیجه گیری ارائه می‌شود.

[1] Vector Autoregresive Moving Average
[2] Outlier
[3] Misspecification
[4] Detection
[5] Genetic Algorithm
[6] Identification

ساختار مدل VARMA در حالت کلی و هنگام آلوده شدن به نقاط دورافتاده

فرض کنید $\boldsymbol{Y}_t = (Y_{1t}, Y_{2t}, \ldots, Y_{mt})'$ برای $t = 0, \pm1, \pm2, \pm3, \ldots$ ، یک فرایند m متغیرهٔ دارای مدل VARMA(p,q) به صورت زیر باشد

$$\Phi(B)\boldsymbol{Y}_t = \Theta(B)\boldsymbol{\varepsilon}_t \qquad (۱)$$

که در آن بردار $\boldsymbol{\varepsilon}_t$، فرایند اغتشاش خالص از توزیع نرمال m-متغیره با میانگین چند متغیره صفر و ماتریس کواریانس Σ و عملگرهای $\Phi(B)$ و $\Theta(B)$ چندجمله‌ای‌های ماتریسی از مرتبه p و q به صورت

$$\Phi(B) = I - \Phi_1 B - \cdots - \Phi_p B^p \quad , \quad \Theta(B) = I - \Theta_1 B - \cdots - \Theta_q B^q$$

می‌باشند [۷]. در این‌جا B عملگر پسرو به صورت $B\boldsymbol{Y}_t = \boldsymbol{Y}_{t-1}$ است.

مثال ساده مدل VARMA(1,1) با دو متغیر به صورت

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \end{bmatrix} - \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \begin{bmatrix} Y_{1(t-1)} \\ Y_{2(t-1)} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} - \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1(t-1)} \\ \varepsilon_{2(t-1)} \end{bmatrix}$$

است.

مدل VARMA برای $p = 0$ به مدل میانگین متحرک چند متغیره (VMA):

$$\boldsymbol{Y}_t = \boldsymbol{\varepsilon}_t - \Theta_1 \boldsymbol{\varepsilon}_{t-1} - \ldots - \Theta_q \boldsymbol{\varepsilon}_{t-q} \qquad (۲)$$

و برای حالت $q = 0$ به مدل اتورگرسیو چند متغیره (VAR):

$$\boldsymbol{Y}_t = \Phi_1 \boldsymbol{Y}_{t-1} + \ldots + \Phi_p \boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t \qquad (۳)$$

تبدیل می‌شود.

شرط کافی برای مانایی فرایند VARMA آن است که برای $|z| \leq 1$، $\det(I - \Phi_1 z - \cdots - \Phi_p z^p) \neq 0$ باشد. در حالت $p = 1$ شرط مانایی این فرایند آن است که قدر مطلق مقادیر ویژه ماتریس Φ_1 کمتر از ۱ باشد.

نمایش میانگین متحرک مدل VARMA مانا به صورت:

$$\boldsymbol{Y}_t = [\Phi(B)]^{-1}\Theta(B)\boldsymbol{\varepsilon}_t = \Psi(B)\boldsymbol{\varepsilon}_t$$

است. همچنین شرط وارون پذیری این فرایند آن است که برای $|z| \leq 1$، $\det(I - \Theta_1 z - \cdots - \Theta_q z^q) \neq 0$ باشد.

نمایش اتورگرسیو فرایند Y_t وارون پذیر به صورت:

$$[\Theta(B)]^{-1}\Phi(B)\boldsymbol{Y}_t = \Pi(B)\boldsymbol{Y}_t = \boldsymbol{\varepsilon}_t$$

بوده و اغلب در یافتن پیش‌بینی فرایند کاربرد دارد.

مدل VARMA آلوده شده به نقاط دورافتاده

فرض کنید $\boldsymbol{Y}_t$ یک سری زمانی m بعدی، مانا و وارون پذیر از مدل VARMA(p,q) باشد. حال فرض کنید سری زمانی چند متغیره $\boldsymbol{Y}_t$ تحت تأثیر یک ضربه یا رویداد خارجی قرار گرفته و مدل آن به صورت

$$\boldsymbol{Z}_t = \boldsymbol{Y}_t + f(t)$$

تغییر یافته است، که در آن $\boldsymbol{Z}_t$ سری آلوده شده و $f(t)$ تابع پارامتری و نشان دهندهٔ اثر مداخله می‌باشد. با توجه به نوع اثرهای مداخله‌ای این تابع ممکن است به صورت مقدار ثابت یا متغیر تصادفی باشد که در عمل توسط تحلیلگر داده‌ها یا براساس اطلاع از اثرهای مداخله‌ای بر روی فرآیند $\boldsymbol{Z}_t$ مشخص می‌شود. اثر مداخله خارجی را می‌توان به صورت زیر نوشت:

$$\boldsymbol{Z}_t = \boldsymbol{Y}_t + \alpha(B)\boldsymbol{\omega}\xi_t^{(h)} \qquad (۴)$$

که در آن $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_m)'$ اثر تداخل اولیه، $\alpha(B)$ بیانگر چگونگی تأثیر تداخل روی زمان‌های دیگر سری زمانی و ξ_t^h متغیر شاخصی است که زمان تداخل h را در سری زمانی نشان می‌دهد. $\boldsymbol{Z}_t$ تعریف شده در (۴) دارای مدل VARMA آلوده شده به نقطه دورافتاده در زمان h می‌باشد. با توجه به ساختار $\alpha(B)$، انواع متفاوت نقاط دورافتاده مشخص می‌شود.

نقطهٔ دورافتاده جمع پذیر (AO[۷]): با توجه به مانایی فرایند $\boldsymbol{Y}_t$، در صورتی که $\alpha(B) = I$ باشد

$$\boldsymbol{Z}_t = \{\Phi(B)\}^{-1}\Theta(B)\boldsymbol{\varepsilon}_t + \boldsymbol{\omega}\xi_t^h$$

در این حالت سری زمانی تنها در نقطه h جهشی به اندازه $\boldsymbol{\omega}$ داشته و در سایر نقاط بدون تغییر باقی می‌ماند.

نقطهٔ دورافتاده نوساز (IO[۸]): اگر در رابطه (۴)

$$\alpha(B) = \Psi(B) = \{\Phi(B)\}^{-1}\Theta(B) = \left(I + \sum_{i=1}^{\infty} \Psi_i B^i\right) \qquad (۵)$$

باشد نقطه دورافتاده را نوساز می‌نامند. در این صورت با توجه به $\alpha(B)$ ارائه شده در رابطه (۵)، $\boldsymbol{Z}_t$ به صورت زیر در می‌آید:

$$\boldsymbol{Z}_t = \{\Phi(B)\}^{-1}\Theta(B)(\boldsymbol{\varepsilon}_t + \boldsymbol{\omega}\xi_t^h). \qquad (۶)$$

با توجه به رابطه (۶) برای نقطه دورافتاده نوساز با پرش فرایند اغتشاش تصادفی در زمان h مواجه هستیم. همچنین با توجه به ساختار $\alpha(B)$ در (۵) و مانایی مدل VARMA، اثر این تداخل برای نقطه hام به اندازه $\boldsymbol{\omega}$ بوده و بعد از زمان h کاهش می‌یابد.

---

[7] Additive Outlier
[8] Innovation Outlier

نقطهٔ دورافتاده تغییر سطح (LS[9]): اگر $\alpha(B) = (1-B)^{-1}I = (I + IB^2 + IB^3 + \cdots)$ باشد مدل سری آلوده شده به صورت

$$\boldsymbol{Z}_t = \{\Phi(B)\}^{-1}\Theta(B)\boldsymbol{\varepsilon}_t + (I-B)^{-1}\boldsymbol{\omega}\xi_t^h$$

در می‌آید. به دلیل تأثیر

$$(1-B)^{-1}\boldsymbol{\omega}\xi_t^h = \boldsymbol{\omega}\xi_t^h + \boldsymbol{\omega}\xi_{t-1}^h + \boldsymbol{\omega}\xi_{t-2}^h + \ldots = \boldsymbol{\omega}\xi_{t>h}$$

بر فرآیند VARMA در رابطه (1) ، از زمان h تا پایان سری، تغییر سطحی به اندازه $\boldsymbol{\omega}$ در سری ایجاد می‌شود. به عبارت دیگر، اثر این نقاط دورافتاده به طور دائمی از زمان h تا انتهای سری ادامه می‌یابد. البته این امکان هم وجود دارد که تغییر سطح تکه‌ای[10] بوده و بعد از چند زمان، سطح سری به حالت قبل برگردد. این حالت تغییر سطح را **نقطه دورافتاده تکه‌ای** هم می‌گویند.

نقطهٔ دورافتاده تغییر موقت (TC[11]): وقتی اثر نقطه دورافتاده به صورت نمایی کاهش یابد، نقطه دورافتاده را تغییر موقت می‌نامند. در این صورت $\alpha(B) = \{D(\delta)\}^{-1}$ که در اینجا $D(\delta)$ یک ماتریس قطری با عناصر $\left(1-\delta_1 B\right), (1-\delta_2 B), \ldots, (1-\delta_k B)$ به طوری که $0 < \delta_i < 1$ است و معمولاً برای سادگی $\delta_1 = \delta_2 \ldots = \delta_k = \delta$ در نظر می‌گیرند. پس $D(\delta) = (1-\delta B)I$ و

$$\boldsymbol{Z}_t = \{\Phi(B)\}^{-1}\Theta(B)\boldsymbol{\varepsilon}_t + (1-\delta B)^{-1}\boldsymbol{\omega}\xi_t^{(h)}, \qquad 0 < \delta < 1$$

توجه داشته باشید که $\left(1-\delta B\right)^{-1}\boldsymbol{\omega}\xi_t^h = \boldsymbol{\omega}\xi_t^h + \boldsymbol{\omega}\delta\xi_{t-1}^h + \boldsymbol{\omega}\delta^2\xi_{t-2}^h + \cdots$. بنابراین اگر در زمان h نقطه دورافتاده تغییر موقت وجود داشته باشد، برای $t \geq h$ داریم:

$$\boldsymbol{Z}_t = \{\Phi(B)\}^{-1}\Theta(B)\boldsymbol{\varepsilon}_t + \boldsymbol{\omega}\delta^{t-h}$$

یعنی از زمان h به بعد سری با تغییر سطحی روبرو است که با افزایش زمان و با نرخ δ کاهش یافته تا جایی که در انتهای سری اثر این نقطه دورافتاده از بین رود. نقطه دورافتاده در حالت جمع پذیر و تغییر سطح به ترتیب حالت حدی تغییر موقت برای $\delta = 1$ و $\delta = 0$ می‌باشند.

شکل ۱ بخش مطالعات شبیه‌سازی، سری زمانی شبیه سازی شده از مدل VARMA(1,1) با نقاط دورافتاده AO ، IO ، LS ، TC را نشان می‌دهد. با توجه به این شکل، در نقطه دورافتاده AO زمان ۲۰ هر دو متغیر Y_1 و Y_1، پرش فقط یک نقطه‌ای به اندازه $\omega_i = 10$ به سمت بالا داشته و سپس سری به حالت گذشته خود برگشته است. در نقطه دورافتاده LS زمان ۱۲۵ هر دو متغیر پرشی به اندازه $\omega_i = 10$ به سمت بالا داشته سپس تا انتهای سری در همان سطح باقیمانده‌اند. همچنین، هر دو متغیر در نقاط دورافتاده IO و TC به ترتیب در زمان‌های ۵۰ و ۱۰۰، پرشی به اندازه $\omega_i = 10$ به سمت بالا داشته سپس سطح سری به آرامی به حالت اولیه خود برگشته است. تفاوت این دو نقطه دورافتاده در این است که در نقطه دورافتاده IO نرخ کاهش بعد از زمان ۵۰ بر اساس پارامترهای نمایش اتورگرسیو (۵) به دست

[9] Level shift Outlier
[10] Patch
[11] Temporary change Outlier

می‌آید ولی در نقطه دورافتاده TC نرخ کاهش بعد از زمان ۱۰۰ به صورت نمایی $10 \times 0/7^{t-100}$ می‌باشد.

کلیات الگوریتم ژنتیک و کاربرد آن در آشکارسازی نقاط دورافتاده

الگوریتم ژنتیک (GA)، الگوریتم بهینه‌سازی عددی است که ایدهٔ آن برگرفته از طبیعت می‌باشد. این الگوریتم قابلیت خود را در حل بسیاری از مسائل پیچیده از قبیل برنامه‌های رایانه‌ای تکاملی، تحلیل داده‌ها و پیش‌بینی، ماکسیمم (مینیمم) سازی توابع و ... نشان داده در حالیکه بسیاری از روش‌های بهینه سازی با مشکل مواجه شده‌اند [۵]. این الگوریتم برای به دست آوردن مقدار بهینهٔ یک تابع به فرض‌های قوی نیاز نداشته و توانایی جستجوی جواب بهینه از فضایی با چندین بهینهٔ محلی را داراست. یعنی مثلا اگر تابعی چندین ماکسیمم (مینیمم) نسبی داشته باشد، این الگوریتم به خوبی ماکسیمم (مینیمم) مطلق این تابع را پیدا می‌کند. برای جزئیات بیشتر GA در دیگر مسائل به [۸] و [۳] مراجعه کنید.

در این مقاله، آشکارسازی نقاط دورافتاده با استفاده از مینیمم سازی معیار اطلاع آکائیک[۱۲] توسط GA انجام می‌شود. چون با توجه به رابطهٔ (۴)، برای هر نقطه دورافتاده یک پارامتر مربوط به اثر $\boldsymbol{\omega}$ این نقطه نیز وجود دارد پس در روش GA با مینیمم‌سازی معیار اطلاع آکائیک سعی در "مینیمم کردن تعداد نقاط دورافتاده" و "ماکسیمم کردن تابع درستنمایی" داریم.

در حالت کلی GA در مسأله مینیمم‌سازی یک تابع بدین صورت عمل می‌کند که ابتدا به طور تصادفی یا اختیاری، چندین جواب مینیمم برای تابع تولید کرده و این مجموعه جواب را جمعیت اولیه و هر جواب را یک کروموزوم می‌نامیم. سپس با انتخاب کروموزوم‌هایی از جمعیت اولیه و استفاده از عملگرهای تولید مثل، این کروموزوم‌ها را باهم ترکیب کرده و جهشی در آنها ایجاد می‌کنیم. در صورتی که مقدار تابع کروموزم‌های تولید شده جدید کمتر از کروموزوم‌های قبلی باشد این کروموزوم‌ها را می‌توان به جمعیت اولیه اضافه کرد یا جایگزین کروموزوم‌هایی با مقدار تابع کمتر در این جمعیت نمود. به عبارت دیگر، جمعیت جدید دارای جواب‌هایی است که از ترکیب و جهش جواب‌های اولیه به دست آمده و مقدار تابع این جواب‌ها کمتر از جواب‌های اولیه می‌باشد. با این رویکرد، مقدار تابع جواب‌های تولید شده ثابت مانده یا کاهش می‌یابد. این فرایند تکرار می‌شود تا زمانی که تعداد تکرار به انتها رسیده یا همگرایی رخ دهد. کروموزم‌های مینیمم کننده هم به عنوان یکی از کروموزم‌های جمعیت نهایی به دست می‌آید. به طور کلی در GA موارد زیر در نظر گرفته می‌شود:

۱) کروموزوم اغلب به صورت کدهای رشته‌ای ۰ و ۱ بیان می‌شود. در مسأله مربوط به وجود نقاط دورافتاده، هر کروموزوم نشانگر طرحی از این نقاط در نمونه بوده و به صورت بردارهایی شبیه زیر تعریف می‌شود:

$$\boldsymbol{\xi}' = (0,0,1,0,\dots,0,0)_{T\times 1}$$

این کروموزوم به معنای وجود نقطه دورافتاده در مشاهده سوم از T مشاهده می‌باشد.

۲) تولید کروموزوم‌ها از ترکیب و جهش اعضای جمعیت اولیه به دست می‌آید. جمعیت اولیه می‌تواند به‌صورت تصادفی یا اختیاری تولید شود. مثلا می‌توان در مسأله مربوط به وجود نقاط دورافتاده چند مورد (ژن[۱۳]) را به تصادف نقطه

[12] Akaike information criterion
[13] Gen

دورافتاده فرض نمود و یا می‌توان جمعیتی را در نظر گرفت که در آن هر نقطهٔ نمونه یک نقطه دورافتاده باشد. یعنی اعضای جمعیت اولیه، T بردار

(۷)

$$\xi_1' = (1,0,0,0,\dots,0,0)_{T\times 1}$$
$$\xi_2' = (0,1,0,0,\dots,0,0)_{T\times 1}$$
$$\vdots$$
$$\xi_T' = (0,0,0,0,\dots,0,1)_{T\times 1}$$

باشند. اگر در ابتدا جواب‌های پیشنهادی مینیمم کننده تابع وجود دارد می‌تواند به جمعیت اولیه اضافه شود. برای مثال اگر نقاط دورافتادهٔ اولیه‌ای با بررسی نمودار داده‌ها مشاهده شوند، آنها را می‌توان به عنوان یک جواب پیشنهادی در جمعیت اولیه استفاده نمود.

۳) تابع هدف تابعی است که با توجه به شرایط مسأله به دنبال بهینه کردن آن می‌باشیم. برای مثال در پژوهش حاضر به دنبال مینیمم کردن معیار آکائیک به عنوان تابع هدف هستیم. مقدار تابع هدف را برای هر کروموزوم، مقدار برازندگی[۱۴] (یا هزینه) آن کروموزوم می‌نامیم. در واقع، GA در هر مرحله با تولید کروموزوم‌ها (یا جواب‌ها) و محاسبهٔ برازندگی آنها سعی در یافتن کروموزوم‌هایی می‌باشد که برازندگی کمتری دارند. با تکرار تولید کروموزم و پیدا کردن جواب‌های بهتر، برازندگی اعضای جمعیت کمتر شده و معمولاً در نهایت به جواب مینیمم می‌رسد.

۴) در هر مرحله از روش GA ، اغلب دو کروموزوم (والدین[۱۵]) از جمعیت انتخاب شده و از آنها، دو کروموزوم جدید (فرزندان[۱۶]) تولید می‌شود و برازندگی آنها محاسبه می‌شود. در صورت کمتر شدن برازندگی کروموزوم‌های تولیدی، این کروموزوم‌ها به جمعیت اولیه اضافه شده یا جایگزین کروموزوم‌های قبلی جمعیت اولیه می‌شوند. انتخاب کروموزوم‌های والدین از جمعیت می‌تواند به صورت کاملا تصادفی یا متناسب با مقدار برازندگی اعضای جمعیت باشد. یعنی کروموزمی که برازندگی (طرح نقاط دورافتاده که مقدار معیار آکائیک) کمتری داشته باشد با احتمال بیشتری انتخاب شود.

۵) تولید کروموزوم‌های جدید از ترکیب کروموزوم‌های والدین بوده و با روش‌های مختلفی امکان پذیر است. در اینجا با توجه به مسأله تعیین نقاط دورافتاده، دو روش جهش[۱۷] و تعویضی[۱۸] (متقاطع) را توضیح می‌دهیم. در روش جهش، یک یا چند نقطه از کروموزوم‌های والدین به صورت تصادفی تغییر می‌کند. ولی در روش تعویضی برخی قطعات به صورت طولی بین دو کروموزوم تبادل می‌شود. الگوریتم GA با استفاده از ترکیب تعویضی و جهشی کروموزوم‌ها می‌تواند تابع مورد نظر را مینیمم کند.

به عنوان مثال ساده در جدول 1، نحوه مینیمم سازی تابع $f(x) = x^2$ با استفاده از ترکیب تعویضی و جهشی کروموزوم‌ها نشان داده شده است. در این جدول برای کاهش مقدار $f(x)$ جمعیت اولیه، کروموزوم‌های ترکیبی جدید با

14 Fitness
15 Parents
16 Offspring
17 Mutation
18 Crossover

مقدار $f(x)$ کمتر، جایگزین کروموزوم‌های با مقدار $f(x)$ بزرگ می‌شوند.

جدول ۱: مثال GA برای مینیمم سازی تابع $f(x)=x^2$ برای اعداد صحیح $0 \le x \le 4095$

	شماره	کروموزوم	x	$f(x)$
جمعیت اولیه	1	101111101110	3054	9326916
	2	010100010111	1303	1697809
	3	110101100100	3428	11751184
	4	010100001100	1292	1669264
	5	011101011101	1885	3553225
	6	101101001001	2889	8346321
	7	101011011010	2778	7717284
	8	010011010101	1237	1530169
ترکیب تعویضی کروموزوم‌های شماره 4 و 8 از ژن چهارم از سمت چپ	——	0101**1**1010101	1493	2229049
	جایگزین کروموزوم شماره 3	0100**00001100**	1036	1073296
جهش ژن دوم کروموزوم شماره 2	جایگزین کروموزوم شماره 1	0**0**0100010111	279	77841
جهش ژن هفتم کروموزوم شماره 5	——	011101**1**11101	1917	3674889

در مسأله مربوط به این تحقیق علاوه بر آشکار ساختن مکان نقاط دورافتاده، نوع نقاط دورافتاده (یعنی IO ، AO، LS و TC) هم شناسایی می‌شود. پس باید کروموزم‌های مربوط به این مسأله برای نمونه‌ای به طول T به صورت $\boldsymbol{\xi}=(\xi_1,\xi_2,\dots,\xi_T)'$ باشد که هر یک از ξ_tها یکی از مقادیر ۰، ۱، ۲، ۳ و ۴ را به ترتیب متناسب با عدم دورافتادگی و یا دورافتاده از نوع IO ، AO ، LS و TC می‌گیرند.

بررسی همه طرح‌های مربوط به نقاط دورافتاده برای یافتن مکان و نوع نقطه دورافتاده بسیار زمان‌بر است. برای مثال وقتی فقط یک نقطه دورافتاده در مشاهدات وجود داشته باشد، باید تعداد $\binom{T}{1}\times 5$ طرح مختلف بررسی شود تا مکان و نوع آن نقطه دورافتاده مشخص گردد. وقتی تعداد مشاهدات و نقاط دورافتاده افزایش یابد، به همین نسبت تعداد حالاتی که باید بررسی شوند افزایش می‌یابد. وقتی بدانیم حداکثر g نقطه دورافتاده از انواع مختلف وجود دارد، باید $\Sigma_{h=.}^{g}\binom{T}{h}\times 5^h$ طرح را بررسی کنیم تا بهترین طرح مکان و نوع نقاط دورافتاده را به دست آوردیم. برای یافتن طرح نقاط دورافتاده (طرح مینیمم کننده معیار آکائیک) با $g=5$ و $T=300$، بررسی تقریبی 6×10^{13} حالت نیاز می‌باشد. ولی در GA برای یافتن طرح نقاط دورافتاده، مقدار معیار آکائیک همه این طرح‌ها محاسبه نشده و سپس طرح مینیمم کننده معیار آکائیک پیدا نمی‌شود، بلکه با ترکیب تصادفی (تعویضی و جهشی) طرح نقاط دورافتاده تکی، طرح کامل و مینیمم کننده معیار آکائیک به دست می‌آید. این الگوریتم تعداد محاسبه را کاهش داده و جواب را سریع‌تر به دست می‌آورد. این الگوریتم طرح نقاط دورافتاده را از طریق مینیمم کردن مقدار معیار آکائیک (مقدار برازندگی) به دست می‌آورد.

وقتی روشی، نقاط دورافتاده را یکی یکی پیدا می‌کند برآورد بد پارامترها، بر آشکارسازی نقاط دورافتاده بعدی اثر می‌کند. یعنی نقاط دورافتاده، برآورد پارامترهای مدل VARMA را به حدی بد می‌کنند که بعضی از داده‌ها به اشتباه به عنوان نقطهٔ دورافتاده در نظر گرفته شوند یا برعکس بعضی از نقاط دورافتاده به عنوان مشاهده‌های بدون آلودگی محسوب شوند. با توجه به [۶]، این دو اثر نقاط دورافتاده به صورت زیر تعریف می‌شوند:

- **اثر برون‌بری**[19]: اگر یک نقطه دورافتاده نقطه دیگری را هم دورافتاده کند آغشتن یا برون‌بری گویند. یعنی نقطهٔ دورافتادهٔ دوم فقط با حضور اولی، دورافتاده در نظر گرفته می‌شود.
- **اثر درون‌آوری**[20]: اگر نقطهٔ دورافتاده‌ای باعث پنهان شدن نقطهٔ دورافتادهٔ دیگر نزدیک به آن شود درون‌آوری گویند. یعنی نقطهٔ دورافتاده دوم به تنهایی نقطهٔ دورافتاده در نظر گرفته شود ولی در کنار اولی، دورافتاده در نظر گرفته نشود.

به عبارت دیگر، وقتی روشی نقاط دورافتاده را یکی یکی پیدا می‌کند طرح ناقص نقاط دورافتاده آشکار شده، بر پیدا کردن صحیح نقاط دورافتاده آشکار نشده، اثر گذاشته و در نتیجه آشکارسازی صحیح و کامل نقاط دورافتاده با اشتباه همراه می‌شود. در حالی‌که در روش GA وقتی می‌خواهیم طرح جدیدی از نقاط بالقوه دورافتاده (طرحی شامل چندین نقطهٔ دورافتاده) برای آشکارسازی را بررسی کنیم، ابتدا اثرات این طرح را حذف می‌کنیم. سپس برآورد دوباره پارامترهای مدل VARMA را به دست آورده و آشکارسازی این طرح نقاط دورافتاده ارزیابی می‌شود. یعنی با تعدیل داده‌ها برای هر طرح پیشنهادی جدید، برآورد پارامترها از طرح ناقص نقاط دورافتاده آشکار شده قبلی اثر کمتری گرفته و در نتیجه، بر پیدا کردن صحیح نقاط دورافتاده آشکار نشده اثر کمتری می‌گذارد. در صورتی که طرح پیشنهادی تولید شده توسط GA، طرح صحیح و کامل نقاط دورافتاده باشد، تأثیر همه نقاط دورافتاده از برآورد پارامترها به درستی حذف شده و این طرح کامل بدون اثرپذیری از نقاط آشکارشدهٔ قبلی به درستی آشکار می‌شود.

با توجه به مطالب ارائه شده، معمولاً در روش‌های شناسایی مرسوم یا نقاط دورافتاده به طور مجزا شناسایی می‌شوند که عموماً زمان‌بری کمتری داشته ولی مشکل درون‌آوری و برون‌بری پابرجاست و یا طرح نقاط دورافتاده یک جا آشکار می‌شوند که با توجه با زیاد بودن تعداد همه طرح‌های ممکن نقاط دورافتاده، بسیار زمان‌بر بوده ولی کمتر مشکل درون‌آوری و برون‌بری به وجود می‌آید. در روش GA هم طرح نقاط دورافتاده به صورت یک پارچه بررسی شده و مشکل درون‌آوری و برون‌بری کمتری وجود دارد. البته در روش GA همه طرح‌های ممکن نقاط دورافتاده بررسی نشده و ترکیب متقاطع و جهش طرح‌ها به کار می‌رود. با این کار زمان‌بری روش GA کاهش می‌یابد. بنابراین در GA طرح نقاط دورافتاده به صورت صحیح‌تر و با زمان‌بری کمتر آشکار می‌شوند.

آشکارسازی نقاط دورافتاده در سری‌های زمانی با روش TPP

روش دیگر آشکارسازی گونه‌های مختلف نقاط دورافتاده چند متغیره به وسیلهٔ تسای و همکاران [۹] تحت عنوان تسای، پنا و پانکراتز (TPP) پیشنهاد شده است. برای تشریح روش TPP، فرض کنید داده‌های $\mathbf{Z}_t = (Z_{t1}, \ldots, Z_{tm})', \ t = 1, \ldots, T$ از مدل VARMA(p,q) آلوده به نقاط دورافتاده تولید شده باشند. آشکارسازی به روش TPP در سه مرحله انجام می‌شود:

مرحله اول: با فرض عدم وجود نقطه دورافتاده، ابتدا بهترین مدل VARMA اولیه را به مشاهدات برازش داده و برآورد ضرایب $\widehat{\Theta}(B)$، $\widehat{\Phi}(B)$ و برآورد باقیمانده‌های $\boldsymbol{a}_t$ را به دست می‌آوریم. سپس برای هر مشاهده، پارامتر اثر هر نوع از دورافتادگی برآورد می‌شود. برای برآورد این پارامتر، با توجه به مدل آلوده شده (۴) به صورت

[19] Swamping
[20] Masking

$$\{\Theta(B)\}^{-1}\Phi(B)\boldsymbol{Z}_t = \{\Theta(B)\}^{-1}\Phi(B)\boldsymbol{Y}_t + \{\Theta(B)\}^{-1}\Phi(B)\alpha(B)\boldsymbol{\omega}\xi_t^h \quad (۸)$$

می‌توان نوشت:

$$\begin{aligned}\boldsymbol{a}_t &= \boldsymbol{\varepsilon}_t + \left\{\widehat{\Theta}(B)\right\}^{-1}\widehat{\Phi}(B)\hat{\alpha}(B)\boldsymbol{\omega}\xi_t^h \\ &= \boldsymbol{\varepsilon}_t + \widehat{\Pi}(B)\hat{\alpha}(B)\boldsymbol{\omega}\xi_t^h \end{aligned} \quad (۹)$$

که در آن $\widehat{\Pi}(B) = \sum_{i=0}^{\infty}\widehat{\Pi}_i B^i$ برآورد ضرایب نمایش اتورگرسیو مدل VARMA می‌باشد. لذا برای هر نقطه با توجه به نوع دورافتادگی، برآوردهای اولیه $\widehat{\Theta}(B)$، $\widehat{\Phi}(B)$ و $\boldsymbol{a}_t$، برآورد پارامتر اثر $\widehat{\boldsymbol{\omega}}$ به وسیلهٔ مدل رگرسیون (۹) برآورد می‌شود. با توجه $\hat{\alpha}(B)$، مدل رگرسیون (۹) برای انواع مختلف نقاط دورافتاده به صورت زیر است:

$$\begin{aligned}
&\boldsymbol{a}_t = \boldsymbol{\varepsilon}_t + \boldsymbol{\omega}\xi_t^h && for \quad IO \\
&\boldsymbol{a}_t = \left(I - \sum_{i=1}^{\infty}\widehat{\Pi}_i B^i\right)\xi_t^h\boldsymbol{\omega} + \boldsymbol{\varepsilon}_t = \left(\xi_t^h - \sum_{i=1}^{\infty}\widehat{\Pi}_i\xi_{t-i}^h\right)\boldsymbol{\omega} + \boldsymbol{\varepsilon}_t && for \quad AO \\
&\boldsymbol{a}_t = \left(I - \sum_{i=1}^{\infty}\widehat{\Pi}_i B^i\right)\xi_{t\geq h}\boldsymbol{\omega} + \boldsymbol{\varepsilon}_t && for \quad LS \\
&\boldsymbol{a}_t = \left(I - \sum_{i=1}^{\infty}\widehat{\Pi}_i B^i\right)\xi_{t\geq h}\delta^{t-h}\boldsymbol{\omega} + \boldsymbol{\varepsilon}_t && for \quad TC.
\end{aligned} \quad (۱۰)$$

حال فرض کنید $\boldsymbol{Y}_t$ها برای $t = 1,2,..,T$ داده‌های سری زمانی باشند. برای نقطهٔ دورافتادهٔ نوساز (IO) در زمان h، همه اطلاعات در بارهٔ نقطهٔ دورافتاده در $\boldsymbol{a}_t$ وجود دارد و اثر نقطهٔ دورافتاده IO به صورت

$$\widehat{\boldsymbol{\omega}}_{IO} = \boldsymbol{a}_h \quad (۱۱)$$

برآورد می‌شود. ماتریس کوواریانس این برآوردگر، $\widehat{\Sigma}$ مربوط به خطاها می‌باشد.

در حالت جمع پذیر (AO) هم با توجه به مدل رگرسیون (۱۰)، برآوردگر اثر نقطهٔ دورافتادهٔ AO به صورت زیر می‌باشد:

$$\widehat{\boldsymbol{\omega}}_{AO} = -\left(\sum_{i=0}^{T-h}\widehat{\Pi}_i'\Sigma_\varepsilon^{-1}\widehat{\Pi}_i\right)^{-1}\sum_{i=0}^{T-h}\widehat{\Pi}_i'\Sigma_\varepsilon^{-1}\boldsymbol{a}_{h+i}, \qquad \Pi_0 = -I \quad \text{که} \quad (۱۲)$$

ماتریس کوواریانس این برآوردگر به صورت $\Sigma_{AO,h} = \left(\sum_{i=0}^{T-h}\widehat{\Pi}_i'\Sigma_\varepsilon^{-1}\widehat{\Pi}_i\right)^{-1}$ می‌باشد.

به طور مشابه، برآورد اثر نقطه دورافتاده LS با توجه به رابطهٔ (۱۰) به صورت زیر به دست می‌آید [۴]:

$$\widehat{\boldsymbol{\omega}}_{LS} = -\left(\sum_{i=0}^{T-h}X_i'\Sigma_\varepsilon^{-1}X_i\right)^{-1}\sum_{i=0}^{T-h}X_i'\Sigma_\varepsilon^{-1}\boldsymbol{a}_{h+i}, \quad (۱۳)$$

که $X_i = \sum_{j=0}^{i}\widehat{\Pi}_j$.

همچنین این اثر برای نقاط دورافتاده TC به صورت زیر می‌شود:

$$\widehat{\boldsymbol{\omega}}_{TC} = -\left(\sum_{i=0}^{T-h}X_i'\Sigma_\varepsilon^{-1}X_i\right)^{-1}\sum_{i=0}^{T-h}X_i'\Sigma_\varepsilon^{-1}\boldsymbol{a}_{h+i}, \quad (۱۴)$$

که $X_i = \delta^i\sum_{j=0}^{i}\widehat{\Pi}_j$.

روش TPP برای آزمون دورافتادگی در زمان h، فرض صفر $H_0: \boldsymbol{\omega} = 0$ (صفر بودن اندازهٔ دورافتادگی) را در مقابل فرض مقابل $H_1: \boldsymbol{\omega} \neq 0$ بررسی می‌کند. برای انجام این آزمون از دو آماره استفاده می‌شود. یکی از این آماره‌ها عناصر $\boldsymbol{\omega}$ را به صورت چند متغیره در نظر می‌گیرد و به صورت

$$J_{i,h} = \hat{\boldsymbol{\omega}}'_{i,h}\hat{\Sigma}^{-1}_{i,h}\hat{\boldsymbol{\omega}}_{i,h} \quad (۱۵)$$

تعریف می‌شود که $i = IO, AO, LS, TC$ نوع دورافتادگی را مشخص می‌کند. این آماره تحت فرض صفر، با فرض معلوم بودن مدل و برای هر زمان ثابت h دارای توزیع مجانبی خی دو با m درجهٔ آزادی می‌باشد [۹]. آماره دوم، ماکسیمم قدرمطلق درایه‌های $\hat{\boldsymbol{\omega}}_h$ را استفاده کرده و برای $\Sigma_{i,h}$ معلوم به صورت

$$C_{i,h} = \max_{1\le j\le m} |\hat{\omega}_{j,i,h}|/\sqrt{\hat{\sigma}_{j,i,h}} \quad (۱۶)$$

تعریف می‌شود که $i = IO, AO, LS, TC$ نوع دورافتادگی، $\hat{\boldsymbol{\omega}}_{j,i,h}$ درایه jام اثر $\hat{\boldsymbol{\omega}}_{i,h}$ و $\sigma_{j,i,h}$ درایه (j,j)ام ماتریس کوواریانس $\Sigma_{i,h}$ می‌باشند. مقادیر بحرانی این آماره را می‌توان با شبیه‌سازی پیدا کرد. آمارهٔ آزمون کلی برای همه نقاط داده به صورت

$$J_{\max}(i) = \max_{1\le h\le T} J_{i,h}, \quad C_{\max}(i) = \max_{1\le h\le T} C_{i,h} \quad (۱۷)$$

تعریف می‌شود. تحت فرض صفر عدم وجود نقطهٔ دورافتاده در نمونه و معلوم بودن مدل $\boldsymbol{Y}_t$، برای نقطه دورافتاده IO با توجه به رابطهٔ (۱۱) داریم:

$$J_{\max}(IO) = \max_{1\le h\le T} a'_h\hat{\Sigma}^{-1}a_h$$

در عمل مدل فرایند داده‌ها معلوم نبوده و مقادیر بحرانی توزیع این دو آماره را می‌توان با شبیه‌سازی به دست آورد. اگر یکی از آماره‌های $J_{\max}(i)$ در زمان h معنی‌دار باشد نقطهٔ دورافتادهٔ چند متغیره نوع i که $i = IO, AO, LS, TC$ در زمان h آشکار می‌شود. اگر چند تا از آماره‌های $J_{\max}(i)$ مربوط به انواع نقاط دورافتاده معنی دار باشند نوع دورافتادگی براساس کوچکترین p-مقدار تجربی مربوط به $J_{\max}(i)$ها به دست می‌آید. وقتی هیچ یک از آماره‌های توام $J_{\max}(i)$ در سطح ۰/۰۵ درصد معنی‌دار نباشند آماره تکی $C_{\max}(i)$ برای جستجوی اضافه‌تر نقاط دورافتاده استفاده می‌شود. این مرحله برای اطمینان از عدم وجود نقاط دورافتاده در اجزاء انجام می‌شود. در بعضی موارد، برآورد پارامتر اثر نقطهٔ دورافتاده $(\hat{\boldsymbol{\omega}}_h)$ برای برخی از اجزاء، آشکارسازی نقاط دورافتاده را پیشنهاد می‌کند.

برای برآورد دقیق‌تر پارامترها، اثر نقطه دورافتاده آشکارشده را می‌توان از سری حذف کرد. در نتیجه سری تعدیل شده[21] با حذف کردن اثر نقاط دورافتاده از سری آلوده شده با استفاده از رابطهٔ (۴) به صورت زیر به دست می‌آید:

$$\mathbf{Y}_t = \mathbf{Z}_t - \boldsymbol{\alpha}(B)\boldsymbol{\omega}\xi_t^h. \quad (۱۸)$$

برای محاسبهٔ این سری، $\boldsymbol{\alpha}(B)$های مربوط به انواع نقاط دورافتاده IO ، AO ، LS و TC را در رابطه (۱۸) قرار می‌دهیم. این $\boldsymbol{\alpha}(B)$ها برای انواع مختلف نقاط دورافتاده در بخش ۳،۱ ارائه شده و در نتیجه سری تعدیل شده به صورت زیر می‌باشد.

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{Z}_t - [\boldsymbol{\Pi}(B)]^{-1}\boldsymbol{\omega}_{IO}\xi_t^h && for\ t > h \\ \boldsymbol{Y}_t &= \boldsymbol{Z}_t - \boldsymbol{\omega}_{AO} && for\ t = h \\ \boldsymbol{Y}_t &= \boldsymbol{Z}_t - \boldsymbol{\omega}_{LS}\xi_t && for\ t > h \end{aligned}$$

[21] Modified

$$\boldsymbol{Y}_t = \boldsymbol{Z}_t - \left[\frac{1}{1-\delta B}\right]^{-1} \boldsymbol{\omega}_{TC}\xi_t^h \qquad for\ t > h. \tag{۱۹}$$

در عمل $0 < \delta < 1$ و مدل VARMA وارون پذیر در نظر گرفته می‌شود. در نتیجه، اثر نقاط دورافتاده IO و TC بعد از چند نقطه‌ی زمانی کم شده و می‌تواند در نظر گرفته نشود. یعنی

$$[\boldsymbol{\Pi}(B)]^{-1} = \sum_{i=0}^{\infty} \Psi_i B^i \approx \sum_{i=0}^{N} \Psi_i B^i \quad و \quad \left[\frac{1}{1-\delta B}\right]^{-1} = \sum_{i=0}^{\infty} \delta^i B^i \approx \sum_{i=0}^{N} \delta^i B^i$$

پس روش TPP ، نقاط دورافتاده را یکی یکی آشکار ساخته و اثر آنها را با استفاده (۱۹) حذف می‌کند. این مرحله تا آشکارسازی هر نقطه دورافتاده‌ای توسط آماره‌های (۱۵) و (۱۶) ادامه می‌یابد.

مرحله دوم: برآورد همزمان پارامترهای مدل و اثرات نقاط دورافتاده آشکارسازی شده در مرحله اول را به دست آورده و سپس نقاط دورافتاده‌ای که آماره اثر آنها معنی‌دار نیست حذف شوند.

مرحله سوم: با توجه به [۴] و [۹] برآورد همزمان مرحله دوم تا معنی داری همه نقاط دورافتاده آشکار شده تکرار شود و با استفاده از این برآورد پارامترها، هیچ نقطه دورافتاده جدیدی شناسایی نشود.

حال با توجه به اینکه روش TPP نقاط دورافتاده را یکی یکی آشکار می‌کند، مشکل درون‌آوری و برون‌بری را دارد. یعنی در این روش نقاط دورافتاده شده به درستی آشکار نشده و برآوردهای مطلوبی برای پارامترهای مدل VARMA به دست نمی‌آید. لذا در این تحقیق برای برآورد مطلوب‌تر پارامترهای این مدل از روش GA نیز استفاده شده است.

شناسایی انواع نقاط دورافتاده در مدل VAR با استفاده از الگوریتم ژنتیک

کاسیانا و همکاران [۶] روش GA را در حالت چند متغیره برای نقاط دورافتاده AO مورد استفاده قرار داده و از معیار آکائیک به عنوان تابع هدف استفاده کردند.. حال با توجه به این‌که سری‌های زمانی چند متغیره ممکن است با گونه‌های متفاوتی از نقاط دورافتاده آلوده شده باشند، در این تحقیق، الگوریتم ژنتیک ارائه شده توسط کاسیانا و همکاران [۶] را برای گونه‌های مختلف نقاط دورافتاده (IO ، AO ، LS و TC) تعمیم می‌دهیم.

برآورد پارامترهای مدل VAR نسبت به مدل VARMA آسان‌تر و سریع‌تر به دست می‌آید. همچنین، تحت شرایط وارون‌پذیری [۷]، مدل VARMA را با مدل VAR(p)، برای p بزرگ، می‌توان تقریب زد. جهت به کارگیری این مدل برای برازش و آشکارسازی نقاط دورافتاده نخست به معرفی تابع درستنمایی مدل VAR(p) می‌پردازیم.

فرض کنید $\mathbf{Y}_t$، سری زمانی m-متغیره (بدون آلودگی) از مدل VAR(p) و به صورت زیر

$$\boldsymbol{Y}_t - \boldsymbol{\mu} = \Pi_1(\boldsymbol{Y}_{t-1} - \boldsymbol{\mu}) + \ldots + \Pi_p(\boldsymbol{Y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t \tag{۲۰}$$

باشد، که در آن بردار $\boldsymbol{\varepsilon}_t$، فرایند اغتشاش خالص از توزیع نرمال m-متغیره با میانگین چند متغیره صفر و ماتریس کواریانس Σ بوده و ضرایب اتورگرسیو Π_iها ماتریس‌های $m \times m$ می‌باشند. با توجه به [۷]، تابع لگاریتم درستنمایی پارامترها در نقطه‌ی $\boldsymbol{Y} = \text{vec}(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_T)$ از طریق تابع چگالی اغتشاش‌های خالص، $\boldsymbol{\varepsilon}_t$ها، به دست می‌آید. چون $\boldsymbol{\varepsilon}_t$ها، دارای توزیع نرمال بوده و برای tهای مختلف، مستقل از هم می‌باشند پس:

$$\boldsymbol{\varepsilon} = \text{vec}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T) = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_T \end{bmatrix} \sim N(0, I_T \otimes \Sigma)$$

به عبارت دیگر، $\boldsymbol{\varepsilon}$ دارای توزیع نرمال mT متغیره بوده و تابع چگالی آن به صورت زیر است:

$$f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi)^{mT/2}} |I_T \otimes \Sigma|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{\varepsilon}'(I_T \otimes \Sigma^{-1})\boldsymbol{\varepsilon}\right]$$
$$= \frac{1}{(2\pi)^{mT/2}} |I_T \otimes \Sigma|^{-1/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{\varepsilon}_t' \Sigma^{-1} \boldsymbol{\varepsilon}_t\right] \quad (۲۱)$$

همچنین

$$\boldsymbol{\varepsilon} = \begin{pmatrix} I_m & 0 & \cdots & 0 & \cdots & \cdots & 0 \\ -\Pi_1 & I_m & & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & & \vdots \\ -\Pi_p & -\Pi_{p-1} & \cdots & I_m & & & 0 \\ 0 & -\Pi_p & & & \ddots & & \vdots \\ \vdots & & \ddots & & & \ddots & \vdots \\ 0 & 0 & \cdots & -\Pi_p & \cdots & \cdots & I_m \end{pmatrix} (\boldsymbol{Y} - \boldsymbol{\mu}^*) + \begin{pmatrix} -\Pi_1 & -\Pi_2 & \cdots & -\Pi_p \\ -\Pi_2 & -\Pi_3 & \cdots & 0 \\ \vdots & & & \vdots \\ -\Pi_p & 0 & \cdots & 0 \\ & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} (\boldsymbol{Y}_0 - \boldsymbol{\mu}_0),$$

که $\boldsymbol{\mu}^* = (\boldsymbol{\mu}', \ldots, \boldsymbol{\mu}')'$ بردار $Tm \times 1$ بعدی، $\boldsymbol{\mu}_0 = (\boldsymbol{\mu}', \ldots, \boldsymbol{\mu}')'$ بردار $mp \times 1$ بعدی شامل میانگین‌ها و $\boldsymbol{Y}_0 = (\mathbf{Y}_{-1}', \mathbf{Y}_{-2}, \cdots, \mathbf{Y}_{-p+1}')'$ می‌باشند. بنابراین، $\partial\boldsymbol{\varepsilon}/\partial\boldsymbol{Y}$ ماتریس پایین مثلثی با عناصر قطری ۱ و دترمینال ۱ می‌باشد. در نتیجه تابع چگالی $\boldsymbol{Y}$ برابر است با

$$f(\boldsymbol{Y}) = \left|\frac{\partial\boldsymbol{\varepsilon}}{\partial Y}\right| f_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) = \frac{1}{(2\pi)^{mT/2}} |I_T \otimes \Sigma|^{-1/2} \times$$
$$\exp\left[-\frac{1}{2}\sum_{t=1}^{T} \left((\mathbf{Y}_t - \boldsymbol{\mu}) - \sum_{i=1}^{p} \Pi_i(\mathbf{Y}_{t-i} - \boldsymbol{\mu})\right)' \Sigma^{-1} \left((\mathbf{Y}_t - \boldsymbol{\mu}) - \sum_{i=1}^{p} \Pi_i(\mathbf{Y}_{t-i} - \boldsymbol{\mu})\right)\right]$$

یعنی تابع لگاریتم درستنمایی در نقطه $\boldsymbol{Y} = \mathrm{vec}(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_T)$ به صورت زیر می‌باشد.

$$\log\ell = -\frac{mT}{2}\ln 2\pi - \frac{T}{2}\ln|\Sigma| -$$
$$\frac{1}{2}\sum_{t=1}^{T} \left((\mathbf{Y}_t - \boldsymbol{\mu}) - \sum_{i=1}^{p} \Pi_i(\mathbf{Y}_{t-i} - \boldsymbol{\mu})\right)' \Sigma^{-1} \left((\mathbf{Y}_t - \boldsymbol{\mu}) - \sum_{i=1}^{p} \Pi_i(\mathbf{Y}_{t-i} - \boldsymbol{\mu})\right) \quad (۲۲)$$

این تابع را می‌توان برای سرعت بخشیدن به انجام محاسبات به صورت زیر نوشت

$$\log\ell = -\frac{mT}{2}\ln 2\pi - \frac{T}{2}\ln|\Sigma| - \frac{1}{2} tr\left[(Y^* - AX)' \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1} (Y^* - AX)\right] \quad (۲۳)$$

که

$$Y^* = (\mathbf{Y}_1 - \boldsymbol{\mu}, \cdots, \mathbf{Y}_T - \boldsymbol{\mu})_{m \times T}$$
$$A = (\Pi_1, \Pi_2, \cdots, \Pi_p)_{m \times mp}$$
$$X = (\mathbf{Y}_0^\circ, \cdots, \mathbf{Y}_{T-1}^\circ)_{mp \times T}$$
$$\mathbf{Y}_t^\circ = \begin{pmatrix} \mathbf{Y}_t - \boldsymbol{\mu} \\ \vdots \\ \mathbf{Y}_{t-p+1} - \boldsymbol{\mu} \end{pmatrix}_{mp \times 1}$$

همچنین $\boldsymbol{\mu}$ میانگین فرایند و Π_i ها ماتریس ضرایب مدل VAR می‌باشند.

به طور کلی تابع درستنمایی (۲۳) با فرض عدم وجود مشاهدات دورافتاده است. حال اگر مشاهدات $\mathbf{Z}_t$ آلوده شده به

نقاط دورافتاده باشند، باید اثر این نقاط از آنها حذف شده و سپس در تابع درستنمایی مورد استفاده قرار گیرد. برای این کار ابتدا GA طرحی از نقاط دورافتاده را تولید می‌کند. سپس با داشتن برآوردهای اولیه از مدل VAR(p) و معلوم بودن مکان نقاط دورافتاده، برآورد اثر نقاط دورافتاده، $\widehat{\boldsymbol{\omega}}$ توسط مدل رگرسیونی (10) و رابطه‌های (۱۱) تا (۱۴) محاسبه می‌شود. برای حذف اثر نقاط دورافتاده از مشاهدات، برآورد اثر نقاط دورافتاده و برآورد پارامترهای مدل و طرح نقاط دورافتاده تولید شده توسط GA را در روابط (۱۸) و (۱۹) جایگزین کرده و سری تعدیل شده را به دست می‌آوریم. سری تعدیل شده تقریبا عاری از اثر نقاط دورافتاده بوده و در محاسبه $\log\ell$ در رابطه (۲۳) به کار می‌رود.

لازم به ذکر است که تابع درستنمایی استفاده شده در (۲۳) برای مدل VAR بوده و برای آشکارسازی نقاط دورافتاده در مدل VARMA (برای تقریب مناسب این مدل) از مدل VAR(p)، با p بزرگ می‌توان استفاده نمود. برای تعیین مرتبه مدل VAR، معیار آکائیک ارائه شده در [۷] را می‌توان به کار برد. البته این معیار برای مشاهدات عاری از نقاط دورافتاده بوده و برای مشاهدات دارای نقاط دورافتاده، باید تحقیقات بیشتری صورت بگیرد.

تابع هدف و مقدار برازندگی

با توجه به رابطهٔ (۴) با افزایش تعداد نقاط دورافتاده، تعداد پارامترهای برآورد شده که شامل برآورد ω ها (اثرات این نقاط) هم می‌باشد افزایش می‌یابد. همچنین برای هر نقطه دورافتاده، Y_t های جدید با استفاده از رابطهٔ (۱۹) تعدیل می‌شوند. پس با افزایش تعداد نقاط دورافتاده، Y_t های بیشتری تعدیل شده و لذا مدل برازش بهتری داشته و در نتیجه تابع درستنمایی (۲۲) هم افزایش می‌یابد. بنابراین برای امساک در تعداد پارامترها و متناظر با معیار آکائیک، تابع $\log\ell$ در رابطهٔ (۲۳) را ماکسیمم کرده و تعداد نقاط دورافتاده را مینیمم می‌کنیم. این تابع همان تابع هدف GA می‌باشد و به صورت زیر است:

$$f(\boldsymbol{\xi}) = -2\log\ell + cs$$
$$\propto tr(Y^* - AX)'\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(Y^* - AX) + cs \qquad (۲۴)$$

که "تعداد نقاط دورافتاده$\times s = m$" و مقدار c از جدول ۲ منبع [6] به دست می‌آید. برای مثال، با توجه به این منبع، مقدار c برای اندازه نمونهٔ $T = ۲۰۰$، $m = ۲$ و خطای نوع اول ۰/۰۵ برابر با ۸/۳ می‌باشد. البته مقدار c را می‌توان (با بررسی مقادیر مختلف در مطالعات شبیه سازی) طوری انتخاب نمود که طرح صحیح نقاط دورافتاده را با درصد بالاتری به دست آورد. مقادیر کوچک (یا بزرگ) c آشکارسازی بیشتر (یا کمتر) نسبت به تعداد واقعی نقاط دورافتاده را باعث می‌شود. تابع $f(\boldsymbol{\xi})$ زمانی مینیمم می‌شود که $\log\ell$ افزایش و تعداد نقاط دورافتاده کاهش یابد.

مقدار برازندگی، مقدار تابع هدف مربوط به هر کروموزوم است و کروموزومی که مقدار $f(\boldsymbol{\xi})$ در (۲۴) را مینیمم کند طرح نقاط دورافتاده می‌باشد. برای به دست آوردن سریع‌تر این مینیمم، (در مرحله انتخاب والدین الگوریتم) می‌توان کروموزم‌هایی با مقدار کمتر برازندگی را با احتمال بیشتری انتخاب نمود و مینیمم را با اصلاح آنها به دست آورد. در این تحقیق، مقدار برازندگی همانند [۶] از رابطهٔ

(۲۵)

$$fitness = \exp\left(\frac{f(\xi)}{b}\right)$$

به دست آمده که b یک اسکالر پیشنهادی بوده و برای خیلی کوچک نشدن توان و صفر نشدن مقدار (۲۵) در محاسبات استفاده شده است. همچنین، احتمال انتخاب کروموزم‌ها متناسب با معکوس مقدار (۲۵) به دست آمده است. با این کار، کروموزم‌هایی که معیار آکائیک کمتری دارند با احتمال بیشتری انتخاب شده و جستجوی طرح نقاط دورافتاده، بیشتر روی این کروموزم‌ها صورت می‌گیرد.

مطالعات شبیه سازی شده نشان می‌دهد که GA بعد از چند تکرار، اکثر نقاط دورافتاده را پیدا می‌کند. پس در ادامه تکرارهای GA، برآورد دوباره پارامترهای مدل VAR روی سری تعدیل شده به دست می‌آید. با این کار، تأثیرپذیری نقاط دورافتاده آشکار شده روی برآورد پارامترها کمتر شده و طرح کامل نقاط دورافتاده به صورت صحیح‌تری آشکار می‌شود. تکرارهای GA تا اتمام مینیمم کردن مقدار (۲۵) ادامه می‌یابد.

مراحل انجام الگوریتم

مراحل انجام الگوریتم GA برای شناسایی انواع نقاط دور افتاده به صورت زیر است:

۱. برازش مدل VAR(p) اولیه روی داده‌ها و محاسبه ماتریس‌های برآورد $\widehat{\Pi}(B)$ و باقیمانده‌های $\widehat{\boldsymbol{a}}_t$.

۲. محاسبهٔ برازندگی کروموزم‌های (۷) برای انواع مختلف نقاط دورافتاده، مرتب کردن این کروموزم‌ها بر اساس برازندگی و انتخاب کروموزم‌هایی با کمترین برازندگی برای تعیین جمعیت اولیه.

۳. در صورتی که کروموزم $\boldsymbol{\xi}' = (0,0,0,0,\dots,0,0)_{T\times 1}$، طرح عدم نقطه دورافتاده در داده‌ها، برازندگی کمتری از کروموزم‌های جمعیت اولیه داشته باشند داده‌ها نقطه دورافتاده نداشته و الگوریتم پایان می‌پذیرد.

۴. شناسایی نوع دورافتادگی کروموزم‌های جمعیت اولیه برای محدود کردن جهش کروموزم‌های انتخابی.

۵. ساخت کروموزومی که همه نقاط دورافتاده جمعیت اولیه داشته و اضافه کردن این کروموزم به جمعیت اولیه به عنوان طرح اولیه همه نقاط دورافتاده.

۶. انتخاب دو کروموزم والدین با احتمالی متناسب با معکوس برازندگی کروموزم‌های جمعیت و ترکیب تعویضی و جهشی این دو کروموزم برای ساخت کروموزم‌های جدید.

۷. ساخت سری موقت تعدیل شده با حذف اثر کروموزم‌های جدید از داده‌ها، برآورد موقت پارامترهای مدل VAR از طریق سری تعدیل شده و محاسبه برازندگی کروموزم‌های جدید.

۸. در صورتی که برازندگی هر کدام از کروموزم‌های جدید کمتر از مینیمم برازندگی کروموزم‌های جمعیت باشد کروموزم جدید جایگزین کروموزم اضافی طرح نقاط دورافتاده می‌شود. با این روش تا آخر الگوریتم، کروموزم‌های اولیه که نقاط دورافتاده تکی اولیه را نشان می‌دهند در جمعیت باقی می‌مانند.

۹. در هر ۱۰۰ تکرار سری تعدیل شده از کروموزم طرح نقاط دورافتاده به دست آمده و برآورد پارامترهای مدل VAR اصلاح می‌شود. تا زمانی که تعداد تکرارها کمتر از M (مثلاً ۱۰۰۰) باشد، مراحل ۶ تا ۸ تکرار می‌شود.

مطالعات شبیه سازی

عملکرد روش GA را می‌توان با شبیه سازی مورد بررسی قرار داد. در این تحقیق، چهار مدل پیشنهاد شده در کاسیانا و همکاران [۶] در نظر گرفته شده است. توجه کنید که مدل‌های به کار رفته در کاسیانا و همکاران [۶] صرفاً برای شناسایی نقاط دورافتاده AO مورد استفاده قرار گرفته ولی در اینجا برای شناسایی انواع مختلف نقاط دورافتاده استفاده می‌شود. این چهار مدل شرایط مانایی و وارون پذیری را دارا هستند. بعنی مقادیر ویژه Φ_1 و Θ_1، قدر مطلق کمتر از یک دارند.

- مدل ۱: دو متغیره VAR(1) با $\Phi_1 = \begin{pmatrix} 0/6 & 0/2 \\ 0/2 & 0/4 \end{pmatrix}$.
- مدل ۲: دو متغیره VARMA(1,1) با $\Phi_1 = \begin{pmatrix} 0/6 & 0/2 \\ 0/2 & 0/4 \end{pmatrix}$ و $\Theta_1 = \begin{pmatrix} -0/7 & 0/2 \\ -0/1 & 0/4 \end{pmatrix}$.
- مدل ۳: سه متغیره VAR(1) با $\Phi_1 = \begin{pmatrix} 0/6 & 0/2 & 0 \\ 0/2 & 0/4 & 0 \\ 0/6 & 0/2 & 0/5 \end{pmatrix}$.
- مدل ۴: سه متغیره VARMA(1,1) با $\Phi_1 = \begin{pmatrix} 0/6 & 0/2 & 0 \\ 0/2 & 0/4 & 0 \\ 0/6 & 0/2 & 0/5 \end{pmatrix}$ و $\Theta_1 = \begin{pmatrix} -0/7 & 0 & 0 \\ -0/1 & -0/3 & 0 \\ -0/7 & 0 & -0/5 \end{pmatrix}$.

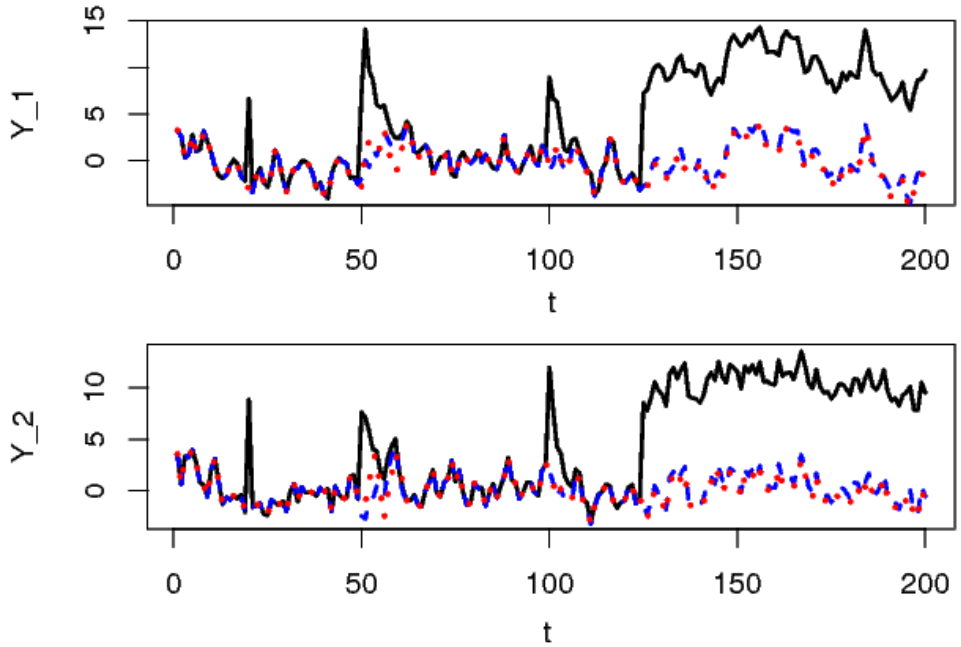
برای هر چهار مدل، مقدار $\boldsymbol{\mu} = \mathbf{0}$ و ماتریس کوواریانس نوفه سفید همانی در نظر گرفته شد.

برای تولید داده‌های با نقطه دورافتاده، ابتدا داده‌ها از مدل‌های VARMA یا VAR شبیه‌سازی شده و سپس با استفاده از تساوی (۴) آلوده شده‌اند. چهار ساختار نقاط دورافتاده در اندازه نمونه $T = ۲۰۰$ و با مکان نقطه دورافتاده AO در زمان ۲۵، IO در زمان ۱۵۰، LS در زمان ۱۰۰ و TC در زمان ۵۰ در نظر گرفته شده است. این ساختارهای نقاط دورافتاده در زیر آورده شده است.

- ساختار نقاط دورافتاده یکتایی: در این ساختار، چهار طرح نقاط دورافتاده وجود دارد که در هر کدام فقط یک نوع نقطه دورافتاده استفاده شده است.
- ساختار نقاط دورافتاده دوتایی: در این ساختار شش ترکیب مختلف طرح نقاط دورافتاده دوتایی آورده شده است (مثل AO,IO و AO,LS و ...).
- ساختار نقاط دورافتاده سه‌تایی: در این ساختار چهار ترکیب مختلف طرح نقاط دورافتاده سه‌تایی آورده شده است (مثل AO,IO,LS و AO,IO,TC و ...).
- ساختار نقاط دورافتاده چهارتایی: در این ساختار یک طرح نقاط دورافتاده که شامل همه نوع‌های نقاط دورافتاده AO,IO,LS,TC می‌باشد آورده شده است.

در شکل ۱، نمونه شبیه‌سازی شده از مدل ۱ با انواع نقاط دورافتاده و اندازهٔ اثر $\boldsymbol{\omega} = (۱۰,۱۰)'$ نشان داده شده است. براساس این شکل آشکارسازی نموداری نقاط دورافتاده غیر ممکن است. ولی با توجه به شکل ۱، سری تعدیل شده

توسط GA با سری اصلی آلوده نشده به نقاط دورافتاده تطابق تقریبی دارد. یعنی GA با آشکارسازی صحیح نقاط دورافتاده، تأثیر آنها را به خوبی از مشاهدات حذف کرده و سری تعدیل شده‌ای نزدیک به سری اصلی آلوده نشده ارائه کرده است.



شکل۱. نمونه دو متغیره از مدل VARMA(1,1) با نقاط دورافتاده AO ، IO ، TC ، LS به ترتیب در زمان‌های 20، 50، 100 و 125 با اندازه اثر $\boldsymbol{\omega} = (10, 10)'$. سری آلوده به نقاط دورافتاده —، سری اصلی بدون نقاط دورافتاده - - و سری تعدیل شده به روش GA ....

در ادامه این تحقیق، اندازه اثر نقاط دورافتاده برای مدل دو-متغیره $\boldsymbol{\omega} = (5,5)'$ و برای سه-متغیره $\boldsymbol{\omega} = (5,5,5)'$ در نظر گرفته می‌شود. سپس عملکرد GA برای ۱۰۰ سری زمانی شبیه سازی شده بررسی می‌شود. البته مطالعات شبیه سازی با دیگر مقادیر پارامترها و ساختارهای نقاط دورافتاده، نتایج مشابه این بخش به‌دست می‌آید که در این‌جا آورده نشده است.

برای نقطه دورافتاده تغییر موقت، مقدار $\delta = ۰/۷$ در نظر گرفته شده است. اندازه جمعیت اولیه در GA برابر $N_{popsize} = ۳۰$ و برای تقریب مناسب مدل VARMA از مدل VAR(10) استفاده شده است.

ارزیابی هر روش آشکارسازی نقاط دورافتاده با بررسی درصد آشکارسازی‌های صحیح آن روش امکان پذیر است. جدول‌های ۲ تا ۶ نتایج مقایسه درصد آشکارسازی‌های صحیح دو روش GA و TPP را نشان می‌دهند. در این جدول‌ها P_{true} به عنوان درصد آشکارسازی دقیق تمام نقاط دورافتاده، $P_{<true}$ درصد آشکارسازی صحیح تعدادی از نقاط دورافتاده و E درصد عدم آشکارسازی کلیه نقاط دورافتاده را به صورت زیر محاسبه می‌شوند

$$P_{true} = \frac{\text{تعداد آشکارسازی کاملا صحیح و دقیق نقاط دورافتاده}}{\text{تعداد کل آشکارسازی‌ها}} \times 100$$

$$P_{<true} = \frac{\text{تعداد آشکارسازی‌های کمتر از کاملا صحیح که حداقل یک نقطهٔ دورافتاده صحیح باشد}}{\text{تعداد کل آشکارسازی‌ها}} \times 100$$

$$E = \frac{\text{تعداد آشکارسازی‌هایی که در آن هیچ کدام از نقاط دورافتاده واقعی آشکار نشده باشند}}{\text{تعداد کل آشکارسازی‌ها}} \times 100$$

جدول ۲: نتایج مقایسه دو روش GA و TPP برای آشکارسازی یک‌تایی انواع نقاط دورافتاده با شبیه‌سازی: P_{true} درصد آشکارسازی دقیق تمام نقاط دورافتاده، $P_{<true}$ درصد آشکارسازی صحیح تعدادی از نقاط دورافتاده و E درصد عدم آشکارسازی کلیه نقاط دورافتاده را نشان می‌دهد.

مقایسه‌های **یک‌تایی**		AO		IO		LS		TC	
		GA	TPP	GA	TPP	GA	TPP	GA	TPP
مدل ۱	P_{true}	99	100	19	21	100	90	89	87
	$P_{<true}$	0	0	0	0	0	0	0	0
	E	1	0	81	79	0	10	11	13
مدل ۲	P_{true}	99	98	87	93	100	71	94	92
	$P_{<true}$	0	0	0	0	0	0	0	0
	E	1	2	13	7	0	29	6	8
مدل ۳	P_{true}	98	100	48	64	100	85	95	97
	$P_{<true}$	0	0	0	0	0	0	0	0
	E	2	0	52	36	0	15	5	3
مدل ۴	P_{true}	100	99	91	100	99	70	98	96
	$P_{<true}$	0	0	0	0	0	0	0	0
	E	0	1	9	0	1	30	2	4

جدول ۳: نتایج مقایسه دو روش GA و TPP برای آشکارسازی دوتایی انواع نقاط دورافتاده با شبیه سازی: P_{true} درصد آشکارسازی دقیق تمام نقاط دورافتاده، $P_{<true}$ درصد آشکارسازی صحیح تعدادی از نقاط دورافتاده و E درصد عدم آشکارسازی کلیه نقاط دورافتاده را نشان می‌دهد.

مقایسه‌های دوتایی		AO,IO		AO,LS		AO,TC		IO,LS		IO,TC		LS,TC	
		GA	TPP	GA	TPP	GA	TPP	GA	TPP	GA	TPP	GA	TPP
مدل ۱	P_{true}	92	94	99	84	87	87	88	68	16	19	85	76
	$P_{<true}$	8	6	1	16	13	13	12	27	71	66	15	24
	E	0	0	0	0	0	0	0	5	13	15	0	0
مدل ۲	P_{true}	97	99	100	90	91	92	96	90	85	90	94	69
	$P_{<true}$	3	1	0	10	9	8	4	10	13	10	5	30
	E	0	0	0	0	0	0	0	0	2	0	1	1
مدل ۳	P_{true}	96	99	100	95	96	93	87	91	34	61	99	77
	$P_{<true}$	4	1	0	5	4	7	13	8	66	37	1	23
	E	0	0	0	0	0	0	0	1	0	2	0	0
مدل ۴	P_{true}	98	100	100	84	95	91	88	92	77	92	90	58
	$P_{<true}$	2	0	0	12	5	9	12	8	23	8	10	40

	E	0	0	0	4	0	0	0	0	0	0	0	2

جدول ۴: نتایج مقایسه دو روش GA و TPP برای آشکارسازی سه‌تایی انواع نقاط دورافتاده با شبیه سازی: P_{true} درصد آشکارسازی دقیق تمام نقاط دورافتاده، $P_{<true}$ درصد آشکارسازی صحیح تعدادی از نقاط دورافتاده و E درصد عدم آشکارسازی کلیه نقاط دورافتاده را نشان می‌دهد.

مقایسه‌های **سه‌تایی**		AO,IO,LS		AO,IO,TC		AO,LS,TC		IO,LS,TC	
		GA	TPP	GA	TPP	GA	TPP	GA	TPP
مدل ۱	P_{true}	94	85	13	17	79	71	15	9
	$P_{<true}$	6	15	87	83	21	29	85	90
	E	0	0	0	0	0	0	0	1
مدل ۲	P_{true}	96	85	79	89	82	75	81	59
	$P_{<true}$	4	15	21	11	18	25	19	40
	E	0	0	0	0	0	0	0	1
مدل ۳	P_{true}	88	78	36	61	95	79	29	43
	$P_{<true}$	12	22	64	39	5	21	71	57
	E	0	0	0	0	0	0	0	0
مدل ۴	P_{true}	75	84	74	92	89	58	68	63
	$P_{<true}$	25	16	26	8	11	41	32	37
	E	0	0	0	0	0	1	0	0

جدول ۵: نتایج مقایسه دو روش GA و TPP با شبیه سازی (تأثیر افزایش گونه‌های نقاط دورافتاده بر شناسایی): P_{true} درصد آشکارسازی دقیق تمام نقاط دورافتاده، $P_{<true}$ درصد آشکارسازی صحیح تعدادی از نقاط دورافتاده و E درصد عدم آشکارسازی کلیه نقاط دورافتاده را نشان می‌دهد.

		میانگین یک‌تایی‌ها		میانگین دوتایی‌ها		میانگین سه‌تایی‌ها		چهارتایی‌ها (AO,IO,LS,TC)	
		GA	TPP	GA	TPP	GA	TPP	GA	TPP
مدل ۱	P_{true}	77	75	78	71	50	46	18	18
	$P_{<true}$	0	0	20	26	50	54	82	82
	E	23	25	2	3	0	0	0	0
مدل ۲	P_{true}	95	89	94	88	85	77	73	65
	$P_{<true}$	0	0	5	12	15	23	27	35
	E	5	11	1	0	0	0	0	0
مدل ۳	P_{true}	85	87	85	86	62	65	43	51
	$P_{<true}$	0	0	15	13	38	35	57	49
	E	15	13	0	1	0	0	0	0
مدل ۴	P_{true}	97	91	91	86	77	74	64	56
	$P_{<true}$	0	0	9	13	23	26	36	43
	E	3	9	0	1	0	0	0	1
میانگین درصدهای مدل‌های ۱ تا ۴	P_{true}	89	85	87	83	69	66	50	48
	$P_{<true}$	0	0	12	16	31	34	50	52
	E	11	15	1	1	0	0	0	0

جدول ۶: نتایج مقایسه دو روش GA و TPP با شبیه سازی (میانگین درصدهای آشکارسازی برای همه مدل‌های ۱ تا ۴ و همه طرح‌های یکتایی، دوتایی، سه‌تایی و چهارتایی).

		GA	TPP
میانگین درصدها برای همه مدل‌ها و همه طرح ها	P_{true}	74	71
	$P_{<true}$	23	25
	E	3	4

با توجه به جدول ۲ از مقایسه P_{true} مربوط به طرح یک‌تایی نقاط دورافتاده می‌توان نتیجه گرفت که روش GA مکان نقطه دورافتاده LS را با دقت بالاتری نسبت به انواع دیگر نقاط دورافتاده آشکار ساخته درحالیکه روش TPP مکان نقاط AO را بهتر آشکار می‌سازد. همچنین روش GA نقاط دورافتاده از نوع LS ، AO و TC را بهتر از روش TPP شناسایی کرده و روش TPP نوع IO را بهتر شناسایی می‌کند. ضعیف‌ترین آشکارسازی هر دو روش مربوط به مدل ۱ بوده و از نوع IO می‌باشد. همچنین جدول ۳ نشان می‌دهد که روش GA ، آشکارسازی دقیق طرح دوتایی LS،AO را با درصد بیشتری نسبت به طرح‌های دوتایی‌های دیگر انجام می‌دهد. ولی روش TPP دوتایی IO،AO را بهتر آشکار می‌سازد. همچنین روش TPP در میان ۶ طرح دوتایی، فقط ۲ طرح IO،AO و IO,TC را با P_{true} بیشتر نسبت به روش GA آشکار می‌شود. در همه مقایسه‌های دوتایی، غیر از طرح TC،IO برای مدل ۱، هر دو روش GA و TPP نقاط دورافتاده را تقریبا به خوبی آشکار می‌سازند. با توجه به جدول ۴ برای طرح‌های سه‌تایی، هر دو روش GA و TPP نقاط دورافتاده AO,IO,LS را با دقت بالاتری نسبت سایر دورافتاده‌ها آشکار می‌سازند.

در حالت کلی با توجه به مقدار پارامترهای مدل و اندازهٔ δ، اثر هر دو نقطه IO و TC بعد از زمان پرش سری کاهش می‌یابد. به همین دلیل اثر این دو نوع نقطه دورافتاده نسبتاً شبیه به هم می‌باشد. البته با مقایسه‌های دستی مشاهده شد که در مدل ۱ بردارهای اثر $\{\Phi(B)\}^{-1}\Theta(B)\boldsymbol{\omega}\xi_t^h$ و $\boldsymbol{\omega}\delta^{t-h}$ مربوط به نقاط IO و TC مقادیر نزدیکی به هم داشته و شناسایی این دو نوع نقطه دورافتاده با مشکل مواجه می‌شود. با توجه به جدول ۲ و ۳ برای مدل ۱، هر جا این دو اثر کنار هم باشند آشکارسازی با دقت کمتری صورت می‌پذیرد.

از مقایسه مقدار E مربوط به طرح‌های یک‌تایی جدول ۲، هر دو روش GA و TPP نقطه دورافتاده نوع IO را با بالاترین درصد عدم آشکارسازی صحیح (به صورت اشتباه) آشکار نمی‌کنند. البته روش TPP نوع LS را نیز با درصد اشتباه نسبتا بالا آشکار نمی‌سازد. برای طرح‌های دوتایی، سه تایی و چهارتایی جدول‌های ۳ تا ۵ مقدار E کوچک بوده و قابل چشم پوشی می‌باشد.

با توجه به سه ردیف آخر جدول ۵ (میانگین درصدهای مدل‌های ۱ تا ۴) ، میانگین P_{true} مربوط به دو روش GA و TPP برای طرح‌هایی یک‌تایی به ترتیب ۸۹ و ۸۵ درصد، طرح‌هایی دوتایی به ترتیب ۸۷ و ۸۳ درصد، طرح‌هایی سه‌تایی به ترتیب ۶۹ و ۶۶ درصد و طرح چهارتایی به ترتیب ۵۰ و ۴۸ درصد است. همچنین با توجه به جدول ۶، میانگین کلی P_{true} روی همه طرح‌ها و روی همه مدل‌ها برای روش GA و روش TPP به ترتیب ۷۴ و ۷۱ بوده و میانگین کلی E برای روش GA و روش TPP به ترتیب ۳ و ۴ می‌باشد. پس، روش GA درصد آشکارسازی دقیق‌تری نسبت به TPP داشته و در هر دو روش با افزایش تعداد گونه‌های نقاط دورافتاده دقت آشکارسازی کمتر می‌شود. همچنین با توجه به جدول‌های ۲ تا ۵، در ۵۶ درصد مقایسه‌هایی یک‌تایی (۹ تا از ۱۶ تا)، در ۵۵ درصد مقایسه‌هایی دوتایی (۱۳ تا از ۲۴

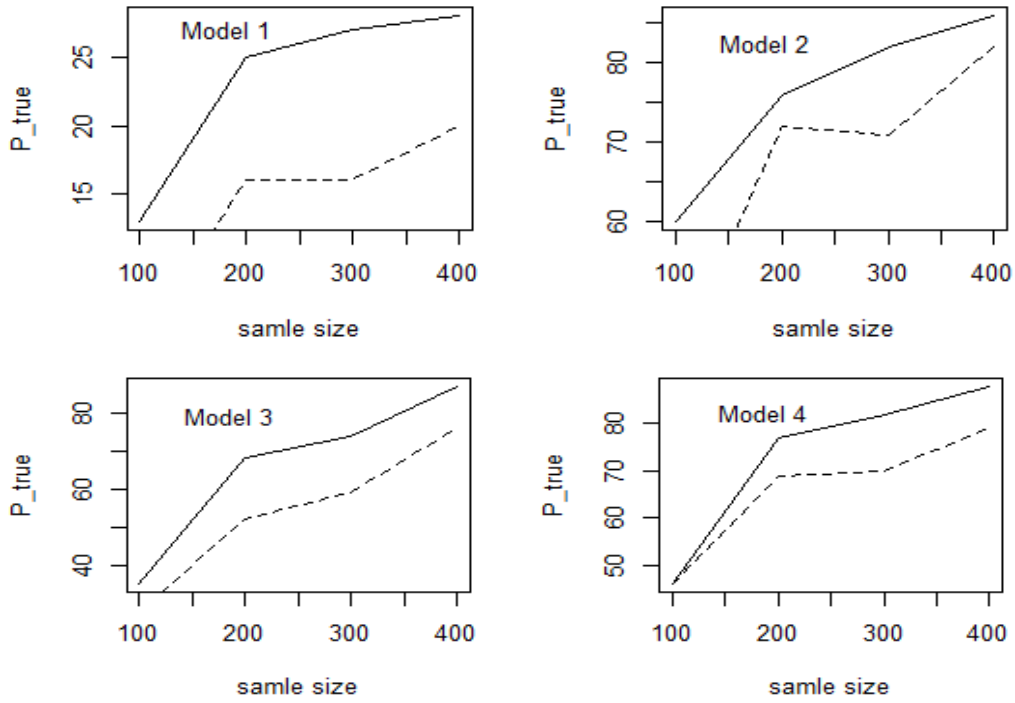
تا)، در ۶۳ درصد مقایسه‌هایی سه‌تایی (۱۰ تا از ۱۶ تا) و ۷۵ درصد مقایسه‌هایی چهارتایی(۳ تا از ۴ تا)، روش GA درصد P_{true} بالاتری نسبت به روش TPP دارد. پس به لحاظ تعداد آشکارسازی صحیح بالاتر هم، دقت روش GA از روش TPP بیشتر بوده و با افزایش تعداد گونه‌های نقاط دورافتاده، کاهش دقت روش TPP نسبت به روش GA بیشتر می‌باشد.

به صورت کلی زیاد بودن تعداد نقاط دورافتاده، آشکارسازی صحیح آنها را سخت‌تر می‌کند. به عبارت دیگر زیاد بودن تعداد نقاط دورافتاده باعث اریبی برآورد پارامترهای مدل VAR شده و چون این برآوردهای اریب خود در آشکارسازی نقاط دورافتاده استفاده می‌شود، آشکارسازی هم با اشتباه بیشتر صورت می‌گیرد. از طرف دیگر افزایش گونه‌های مختلف نقاط دورافتاده نیز باعث افزایش اثرات آنها و اریبی برآورد پارامترهای مدل VAR شده و در نتیجه آشکارسازی در هر دو روش GA و TPP با اشتباه بیشتر همراه می‌شود. درصد پایین آشکارسازی صحیح مربوط به بیشترین آلودگی تعداد و نوع نقاط دورافتاده در ستون‌های ۵ و ۶ جدول ۵ نشان داده شده است. اغلب آشکارسازی‌های اشتباه این دو ستون، شناسایی‌های اشتباه گونه‌های مختلف نقاط دورافتاده می‌باشد. برای مثال روش TPP نقاط دورافتاده TC یا IO را به جای LS شناسایی کرده است.

با توجه به سه ردیف آخر جدول ۵، الگوریتم GA خصوصاً برای حالتی که تعدد گونه‌های مختلف نقاط دورافتاده وجود دارد بهتر از روش TPP عمل می‌کند. علت این موضوع آن است که در TPP، برای آشکارسازی یک نقطه دورافتاده جدید از برآوردهایی استفاده می‌شود که از سری تعدیل شده از نقاط آشکار شده قبلی به دست آمده‌اند. یعنی طرح ناقص نقاط دورافتاده آشکار شده با تأثیر در برآورد پارامترها و آماره‌های (۱۵) و (۱۶)، در آشکارسازی طرح کامل و صحیح نقاط دورافتاده اثر می‌گذارند. ولی در GA، برای آشکارسازی هر طرح جدید نقاط دورافتاده، ابتدا تعدیل سری این طرح جدید صورت می‌گیرد و برآورد دوباره مدل VAR روی آن به دست می‌آید. سپس مقدار برازندگی این طرح جدید به دست آمده و آشکارسازی آن بررسی می‌شود. این کار باعث کاهش تأثیر نقاط دورافتاده آشکار شده قبلی در برآورد پارامترها، مقدار برازندگی و آشکارسازی طرح جدید می‌شود. یعنی طرح ناقص نقاط دورافتاده آشکارشده، تأثیر کمتری در آشکارسازی طرح کامل و صحیح نقاط دورافتاده دارند. در واقع، اگر GA طرح نقاط دورافتاده را به درستی تولید کند، تقریبا مدل VAR بدون اثر این نقاط برآورد شده و آشکارسازی به درستی انجام می‌شود.

با توجه به جدول ۲ و ۳، در چندین مورد از مقایسه‌های GA و TPP، خصوصاً هنگامی که تعداد گونه‌های دورافتاده‌ها کم است، روش TPP نقاط دورافتاده را به صورت صحیح‌تری آشکار می‌سازد. در اغلب این حالات، با تعداد نقاط دورافتاده کمتر، برآورد مدل VAR دقیق‌تر بوده و مشکل برون‌بری و درون‌آوری روش TPP کمتر می‌شود. بنابراین درصد آشکارسازی صحیح روش TPP هم زیاد می‌شود. همچنین GA بر اساس جستجوی تصادفی طرح نقاط دورافتاده کار کرده و ممکن است طرح درست را در تعداد محدود تکرار به صورت تصادفی آشکار نسازد.

همچنین در شکل ۲، تأثیر اندازهٔ حجم نمونه در میزان آشکارسازی روش‌های GA و TPP برای چهار مدل نشان داده شده است. با توجه به این شکل برای هر چهار مدل، با افزایش حجم نمونه، درصد آشکارسازی صحیح P_{true} بالا رفته و البته بهتر بودن روش GA نسبت به TPP به خوبی نشان داده شده‌است. لازم به ذکر است که این مقایسه برای ساختار نقاط دورافتاده چهارتایی انجام شده است.



شکل ۲: تأثیر اندازه حجم نمونه در میزان آشکارسازی روش GA با ـــــ و روش TPP با - - - نشان داده شده‌است.

لازم به ذکر است که در دو روش GA و TPP برای داده‌های شبیه سازی شده مدل‌های VAR(1) و VARMA(1,1) از برازش مدل VAR(10) استفاده شده و آشکارسازی نقاط دورافتاده به خوبی صورت گرفته است.

چون روش GA در هر تکرار، برآورد پارمترهای مدل VAR را دو بار (برای دو کرموزوم تولیدی) به دست می‌آورد، به زمان بیشتری نسبت به روش TPP نیاز دارد. بنابراین برای آشکارسازی سریع‌تر نقاط دورافتاده در GA، مکان و نوع جهش کروموزم‌ها فقط به مکان و نوع دورافتادگی که در جمعیت اولیه آن نقطه وجود داشت محدود شد (برای محدود شدن دامنهٔ جستجو). با این حال برای نمونه ۲۰۰ تایی از مدل ۲، زمان انجام GA حدود ۳۶ ثانیه بوده و در روش TPP حدود ۴ ثانیه می‌شود.

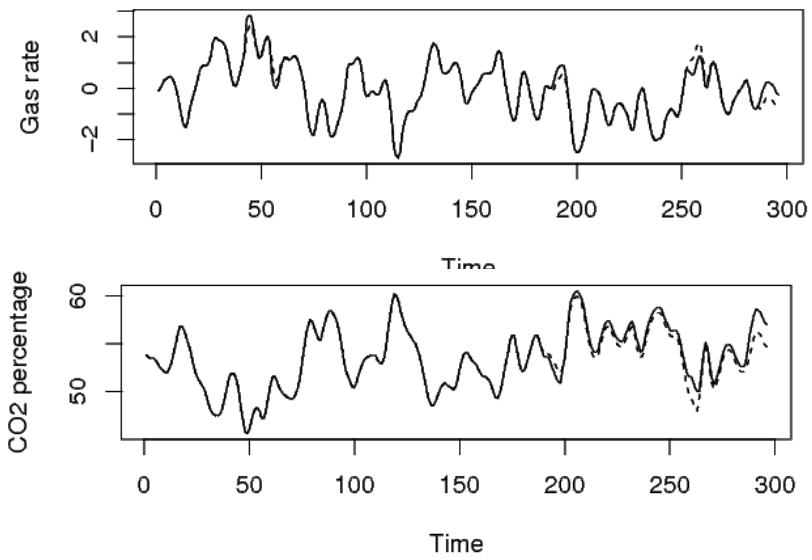
نتایج روش‌های GA و TPP برای داده‌های گاز-کوره

در این بخش، سری زمانی مربوط به گاز-کوره[۲۲] را بررسی می‌کنیم. این داده‌ها در بسته نرم‌افزاری "arfima" از نرم افزار R موجود بوده و شامل ۲۹۶ زوج از مشاهدات است که هر ۹ ثانیه اندازه گیری شده‌اند. متغیرها شامل میزان گاز ورودی و درصد دی اکسید کربن در بین گازهای خروجی است [۲]. این داده‌ها به صورت دو-متغیره در نظر گرفته شده و در شکل ۳ همراه با سری تعدیل شده بوسیله GA نشان داده شده است.

[22] Gas-Furnace

نقاط دورافتاده این داده‌ها در [۹] و [۶] با برازش مدل VAR(6) بررسی شده است. روش TPP با استفاده از آماره J_{max} نقاط دورافتاده را در زمان‌های ۴۳, ۵۵, ۱۱۳, ۱۹۹, ۲۳۶, ۲۶۵, ۲۸۷ و ۲۸۸ آشکار کرده است [۹]. برای روش GA هم، نقاط دورافتاده آشکارسازی شده و اندازه اثر آنها، $\hat{\omega}$، در جدول ۷ نشان داده شده است. این جدول نشان می‌دهد که روش GA نتایج مشابهی در زمان و نوع نقاط دورافتاده به دست آورده است. تنها تفاوت روش‌های GA و TPP در زمان‌های ۱۸۸ و ۱۹۹ می‌باشد. نقطه دورافتاده زمان ۱۸۸ فقط بوسیله GA و نقطه دورافتاده زمان ۱۹۹ فقط با روش TPP آشکارسازی شده است. همچنین مشاهده در زمان ۲۶۵ بوسیله TPP به عنوان IO شناسایی شده است ولی مشاهده در زمان ۲۶۴ با استفاده از روش GA به عنوان AO. روش GA این نقاط دورافتاده را با ۱۰۰۰ تکرار و $c = ۶$ به دست آورده است.

با حذف اثر نقاط دورافتاده آشکار شده از داده‌های گاز-کوره، داده‌های تعدیل شده به دست می‌آید. برازش مدل VAR(6) به این داده‌ها نشان می‌دهد که واریانس خطای مربوط به گاز ورودی در داده‌های تعدیل شده روش GA نسبت به TPP، ۱۷ درصد و واریانس خطای مربوط به دی اکسید کربن در داده‌های تعدیل شده روش GA نسبت به TPP، ۴۳ درصد کمتر می‌باشد.



شکل ۳: سری زمانی گاز-کوره با خط توپر و سری تعدیل شده با خط نقطه‌چین نشان داده شده است.

جدول ۷: آشکارسازی نقاط دورافتاده چند متغیره داده‌های گاز-کوره با استفاده از GA.

	1	2	3	4	5	6	7
زمان	43	55	113	188	235	264	288
نوع نقطه دورافتاده	TC	TC	TC	IO	LS	AO	LS
$\hat{\omega}_1$	0/707	−0/642	−0/397	0/114	−0/068	0/122	0/191
$\hat{\omega}_2$	−0/077	0/145	0/030	−0/134	0/770	−0/512	0/590

نتیجه‌گیری

وجود نقاط دورافتاده در داده‌های سری زمانی، ناقض فروض مانایی بوده و معمولا منجر به شناسایی غلط مدل، اریبی برآورد پارامترها و پیش‌بینی‌های ضعیف می‌شود. لذا در هنگام مواجهه با این نقاط ناخواسته می‌توان مکان و نوع نقاط دورافتاده را آشکار ساخته و سپس برآورد پارامترها و اندازه اثر این نقاط با استفاده از سری تعدیل شده از این نقاط به دست آورد. البته گاهی اوقات آشکارسازی نقاط دور افتاده از آن لحاظ اهمیت پیدا می‌کند که چنین نقاطی، وجود یک رویداد خارجی را در زمان داده مورد نظر نشان می‌دهد [۹].

در این تحقیق، پس از معرفی انواع نقاط دورافتاده در حالت چند متغیره، نخست روش شناسایی گونه‌های مختلف نقاط دورافتاده برای سری‌های زمانی چند متغیره بوسیله الگوریتم ژنتیک (GA) پیشنهاد شده است. این الگوریتم طرح نقاط دورافتاده را به صورت کدهای رشته‌ای به اندازه طول مشاهدات در نظر می‌گیرد. کد مربوط به هر نقطه زمانی، دورافتاده بودن یا نبودن آن نقطه را مشخص می‌کند. این روش طرح‌های تصادفی از مکان‌ها و گونه‌های نقاط دورافتاده را برای مینیمم کردن معیار آکائیک کنترل می‌کند. روش تسای، پنا و پانکراتز (TPP) روش دیگر آشکارسازی نقاط دورافتاده در این تحقیق می‌باشد. در هر تکرار این روش یک نقطه دورافتاده آشکار شده و تعدیل اثر این نقطه انجام می‌شود. سپس برآورد پارامترها از سری تعدیل شده به دست آمده و آشکارسازی نقطه بعدی با استفاده از این برآوردها ادامه می‌یابد. این کار ممکن است منجر به اریبی برآوردها و آشکارسازی اشتباه نقطه بعدی شود. به عبارت دیگر، در روش TPP یک نقطه دورافتاده آشکارشده باعث پنهان شدن نقطه دورافتاده دیگر می‌شود (درون‌آوری) یا یک نقطه دورافتاده آشکارشده باعث آشکارسازی مشاهده معمولی به عنوان نقطه دورافتاده می‌شود (برون‌بری). این روش، اغلب، گونه دورافتادگی را اشتباه آشکار می‌سازد. اما در هر تکرار GA، ابتدا طرحی تصادفی از همه نقاط دورافتاده (برای بررسی) تولید شده و سری تعدیل شده از همین طرح به دست می‌آید. سپس برآورد پارامترها و آشکارسازی همین طرح بررسی می‌شود. این کار تأثیر نقاط دورافتاده آشکار شده قبلی را بر آشکارسازی طرح کامل نقاط دورافتاده کاهش می‌دهد. در حقیقت، اگر طرح تصادفی همه نقاط دورافتاده به درستی تولید شود، تقریبا در سری تعدیل شده اثر همهٔ آنها حذف می‌شود. بنابراین GA با این سری تعدیل شده برآوردهای دقیق‌تری به دست آورده و نقاط دورافتاده را با صحت بیشتری آشکار می‌سازد. نتایج شبیه سازی صحت روش GA را تایید کرده و درصد آشکارسازی درست نقاط دورافتاده در این روش بیشتر از رویکرد TPP می‌باشد. البته GA زمان بیشتری برای محاسبات نیاز دارد.

References

1. Baragona R and Battaglia F (2007). Outliers Detection in Multivariate Time Series by Independent Component Analysis. Neural Computation, 19: 1962–1984.

2. Box GEP, Jenkins GM and Reinsel GC (1994). Time Series Analysis: Forecasting and Control. 3rd edition. Prentice Hall, Englewood Cliffs, New Jersey.

3. Chambers L. (2001). The Practical Handbook of Genetic Algorithms: Applications. 2nd edition, Chapman and Hall, Boca Raton, Fla.

4. Chen C. and Liu L. (1993). Joint Estimation of Model Parameters and Outlier Effects in Time Series. Journal of the American Statistical Association, 88(421):284–297.

5. Coley D.A (1999). An Introduction to Genetic Algorithms for Scientists and Engineers. World Scientific Publishing Co. Pte. Ltd., River Edge.

6. Cucina D., Salvatore A. and Protopapas MK (2014). Outliers Detection in Multivariate Time Series using Genetic Algorithms. Chemometrics and Intelligent Laboratory Systems, 132: 103-110.

7. Lutkepohl H. (2006). New Introduction to Multiple Time Series Analysis. Berlin and Heidelberg, Springer.

8. Melanie M. (1998). An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA.

9. Tsay R., Pena D. and Pankratz A. (2000). Outliers in Multivariate Time Series. Biometrika, 87: 789-804.