بسم الله الرحمن الرحيم

دانشکده علوم ریاضی

گروه ریاضی کاربردی

عنوان:

تعیین نقطه تغییر روی داده‌های سری زمانی

دانشجو:

امه کلثوم همتی راد

اساتید راهنما:

دکتر داود شاهسونی

دکتر احمد نزاکتی رضازاده

پایان‌نامه ارشد جهت اخذ درجه کارشناسی ارشد

بهمن ماه ۱۳۹۱

دانشگاه صنعتی شاهرود

تقدیم به

خورشید پنهانم

امام زمانم

که جهان از نور وجودش روشن است و شقایق‌ها به بهانه‌ی دیدنش جمعه‌ها را انتظار می‌کشد

پیشکش به:

مادر مهربانم

که شنا در اقیانوس وسیع زندگی و پرواز در آسمان آبی محبت را به من آموخت

و پدر عزیزم

که کلامش همچون مشعلی روشن و فروزان، در تاریکی راه‌های پرپیچ و خم زندگی‌ام است.

همچنین تقدیم به

دو خواهر مهربانم عاطفه و فاطمه همتی راد و برادر عزیزم عقیل همتی راد که وجودشان گرمابخش زندگی‌ام

است.

سپاس

سپاس می گویم خدای مهربانم را که زیباترین راه زندگی را به من آموخت و با در نظر گرفتن دستان ناتوانم قدم های مرا در این راه محکم کرد، مشکلات و سختی ها را برایم آسان نمود و درهای بسته ی دانش را با نور وجودش به رویم گشود. خدای مهربانم را شاکرم که تا این مرحله از مسیر علم و دانش را با موفقیت به اتمام رساندم و از او توفیق ادامه ی مسیر و حرکت در راستای خشنودیش را مسئلت دارم.

بر خود لازم می دانم از تمام کسانی که من را در نوشتن این پایان نامه یاری کردند، قدردانی و تشکر نمایم. تشکر می کنم از اساتید راهنمای محترمم جناب آقای دکتر داود شاهسونی و آقای دکتر احمد نزاکتی رضازاده که با خلوص نیت و در کمال صبر و آرامش پاسخگوی سوالات من بودند، و پیشنهادات ارزنده ای که در این امر داشتند و سپاسگزارم از جناب آقای دکتر حسین باغیشنی و سرکار خانم دکتر الهه ظهوریان که داوری این پایان نامه را بر عهده داشتند.

از خانواده محترمم به جهت تحمل شرایط سختی که در این مقطع تحصیلی داشتم، تشکر می کنم و برای تمامی دوستانم آرزوی موفقیت در تمامی مراحل زندگی را دارم.

تعهد نامه

اینجانب **امه کلثوم همتی راد** دانشجوی دوره کارشناسی ارشد رشته **آمار ریاضی** دانشکده **ریاضی** دانشگاه صنعتی شاهرود نویسنده پایان نامه **تعیین نقطه تغییر روی داده‌های سری زمانی** تحت راهنمائی **دکتر داود شاهسونی و دکتر احمد نزاکتی رضازاده** متعهد می شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ ۱۳۹۱/۱۱/۲۹

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

در یک سری زمانی که اساسا به صورت دنباله‌ای از مشاهدات مرتب‌شده بر اساس زمان به صورت $x_1,\ldots,x_n$ می‌باشد، ممکن است نقاطی وجود داشته باشندکه در آن نقاط، خصوصیات آماری متغیر تصادفی دست‌خوش تغییر قرار گرفته باشند. برای مثال، ممکن است n_0 مشاهده اول، از توزیعی مانند F_0 و مشاهدات باقی‌مانده از توزیع دیگری مانند F_1 تبعیت کنند. نقطه n_0 را نقطه تغییر (Change Point) نامند و یافتن این نقطه که در آن، ویژگی سری زمانی تغییر یافته است طیف گسترده‌ای از مسایل جهان واقعی را در بر می‌گیرد. موضوع یافتن نقطه تغییر یکی از چالش برانگیزترین مسایل آماری است، زیرا تعداد و محل این نقاط ناشناخته هستند. بدین منظور روش‌های متعددی با قابلیت‌های مختلف پیشنهاد شده‌اند. در این پایان‌نامه، چهار روش مجموع تجمعی (CUSUM) مبتنی بر نمونه‌های خودگردان، تحلیل طیفی منفرد (SSA)، تشخیص بیزی برخط (BOCPD) و مدل بیزی افراز ضربی (PPM) مورد مطالعه قرار گرفته اند. به منظور بررسی دقت و توانایی هر یک از این روش‌ها در تشخیص نقطه تغییر، از سه سری زمانی شبیه‌سازی شده و یک سری زمانی واقعی استفاده شده به طوری‌که سری‌های زمانی شبیه‌سازی‌شده به ترتیب مبین تغییر در سطح میانگین، تغییر در سطح واریانس و تغییر در خودهمبستگی هستند. نتایج داده‌های شبیه‌سازی‌شده نشان داد که از بین روش‌های ارایه‌شده، روش تشخیص برخط نقطه تغییر بیزی، نسبت به سه روش دیگر، عملکرد مناسب‌تری داشته و توانسته است هر گونه از تغییرات ذکر شده را به‌طور مناسبی برآورد نماید. با انتخاب این روش به عنوان روش برتر، داده‌های سری زمانی واقعی مورد بررسی قرار گرفته و عملکرد سایر روش‌ها با آن مقایسه شده اند. به دلیل هزینه زمانی صرف شده برای تحلیل داده‌ها و تشخیص نقاط تغییر از اهمیت بسزایی برخوردار می‌باشد، زمان اجرای روش‌ها نیز مورد مقایسه قرار گرفته اند. نتایج اخذ شده حاکی از مقرون به صرفه بودن روش تشخیص برخط نقطه تغییر بیزی در مجموعه داده‌های بزرگ است.

کلمات کلیدی: سری زمانی، نقطه تغییر، مجموع تجمعی (CUSUM)، تحلیل طیفی منفرد (SSA)، توابع تشخیص، ماتریس ناهمگنی، تشخیص برخط نقطه تغییر بیزی (BOCPD)، طول گشت، مدل افراز ضربی (PPM)

فهرست

عنوان **صفحه**

فهرست شکل‌ها

فصل اول

فصل سوم

فصل چهارم

فصل پنجم

فهرست جدول‌ها

فصل سوم

فصل چهارم

فصل پنجم

فصل اول

مقدمه و مفاهیم نقطه تغییر

۱-۱ مقدمه

آمار صنعتی نشان می‌دهد که حدود ۷۰ درصد از حوادث صنعتی توسط خطاهای انسانی ایجاد می‌شوند. اگر چه وقوع حوادث صنعتی معمول نیست ولی حوادث جزیی که به‌طور مکرر و روزانه رخ می‌دهند، موجب بسیاری از آسیب‌ها شده و باعث میلیاردها دلار هزینه در هر سال می‌شوند. بنابراین باید به‌منظور کاهش حوادث و بهبود ایمنی فعالیت‌های بسیاری انجام پذیرند. یکی از این فعالیت‌ها، روش‌های *بازنگری فرآیند*[۱] است که به بررسی فرآیند می‌پردازد (ونکاتاسوبرامانیان[۲] و همکاران، ۲۰۰۳).

بازنگری فرآیند، عملی زمان‌پیوسته برای تشخیص ناهنجاری‌ها در رفتار سیستم‌های ایستا[۳]، پویا[۴] و اتفاقی است و به معنی تشخیص اختلالات ناخواسته‌ی فرآیند بر حسب زمان است. در واقع مشکل اساسی، تشخیص اختلالاتی است که به‌طور مستقیم قابل مشاهده نبوده و با انواع دیگر اختلالات اندازه‌گیری می‌شوند، به عنوان مثال بسیاری از مسایل بازنگری تحت عنوان مساله تشخیص تغییر در پارامترهای سیستم بیان شده‌اند. از جمله روش‌های سنتی که برای بازنگری فرآیند و بهبود کیفیت محصول استفاده می‌شوند می‌توان به نمودارهای کنترل فرآیند آماری[۵] (SPC) مانند شوهارت[۶]، مجموع تجمعی (CUSUM)[۷] و نمودارهای میانگین متحرک موزون نمایی (EWMA)[۸] اشاره کرد (لی[۹] و همکاران، ۲۰۰۴). در دو دهه گذشته جستجوی روش‌های موثرتر برای بهبود

---

[1] Process monitoring

[2] Venkatasubramanian

[۳] ایستایی یعنی این‌که پارامترهای سیستم (یا متغیرهای حالت) با گذر زمان تغییر نمی‌کنند. نمونه یک سیستم ساده را می‌توان یک ساعت فرض کرد. برای نمونه، سرعت چرخش عقربه‌های ساعت در طول یک دور گردش بر اثر جاذبه زمین، تغییر بسیار ناچیزی دارند و قابل صرف نظر هستند.

[۴] سیستم‌های پویا، سیستم‌هایی هستند که دارای رفتار مختلفی می‌باشند، به‌طوری‌که رفتار آن‌ها بر اثر جزیی‌ترین تغییر در هر یک از پارامترهایش دچار نوسان‌های شدید می‌شود.

[5] Statistical process control

[6] Shewhart

[۷] Cumulative sum

[۸] Exponentially weighted moving average

[8] Lie

بازنگری فرآیند افزایش چشم‌گیری داشته است و با پیشرفت علوم کامپیوتری، روش‌های بازنگری فرآیند نیز بهبود یافته‌اند که از آن جمله می‌توان به تحلیل طیفی منفرد[۱] (SSA)، جنگل‌های تصادفی (آئورت و آلدریچ[۲]، ۲۰۱۰)، شبکه‌های عصبی مصنوعی[۳] (عبدالله و همکاران، ۲۰۰۵) و ..... اشاره کرد.

لازم به ذکر است داده‌هایی که برای بازنگری فرآیند استفاده می‌شوند، غالبا به شکل *سری زمانی* هستند. یک سری زمانی در واقع دنباله‌ای از اعداد است که در فواصل زمانی معین در خروجی سیستم مشاهده شده و به عنوان یک منبع مفید برای استخراج ویژگی و مطالعه‌ی رفتار سیستم استفاده می‌شود (حاجی زاده و همکاران، ۲۰۱۰). بیشتر طرح‌های بازنگری به زمان دقیق رخ دادن یک رویداد غیر طبیعی اهمیت نمی‌دهند، در حالی که زمان دقیق رخ دادن اختلال صورت‌گرفته در تشخیص علت اختلال کمک می‌کند (کاواهارا[۴] و همکاران ۲۰۰۹، باسسیویلی و نیکیفورم،[۵] ۱۹۹۳).

مساله کشف نقاط زمانی که در آن‌ها تغییری در ویژگی داده‌های سری زمانی رخ داده است، به عنوان "تشخیص نقطه تغییر[۶]" یا تشخیص رویداد، شناخته شده و طیف گسترده‌ای از مسایل جهان واقعی مانند تشخیص تقلب در سیستم‌های سلولی (کاواهارا و سوگی یاما، ۲۰۰۹[۷])، تشخیص نفوذ در شبکه‌های کامپیوتری (چن[۸] و همکاران، ۲۰۰۷)، تشخیص گسستگی در سیستم‌های مهندسی (سای[۹] و همکاران، ۲۰۰۵)، مطالعات زیست محیطی (فرایسین[۱۰] و همکاران، ۲۰۰۳)، رباتیک، کنترل فرآیند، امور مالی،تحلیل EEG[۱۱]، تقسیم‌بندی DNA، اقتصاد، جمعیت‌شناسی، داروشناسی، روان‌شناسی، زمین‌شناسی، هواشناسی و غیره را در بر می‌گیرد. به‌طور مثال، این مساله در زمین‌شناسی برای

---

[1] Singular spectrum analysis
[2] Auret and Aldrich
[3] Artificial neural networks
[۴] Kawahara
[5] Basseville and Nikiforov
[6] Change point detection
[۷] Kawahara and Sugiyama
[۸] Chen
[۹] Xie
[۱۰] Frisen
[11] Electrocardiogram

تشخیص زمان زلزله، در هواشناسی برای تشخیص زمان حوادث غیرطبیعی مانند سیل و طوفان به کار برده می‌شود. به این ترتیب، مساله‌ی تشخیص نقاط تغییر یکی از چالش‌برانگیزترین مسایل آماری است، زیرا تعداد و محل آن‌ها ناشناخته هستند (چیون[۱] و همکاران، ۲۰۰۹).

با توجه به توضیحات ذکرشده، تشخیص نقاط تغییر در بازنگری فرآیند و تشخیص دقیق اختلال و تشخیص علت آن، بسیار حایز اهمیت است. در نتیجه ضرورت معرفی و شناخت روش‌هایی برای تشخیص نقاط تغییر، احساس می‌شود. در راستای این پایان‌نامه، قصد داریم تعدادی از این روش‌ها را در زمینه‌های مختلف کنترلی، زیرفضایی و بیزی معرفی کنیم، مقایسه‌ای بین آن‌ها انجام دهیم و روشی را که دارای عملکرد مناسب‌تری برای تشخیص نقطه تغییر است، انتخاب کنیم.

۱-۲ تعاریف و مفاهیم مورد نیاز

طراحی الگوریتم تشخیص نقطه تغییر با داده‌های سری زمانی و اندازه‌گیری داده‌های حاصل از مقدارهای واقعی سرو کار دارد. در این بخش، تعاریف و مفاهیمی را که در فصل‌های آتی مورد استفاده قرار می‌گیرند، بیان می‌کنیم.

۱-۲-۱ مساله نقطه تغییر

مساله‌ی نقطه تغییر را می‌توان این‌گونه بیان کرد:

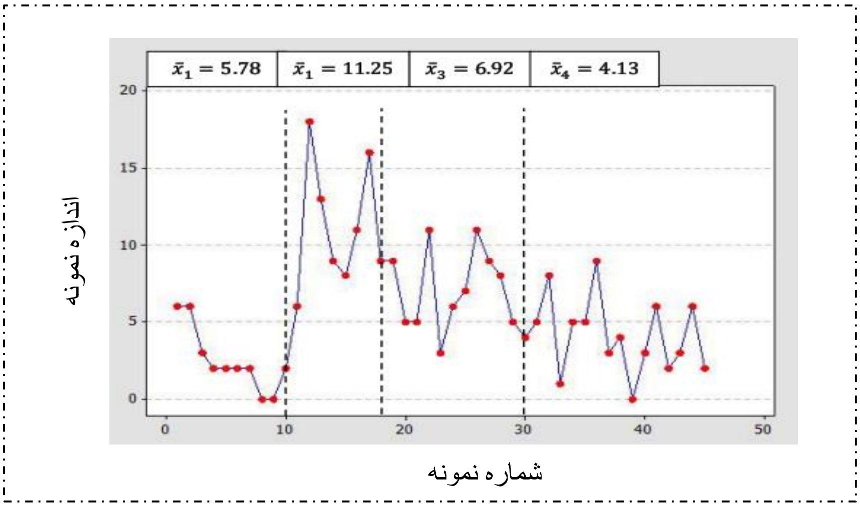
فرض کنید مجموعه‌ای از داده‌ها دارید که آن‌ها را با هدف تهیه مدل‌بندی ریاضی یا تعیین پارامترهای آن، تحلیل می‌کنید. از لحاظ آماری، فقط آن دسته از پارامترهایی که در طول فرآیند تغییر نکرده‌اند، قابل تشخیص هستند. بنابراین پیش از تشخیص پارامترها و معرفی مدل باید پذیره همگن بودن تصادفی[۲] داده‌ها را آزمون کرد. در صورت رد این پذیره، باید بخش‌هایی را که دارای داده-

---

[۱] Cheon

[۲] Stochastic homogeneity

های همگن تصادفی هستند، برای تشخیص پارامترهای فرآیند آن بخش شناسایی کرد. روش مناسب در این‌جا روش غیردنباله‌ای (واخط[۱]) نامیده می‌شود که در شکل ۱-۱ مثالی در خصوص بخش‌بندی داده‌های همگن در روش غیردنباله‌ای نشان داده شده است (کمالی، ۱۳۸۹).

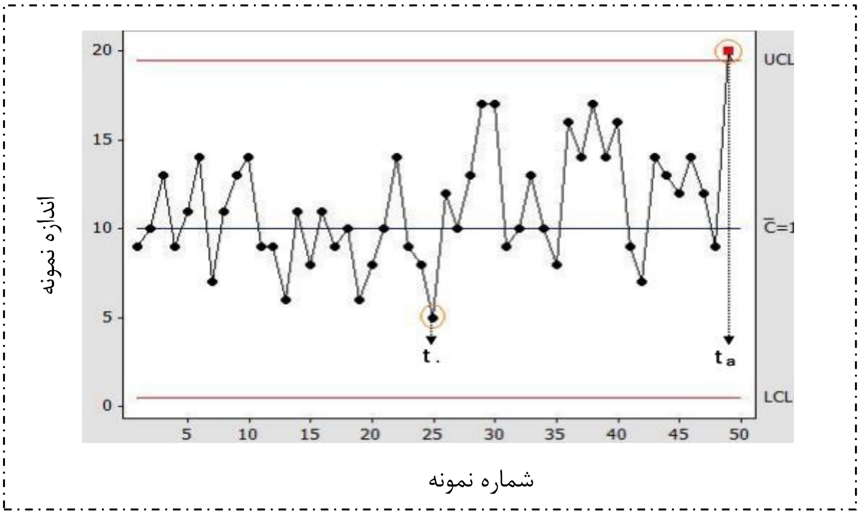


شکل ۱-۱: نمودار مشخص‌سازی بخش‌های همگن در روش غیر دنباله‌ای

اکنون فرض کنید، مشاهدات به ترتیب و یکی یکی جمع‌آوری می‌شوند و قطع همگنی تصادفی داده‌ها می‌تواند به عنوان اخطاری برای وقوع تغییرات اساسی (نقص یا خرابی دستگاه) در فرآیند مورد بررسی باشد. روش آماری مناسب در این‌جا روش دنباله‌ای[۲] (برخط[۳]) نامیده می‌شود. در واقع هدف روش‌های دنباله‌ای، شناسایی هر چه زودتر تغییر پس از رخ دادن آن است. به‌طور مثال، چنانچه t_a را زمانی در نظر بگیریم که نمودار کنترل هشدار می‌دهد، t_0 زمان واقعی تغییر است، به‌طوری‌که $t_0 \le t_a$. در نمودار شکل ۱-۲، زمان اعلام اخطار و همچنین زمان واقعی رخ دادن تغییر در روش دنباله‌ای مشخص شده‌اند.

---

[۱] off line
[۲] Sequential method
[۳] on line



شکل ۱-۲: نمودار تشخیص نقطه تغییر در روش دنباله‌ای

در نتیجه مساله تشخیص تغییرات در مشخصه‌های فرآیندهای تصادفی پدید می‌آید که این شاخه از تحقیقات را "تشخیص نقطه تغییر" می‌نامند. از دیگر نام‌های پر کاربرد آن، می‌توان به "مساله بی‌نظمی[۱]" و "تشخیص احتمالی[۲]" اشاره کرد (کمالی، ۱۳۸۹).

در واقع مساله شناسایی تغییرات در زمان‌های نامعلوم و برآورد مکان تغییرات در فرآیندهای تصادفی به مساله نقطه تغییر اشاره دارند. مساله نقطه تغییر دارای چندین جنبه است. اول آن‌که تشخیص دهیم که آیا تغییر وجود دارد و سپس یافتن نقطه تغییری که ممکن است رخ داده باشد و همچنین برآورد تعداد تغییرات، مکان‌های متناظر با آن‌ها و میزان تغییر رخ داده در مشخصه کیفی مورد بررسی از جنبه‌های این مساله می‌باشند. مسایل تشخیص نقطه تغییر را می‌توان بر اساس نوع تغییر به شکل زیر دسته‌بندی کرد:

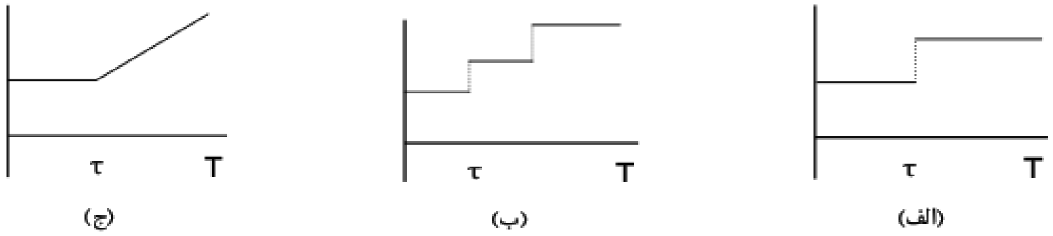
الف) تغییر پله‌ای یگانه[۳]   ب) تغییر پله‌ای چندگانه[۴]   ج) تغییر با روند خطی[۵]

---

[1] Disorder problem
[2] Probabilistic diagnostics
[3] Single point step change (Single abrupt change)
[4] Multiple point step change (Multiple abrupt change)
[5] Linear trend change

تغییر پله‌ای به نوعی از تغییر گویند که پارامتر نرخ فرآیندی که تحت کنترل است در نقطه نامعلومی از زمان از حالت کنترل $\theta = \theta_\circ$ به حالت خارج از کنترل مجهول $\theta = \theta_a$ $(\theta_a \neq \theta_\circ)$ می‌رود و فرآیند در همان حالت باقی می‌ماند تا زمانی‌که علل ایجاد انحراف شناسایی و حذف شوند. چنانچه تغییر پله‌ای رخ‌داده در یک نقطه باشد، به آن تغییر پله‌ای یگانه و چنانچه در بیش از یک نقطه از زمان رخ دهد به آن تغییر پله‌ای چندگانه گویند.

تغییر خطی به نوعی از تغییر گویند که پارامتر نرخ فرآیندی که تحت کنترل است تا زمان τ در حالت تحت کنترل $\theta = \theta_\circ$ می‌باشد و بعد از آن به حالت خارج از کنترل مجهول $\theta = \theta_i$ برای $i = \tau + 1, \dots$ می‌رود، به طوری که $\theta < \theta_i$ و تابع θ_i به صورت $\theta_i = \theta_\circ + \beta(i + \tau)$ است که در آن، τ نقطه تغییر است. در شکل ۱-۳ انواع تغییر نشان داده شده است.



شکل ۱-۳: انواع تغییر در پارامتر- (الف) تغییر پله‌ای یگانه – (ب) تغییر پله‌ای چندگانه – (ج) تغییر با روند خطی

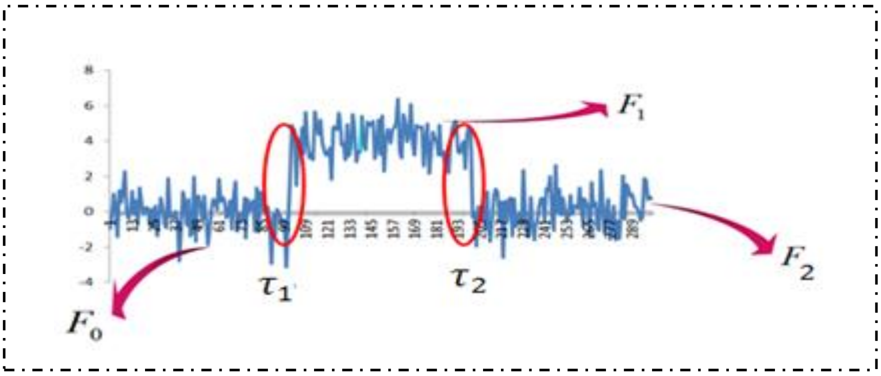
۱-۲-۲ تعریف ریاضی مساله نقطه تغییر

همان طور که گفته شد، نقطه تغییر با تحلیل داده‌های تولید شده توسط فرآیندهایی که دستخوش تغییرات ناگهانی می‌شوند و برآورد محل آن‌ها، مرتبط است. یک سری زمانی را می‌توان به صورت ریاضی مدل‌بندی کرد به طوری که هر مدل توسط مجموعه‌ای از پارامترها مشخص می‌شود. فرض کنید $\{x_t : t = 1, \dots, n\}$ مجموعه داده‌های مشاهده شده باشد که در آن t پارامتر زمان بوده و به‌طور

پیوسته یا گسسته در بازه زمانی $[0\,,T]$ تغییر می‌کند. فرض کنید x_tها مستقل و دارای توزیع زیر باشند:

$$x_t \sim F_j\,, \qquad \tau_j < t < \tau_{j+1}\,, \qquad j = 0, \dots, m\,.$$

که در آن $0 = \tau_0 < \tau_1 < \cdots < \tau_{m+1} = T$. در این صورت نقاط $\tau = (\tau_1, \dots, \tau_m)$ نقاط تغییر فرآیند بوده و در شکل ۱-۴ نمایش داده شده‌اند.



شکل ۱-۴: نمایش مدل ریاضی نقطه تغییر در سری زمانی

در بیشتر کاربردها تعداد نقاط تغییر (m) نامعلوم بوده و از داده‌ها استنباط می‌شود.

در این بخش، سری زمانی و مفهوم نقطه تغییر بیان شد. روش‌های مختلفی برای تشخیص این نقطه وجود دارند که در این پایان‌نامه به برخی از این روش‌ها از دو دیدگاه بیزی و غیربیزی پرداخته شده است. بنابراین ضرورت معرفی مطالبی در مورد آمار بیزی احساس می‌شود.

۱-۳ آمار بیزی

نظریه اصلی در تمام آمار بیزی، تفسیرهای مختلف از قضیه بیز است که در زیر تعریف می‌شود:

قضیه بیز: چنانچه $E_1, E_2, \dots, E_n$ پیشامدهای دو به دو ناسازگار باشند (یعنی $E_i \cap E_j = \Phi$, $i \neq j$) به گونه‌ای که اجتماع آن‌ها فضای نمونه‌ی S را تشکیل دهند (یعنی $S = \cup_{i=1}^{n} E_i$) و چنانچه A هر پیش‌آمد دلخواهی باشد و r یک عدد صحیح و مثبت کوچکتر یا مساوی با n باشد، آنگاه

$$P(E_r|A) = \frac{P(E_r)P(A|E_r)}{\sum_{i=1}^{n} P(E_i)P(A|E_i)}.$$

قضیه بیز در توزیع‌ها به صورت زیر بیان می‌شود:

$P(A)$ به معنی احتمال وقوع پیشامد A و عبارت $p(x)$ خلاصه‌ی عبارت تابع چگالی احتمال $pdf_X(x)$ است. برای مثال اگر y مجموعه داده باشد، برای محاسبه توزیع پارامتر θ، یا مجموعه پارامترهای فرآیند که y را تولید می‌کنند از قضیه بیز به صورت زیر استفاده می‌شود:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$p(\theta)$، توزیع پیشین θ و $p(\theta|y)$، توزیع پسین پارامتر است (دونی[1] ۲۰۰۸).

در بخش بعد ساختار اصلی پایان‌نامه معرفی می‌شود.

۴-۱ ساختار پایان‌نامه

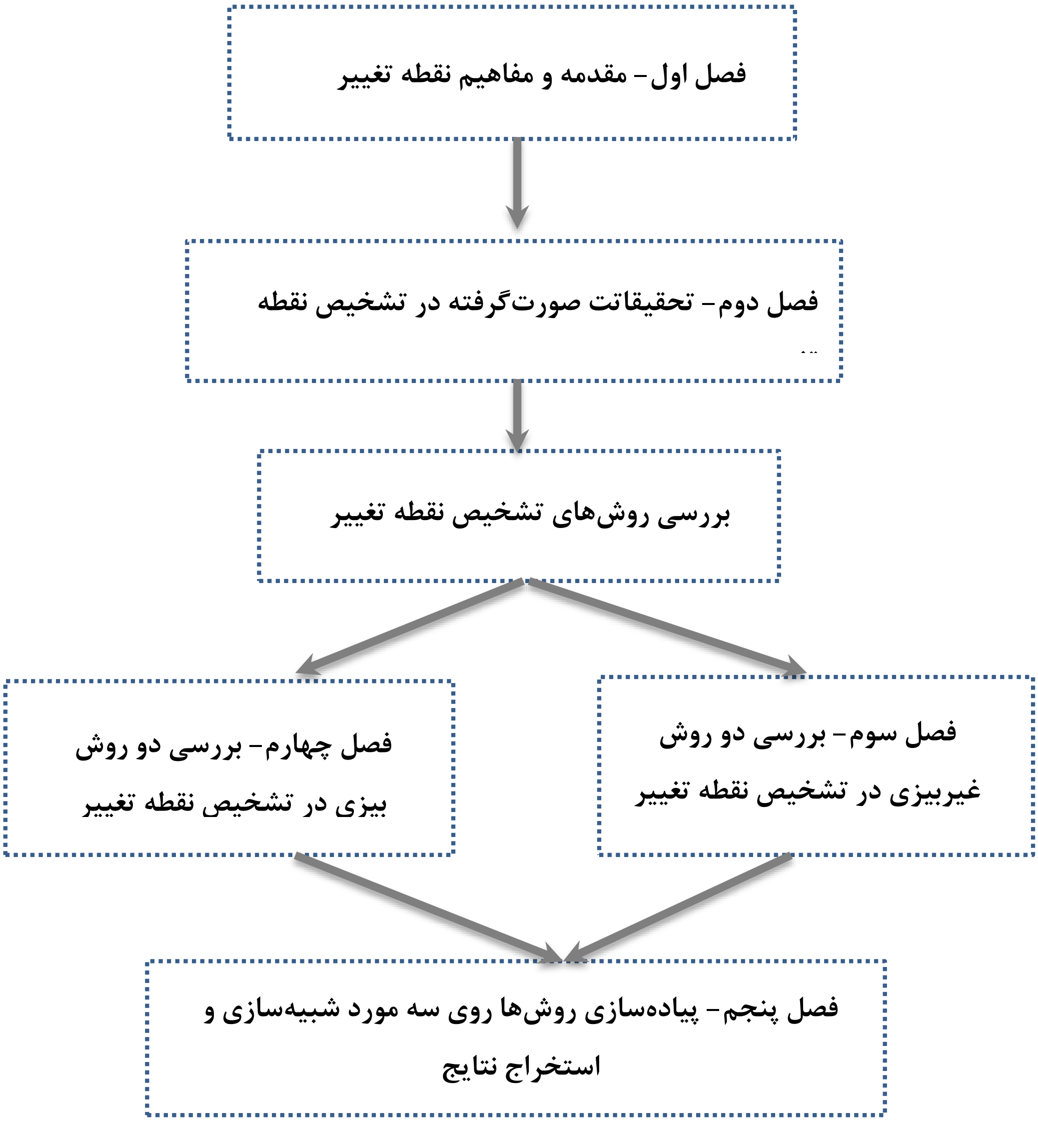
بخش‌های مختلف این پایان‌نامه به صورت زیر تهیه و تنظیم شده‌اند: در فصل دوم، بخشی از تحقیقات صورت گرفته در مورد تشخیص نقطه تغییر معرفی گردیده‌اند. در فصل سوم، به معرفی دو روش غیر-بیزی در تشخیص نقطه تغییر به نام‌های روش مجموع تجمعی CUSUM مبتنی بر نمونه‌های خودگردان و روش تحلیل طیفی منفرد (SSA) پرداخته شده است. در فصل چهارم نیز دو روش بیزی برای تشخیص نقطه تغییر، با دو رویکرد دنباله‌ای (برخط) و غیردنباله‌ای(واخط) هستند به ترتیب به نام‌های روش تشخیص برخط نقطه تغییر بیزی[2] (BOCPD) و "مدل بیزی افراز ضربی (PPM) بیزی"، معرفی شده‌اند.

---

[1] Downey

[2] Bayesian Online Change point Detection

در فصل پنجم، ابتدا به تحلیل داده و ارایه نتایج پرداخته شده و سپس مقایسه‌ای بین روش‌ها انجام گرفته است. در زیر یک نمودار شماتیکی از ساختار پایان‌نامه آورده شده است.



شکل ۱-۵: ساختار کلی پایان‌نامه

فصل دوم

تحقیقات صورت‌گرفته در تشخیص نقطه تغییر

۲-۱ مقدمه

تحقیق روی مساله نقطه تغییر به طور فعال از سال ۱۹۵۰ آغاز شد. هدف این مساله تشخیص تغییرات ناگهانی در روند تولید داده‌ها است، که این تغییرات در زمان‌های ناشناخته اتفاق می‌افتند (باسسیویلی و نیکفوروم، ۱۹۹۳؛ برودسکی[1] و همکاران، ۲۰۰۰).

همان‌طور که در فصل ۱ بیان شد، تشخیص نقطه تغییر در تحلیل سری‌های زمانی در زمینه‌های مختلف کاربردی مورد مطالعه قرار گرفته است. در این فصل، به بررسی برخی از مطالعات صورت‌گرفته در این زمینه پرداخته و در فصل‌های آینده به بررسی چندین روش تشخیص نقطه تغییر می‌پردازیم.

۲-۲ تاریخچه تحقیق

نویسندگان مختلفی مسایل نقطه تغییر را با استفاده از روش‌ها و الگوریتم‌های متفاوت مورد مطالعه قرار داده‌اند. این روش‌ها و الگوریتم‌ها روی انواع خاصی از سری‌های زمانی قابل اجرا هستند، اما یک روش کلی برای همه آن‌ها وجود ندارد. همان طور که در فصل قبل بیان شد، مسایل نقطه تغییر از دو دیدگاه تحلیل دنباله‌ای و غیردنباله‌ای مورد بررسی قرار می‌گیرند. در حالت کلی برخی از روش‌ها در تشخیص نقطه تغییر با فرض مشخص بودن توزیع داده‌ها، به طور مثال هندسی، نرمال، پواسون و غیره (پارامتری)، و برخی دیگر بدون در نظر گرفتن فرضی برای توزیع داده‌ها (ناپارامتری)، انجام می‌شوند (برودسکی و همکاران، ۲۰۰۰).

رویکردهای اصلی که برای تشخیص نقطه تغییر توسط محققان پیشنهاد شده‌اند را می‌توان به سه دسته کلی زیر تقسیم کرد:

---

[1] Brodsky

الف) استفاده از روش درستنمایی ماکسیمم[1](ML) و آزمون نسبت درستنمایی،

ب) استفاده از نمودارهای CUSUM و $EWMA$،

ج) استفاده از روش‌های دیگر مانند تحلیل بیزی، شبکه‌های عصبی، تحلیل طیفی منفرد، خوشه بندی.

مطالعات اولیه برای تشخیص نقطه تغییر، بر فرض نرمال بودن مشخصه مورد بررسی استوار است. یکی از اولین روش‌ها برای یافتن تغییرات نامعلوم در میانگین و واریانس، نمودارهای کنترل است که توسط شوهارت ارایه شد. پیج در اواسط دهه پنجاه (۱۹۵۴، ۱۹۵۵، ۱۹۵۷) مساله نقطه تغییر منفرد را تعریف و با بررسی مساله کنترل کیفیت در تولید مداوم، مساله را به صورت زیر فرمول‌بندی و پیشنهاد کرد.

فرض کنید $x_1, x_2, \dots, x_N$ دنباله‌ای از متغیرهای تصادفی مشاهده‌شده باشد که تغییر پارامتری در نقطه مجهول m در این دنباله رخ داده است و مقدار اصلی پارامتر قبل از تغییر، معلوم و برابر θ است. پیج[2] محاسبه مجموع تجمعی $CUSUM$ را به‌صورت

$$S_{\circ} = 0, \qquad S_r = \sum_{i=1}^{r} (x_i - \theta)\,,$$

پیشنهاد می‌کند. در این روش، اگر اختلاف می‌نیمم مجموع تجمعی از مجموع تجمعی فعلی بیشتر از مقدار خاص h باشد، تغییر رخ داده است (پیج ۱۹۵۵):

$$S_r - min_{\circ \le i \le r} S_i \ge h.$$

---

[1] Maximum likelihood

[2] Page

الگوریتم مجموع تجمعی $CUSUM$ یکی از الگوریتم‌های تشخیص نقطه تغییر است، که به طور گسترده مورد استفاده قرار می‌گیرد. انواع مختلفی از این الگوریتم وجود دارند. باسسیویلی و نیکیفورم (۱۹۹۳) چهار نوع از این الگوریتم را معرفی کردند.

در سال‌های اخیر الگوریتم مجموع تجمعی، تعمیم یافته است، به طور مثال تیلور[1] (۲۰۰۰) روش مجموع تجمعی مبتنی بر نمونه‌های خودگردان[2] را برای تشخیص نقطه تغییر پیشنهاد داد. نمودارهای مجموع تجمعی معرفی‌شده توسط وی، همان نمودارهای مجموع تجمعی است که توسط پیج (۱۹۵۵) و هینکلی[3] (۱۹۷۱) ارایه شده بود. در این پایان‌نامه، به بررسی این روش می‌پردازیم.

اساس الگوریتم $CUSUM$ بر مبنای مدل فرض آماری در مورد پارامتر توزیع، θ، است و به صورت

$$\begin{cases} H_\circ : & \theta = \theta_\circ \\ H_1 : & \theta = \theta_1 \end{cases},$$

تعریف می‌شود. طبق این مدل تا زمانی‌که تصمیم موافق پذیره $H_\circ$ باشد، هیچ تغییری در پارامتر وجود ندارد ولی اگر تصمیمی موافق پذیره H_1 صورت گیرد، تغییر پارامتر وجود خواهد داشت. آزمون برای تشخیص تغییر از نسبت لگاریتم حداکثر درستنمایی[4] استفاده می‌کند، که به صورت

$$S_k = \sum_{i=1}^{k} s_i = \sum_{i=1}^{k} ln \frac{p_{\theta_1}(x_i)}{p_{\theta_\circ}(x_i)} \ ,$$

قابل محاسبه است، که در آن k بیان‌کننده زمان حاضر، x_i سری زمانی، $\theta_\circ$ و θ_1 مقدار پارامترها تحت دو فرضیه $H_\circ$ و H_1 و s_i نسبت لگاریتم حداکثر درستنمایی سری است. این نسبت در نقطه kام، یعنی S_k یک مقدار منفی را قبل از تغییر و یک مقدار مثبت را بعد از تغییر نشان می‌دهد. بنابراین اختلاف

---

[1] Taylor
[2] Bootstrapping
[3] Hinkley
[4] Maximum log likelihood ratio

بین مقدار نسبت لگاریتم حداکثر درستنمایی در نقطه kام و مقدار می‌نیمم نسبت‌های درستنمایی، بیان‌کننده تغییر است (باسسیویلی و نیکفوروم، ۱۹۹۳). به این ترتیب، شرط تشخیص تغییر برای الگوریتم CUSUM به صورت

$$g_k = S_k - m_k \geq h\,, \qquad m_k = \min_{1 \leq j \leq k} S_j\,,$$

می‌باشد، که در آن h پارامتر آستانه بوده و زمان تغییر نیز به‌صورت

$$t_a = \min\{k:\ g_k \geq h\}\,,$$

به دست می‌آید. در نتیجه زمان تغییر، اولین نقطه‌ای است که اختلاف بین مجموع لگاریتم نسبت درستنمایی در این نقطه و می‌نیمم همه مجموع‌ها تا این نقطه از حد آستانه در نظر گرفته‌شده، بیشتر باشد. از این الگوریتم در زمینه‌های کاربردی مختلفی مانند صنعت داروسازی (جاویت[1] و همکارانش، ۲۰۰۹)، و همین طور برای تشخیص نقطه تغییر در دنباله‌های ناایستا برای نظارت بر داده‌های شبکه (دی اوسا[2] و همکارانش، ۲۰۱۰) استفاده می‌شود.

یکی دیگر از روش‌های تشخیص نقطه تغییر، استفاده از برآوردگر حداکثر درستنمایی است، که برای تشخیص زمان تغییر پله‌ای در میانگین توزیع نرمال به کار می‌رود. ساموئل[3] و همکارانش (۱۹۹۸) آن را بر روی داده‌هایی که از نمودار کنترل $\bar{x}$ گرفته شده، به کار بستند. پینگاتیلو[4] و ساموئل (۲۰۰۱) عملکرد برآوردگر ارایه‌شده را نسبت به تشخیص‌های $CUSUM$ و $EWMA$ مقایسه کرده و به این نتیجه رسیدند که این برآوردگر عملکرد مناسب‌تری را از زمان تغییر ارایه می‌دهد (پینگاتیلو و سیمپسون[5]، ۲۰۰۲).

---

[1] Gavit
[2] De.oca
[3] Samuel
[4] Pignatiello
[5] Simpson

روش‌های CUSUM بیان‌شده در بالا، جزء روش‌های غیردنباله‌ای (واخط) تشخیص نقطه تغییر هستند. یکی دیگر از روش‌های غیردنباله‌ای تشخیص نقطه تغییر که رویکردی بیزی نیز دارد و در این پایان‌نامه به آن پرداخته می‌شود روش مدل افراز ضربی است که توسط بری و هارتیگن[۱] (۱۹۹۳) معرفی شده است.

همان طور که قبلا اشاره شد، رویکرد دیگر روش‌های آماری مسایل نقطه تغییر، روش‌های دنباله‌ای (برخط) است. یکی از روش‌های آزمون دنباله‌ای در فرآیندهای تصادفی، مدل‌سازی میانگین متحرک اتورگرسیو[۲] (ARMA) است که توسط نازاریو[۳] و همکارانش (۱۹۹۷) توسعه یافته است. تحلیل اولیه نشان می‌دهد که اگر سری زمانی $ARMA(p,q)$ رفتار ناپایداری را نشان دهد، رفتار باقیمانده‌ها همانند $ARMA(Q,Q)$ خواهد شد، به طوری که $Q \le p+q$ و باقیمانده‌های مدل، قبل از تغییر پارامتر تقریباً به‌صورت دنباله ای از متغیرهای تصادفی مستقل رفتار می‌کنند و بعد از تغییر پارامتر، باقیمانده‌ها همبسته می‌شوند. بر این اساس آن‌ها یک آزمون دنباله‌ای جدید برای تشخیص رفتار ناپایدار در سری زمانی ARMA به دست آوردند. ایراد آزمون این است که فقط قادر به تشخیص تغییرات پارامتری است که باعث تغییرات قابل ملاحظه‌ای در تابع خودهمبستگی[۴] (ACF) باقیمانده‌ها می‌شود، ولی اگر تغییر در پارامتر منجر به تغییر کوچکی در ACF باقیمانده‌ها شود، ممکن است این تغییر تشخیص داده نشود.

آدامز و مک کی[5] (۲۰۰۷) روش تشخیص برخط نقطه تغییر بیزی را ارایه کردند که در این پایان‌نامه به طور مفصل مورد بررسی قرار خواهد گرفت. یکی از کاربردهای روش‌های دنباله‌ای، تحلیل

---

[1] Barry and Hartigan’s

[1] Autoregressive moving average

[2] Nazario

[3] Autocorrelation function

[4] Adams and MacKay

برخط تغییرپذیری ضربان قلب در خواب است (اسایودچر[۱] و همکارانش، ۲۰۰۵). این روش از تحلیل نوسان روندزدایی شده[۲] (DFA) استفاده می‌کند. روش جدید بیان‌شده، تحلیل نوسان خروجی پیش رونده[۳] (PDFA) نامیده می‌شود. اگر چه این روش ابزاری برای ارزیابی عددی خواب بر مبنای مطالعه تغییرپذیری ضربان قلب از طریق الکتروکاردیوگراف (ECG) می‌باشد، ولی ثابت شده است که این روش در بیشتر سری‌های زمانی قابل اجرا است.

برخی از محققان روش‌هایی از تشخیص نقطه تغییر را بیان کردند که رویکردی ریاضی داشته و وابسته به تشخیص زیرفضایی است. روش‌های اصلی زیرفضایی شامل تحلیل طیفی منفرد (SSA) بر مبنای روش گولیاندینا و همکاران[۴] (۲۰۰۱) و موسکوینا و همکاران[5] (۲۰۰۳) و مدل فضای حالت[۶] (SSM) مبتنی بر روش کاواهارا و همکاران (۲۰۰۷) است.

روش SSM، روشی جدیدتر و کلی‌تر برای مساله تشخیص نقطه تغییر با استفاده از تشخیص زیرفضایی است. این روش از SSMهای کلی به عنوان مدلی برای سری‌های زمانی استفاده می‌کند. به این ترتیب با انواع بیشتری از سری‌های زمانی سروکار خواهد داشت. عیب این روش محاسبات حجیم آن است (کاواهارا و همکاران ۲۰۰۷). بنابراین از روشی دیگر که مبتنی بر تحلیل طیفی منفرد SSA بوده و یک الگوریتم تشخیصی زیرفضایی بسامدی و قدیمی‌تر است، استفاده می‌شود. گرچه این روش عمومیت کمتری نسبت به روش SSM دارد، ولی فرآیند حل آن ساده‌تر می‌باشد. روش SSA، اجرای تجزیه مؤلفه اصلی از ماتریس مسیر، بر مبنای تعبیه‌سازی تیکنز[۷] (تیکنز، ۱۹۸۱) از سری زمانی اصلی و سپس تحلیل سری با استفاده از این مؤلفه‌ها است.

---

[5] Staudacher

[۱] Detrended fluctuation analysis

[۲] Progressive detrended fluctuation analysis

[۳] Goljandina

[۴] Moskvina

[۵] State space model

[۶] Takens

[۷] Bhattacharya and Frierson

از دیدگاهی دیگر، روش‌های تشخیص نقطه تغییر را می‌توان به روش‌های پارامتری و ناپارامتری

تقسیم‌بندی کرد. روش‌های ذکرشده، از روش‌های پارامتری بودند. اولین روش ناپارامتری توسط باتاچاریا و فرایرسون[1] (۱۹۸۱) ارایه شد. برودسکی و دارخوسکی[2] (۲۰۰۰، ۱۹۹۹) استفاده از آماره اسمیرنوف-کولموگروف[3] را پیشنهاد کردند که به صورت

$$Y_N(n,\delta) = [\left(1-\frac{n}{N}\right)\frac{n}{N}]^{\delta}\left[\frac{1}{n}\sum_{k=1}^{n} x^N(k) - \frac{1}{N-n}\sum_{k=n+1}^{N} x^N(k)\right],$$

تعریف می‌شود، که در آن $1 \le n \le N-1$, $۰ \le \delta \le 1$ و x^N دنباله‌ی تشخیصی[4] برای سیگنال و δ، احتمال اعلان خطرهای اشتباه است. دنباله تشخیصی به شکل تابعی مرحله‌ای[5] از زمان و نوفه فرض شده و آماره‌ی Y_N برای این دنباله محاسبه و سپس مقدار ماکزیمم مطلق آن تعیین می‌شود. اگر این مقدار از آستانه‌ی تعیین شده‌ی داده‌ها بزرگتر باشد، مکان این مقدار ماکزیمم را به عنوان نقطه تغییر در نظر می‌گیریم. برای تشخیص نقاط تغییر بیشتر، دنباله به دو بخش قبل و بعد از تغییر تقسیم‌بندی شده و نقاط تغییر در هر بخش تعیین می‌شوند. فرآیند تقسیم‌بندی تا زمانی که دیگر هیچ نقطه‌ی تغییری وجود نداشته باشد، تکرار می‌شود (برودسکی و همکاران، ۱۹۹۹، ۲۰۰۰).

حبیبی و همکارانش (۲۰۰۵) آماره‌ی آزمونی را برای تشخیص نقطه تغییر در کلاس کلی توزیع‌ها به‌دست آوردند. روش ناپارامتری جدیدی که برای تشخیص تغییر در توزیع‌های احتمال دنباله‌ی داده استفاده می‌شود، توسط کاواهارا و سوگی‌یاما (۲۰۰۹) ارایه گردید که نسبت چگالی احتمال را به طور مستقیم و بدون برآورد چگالی معمول، برآورد می‌کند.

---

[1] Brodsky and Darkhovsky
[2] Kolmogorov smirnov statistics
[3] Diagnostic sequence
[4] Step function

روش‌ها و رویکردهای فراوان دیگری نیز برای تشخیص نقطه تغییر وجود دارند و در این‌جا به چند نمونه از آن‌ها اشاره می‌کنیم. علاوه بر روش‌های ذکرشده، انواع روش‌های رگرسیونی (رگرسیون لجستیک، رگرسیون خطی ماتریس مشخصه و غیره) نیز برای تشخیص این نقطه وجود دارند که بررسی کاملی از این روش‌ها توسط خدادادی و اصغری انجام شده است. در این پایان‌نامه، چندین روش از رویکردهای مختلف (کنترل کیفی، زیرفضایی و بیزی) مورد بررسی و مقایسه قرار گرفته اند.

فصل سوم

بررسی دو روش غیربیزی

در تشخیص نقطه تغییر

در این فصل به معرفی دو روش مجموع تجمعی مبتنی بر نمونه‌های خودگردان و تحلیل طیفی منفرد می‌پردازیم که به ترتیب دارای دو رویکرد کنترلی و زیرفضایی هستند. این دو روش از جمله روش‌های غیر بیزی برای تشخیص نقطه تغییر می‌باشند.

۳-۱ روش مجموع تجمعی ($CUSUM$) مبتنی بر نمونه‌های خود گردان[۱]

همان طور که در فصل اول گفته شد، نمودارهای کنترل از مهمترین ابزارهای کنترل کیفیت آماری در کشف انحرافات با دلیل[۲] در فرآیندها هستند. یکی از معایب این نمودارها، عدم توانایی آن‌ها در کشف نقاط تغییری است که خارج از حدود کنترل نیستند. برای حذف منابع اصلی ایجاد خطا، تشخیص زمان واقعی آغاز انحراف در فرآیند که به آن نقطه تغییر گفته می‌شود، اهمیت بسزایی دارد. تشخیص زمان واقعی تغییر، محدوده‌ی جستجو در خصوص علل بروز انحرافات را محدودتر کرده و شانس یافتن منابع اصلی ایجاد انحراف را به حداکثر می‌رساند و موجب صرفه‌جویی در زمان و هزینه خواهد شد. یکی از روش‌های تحلیل نقطه تغییر که بسیار ساده بوده و رویکرد کنترل کیفی دارد، نمودار $CUSUM$ است. این روش توسط تیلور (۲۰۰۰) تعمیم داده شده است و از نمودارهای $CUSUM$ مبتنی بر نمونه‌های خودگردان استفاده می‌کند. این روش می‌تواند تغییراتی را که توسط نمودار کنترل تشخیص داده نشده است را شناسایی کند و نسبت به نقاط پرت انعطاف‌پذیر است.

۳-۱-۱ فرآیند اجرای تحلیل نقطه تغییر

در این روش برای اجرای تحلیل نقطه تغییر، به‌طور مکرر از نمودارهای $CUSUM$، مبتنی بر نمونه‌هایی خودگردان از تمام داده‌ها استفاده می‌شود. در ابتدا نمودار $CUSUM$ را معرفی می‌کنیم.

---

[1] Bootstrap samples
[2] Assignable causes

روش ساخت نمودارهای $CUSUM$

فرض کنید $x_1, x_2, \dots, x_n$ یک سری مشاهده شده به طول n باشد. $CUSUM$ در هر نقطه زمانی t با S_t نشان داده شده و برابر است با مجموع انحراف از میانگین مشاهدات اول تا t ام یعنی

$$S_t = \sum_{i=1}^{t}(x_i - \bar{x}), \qquad t = 1, \dots, n\,,$$

که در آن $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. همواره داریم $S_n = \sum_{i=1}^{n}(x_i - \bar{x}) = 0$. تیلور (۲۰۰۰a)، برآوردگری به نام اندازه‌ی بزرگی تغییر که با نماد S_{diff} نشان داده می‌شود و برای تشخیص تغییر در میانگین و واریانس مفید است، به صورت

$$S_{diff} = S_{max} - S_{min}\,,$$

تعریف کرده است، که در آن $S_{max} = \max_{i=1,\dots,n} S_i$ و $S_{min} = \min_{i=1,\dots,n} S_i$.

نمودار مقادیر S_t نسبت به زمان را نمودار $CUSUM$ نامند. روند افزایشی نمودار $CUSUM$ در یک بازه زمانی نشان می‌دهد که در آن بازه، مقدار همه مشاهدات بیشتر از میانگین بوده و روند کاهشی آن حاکی از آن است که همه مشاهدات، دارای مقداری کمتر از میانگین هستند. بنابراین نقطه مرزی بین روند افزایشی و کاهشی را می‌توان به عنوان نقطه تغییر در سری زمانی تلقی کرد. وی همچنین برای اطمینان از درستی تشخیص تغییر، از یک سطح اطمینان استفاده می‌کند که با انجام نمونه‌گیری خودگردان از تمام داده‌ها به دست می‌آید.

فرض کنید N نمونه‌ی خودگردان به حجم n از مشاهدات $x_1, x_2, \dots, x_n$ انتخاب شود. باید توجه داشت که ممکن است بعضی از مشاهدات، تکراری باشند و برخی دیگر نیز انتخاب نشوند. برای هر یک از N نمونه گرفته‌شده، مقادیر مجموع تجمعی S_i^k، S_{diff}^k، S_{max}^k و S_{min}^k محاسبه

می‌شوند که $k = 1, \dots N$ معرف k امین نمونه خودگردان است. سطح اطمینان برای نقطه تغییر از رابطه (۳-۱) به دست می‌آید(هینکلی و اسچیچمن[1] ۱۹۸۷).

$$\text{سطح اطمینان} = 100\,\frac{X}{N}\,\%. \tag{1-3}$$

که در آن X تعداد نمونه‌های خودگردان است که برای آن‌ها $S_{diff}^{k} < S_{diff}$. بهترین سطح اطمینان برای معنی‌دار بودن تغییر، سطح اطمینانی بالاتر از ۹۰٪ یا ۹۵٪ است. با توجه به محاسبه سطح اطمینان می‌توان فاصله اطمینانی نیز برای تغییر تشخیص داده شده به دست آورد، به این صورت که با محاسبه واریانس S_{diff} های به دست آمده از نمونه‌های خودگردان، یک فاصله اطمینان برای S_{diff} اصلی محاسبه می‌شود. با توجه به فاصله اطمینان به دست آمده برای S_{diff}، فاصله اطمینان متناظر با آن برای نقطه تغییر محاسبه می‌شود (تیلور a ۲۰۰۰).

قبل از بیان جزییات بیشتر در مورد تحلیل نقطه تغییر، طریقه‌ی رسم نمودار $CUSUM$، به‌دست آوردن اندازه بزرگی و سطح اطمینان، توسط مثالی در زیر نشان داده شده است.

مثال ۳-۱: داده‌های مربوط به کسری بودجه ایالات متحده در طی دو سال بر حسب میلیون دلار را در نظر بگیرید (ویلر، ۱۹۹۳).

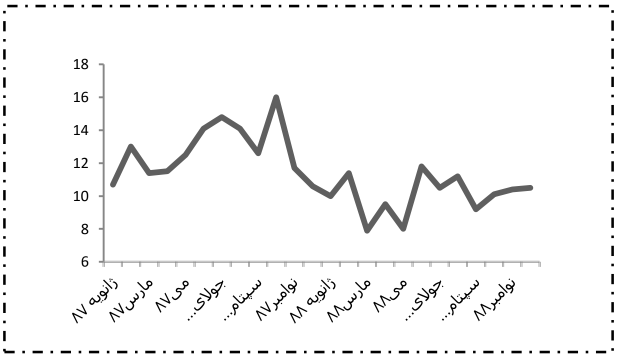
جدول۳-۱: داده‌های کسری بودجه ایالات متحده در طی دو سال

سال	ژانویه	فوریه	مارس	آوریل	می	ژوئن	جولای	اگوست	سپتامبر	اکتبر	نوامبر	دسامبر
۱۹۸۷	۱۰٫۷	۱۳	۱۱٫۴	۱۱٫۵	۱۲٫۵	۱۴٫۱	۱۴٫۸	۱۴٫۱	۱۲٫۶	۱۶	۱۱٫۷	۱۰٫۶
۱۹۸۸	۱۰	۱۱٫۴	۷٫۹	۹٫۵	۸	۱۱٫۸	۱۰٫۵	۱۱٫۲	۹٫۲	۱۰/۱	۱۰٫۴	۱۰٫۵

نمودار زمانی این داده‌ها در شکل ۳-۱ رسم شده است.

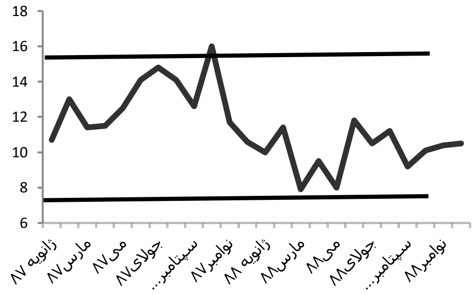
---

[1] Hinkley and Schechtman



شکل ۳-۱: نمودار داده‌های کسری بودجه ایالات متحده در طی دو سال

همان طور که در نمودار ملاحظه می‌شود، سطح کسری بودجه از سال ۸۷ به ۸۸ کاهش یافته است. در ابتدا برای تشخیص نقطه تغییر از نمودار کنترل استفاده می‌کنیم که در آن، حدود کنترل به صورت $\bar{x} \pm s$ تعریف شده و نقاطی که خارج از حدود کنترل هستند، به عنوان نقاط تغییر شناسایی می‌شوند.



$$\bar{x} + \sigma = 11{/}39583 + 4{/}020417$$

$$\bar{x} - \sigma = 11{/}39583 - 4{/}020417$$

این نمودار، تغییری در اکتبر سال ۸۷ را نشان می‌دهد. این سوال مطرح می‌شود که آیا تغییر تشخیص داده‌شده، صحیح است؟ و آیا تنها این تغییر در داده‌ها وجود دارد؟ در خصوص پاسخ به این سوالات و با توجه به داده‌های کسری بودجه داریم:

$\bar{x} = 11{/}39583,$

$S_{\circ} = 0,$

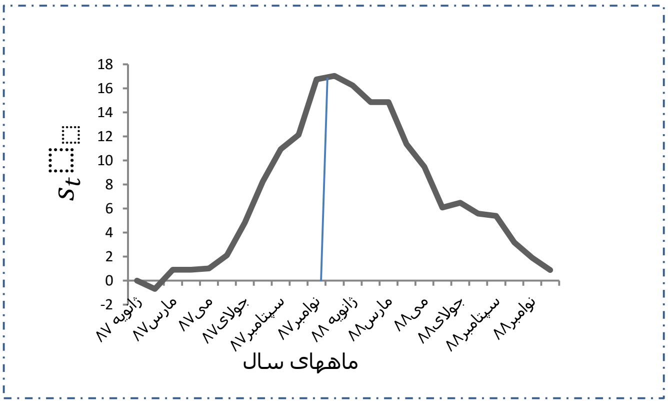
$S_1 = S_{\circ} + (x_1 - \bar{x}) = 0 + (10{/}7 - 11{/}39583) = -0{/}69583,$

$S_2 = S_1 + (x_2 - \bar{x}) = -0{/}69583 + \left(13 - \frac{11}{39583}\right) = 0{/}908333,$

$$\vdots$$

$$S_{24} = S_{23} + (x_{24} - \bar{x}) = 0/89583 + (10/5 - 11/39583) = 0.$$

نمودار مقادیر S_i برای $i = 1, \ldots, n$ در شکل ۳-۲ رسم شده است.



شکل ۳-۲: نمودار $CUSUM$ داده‌های کسری بودجه ایالات متحده در طی دو سال

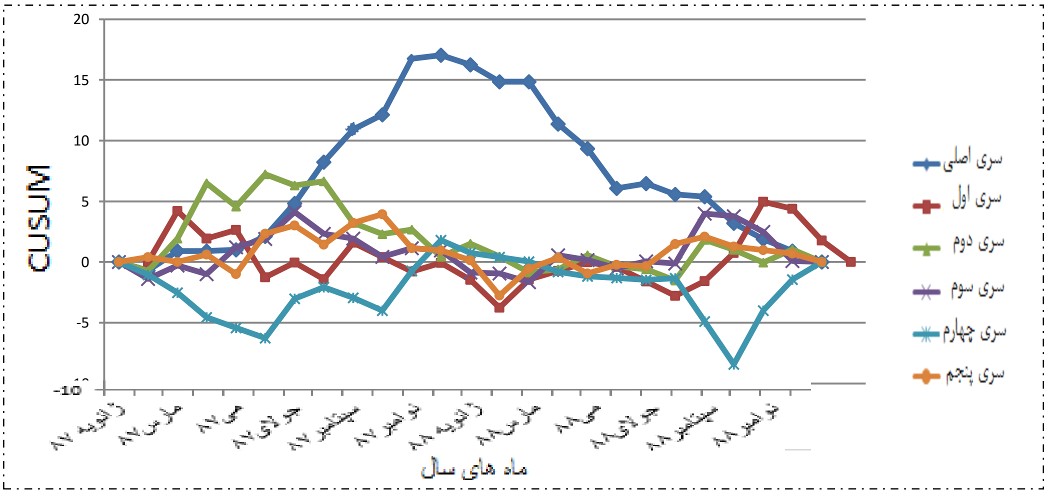
با توجه به نمودار $CUSUM$ در شکل ۳-۲ ملاحظه می‌شود که شیب نمودار در اطراف نوامبر۸۷ که دارای بزرگ‌ترین مجموع تجمعی است، تغییر یافته است. بنابراین می‌توان گفت که تغییری در این ناحیه وجود دارد، اما تشخیص تغییرات دیگر و اطمینان از وجود تغییر و تشخیص مکان آن با استفاده از نمودار $CUSUM$ کار دشواری است. زیرا مجموع تجمعی تا ماه می ۸۷ روند افزایشی کندی داشته و در ماه ژوئن این روند رشد سریعی پیدا کرده و در ماه جولای دوباره مجموع تجمعی به کندی افزایش یافته است. بنابراین بین ماه می۸۷ و جولای ۸۷ تغییری در داده‌ها رخ داده ولی چون تغییری در جهت نمودار $CUSUM$ ملاحظه نشده است، لذا این تغییر تشخیص داده نشده است. بنابراین استفاده از نمودار $CUSUM$ نیازمند مهارت زیادی در این زمینه است. برای تشخیص ساده‌تر تغییر و تشخیص مکان آن، از یک سطح و فاصله اطمینان استفاده می‌شود. برای به‌دست آوردن سطح و فاصله اطمینان برای این نقطه تغییر از نمونه‌های خودگردان ساز به صورت آن‌چه در قبل بیان شد، استفاده می‌شود. برای به دست آوردن سطح اطمینان معقولی برای داده‌ها، حداقل به ۱۰۰۰ نمونه خودگردان نیاز است. برای داده‌های کسری

بودجه و به منظور توضیح مطلب، فقط از $N = 5$ نمونه خودگردان به شرح جدول ۳-۲ استفاده کرده‌ایم.

جدول ۳-۲: داده‌های اصلی مثال کسری بودجه و ۵ نمونه خودگردان

سال	ژانویه	فوریه	مارس	آپریل	می	ژوئن	جولای	اگوست	سپتامبر	اکتبر	نوامبر	دسامبر
۱۹۸۷	۱۰٫۷	۱۳	۱۱٫۴	۱۱٫۵	۱۲٫۵	۱۴٫۱	۱۴٫۸	۱۴٫۱	۱۲٫۶	۱۶	۱۱٫۷	۱۰٫۶
۱۹۸۸	۱۰	۱۱٫۴	۷٫۹	۹٫۵	۸	۱۱٫۸	۱۰٫۵	۱۱٫۲	۹٫۲	۱۰٫۱	۱۰٫۴	۱۰٫۵
نمونه اول	۱۶	۹٫۵	۱۲٫۵	۷٫۹	۱۳	۱۰٫۴	۱۴٫۸	۱۰٫۶	۱۰٫۶	۱۲٫۵	۱۰٫۴	۹٫۵
	۱۴٫۱	۱۲٫۵	۱۲٫۵	۱۱٫۴	۱۰٫۶	۱۰٫۶	۱۳	۱۴٫۱	۱۶	۱۱٫۲	۹٫۲	۱۰
نمونه دوم	10٫7	۱۴٫۱	۱۶	۹٫۵	۱۴٫۱	۱۰٫۵	۱۱٫۸	۸	۱۰٫۵	۱۱٫۸	۹٫۲	۱۲٫۵
	۱۰٫۴	۱۰٫۱	۱۱٫۷	۱۲٫۶	۱۰٫۵	۱۱٫۲	۱۰٫۵	۱۴٫۸	۱۰٫۶	۱۰٫۴	۱۲٫۵	۱۰٫۴
نمونه سوم	۱۰٫۵	۱۳	۱۱٫۲	۱۴٫۱	۱۲٫۶	۱۴٫۱	۱۰٫۱	۱۰٫۴	۱۰٫۴	۱۲٫۶	۱۱٫۷	۱۰٫۱
	۱۱٫۸	۱۱٫۲	۱۴٫۱	۱۱٫۵	۱۱٫۲	۱۲٫۵	۱۱٫۷	۱۶	۱۱٫۷	۱۰٫۶	۹٫۵	۱۱٫۸
نمونه چهارم	۱۰٫۵	۱۰٫۱	۹٫۵	۱۰٫۷	۱۰٫۷	۱۴٫۸	۱۲٫۵	۱۰٫۷	۱۰٫۵	۱۴٫۸	۱۴٫۱	۱۰٫۵
	۱۱٫۲	۱۱٫۲	۱۰٫۷	۱۱٫۲	۱۱٫۴	۱۱٫۴	۱۱٫۷	۸	۸	۱۶	۱۴٫۱	۱۳
نمونه پنجم	۱۱٫۲	۱۰٫۴	۱۱٫۴	۹٫۲	۱۴٫۱	۱۱٫۵	۹٫۲	۱۲٫۶	۱۱٫۵	۸	۱۰٫۶	۱۰
	۷٫۹	۱۳	۱۱٫۷	۹٫۵	۱۱٫۵	۱۰٫۷	۱۲٫۶	۱۱٫۴	۱۰	۱۰/۵	۱۰٫۵	۱۰٫۱

مجموع تجمعی $S_1^k, S_2^k, \dots, S_n^k$ برای $k = 1, \dots, 5$، محاسبه شده و نمودار $CUSUM$ آن‌ها به صورت همزمان در شکل ۳-۳ رسم شده است.



شکل ۳-۳: نمودار $(CUSUM)$ برای داده‌های اصلی و ۵ نمونه خودگردان

در جدول ۳-۳، مقادیر S_{diff}، S_{max}، S_{min} نمونه‌های مندرج در جدول ۳-۲ ذکر شده‌اند.

جدول ۳-۳: مقادیر S_{diff}، S_{max}، S_{min} برای N=5 نمونه خودگردان از نمونه‌ی داده کسری بودجه

نمونه پنجم	نمونه چهارم	نمونه سوم	نمونه دوم	نمونه اول	نمونه اصلی	
۳٫۹۳۷۵	۱٫۸۰۴۱	۴٫۱۲۵۰	۷٫۲۳۴	۴٫۹۶۳	۱۷٫۰۴۵۹	S_{max}
- ۲٫۷۴۵۸	- ۸٫۴۳۷۶	-۱٫۶۴۱۶	-۱٫۵۳۳	- ۳٫۷۵	-۰٫۶۹۵۸	S_{min}
۶٫۶۸۳۳	۱۰٫۲۴۱۷	۵٫۷۶۶۷	۸٫۱۸۲۸	۸٫۷۱۳	۱۷٫۷۴۲	S_{diff}

با توجه به مقادیر S_iهای به دست آمده برای نمونه اصلی (شکل ۳-۳)، ملاحظه می‌شود که بزرگ‌ترین S_i متعلق به داده نوامبر ۸۷ است و از آن‌جا که مقدار S_{diff} نمونه‌ها، کمتر از مقدار S_{diff} داده‌های کمتر است (جدول ۳-۳)، لذا نقطه نوامبر ۸۷ به عنوان نقطه تغییر تشخیص داده می‌شود. همچنین از مقایسه نمودار $CUSUM$ داده‌های اصلی با نمودارهای برگرفته از ۵ نمونه خودگردان (شکل ۳-۳) ملاحظه می‌شود که نمونه‌های گرفته‌شده، رفتار نمودار $CUSUM$ را که دارای تغییر نیست تقلید کرده و در اطراف خط صفر تغییر می‌کنند، و با نمودار اصلی متفاوت بوده و نشان‌دهنده تغییر در نمودار اصلی است.

در عمل باید مراحل بالا با تعداد نمونه‌های خیلی بزرگتری انجام شود. زیرا در حجم نمونه کوچک، یک مقدار اشتباه اندازه گیری شده یا یک نقطه نادیده گرفته شده ممکن است تحلیل مساله را تغییر دهد ولی در حجم نمونه بالا تحلیل دقیق‌تری خواهیم داشت. با توجه به مطالب گفته‌شده و رابطه (۳-۱)، سطح و فاصله اطمینان برای نقطه تغییر به دست آمده است (جدول ۳-۴) و با توجه به سطح اطمینان به دست آمده، نقاط تغییر رتبه‌بندی می‌شوند.

جدول ۳-۴: تشخیص تغییر داده‌های کسری بودجه

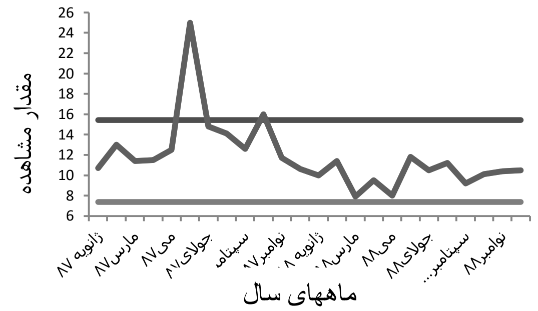
ماه	**فاصله اطمینان**	**سطح اطمینان**	**میانگین قبل تغییر**	**میانگین بعد تغییر**	**سطح تغییر**
ژوئن۸۷	**(می۸۷,جولای۸۷)**	**۹۱٪**	**۱۱٫۸۲**	**۱۴٫۳۲**	**۲**
نوامبر۸۷	**(نوامبر۸۷,نوامبر۸۷)**	**۱۰۰٪**	**۱۴٫۳۲**	**۱۰٫۲**	**۱**

با توجه به جدول ۳-۴، ملاحظه می‌شود که دو تغییر در داده‌ها با دو سطح اطمینان رخ داده، که تغییر اول با ۱۰۰٪ اطمینان مربوط به ماه نوامبر ۸۷ (بین نوامبر۸۷ و نوامبر۸۷) است که کسری بودجه از ۱۴٫۳۲ به ۱۰٫۲ کاهش یافته است و تغییر دوم با ۹۱٪ اطمینان در حدود ماه ژوئن (بین ماه می و جولای ۸۷) است که کسری بودجه از ۱۱٫۸ به ۱۴٫۳۲ افزایش یافته است. همان طور که مشاهده شد، نمودار کنترل تنها قادر به تشخیص یک تغییر بود، ولی تحلیل انجام‌شده توانست تغییر دیگری را که نمودار کنترل قادر به تشخیص آن نبود را تشخیص داده و اطلاعات بیشتری از تعداد و محل تغییر و اطمینان از وجود تغییر ارایه دهد (صمدی، ۲۰۰۶).

در این‌جا این سوال مطرح می‌شود که آیا این روش تمام نقاط خارج از حدود کنترل را به عنوان نقطه تغییر تشخیص می‌دهد؟ این روش در مقابل داده‌هایی که دارای مقادیر پرت هستند چه عملکردی دارد؟ به این سوالات با بررسی تحلیل در بخش بعد پاسخ داده می‌شود.

۳-۱-۲ بررسی نقاط پرت

در شناسایی نقاط تغییر، تمایز بین نقطه پرت و نقطه تغییر بسیار حایز اهمیت است. زیرا وجود نقطه پرت در داده‌ها باعث تغییرات اضافی در داده‌ها شده و تشخیص دقیق تغییر را مشکل می‌سازد. برای بررسی عملکرد روش تیلور نسبت به نقاط پرت، فرض کنید در مثال قبل، داده ماه ژوئن۸۷ برابر ۲۵ باشد که یک مقدار پرت است. نمودار کنترل داده‌های جدید به صورت شکل ۳-۴ خواهد بود.



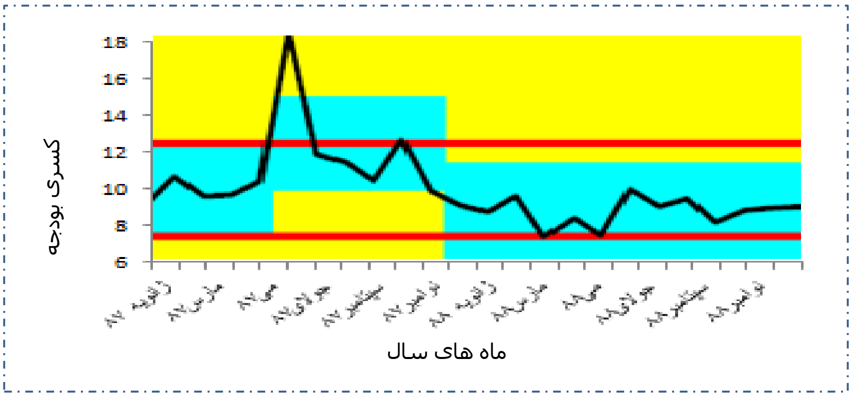
شکل ۳-۴: نمودار کنترل داده کسری بودجه با تغییر داده ماه ژوئن۸۷ به مقدار ۲۵

با توجه به نمودار، ملاحظه می‌شود که علاوه بر ماه نوامبر ۸۷، داده ماه ژوئن۸۷ نیز یک نقطه خارج از کنترل است. حال مراحل تحلیل را برای شناسایی نقاط تغییر روی داده‌های جدید به‌کار می‌بریم و نتایج زیر حاصل می‌شود.

جدول ۳-۵: تشخیص نقطه تغییر برای داده کسری بودجه با تغییر داده ماه ژوئن۸۷ به مقدار ۲۵

ماه	فاصله اطمینان	سطح اطمینان	میانگین قبل تغییر	میانگین بعد تغییر	سطح تغییر
ژوئن۸۷	(آوریل۸۷,آگوست۸۷)	٪۹۱	۱۱٫۸۲	۱۵٫۷	۲
دسامبر۸۷	(نوامبر۸۷,فوریه ۸۸)	٪۱۰۰	۱۵٫۱	۱۰٫۰۸۵	۱

با توجه به جدول ۳-۵، ملاحظه می‌شود که نقطه ژوئن ۸۷ همانند قبل به عنوان دومین نقطه تغییر تشخیص داده شده و افزایش مقدار آن به عنوان نقطه پرت در تشخیص نقطه تغییر موثر نبوده است. در نمودار شکل ۳-۵، نقطه ژوئن ۸۷ خارج از حدود کنترل است. در این روش نقاطی که خارج از حدود کنترل هر بخش از داده‌ها قرار می‌گیرند، به عنوان نقاط پرت تشخیص داده می‌شوند زیرا این روش برای هر بخش از داده‌ها که به وسیله نقاط تغییر از هم جدا می‌شوند، حدود کنترل را محاسبه می‌کند و نقاطی که خارج از این حدود باشند را به عنوان نقاط پرت تشخیص می‌دهد.



شکل ۳-۵: نمودار تشخیص حدود کنترل برای نقاط تغییر به دست آمده و نقاط پرت

بنابراین روش تحلیل نقطه تغییر که در این بخش بیان شد نسبت به نقاط پرت قوی بوده و قادر به تشخیص نقاط پرت است و یک ابزار مفید برای تشخیص تغییراتی است که توسط نمودار کنترل شناسایی نشده‌اند. روش‌های بیان‌شده در بخش‌های قبل برای تشخیص تغییر در میانگین داده مفید است ولی برای تشخیص تغییر در واریانس، تیلور (۲۰۰۰b) روش زیر را ارایه می‌دهد.

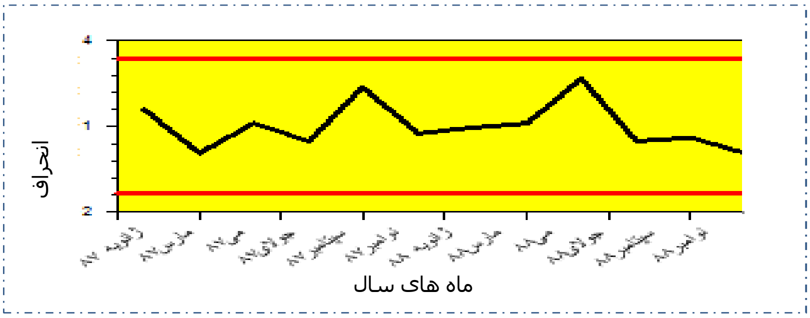
۳-۱-۳ تشخیص تغییر در واریانس و ویژگی‌های دیگر

روش ارایه‌شده در بخش ۳-۱ برای تشخیص تغییر در هر ویژگی از داده‌های سری زمانی نیز به‌کار برده می‌شود. این روش می‌تواند روی میانگین‌ها، انحراف استانداردها، رتبه‌ها، سطوح نقص‌ها و انواع دیگر داده به‌کار رود. برای تشخیص تغییر در واریانس داده کسری بودجه از اختلاف بین مشاهدات برای تحلیل نقطه تغییر به صورت زیر استفاده می‌شود.

فرض کنید $x_1, x_2, \dots, x_{24}$ داده‌های کسری بودجه برای دو سال ۸۷ و ۸۸ باشند. این ۲۴ نقطه دارای ۱۲ اختلاف $D_1, D_2, \dots, D_{12}$ هستند که به‌صورت

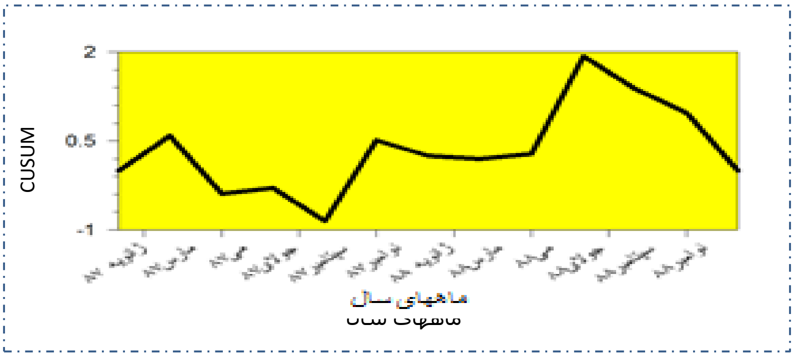
$$D_i = |x_{2i} - x_{2i-1}|, \qquad i = 1,2, \dots ,12 ,$$

محاسبه می‌شوند. نقاط مشترک در D_i ها وجود ندارد و هر D_i مستقل از دیگری است. نمودار کنترل برای دنباله‌ی D_i ها به صورت شکل ۳-۶ رسم می‌شود.



شکل ۳-۶: نمودار کنترل انحراف استاندارد داده‌های کسری بودجه

با توجه به نمودار شکل ۳-۶ ملاحظه می‌شود که هیچ نقطه خارج از کنترلی وجود ندارد. بنابراین تغییری در واریانس داده‌ها رخ نداده است. برای اطمینان از وجود تغییر، مجموع تجمعی D_i ها با استفاده از روش ارایه‌شده در بخش ۳-۲ محاسبه شده و نمودار آن در شکل ۳-۷ رسم گردیده است.



شکل ۳-۷: نمودار مجموع تجمعی D_i ها

نمودار رسم شده، نشان‌دهنده عدم تغییر در واریانس داده‌ها است.

۳-۱-۴ مزایای تحلیل نقطه تغییر تیلور نسبت به نمودار کنترل

با توجه به مطالب بیان شده در مورد تشخیص نقطه تغییر با استفاده از دو روش نمودار کنترل و روش مجموع تجمعی مبتنی بر نمونه‌های خودگردان بیان شده توسط تیلور ملاحظه شد که این این دو روش دارای مزایا و معایبی نسبت به یگدیگر هستند که در ادامه بیان می‌شوند.

۱- در تشخیص نقاط تغییر کوچک مفیدتر است.

۲- تغییرات چندگانه را بهتر مشخص کرده و یک سطح و فاصله اطمینان برای زمان‌های تغییر فراهم می‌کند.

۳- تعداد اعلام خطرهای اشتباه را کاهش داده و با اطمینان بالاتر تغییرات را مشخص می‌کند.

۴- تحلیل نسبت به نقاط پرت انعطاف‌پذیر بوده و با اجرای آن بر روی رتبه‌ی داده‌ها، تحلیل قوی‌تر می‌شود.

۵- نسبت به نوع داده‌ها انعطاف‌پذیر است و می‌توان همین روند را برای انواع داده‌ها (داده‌های صفت، مقدارهای منفرد، شمارشی و غیره) به‌کار برد.

۶- برای داده‌های بزرگ و تغییرات چندگانه دارای تفسیر ساده‌تری بوده و استفاده از آن آسان‌تر است.

این تحلیل علاوه بر مزایای گفته‌شده، دارای دو عیب زیر است:

۱- این تحلیل نمی‌تواند نقاط غیرطبیعی را جدا کند و باید با یک نمودار کنترل شوهارت تکمیل شود.

۲- روش خودگردان ساز در هر اجرا، به دلیل تصادفی انتخاب کردن داده‌ها نتایج یکسانی را نمی‌دهد. بنابراین برای برطرف کردن این عیب باید از تعداد زیادی نمونه استفاده کرد.

۳-۲ تحلیل طیفی منفرد

تحلیل طیفی منفرد که از این پس با نماد اختصاری SSA نمایش داده می‌شود، روشی مفید در تحلیل سری‌های زمانی است که در علوم مختلفی مانند ریاضیات، فیزیک، اقتصاد و همچنین در ریاضیات مالی، اقیانوس‌شناسی، علوم اجتماعی و تحقیقات علمی در بازار داد و ستد کاربرد دارد. روش SSA در تحلیل سری‌های زمانی بسامدی، آمار چندمتغیره[1]، هندسه چندمتغیره[2]، سیستم‌های پویا و پردازش سیگنال نیز به‌کار می‌رود. این روش از دهه ۸۰ با توجه به تحقیقات بروم هید و کینگ[3]، (۱۹۸۶؛ ۱۹۸۷) شناخته شد و وائوتارد و همکاران (۱۹۹۲) آن را برای تحلیل سری‌های کوتاه و بی‌نظم به‌کار برد. گولیاندینا و همکاران، (۲۰۰۱) کتابی در شرح کامل این روش با حل مثال‌های واقعی نوشته و گولیاندینا، نکروتکین و برولوو[4](۲۰۰۰) نرم‌افزار Catterpiller را برای کاربرد این روش روی داده‌های سری زمانی، ارایه نمودند. ژیگل جاوسکی و موسکوینا (۲۰۰۳)، نیز الگوریتمی را برای تشخیص نقطه تغییر توسط روش SSA ارایه نمودند که توسط حسنی (۲۰۰۷) برای تحلیل انواع داده‌های واقعی استفاده شده است.

در تحلیل داده‌ها همواره خطا باعث عدم قطعیت در نتایج به دست‌آمده می‌شود. بنابراین حذف خطا یا کم کردن اثر خطا، موضوعی حایز اهمیت است. کار اصلی SSA، تجزیه مقدار منفرد[5] (SVD) ماتریس مسیری[6] است که از سری زمانی اصلی با بازسازی زیردنباله‌های سری به دست می‌آید. به این ترتیب، این روش با تجزیه سری به سری‌های مختلف و جداسازی خطا، تحلیل نسبتا دقیق‌تری از داده‌ها فراهم می‌کند. الگوریتم این روش، شامل چهار مرحله تعبیه (جاسازی کردن)[7]، تجزیه مقدار

---

[1] Multivariate statistics
[2] Multivariate geometry
[3] Broomhead and King
[4] Goljandina and Nekruktin and Braulov
[5] Singular value decomposition
[6] Trajectory matrix
[7] Embedding

منفرد ماتریس مسیر (SVD)، گروه‌بندی ماتریس‌ها[1] و بازسازی سری[2] است که در زیربخش ۳-۲-۱ توضیح داده شده است (برای توضیحات بیشتر به کتاب گولیاندینا و ژیگل جاوسکی (۲۰۰۱) مراجعه کنید). کاربرد این روش برای تشخیص نقطه تغییر در بخش ۳-۲-۲ بیان شده است. برای تشخیص نقطه تغییر، ابتدا دو ماتریس مسیر پایه و آزمون ساخته شده و سپس مراحل کلی روش SSA بر روی ماتریس پایه، اجرا می‌شود. سپس سری خطا را جدا کرده و فاصله اقلیدسی بین زیرفضای سازنده سری پایه، بدون وجود خطا و ماتریس آزمون را محاسبه کرده و به عنوان آماره تشخیص نقطه تغییر در نظر می‌گیرند. طبق مطالب گفته‌شده، برای به‌کارگیری روش SSA جهت تشخیص نقطه تغییر، ابتدا باید این روش را در حالت کلی مورد بررسی قرار داد.

۳-۲-۱ SSA در حالت کلی

همان‌طور که گفته شد این روش شامل چهار مرحله است که در زیر توضیح داده می‌شود.

۳-۲-۱-۱ تعبیه یا جاسازی کردن

تعبیه یا جاسازی کردن، در واقع ساختن ماتریس مسیر با توجه به داده‌های سری است. فرض کنید $x_1, x_2, \dots, x_N$ سری زمانی به طول N باشد. پارامتر تاخیر $M \leq \frac{N}{2}$ را که مقداری صحیح است و همچنین K بردار تأخیر X_j که به صورت

$$X_j = (x_j, \dots, x_{j+M-1})^T, \qquad j = 1,2,\dots,K \tag{۳-۲}$$

معرفی می‌شوند را در نظر می‌گیریم که در آن $K = N - M + 1$. با استفاده از K بردار تأخیر M–بعدی X_j، ماتریس مسیر $(X_{M\times K})$ به صورت زیر ساخته می‌شود.

---

[1] Grouping
[2] Restructuring

$$X_{M\times K}=[X_1,\dots,X_K]=\begin{pmatrix} x_1 & x_2 & x_3 & \dots x_K \\ x_2 & x_3 & x_4 & \dots x_{K+1} \\ & . & & \\ & . & & \\ x_M & x_{M+1} & x_{M+2} & \dots x_N \end{pmatrix},$$

ماتریس مسیر X یک *ماتریس هنکل*[1] است، زیرا تمام عناصر آن در امتداد قطرهای فرعی با هم برابر هستند. ساختن ماتریس مسیر، فرآیند "هنکل‌سازی[2]" نیز نامیده می‌شود. در زیر با ارایه مثالی، ساخت ماتریس مسیر را برای یک سری به طول ۸ بیان می‌کنیم.

مثال ۳-۲:

سری زمانی $x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8$ را در نظر بگیرید. داریم

$$N=8 \qquad M=\frac{8}{2}=4 \qquad K=8-4+1=5$$

بنابراین ۵ بردار تاخیر ۴-بعدی X_j با توجه به رابطه (۳-۲) به صورت

$$X_1=[x_1,x_2,x_3,x_4,x_5],$$
$$X_2=[x_2,x_3,x_4,x_5,x_6],$$
$$X_3=[x_3,x_4,x_5,x_6,x_7],$$
$$X_4=[x_4,x_5,x_6,x_7,x_8],$$

ساخته می‌شوند و ماتریس مسیر به صورت زیر به دست می‌آید:

$$X=H=\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ x_2 & x_3 & x_4 & x_5 & x_6 \\ x_3 & x_4 & x_5 & x_6 & x_7 \\ x_4 & x_5 & x_6 & x_7 & x_8 \end{bmatrix}.$$

با تغییر پارامتر تأخیر M، مقدار K که تعداد ستون‌های ماتریس مسیر است، تغییر خواهد کرد. بنابراین این سوال مطرح خواهد شد که M چه مقدار باشد؟ در پاسخ باید گفت که برای سری‌های زمانی کوتاه، به طور قراردادی مقدار $M\leq\frac{N}{2}$ و در صورت طولانی بودن سری، قسمتی از آن به طول

---

[1] Hankel
[2] Hankelization

m به عنوان زیرسری از سری اصلی (عرض پنجره) انتخاب شده و پارامتر تأخیر M با شرط $M \leq \frac{m}{2}$ در نظر گرفته می‌شود. شرح کامل انتخاب مقدار M در بخش ۳-۲-۳ آمده است (گولیاندینا و همکاران، ۲۰۰۱).

۳-۲-۱-۲ تجزیه مقدار منفرد (*SVD*) ماتریس مسیر

مقدار منفرد در واقع همان جذر مقدار ویژه است و تجزیه مقدار منفرد (SVD) به معنی به‌دست آوردن مقادیر و بردارهای ویژه از ماتریس کواریانس تأخیر[1] است که این ماتریس از حاصل‌ضرب ماتریس مسیر در ترانهاده‌اش به‌صورت

$$S_{M\times M} = H\, H^T,$$

محاسبه می‌شود. مقادیر ویژه ماتریس کواریانس تاخیر S یعنی $\lambda_1, \lambda_2, \ldots, \lambda_M$ از حل معادله $|S - \lambda I| = ۰$ به دست می‌آیند که در آن I ماتریسی همانی، هم‌رتبه ماتریس کواریانس S است. همچنین بردارهای ویژه ماتریس S یعنی $U_1, U_2, \ldots, U_M$ از حل معادله $(S - \lambda I)U = ۰$ محاسبه می‌شوند. با توجه به M مقدار ویژه λ_i و M بردار ویژه U_i به دست‌آمده از ماتریس کواریانس تأخیر، *بردار مؤلفه‌های اصلی*[2] به ازای هر $i = 1, \ldots, M$ توسط رابطه زیر محاسبه می‌شوند:

$$V_i = H^T U_i / \sqrt{\lambda_i}\ ,$$

که مقادیر $\sqrt{\lambda_i}$ ها را مقادیر منفرد ماتریس مسیر H نامند. البته این نکته قابل ذکر است که ممکن است برخی از مقادیر ویژه λ_i برابر صفر باشند. در این صورت به ازای d مقدار ویژه‌ی مخالف صفر λ_i و با استفاده از بردارهای ویژه متناظر و بردار مؤلفه‌های اصلی به دست آمده از مرحله قبل، ماتریس‌های مسیر به ازای هر $i = 1,2, \ldots, d$ با توجه به رابطه زیر، بازسازی می‌شوند:

---

[1] Lag-covariance matrix
[2] Principal component

$$H_i = \sqrt{\lambda_i}\, U_i\, V_i^T. \qquad (3-3)$$

مجموع ماتریس‌های به دست آمده‌ی H_i در واقع تقریبی از ماتریس مسیر اولیه H هستند، یعنی $H \simeq H_1 + \cdots + H_d$. بنابراین با استفاده از سه تایی $(\sqrt{\lambda_i}, U_i, V_i)$، می‌توان تقریبی از ماتریس مسیر اولیه را به‌دست آورد. این سه مؤلفه، *سه گانه*[1] i ام ماتریس H نامیده می‌شوند. در مثال 3-3 این سه مؤلفه برای سری به طول ۶ با استفاده از روابط بالا ساخته شده‌اند.

مثال ۳-۳: با توجه به سری ۶،۵،۷،۱۰،۴،۳، داریم

$$H = \begin{bmatrix} 6 & 5 & 7 & 10 \\ 5 & 7 & 10 & 4 \\ 7 & 10 & 4 & 3 \end{bmatrix}, \qquad S = \begin{bmatrix} 210 & 175 & 150 \\ 175 & 190 & 157 \\ 150 & 157 & 174 \end{bmatrix},$$

$$\lambda_1 = 22/668 \qquad \lambda_2 = 6/411 \qquad \lambda_3 = 4/387 \qquad \lambda_4 = ۰$$

$$U = (u_1, u_2, u_3\,) = \begin{bmatrix} 0/۶۰۴۲ & -0/۶۸۵۲ & 0/۴۰۶۷ \\ 0/۵۸۷۶ & 0/۰۳۸۴ & -0/۸۰۸۲ \\ 0/۵۳۸۲ & 0/۷۲۷۳ & 0/۴۲۵۸ \end{bmatrix},$$

$$V = (v_1, v_2, v_3\,) = \begin{bmatrix} 0/۱۳۴۴ & 0/۱۰۱۱ & 0/۰۴۳۹ & 0/۱۹۹۵ \\ 0/۳۰۳۲ & 0/۷۶۱۸ & 0/۱۹۷۵ & 0/۵۶۳۱ \\ 2/۲۴۴۸ & 2/۷۴۴۹ & 2/۹۰۵۳ & 2/۱۸۱۵ \end{bmatrix}, .$$

در ادامه بحث، با استفاده از مقادیر و بردارهای ویژه در این بخش، ماتریس‌های مسیر متناظر با هر مقدار ویژه، با توجه به تاثیرپذیری این مقادیر گروه‌بندی می‌شوند.

---

[1] Eigentriple

۳-۲-۱-۳ گروه‌بندی ماتریس‌ها

در این مرحله، برای جداسازی خطا، ماتریس‌های متناظر با d مقدار ویژه مخالف صفر که با مجموعه اندیس $D = \{1, \dots, d\}$ نشان داده می‌شوند، به دو گروه تقسیم‌بندی خواهند شد که سری بازسازی‌شده گروه اول، نشان‌دهنده سری اصلی داده‌ها بدون وجود خطا و سری بازسازی‌شده‌ی گروه دوم، القاکننده سری خطا است. در حالت کلی مجموعه اندیس D ممکن است به m گروه مجزای $I_1, I_2, \dots, I_m$ تقسیم‌بندی شود. تعداد این تقسیم‌بندی‌ها وابسته به کاربردهای خاص تحلیل سری (به‌طور مثال جداسازی خطا، روند، دوره و غیره) می‌باشد.

تقسیم‌بندی به این صورت انجام خواهد شد که ابتدا مقادیر ویژه وزن‌دهی می‌شوند، به طوری‌که به هر مقدار ویژه وزن $w_i = \frac{\lambda_i}{\sum_{i=1}^{d} \lambda_i}$ داده می‌شود. ماتریس‌های ساخته‌شده بر اساس مقادیر ویژه‌ای که دارای بیشترین وزن هستند (به‌طور مثال ۹۵٪ که این درصد به تحلیل‌گر بستگی دارد)، در گروهی قرار می‌گیرند که برای بازسازی سری اصلی به‌کار می‌رود. ماتریس‌هایی که از بقیه مقادیر ویژه تولید می‌شوند در گروهی قرار خواهند گرفت که سری خطا را بازسازی می‌کنند. در مثال ۳-۳، وزن مقادیر ویژه عبارت است از:

$w_1 = 0/۸۹۵۱$, $w_2 = 0/۰۷۱۴$, $w_3 = 0/۰۳۳۵$.

بنابراین با توجه به وزن‌های به دست آمده، می‌توان گفت که 89/51٪ تاثیرپذیری سری، مربوط به مولفه اصلی اول است که از مقدار ویژه اول تولید می‌شود، و تاثیر پذیری سری، از مولفه دوم و سوم معادل ۱۰/۴۹٪ می‌باشد. بنابراین می‌توان ماتریس تولیدشده از مولفه اصلی اول را برای بازسازی سری اصلی و ماتریس‌های تولیدشده از مولفه‌های اصلی دوم و سوم را برای بازسازی سری خطا استفاده نمود. ماتریس‌های متناظر هر مولفه، با استفاده از رابطه (۳-۳) در زیر محاسبه شده است.

$$H_1 = \begin{bmatrix} 1/\text{۸۴۰۴} & 1/\text{۳۸۴۵} & -0/\text{۶۰۱۸} & 2/\text{۷۳۲۰} \\ -2/\text{۰۸۷۱} & -1/\text{۵۷۰۰} & 0/\text{۶۸۲۴} & -3/\text{۰۹۸۲} \\ 1/\text{۲۳۸۷} & 0/\text{۹۳۱۸} & -0/\text{۴۰۵۰} & 1/\text{۸۳۹} \end{bmatrix}$$

$$H_2 = \begin{bmatrix} -1/\text{۱۴۰۲} & -2/\text{۸۶۵} & 0/\text{۷۴۲۶} & 2/\text{۱۱۷۸} \\ -0/\text{۰۷۴۵} & -0/\text{۱۸۷۳} & 0/\text{۰۴۸۵۵} & 0/\text{۱۳۸ ۴} \\ 1/\text{۸۳۹} & 3/\text{۹۴۱۲} & -1/\text{۰۲۱} & -2/\text{۹۱۳۳} \end{bmatrix}$$

$$H_3 = \begin{bmatrix} 5/\text{۳۰۰} & 6/\text{۴۸۱} & 6/\text{۸۵۹} & 5/\text{۱۵۰} \\ 7/\text{۱۶۲} & 8/\text{۷۵۷} & 9/\text{۲۶۹} & 6/\text{۹۶۰} \\ 4/\text{۱۹۳} & 5/\text{۱۲۷} & 5/\text{۴۲۷} & 4/\text{۰۷۵} \end{bmatrix}$$

بنابراین می‌توان ماتریس H را به صورت مجموع $H = H_1 + (H_2 + H_3)$ نوشت. به این ترتیب اگر در حالت کلی مقادیر ویژه به دو گروه تقسیم‌بندی شوند و فرض کنید $I_1 = \{i_1, i_2, \dots, i_l\}$ گروه اول از مجموعه شاخص D باشد (l تعداد مقادیر ویژه با بیشترین تاثیر هستند)، ماتریس متناظر با این گروه برابر است با $H_{I_1} = H_{i_1} + H_{i_2} + \cdots + H_{i_l}$. بنابراین اگر مجموعه شاخص به m گروه تقسیم شود، آن‌گاه ماتریس متناظر با ماتریس مسیر H برابر مجموع ماتریس تمام گروه‌ها است یعنی $H = H_{I_1} + H_{I_2} + \cdots + H_{I_m}$ که این ماتریس تقریبی از ماتریس مسیر اولیه است.

۳-۲-۱-۴ بازسازی سری

همان‌طور که قبلا ذکر شد، فرآیند ساخت ماتریس مسیر از سری داده‌ها را، فرآیند هنکل‌سازی گویند. در این‌جا می‌خواهیم با استفاده از ماتریس‌های مسیر گروه‌بندی‌شده، سری را بازسازی کنیم که به این عمل، فرآیند "هنکل‌سازی معکوس[1]" گویند. این فرآیند شامل عمل میانگین‌گیری قطری[2] روی عناصر ماتریس‌های گروه‌بندی‌شده مرحله قبل است که با استفاده از رابطه (۳-۴) انجام می‌شود (موسکوینا و همکاران، ۲۰۰۱، بولونگا، ۲۰۱۲):

---

[1] Inverse hankelization

[2] Diagonal averaging

$$\tilde{X}_s=\begin{cases}\dfrac{1}{s-1}\displaystyle\sum_{j=1}^{s-1}h_{j,s-j}\,, & 2\le s\le M\,,\\ \dfrac{1}{M}\displaystyle\sum_{j=1}^{N}h_{j,s-j}, & M+1\le s\le K+1,\\ \dfrac{1}{N-s+2}\displaystyle\sum_{j=s-K}^{N-s+2}h_{j,s-j}\,, & K+2\le s\le N+1\,,\end{cases}\qquad (4-۳)$$

به طوری که به ازای هر گروه I یک سری با همان طول سری اولیه بازسازی شده و از مجموع این سری‌ها، تقریبی از سری اولیه به دست می‌آید. عمل میانگین‌گیری قطری و بازسازی سری برای داده‌های مثال ۳-۳ در زیر نمایش داده شده است.

$$H_{I_1}=H_1=\begin{bmatrix}1٫۸۴۰۴ & 1٫۳۸۴۵ & -0٫۶۰۱۸ & 2٫۷۳۲۰\\ -2٫۰۸۷۱ & -1٫۵۷۰۰ & 0٫۶۸۲۴ & -3٫۰۹۸۲\\ 1٫۲۳۸۷ & 0٫۹۳۱۸ & -0٫۴۰۵۰ & 1٫۸۳۹\end{bmatrix}$$

$\tilde{X}_2=h_{11}=1٫۸۴۰۴$ $\qquad$ $\tilde{X}_3=\frac{(h_{12}+h_{21})}{2}=\frac{۱}{۲}(1٫۳۸۴۵-2٫۰۸۷۱)=-0٫۳۵۱۳$

$$\tilde{X}_4=\frac{(h_{31}+h_{22}+h_{13})}{3}=\frac{۱}{۳}(-0٫۶۰۱۸-1٫۵۷۰۰+1٫۲۳۸۷)=-0٫۳۱۱$$

$$\tilde{X}_5=\frac{(h_{32}+h_{23}+h_{14})}{3}=\frac{۱}{۳}(2٫۷۳۲۰+0٫۶۸۲۴+0٫۹۳۱۸)=1٫۴۴۸۷$$

$\tilde{X}_6=\frac{(h_{33}+h_{24})}{2}=\frac{۱}{۲}(-3٫۰۹۸۲-0٫۴۰۵۰)=-1٫۷۵۱۶$ $\qquad$ $\tilde{X}_7=h_{34}=1٫۸۳۹$

در نتیجه سری ساخته‌شده با نام جدید z_t به صورت زیر می‌باشد:

$z_1=\tilde{X}_2=1٫۸۴۰۴,\qquad z_2=-0٫۳۵۱۳,\qquad z_3=-0٫۳۱۱,\qquad z_4=1٫۴۴۸۷,$

$z_5=-1٫۷۵۱۶\,,\qquad z_6=1٫۸۳۹.$

حال با استفاده از ماتریس گروه دوم که از مجموع دو ماتریس H_2 و H_3 تولید می‌شود ($H_{I_2}=H_2+H_3$)، سری خطا بازسازی می‌شود.

$$H_{I_2} = H_2 + H_3 = \begin{bmatrix} 4/۱۵۹ & 3/۶۱۶ & 7/۶۰۲ & 7/۲۶۸ \\ 7/۰۸۸ & 8/۵۷۰ & 9/۳۱۸ & 7/۰۹۸ \\ 6/۰۳۲ & 9/۰۶۸ & 4/۴۰۶ & 1/۱۶۲ \end{bmatrix}$$

$$\tilde{X}_2 = h_{11} = 4/۱۵۹ \qquad \tilde{X}_3 = \frac{(h_{12}+h_{21})}{2} = \frac{۱}{۲}(3/۶۱۶ + 7/۰۸۸) = 5/۳۵۲$$

$$\tilde{X}_4 = \frac{(h_{31}+h_{22}+h_{13})}{3} = \frac{۱}{۳}(6/۰۳۲ + 8/۵۷۰ + 7/۶۰۲) = 7/۴۰۱$$

$$\tilde{X}_5 = \frac{(h_{32}+h_{23}+h_{14})}{3} = \frac{۱}{۳}(9/۰۶۸ + 9/۳۱۸ + 7/۲۶۸) = 8/۵۵۱$$

$$\tilde{X}_6 = \frac{(h_{33}+h_{24})}{2} = \frac{۱}{۲}(4/۴۰۶ + 7/۰۹۸) = 5/۷۵۲ \qquad \tilde{X}_7 = h_{34} = 1/۱۶۲$$

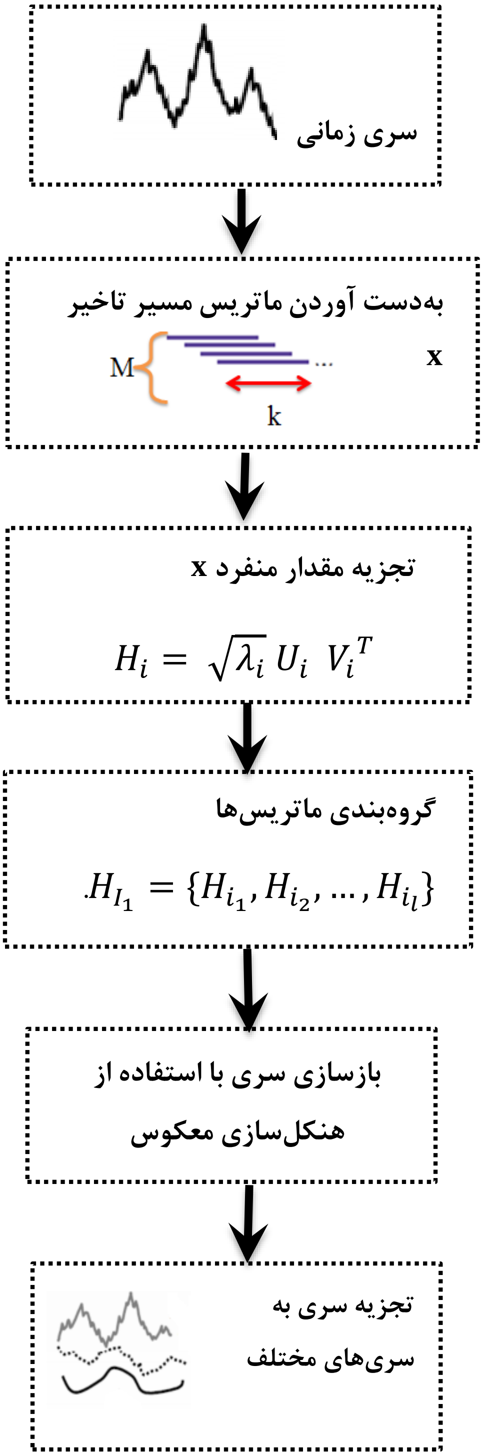
در نتیجه سری ساخته شده با نام جدید e_t به صورت زیر می‌باشد:

$$e_1 = \tilde{X}_2 = 4/۱۵۹, \quad e_2 = 5/۳۵۲, \quad e_3 = 7/۴۰۱\,, \quad e_4 = 8/۵۵۱,$$
$$e_5 = 5/۷۵۲ \quad e_6 = 1/۱۶۲.$$

بنابراین سری x_t به دو سری مجزای z_t و e_t جداسازی می‌شود. در واقع مجموع این دو سری ساخته‌شده تقریبی از سری اولیه است یعنی $x_t \simeq z_t + e_t$. به طور کلی مجموع سری‌های بازسازی شده متناظر با ماتریس‌های مسیر m گروه، تقریبی از سری اولیه x_t را تولید می‌کند، یعنی

$$x_t \simeq z_{I_1} + z_{I_2} + \cdots + z_{I_m}\,.$$

ملاحظه می‌شود که در روش SSA انتخاب پارامترهای تأخیر M و گروه I اهمیت بسیاری در تجزیه سری دارد که در بخش ۳-۲-۳ مورد بررسی قرار می‌گیرد. الگوریتم این روش در شکل ۳-۸ نشان داده شده است و با استفاده از آن‌چه در بخش بعد بیان خواهد شد، می‌توان این روش را برای تشخیص نقطه تغییر به‌کار برد.



شکل ۳-۸: ساختار اصلی الگوریتم SSA

۳-۲-2 الگوریتم SSA برای تشخیص نقطه تغییر

الگوریتم SSA را می‌توان با اعمال تغییراتی، برای تشخیص نقطه تغییر به شرح زیر به‌کار برد. فرض کنید $x_1, x_2, \dots, x_N$ سری زمانی به طول N باشد. پارامترهای M و K برای ساخت ماتریس مسیر همانند قبل انتخاب می‌شوند. در این‌جا ماتریس مسیر ساخته‌شده را *ماتریس پایه*[1] نامند. تنها تفاوت در این‌جا است که در الگوریتم اصلی SSA تنها به یک ماتریس مسیر نیاز داریم که روی فاصله زمانی $[1\,,N]$ ساخته می‌شود، ولی برای تشخیص نقطه تغییر باید ماتریس‌های مسیر به ازای هر n $(\mathrm{n} = 0,1,\dots,\mathrm{N}-\mathrm{m}-\mathrm{M})$ با استفاده از رابطه (۳-۲)، روی فواصل زمانی $[n+1\,,n+m]$ به صورت

$$X_B^{(n)} = H_B^{(n)} = \begin{pmatrix} x_{n+1} & x_{n+2} & \cdots & x_{n+K} \\ x_{n+2} & x_{n+3} & \cdots & x_{n+K+1} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{n+M} & x_{n+M+1} & \cdots & x_{n+m} \end{pmatrix},$$

ساخته شوند. در واقع ستون‌های این ماتریس به‌صورت

$$X_j^{(n)} = (x_{n+j}, \dots, x_{n+j+M-1})^T\,, \qquad\qquad j = 1, \dots, K,$$

می‌باشند. تحلیل SSA علاوه بر ماتریس پایه، نیازمند ماتریسی برای تشخیص نقطه تغییر است که *ماتریس آزمون*[2] نامیده می‌شود. این ماتریس همانند ماتریس پایه در فواصل $[n+p+1\,,n+q+M-1]$ که n $(\mathrm{n} = 0,1,\dots,\mathrm{N}-\mathrm{m}-\mathrm{M})$به صورت زیر ساخته می‌شود.

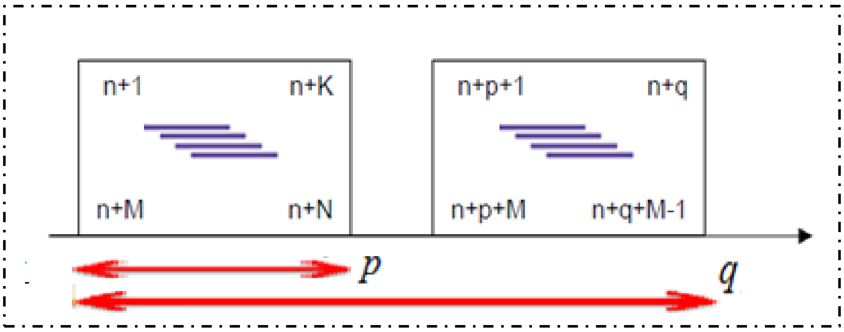
$$\mathrm{x}_T^{(n)} = \mathrm{H}_T^{(n)} = \begin{pmatrix} x_{n+p+1} & x_{n+p+2} & \cdots & x_{n+q} \\ x_{n+p+2} & x_{n+p+3} & \cdots & x_{n+q+1} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{n+p+M} & x_{n+M+p+1} & \cdots & x_{n+q+M-1} \end{pmatrix}.$$

---

[1] Base matrix

[2] Test matrix

در این‌جا پارامتر p، عرض پنجره پایه (طول ماتریس پایه) و پارامتر q، عرض پنجره پایه و آزمون است. ساختار ماتریس اصلی و آزمون در شکل ۳-۹ نشان داده شده است (وایسمن، ۲۰۰۸؛ ۲۰۱۰).



شکل ۳-۹: ساختار ماتریس مسیر و آزمون

در این‌جا این سوال مطرح خواهد شد که مقدار p و q چه اندازه باید انتخاب شوند؟ در جواب به این سوال باید گفت که پیشنهاد عمومی، انتخاب $p \geq k$ است که باعث متفاوت بودن ستون‌های ماتریس پایه و آزمون می‌شود. اگر $p \geq N = M + K - 1$ باشد، آن‌گاه ماتریس‌های پایه و آزمون دارای عناصر مشترک خواهند شد. باید در نظر داشت که اختلاف این دو پارامتر نیز باید $1 \leq q - p \leq M$ باشد (ژیگل جاوسکی و موسکوینا، ۲۰۰۳). اگر این اختلاف بیش از حد زیاد باشد، ممکن است بعضی از تغییرات تشخیص داده نشوند و اگر بیش از حد کوچک باشد ممکن است نقطه‌ای که دارای تغییر نیست به عنوان نقطه تغییر تشخیص داده شود. اگر سری را با طول $N = \infty$ فرض کنیم، آن‌گاه نمایش ماتریس پایه و آزمون روی زیرسری از این سری با تنظیمات زیر به وجود می‌آید:

$$x_1, x_2, \dots, x_{100} \qquad m = 100 \qquad M \leq \frac{100}{2} = 50 \qquad \begin{cases} K = 51 \\ p \geq k \geq 51 \\ 1 \leq q - p \leq M. \end{cases}$$

برای مثال با $n = 0$ و انتخاب‌های متفاوت پارامترهای K, p, q داریم:

$$\begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_{51} \\ x_2 & x_3 & x_4 & \dots & x_{52} \\ & & \cdot & & \\ & & \cdot & & \\ x_{50} & x_{52} & x_{53} & \dots & x_{100} \end{bmatrix}, \quad \begin{bmatrix} x_{52} \\ x_{53} \\ \cdot \\ \cdot \\ x_{101} \end{bmatrix}$$

$$\begin{cases} K = 51 \\ p = 51 \\ q = 52 \\ q - p = 1 \end{cases}$$

$$\begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_{51} \\ x_2 & x_3 & x_4 & \dots & x_{52} \\ & & \cdot & & \\ & & \cdot & & \\ x_{50} & x_{52} & x_{53} & \dots & x_{100} \end{bmatrix}, \quad \begin{bmatrix} x_{60} & x_{61} & x_{62} & \dots & x_{71} \\ x_{61} & x_{62} & x_{63} & \dots & x_{72} \\ & & \cdot & & \\ & & \cdot & & \\ x_{110} & x_{111} & x_{112} & \dots & x_{121} \end{bmatrix}$$

$$\begin{cases} K = 51 \\ p = 51 \\ q = 71 \\ q - p = 20 \end{cases}$$

$$\begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_{51} \\ x_2 & x_3 & x_4 & \dots & x_{52} \\ & & \cdot & & \\ & & \cdot & & \\ x_{50} & x_{52} & x_{53} & \dots & x_{100} \end{bmatrix}, \quad \begin{bmatrix} x_{52} & x_{53} & x_{54} & \dots & x_{101} \\ x_{53} & x_{54} & x_{55} & \dots & x_{102} \\ & & \cdot & & \\ & & \cdot & & \\ x_{101} & x_{102} & x_{103} & \dots & x_{151} \end{bmatrix}$$

$$\begin{cases} K = 51 \\ p = 51 \\ q = 101 \\ q - p = 50 \end{cases}$$

p

q

توضیحات بیشتر در مورد انتخاب این دو پارامتر در بخش ۳-۲-۳ بیان شده‌اند. بعد از ساخت ماتریس مسیر پایه و آزمون، تحلیل SSA برای تشخیص نقطه تغییر طبق الگوریتم زیر انجام می‌شوند:

۱- همانند الگوریتم اصلی SSA، ماتریس کوواریانس تأخیر به ازای هر $n = 0,1,\dots,N-m-M,$ محاسبه شود، یعنی $S_n = H^{(n)}(H^{(n)})^T$.

۲- برای ماتریس کوواریانس تأخیر S_n، با استفاده از SVD، M مقدار و بردار ویژه محاسبه گردد.

۳- مقادیر ویژه، وزن‌دهی‌شده و مؤلفه‌های اصلی متناظر با مقدارهای ویژه مخالف صفر که با درصد بالاتری بیان‌کننده سری هستند، محاسبه شوند.

۴- مؤلفه‌ها گروه‌بندی شوند: گروه اول متناظر با l مؤلفه‌ای که دارای موثرترین مقادیر ویژه بوده و بیان‌کننده سری واقعی داده‌ها است و تشکیل یک زیرفضای l-بعدی $\mathcal{L}_{n,I}$ را خواهند داد و گروه دوم شامل بقیه مؤلفه‌ها است که بیان‌کننده سری خطا می‌باشند. همانند قبل، سری‌ها بازسازی می‌شوند. حال سوال این‌جا است که نقاط تغییر را چطور محاسبه کنیم؟ به این سوال در بخش بعد پاسخ داده می‌شود.

۳-۲-۲-۱ محاسبه آماره تغییر

همان‌طور که در مرحله ۴ بخش ۳-۲-۱ گفته شد، بعد از گروه‌بندی مقادیر ویژه، l مؤلفه‌ای که دارای موثرترین مقادیر ویژه هستند تشکیل یک زیرفضای l-بعدی $\mathcal{L}_{n,I}$ می‌دهند. اگر آماره $D_{n,I,p,q}$ را به صورت مجموع توان دوم فاصله اقلیدسی[1] بین زیرفضای $\mathcal{L}_{n,I}$ و ماتریس آزمون تعریف کنیم، داریم

$$D_{n,I,p,q} = \sum_{j=p+1}^{q} ((X_j)^T X_j - (X_j)^T U U^T X_j)^2. \qquad (۳-۳)$$

آماره $D_{n,I,p,q}$ را آماره تشخیص تغییر می‌نامند که در آن X_j $(j = 1,2,\dots,K)$ ستون‌های ماتریس آزمون است. آماره تغییر $D_{n,I,p,q}$ به ازای هر n محاسبه می‌شود و این مقدار با تغییر ماتریس‌های پایه و آزمون تغییر می‌کند، به طوری‌که به ازای هر n مشخص، ابتدا ماتریس‌های مسیر ساخته‌شده و سپس مقدار آماره آزمون $D_{n,I,p,q}$ محاسبه می‌شود. آماره متناظر چند نمونه از n، برای سری به طول ۱۰۰ در زیر محاسبه شده‌اند:

$$x_1, x_2, x_3, x_4, x_5, x_6, \dots, x_{51}, x_{52}, x_{53}, x_{54}, \dots, x_{100}$$

---

[1] Sum of squared euclidean distances

$$\begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_{51} \\ x_2 & x_3 & x_4 & \dots & x_{52} \\ & & . & & \\ & & . & & \\ x_{50} & x_{51} & x_{52} & \dots & x_{100} \end{bmatrix}, \quad \begin{bmatrix} x_{52} \\ x_{53} \\ . \\ . \\ x_{101} \end{bmatrix} \qquad n = 0 \qquad D_{0,I,p,q}$$

$$\begin{bmatrix} x_2 & x_3 & x_4 & \dots & x_{52} \\ x_3 & x_4 & x_5 & \dots & x_{53} \\ & & . & & \\ & & . & & \\ x_{51} & x_{52} & x_{53} & \dots & x_{101} \end{bmatrix}, \quad \begin{bmatrix} x_{52} \\ x_{53} \\ . \\ . \\ x_{101} \end{bmatrix} \qquad n = 1 \qquad D_{1,I,p,q}$$

$$\begin{bmatrix} x_3 & x_4 & x_5 & \dots & x_{53} \\ x_4 & x_5 & x_6 & \dots & x_{54} \\ & & . & & \\ & & . & & \\ x_{52} & x_{53} & x_{54} & \dots & x_{102} \end{bmatrix}, \quad \begin{bmatrix} x_{52} \\ x_{53} \\ . \\ . \\ x_{101} \end{bmatrix} \qquad n = 2 \qquad D_{2,I,p,q}$$

بنابراین ماتریس مسیر به ازای هر n یک حرکت به جلو داشته و به ازای ماتریس مسیر جدید، آماره $D_{n,I,p,q}$ محاسبه شده و تشکیل دنباله‌ی $D_{1,I,p,q}, D_{2,I,p,q}, \dots, D_{\mathrm{N-m-M},I,p,q}$ را خواهد داد. نقطه‌ای که آماره $D_{n,I,p,q}$ شروع به افزایش کند، همان نقطه تغییر است. به عبارت دیگر افزایش در نقطه‌ای مانند τ به دلیل تغییری است که در مکانیسم تولید x_t، در نقطه τ روی می‌دهد که در این-صورت، فاصله‌ی بین زیرفضای l-بعدی $\mathcal{L}_{n,I}$ و بردارهای X_j به ازای $j > \tau$، بیشتر از این فاصله به ازای $j \leq \tau$ می‌باشد.

در این‌جا برای تشخیص نقطه تغییر، شاخصی را به نام *شاخص ناهمگنی*[1] (عدم تجانس) تعریف می‌کنیم که با استفاده از آماره تغییر $D_{n,I,p,q}$ به صورت

$$g(z_t, x_t) = \frac{\sum_{j=p+1}^{q} \left((X_j)^T X_j - (X_j)^T U U^T X_j\right)^2}{\sum_{j=p+1}^{q} ||(X_j)^T X_j||^2} = \frac{\sum_{j=p+1}^{q} \left(D_{n,I,p,q}\right)^2}{\sum_{j=p+1}^{q} ||(X_j)^T X_j||^2} \;, \qquad (۳ - ۴)$$

---

[1] Heterogeneity index

بیان می‌شود. مقدار g متعلق به بازه $[0,1]$ است. اگر همه بردارهای تاخیر ماتریس آزمون در زیر فضای $\mathcal{L}_{n,I}$ قرار گیرند، آن‌گاه $g(z_t, e_t) = ۰$، و اگر همه بردارهای تاخیر ماتریس آزمون و زیر فضای $\mathcal{L}_{n,I}$ نسبت به هم متعامد باشند، آن‌گاه $g(z_t, e_t) = 1$ می‌شود. حال ماتریس ناهمگنی سری اولیه را تعریف می‌کنیم که عناصر این ماتریس، مقادیر شاخص‌های ناهمگنی g برای جفت‌های مختلف زیرسری‌های پایه از سری اولیه است. بنابراین مولفه‌های این ماتریس به صورت

$$g_{ij} = \left(X_{i\,,i+B-1}\colon X_{j\,,j+T-1}\right), \qquad i = 1, \dots, N-B+1, \quad j = 1, \dots, N-T+1,$$

تعیین می‌شوند، که در آن B و T را به ترتیب طول ماتریس پایه و آزمون نامند. مقادیر مختلف g_{ij}، تشکیل ماتریس ناهمگنی G را می‌دهند. این نکته قابل ذکر است که ماتریس ناهمگنی G حتی در حالتی که B = T باشد نیز متقارن نیست، زیرا زیرسری‌های پایه و آزمون نقش متفاوتی در ساخت شاخص ناهمگنی دارند. حال برای تشخیص نقطه تغییر با استفاده از این ماتریس، لازم است توابعی را روی ماتریس ناهمگنی به عنوان توابع ناهمگنی تعریف کنیم.

۱ – توابع ناهمگنی سطری

توابع ناهمگنی سطری ماتریس G، منعکس‌کننده همگنی سری اصلی است که این توابع منطبق بر سطر i ام ماتریس ناهمگنی است. در این توابع که در زیر نمایش داده می‌شوند، زیرسری‌های پایه ثابت هستند ولی ماتریس آزمون تغییر می‌کند.

$$g_{in} = \left(X_{i\,,i+B-1}\colon X_{n\,,n+T-1}\right), \qquad n = 1, \dots, N-T+1. \qquad (۵\text{-}۳)$$

۲ – توابع ناهمگنی ستونی

علاوه بر توابع ناهمگنی سطری می‌توانیم توابع ناهمگنی ستونی را نیز در ماتریس G مشاهده کنیم که منطبق بر ستون‌های این ماتریس هستند. در این مورد، زیر سری‌های آزمون ثابت بوده اما زیرسری-

های پایه تغییر می‌کنند. در واقع ستون‌های این ماتریس منعکس‌کننده همگنی زیرفضا نسبت به زیر-سری‌های آزمون هستند. این توابع در زیر نشان داده شده‌اند:

$$g_{nj} = \left(X_{n,n+B-1}: X_{j,j+T-1}\right), \qquad j = 1, \dots, N-T+1. \tag{۶-۳}$$

تغییر در مقدار توابع ناهمگنی ستونی و سطری ممکن است در دو ناحیه اتفاق بیافتد، یکی هنگامی است که فاصله آزمون "*وارد*" فاصله ناهمگنی سری پایه می‌شود و دیگری در ناحیه‌ای است که از این فاصله "*خارج*" می‌شود.

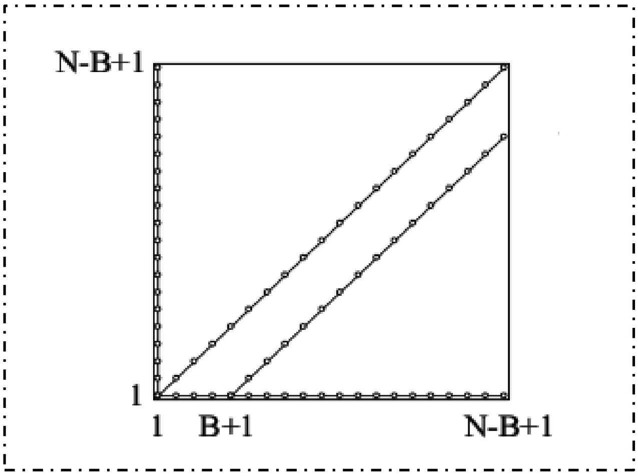
۳ – توابع ناهمگنی قطری

علاوه بر توابع ناهمگنی سطری و ستونی، توابع ناهمگنی وجود دارند که منطبق بر قطرهای ماتریس ناهمگنی هستند. در واقع قطرهای این ماتریس بیان‌کننده همگنی قطری زیرفضا و زیرسری‌های آزمون است. اگر $۰ \le \delta \le N-T+1$، آن‌گاه توابع ناهمگنی قطری به صورت

$$g_{n,n+\delta} = \left(X_{n,n+B-1}: X_{n+\delta,n+\delta+T-1}\right), \qquad n = 1, \dots, N-T+\delta+1,$$

است. توابع ناهمگنی دیگری نیز وجود دارند که برای کسب اطلاعات بیشتر در این مورد می‌توان به گولیاندینا و همکاران (۲۰۰۱) مراجعه نمود.

روش *SSA* برای تشخیص نقطه تغییر با استفاده از توابع ناهمگنی بیان‌شده، توابعی را به نام *توابع تشخیص* معرفی می‌کند که در شکل ۳-۱۰ به همراه ماتریس ناهمگنی نمایش داده شده است.



شکل ۳-۱۰: نمودار ماتریس ناهمگنی و توابع تشخیص

با توجه به مطالب گفته‌شده، همان‌طور که در شکل ملاحظه می‌شود، توابع تشخیص سطری منطبق بر اولین سطر ماتریس ناهمگنی، یعنی $D_{T,N} = g(X_{1\,,B}\colon X_{,n-T+1,n})$، و توابع تشخیص ستونی منطبق بر اولین ستون ماتریس ناهمگنی، یعنی $D_{B,N} = (X_{n-B+1,n}\colon X_{1\,,T})$، می‌باشند. قطر اصلی این ماتریس نشان‌دهنده تابع تشخیص متقارن و قطرهای موازی آن نشان‌دهنده توابع تشخیص قطری هستند. به این ترتیب می‌توان از این توابع برای تشخیص نقطه تغییر استفاده کرد، یعنی محلی که در آن روند همگنی بین مشاهدات تغییر کند به عنوان نقطه تغییر در نظر گرفته می‌شود.

۳-۲-3 انتخاب پارامترها

انتخاب مناسب پارامترهایی که در الگوریتم اصلی SSA (M و I) و الگوریتم SSA برای تشخیص نقطه تغییر به‌کار می‌روند (q , p) بسیار حایز اهمیت است، به طوری‌که تشخیص دقیق‌تر نقاط تغییر در سری با مقادیر قابل قبولی برای پارامترها امکان‌پذیر می‌باشد. بنابراین تشخیص تغییرات کوچک در سری دارای نوفه، نیازمند انتخاب مقادیر دقیق هر یک از پارامترها است. در مورد انتخاب پارامترهای عرض پنجره، m، طول و محل نمونه آزمون، p و q، پارامترهای الگوریتم SSA شامل پارامتر تأخیر M و گروه I، پیشنهادات زیادی وجود دارند که در بخش‌های زیر ارایه شده‌اند.

۳-۲-۳-۱ عرض پنجره m

عرض پنجره، به عنوان زیرسری از سری اولیه در نظر گرفته شده و وابسته به نوع تغییرات ساختاری است که جستجو می‌شود. مطلوب‌تر آن است که m بزرگ انتخاب شود ولی اگر تغییرات تدریجی کوچک در سری وجود داشته باشد، نمی‌توان m را بیش از حد بزرگ انتخاب کرد. اگر m بیش از حد بزرگ انتخاب شود، رفتار آماره‌ی تغییر بسیار هموار شده و ممکن است برخی از تغییرات ساختاری به خوبی تشخیص داده نشوند. بنابراین بهتر است m طوری انتخاب شود که در هر زیر سری به طول m تقریبا یک تغییر وجود داشته باشد. روش SSA برای انتخاب m بیش از حد کوچک، دارای دقت ناچیز بوده و دنباله آماره تغییر، تعداد زیادی از نقاط و همین طور نقاط پرت را به عنوان نقطه تغییر تشخیص می‌دهد (موسکوینا، ۲۰۰۰؛ ۲۰۰۱).

۳-۲-۳-۲ پارامترهای الگوریتم اصلی SSA (پارامترهای تأخیر M و گروه I)

در الگوریتم SSA ، انتخاب پارامترهای تأخیر M و گروه I بسیار مهم است. به طوری‌که اگر طول سری، N، بیش از حد بزرگ نباشد، مقدار قراردادی $M=\left[\frac{N}{2}\right]$ انتخاب شده و گروه I برابر مجموعه $\{1,2,\dots,l\}$ است که عناصر گروه I، l مؤلفه اولی هستند که سری را به خوبی توصیف می‌کنند و $M-l$ مؤلفه‌ی باقیمانده، متناظر با سری خطا در نظر گرفته می‌شوند. بنابراین سری اولیه x_t، به دو سری، که یکی بیان‌کننده سری واقعی داده‌ها، z_t، و دیگری نمایانگر سری خطا ε_t، است تجزیه می‌شود به طوری که

$$x_t \simeq z_t + \varepsilon_t\ .$$

در تجزیه سری، تعداد مؤلفه‌های گروه I طوری انتخاب می‌شوند که دو سری از هم مستقل و تفکیک‌پذیر باشند. این تفکیک‌پذیری بخش مهمی از تجزیه SSA است و یکی از معیارهای تفکیک‌پذیری، مقدار W-همبستگی است که به صورت

$$corr_w(z_t, \varepsilon_t) = \frac{\sum_{t=1}^{N} w_t z_t \varepsilon_t}{\left(\sum_{t=1}^{N} w_t z_t^2 \sum_{t=1}^{N} w_t \varepsilon_t^2\right)^{\frac{1}{2}}} ,$$

تعریف می‌شود، که در آن w_t به صورت

$$w_t = \begin{cases} t & 1 \le t \le M, \\ M & M \le t \le K, \\ K + M - t & k < t \le N = K + M - 1, \\ ۰ & \text{جاهای دیگر}, \end{cases}$$

محاسبه می‌شود (گولیاندینا و همکاران، ۲۰۰۱).

W-همبستگی در واقع نمایشی از ماتریس خودهمبستگی سری است که نمایانگر تعداد l مقدار ویژه‌ای است که تفکیک‌پذیری مناسبی از دو سری را بیان می‌کنند (گولیاندینا و همکاران، ۲۰۰۱).

یکی دیگر از روش‌های انتخاب مؤلفه‌های l، استفاده از بازرسی بصری تجزیه SSA است، به این ترتیب که به هر مقدار ویژه یک وزن داده شده و آن تعداد از مقادیر ویژه که دارای درصد قابل قبولی باشند، به عنوان گروه I در نظر گرفته می‌شوند. به این نکته باید توجه داشت که انتخاب l کوچک باعث کم‌برازشی[1] در تشخیص نقطه تغییر شده و ممکن است تغییری که در مؤلفه‌های تخمین‌زده‌نشده موجود است را شناسایی نکند و به طور متقابل اگر l بیش از حد بزرگ باشد، ممکن است بخشی از خطا با سری پایه تقریب شده و تشخیص نقطه تغییر دچار مشکل شده و بیش‌برازشی[2] رخ دهد. روش کلی که درکتاب‌ها و مقالات مختلف SSA برای انتخاب l پیشنهاد شده است، انتخاب l مقدار ویژه‌ای است که دارای بزرگ‌ترین مقادیر هستند (گولیاندینا و همکاران، ۲۰۰۱). شرح کاملی از انتخاب مقدارهای تأخیر M، گروه l از مجموعه شاخص و بردارهای ویژه را می‌توانید در کتاب گولیاندینا و همکاران (۲۰۰۱) ملاحظه کنید.

---

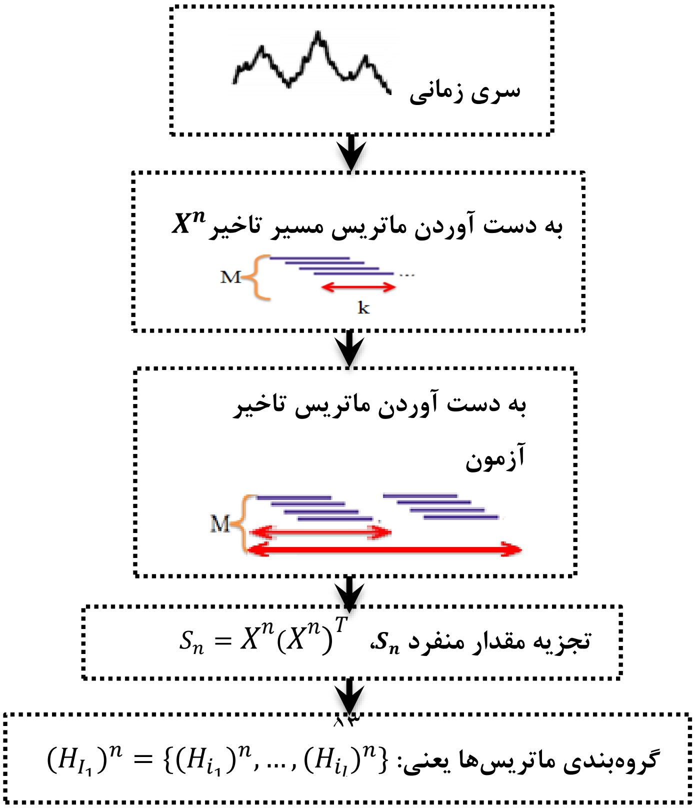
[1] Under dispersion

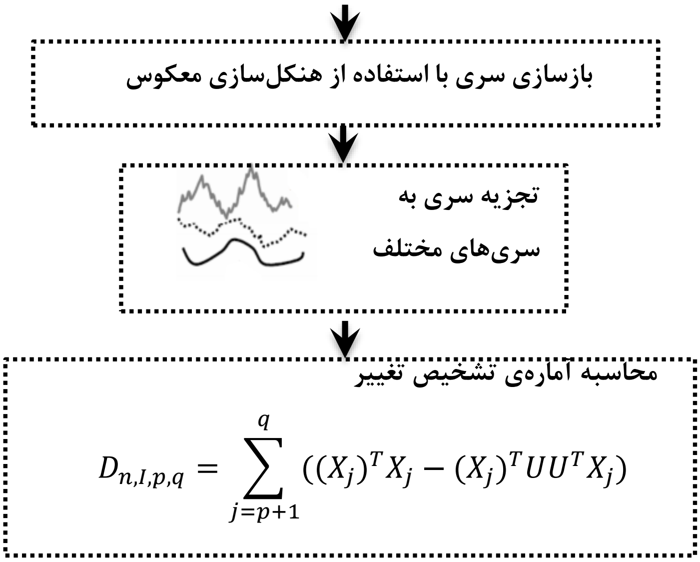
[2] Over dispersion

۳-۲-۳-۳ طول و محل نمونه آزمون: q,p

انتخاب طول و محل نمونه آزمون در تشخیص نقاط تغییر بسیار حایز اهمیت است و پیشنهادات زیادی در رابطه با مقدار انتخابی آن‌ها وجود دارند. پیشنهاد کلی، انتخاب $p \geq K$ است، زیرا باعث متمایز شدن ستون‌های ماتریس اصلی و آزمون شده و ماتریس‌ها دارای عناصر متفاوت خواهند شد. طبق شبیه‌سازی‌های عددی صورت‌گرفته توسط ژیگل جاوسکی و موسکوینا (۲۰۰۳) انتخاب مقادیر p و q که برای آن‌ها $q - p = 1$ باشد، انتخابی قابل قبول و حتی بهینه خواهد بود. ولی برای انتخاب $q - p > 1$، آماره تغییر دارای رفتاری هموارتر می‌شود و اگر این اختلاف بیش از حد بزرگ باشد، هموار شدن بیش از حد آماره تغییر، تشخیص نقطه تغییر را مشکل خواهد کرد. همیشه $q - p \leq M$، انتخاب می‌شود (ژیگل جاوسکی و موسکوینا، ۲۰۰۳).

با توجه به مطالب گفته‌شد، می‌توان روش SSA را برای تشخیص نقطه تغییر به‌کار برد. ساختار الگوریتمی روش SSA که برای تشخیص نقطه تغییر به‌کار برده می‌شود، در شکل ۳-۱۲ نشان داده شده است.





شکل ۳-۱۱: ساختار الگوریتم *SSA* برای تشخیص نقطه تغییر

فصل چهارم

بررسی دو روش بیزی درتشخیص نقطه تغییر

در این فصل به بررسی دو روش بیزی برای تشخیص نقطه تغییر، پرداخته می‌شود که شامل مدل افراز ضربی (PPM) و روش تشخیص برخط نقطه تغییر بیزی ($BOCPD$) می‌باشد. این دو روش به ترتیب در دو رویکرد غیر دنباله‌ای و دنباله‌ای مورد بررسی قرار می‌گیرند.

۴-۱ مدل افراز ضربی

یکی دیگر از روش‌های موجود برای تشخیص نقطه تغییر که دارای رویکردی بیزی است، روش مدل افراز ضربی می‌باشد که از این پس به اختصار آن را PPM می‌نامیم. همان طور که در فصل دوم گفته شد، PPM یکی از روش‌های غیردنباله‌ای (واخط) تشخیص نقطه تغییر است که کاربردهای فراوانی در زمینه‌های مختلف علوم، مانند پزشکی و مهندسی، کنترل صنعتی، اقتصاد و غیره دارد. مدل PPM اولین بار در سال ۱۹۹۰ توسط بری و هارتیگن ارایه شد و به نام این دو محقق شناخته می‌شود. بری و هارتیگن (۱۹۹۲، ۱۹۹۳) و کرولی[1] (۱۹۹۷)، PPM را برای تشخیص نقاط تغییر در میانگین و واریانس داده‌های نرمال به‌کار بردند و از روش نمونه‌گیری گیبز[2] برای محاسبه پارامترهای توزیع‌هایی که دارای شکل بسته‌ای نیستند استفاده کردند. لوسچی و کروز[3] (۲۰۰۱) این روش را بهبود بخشیدند و آن را برای انواع داده‌ها (پواسون، دوجمله‌ای و غیره) به‌کار بردند.

فرض کنید X_i، $(i = 1, \dots, n)$ متغیرهای تصادفی با تابع چگالی احتمال $f_{X_i}(x_i|\theta_i)$ باشند که بر حسب زمان مرتب شده‌اند و پارامترهای $\theta_i \in \Theta \ (i = 1, \dots, n)$ ممکن است از یک مشاهده به مشاهده دیگر تغییر کنند. بنابراین می‌توان X_i ها را یک سری زمانی فرض نمود که ممکن است دارای نقاط تغییر مجهولی باشد و هدف ما تشخیص این نقاط است. در این بخش، روشی را بیان می‌کنیم که برای تشخیص این نقاط، از تقسیم‌بندی یا افراز مجموعه داده‌ها استفاده می‌کند. در واقع افراز یک

---

[1] Crowley
[2] Gibs sampling
[3] Loschi and Cruz

مجموعه، تقسیم‌بندی آن به زیرمجموعه‌های کوچکتر و غیر تهی است، به طوری‌که اشتراک آن‌ها برابر مجموعه مرجع بوده و همچنین هیچ دو زیرمجموعه‌ای دارای عضو مشترک نباشند.

فرض کنید مجموعه داده $\{x_1,\dots,x_n\}$ را بتوان با یک نمادگذاری مجدد، به صورت

$$\{x_1,\dots,x_{i_1},x_{i_1+1},\dots,x_{i_2},\dots,x_{i_{j-1}+1},\dots,x_{i_j},\dots,x_{i_{b-1}},x_{i_b}\}, \tag{۱-۴}$$

نوشت که در آن $1 = i_0 < \cdots < i_1 < \cdots < i_b = n$. اگر فرض کنیم $s_j = \{x_{i_{j-1}+1},\dots,x_{i_j}\}$، آن‌گاه $\rho = \{s_1,s_2,\dots,s_b\}$ را یک افراز مجموعه داده مرجع نامیده و به هر s_j یک "بلوک" می‌گویند. از آن‌جا که b می‌تواند یکی از مقادیر $1,\dots,n$ را اختیار کند، لذا با تعریف متغیر تصادفی B به عنوان "تعداد بلوک‌ها" می‌توان افراز ρ را یک *افراز تصادفی* قلمداد کرد. بدین ترتیب می‌توان افرازهای متفاوتی را روی مجموعه داده مرجع در نظر گرفت. به عنوان مثال، چند نمونه از افرازهای مختلفی را که می‌توان روی مجموعه مشاهدات $x_1,\dots,x_{20}$ در نظر گرفت، در زیر آورده شده‌اند:

افراز سه بلوکی $\rho_1 = \{\{x_1,\dots,x_5\},\{x_6,\dots,x_8\},\{x_9,\dots,x_{20}\}\},$

افراز سه بلوکی $\rho_2 = \{\{x_1,\dots,x_7\},\{x_8,\dots,x_{14}\},\{x_{15},\dots,x_{20}\}\},$

افراز چهار بلوکی $\rho_3 = \{\{x_1,\dots,x_3\},\{x_4,\dots,x_6\},\{x_7,\dots,x_{12}\},\{x_{13},\dots,x_{20}\}\},$

فرض کنید داده‌های هر بلوک، مصداقی از مقادیر متغیرهای تصادفی باشند که پارامتر توزیع آن‌ها یکسان است یعنی در هر بلوک مانند s_j داریم $\theta_{i_{j-1}+1} = \cdots = \theta_{i_j} = \theta_j$. همچنین فرض کنید با انتقال از یک بلوک به بلوک دیگر، پارامتر توزیع تغییر کند. بنابراین مشاهدات واقع در اول هر بلوک را می‌توان "*نقاط تغییر*" تلقی نمود، یعنی با توجه به تعریف $s_j\,;\, j = 1,\dots,b$ نقاط تغییر عبارتند از $x_{i_1+1},x_{i_2+1},\dots,x_{i_{b-1}+1}$. شایان ذکر است که افرازهای متفاوت، القاکننده بلوک‌های متفاوت و در

نتیجه نقاط تغییر متفاوتی هستند. برای مثال، در افراز ρ_1، نقاط x_6 و x_9، در افراز ρ_2، نقاط x_8 و x_{15} و در افراز ρ_3، نقاط x_7, x_4 و x_{13} به عنوان نقاط تغییر در نظر گرفته می‌شوند.

این نکته قابل ذکر است که مشاهده x_1 به عنوان نقطه تغییر در نظر گرفته نمی‌شود، ولی هر مشاهده دیگر می‌تواند نقطه تغییر باشد. اکنون دو سوال اساسی پیش می‌آید:

۱- محتمل‌ترین تعداد نقاط تغییری که هر سری داده می‌تواند داشته باشد، چه تعداد است.

۲- اگر تعداد نقاط تغییر، داده شده باشد $(B = b)$، محتمل‌ترین این نقاط کدامند.

به عبارت دیگر اگر تابع احتمال B معلوم باشد، ابتدا می‌خواهیم محتمل‌ترین تعداد نقاط تغییر (K) در آن سری داده یعنی $K = \underset{i=1,\ldots,n}{\arg max}\, P(B = i)$ را یافته و سپس $K - 1$ مشاهده از مجموعه داده $\{ x_1, \ldots, x_n \}$ را به گونه‌ای انتخاب کنیم که بتوان آن‌ها را محتمل‌ترین نقاطی دانست که سری زمانی در آن‌جا دچار تغییر شده است. برای روشن شدن مطلب، افرازهای متناظر با پیشامدهای $(b = 1, \ldots, n)$ $B = b$ در جدول ۴-۱ را در نظر بگیرید.

جدول ۴-۱: دسته‌بندی افرازهای ممکن مجموعه داده $\{x_1, \ldots, x_n\}$ بر اساس تعداد بلوک

افراز n بلوکی $B = n$	...	افرازهای ۳ بلوکی $B = 3$	افرازهای ۲ بلوکی $B = 2$	افرازهای ۱ بلوکی $B = 1$
هر مشاهده در یک بلوک	...	$\{\{x_1\}, \{x_2\}, \{x_3, \ldots, x_n\}\}$ $\{\{x_1\}, \{x_2, x_3\}, \{x_4, \ldots, x_n\}\}$ $\{\{x_1, x_2\}, \{x_3, x_4\}, \{x_5, \ldots, x_n\}\}$ ...	$\{\{x_1\}, \{x_2, \ldots, x_n\}\}$ $\{\{x_1, x_2\}, \{x_3, \ldots, x_n\}\}$ $\{\{x_1, x_2, x_3\}, \{x_4, \ldots, x_n\}\}$ ...	$\{[x_1, \ldots, x_n]\}$

در بخش ۴-۱-۴ نشان داده شده است که چگونه می‌توان با فرض ثابت بودن تعداد بلوک، احتمال افرازهای هر ستون جدول ۴-۱، را محاسبه نمود و سرانجام مقادیر احتمال افرازهای ۱، ۲،۳،...، n بلوکی را پیدا کرد. برای یافتن نقاط تغییر و پاسخ سوال دوم باید افرازهای جدول ۴-۱ را به

گونه‌ای دیگر، همانند جدول ۴-۲، مرتب نمود که ستون i ام آن مبین افرازهایی است که در آن،مشاهده i ام به عنوان نقطه تغییر تلقی می‌گردد.

جدول ۴-۲:دسته‌بندی کل افرازهای ممکن مجموعه داده $\{x_1, \dots, x_n\}$ بر اساس نقطه تغییر

مشاهده x_n نقطه تغییر است (پیشامد A_n)		مشاهده x_3 نقطه تغییر است (پیشامد A_3)	مشاهده x_2 نقطه تغییر است (پیشامد A_2)
افرازهای ۲بلوکی $\{\{x_1, \dots, x_{19}\}, \{x_n]\}$	…	افرازهای ۲بلوکی $\{\{x_1, x_2\}\{x_3, \dots, x_n]\}$	افرازهای ۲بلوکی $\{\{x_1\}, \{x_2, \dots, x_n\}\}$
افرازهای ۳بلوکی $\{\{x_1\}, \{x_2, \dots, x_3\}, \{x_n\}\}$ $\{\{x_1, x_2\}, \{x_3, \dots, x_{19}\}, \{x_n\}\}$ …	..	افرازهای ۳بلوکی $\{\{x_1\}, \{x_2\}, [x_3, \dots, x_n\}\}$ $\{\{x_1, x_2\}, \{x_3, \dots, x_{19}\}, \{x_n\}\}$ …	افرازهای ۳بلوکی $\{\{x_1\}, \{x_2\}, \{x_3, \dots, x_n\}\}$ $\{\{x_1\}, \{x_2, x_3\}, \{x_4, \dots, x_n\}\}$ …
⋮	⋮	⋮	⋮
افرازهای n بلوکی $\{\{x_1\}, \{x_2\}, \dots, \{x_n\}\}$	…	افرازهای n بلوکی $\{\{x_1\}, \{x_2\}, \{x_3\} \dots, \{x_n\}\}$	افرازهای n بلوکی $\{\{x_1\}, \{x_2\}, \dots, \{x_n\}\}$

برای مثال فرض کنید $\arg\max_{i=1,\dots,n} P(B = i) = 3$، یعنی احتمال $P(B = b)$ به ازای $b = 3$ به حداکثر مقدار خود می‌رسد. بنابراین محتمل‌ترین تعداد نقاط تغییری که این سری داده می‌تواند داشته باشد، برابر ۲ است. در این‌جا برای پاسخ به سوال دوم مبنی بر این‌که کدام دو نقطه می‌توانند نقطه تغییر باشند، پیشامد A_k را به‌صورت

$A_k =$ "پیشامد آن‌که نقطه زمانی x_k نقطه تغییر باشد"، $\qquad k = 1, \dots, n,$

تعریف می‌کنیم. سپس بایستی در ردیف افرازهای ۳-بلوکی جدول ۴-۳، احتمال‌های $P(A_k | x_1, \dots, x_n)$، را به ازای $k = 1, \dots, n$ محاسبه نموده و سرانجام دو نقطه‌ای را که دارای

بیشترین احتمال هستند به عنوان نقاط تغییر در نظر بگیریم. نحوه محاسبه این احتمال‌ها در بخش ۴-۱-4 آمده است. بسط این احتمال‌ها برای تعدادی از x_k ها $(k = 2,3,4)$ با فرض $b = 3$ در زیر نشان داده شده است.

$$P(A_2|X_1,\dots,X_n) = \begin{cases} P(\rho = \rho_1 = \{[x_1], [x_2], [x_3,\dots,x_{20}]\}) \\ +P(\rho = \rho_2 = \{[x_1], [x_2,x_3], [x_4,\dots,x_{20}]\}) \\ +P(\rho = \rho_3 = \{[x_1], [x_2,x_3,x_4], [x_5,\dots,x_{20}]\}) \\ \vdots \\ +P(\rho = \rho_{M_2} = \{[x_1], [x_2,\dots,x_{19}], [x_{20}]\}) \end{cases} \qquad (۲\text{-}۴)$$

$$P(A_3|X_1,\dots,X_n) = \begin{cases} P(\rho = \rho_1 = \{[x_1], [x_2], [x_3,\dots,x_{20}]\}) \\ +P(\rho = \rho_2 = \{[x_1,x_2], [x_3], [x_4,\dots,x_{20}]\}) \\ +P(\rho = \rho_3 = \{[x_1,x_2], [x_3,x_4], [x_5,\dots,x_{20}]\}) \\ +P(\rho = \rho_4 = \{[x_1,x_2], [x_3,x_4,x_5], [x_5,\dots,x_{20}]\}) \\ \vdots \\ +P(\rho = \rho_{M_3} = \{[x_1,x_2], [x_3,\dots,x_{19}], [x_{20}]\}) \end{cases}$$

$$P(A_4|X_1,\dots,X_n) = \begin{cases} P(\rho = \rho_1 = \{[x_1,x_2], [x_3], [x_4,\dots,x_{20}]\}) \\ +P(\rho = \rho_2 = \{[x_1], [x_2,x_3], [x_4,\dots,x_{20}]\}) \\ +P(\rho = \rho_3 = \{[x_1,x_2,x_3], [x_4], [x_5,\dots,x_{20}]\}) \\ +P(\rho = \rho_4 = \{[x_1,x_2,x_3], [x_4,x_5], [x_6,\dots,x_{20}]\}) \\ \vdots \\ +P(\rho = \rho_{M_4} = \{[x_1,x_2,x_3], [x_4,\dots,x_{19}], [x_{20}]\}) \end{cases}$$

که در آن M_k تعداد کل افرازهای سه بلوکی است که در آن‌ها نقطه k ام $(k = 2,\dots,n)$، نقطه تغییر است. این روند تا محاسبه $P(A_n|x_1,\dots,x_n)$ و تا انتخاب بهترین دو نقطه ادامه خواهد یافت. محاسبه احتمال‌های $P(B = i)$ و $P(A_k|x_1,\dots,x_n)$ به ترتیب در بخش‌های ۴-۱-۳ و ۴-۱-۴ آمده‌اند اما قبل از آن لازم است که با فرض ثابت بودن تعداد بلوک‌ها، احتمال یک افراز مشخص، یعنی

$P(\rho=\rho_j)$، و تابع چگالی توام مشاهدات با این شرط که افراز تصادفی ρ به صورت ρ_j باشد، یعنی $f(x_1,\dots,x_n|\rho_j)$، را محاسبه کنیم.

۴-۱-۱ تابع احتمال یک افراز مشخص با فرض ثابت بودن تعداد بلوک

اگر برای هر مشاهده، "موفقیت" را معادل با "نقطه تغییر بودن" آن مشاهده تعریف کنیم و در نظر بگیریم $p=P(\text{موفقیت})$، آن‌گاه با توجه به مفهوم تعریف‌شده برای بلوک می‌توان گفت که در مشاهده اول هر بلوک، موفقیتی رخ داده و سایر مشاهدات آن بلوک به عنوان "شکست" تلقی می‌شوند. لذا برای بلوک $s_j=[x_{i_{j-1}+1},\dots,x_{i_j}]$ داریم

$$P(s_j)=c_{[i_{j-1}i_j]}=\begin{cases}p(1-p)^{j-i-1}\,, & j<n,\\ (1-p)^{j-i-1}\,, & j=n.\end{cases}\qquad (۳\text{-}۴)$$

حال با فرض این‌که تعداد بلوک‌ها (b) و مقدار p داده شده باشد، می‌توان نوشت

$$P(\rho_j|p)=\frac{\prod_{j=1}^{b}c_{[i_{j-1}i_j]}}{\sum_{C}\prod_{j=1}^{b}c_{[i_{j-1}i_j]}}\propto\ p^{b-1}(1-p)^{n-b}\,,$$

که در آن، C همه افرازهای ممکن مجموعه داده و ρ_j یکی از افرازهای ممکن b-بلوکی است. برای روشن‌تر شدن مطلب در حالت $b=3$ افرازهای ممکن سه بلوکی (رابطه (۴-2)) و احتمال‌های متناظر به‌صورت زیر در نظر گرفته می‌شوند:

$$P(\,\rho=\{\,[\,x_1],\ [x_2],\ [x_3,\dots,x_{20}]\,\}\,)=p.p.p(1-p)^{20-3}=p^3(1-p)^{17}$$
$$P(\,\rho=\{\,[\,x_1,\ x_2],\ [x_3],\ [x_4,\dots,x_{20}]\,\}\,)=p(1-p).p.p(1-p)^{20-4}=p^3(1-p)^{17}$$
$$P(\,\rho=\{\,[\,x_1,\ x_2,\ x_3],\ \ [x_4,x_5],\ [x_6,\dots,x_{20}]\,\}\,)=p(1-p)^2.p(1-p).p(1-p)^{20-5}=p^3(1-p)^{17}$$
$$\vdots$$
$$P(\,\rho=\{\,[\,x_1],\ \ [x_2,\dots,\ x_{19}],\ [x_{20}]\,\}\,)=p.p(1-p)^{19-2}\,.p=p^3(1-p)^{17}$$

حال اگر p را یک متغیر تصادفی با توزیع پیشین $\pi(p)$ در نظر بگیریم، که $\pi(p)$ یک توزیع پیوسته در بازه $(0,1)$ باشد، آن‌گاه با توجه به قانون احتمال کل داریم

$$P(\rho_j) = \int_{\circ}^{1} p^{b-1}(1-p)^{n-b}\pi(p)dp. \qquad (۴-۴)$$

به عنوان مثال اگر توزیع پیشین p را توزیع بتا در نظر بگیریم ($p \sim B(\alpha,\beta)$)، آن‌گاه خواهیم داشت

$$P(\rho_j) = \int_{\circ}^{1} p^{b-1}(1-p)^{n-b}.\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}dp$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.\int_{\circ}^{1} p^{\alpha+b-2}(1-p)^{n-b+\beta-1}dp$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.\frac{\Gamma(\alpha+b-1)\Gamma(n+\beta-b)}{\Gamma(n+\alpha+\beta-1)}.$$

۴-۱-۲ تابع چگالی توام مشاهدات در هر افراز با فرض ثابت بودن تعداد بلوک

قبلا گفته شد، که هر X_i دارای تابع چگالی $f(x_i|\theta_i)$ می‌باشد و با توجه به تعریف افراز، می‌توان بلوک‌ها را نسبت به هم مستقل فرض کرد. از آنجا که هر مشاهده x_i می‌تواند درون یک بلوک قرار گیرد (با توجه به بزرگ‌ترین افراز ممکن)، می‌توان x_iها را نسبت به هم مستقل در نظر گرفت و به دلیل این که فرض کردیم مشاهدات واقع در هر بلوک، همگن و پارامتر توزیع آن‌ها یکسان است، لذا پارامتر مربوط به بلوک s_j را به‌صورت $\theta_j = \theta_{i_{j-1}+1} = \theta_{i_{j-1}+2} \dots = \theta_{i_j}$ نمایش می‌دهیم. بنابراین تابع چگالی توام مشاهدات هر بلوک مشخص s_j به صورت

$$f(s_j|\theta_j) = f\left(x_{i_{j-1}+1},\dots,x_{i_j}\middle|\theta_j\right) = \prod_{k=i_{j-1}+1}^{i_j} f(x_k|\theta_j),$$

به دست می‌آید. به این ترتیب تابع چگالی توام مشاهدات به ازای هر افراز مشخصی مانند ρ_j و با فرض ثابت بودن تعداد بلوک‌ها به‌صورت

$$f\left(x_1,\dots,x_n\middle|\rho_j\right)=\prod_{j=1}^{b} f\left(s_j\middle|\theta_j\right), \qquad (5-4)$$

تعریف می‌شود.

۴-۱-۳ تابع احتمال تعداد بلوک

در این‌جا $P(\rho_j)$ و $f\left(x_1,\dots,x_n\middle|\rho_j\right)$ برای تعداد بلوک مشخص (b) را که در بخش‌های ۴-۱-۱ و ۴-۱-۲ محاسبه شدند، برای محاسبه احتمال تعداد بلوک به‌کار می‌بریم. با توجه به ستون‌های جدول ۴-۱، پیشامد $B=b$، در واقع مجموع احتمال‌های تمام افرازهای b-بلوکی $(b=1,\dots,n)$ آن ستون بوده و به صورت

$$P(B\;=\;b|x_1,\dots,x_n\,)=\sum\nolimits_{\rho\epsilon C_b} P(\rho=\rho_j|x_1,\dots,x_n), \qquad (6-4)$$

محاسبه می‌شود، که در آن C_b مجموعه کل افرازهای ممکن b-بلوکی است. برای محاسبه جمله عمومی مجموع (۴-۶)، با استفاده از قضیه بیز داریم

$$P(\rho_j=\{s_1,\dots,s_b\}|x_1,\dots,x_n\}=\frac{f\left(x_1,\dots,x_n\middle|\rho_j\right)P\left(\rho=\rho_j\right)}{f(x_1,\dots,x_n)}.$$

در نتیجه

$$P(\mathrm{B}\;=\;\mathrm{b}|x_1,\dots,x_n\,)\propto k\sum\nolimits_{\rho_j\epsilon C_b} f(x_1,\dots,x_n|\rho_j)P(\rho=\rho_j),$$

که در آن $k=\frac{1}{f(x_1,\dots,x_n)}$ و $P(\rho=\rho_j)$ و $f(x_1,\dots,x_n|\rho_j)$ به ترتیب در رابطه‌های (۴-۴) و (۴-۵) تعریف شده‌اند، بنابراین داریم

$$P(\mathrm{B}\;=\;\mathrm{b}|x_1,\dots,x_n\,)\propto\sum\nolimits_{\rho_j\epsilon C_b}\prod_{j=1}^{b} f\left(s_j\middle|\theta_j\right)\int_{\circ}^{1} p^{b-1}(1-p)^{n-b}\pi(p)dp. \qquad (7-4)$$

لازم به ذکر است که مقدار k ثابت بوده و از آنجا که هدف ما تشخیص محتمل‌ترین مقدار $P(\mathrm{B} = \mathrm{b}|x_1, \dots, x_n)$، با انجام مقایسه این احتمال‌ها، است، لذا مقدار k در محاسبات نقشی ایفا نمی‌کند.

واضح است که با مشخص شدن محتمل‌ترین تعداد بلوک در افراز مجموعه داده $x_1, \dots, x_n$، یعنی $K = \arg\max_i P(B = i)$، تعداد نقاط تغییر نیز مشخص می‌شوند. اما نمی‌دانیم کدامیک می‌توانند نقطه تغییر باشند که در بخش بعدی این موضوع بررسی خواهد شد.

۴-۱-۴ احتمال نقطه تغییر بودن هر مشاهده با فرض ثابت بودن تعداد بلوک

مطلب را با ادامه مثال قبل آغاز می‌کنیم که در آن $\arg\max_i P(B = i) = 3$. برای محاسبه $P($ نقطه تغییر باشد $x_2)$، با توجه به ردیف دوم از ستون اول جدول ۴-۲ که در آن بعضی از افرازهای سه بلوکی آمده است، داریم:

$$\begin{aligned}
P(\,B = 3 \text{ و } x_2 \text{ نقطه تغییر است}) &= P(\rho = \{\,[\,x_1],\ [x_2], [x_3, \dots, x_{20}]\,\}\,) \\
&+ P(\rho = \{\,[\,x_1],\ [x_2, x_3], [x_4, \dots, x_{20}]\,\}\,) \\
&+ P(\rho = \{\,[\,x_1],\ [x_2, x_3, x_4], [x_5, \dots, x_{20}]\,\}\,) \\
&\quad\vdots \\
&+ P(\rho = \{\,[\,x_1],\ [x_2, \dots, x_{19}], [x_{20}]\,\}\,) \\
&= \sum_{\rho_j \epsilon C_{2,3}} P\big(\rho = \rho_j | x_1, \dots, x_n\big).
\end{aligned}$$

که در آن $C_{2,3}$ مجموعه افرازهای سه بلوکی است که در آن x_2 نقطه تغییر است. بنابراین اگر b، ایده‌آل‌ترین تعداد بلوک‌ها (یعنی اگر $(b-1)$، ایده‌آل‌ترین تعداد نقاط تغییر) باشد، لذا می‌توان احتمال این که مشاهده kام بتواند نقطه تغییر باشد را به صورت

$$P(A_k|x_1,\dots,x_n\ ) \ = \sum\nolimits_{\rho_j \epsilon C_{k,b}} P(\rho=\rho_j|x_1,\dots,x_n), \qquad (8-4)$$

تعیین نمود، که در آن $C_{k,b}$ مجموعه همه افرازهای b-بلوکی است که در همه آن‌ها k امین داده، نقطه تغییر است. در واقع $C_{k,b} \subset C_b$ که C_b مجموعه همه افرازهای b -بلوکی است. محاسبه جمله عمومی رابطه (۴-۸)، همانند محاسبه جمله عمومی (۴-۶) می‌باشد. بنابراین داریم

$$P(A_k|x_1,\dots,x_n\ ) \ = \sum\nolimits_{\rho_j \epsilon C_{k,b}} f\big(x_1,\dots,x_n|\rho_j\big)P\big(\rho_j\big).$$

با توجه به محاسبه $P(\rho_j)$ و $f(x_1,\dots,x_n|\rho_j)$ که به ترتیب در رابطه‌های (۴-۴) و (۴-۵) بیان شده‌اند، داریم

$$P(A_k|x_1,\dots,x_n\ ) \ \propto \sum\nolimits_{\rho_j \epsilon C_{k,b}} \prod_{j=1}^{b} f\big(s_j|\theta_j\big) \int_0^1 p^{b-1}(1-p)^{n-b}\pi(p)dp. \qquad (9-4)$$

سرانجام با محاسبه احتمال‌های (۴-۸) برای هر مشاهده (به ازای $k=1,2,\dots,n$)، b-۱ تا از محتمل‌ترین آن‌ها، مکان نقاط تغییر را تعیین می‌کنند.

۴-۱-۵ تشخیص نقطه تغییر از دیدگاه تغییر در میانگین پارامترهای توزیع

در بخش ۴-۱-۱ فرض کردیم که $X_i\ (i=1,\dots,n)$ ها متغیرهای تصادفی با تابع چگالی $f_{X_i}(x_i|\theta_i)$ هستند که در آن $\theta_i \in \Theta\ (i=1,\dots,n)$. در آن بخش با توجه به مفهوم افراز پس از بلوک‌بندی دنباله $x_1,\dots,x_n$، محتمل‌ترین تعداد نقطه تغییر مشخص گردید و سرانجام احتمال این که هر مشاهده بتواند نقطه تغییر باشد، محاسبه شد. اما با توجه به این فرض که مشاهدات درون هر بلوک، همگن و دارای پارامتر یکسان هستند و با انتقال از یک بلوک به بلوک دیگر، پارامتر توزیع تغییر می‌کند، می‌توان از دیدگاه بلوک‌بندی پارامترهای θ_i نیز به مساله نگاه کرد. بنابراین می‌توان مقدار مورد انتظار هر پارامتر $\theta_k\ (k=1,\dots,n)$ یعنی $E(\theta_k|x_1,\dots,x_n)$ را محاسبه کرده و محل‌هایی (زمان-

هایی) که این مقادیر دارای تغییرات محسوسی می‌باشند را به عنوان نقاط تغییر در نظر گرفت. باید توجه نمود که θ_k پارامتر توزیع X_k است و صرف نظر از اینکه افراز تصادفی چگونه در نظر گرفته شود، بالاخره x_k در یکی از بلوک‌های آن افراز واقع می‌شود که این بلوک‌ها می‌توانند تکی (تنها شامل x_k)، دوتایی، سه‌تایی و "r-تایی، $1 \le r \le n$ ، باشند. به طور کلی x_k می‌تواند متعلق به بلوک‌های $(j-i)$-تایی $s_j = [x_{i+1},\dots,x_j]$ باشد که در آن $(i=0,\dots,n-1\,, j=1,\dots,n\,,\ i<k\le j)$. در نتیجه

$$E(\theta_k|x_1,\dots,x_n)=\sum\nolimits_{i<k\le j}E(\theta_k|\ \text{بلوک } s_j \text{ جزء افراز تصادفی باشد})\ P(\text{بلوک } s_j \text{ جزء افراز تصادفی})$$

لازم به ذکر است که s_j وابسته به i نیز می‌باشد، لذا با یک نمادگذاری مجدد داریم

$$E(\theta_k|x_1,\dots,x_n)=\sum_{i<k\le j}E(\theta_k|s_j)\,r_{ij}(X)$$

$$=\sum\nolimits_{i=0}^{k-1}\sum\nolimits_{j=k}^{n}\Big(\int_{\Theta}\theta_k f(\theta_k|\,s_j)d\theta_k\Big)\ r_{ij}(X), \qquad (10-4)$$

که در آن $r_{ij}(X)=P($ بلوک s_j جزء افراز تصادفی باشد $)$. بنابراین میانگین θ_k در هر نقطه زمانی x_k با توجه به رابطه (۴-۱۰) محاسبه می‌شود. اما قبل از آن لازم است که $f(\theta_k|\,s_j)$ و $r_{ij}(X)$ را محاسبه نماییم (لوسچی و کروز، ۲۰۰۳).

۱ – چگالی پسین پارامتر

همان طور که ملاحظه شد، برای محاسبه $E(\theta_k|s_j)$ نیازمند محاسبه چگالی پسین $f(\theta_k|\,s_j)$ هستیم و با توجه به داشتن بلوک $s_j = [x_{i+1},\dots,x_j]$، می‌دانیم $\theta_{i+1}=\theta_{i+2}=\dots=\theta_k=\dots=\theta_j$ لذا با استفاده از قضیه بیز داریم

$$f(\theta_k|s_j)=\frac{f(\theta_k)\,f(\,s_j|\theta_j)}{f(\,s_j)}=kf(\theta_k)\,f(\,s_j|\theta_j),$$

که در آن $k = \frac{1}{f(s_j)} = \frac{1}{f_{x_{i+1},\ldots,x_j}(x_{i+1},\ldots,x_j)}$ است و $f(\theta_k)$ چگالی پیشین بلوکی است که شامل پارامتر θ_k است و با توجه به بخش ۴-۱-۳، تابع چگالی توام مشاهدات بلوک s_j به صورت زیر به دست می‌آید:

$$f(s_j|\theta_j) = f(x_{i+1},\ldots,x_j|\theta_j) = \prod_{m=i+1}^{j} f(x_m|\theta_j)\ . \qquad (11-4)$$

۲ – احتمال تعلق یک بلوک خاص به افراز

لازم به یادآوری است که طبق رابطه (۴-۱۰) احتمال این که بلوک s_j متعلق به افراز ρ باشد را با کمیت r_{ij} نمایش دادیم. در واقع برای تمام افرازهای شامل بلوک s_j، با ثابت در نظر گرفتن این بلوک، بلوک‌بندی قبل و بعد آن در افرازهای مختلف، تغییر خواهد کرد. بنابراین احتمال تعلق بلوک s_j به افراز تصادفی ρ برابر مجموع احتمال تمام این افرازها است. با توجه به بخش ۴-۱-۱ و مستقل بودن بلوک‌ها، احتمال رخداد هر افراز برابر حاصلضرب احتمال‌های رخداد بلوک‌های آن می‌باشد. بنابراین احتمال رخداد هر افراز از مجموعه ρ به صورت

$$P(\rho = \rho_j) = P(\text{بلوک‌های قبل از بلوک ثابت } s_j)\,P(s_j)\,P(\text{بلوک‌های بعد از بلوک ثابت } s_j)$$

$$= \prod_{k=1}^{b_1} C_{i_{k-1}i_k}\,(C_{i\,j})\,\prod_{k=1}^{b_2} C_{i_{k-1}i_k}\,,$$

است، که در آن b_1 تعداد بلوک‌های ممکن مشاهدات قبل از بلوک s_j و b_2 تعداد بلوک‌های ممکن مشاهدات بعد از بلوک s_j است. بنابراین برای کلیه افرازهای ممکن داریم

$$\sum_{\rho_j} P(\rho = \rho_j) = \left(\sum_{0<k\le i}\prod_{k=1}^{b_2} C_{i_{k-1}i_k}\right)(C_{i\,j})\left(\sum_{j<k\le n}\prod_{k=1}^{b_2} C_{i_{k-1}i_k}\right).$$

که در آن $\lambda_{i\,j} = \sum_k \prod_{k=1}^{b} C_{i_{k-1}i_k}$ تعریف می‌کنیم. بنابراین با توجه به احتمال‌های نوشته‌شده، می‌توان احتمال‌های بلوک‌های قبل از بلوک $[i\,j]$ را با استفاده از فرمول بازگشتی زیر نوشت.

$$\lambda_{\circ\circ} = 1, \qquad \lambda_{\circ 1} = c_{\circ 1}, \qquad \lambda_{\circ\; i+1} = c_{\circ\, i+1} + \sum_{k=1}^{i} \lambda_{\circ\, k}\, c_{k\; i+1}.$$

محاسبه احتمال‌های بلوک‌های بعد از بلوک مشخص $[i\, j]$ نیز همانند بالا به صورت بازگشتی زیر نوشته می‌شود:

$$\lambda_{j\, n} = c_{j\, n} + \sum_{k=1}^{n-1} \lambda_{j\, k}\, c_{k\, n}\,, \qquad \lambda_{n-1\, n} = c_{n-1\, n}\,, \qquad \lambda_{n\, n} = 1\,.$$

به این ترتیب با استفاده از فرمول‌های بازگشتی محاسبه‌شده در بالا می‌توان احتمال این که بلوک $\mathrm{s_j}$ متعلق به افراز تصادفی ρ باشد را به صورت رابطه

$$r_{i\, j}(X) = \frac{\lambda_{\circ i}\ \ C_{ij}\ \lambda_{jn}}{\lambda_{\circ n}}\,, \qquad (12-4)$$

نوشت. برای مثال، فرض کنید بخواهیم احتمال یک افراز تصادفی را که شامل بلوک مشخص $\mathrm{s_j} = [x_4] = [3\ 4]$ است، محاسبه کنیم. طبق مطالب گفته‌شده برای محاسبه احتمال رخداد بلوک بندی قبل از بلوک $[x_4]$ داریم

$\rho = \{\, \{\, x_1, x_2, x_3\},\ \{x_4\}, \{x_5, \dots, x_{20}\}\}$

$P(\{\, \{\, x_1\}, \{x_2\}, \{x_3\},\ \{x_4\}, \{x_5, \dots, x_{20}\}\}) = C_{\circ 1}\ C_{12}\ C_{2\,3}\ C\ _{34}\ C_{4\,20}\ ,$

$P(\{\, \{\, x_1\}, \{x_2\ x_3\},\ \{x_4\}, \{x_5, \dots, x_{20}\}\}) = C_{\circ 1}. C_{13}. C_{3\,4}. C_{4\,20}\ ,$

$P\big(\{\, \{\, x_1, x_2\}, \{\, x_3\}, \{x_4\}, \{x_5, \dots, x_{20}\}\}\big) = C_{\circ 2}. C_{23}. C_{3\,4}. C_{4\,20}\ ,$

$P(\{\, \{\, x_1, x_2\ x_3\},\ \{x_4\}, \{x_5, \dots, x_{20}\}\}) = C_{\circ 3}. C_{34}. C_{4\,20}\ .$

سرانجام با جایگذاری $f(\theta_k \big| s_j\big)$ و $r_{ij}(X)$ در رابطه (۴-۱۰) که به ترتیب در روابط (۴-۱۱) و (۴-۱۲) محاسبه شدند، مقدار میانگین پارامترهای توزیع به دست آمده و می‌توان نقاطی را که در آن‌ها مقدار میانگین پارامتر تغییر محسوسی دارند، به عنوان نقاط تغییر در نظر گرفت.

۴-۱-۶ مدل افراز ضربی برای توزیع نرمال

در بخش‌های قبل روند محاسباتی تشخیص نقطه تغییر در مدل افراز ضربی از دو دیدگاه بررسی شد. در این بخش، روند محاسبات تشخیص نقاط تغییر در پارامتر میانگین مجموعه داده $\{x_1, \dots, x_n\}$ که دارای توزیع نرمال هستند، بیان می‌شود. فرض کنید $(i = 1, \dots, n)$ $X_i \sim N(\mu_i, \sigma^2)$ (پذیره نرمال بودن می‌تواند با هر پذیره دیگری جایگزین شود) و پارامتر میانگین هر بلوک s_j دارای چگالی پیشین $\mu \sim N(\mu_{\circ}, \sigma_{\circ}^2/(j-i))$ باشد بنابراین $\bar{X}_j \sim N(\mu_{\circ}, (\sigma_{\circ}^2 + \sigma^2)/(j-i))$ است. روند محاسباتی بیان‌شده در بخش‌های ۴-۱-۲ تا ۴-۱-۵ برای تشخیص نقطه تغییر از دیدگاه اول (تقسیم‌بندی مشاهدات) در حالت نرمال بودن توزیع داده‌ها به ترتیب در بخش‌های زیر (الف-د) آمده است. همچنین تشخیص نقطه تغییر از دیدگاه دوم (تغییر در میانگین پارامتر) در بخش (ه) بیان می‌شود.

همان‌طور که قبلا در اجرای مدل افراز ضربی مشاهده شد، قبل از محاسبه محتمل‌ترین تعداد نقاط تغییر و تشخیص محل این نقاط، نیاز به محاسبه تابع چگالی مشاهدات به شرط هر افراز ممکن $f(x_1, \dots, x_n | \rho_j)$ و احتمال یک افراز مشخص $P(\rho_j)$ (هر دو با فرض ثابت بودن تعداد بلوک)، داریم که به ترتیب در (الف) و (ب) محاسبه می‌شوند.

الف- تابع چگالی پیشین مشاهدات به شرط هر افراز با فرض ثابت بودن تعداد بلوک

$$f(s_j) = f\left(x_{i_{j-1}+1}, \dots, x_{i_j}\right) = f\left(x_{i_{j-1}+1}, \dots, x_{i_j} \middle| \mu = \bar{X}_j\right) f(\bar{X}_j)$$

$$= \prod_{l=i+1}^{j} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{\frac{-1}{2\sigma^2}(x_l - \bar{X}_j)^2} \frac{(j-i)^{1/2}}{(2\pi(\sigma_{\circ}^2 + \sigma^2))^{1/2}} e^{\frac{-(j-i)}{2(\sigma_{\circ}^2+\sigma^2)}(\bar{X}_j - \mu_{\circ})^2}$$

$$\propto \frac{1}{(2\pi\sigma^2)^{(j-i)/2}} \left(\frac{\sigma^2}{\sigma_{\circ}^2 + \sigma^2}\right)^{1/2} \exp(V_{ij}),$$

که در آن $V_{ij} = -\frac{\sum_{l=i+1}^{j}(x_l - \bar{X}_j)^2}{2\sigma^2} - \frac{(j-i)(\bar{X}_j - \mu_{\circ})^2}{2(\sigma_{\circ}^2+\sigma^2)}$، $\bar{X}_j = \sum_{l=i+1}^{j}(x_l)/(j-i)$ و $w = \frac{\sigma^2}{\sigma_{\circ}^2+\sigma^2}$.

بنابراین با محاسبه تابع چگالی بلوک طبق روابط (۴-۲)، تابع چگالی مشاهدات به شرط هر افراز ممکن به‌صورت زیر محاسبه می‌شود که وابسته به پارامترهای $\mu_\circ$، $\sigma_\circ^2$، σ^2 و ρ است:

$$f(x_1,\dots,x_n|\rho,\mu_\circ,w)=\prod_{j=1}^{b} f(s_j)\propto\prod_{j=1}^{b}\frac{1}{(2\pi\sigma^2)^{(j-i)/2}}\left(\frac{\sigma^2}{\sigma_\circ^2+\sigma^2}\right)^{1/2}\exp(V_{ij})\,.$$

ملاحظه می‌شود که تابع چگالی محاسبه‌شده وابسته به پارامترهایی است که هیچ اطلاعی در مورد این پارامترها وجود ندارد. یائو[1] (۱۹۸۴) توزیع‌های پیشین خاصی را برای پارامترهای $\mu_\circ, w$ و σ^2 به صورت

$$\pi(\mu_\circ)=1 \qquad -\infty\le\mu_\circ\le\infty\,,$$

$$\pi(\sigma^2)=\frac{1}{\sigma^2} \qquad \circ\le\sigma^2\le\infty\,,$$

$$\pi(w)=\frac{1}{w_\circ} \qquad \circ\le w\le w_\circ\,,$$

معرفی کرده است. بنابراین با توجه به این توزیع‌های پیشین، برای محاسبه تابع چگالی مشاهدات به شرط هر افراز ممکن ρ، توزیع حاشیه‌ای تابع چگالی را به ازای هر پارامتر به دست می‌آوریم که روند محاسبه در زیر بیان شده است.

$$f(x_1,\dots,x_n|\rho,\mu_\circ,w)\propto\int_\circ^\infty\frac{1}{\sigma^2}\prod_{j=1}^{b}f(s_j)d\sigma^2$$

$$=\int_\circ^\infty\frac{1}{\sigma^2}\prod_{j=1}^{b}\frac{1}{(2\pi\sigma^2)^{\frac{j-i}{2}}}\left(\frac{\sigma^2}{\sigma_\circ^2+\sigma^2}\right)^{\frac{1}{2}}\exp(-\frac{\sum_{l=i+1}^{j}(x_l-\bar{X}_j)^2}{2\sigma^2}-\frac{(j-i)(\bar{X}_j-\mu_\circ)^2}{2(\sigma_\circ^2+\sigma^2)})d\sigma^2$$

$$=\int_\circ^\infty\frac{1}{\sigma^2}\left(\prod_{j=1}^{b}\left(\frac{\sigma^2}{\sigma_\circ^2+\sigma^2}\right)^{\frac{1}{2}}\frac{1}{(2\pi\sigma^2)^{(j-i)/2}}\right)\exp(-\frac{\sum_{j=1}^{b}\sum_{l=i+1}^{j}(x_l-\bar{X}_j)^2}{2\sigma^2}-\frac{\sigma^2}{\sigma^2}$$

---

[1] Yao

$$\times\frac{\sum_{j=1}^{b}(j-i)(\bar{X}_j-\overline{X})^2+n(\overline{X}-\mu_\circ)^2}{2(\sigma_\circ^2+\sigma^2)})d\sigma^2, \qquad (11-4)$$

که در آن اگر تعداد بلوک‌ها در افراز ρ برابر b باشد، با توجه به تعاریف $\overline{X}=\sum_{i=1}^{n}\frac{x_i}{n}$، $B=\sum_{j=1}^{b}(j-i)(\bar{X}_j-\overline{X})^2$، $W=\sum_{j=1}^{b}\sum_{l=i+1}^{j}(x_l-\bar{X}_j)^2$ و $w=\frac{\sigma^2}{\sigma_\circ^2+\sigma^2}$، رابطه (۴-۱۱) به صورت زیر باز نویسی می‌شود.

$$\left(w^{\frac{b}{2}}\right)\int_\circ^\infty\frac{1}{\sigma^2}\left(\prod_{j=1}^{b}\frac{1}{(2\pi\sigma^2)^{\frac{j-i}{2}}}\right)\exp(-\frac{1}{2\sigma^2}\left[\mathrm{W}+\mathrm{Bw}+\mathrm{wn}(\overline{X}-\mu_\circ)^2\right])d\sigma^2$$

$$=\left(w^{\frac{b}{2}}\right)\int_\circ^\infty\frac{1}{(2\pi)^{\frac{n}{2}}}\frac{1}{(\sigma^2)^{\frac{n}{2}+1}}\exp(-\frac{1}{2\sigma^2}\left[\mathrm{W}+\mathrm{Bw}+\mathrm{wn}(\overline{X}-\mu_\circ)^2\right])d\sigma^2.$$

همچنین با تغییر متغیر $u=\frac{1}{\sigma^2}$ و $a=\left[\mathrm{W}+\mathrm{Bw}+\mathrm{wn}(\overline{X}-\mu_\circ)^2\right]$، خواهیم داشت

$$\left(w^{\frac{b}{2}}\right)\frac{1}{(2\pi)^{\frac{n}{2}}}\int_\circ^\infty u^{(\frac{n}{2}+1)}\exp(-\frac{u}{2}a)\frac{-1}{u^2}du\,.$$

این انتگرال، یک انتگرال گاما است، بنابراین داریم

$$\left(w^{\frac{b}{2}}\right)\frac{-1}{(2\pi)^{\frac{n}{2}}}\Gamma\left(\frac{n}{2}\right)(\frac{a}{2})^{\frac{n}{2}}\int_\circ^\infty\frac{1}{\Gamma\left(\frac{n}{2}\right)(\frac{a}{2})^{\frac{n}{2}}}u^{(\frac{n}{2})}\exp\left(-\frac{u}{2}a\right)du\,.$$

با جایگذاری مقدار a، نتیجه به صورت

$$f(x_1,\dots,x_n|\rho,\mu_\circ,w)\propto\frac{w^{\frac{b}{2}}}{[W+Bw+wn(\mu_\circ-\overline{X})^2]^{\frac{n}{2}}},$$

حاصل می‌شود. توزیع حاشیه‌ای روی $\mu_\circ$ و w نیز به صورت

$$f(x_1,\dots,x_n|\rho,w)\propto\int_\circ^{\mu_\circ}\frac{w^{\frac{b}{2}}}{[W+Bw+wn(\mu_\circ-\overline{X})^2]^{n/2}}d\mu_\circ\,,$$

به دست می آید که با تغییر متغیر $wn(\mu_\circ - \overline{X})^2 = (W + Bw)sinh^2 t$، داریم

$$\frac{w^{\frac{b}{2}}}{(W+Bw)^{\frac{n}{2}}}\int_{-\infty}^{+\infty}\sqrt{\frac{W+Bw}{wn}}\,\frac{\cos ht}{cosh^n t}dt = \frac{w^{\frac{b}{2}}}{(W+Bw)^{\frac{n}{2}}}\,\frac{(W+Bw)^{\frac{1}{2}}}{(wn)^{\frac{1}{2}}}\int_{-\infty}^{+\infty}\frac{dt}{cosh^{n-1}t}\,.$$

با توجه به اینکه $\int_{-\infty}^{+\infty}\frac{dt}{cosh^{n-1}t} \le \int_{-\infty}^{+\infty}\frac{dt}{cosh\,t} = \frac{\pi}{n-2}$، آن‌گاه

$$f(x_1,\dots,x_n|\rho,w) \propto \frac{\pi}{n-2}\,\frac{w^{\frac{b}{2}}}{(W+Bw)^{\frac{n}{2}}}\,\frac{(W+Bw)^{\frac{1}{2}}}{(wn)^{\frac{1}{2}}} \propto \frac{w^{\frac{(b-1)}{2}}}{[W+Bw]^{\frac{n-1}{2}}}.$$

و در نتیجه توزیع حاشیه‌ای افراز ρ به صورت

$$f(x_1,\dots,x_n|\rho) \propto \int_{\circ}^{w_\circ}\frac{w^{\frac{(b-1)}{2}}}{[W+Bw]^{(n-1)/2}}dw\,, \qquad (13-4)$$

محاسبه خواهد شد.

ب- محاسبه احتمال یک افراز مشخص با فرض ثابت بودن تعداد بلوک

برای محاسبه احتمال یک افراز مشخص، همانند قبل (بخش ۴-۱-۱) ابتدا باید احتمال هر بلوک $P(s_j)$ را محاسبه کرد که یائو (۱۹۸۴) همان احتمال بلوک محاسبه‌شده در رابطه (۴-۳) را به عنوان احتمال هر بلوک در حالت نرمال پیشنهاد می‌کند. به این ترتیب $P(\rho = \rho_j)$ همانند رابطه (۴-۴) محاسبه می‌گردد، یعنی

$$P(\rho = \rho_j) = \int_{\circ}^{1} p^{b-1}(1-p)^{n-b}\pi(p)dp\,.$$

یائو برای متغیر تصادفی p توزیع پیشین $\pi(p)$ را به صورت

$$\pi(p) = \frac{1}{p_\circ} \qquad \circ \le p \le p_\circ\,, \qquad (14-4)$$

تعریف می‌کند. در نتیجه خواهیم داشت

$$P(\rho = \rho_j) = \frac{1}{p_\circ}\int_\circ^{p_\circ} p^{b-1}(1-p)^{n-b}\,dp\,.$$

ج- تابع احتمال تعداد بلوک

در این‌جا احتمال‌های $f(x_1,\dots,x_n|\rho_j)$ و $P(\rho_j)$ برای تعداد بلوک مشخص (b) که برای مدل نرمال در قسمت‌های الف و ب محاسبه شدند را برای محاسبه احتمال تعداد بلوک در این مدل به‌کار می‌بریم. با توجه به رابطه (۴-۷) داریم

$$P(B\ =\ b|x_1,\dots,x_n\ ) \propto \sum\nolimits_{\rho \epsilon C_b}\ \left[\int_\circ^{w_\circ}\frac{w^{\frac{b}{2}}}{(W+Bw)^{\frac{n-1}{2}}}\,dw\right]\left[\int_\circ^{p_\circ} p^{b-1}(1-p)^{n-b}\,dp\right],$$

که در آن، C_b مجموعه کل افرازهای ممکن b-بلوکی است. واضح است که با مشخص شدن تعداد بلوک، تعداد نقاط تغییر نیز مشخص می‌شود. در بخش بعد نحوه محاسبه این که با تعداد بلوک مشخص شده کدام نقاط ممکن است نقاط تغییر باشند، آورده شده است.

د- احتمال نقطه تغییر بودن هر مشاهده با فرض ثابت بودن تعداد بلوک

اگر b، تعداد بلوک‌های تعیین شده باشد، می‌توان احتمال این که مشاهده kام بتواند نقطه تغییر باشد را با استفاده از رابطه (۴-9) و قسمت‌های الف و ب به صورت زیر تعیین کرد:

$$P(A_k|x_1,\dots,x_n\ )\ \propto \sum\nolimits_{\rho_j \epsilon C_{k,b}}\left[\int_\circ^{w_\circ}\frac{w^{b/2}}{(W+Bw)^{(n-1)/2}}\,dw\right]\left[\int_\circ^{p_\circ} p^{b-1}(1-p)^{n-b}\,dp\right],$$

که در آن، $C_{k,b}$ مجموعه همه افرازهای b-بلوکی است که در همه آن‌ها k امین داده، نقطه تغییر است. با محاسبه این احتمال‌ها، b-۱ تا از محتمل‌ترین آن‌ها مکان نقاط تغییر را مشخص می‌کنند.

ه- تشخیص نقطه تغییر از دیدگاه تغییر در میانگین پارامتر توزیع

با توجه به مطالب گفته شده در بخش ۴-۱-۵، در خصوص تشخیص نقطه تغییر از دیدگاه تقسیم‌بندی پارامترها و تشخیص تغییر در پارامتر توزیع، مجموع مقدار مورد انتظار پارامتر توزیع روی بلوک‌های مختلفی از افراز، که شامل آن پارامتر هستند محاسبه شده و به عنوان مقدار مورد انتظار پارامتر توزیع در آن نقطه زمانی معرفی می‌گردد. بنابراین برای محاسبه مقدار مورد انتظار پارامتر میانگین $(1 \le k \le n)$ μ_k به شرط مشاهدات $x_1, \dots, x_n$، طبق رابطه (۱۰-۴)، داریم:

$$E(\mu_k \,|x_1, \dots, x_n) = \sum\nolimits_{\rho} E(\mu_k \,|x_1, \dots, x_n, \rho) r_{i\,j}(X), \qquad (15-4)$$

که در آن، ρ کل افرازهایی است که شامل بلوک مشخص s_j هستند. با توجه به μ_k محاسبه‌شده در بخش ۴-۱-۶، $\mu_k = (1-w)\bar{X}_j + w\mu_0$، و با جایگذاری آن در رابطه (۱۵-۴) داریم

$$E(\mu_k \,|x_1, \dots, x_n) = \sum\nolimits_{\rho} E\left((1-w)\bar{X}_j + w\mu_0 \middle| x_1, \dots, x_n, \rho\right) r_{i\,j}(X).$$

برای محاسبه عبارت فوق ابتدا عبارت‌های $E(\mu_0|w, x_1, \dots, x_n \,, \rho)$ و $E(w|x_1, \dots, x_n, \rho)$ محاسبه می‌گردند که در زیر آمده‌اند:

$$E(\mu_0|w, x_1, \dots, x_n \,, \rho) = \int_{-\infty}^{+\infty} \mu_0 \, f(\mu_0|w, x_1, \dots, x_n \,, \rho) \, d\mu_0.$$

برای محاسبه $f(\mu_0|w, x_1, \dots, x_n \,, \rho)$ داریم

$$f(\mu_0|w, x_1, \dots, x_n \,, \rho) = \frac{f(w, \mu_0, x_1, \dots, x_n, \rho)}{f(w, x_1, \dots, x_n, \rho)} = \frac{f(w)f(\mu_0)f(\rho)f(x_1, \dots, x_n|w, \mu_0, \rho)}{f(w)f(\rho)f(x_1, \dots, x_n|\rho, w)}$$

$$= \frac{w^{\frac{b}{2}}}{[W + Bw + wn(\mu_0 - \bar{X})^2]^{n/2}} \; \frac{(W + Bw)^{(n-1)/2}}{w^{(b-1)/2}}.$$

در نتیجه

$$E(\mu_0|w, x_1, \dots, x_n \,, \rho) = \int_{-\infty}^{+\infty} \mu_0 \, \frac{w^{\frac{b}{2}}}{[W + Bw + wn(\mu_0 - \bar{X})^2]^{n/2}} \cdot \frac{(W + Bw)^{(n-1)/2}}{w^{(b-1)/2}} \, d\mu_0$$

$$= \frac{1}{2wn} \int_{-\infty}^{+\infty} 2wn\, (\mu_{\circ} + \bar{X} - \bar{X}) \frac{w^{\frac{1}{2}}\, (W + Bw)^{\frac{n-1}{2}}}{[W + Bw + wn(\mu_{\circ} - \overline{X})^2]^{\frac{n}{2}}}\ d\mu_{\circ}$$

$$= w^{\frac{1}{2}}\, (W + Bw)^{\frac{n-1}{2}}\ \frac{u^{-\frac{n}{2}+1}}{-\frac{n}{2}+1} |_{-\infty}^{+\infty} + \bar{X} \,.\frac{(W + Bw)^{\frac{n-1}{2}}}{w^{\frac{b-1}{2}}} . \frac{w^{\frac{b-1}{2}}}{(W + Bw)^{\frac{n-1}{2}}} = \bar{X}\,.$$

برای محاسبه عبارت دوم داریم

$$E(w|x_1, \dots, x_n, \rho) = \int_{\circ}^{w_{\circ}} w\, f(w|x_1, \dots, x_n, \rho)\, dw\,.$$

برای محاسبه $f(w|x_1, \dots, x_n, \rho)$ داریم

$$f(w|x_1, \dots, x_n, \rho) = \frac{f(w, x_1, \dots, x_n, \rho)}{f(x_1, \dots, x_n, \rho)} = \frac{f(w) f(\rho) f(x_1, \dots, x_n|w, \rho)}{f(\rho) f(x_1, \dots, x_n|\rho)}$$

$$= \frac{\frac{1}{w_{\circ}}\ \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}}{\int_{\circ}^{w_{\circ}} \frac{1}{w_{\circ}}\ \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}\, dw}\,.$$

بنابراین خواهیم داشت:

$$E(w|x_1, \dots, x_n, \rho) = \int_{\circ}^{w_{\circ}} w\ \frac{\frac{1}{w_{\circ}}\ \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}}{\int_{\circ}^{w_{\circ}} \frac{1}{w_{\circ}}\ \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}\, dw}\ dw$$

$$= \frac{1}{\int_{\circ}^{w_{\circ}} \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}\, dw} \int_{\circ}^{w_{\circ}} w\ \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}\ dw$$

$$= \frac{\int_{\circ}^{w_{\circ}} \frac{w^{(b+1)/2}}{(W + Bw)^{(n-1)/2}}\, dw}{\int_{\circ}^{w_{\circ}} \frac{w^{(b-1)/2}}{(W + Bw)^{(n-1)/2}}\, dw} = w^{*}\,.$$

درنتیجه:

$$E(\mu_k|x_1,\dots,x_n)=\sum_{\rho}\left((1-w^*)\bar{X}_j+w^*\bar{X}\right)r_{i\,j}(X)\,.$$

به این ترتیب مقدار مورد انتظار پارامتر میانگین به ازای هر مشاهده به دست آمده و نقاط زمانی که مقدار پارامتر در آن‌ها تغییر محسوسی کرده باشد به عنوان نقاط تغییر معرفی می‌شوند. این محاسبات برای پارامتر واریانس نیز به طور مشابه انجام می‌شود. همان طور که ملاحظه شد، تعداد محاسبات مقدار مورد انتظار پارامتر ($E(\mu_k|x_1,\dots,x_n)$) و احتمال این که هر نقطه زمانی نقطه تغییر باشد ($P(A_k|x_1,\dots,x_n)$) بسیار طولانی و وقت‌گیر (از مرتبه $O(n^3)$) بوده و محاسبه آن‌ها بسیار پیچیده است. بنابراین برای کاهش تعداد محاسبات (تا مرتبه $O(n)$) و تقریب عددی آسان‌تر و دقیق‌تر، استفاده از روش‌های نمونه‌گیری و زنجیر مارکوف مونت کارلویی[1] (*MCMC*) پیشنهاد می‌گردد که در ادامه معرفی می‌شوند (بری و هارتیگن ۱۹۹۳).

۴-۱-۷ برآورد بیزی با استفاده از روش‌های نمونه‌گیری *MCMC*

۴-۱-۷-۱ روش نمونه گیری مارکف

روش نمونه‌گیری مارکوف در ابتدا از فیزیک آماری توسط هامرسلی و هند سکامب[2] (۱۹۶۴)، متروپولیس[3] و همکاران (۱۹۵۳) آغاز شد. در واقع زنجیر مارکوف دنباله‌ای از متغیرهای تصادفی است که همگی این متغیرهای تصادفی دارای فضای نمونه‌ای یکسانی هستند و هر متغیر تصادفی در یک زنجیر مارکوف، تنها به متغیر قبل از خود وابسته است. در زیر روش کلی این نمونه‌گیری و طرح نمونه‌گیری گیبز بیان شده و در قسمت بعد الگوریتم نمونه‌گیری مارکوف برای مثال بالا بیان می‌شود (بری و هارتیگن، ۱۹۹۳).

---

[1] Mrakov chain monte carlo

[2] Hammersley and Handscomb

[3] Metrpolis

۴-۱-۷-۲ روش نمونه‌گیری گیبز در حالت کلی

فرض کنید z یک نمونه متغیر تصادفی با تابع چگالی f باشد، بنابراین یک زنجیره مارکف با احتمالات انتقال $f(y|z)$ وجود دارد، که متناظر با احتمالات ایستای $f(z)$ است. زنجیره با مقدار دلخواه z_0 شروع شده و دنباله مارکف $z_1, z_2, \dots, z_m$ با احتمالات متناظر $f(y|z)$ به‌دست می‌آید.

۴-۱-۷-۳ طرح نمونه گیری گیبز

بسیاری از مسایل استنباط بیزی با استفاده از روش‌های نمونه‌گیری *MCMC* (اسمیت و جیلفند[1]، ۱۹۹۰) حل می‌شوند. مدل‌های پیچیده بیزی با استفاده از این روش به آسانی تقریب می‌شوند. فرض کنید X و Y متغیرهای تصادفی با تابع چگالی توام $f(x,y)$ باشند که دارای چگالی‌های شرطی کامل به صورت $f(x|y)$ و $f(y|x)$ هستند. حالت‌های اولیه‌ی متغیرهای X و Y به‌صورت x^0 و y^0 در نظر گرفته می‌شوند. ماتریس انتقال برای حالت‌های جدید x^i و y^i با استفاده از حالت‌های قبل x^{i-1} و y^{i-1} به صورت

$$f\left(x^i, y^i \middle| x^{i-1}, y^{i-1}\right) = f(x^i | y^{i-1}) f(y^i | x^i),$$

تولید می‌شود. چگالی توأم $f(x,y)$ یک توزیع مانا برای این ماتریس انتقال است. آن‌چه که بیان شد الگوریتم نمونه‌گیری گیبز دومرحله‌ای است. تعمیم آن به حالت چندمرحله‌ای به صورت زیر است.

فرض کنید $X = (X_1, X_2, \dots, X_n)$ یک دنباله از متغیرهای تصادفی باشد و بتوان از چگالی‌های شرطی کامل آن‌ها نمونه تولید کرد، به طوری که

$$X_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n \sim f(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n).$$

به این ترتیب الگوریتم نمونه‌گیری گیبز به صورت زیر قابل نمایش است.

---

[1] Smith and Gilfand

اگر در مرحله $t = 1,2, \dots$، بردار $x^{(t)} = (x_1^{(t)}, x_2^{(t)}, \dots, x_n^{(t)})$ را داشته باشیم. آن‌گاه نمونه به صورت زیر تولید می‌شود:

$$X_1^{(t+1)} \sim f_1\left(x_1 \middle| x_2^{(t)}, x_3^{(t)}, \dots, x_n^{(t)}\right),$$

$$X_2^{(t+1)} \sim f_2\left(x_2 \middle| x_1^{(t+1)}, x_3^{(t)}, \dots, x_n^{(t)}\right),$$

$$\vdots$$

$$X_n^{(t+1)} \sim f_n\left(x_n \middle| x_1^{(t+1)}, \dots, x_n^{(t+1)}\right).$$

با توجه به تعاریف بیان‌شده، در بخش بعد، نمونه‌گیری گیبز برای تشخیص نقطه تغییر در مدل افراز ضربی به‌کار برده می‌شود.

۴-۱-۸ طرح نمونه‌گیری گیبز برای مدل افراز ضربی

با توجه به توضیحات ارایه شده، اگر n متغیر تصادفی $U_1, U_2, \dots, U_n$ داشته باشیم به طوری‌که تشکیل یک زنجیر مارکوف دهند و بتوان با احتمال انتقال مثبت از یک حالت به حالت دیگر رفت، آن‌گاه با استفاده از توزیع $f(U_1, U_2, \dots, U_n)$ و مقادیر اولیه $U_1^{\circ}, U_2^{\circ}, \dots, U_n^{\circ}$، می‌توان یک نمونه تولید کرد. به‌طور مثال، نمونه kام در مرحله iام با چگالی $f(U_k | U_1^{(i)}, \dots, U_{k-1}^{(i)}, U_{k+1}^{(i-1)}, \dots, U_n^{(i-1)})$ تولید می‌شود. حال اگر فرض کنید $\rho = (U_1, U_2, \dots, U_n)$ بزرگ‌ترین افراز ممکن مجموعه مشاهدات $x_1, \dots, x_n$ باشد، به طوری‌که هر مشاهده درون یک بلوک قرار گیرد و با توجه به این که هر مشاهده می‌تواند نقطه تغییر باشد، U_i دو مقدار $U_i = 1$، در صورتی‌که $\theta_i = \theta_{i+1}$ باشد، یعنی مشاهده i ام نقطه تغییر نیست، و $U_i = ۰$، در صورتی‌که $\theta_i \neq \theta_{i+1}$ باشد، یعنی مشاهده i ام نقطه تغییر است، را انتخاب می‌کند. یعنی

$$U_i = \begin{cases} 1 & \theta_i = \theta_{i+1} \\ \circ & \theta_i \neq \theta_{i+1} \end{cases},$$

که در آن $i = 1, \dots, n-1$ زیرا $U_n = 1$ است. با توجه به این تعریف، اگر افراز تصادفی ρ را یک بردار $(U_1, \dots, U_{n-1})$ در نظر بگیریم، آن‌گاه با استفاده از نمونه‌گیری گیبز بردار $(U_1^s, \dots, U_{n-1}^s)$ برای $s \geq 1$ با مقدار اولیه $(U_1^0, \dots, U_{n-1}^0)$ تولید می‌شود. به عنوان مثال عنصر r ام در مرحله s ام بردار یعنی U_r^s به‌صورت

$$U_r | U_1^{(s)}, \dots, U_{r-1}^{(s)}, U_{r+1}^{(s-1)}, \dots, U_{n-1}^{(s-1)}: x_1, \dots, x_n \,,$$

تولید می‌شود، که در آن $r = 1, \dots, n-1$. به منظور تولید نمونه‌های بالا کافی است نسبت زیر محاسبه شود:

$$R_r = \frac{P(U_r = 1 | U_1^{(s)} = u_1, \dots, U_{r-1}^{(s)} = u_{r-1}, U_{r+1}^{(s-1)} = u_{r+1}, \dots, U_{n-1}^{(s-1)} = u_{n-1}: x_1, \dots, x_n)}{P(U_r = 0 \, | U_1^{(s)} = u_1, \dots, U_{r-1}^{(s)} = u_{r-1}, U_{r+1}^{(s-1)} = u_{r+1}, \dots, U_{n-1}^{(s-1)} = u_{n-1}: x_1, \dots, x_n)}.$$

معیار انتخاب $U_i^{(s)}$ برای $i = 1, \dots, n-1$ در زیر بیان شده که در آن $u \sim u(0,1)$ است

$$U_i^{(s)} = \begin{cases} 1 & R_r \geq \dfrac{1-u}{u} \\ 0 & \text{باقی نقاط} \end{cases} .$$

همان طور که قبلا گفته شد، اگر p دارای توزیع پیشین $\pi(p)$ باشد، آن‌گاه نسبت بالا به صورت زیر محاسبه می‌شود:

$$R_r = \frac{f_{[xy]}\big(X_{[xy]}\big) \;\left[\int_0^{p_0} p^{b-2}(1-p)^{n-b+1} dp\right]}{f_{[xr]}\big(X_{[xr]}\big)\, f_{[ry]}\big(X_{[ry]}\big)\left[\int_0^{p_0} p^{b-1}(1-p)^{n-b} dp\right]},$$

که در آن، x و y برای $i \epsilon \{1, \dots, r-1\}$ به صورت زیر تعریف می‌شوند:

$$x = \begin{cases} \max\left\{i,\ \ s.t{:}0 < i < r \ , U_i^{(s)} = 0\right\}, & U_i^{(s)} = 0 \ , \\ 0\,, & \text{باقی جاها} \end{cases}$$

و

$$y=\begin{cases}\min\left\{i,\ \ s.t{:}\, r<i<n\ ,U_i^{(s-1)}=\circ\right\}, & U_i^{(s-1)}=\circ\\ \circ\,, & \text{باقی جاها}\end{cases}$$

با توجه به مطالب گفته‌شده، این نسبت در حالت نرمال به صورت زیر محاسبه می‌شود:

$$R_r=\frac{\left[\int_\circ^{w_\circ}\frac{w^{(b-2)/2}}{(W_1+B_1w)^{(n-1)/2}}dw\right]\left[\int_\circ^{p_\circ}p^{b-2}(1-p)^{n-b+1}dp\right]}{\left[\int_\circ^{w_\circ}\frac{w^{(b-1)/2}}{(W_\circ+B_\circ w)^{(n-1)/2}}dw\right]\left[\int_\circ^{p_\circ}p^{b-1}(1-p)^{n-b}dp\right]}\ .$$

انتگرال‌های صورت و مخرج کسر، انتگرال بتای بریده‌شده می‌باشد که بسط آن به صورت

$$\propto\left(\frac{W_\circ}{W_1}\right)^{\frac{n-b-2}{2}}\left(\frac{B_\circ}{B_1}\right)^{\frac{b+1}{2}}\sqrt{\frac{W_1}{B_1}}\left[\frac{\int_\circ^{\frac{B_1p_O/W_1}{1+B_1p_O/W_1}}p^{(b+2)/2}(1-p)^{(n-b-3)/2}dp}{\int_\circ^{\frac{B_\circ p_O/W_\circ}{1+B_\circ p_O/W_\circ}}p^{(b+1)/2}(1-p)^{(n-b-2)/2}dp}\right]\cdot\frac{\left[\int_\circ^{p_\circ}p^{b}(1-p)^{n-b-1}dp\right]}{\left[\int_\circ^{p_\circ}p^{b-1}(1-p)^{n-b}dp\right]},$$

است. با توجه به مطالبی که در این زیربخش گفته شد، به راحتی می‌توان احتمال نقطه تغییر بودن هر مشاهده را محاسبه کرد. قابل ذکر است تعداد نمونه‌ها بین ۱۰۰۰۰ تا ۲۰۰۰۰ انتخاب می‌شود. با استفاده از این روش، حجم محاسبات بسیار کاهش می‌یابد.

۴-۲ تشخیص برخط نقطه تغییر بیزی (BOCPD)

به گفته آدامز و مک کی (۲۰۰۷)، بیشتر مقاله‌های بیزی بر مساله تقسیم‌بندی بازنگرانه (استنباط واخط) متمرکز شده‌اند. همان‌طور که گفته شد استفاده از استنباط واخط برای تشخیص نقاط تغییر دارای محدودیت زمانی است. بنابراین استفاده از استنباط برخط مطرح می‌شود. تشخیص برخط نقطه تغییر بیزی در مدل سازی و پیش‌بینی سری‌های زمانی در زمینه‌های کاربردی مانند امور مالی, زیست شناسی، رباتیک و غیره مفید است. آدامز و مک کی (۲۰۰۷)، روشی برای استنباط‌های دقیق برخط سری زمانی پیشنهاد کردند، که محدودیت‌ها را حذف کرده و در آنها پارامترهای مدل قبل و بعد از نقطه تغییر مستقل هستند. محققانی مانند ترنر و راسموسن [1]، تورنر (۲۰۱۰)، اندرسون[2] (۲۰۰۸)، ویلسون[3] و همکاران (۲۰۱۰)، این روش را در زمینه‌های مختلف علوم استفاده نمودند.

۴-۲-۱ تشخیص برخط تغییر

فرض کنید x_t $(t = 1, \dots, n)$ دنباله‌ای از متغیرهای تصادفی مشاهده‌شده با تابع چگالی شرطی $p_\theta(x_t|x_{t-1}, \dots, x_1)$ باشد. اگر $t = t_i$ که $i = 1, \dots, n$ زمان نامعلوم تغییر باشد، آن‌گاه پارامتر چگالی شرطی قبل از زمان تغییر $\theta = \theta_0$ و بعد از تغییر $\theta = \theta_1$ است.

مساله برخط تغییر، تشخیص رخ دادن تغییر در کوتاه‌ترین زمان ممکن با نرخ ثابتی از اعلان خطرهای اشتباه قبل از زمان تغییر t_i است. در برآورد زمان تغییر t_i، برآورد پارامترهای θ_0 و θ_1 مورد نیاز نیستند اما ممکن است این پارامترها در برخی از الگوریتم‌های تشخیص تغییر مورد استفاده قرار گیرند. در این‌جا این سوال مطرح خواهد شد که در تشخیص برخط تا چه موقع باید به دنبال نقاط

---

[1] Turner and Rasmussen
[2] Anderson
[3] Wilson

تغییر گشت؟ در پاسخ به این سوال می‌توان گفت که در چهارچوب الگوریتم برخط، قانون توقف برای تشخیص تغییر به صورت

$$t_a = \inf\{\, n: g_n(x_1, \dots, x_n) \ge \lambda\},$$

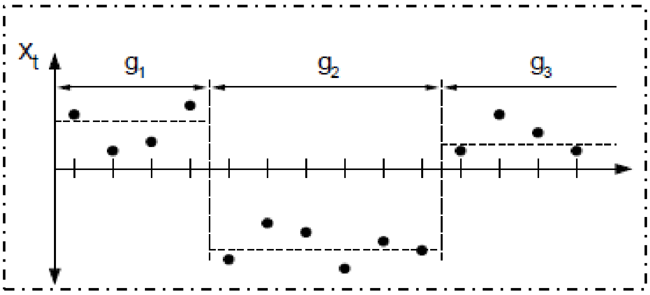
است، که در آن λ مقدار آستانه و $(g_n)_{n\ge1}$ خانواده‌ای از توابع با n مشاهده است. t_a زمان اعلان خطر است به طوری‌که اگر $t_a = n$ باشد، مشاهده نمونه تا زمان n کافی است (باسسیویلی و نیکفوروم، ۱۹۹۳).

۴-۲-2 الگوریتم تشخیص برخط نقطه تغییر

فرض کنید $x_1, \dots, x_n$ یک سری مشاهده باشد که دارای نقاط تغییر نامعلومی هستند. برای تشخیص و تعیین این نقاط تغییر، ابتدا با استفاده از افرازهای ضربی که در بخش قبل بیان شد، سری مشاهدات تقسیم‌بندی شود. هر افراز ρ، سری را به بلوک‌های مجزا تقسیم می‌کند که این بلوک‌ها مستقل بوده و دارای توزیع احتمال $p(x_t|\eta_\rho)$ هستند. پارامترهای η_ρ نیز متغیرهای تصادفی مستقل و هم‌توزیع می‌باشند. مشاهدات درون بلوک [a,b] یعنی مشاهداتی که بین زمان‌های a وb قرار می‌گیرند را با $x_{a:b}$ نشان می‌دهیم. همان طور که گفته شد بلوک‌ها مستقل هستند ولی پارامتر توزیع و گاهی خود توزیع برای هر بلوک متفاوت است. یعنی با انتقال از یک بلوک به بلوک دیگر، تغییری در پارامتر توزیع رخ خواهد داد. بنابراین می‌توان یک توزیع احتمال پیشین گسسته برای بلوک‌ها یا فاصله بین نقاط تغییر مشخص، متصور شد که در این‌جا آن را با نماد $P_{gap}(B_i)$، نشان می‌دهیم که B_i نشان-دهنده بلوک iام است. هر بلوک را یک گشت (r) نامیده و فاصله‌ی بین نقطه x_t از آخرین نقطه تغییر را یک طول گشت تعریف می‌کنیم و با نماد r_t نمایش می‌دهیم. مجموعه نقاط داخل هر بلوک که متعلق به طول گشت r_t هستند را با نماد $x_t^{(r)}$ نشان می‌دهیم. بنابراین نقاط ابتدایی هر بلوک یا

هر طول گشت، نقاط تغییر هستند. برای درک بیشتر مفهوم طول گشت و نقاط تغییر، مثالی در زیر آورده شده است.

مثال ۴-۱- فرض کنید در یک سری زمانی با ۱۴ مشاهده، سطح میانگین در مشاهده پنجم (x_5) و یازدهم (x_{11}) تغییر یافته است.



شکل ۴-۱: نمودار تقسیم‌بندی سری به طول ۱۴ با دو تغییر در میانگین داده

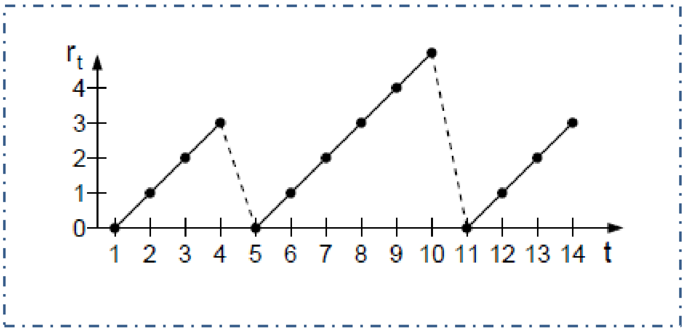
اگر این سری را به سه بلوک B_1، B_2 و B_3 (شکل 4-1) تفکیک نماییم، طول گشت‌ها در هر بلوک عبارتند از:

$x_t^{(1)} = [x_1, x_2, x_3, x_4],$ $r_1 = 0,$ $r_2 = 1,$ $r_3 = 2,$ $r_4 = 3,$

$x_t^{(2)} = [x_5, x_6, x_7, x_8, x_9, x_{10}],$ $r_5 = 0,$ $r_6 = 1,$ $r_7 = 2,$ $r_8 = 3,$ $r_9 = 4,$
$r_{10} = 5,$

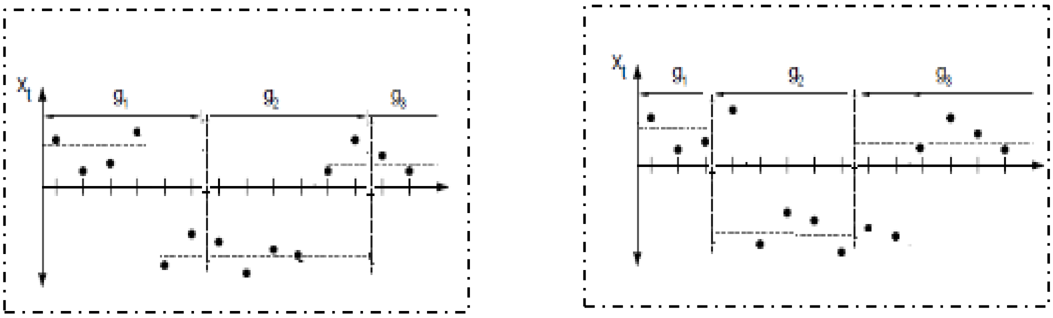
$x_t^{(3)} = [x_{11}, x_{12}, x_{13}, x_{14}],$ $r_{11} = 0,$ $r_{12} = 1,$ $r_{13} = 2,$ $r_{14} = 3,$

مقادیر طول گشت در شکل ۴-۲ با دایره‌های توپر نمایش داده شده است.



شکل ۴-۲: نمودار طول گشت سری با توجه به داده‌های تقسیم‌بندی شده‌ی شکل ۴-۱

همان‌طور که ملاحظه می‌شود در نقاطی که سطح میانگین داده‌ها تغییر کرده، مقدار r_t برابر صفر است (یعنی نقاط r_5 و r_{11}) که این نقاط به عنوان نقاط تغییر شناسایی می‌شوند. در این مثال، تقسیم‌بندی و طول گشت برای نقاط تغییر معلوم به دست آمده و رسم شده است، ولی در عمل نقاط تغییر نامعلوم بوده و ما به دنبال تشخیص و تعیین این نقاط هستیم. واضح است که با در نظر گرفتن افرازها یا بلوک‌بندی‌های متفاوت دیگر، مقادیر r_t تغییر خواهند کرد. به عنوان مثال دو افراز (الف) و (ب) نشان داده شده در شکل ۴-۳ را در نظر بگیرید.



(الف) (ب)

شکل ۴-۳: نمودار زمانی ۱۴ مشاهده مثال ۴-۱ با دو بلوک‌بندی متفاوت

طول گشت‌های این دو افراز به صورت زیر به دست می‌آیند. در حالت (الف) داریم:

$x_t^{(1)} = [x_1, x_2, x_3],$ $r_1 = 0,$ $r_2 = 1,$ $r_3 = 2,$

$x_t^{(2)} = [x_4, x_5, x_6, x_7, x_8],$ $r_4 = 0,$ $r_5 = 1,$ $r_6 = 2,$ $r_7 = 3,$ $r_8 = 4$

$x_t^{(3)} = [x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}],$ $r_9 = 0,$ $r_{10} = 1,$ $r_{11} = 2,$ $r_{12} = 3,$ $r_{13} = 4,$ $r_{14} = 5.$

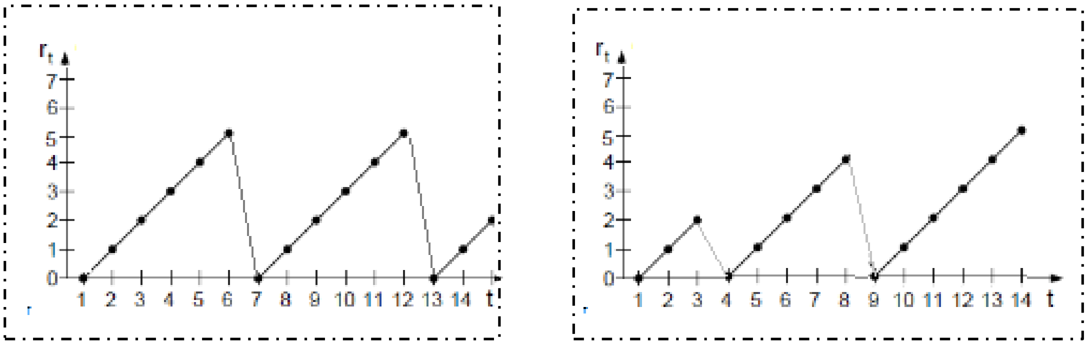
و در حالت (ب) داریم:

$x_t^{(1)} = [x_1, x_2, x_3, x_4, x_5, x_6],\qquad r_1 = 0,\quad r_2 = 1,\quad r_3 = 2,\quad r_4 = 3,\quad r_5 = 4,\quad r_6 = 5,$

$x_t^{(2)} = [x_7, x_8, x_9, x_{10}, x_{11}, x_{12}],\qquad r_7 = 0,\quad r_8 = 1,\quad r_9 = 2,\quad r_{10} = 3,\quad r_{11} = 4,\quad r_{12} = 5,$

$x_t^{(3)} = [x_{13}, x_{14}],$,,,,,,,,,,,,,,, $r_{13} = 0,\quad r_{14} = 1.$

نمودار مربوط به مقادیر به دست آمده در شکل ۴-۴ نمایش داده شده است.



(ب) (الف)

شکل ۴-۴: (الف- ب) نمودار طول گشت حاصل از بلوک‌بندی (الف-ب) در شکل ۴-۳

اگر صفر بودن مقدار r_t را به معنای این بدانیم که x_t یکی از نقاط تغییر است، افرازهای متفاوت، القاکننده نقاط تغییر متفاوت بوده و می‌توان گفت هر مشاهده می‌تواند به عنوان نقطه تغییر شناسایی شود. به این ترتیب به ازای هر مشاهده x_t دو انتخاب وجود دارد: یا مشاهده در ابتدای طول گشت قرار گرفته ($r_t =$ ۰) و به عنوان نقطه تغییر شناسایی می‌شود، یا مقدار آن یک واحد از طول گشت قبلی بیشتر است ($r_t = r_{t-1} + 1$) و نقطه تغییر تلقی نمی‌گردد. برای مثال، در خصوص بررسی این که آیا x_5 نقطه تغییر است، دو حالت وجود دارد: $r_5 =$ ۰ یا $r_5 = r_4 + 1$ (نمودار (الف) و (ب) شکل ۴-۴). در این‌جا این سوال مطرح خواهد شد که کدام یک از این دو حالت محتمل‌تر است.

پس باید $P(r_5 = 0|x_{1:5})$ و $P(r_5 = r_4 + 1|x_{1:5})$ محاسبه شوند، ولی $P(r_5 = r_4 + 1|x_{1:5})$ و

$P(r_5 = r_4 + 1|x_{1:5})$ به طول گشت r_4 وابسته بوده و در این‌جا حالت مشابهی برای r_4 پیش می‌آید که آیا $r_4 = ۰$ است یا $r_4 = r_3 + 1$. بنابراین محاسبه $P(r_4 = 0|x_{1:4})$ و $P(r_4 = r_3 + 1|x_{1:4})$ نیز مورد نیاز است. به همین ترتیب برای محاسبه $P(r_4 = r_3 + 1|x_{1:4})$، نیاز داریم تا احتمال‌های $P(r_3 = 0|x_{1:3})$ و $P(r_3 = r_2 + 1|x_{1:3})$ را محاسبه نماییم. پر واضح است که این روند تا محاسبه $P(r_1 = 0|x_1)$ ادامه خواهد یافت. شکل ۴-۵ نمودار تمامی طول گشت‌های ممکن مشاهدات اول تا چهارم را نشان می‌دهد به طوری که در هر ستون می‌توان نحوه به وجود آمدن گشت‌های مختلف برای هر مشاهده را ملاحظه نمود. همچنین وابستگی احتمال‌های طول گشت‌ها، به صورت نمادین در شکل ۴-۶ نمایش داده شده است. با توجه به نمودارهای شکل ۴-۵ در خصوص ستون سوم، مشاهده می‌شود که مقادیر ممکن r_3 عبارتند از $r_3 = \begin{cases} 0 \\ r_2 + 1 = \begin{cases} 1 \\ 2 \end{cases} \end{cases}$ همان طور که مشاهده می‌شود:

- $r_3 = ۰$ در دو حالت رخ می‌دهد که یکی وابسته به $r_2 = ۰$ و دیگری وابسته به $r_2 = 1$ است.
- تعداد دفعات رخداد $r_3 = 1$ یکی است که وابسته به $r_2 = ۰$ است.
- تعداد دفعات رخداد $r_3 = 2$ یکی است که وابسته به $r_2 = 1$ است.

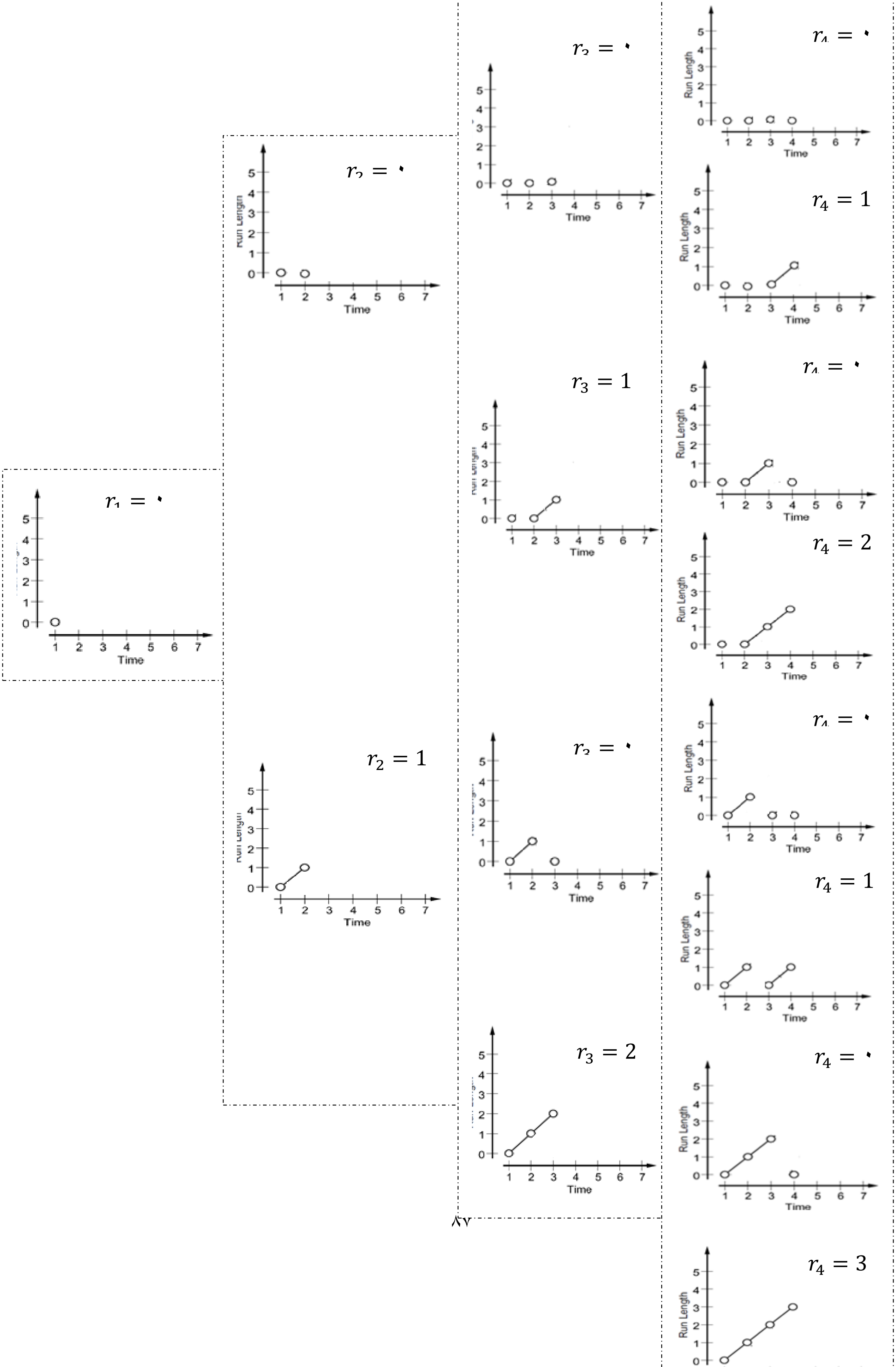
در خصوص ستون چهارم، مقادیر ممکن r_4 عبارتند از $r_4 = \begin{cases} 0 \\ r_3 + 1 = \begin{cases} 1 \\ 2 \\ 3 \end{cases} \end{cases}$ به‌طوری‌که:

- $r_4 = ۰$ یا از $r_3 = ۰$ نتیجه شده است که مبدا آن $r_2 = ۰$ یا $r_2 = 1$ است.

  یا از $r_3 = 1$ نتیجه شده است که مبدا آن $r_2 = ۰$ بوده

  یا $r_4 = ۰$ از $r_3 = 2$ نتیجه شدهاست که مبدا آن $r_2 = 1$ بوده است.
- $r_4 = 1$ یا از $r_3 = ۰$ نتیجه شده است که مبدا آن $r_2 = ۰$ یا $r_2 = 1$ بوده است.
- $r_4 = 2$ از $r_3 = 1$ نتیجه شده است که مبدا آن از $r_2 = ۰$ است.

- $r_4 = 3$ از $r_3 = 2$ نتیجه شده است که مبدا آن از $r_2 = 1$ است.



شکل ۴-۵: نمودار طول گشت‌های متفاوت تا زمان $t = 4$

$$P(r_1 = 0|x_1)\begin{cases} P(r_2 = 0|x_{1:2})\begin{cases} P(r_3 = 0|x_{1:3})\begin{cases} P(r_4 = 0|x_{1:4}) \\ P(r_4 = 1|x_{1:4}) \end{cases} \\ P(r_3 = 1|x_{1:3})\begin{cases} P(r_4 = 0|x_{1:4}) \\ P(r_4 = 2|x_{1:4}) \end{cases} \end{cases} \\ P(r_2 = 1|x_{1:2})\begin{cases} P(r_3 = 0|x_{1:3})\begin{cases} P(r_4 = 0|x_{1:4}) \\ P(r_4 = 1|x_{1:4}) \end{cases} \\ P(r_3 = 2|x_{1:3})\begin{cases} P(r_4 = 0|x_{1:4}) \\ P(r_4 = 3|x_{1:4}) \end{cases} \end{cases} \end{cases}$$

شکل ۴-۶: نمودار احتمال‌های طول گشت‌های شکل ۴-۵ تا زمان $t = 4$

با توجه به نمودارهای طول گشت (شکل ۴-۵) و نمودارهای احتمال‌های (شکل ۴-۶می‌توان گفت که هر مشاهده می‌تواند اول طول گشت قرار گرفته و نقطه تغییر باشد. بنابراین برای یافتن نقطه تغییر، باید بررسی از مشاهده ابتدایی صورت گیرد. در ابتدا احتمال $P(r_1 = 0|x_1)$ محاسبه می‌شود. در این‌جا ممکن است این سوال مطرح شود که آیا نقطه ابتدایی مشاهدات همواره اول طول گشت قرار گرفته و در نتیجه $r_1 =$ ۰، و x_1 به عنوان نقطه تغییر شناسایی می‌شود؟ با مطرح شدن این سوال، نیاز به داشتن شرایط مرزی برای ابتدای سری احساس می‌شود که در بخش ۴-۲-۴ به تفصیل بیان شده است. بعد از محاسبه احتمال $P(r_1 = 0|x_1)$، به بررسی نقطه تغییر بودن مشاهده x_2، می‌پردازیم. با توجه به نمودارهای شکل‌های ۴-۵ و ۴-۶، ملاحظه می‌گردد که باید دو احتمال $P(r_2 = 0|x_{1:2})$ و $P(r_2 = 1|x_{1:2})$ محاسبه شوند و هر حالت که محتمل‌تر باشد رخ می‌دهد. فرض کنید همانند بلوک‌بندی شکل ۴-۲، $P(r_2 = 1|x_{1:2})$ محتمل‌تر بوده و در نتیجه x_2 نقطه تغییر نیست.

برای بررسی آیا x_3، نقطه تغییر است، با توجه به ستون سوم نمودارهای r_t (شکل4-۵) و احتمال‌های متناظر هر طول گشت (شکل4-۶)، تمام احتمال‌های $P(r_3 = 0|x_{1:3})$، $P(r_3 = 1|x_{1:3})$ و $P(r_3 = 2|x_{1:3})$ که وابسته به $P(r_2 = 0|x_{1:2})$ و $P(r_2 = 1|x_{1:2})$ هستند، محاسبه می‌شوند. ولی از آن‌جا که در مرحله قبل، احتمال $P(r_2 = 1|x_{1:2})$ بیشتر بوده، لذا احتمال دو حالت $P(r_3 = 0|x_{1:3})$ و $P(r_3 = 2|x_{1:3})$، که وابسته به $r_2 = 1$ هستند، محتمل‌تر می‌باشند. با محاسبه این دو احتمال ملاحظه می‌شود که مقدار $P(r_3 = 2|x_{1:3})$ بیشتر بوده و به این ترتیب x_3 نیز نقطه تغییر نمی‌باشد.

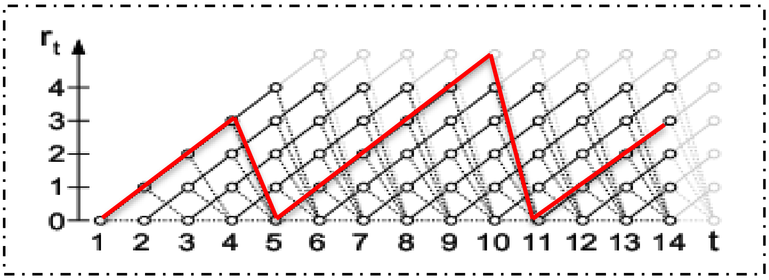
روند بررسی نقطه تغییر بودن مشاهده x_4 همانند قبل است. بنابراین باید احتمال‌های $P(r_4 = 0|x_{1:4})$، $P(r_4 = 1|x_{1:4})$، $P(r_4 = 2|x_{1:4})$ و $P(r_4 = 3|x_{1:4})$ محاسبه شوند و با توجه به این‌که در مرحله قبل احتمال $P(r_3 = 2|x_{1:3})$ دارای بیشترین مقدار بود، به طور قطع احتمال $P(r_4 = 3|x_{1:4})$ و $P(r_4 = 0|x_{1:4})$ محتمل‌تر می‌شوند که با توجه به تقسیم‌بندی مثال ۴-۱ (شکل۴-۲)، $r_4 = 3$ دارای بالاترین احتمال بوده و مشاهده x_4 نقطه تغییر نیست.

برای مشاهده x_5 احتمال‌های $P(r_5 = 0|x_{1:5})$، $P(r_5 = 1|x_{1:5})$، $P(r_5 = 2|x_{1:5})$، $P(r_5 = 3|x_{1:5})$ و $P(r_5 = 4|x_{1:5})$ محاسبه می‌شوند، و از آن‌جا که در مرحله قبل طول گشت $r_4 = 3$ رخ داد، بنابراین $P(r_5 = 0|x_{1:5})$ و $P(r_5 = 4|x_{1:5})$ محتمل‌تر هستند، و همان طور که مشاهده می‌شود احتمال $P(r_5 = 0|x_{1:5})$ دارای بیشترین مقدار بوده و مشاهده پنجم دارای طول گشت $r_t =$ ۰ بوده و به عنوان نقطه تغییر تشخیص داده می‌شود.

این روند برای هر یک از مشاهدات انجام شده و نقاط تغییر مشخص می‌شوند. برای نمایش نقاط تغییر روی نمودار، می‌توان تمام طول گشت‌های ممکن r_t به ازای هر مشاهده را، روی یک نمودار رسم کرد.

آن‌گاه مقادیر احتمال هر طول گشت به ازای هر مشاهده را که در بالا تا مشاهده x_5 نمایش داده شده است، با یک طیف رنگی مشخص کرده و روی نمودار طول گشت r_t نشان داد.

فرض کنید تمام طول گشت‌های ممکن مثال ۴-۱ (شکل ۴-۱) به صورت یک نمودار شبکه‌ای در شکل ۴-۷ نشان داده شده باشد. طول گشت‌هایی که دارای بالاترین احتمال بودند (شکل ۴-۲) با خطوط مورب تیره مشخص شده، و نقاط تغییر را به راحتی می‌توان در نمودار ملاحظه کرد (پکویت[1]، ۲۰۰۸).



شکل ۴-۷: نمودار شبکه‌ای طول گشت‌های ممکن مثال ۴-۱ با ۱۴ مشاهده، نقاط تغییر طول گشت‌هایی که دارای بیشترین احتمال رخداد هستند و با خطوط مورب تیره مشخص شده‌اند

برای تصور بهتر مطالب فوق، سری مثال ۱ را به صورت زیر در نظر می‌گیریم و نقاط تغییر آن (x_{11}, x_5) را با محاسبه مقادیر طول گشت تشخیص می‌دهیم.

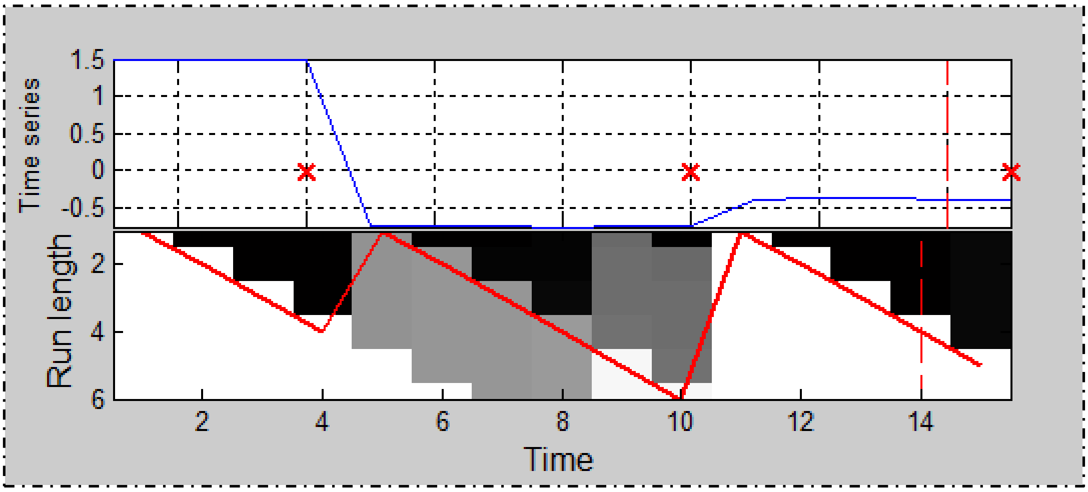
x_{14}	x_{13}	x_{12}	x_{11}	x_{10}	x_9	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1
۱٫۱	۱٫۲	۱٫۵	۱	-۱۰٫۲	-۱۰٫۱	-۱۰٫۷	-۱۰٫۳	-۱۰٫۴	-۱۰٫۱	۶۰٫۱	۶۰٫۲	۶۰٫۵	۶۰

در شکل ۴-۸ که خروجی یک نرم افزار است، مقادیر احتمال هر طول گشت و نمودار محتمل‌ترین طول گشت (خطوط شکسته قرمز) نشان داده شده است. قابل ذکر است که کد این روش برای مجموعه داده‌های که دارای مقادیر بزرگ می‌باشند از لگاریتم مشاهدات برای تعیین نقطه تغییر استفاده می‌کند. در واقع نمودار ۴-۸ معکوس نمودار ۴-۷ می‌باشد، که تمام طول گشت‌های ممکن

---

[1] Paquet

مشاهدات را در نظر گرفته، احتمال هر طول گشت را محاسبه کرده و مقدار این احتمال را با یک طیف رنگی مشکی نمایش داده است. بنابراین برای یافتن نقاط تغییر باید احتمال نقطه تغییر بودن هر مشاهده یا احتمال هر طول گشت به شرط مشاهدات یعنی $P(r_t|x_{1:t})$ محاسبه شود که هدف اصلی در روش تشخیص برخط نقطه تغییر، محاسبه این احتمال می‌باشد. نحوه محاسبه این احتمال در بخش بعدی بیان می‌شود.



شکل۴-۸: نمودار ۱۴ داده شبیه‌سازی‌شده و نمودار پایین نمودار طول گشت و نمایش احتمال آن با طیف رنگی

۴-۲-۲-۱ محاسبه احتمال طول گشت ($P(r_t|x_{1:t})$)

همان‌طور که در بالا ملاحظه شد، مساله تشخیص نقطه تغییر منجر به محاسبه‌ی احتمال طول گشت به شرط داشتن مشاهدات قبلی آن، یعنی $P(r_t|x_{1:t})$، گردید. این احتمال یک توزیع پسین بوده و محاسبه‌ی آن به صورت زیر است:

$$P(r_t|x_{1:t}) = \frac{P(r_t, x_{1:t})}{P(x_{1:t})}.$$

در این رابطه $P(r_t, x_{1:t})$ توزیع توام طول گشت و مشاهدات بوده و به صورت

$$P(r_t, x_{1:t}) = \sum_{r_{t-1}} P(r_t, x_t|r_{t-1}, x_{1:t-1}) P(r_{t-1}, x_{1:t-1})$$

$$=\sum\nolimits_{r_{t-1}} P(r_t|r_{t-1})P(x_t|r_{t-1},x_{t-1}^{(r_{t-1})})P(r_{t-1},x_{1:t-1}), \qquad (15-4)$$

محاسبه می‌شود. در رابطه (۴-۱۵)، احتمال طول گشت به ازای طول گشت‌های قبلی، یعنی $P(r_t|r_{t-1})$، به نرخ مخاطره[1] معروف است که نحوه‌ی محاسبه‌ی آن به طور مفصل در بخش ۴-۲-۲-۲ بیان می‌شود. $P(x_t|r_{t-1},x_{t-1}^{(r_{t-1})})$ یک توزیع پیشگوی زیربنایی به ازای هر مشاهده است که روی طول گشت قبلی و داده‌های متعلق به آن $(x_{t-1}^{(r_{t-1})})$ شرطی شده و به سادگی قابل محاسبه است. همچنین $P(r_{t-1},x_{1:t-1})$ توزیع توام، طول گشت‌ها و مشاهدات قبلی است که برای محاسبه $P(r_{t-1}|x_{1:t-1})$ به‌کار برده می‌شود. در واقع محاسبه توزیع توام یک الگوریتم بازگشتی است (آدامز و مک کی ۲۰۰۷).

طبق مطالب بیان‌شده در بخش ۴-۲-۲، احتمال طول گشت $P(r_t|x_{1:t})$، در واقع برای محاسبه یک توزیع احتمال پیشگوی شرطی[2]، به صورت زیر به‌کار برده می‌شود:

$$P(x_{t+1}|x_{1:t})=\sum\nolimits_{r_t} P\left(x_{t+1}\middle|r_t,x_t^{(r_t)}\right)\,P(r_t|x_{1:t}). \qquad (16-4)$$

بنابراین با توجه به رابطه (۴-۱۶) می‌توان احتمال رخداد هر مشاهده را به شرط مشاهدات قبل به‌دست آورد. ولی همان طور که در بخش قبل توضیح داده شد، ملاک اصلی برای تشخیص نقطه تغییر، محاسبه احتمال $P(r_t|x_{1:t})$ است. به طور مثال برای تشخیص رخ دادن مشاهده x_5 به شرط مشاهدات قبل، یعنی $P(x_5|x_{1:4})$ با توجه به رابطه (4-۱۶) داریم:

$$P(x_5|x_{1:4})=\sum\nolimits_{r_4} P\left(x_5\middle|r_4,x_4^{(r_4)}\right)\,P(r_4|x_{1:4})$$

$$=P(x_5|r_4=3)\,P(r_4=3|x_{1:4})+P(x_5|r_4=2)\,P(r_4=2|x_{1:4})$$

$$+\;P(x_5|r_4=1)\,P(r_4=1|x_{1:4})+P(x_5|r_4=0)\,P(r_4=0|x_{1:4}).$$

---

[1] Hazard rate

[2] Predictive distribution conditional

در محاسبات بالا برای محاسبه $k = 0,1,2,3,4$ ،$P(r_4 = k|x_{1:4})$، در مرحله چهارم، نیاز به محاسبه احتمال طول گشت‌های قبل، یعنی مراحل اول تا سوم که در زیر نمایش داده شده‌اند، داریم:

مرحله اول

$$P(r_1 = 0|x_1) = \frac{P(r_1=0,x_1)}{P(x_1)} =$$

$$\frac{P(r_1 = 0|r_\circ = 0).P(x_1|r_\circ = 0)}{P(x_1)}$$

مرحله دوم

$$\begin{cases} P(r_2 = 0|x_{1:2}) = \frac{P(r_2 = 0|r_1 = 0).P(x_2|r_1 = 0)P(r_1=0,x_1)}{P(x_{1:2})} \\ P(r_2 = 1|x_{1:2}) = \frac{P(r_2 = 1|r_1 = 0).P(x_2|r_1 = 0)P(r_1=0,x_1)}{P(x_{1:2})} \end{cases}$$

مرحله سوم

$$\begin{cases} P(r_3 = 0|x_{1:3}) = \frac{P(r_3 = 0|r_2 = 0).P(x_3|r_2 = 0)P(r_2=0,x_{1:2})}{P(x_{1:3})} \\ P(r_3 = 1|x_{1:3}) = \frac{P(r_3 = 1|r_2 = 0).P(x_2|r_2 = 0)P(r_2=0,x_{1:2})}{P(x_{1:3})} \\ P(r_3 = 0|x_{1:3}) = \frac{P(r_3 = 0|r_2 = 1).P(x_3|r_2 = 1)P(r_2=1,x_{1:2})}{P(x_{1:3})} \\ P(r_3 = 2|x_{1:3}) = \frac{P(r_3 = 2|r_2 = 1).P(x_3|r_2 = 1)P(r_2=1,x_{1:2})}{P(x_{1:3})} \end{cases}$$

مرحله چهارم

$$\begin{cases} P(\,r_4 = 0|x_{1:4}) = \frac{P(r_4 = 0|r_3 = 0).P(x_4|r_3 = 0)P(r_3=0,x_{1:3})}{P(x_{1:4})} \\ P(r_4 = 1|x_{1:4}) = \frac{P(r_4 = 1|r_3 = 0).P(x_4|r_3 = 0)P(r_3=0,x_{1:3})}{P(x_{1:4})} \\ P(r_4 = 0|x_{1:4}) = \frac{P(r_4 = 0|r_3 = 1).P(x_4|r_3 = 1)P(r_3=1,x_{1:3})}{P(x_{1:4})} \\ P(r_4 = 2|x_{1:4}) = \frac{P(r_4 = 2|r_3 = 1).P(x_4|r_3 = 2)P(r_3=1,x_{1:3})}{P(x_{1:4})} \\ P(r_4 = 0|x_{1:4}) = \frac{P(r_4 = 0|r_3 = 2).P(x_4|r_3 = 2)P(r_3=2,x_{1:3})}{P(x_{1:4})} \\ P(r_4 = 1|x_{1:4}) = \frac{P(r_4 = 1|r_3 = 0).P(x_4|r_3 = 0)P(r_3=0,x_{1:3})}{P(x_{1:4})} \\ P(r_4 = 0|x_{1:4}) = \frac{P(r_4 = 0|r_3 = 0).P(x_4|r_3 = 0)P(r_3=0,x_{1:3})}{P(x_{1:4})} \\ P(r_4 = 3|x_{1:4}) = \frac{P(r_4 = 3|r_3 = 2).P(x_4|r_3 = 2)P(r_3=2,x_{1:3})}{P(x_{1:4})} \end{cases}$$

با محاسبه احتمال‌های بالا و جایگذاری در احتمال $P(x_5|x_{1:4})$، احتمال داده مشاهده شده x_5 به شرط تمام مشاهدات قبلی به دست می‌آید. این نکته قابل ذکر است که محاسبه‌ی احتمال $P(x_{t+1}|x_{1:t})$ به شناسایی نقطه تغییر کمکی نمی‌کند و همان‌طور که گفته شد ملاک اصلی در تشخیص نقطه تغییر، محاسبه $P(r_t|x_{1:t})$ است که این احتمال وابسته به محاسبه نرخ مخاطره یعنی $P(r_t|r_{t-1})$ می‌باشد که در بخش بعد نحوه محاسبه آن بیان خواهد شدو. محاسبه‌ی تمام این احتمال‌ها وابسته به محاسبه تابع توزیع است. بنابراین انتخاب نوع تابع توزیع که در بخش ۴-۲-۴ بیان خواهد شد نیز حایز اهمیت

۴-۲-۲ محاسبه توزیع پیشین طول گشت $P(r_t|r_{t-1})$

قبلا گفته شد که قانون تصمیم در مورد وجود نقطه تغییر، رسیدن طول گشت به نقطه صفر یعنی $r_t =$ ۰ است، در غیر این صورت طول گشت افزایش می‌یابد، یعنی $r_t = r_{t-1} + 1$. در صورتی‌که $r_t =$ ۰ باشد و تغییر رخ دهد فرض می‌شود که $P(r_t|r_{t-1})$ برابر یک مقدار ثابت یا یک تابع ثابت

است. به این مقدار ثابت یا تابع، نرخ مخاطره گفته شده و به اختصار آنرا با نماد h نشان می‌دهیم. احتمال پیشین شرطی روی طول گشت، یعنی $P(r_t|r_{t-1})$، دارای یک الگوریتم محاسباتی کارا بوده و به صورت

$$P(r_t|r_{t-1}) = \begin{cases} 1 - h(r_{t-1} + 1) & r_t = r_{t-1} + 1, \\ h(r_{t-1} + 1) & r_t = 0, \\ 0 & \text{در غیر این صورت} \quad ow\,, \end{cases}$$

محاسبه می‌شود. لذا این احتمال پیشین فقط روی دو برآمد $r_t = ۰$ و $r_t = r_{t-1} + 1$ دارای مقدار غیرصفر بوده و در بقیه نقاط دارای مقدار صفر است. محاسبه‌ی این احتمال پیشین یا نرخ مخاطره h، در نقطه‌ای مانند τ در زیر آورده شده است.

$$h(\tau) = \frac{P_{gap}(B = \tau)}{\sum_{t=\tau}^{\infty} P_{gap}(B = t)}\ ,$$

که در آن همان‌طور که قبلا گفته شد $P_{gap}(B)$ توزیع پیشینی است که به داده‌های هر بلوک B نسبت داده می‌شود و در موارد خاص اگر $P_{gap}(B)$ دارای توزیع نمایی گسسته (آدامز و مک کی، ۲۰۰۷) با پارامتر مقیاس λ باشد، آن‌گاه $h(\tau) = 1/\lambda$. به این نکته باید توجه کرد که نرخ مخاطره به دست آمده با توجه به افرازهای متفاوتی که روی سری انجام می‌شود، تغییر می‌کند. بنابراین نرخ مخاطره h هنگامی دارای مقدار درست است که محل نقاط تغییر معلوم بوده و افراز سری به درستی انجام شده باشد. به دلیل وابسته بودن الگوریتم تشخیص برخط نقطه تغییر به نرخ مخاطره، تشخیص نرخ مخاطره صحیح در تشخیص صحیح نقاط تغییر بسیار مهم است. در زیر دو استنباط برخط نرخ مخاطره بیان شده‌اند.

الف) استنباط نرخ مخاطره در صورت معلوم بودن محل‌های نقاط تغییر

در این استنباط، محل نقاط تغییر، معلوم فرض می‌شود. به دلیل این که نقاط تغییر، پیشامد رخ دادن تغییر هستند، می‌توان فرض کرد که به صورت نمونه‌هایی از یک فرآیند برنولی (زمان گسسته) با نرخی برابر نرخ مخاطره h، تولید می‌شوند. بنابراین نرخ مخاطره می‌تواند به عنوان نرخی از فرآیند برنولی به شرط مشاهدات دودویی استنباط شود، اگر فرض کنیم y_t یک متغیر دودویی بر حسب زمان باشد که وجود نقطه تغییر در زمان t را با $y_t = 1$ و فقدان آن را با $y_t = ۰$ نمایش دهیم، به این ترتیب پیش‌بینی نرخ مخاطره برای زمان $t + 1$ به صورت $\tilde{h}_{t+1} = p(y_{t+1} = 1|y_{1:t})$ تعریف می‌شود، در این صورت برای تشخیص نرخ مخاطره در زمان $t + 1$ نیاز به محاسبه‌ی $p(y_{t+1}|y_{1:t})$ داریم که به صورت

$$p(y_{t+1}|y_{1:t}) = \int_0^1 p(y_{t+1}|h)\, p(h|y_{1:t})\, dh = \frac{\int_{\circ}^1 p(h)\, \prod_{i=1}^{t+1} p(y_i|h)\, dh}{\int_{\circ}^1 p(h)\, \prod_{i=1}^{t} p(y_i|h)\, dh} \ ,$$

محاسبه می‌شود، که در آن $p(h)$ توزیع پیشین نرخ مخاطره است (دونی، ۲۰۰۸).

اگر تعریف کنیم $p(y_i = 1|h) = h$ و $p(y_i = 0|h) = 1 - h$ و توزیع پیشین نرخ مخاطره را توزیع بتا

$$p(h|a_{\circ}, b_{\circ}) = Beta\,(h; a_{\circ}, b_{\circ}) = \frac{\Gamma(a_{\circ} + b_{\circ})}{\Gamma(a_{\circ})\Gamma(b_{\circ})} h^{a_{\circ}-1}(1-h)^{b_{\circ}-1} \,,$$

در نظر بگیریم، آن‌گاه نرخ مخاطره به صورت

$$p(y_{t+1}|y_{1:t}) = \frac{\int_{\circ}^1 h^{a_{t+1}+a_{\circ}-1}(1-h)^{b_{t+1}+b_{\circ}-1}\, dh}{\int_{\circ}^1 h^{a_t}\, (1-h)^{b_t}\, dh}$$

$$= \frac{\Gamma(a_{t+1} + a_{\circ})\, \Gamma(b_{t+1} + b_{\circ})\, \Gamma(a_t + b_t + a_{\circ} + b_{\circ} - 1)}{\Gamma(a_t + a_{\circ})\, \Gamma(b_t + b_{\circ})\, \Gamma(a_{t+1} + b_{t+1} + a_{\circ} + b_{\circ} - 1)} ,$$

به دست می‌آید، که در آن a_t تعداد نقاط تغییر تا زمان t و $b_t = t - a_t$ تعداد نقاطی است که نقطه تغییر نیستند. در این صورت، اگر زمان بعدی $(t+1)$، نقطه تغییر باشد آن‌گاه $a_{t+1} = a_t + 1$ و در غیر این صورت $a_{t+1} = a_t$. لذا داریم

$$p(y_{t+1} = 1|y_{1:t}) = \tilde{h}_{t+1} = \frac{a_t + a_0}{a_t + b_t + a_0 + b_0},$$

$$p(y_{t+1} = 0\,|y_{1:t}) = 1 - \tilde{h}_{t+1} = \frac{b_t + b_0}{a_t + b_t + a_0 + b_0},$$

که در آن $b_t = t - a_t + b_0$. بنابراین با معلوم بودن تعداد نقاط تغییر، نرخ ثابت مخاطره به ترتیبی که گفته شد محاسبه می‌شود. در قسمت بعد نرخ مخاطره را در صورتی که محل‌های نقاط تغییر نامعلوم هستند، محاسبه می‌کنیم.

ب) استنباط نرخ مخاطره در صورت نامعلوم بودن محل‌های نقاط تغییر

وقتی محل‌های نقاط تغییر نامعلوم باشند، تعداد این نقاط یعنی a_t نامعلوم بوده و نمی‌توان نرخ مخاطره را مانند قبل به دست آورد. در این حالت توزیع احتمال پیشگو علاوه بر r_t به a_t نیز وابسته بوده و احتمال‌ها به صورت زیر نوشته می‌شوند (ویلسون و همکاران، ۲۰۱۰).

$$\begin{aligned} P(x_{t+1}|x_{1:t}) &= \sum_{r_t} \sum\nolimits_{a_t} P(x_{t+1}|r_t, a_t, x_{1:t})\ P(r_t, a_t|x_{1:t}) \\ &= \sum_{r_t} \sum\nolimits_{a_t} P\left(x_{t+1}\middle|x_t^{(r_t)}\right)\ P(r_t, a_t|x_{1:t}), \end{aligned}$$

که در آن، توزیع پسین $P(r_t, a_t|x_{1:t})$ به صورت

$$P(r_t, a_t|x_{1:t}) = \frac{P(r_t, a_t, x_{1:t})}{\sum_{r_t} \sum_{a_t} P(r_t, a_t, x_{1:t})},$$

محاسبه می‌گردد. به این ترتیب برای محاسبه توزیع توام $P(r_t, a_t, x_{1:t})$ داریم

$$P(r_t, a_t, x_{1:t}) = \sum_{r_{t-1}} \sum_{a_{t-1}} P(r_t, r_{t-1}, a_t, a_{t-1}, x_{1:t})$$

$$= \sum_{r_{t-1}} \sum_{a_{t-1}} P(r_t, a_t, x_t | r_{t-1},, a_{t-1}, x_{1:t-1})\, P(r_{t-1}, a_{t-1}, x_{1:t-1})$$

$$= \sum_{r_{t-1}} \sum_{a_{t-1}} \begin{matrix} P(r_t, a_t | r_{t-1}, a_{t-1}) p\left(x_t \middle| x_{t-1}^{(r_{t-1})}\right)\ P \\ (r_{t-1}, a_{t-1}, x_{1:t-1}). \end{matrix}$$

بنابراین توزیع پیشین طول گشت $P(r_t, a_t | r_{t-1},, a_{t-1})$، به صورت توزیع حاشیه‌ای از نرخ مخاطره که در صفحه بعد بیان شده، نوشته می‌شود:

$$P(r_t, a_t | r_{t-1}, a_{t-1}) = \int_{۰}^{1} P(r_t, a_t | h, r_{t-1}, a_{t-1})\, P(h | r_{t-1}, a_{t-1}) dh$$

$$= \frac{\Gamma(a_{t-1}+1)\,\Gamma(b_{t-1}+1)}{\Gamma(a_{t-1}+b_{t-1}+1)} \times \int_{۰}^{1} P(r_t, a_t | h, r_{t-1}, a_{t-1})\, h^{a_{t-1}} (1-h)^{b_{t-1}}\, dh\,.$$

بنابراین داریم

$$P(r_t, a_t | h, r_{t-1},, a_{t-1}) = \begin{cases} 1-h & \text{اگر } r_t = r_{t-1}+1 \quad \text{و} \quad a_t = a_{t-1} \\ h & \text{اگر } \ r_t = ۰ \quad \text{و} \quad a_t = a_{t-1}+1 \\ ۰ & \text{در غیر این صورت} \end{cases}.$$

با محاسبه روابط بالا، توزیع پیشین طول گشت به‌صورت

$$P(r_t, a_t | h, r_{t-1},, a_{t-1}) = \begin{cases} \dfrac{b_{t-1}+1}{a_{t-1}+b_{t-1}+2} = 1-\tilde{h}_t & \text{اگر } r_t = r_{t-1}+1 \quad \text{و} \quad a_t = a_{t-1} \\ \dfrac{a_{t-1}+1}{a_{t-1}+b_{t-1}+2} = \tilde{h}_t & \text{اگر } \ r_t = ۰ \quad \text{و} \quad a_t = a_{t-1}+1 \\ ۰ & \text{در غیر این صورت} \end{cases} ،$$

به‌دست می‌آید. بنابراین توزیع پیشین طول گشت برای نقاط تغییر نامعلوم به ترتیبی که گفته شد محاسبه می‌شود (ویلسون و همکاران، ۲۰۱۰).

۴-۲-۳ شرایط مرزی

همان‌طور که در بخش۴-۲-۲ گفته شد، نقطه‌ی ابتدایی سری همواره نقطه تغییر نیست. زیرا در سری‌های زمانی واقعی داده‌های مورد بررسی غالبا به صورت بازه‌ی زمانی از رخداد یک فرآیند هستند و در اکثر موارد، نقاط ابتدایی آن‌ها نقطه تغییر نمی‌باشند. بنابراین الگوریتم تشخیص برخط نقطه تغییر برای نقطه ابتدایی دارای یک مقداردهی اولیه است. به طوری‌که اگر $t = ۰$ در نظر گرفته شود، احتمال طول گشت $r_۰$ به صورت زیر محاسبه می‌شود:

۱- اگر نقطه تغییر پیشینی، قبل از اولین داده رخ دهد، در این صورت تمام مقدار احتمال طول گشت ابتدایی (آغازی) در صفر واقع می‌شود، یعنی $P(r_۰ = 0) = 1$ و به این معنی است که مشاهده‌ی اول در ابتدا طول گشت قرار گرفته لذا نقطه تغییر است.

۲- وقتی مجموعه داده در دسترس، یک بازه زمانی به صورت زیردنباله‌ای از سری زمانی اصلی است، که اکثر سری‌های زمانی مورد بررسی همین طور هستند، در این صورت احتمال پیشین

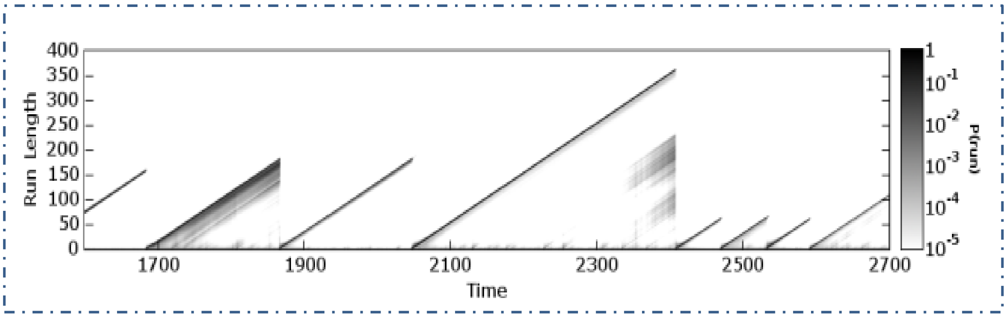
روی طول گشت آغازی یعنی احتمال این که طول گشت در نقطه زمانی صفر برابر مقدار τ باشد به‌صورت

$$P(r_۰ = \tau) = \frac{1}{Z} S(\tau),$$

محاسبه می‌شود، که در آن Z یک ثابت نرمال‌ساز مناسب می‌باشد و $S(\tau)$ به صورت زیر قابل محاسبه است.

$$S(\tau) = \sum_{t=\tau+1}^{\infty} P_{gap}(B = t).$$

این بدان معنی است که نقطه ابتدایی سری مشاهدات، در ابتدای طول گشت قرار نمی‌گیرد بلکه به گشت قبل از خود افزوده می‌شود. به عنوان مثال، زیرمجموعه ۱۵۰۰ تا ۲۷۰۰ داده از ۴۰۵۰ داده مربوط به حفاری چاه را در نظر گرفته و نقاط تغییر آنرا بررسی می‌کنیم. با توجه به محاسبات طول گشت و رسم نمودار آن در شکل ۴-۹، ملاحظه می‌شود که نقطه ابتدایی نقطه تغییر نبوده و به طول گشت قبل از خود افزده شده است (آدامز و مک کی، ۲۰۰۷).



شکل۴-۹:نمودار طول گشت برای زیر مجموعه ۱۵۰۰تا ۲۷۰۰داده از ۴۰۵۰ داده مربوط به حفاری چاه برای تشخیص نوع سنگ (آدامز و مک کی، ۲۰۰۷)

در الگوریتم تشخیص برخط نقطه تغییر، محاسبه احتمال نقطه تغییر بودن، احتمال طول گشت و نرخ مخاطره و همین‌طور مقداردهی اولیه برای نقطه ابتدایی بیان شد. ولی همان طور که گفته شد، باید به این نکته بسیار مهم توجه داشت که محاسبه این توزیع احتمال‌ها وابسته به پارامترهای توزیع‌ها است، یعنی احتمال $P(x_t|r_{t-1},x_{t-1}^{(r_{t-1})})$ به پارامترهای توزیع نیز وابسته بوده و به‌صورت $P(x_t|r_{t-1},x_{t-1}^{(r_{t-1})},\theta_m)$ نوشته می‌شود. محاسبه‌ی نرخ مخاطره نیز وابسته به پارامتر توزیع h بوده و به صورت $P(r_t|r_{t-1},\theta_h)$ نوشته می‌شود. بنابراین با بیان این مطالب این سوال پیش می‌آید که آیا می‌توان هر توزیع پیشینی را برای داده‌ها و طول گشت‌ها در نظر گرفت یا باید از توزیع‌های خاصی استفاده کرد؟ به این سوال در بخش بعد پاسخ داده می‌شود (ساتسی[1] و همکاران، ۲۰۱۰).

---

[1] Saatci

۴-۲-۴ انتخاب نوع توزیع‌های پیشین

انتخاب نوع توزیع برای الگوریتم بیزی تشخیص برخط نقطه تغییر بسیار مهم است، زیرا محاسبه‌ی احتمال‌ها برای شناسایی نقاط تغییر وابسته به پارامترهای توزیع بوده و انتخاب توزیعی که بتوان این احتمال‌ها را به آسانی تعیین کرد حایز اهمیت است. در این روش غالبا مدل‌های نمایی توام پیشنهاد شده و بسیار مناسب هستند، زیرا از آن‌جا که برای به دست آوردن توزیع پیشین و پسین نمایی به شرط پارامتر η، می‌توان پارامتر η را به وسیله پارامترهای اضافه‌ی χ و υ خلاصه‌سازی نمود، استفاده از این مدل‌ها مناسب است. در زیر خلاصه‌سازی مدل‌های نمایی توام برای محاسبه احتمال‌ها توزیع پیشگو آورده شده است. همان طور که می‌دانیم درستنمایی خانواده نمایی به صورت

$$P(x|\eta) = h(x) \exp\left(\eta^T U(x) - A(\eta)\right),$$

است، که در آن داریم

$$A(\eta) = \log \int h(x) \exp(\eta^T U(x))\, d\eta.$$

بنابراین با توجه به توزیع پیشین پارامتر η، توزیع خلاصه‌سازی شده به صورت

$$P_\pi(\eta|\chi, \upsilon) = f(\chi, \nu) \exp\left(\eta^T \chi - \upsilon A(\eta)\right),$$

محاسبه می‌شود. یا به طور هم ارز داریم

$$P_\pi(\eta|\chi, \upsilon) = f(\chi, \nu) g(\eta)^n \exp(\eta^T \chi).$$

حال با توجه به توزیع پیشین پارامتر η نسبت به پارامترهای اضافه[1] (χ, υ) و توزیع احتمال مشاهدات به شرط پارامتر η، توزیع مشاهدات نسبت به پارامتر η به صورت

---

[1] Hyper parametres

$$P(x_1,\dots,x_n|\eta)=(\prod_{i=1}^{n} h(x_i))\exp(\eta^T\sum_{i=1}^{n}U(x_i)-A(\eta)^n),$$

به دست می‌آید. توزیع پسین پارامتر η نیز نسبت به پارامترهای اضافه به صورت

$$P(\eta|\chi,\upsilon)\propto P(x_1,\dots,x_n|\eta)\,P_\pi(\eta|\chi,\upsilon)$$

$$=\left(\prod_{i=1}^{n} h(x_i)\right)\,g(\eta)^n\exp(\eta^T(\sum_{i=1}^{n}U(x_i)))\,f(\chi,\nu)g(\eta)^n\exp(\eta^T\chi)$$

$$\propto g(\eta)^n\exp(\eta^T(\sum_{i=1}^{n}U(x_i)))\,g(\eta)^n\exp(\eta^T\chi)$$

$$\propto g(\eta)^{\nu+n}\exp(\eta^T(\chi+\sum_{i=1}^{n}U(x_i))),$$

به دست می‌آید. توزیع پارامتر η به شرط مشاهدات و پارامترهای اضافه نیز به صورت

$$P(\eta|x_1,\dots,x_n,\chi,\upsilon)=P_\pi(\eta|\chi+\textstyle\sum_{i=1}^{n}U(x_i),\upsilon+n)\,,$$

به‌دست می‌آید. در این حالت اگر در الگوریتم تشخیص برخط نقطه تغییر بیزی، داده‌های متعلق به یک طول گشت r_t، دارای بردار پارامتر η باشند و به‌صورت $\eta_t^{(r)}$ نمایش داده شوند، بعد از محاسبه توزیع پسین پارامتر نسبت به مشاهدات و طول گشت، یعنی $P(\eta_t^{(r)}|r_t,x_t^{(r)})$، توزیع پیشگو به صورت

$$P(x_{t+1}|x_{1:t})=\int P(x_{t+1}|\eta)\,P\left(\eta_t^{(r)}=\eta\middle|r_t,x_t^{(r)}\right)d\eta\,,$$

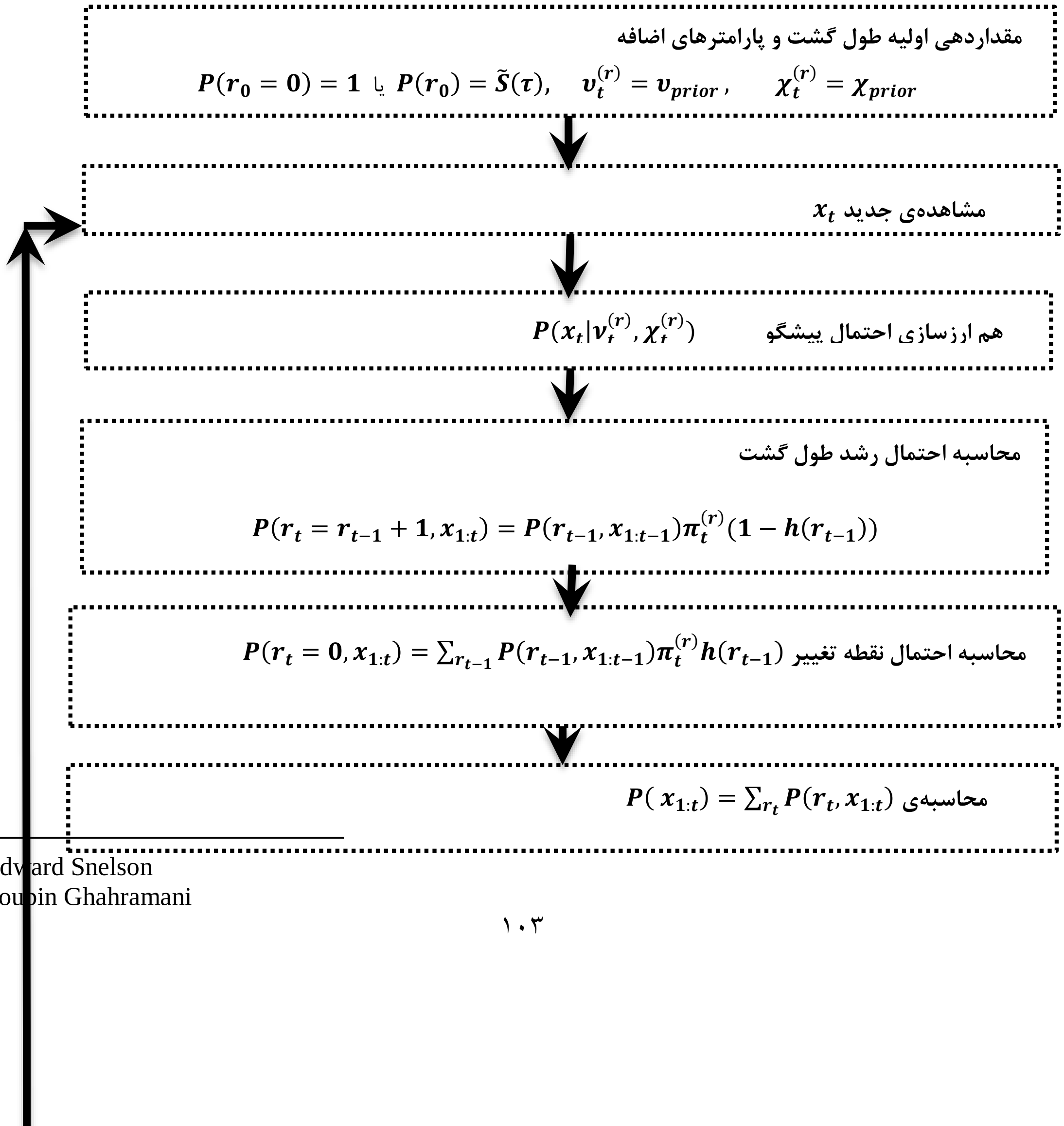
محاسبه می‌شود. این توزیع پیشگو حاشیه‌ای، توزیع نمایی با بردار پارامتر $\eta_t^{(r)}$ است که وابسته به طول گشت r_t می‌باشد. طبق روابط بالا این توزیع احتمال با استفاده از پارامترهای اضافه $(\chi_t^{(r)},\upsilon_t^{(r)})$، خلاصه‌سازی شده و توزیع پیشگو، تبدیل به تابع ساده‌ای از آماره‌های بسنده می‌شود، بنابراین پارامتر مجموعه مشاهدات در طول گشت به صورت

$$v_t^{(r)} = v_{prior} + r_t \ ,$$

$$\chi_t^{(r)} = \chi_{prior} + \sum_{t \epsilon r_t} u(x_t) \ ,$$

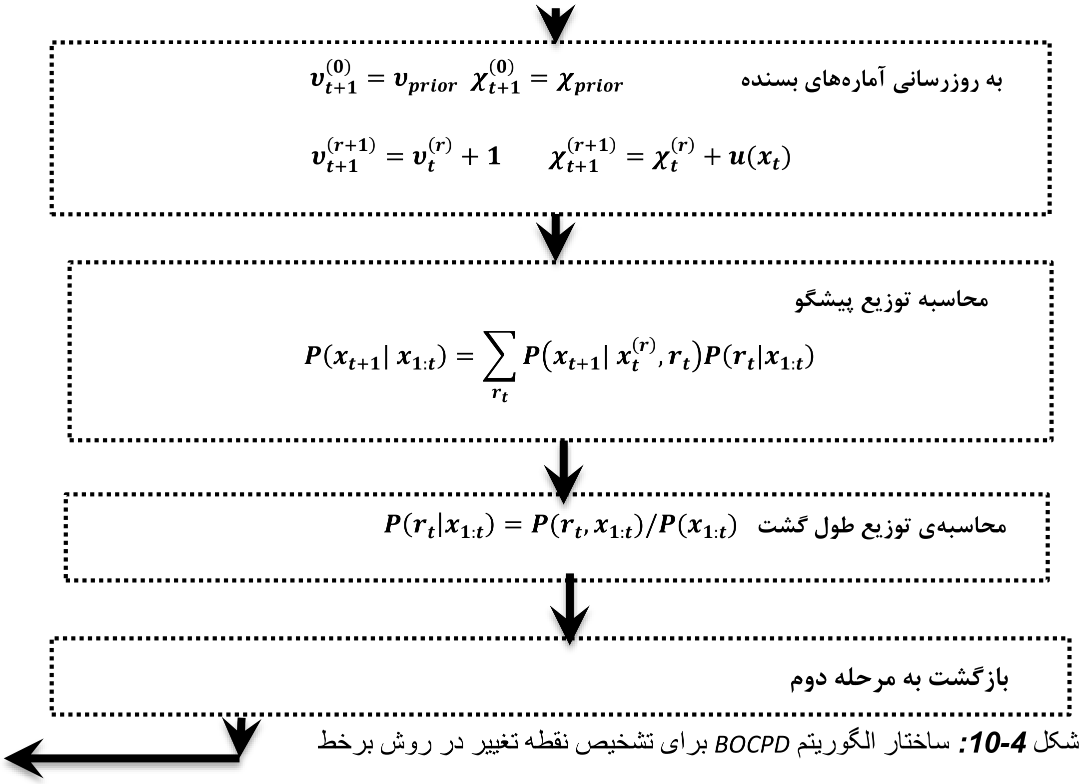
تعریف می‌شود. این نکته قابل ذکر است، در حالتی‌که توزیع‌های دقیقی برای مشاهدات در دسترس نباشند می‌توان از تقریب‌هایی که توسط اسنیلسون[1] و قهرمانی[2] (۲۰۰۵) بیان شدند، استفاده کرد.

ساختار الگوریتم تشخیص برخط نقطه تغییر بیزی برای خانواده نمایی که قبلا توضیح دادیم در شکل ۴-۱۰ بیان شده است.



---

[1] Edward Snelson
[2] Zoubin Ghahramani



شکل **10-4:** ساختار الگوریتم *BOCPD* برای تشخیص نقطه تغییر در روش برخط

فصل پنجم

پیاده‌سازی روش‌ها و استخراج نتایج

در این فصل، ابتدا داده‌های مورد استفاده در پایان‌نامه معرفی شده و سپس روش‌های بیان شده در فصل‌های ۳ و ۴ را بر روی هر یک از مجموعه داده‌ها، برازش داده و نقاط تغییر تعیین می‌شوند.

۵-۱ معرفی موارد مطالعاتی و نرم‌افزارهای مورد استفاده

داده‌های مورد استفاده در این پایان‌نامه، شامل سه سری زمانی شبیه‌سازی‌شده T_1، T_2، T_3 است که در آن‌ها به ترتیب سطح میانگین، سطح واریانس و خودهمبستگی تغییر یافته است. همچنین از داده‌هایی حقیقی به نام T_4 که مربوط به اندازه‌گیری واکنش مغناطیسی هسته‌ای در حفاری چاه هستند، استفاده شده است. در این فصل می‌خواهیم توانایی روش‌های بیان‌شده در فصل‌های ۳ و ۴ را در خصوص تشخیص نقاط تغییر در این سری‌های زمانی، مورد بررسی قرار دهیم. بدین منظور در زیربخش‌های بعدی روش‌های معرفی‌شده را روی هر یک از مجموعه داده‌های شبیه‌سازی‌شده و واقعی اجرا می‌نماییم. نقاط تغییر را با استفاده از هر کدام از روش‌ها مشخص نموده و نتایج را با هم مقایسه می‌کنیم. قبل از اجرای این روش‌ها لازم است که نرم‌افزارهای مورد استفاده برای هر روش معرفی شوند.

۱- روش $CUSUM$ مبتنی بر نمونه‌های خودگردان

کدهای روش $CUSUM$ در نرم‌افزارهای *R*، *Matlab*، *C++*، *Fortran* موجود است و در انواع نرم‌افزارهای آماری و مدیریتی نیز قابل اجرا است. روش به‌کار برده شده در این پایان‌نامه، روش $CUSUM$ مبتنی بر نمونه‌های خودگردان است، که با استفاده از نرم‌افزار $Change - Point\ Analyzer$ اجرا می‌گردد. این روش و نرم افزار مربوطه، توسط تیلور (۲۰۰۰) ارایه شد، که برای دانلود و استفاده از آن می‌توان به سایت *http://www.cf.ac.uk/maths/subsites/stats/change poin* مراجعه کرد.

۲- روش تحلیل طیفی منفرد

کد مربوط به این روش در نرم‌افزارهای R و *Matlab* موجود می‌باشد. البته قابل ذکر است که نرم‌افزار تخصصی *Catterpiller*، مختص این روش وجود دارد که از طریق سایت *http://www.cf.ac.uk/maths/subsites/stats/change point* قابل دستیابی است. در این‌جا ما برای تحلیل داده‌ها از بسته $Rssa$ در نرم‌افزار R استفاده کرده‌ایم.

۳- روش مدل افراز ضربی

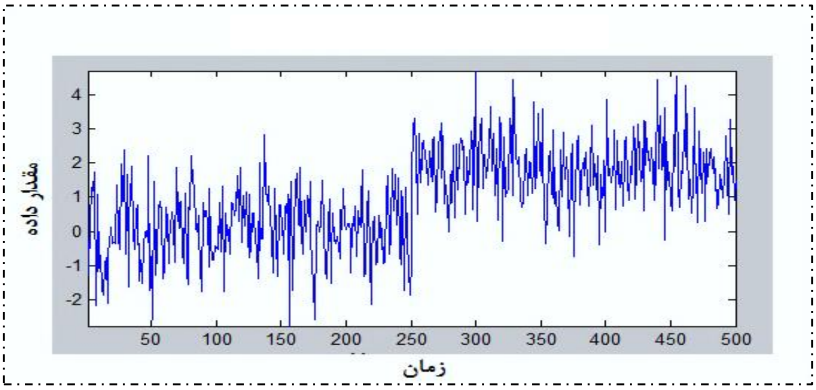
کد مربوط به روش PPM در نرم‌افزارهای R، *Matlab* ، *C++* و *Fortran* موجود است. در این پایان‌نامه، روش PPM با استفاده از بسته bcp در نرم‌افزار R اجرا گردیده است.

۴- روش تشخیص برخط نقطه تغییر بیزی

کدهای مربوط به روش $BOCPD$ در نرم‌افزارهای *Fortran*، *C++*، *S-PLUS* و *Matlab* موجود می‌باشد. در این پایان‌نامه از نرم‌افزار *Matlab* برای تحلیل داده‌ها استفاده شده‌است، که برخی از کدهای این روش در سایت‌های *http://mlg.eng.cam.ac.uk* و *http://www.inference.phy.cam.ac.uk* قابل دسترس است.

۵-۲ نتایج تشخیص تغییر درسطح میانگین سری زمانی

در این بخش، سری شبیه‌سازی‌شده‌ای را در نظر گرفتیم که در آن سطح میانگین سری تغییر کرده است. بدین منظور سری T_1، از توزیع نرمال با طول ۵۰۰ را بدین صورت در نظر گرفتیم که ۲۵۰ مقدار اول آن از توزیع $N(0,1)$ و ۲۵۰ مقدار دوم از توزیع $N(2,1)$ شبیه‌سازی‌شده باشند. نمودار این سری زمانی شبیه سازی شده در شکل ۵-۱ رسم شده است. به وضوح مشخص است که سطح میانگین در نقطه زمانی t=۲۵۱ تغییر یافته است.

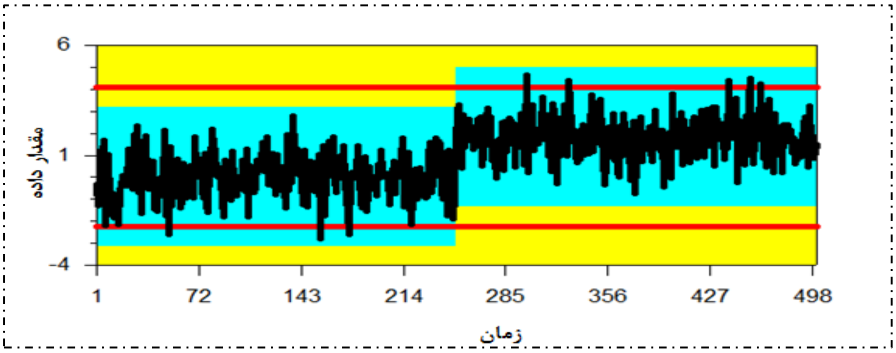


شکل ۵-۱: نمودار سری زمانی T_1 که میانگین آن در نقطه زمانی ۲۵۱ تغییر یافته است

در این‌جا می‌خواهیم بدانیم کدامیک از این روش‌ها، قابلیت تشخیص این نوع تغییر را دارند. لذا نتایج اجرای هر روش بر این سری داده را مطالعه می‌کنیم.

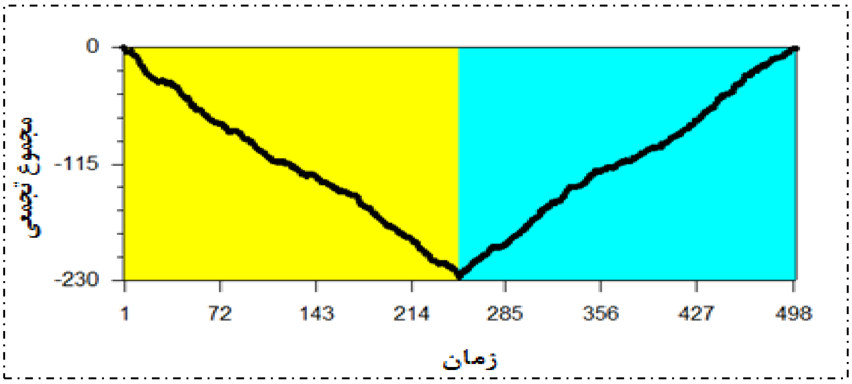
۵-۲-۱ نتایج روش $CUSUM$ مبتنی بر نمونه‌های خودگردان

قبل از بیان نتایج این روش، نمودار کنترل سری زمانی T_1 که در شکل ۵-۲ نشان داده شده است را در نظر بگیرید. همان طور که ملاحظه می‌شود برخی از مشاهدات، خارج از حدود کنترل بوده و به عنوان نقطه تغییر شناسایی شده‌اند، که در واقع نقطه تغییر نیستند. در عوض نقطه تغییر واقعی (t=۲۵۱)، خارج از حدود کنترل واقع نشده است. بنابراین به سادگی می‌توان دریافت که نمودار کنترل قادر به تشخیص نقطه تغییر نیست.



شکل ۵-۲: نمودار کنترل سری زمانی T_1

در شکل ۵-۳ نمودار $CUSUM$ این داده‌ها نشان داده شده است. مشاهده می‌شود، که شیب این نمودار در اطراف نقطه زمانی ۲۵۱ تغییر کرده و این بدان معنا است که در این نقطه تغییری در میانگین مشاهدات رخ داده است. برای حصول اطمینان از وجود تغییر در این نقطه و به دست آوردن سطح و فاصله اطمینان برای تغییر، ۱۰۰۰ نمونه خودگردان از داده‌ها تولید شده‌اند و همان‌گونه که در بخش ۳-۱ توضیح داده شد، نمودار $CUSUM$ آن‌ها با نمودار اصلی مقایسه گردید. نتایج در جدول ۵-۱ آمده است.



شکل ۵-۳: نمودار $CUSUM$ سری زمانی T_1

جدول ۵-۱: نقطه تغییر تعیین شده، سطح اطمینان و فاصله اطمینان سری T_1

شماره داده	فاصله اطمینان	سطح اطمینان	تغییر میانگین از	تغییر میانگین به
۲۵۱	(۲۴۷,۲۵۴)	٪۱۰۰	۰٫۰۵۲۴۵۳	۱٫۸۵۸۶

جدول ۵-۱ بیان می‌کند که این روش، تغییری را در نقطه زمانی ۲۵۱ تشخیص داده، به طوری که در این نقطه زمانی سطح میانگین از ۰/۰۵۲۴۵۳ به ۱/۸۵۸۶ تغییر یافته است. فاصله اطمینان برای نقطه تغییر تعیین شده، با سطح اطمینانی برابر ٪۱۰۰ (که از رابطه (۳-۱) به دست می‌آید)، بین نقاط (۲۵۴ , ۲۴۷) تعیین می‌شود و بدین معنا است که در این فاصله با اطمینان ٪۱۰۰، حتما یک تغییر وجود دارد. برای تشخیص تغییر در واریانس این سری، روش بیان‌شده در بخش ۳-۱-۳ اعمال گردیده

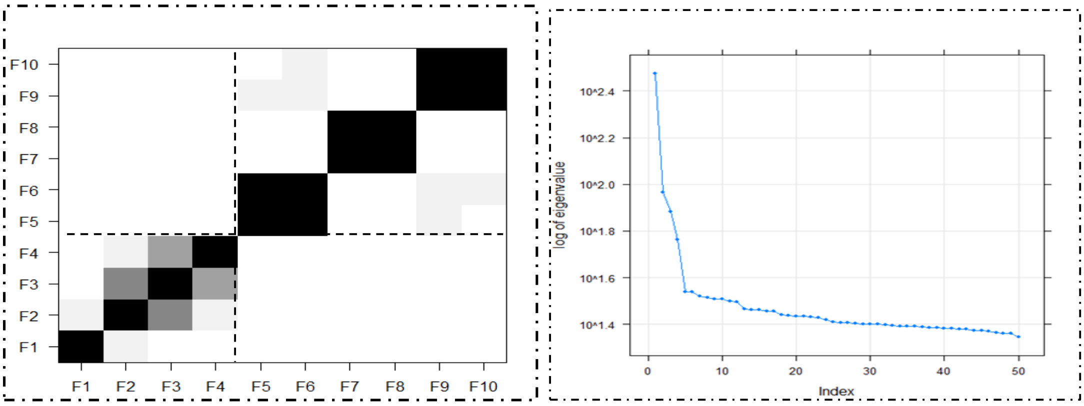
و مشخص شده که مقدار واریانس برای تمامی مشاهدات برابر $1/112275$ بوده و لذا می‌توان گفت، تغییری در واریانس سری رخ نداده که دقیقا مطابق با طبیعت سری مورد مطالعه است.

۵-۲-۲ نتایج روش تحلیل طیفی منفرد

طبق مطالب بیان شده بخش ۳-۲ در خصوص روش تحلیل طیفی منفرد SSA، دقت تحلیل این روش به عوامل مختلفی همانند انتخاب M و l که پارامترهای اصلی این روش هستند و طول ماتریس آزمون (B) و اصلی (T) که پارامترهای این روش برای تشخیص نقطه تغییر می‌باشند، بستگی دارد. تغییر هر یک از این عوامل باعث ایجاد نتایج متفاوتی در تشخیص نقطه تغییر می‌شود. در این‌جا $M = \frac{B}{2}$ در نظر گرفته شده و مقدار l با استفاده از دو روش بیان‌شده در بخش ۳-۲-۳-۲ انتخاب می‌شود که این دو روش در ادامه توضیح داده می‌شوند.

روش کلی که درکتاب‌ها و مقالات مختلف (کتاب آنالیز سری‌های زمانی: SSA و روش‌های مربوطه، گلویندینا و همکاران، ۲۰۰۱) برای انتخاب l پیشنهاد شده است، بر اساس l بردار ویژه‌ای است که دارای بزرگ‌ترین مقادیر ویژه بوده و درصد بالایی از تاثیرپذیری سری، را به خود اختصاص می‌دهند (این درصد بستگی به تحلیل‌گر دارد). لذا برای انتخاب l مولفه سری داده T_1، نمودار ۵۰ مقدار ویژه این سری که به طور نزولی مرتب شده‌اند در شکل ۵-۴ (الف) رسم گردیده است. ملاحظه می‌شود که ۵ مقدار ویژه اول می‌توانند به عنوان مولفه‌های گروهی انتخاب شوند که به خوبی و با درصد بالایی (بالاتر از ۸۵٪) بیان‌کننده سری اصلی هستند. روش دیگری که برای انتخاب l مولفه این گروه به‌کار برده می‌شود، استفاده از ماتریس W-همبستگی است که در شکل ۵-۴ (ب) نمایش داده شده است. همانند روش قبل، این نمودار نیز ۵ مولفه اول را به عنوان گروهی که می‌توانند به خوبی بیان‌کننده سری اصلی باشند، معرفی می‌کند. زیرا همبستگی بین مشاهدات در این مولفه (F)

تغییر کرده و مشاهدات نسبت به هم همبسته می‌شوند. در این‌جا این مولفه، ماتریس را به دو قسمت تفکیک می‌کند که در شکل ۵-۴(ب) با خطوط نقطه چین مشخص شده است.



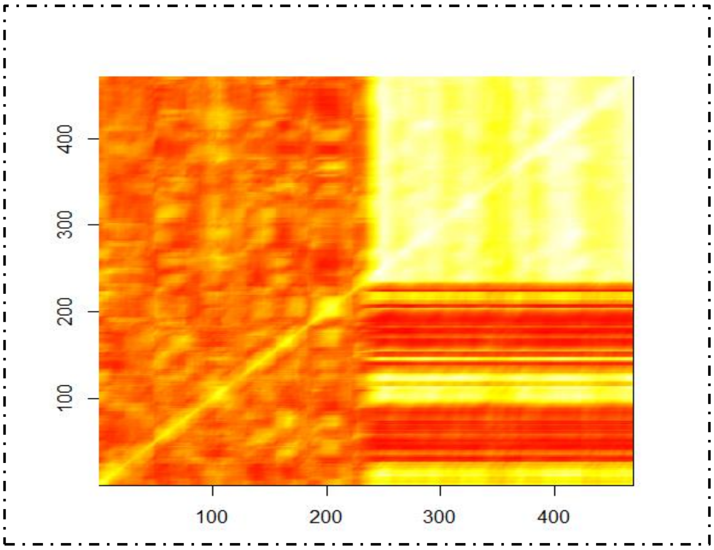
(الف)

(ب)

شکل ۵-۴: (الف) نمودار مقادیر ویژه و (ب) نمودار ماتریس W-همبستگی سری T_1 برای تشخیص l مولفه گروه I

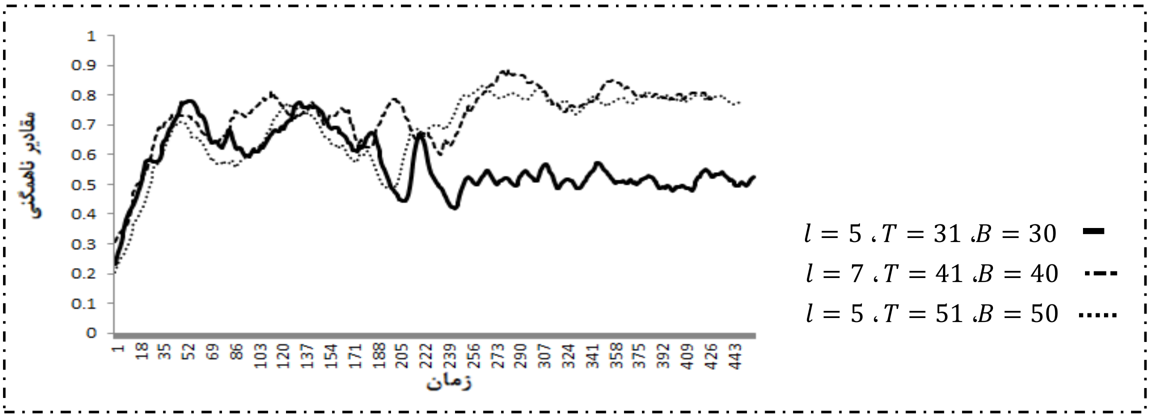
بعد از انتخاب پارامترهای موثر در روش SSA نیازمند انتخاب پارامترهای B و T هستیم که مقدار قراردادی برای انتخاب این پارامترها وجود ندارد. بنابراین، تحلیل برای مقادیر مختلف این پارامترها انجام شده و بهترین انتخاب، حالت $B - T = 1$ است. در این‌جا برای تشخیص نقاط تغییر از توابع تشخیص ماتریس ناهمگنی (*Heterogeneity*) بین مشاهدات استفاده می‌شود که مقادیر این ماتریس، شاخص ناهمگنی بین مشاهدات است که طبق رابطه (۳-۵) محاسبه می‌شوند. مقادیر این ماتریس ناهمگنی در شکل ۵-۵ با یک طیف رنگی نمایش داده شده است. به طوری که مقدار هر مولفه ماتریس ناهمگنی نشان‌دهنده یک رنگ در نمودار ماتریس بوده و نقاط پررنگ‌تر در نمودار، نشان‌دهنده مقادیر بالای ناهمگنی بین مشاهدات است. در محل‌هایی که نمودار ماتریس کم‌رنگ‌تر می‌شود، در واقع مقدار ناهمگنی بین مشاهدات کاهش یافته است. بنابراین می‌توان محلی را که نمودار ماتریس تغییر رنگ داده به عنوان نقطه تغییر در نظر گرفت. با توجه به شکل ۵-۵، مشاهده می‌شود که در اطراف نقطه ۲۵۰، تغییر رنگی در نمودار رخ داده و بدان معنا است که در این محل روند

همگنی بین مشاهدات تغییر کرده است. برای تشخیص محل تغییر باید از توابع تشخیص سطری یا ستونی این ماتریس که در بخش ۳-۲-۲-۱ فصل سوم بیان شد، استفاده کرد.



شکل۵-۵:نمایش ماتریس ناهمگنی سری T_1

ماتریس ناهمگنی فوق برای مقادیر مختلف B، T و l رسم شده و در هر حالت نقاط تغییر توسط توابع تشخیص این ماتریس مشخص گردیدند که چند نمونه از توابع تشخیص که عملکرد بهتری برای تشخیص نقطه تغییر واقعی ($t = 251$) داشتند، در شکل ۵-۶ در کنار هم رسم شده‌اند. نقاط تغییر تعیین شده متناظر با هر یک از این تنظیمات در جدول ۵-۲ آمده است.



شکل ۵-۶: نمودار توابع تشخیص ستونی ماتریس ناهمگنی

جدول ۵-۲: مقدار ناهمگنی سری T_1

	مقدار ناهمگنی به ازای $B = 30 \quad, T = 31 \quad, l = 6$	مقدار ناهمگنی به ازای $B = 40 \quad, T = 41 \quad, l = 7$	مقدار ناهمگنی به ازای $B = 50 \quad, T = 51 \quad, l = 5$
شماره نقطه تغییر	۲۴۵	۲۳۷	۲۴۲
مقدار ناهمگنی	۰٫۰۴۲۰۳	۰٫۰۶۹۹	۰٫0713

با توجه به نمودار ۵-۶ و جدول ۵-۲، ملاحظه می‌شود که با انتخاب $B = 30$، $T = 31$ و $l = 5$، نقطه زمانی ۲۴۵ به عنوان نقطه تغییر مشخص می‌شود که دارای کم‌ترین مقدار ناهمگنی (بیشترین همگنی) برابر ۰٫۰۴۲۰۳ است. در واقع از این نقطه به بعد روند نمودار ناهمگنی (شکل ۵-۶) تغییر کرده است و به منزله آن است که در این نقطه حتما تغییری در سری زمانی رخ داده است. بنابراین روش *SSA* در بهترین حالت انتخاب پارامترها توانسته است، نقطه ۲۴۵ را که فاصله کمی تا نقطه تغییر واقعی ($t = 251$) دارد به عنوان نقطه تغییر تعیین کند.

۵-۲-۳ نتایج روش افراز ضربی

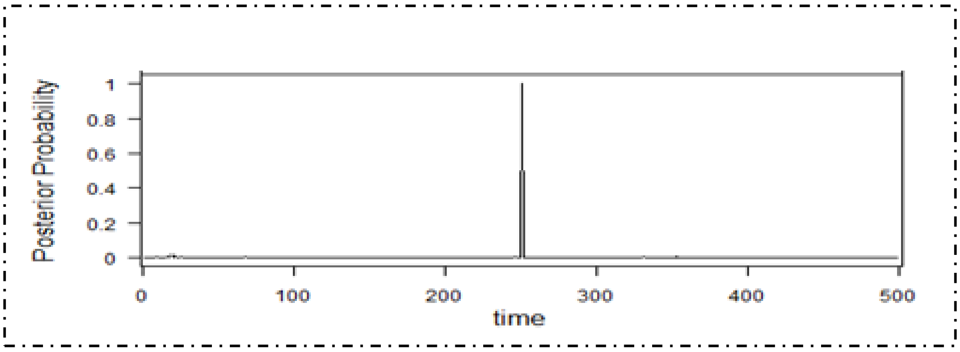
طبق مطالب بیان شده در بخش ۴-۱ در خصوص تشخیص نقطه تغییر با استفاده از روش *PPM*، می‌دانیم که در این روش ابتدا محتمل‌ترین تعداد نقاط تغییر، برآورد شده و سپس احتمال این که هر مشاهده بتواند نقطه تغییر باشد (بخش ۴-۱-۴ را ببینید)، تعیین می‌گردد. در دیدگاه دوم (بخش ۴-۱-۵) مقدار مورد انتظار پارمتر توزیع هر مشاهده محاسبه شده و نقاطی را که در آن‌ها پارامتر توزیع تغییر می‌کند به عنوان نقاط تغییر شناسایی می‌گردند.

در این روش عوامل مختلفی مانند انتخاب توزیع‌های پیشین، هم‌ارزهای پسین، پارامترهای p_0 و w_0 و همین طور تعداد تکرار در روش نمونه‌گیری گیبز که دقت تقریب به آن بستگی دارد، دخالت دارند و با تغییر هر مورد، نتایج متفاوتی به دست خواهد آمد. در روش *PPM* به‌کار برده شده در این‌جا پیشین‌ها و هم‌ارزهای یائو (۱۹۸۴) مورد استفاده قرار گرفته و مقدار قراردادی تعداد تکرار نمونه‌گیری گیبز برای برآورد احتمال پسین (احتمال نقطه تغییر بودن هر مشاهده) ۲۰۰۰۰ تکرار و

برای برآورد میانگین پسین (مقدار مورد انتظار پارامتر) ۱۰۰۰۰ تکرار در نظر گرفته شده است. مقدار بهینه برای پارامترهای p_0 و w_0 در روش بری و هارتیگن برابر ۰/۲ ذکر شده ولی ما علاوه بر آن تحلیل را برای مقادیر مختلف p_0 و w_0 انجام دادیم که فقط دو نمونه از آن برای مقادیر ۰/۱ و ۰/۳ در جدول ۵-۳ آمده است. با توجه به مندرجات جدول، ملاحظه می‌شود که بهترین مقدار برای پارامترهای p_0 و w_0 برابر همان مقدار ۰/۲ می‌باشد، زیرا با این انتخاب، نقطه زمانی ۲۵۰ با احتمال ۱/۰۰۰ به عنوان نقطه تغییر تشخیص داده می‌شود که نقطه تغییر واقعی نیست. همچنین در حالت $w_0 = p_0 = 0/1$ و $w_0 = p_0 = 0/3$ نقطه زمانی ۲۵۰ به ترتیب با احتمال ۰/۸۰۲ و ۰/۶۴۶ که خیلی کمتر از حالت $w_0 = p_0 = 0/2$ است به عنوان نقطه تغییر تعیین می‌شود. نمودار احتمال پسین برای تمام مشاهدات در حالت $w_0 = p_0 = 0/2$ در شکل ۵-۷ رسم شده است.

جدول ۵-۳ مقدار احتمال پسین برای سه داده سری T_1 در حالت $w_0 = p_0 = 0/3, 0/2, 0/1$

شماره داده	$w_0 = 0/2$ $p_0 = 0/2$	$w_0 = 0/1$ $p_0 = 0/1$	$w_0 = 0/3$ $p_0 = 0/3$
۲۴۹	۰/۰۰۰	۰/۰۰۰	۰/۱۵۶
۲۵۰	۱/۰۰۰	۰/۸۰۲	۰/۶۴۶
۲۵۱	۰	۰/۲	۰/۲



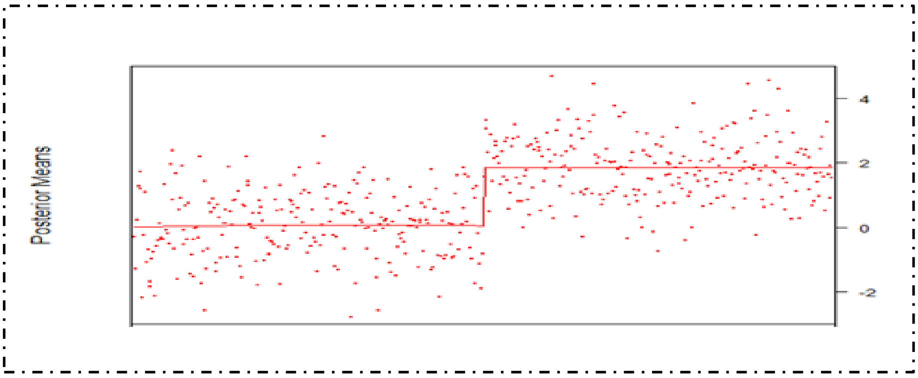
شکل ۵-۷: نمودار احتمال پسین نقطه تغییر بودن هر مشاهده با استفاده از مدل *PPM*

روش *PPM* از دیدگاه دوم یعنی تغییر در مقدار مورد انتظار پارامتر توزیع (میانگین)، نیز به بررسی مساله می‌پردازد. نمودار این مقادیر به ازای هر مشاهده یعنی $E(\mu_1), E(\mu_2), \ldots, E(\mu_n)$ برای

حالت $w_{\circ} = p_{\circ} = 0/2$ در شکل ۵-۸ رسم شده است. با توجه به نمودار ملاحظه می‌شود که تغییری در اطراف نقطه ۲۵۰ رخ داده است. بنابراین برای تشخیص محل تغییر، مقادیر مورد انتظار پارامتر میانگین مشاهدات مورد بررسی قرار می‌گیرند که سه داده متوالی آن در اطراف نقطه ۲۵۰ در جدول ۵-۴ ذکر شده‌اند. با توجه به جدول، ملاحظه می‌شود که در نقطه ۲۵۰ مقدار میانگین، تغییر چشم‌گیری از مقدار ۰٫۰۵۲۵ به ۱٫۸۵۸۶ داشته و بنابراین به عنوان نقطه تغییر در نظر گرفته می‌شود.

جدول ۵-۴: تغییر در میانگین پسین برای سه داده متوالی در حالت $w_{\circ} = p_{\circ} = 0/2$

شماره داده	میانگین پسین از مقدار $w_{\circ} = p_{\circ} = 0/2$	به مقدار $w_{\circ} = p_{\circ} = 0/2$
۲۴۹	۰٫۰۳۶۳	۰٫۰۵۲۵
۲۵۰	۰٫۰۵۲۵	۱٫۸۵۸۶
۲۵۱	۱٫۸۵۸۶	۱٫۸۶۷۹



شکل ۵-۸: نمودار مقادیر مورد انتظار میانگین برای هر مشاهده با استفاده از مدل PPM به ازای $w_{\circ} = p_{\circ} = 0/2$

با توجه به جدول ۵-۳ و ۵-۴، ملاحظه می‌شود که روش PPM، با استفاده از هر دو دیدگاه توانسته است، تغییر در سطح میانگین سری را به خوبی تشخیص دهد و نقطه زمانی ۲۵۰ را که نقطه تغییر واقعی $(t = ۲۵۱)$ نیست ولی خیلی نزدیک به آن است به عنوان نقطه تغییر تعیین کند.

ذکر این نکته حایز اهمیت است که اگر این روش روی همین سری داده تکرار شود آن‌گاه مقادیر احتمال‌ها، کمی متفاوت‌تر به دست می‌آیند ولی در تشخیص نقطه تغییر تاثیری ندارد و با هر بار تکرار باز هم نقطه ۲۵۰ به عنوان نقطه تغییر تشخیص داده می‌شود.

۵-۲-۴ نتایج روش برخط بیزی

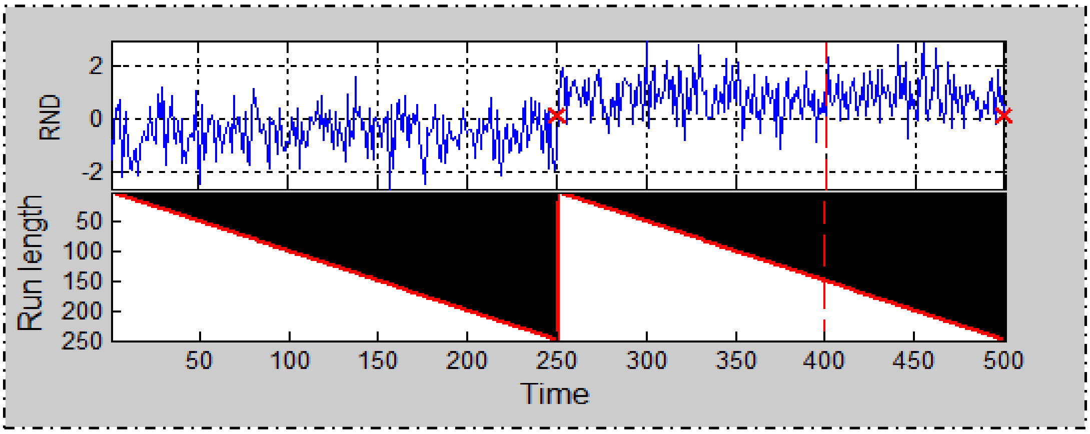
با توجه به این نکته که در روش برخط نقطه تغییر، می‌توان از توابع مختلفی برای نرخ مخاطره استفاده کرد یا توزیع‌های پیشین مختلفی برای مشاهدات در نظر گرفت، لذا نتایج متفاوتی به‌دست خواهد آمد. در این‌جا از تابع مخاطره لجستیک استفاده شده و مشاهدات از توزیع پیشین نرمال، پیروی می‌کنند.

نتایج برازش روش $BOCPD$ روی سری داده T_1 در جدول ۵-۵ آمده است. ملاحظه می‌شود که مقدار احتمال طول گشت $P(r_t|x_{1:t})$، با افزایش هر نقطه به طول گشت r_t، از نقطه صفر با احتمال ۰/۱۰۰۰ شروع شده (با توجه به تعریف شرایط مرزی $P(r_۰ = 0) = 1$)، در نقطه زمانی ۲۴۸ به مقدار ۰/۹۹۰۲ و در نقطه زمانی ۲۴۹ به مقدار ۰/۹۹۷۳ افزایش یافته، اما در نقطه زمانی ۲۵۰ به یکباره به مقدار صفر کاهش می‌یابد. سپس از این نقطه به بعد نیز این کاهش ادامه داشته و به مقدار ۹/۹۶۴۶ e-۸ نزول می‌کند. با توجه به توضیحات ذکر شده، ملاحظه می‌شود که مقدار احتمال طول گشت در نقطه ۲۵۰ به یکباره کاهش یافته و این بدان معنا است که این نقطه با احتمال بسیار پایین (که نرم‌افزار مقدار صفر در نظر می‌گیرد)، متعلق به این گشت است ولی در عوض با احتمال بسیار بالایی (تقریبا یک) متعلق به گشت بعدی می‌باشد.

جدول۵-۵: احتمال طول گشت $P(r_t|x_{1:t})$، احتمال توزیع پیشگو $P(x_{t+1}|x_{1:t})$، و مقدار تابع مخاطره $P(r_{t-1}|r_t)$

شماره داده	احتمال طول گشت $P(r_t\|x_{1:t})$	احتمال توزیع پیشگو $P(x_{t+1}\|x_{1:t})$	مقدار پارامترهای تابع مخاطره (توزیع لجستیک $P(r_{t-1}\|r_t)$
۲۴۸	۰٫۹۹۰۲	۰٫۲۲۰	-۸٫۸۴۹۶۳
۲۴۹	۰٫۹۹۷۳	۰٫۲۲۱۵	۰٫۹۸۹۰
۲۵۰	۰	۰٫۲۶۲۲	۰٫۹۹۴۱

در شکل ۵-۹، نمودار سری زمانی T_1 (قاب بالا) و مقادیر احتمال طول گشت $P(r_t|x_{1:t})$ (قاب پایین) به ازای تمام طول گشت‌های مختلف با یک طیف رنگی مشخص شده‌اند. احتمال‌های بزرگتر، پررنگ‌تر هستند.



شکل ۵-۹: نمودار سری T_1 (قاب بالا) و نمودار طول گشت‌ها و مقادیر احتمال آن با یک طیف رنگی (قاب پایین)

با توجه به نمودار (قاب پایین)، ملاحظه می‌شود که مقادیر احتمال طول گشت تا نقطه زمانی ۲۴۹ پررنگ بوده و طول گشت تا این نقطه افزایش یافته ولی در نقطه زمانی ۲۵۰ مقدار احتمال به صفر رسیده و در این نقطه، گشت جدیدی با طول صفر که دارای احتمال بالاتری است شروع می‌شود. طول گشت نهایی در نمودار (قاب پایین) با خط مورب قرمز نشان داده شده است. بنابراین با توجه به جدول ۵-۵ و نمودار ۵-۹، نقطه زمانی ۲۵۰ به عنوان نقطه تغییر برآورد می‌شود.

۵-۲-۵ خلاصه نتایج برای سری شبیه‌سازی شده T_1

پس از اعمال هریک از روش‌های ذکرشده بر روی سری شبیه‌سازی T_1، می‌توان در خصوص تعیین مناسب‌ترین روش برای تشخیص نقطه تغییر بحث کرد. با توجه به جدول ۵-۶ ملاحظه می‌شود که سه روش تحلیل طیفی منفرد، روش تشخیص برخط نقطه تغییر بیزی و مدل افراز ضربی توانسته‌اند نقطه زمانی ۲۵۰ را که به نقطه تغییر واقعی $(t = ۲۵۱)$ بسیار نزدیک است، به عنوان نقطه تغییر تشخیص دهند. ولی روش مجموع تجمعی مبتنی بر نمونه‌های خودگردان، عملکرد بهتری داشته و دقیقا نقطه زمانی ۲۵۱ را به عنوان نقطه تغییر تشخیص می‌دهد.

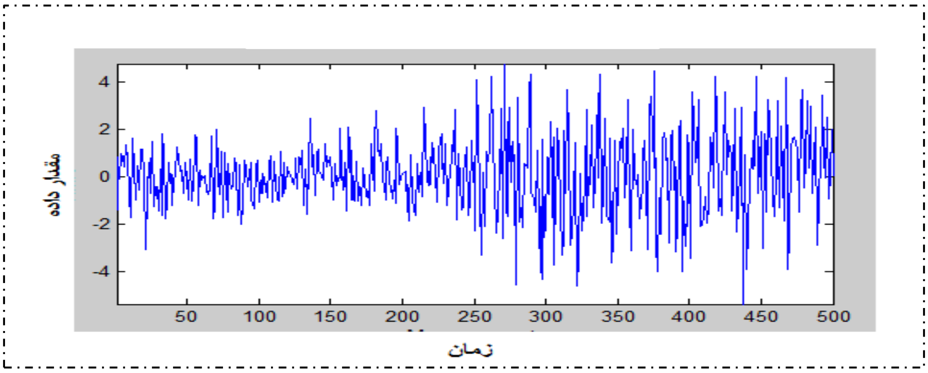
جدول ۵-۶: نقطه تغییر واقعی سری شبیه‌سازی‌شده T_1 و مقادیر برآورد آن توسط روش‌های مختلف

روش	نام روش	نقطه تغییر برآورد شده	نقطه تغییر اصلی
۱	روش *CUSUM*	۲۵۱	۲۵۱
۲	روش *SSA*	۲۴۵	
۳	روش *PPM*	۲۵۰	
۴	روش *BOCPD*	۲۵۰	

بنابراین با مطالب بیان‌شده، برای تشخیص تغییر در سطح میانگین سری، هر چهار روش ذکر شده، عملکرد مناسبی داشتند ولی روش مجموع تجمعی نسبت به بقیه ارجحیت دارد.

۵-۳ نتایج تشخیص تغییر در سطح واریانس سری زمانی

سری داده T_1 در بخش قبل به‌صورتی در نظر گرفته شده بود که تغییر، فقط در سطح میانگین سری رخ داده باشد. در این بخش سری شبیه‌سازی‌شده‌ای را در نظر می‌گیریم که سطح میانگین سری ثابت ولی واریانس آن تغییر کند. بدین منظور، سری T_2، با طول ۵۰۰ را بدین صورت در نظر می‌گیریم که ۲۵۰ مقدار اول آن از توزیع $N(0,1)$ و ۲۵۰ مقدار دوم از توزیع $N(0,2)$ شبیه سازی شده باشند. نمودار سری زمانی شبیه سازی شده T_2 در شکل ۵-۱۰ نمایش داده‌شده‌است.

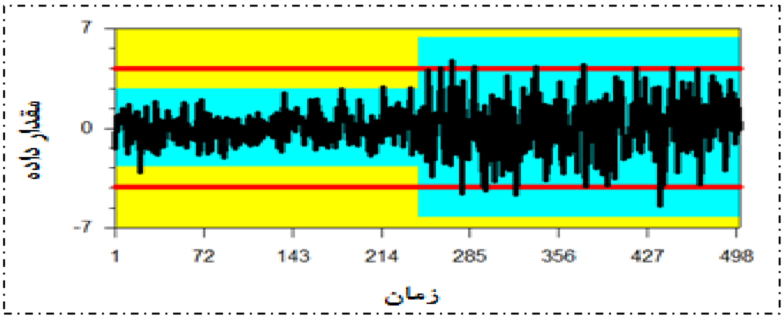


شکل ۵-۱۰: نمودار سری زمانی T_2 که واریانس آن در نقطه زمانی ۲۵۱ تغییر یافته است

پس از معرفی سری داده T_2، نتایج اعمال هر روش بر این سری داده را مطالعه می‌کنیم.

۵-۳-۱ نتایج روش $CUSUM$ مبتنی بر نمونه‌های خودگردان

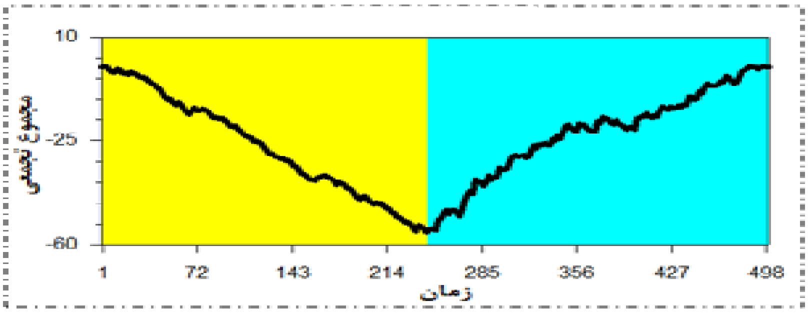
ابتدا نمودار کنترل سری زمانی T_2 را که در شکل ۵-۱۱ نشان داده شده، در نظر بگیرید. مشاهده می‌شود که نمودار کنترل، همانند قبل، قادر به تشخیص تغییر در واریانس در نقطه مورد نظر نیست. زیرا نقاط بسیاری خارج از حدود کنترل قرار گرفته‌اند ولی نقطه تغییر واقعی $(t = 251)$ درون حدود کنترل است.



شکل ۵-۱۱: نمودار کنترل سری زمانی T_2

بنابراین از نمودار $CUSUM$ که در شکل ۵-۱۲ نمایش داده شده است، استفاده می‌کنیم.، شیب این نمودار در اطراف نقطه زمانی ۲۴۵ تغییر کرده و بدان معنا است که این روش نقطه زمانی $t = 245$ را به عنوان نقطه تغییر تعیین می‌کند. همچنین برای نقطه تغییر مشخص شده، سطح و فاصله

اطمینان مربوطه در جدول ۵-۷ آمده است. با توجه به این جدول ملاحظه می‌شود که واریانس مشاهدات در نقطه زمانی ۲۴۵ از مقدار ۰٫۹۱۴۳ به مقدار ۲٫۰۹۷۸ افزایش یافته و این بدان معنا است که فاصله اطمینان، بین نقاط (۲۷۹ , ۲۳۵) با سطح اطمینان ۱۰۰٪ تعیین می‌شود. قابل ذکر است که تغییری در میانگین مشاهدات رخ نداده و مقدار میانگین برای تمامی مشاهدات برابر ۰٫۰۷۷۴۴ برآورد شده است که تفاوت بسیار اندکی با میانگین واقعی صفر دارد.



شکل۵-۱۲: نمودار CUSUM سری T_2

جدول ۵-۷: نقطه تغییر تعیین‌شده، سطح اطمینان و فاصله اطمینان سری T_2

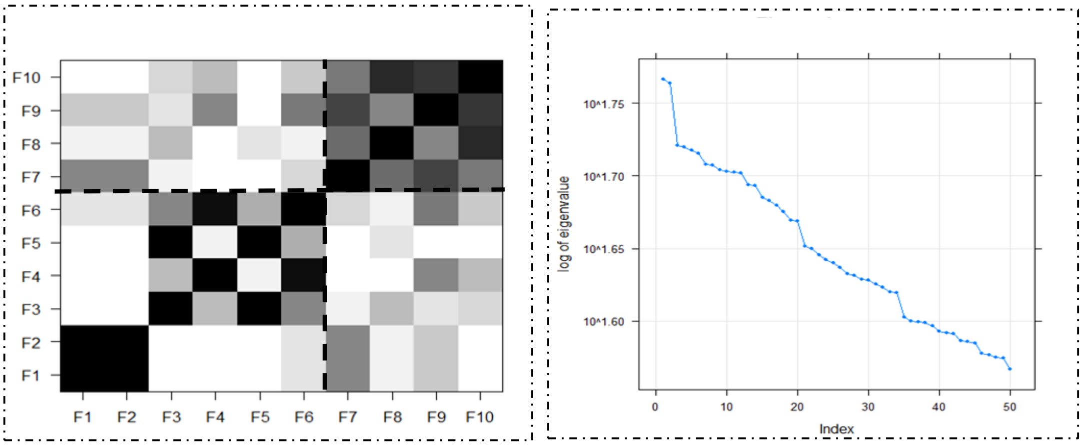
شماره داده	فاصله اطمینان	سطح اطمینان	تغییر واریانس از	تغییر واریانس به
۲۴۵	(۲۳۵ ,۲۷۹)	٪۱۰۰	۰٫۹۱۴۳	۲٫۰۹۷۸

بنابراین با توجه به توضیحات ذکر شده، این روش توانسته است تغییری را در واریانس سری تشخیص دهد اما قادر به تعیین دقیق نقطه تغییر $(t = ۲۵۱)$ نبوده است.

۵-۳-۲ نتایج روش تحلیل طیفی منفرد

همان‌طور که در زیر بخش ۵-۲-۲ توضیح داده شد، پارامترهای B, T, M و l پارامترهای موثر در این روش هستند و تحلیل در این روش باید برای مقادیر مختلف پارامترهای B و T انجام شود. پارامتر M همانند قبل برابر $\frac{B}{2}$ انتخاب می‌شود. پارامتر l (عناصر گروهی که به خوبی بیان‌کننده سری واقعی هستند)، نیز از دو روش، یکی انتخاب l مقدار ویژه بزرگتر و دیگری ماتریس W-همبستگی به دست

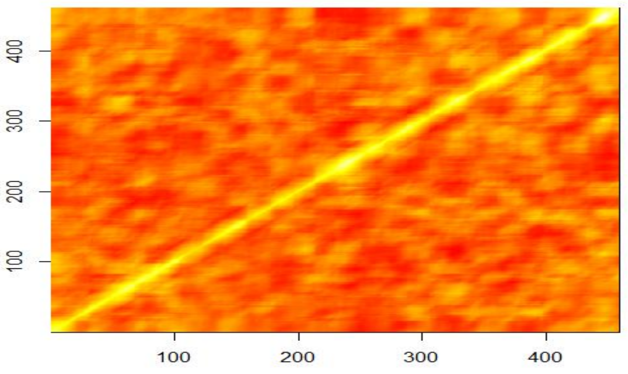
می‌آید. نمودار هر دو روش در شکل ۵-۱۳ رسم شده‌است. با توجه به نمودار (الف)، ۷ مقدار ویژه اول با درصد بالایی(بالاتر از ۷۰٪)، بیان کننده سری اصلی است. همچنین با استفاده از روش W-همبستگی که در نمودار (ب) رسم شده‌است، تفکیک ماتریس W-همبستگی که با خطوط خطچین در نمودار (ب) مشخص شده نشان می‌دهد که ۷ مولفه اول به خوبی بیان‌کننده سری واقعی داده‌ها خواهند بود.



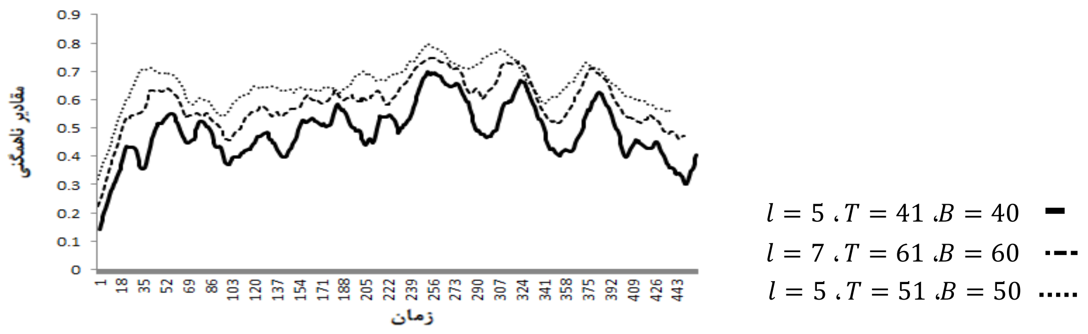
(ب) (الف)

شکل ۵-۱۳: (الف) نمودار مقادیر ویژه، (ب)نمودار W-همبستگی سری T_2 برای تشخیص l مولفه گروه I

پس از انتخاب پارامترها بایستی ماتریس ناهمگنی، با استفاده از مقادیر شاخص ناهمگنی به‌دست آیند که نمودار این مقادیر در شکل ۵-۱۴ رسم شده‌است. با توجه به شکل ملاحظه می‌شود که تغییر ناحیه رنگ در اطراف نقطه زمانی ۲۰۰ رخ داده و این بدان معنا است که در این ناحیه روند ناهمگنی بین مشاهدت تغییر کرده است. بنابراین همانند قبل، برای تشخیص نقاط تغییر، مقادیر ماتریس ناهمگنی برای مقادیر مختلف B، T و l رسم شده و در هر حالت نقاط تغییر توسط توابع تشخیص این ماتریس مشخص می‌شوند. چند نمونه از توابع تشخیص برای انتخاب مقادیر مختلف پارامترهای موثر (B، T و l) که دارای بهترین عملکرد بودند در شکل (۵-۱۵) رسم شده‌اند و نقاط تغییر برآورد شده توسط هر یک در جدول (۵-۸) آمده‌است.



شکل ۵-۱۴:نمایش ماتریس ناهمگنی سری T_2



شکل ۵-۱۵: نمودار توابع تشخیص ستونی ماتریس ناهمگنی سری T_2

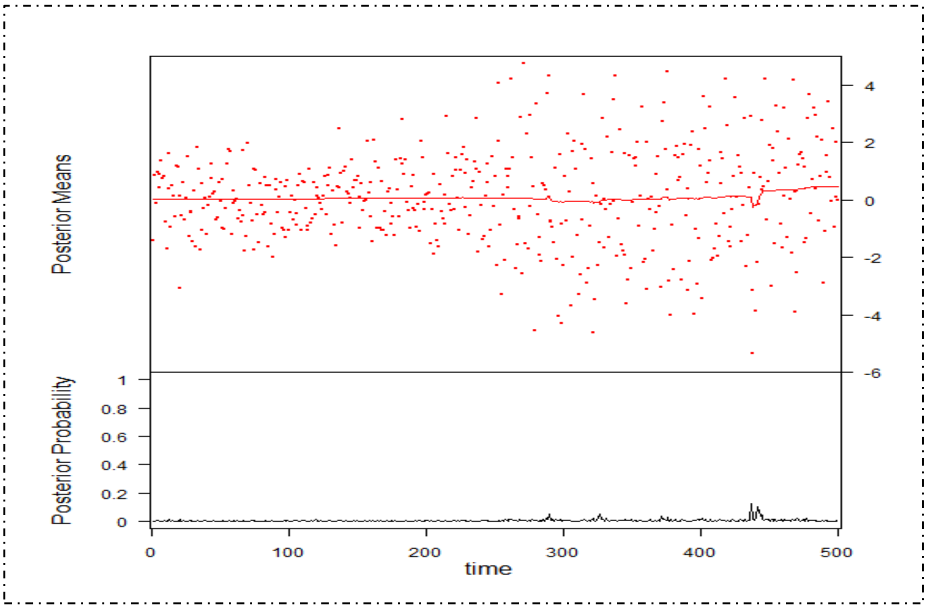
جدول ۵-۸: مقدار ناهمگنی داده‌های سری T_2

	مقدار ناهمگنی به ازای B=40 ,T=41 , l=6	مقدار ناهمگنی به ازای B=50 ,T=51 , l=10	مقدار ناهمگنی به ازای B=60 ,T=61 , l=7
شماره نقطه تغییر	۲۳۲	۲۱۵	۲۲۳
مقدار ناهمگنی	۰/۱۵۴۸	۰/۲۶۱۱	۰/۲۱۱۷

با توجه به نمودار ۵-*15* و جدول ۵-*8* ملاحظه می‌شود که با انتخاب $B = 40$، $T = 41$ و $l = 6$ نقطه زمانی ۲۳۲ به عنوان نقطه تغییر مشخص می‌شود که دارای کمترین مقدار ناهمگنی یعنی ۰/۱۵۴۸ بوده و از این نقطه به بعد روند نمودار ناهمگنی تغییر کرده، یعنی در این نقطه حتما تغییری در مشاهدات رخ داده است. پس این روش در بهترین حالت انتخاب پارامترها توانسته نقطه ۲۳۲ را که نقطه تغییر واقعی $(t = 251)$ نیست به عنوان نقطه تغییر تشخیص دهد.

۵-۳-۳ نتایج روش افراز ضربی

در این‌جا تمام شرایط اعمال شده برای سری T_1 (بخش ۵-۲-۳) نیز، برای سری T_2 در نظر گرفته شده است. با توجه به شکل ۵-۱۶ می‌توان گفت که این روش قادر به تشخیص تغییر در واریانس مشاهدات نیست. زیرا هیچ نقطه‌ای را به عنوان نقطه تغییر تشخیص نمی‌دهد. قابل ذکر است که مقدار میانگین پسین نیز برای مشاهدات برابر ۰٫۰۷۷۴۴ به دست می‌آید که تقریبا برابر میانگین در نظر گرفته شده صفر است.



شکل ۵-۱۶: نمودار واریانس پسین (قاب بالا)و نمودار احتمال پسین نقطه تغییر بودن هر مشاهده (قاب پایین) با استفاده از مدل PPM به ازای $w_\circ = p_\circ = 0/2$

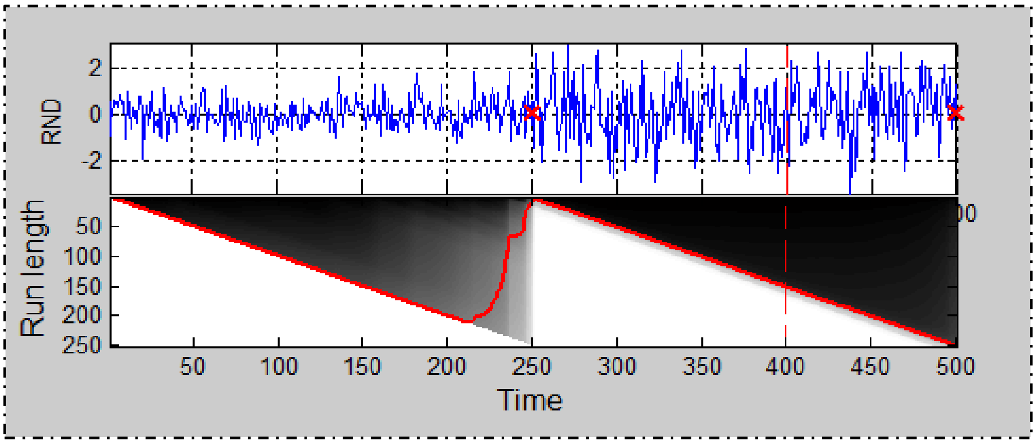
۵-۳-۴ نتایج روش برخط بیزی

با توجه به نتایج مندرج در جدول ۵-۹ می‌توان گفت که مقدار احتمال طول گشت $P(r_t|x_{1:t})$ از نقطه صفر با احتمال ۱٫۰۰۰ شروع شده، در نقطه زمانی ۲۴۸ به مقدار ۰٫۹۸۸۸ رسیده و در نقطه زمانی ۲۴۹ به بیشترین مقدار خود یعنی ۰٫۹۹۳۹ افزایش پیدا می‌کند. اما مقدار این احتمال در نقطه ۲۵۰ کاهش چشم‌گیری داشته و به مقدار صفر می‌رسد و از این نقطه به بعد نیز این کاهش ادامه یافته و به مقدار ۴/۳۵۶۸ e-۶ نزول می‌کند.

جدول۵-۹: احتمال طول گشت $P(r_t|x_{1:t})$، احتمال توزیع پیشگو $P(x_{t+1}|x_{1:t})$، و مقدار تابع مخاطره $P(r_{t-1}|r_t)$

شماره داده	احتمال طول گشت $P(r_t\|x_{1:t})$	احتمال توزیع پیشگو $P(x_{t+1}\|x_{1:t})$	مقدار پارامترهای تابع مخاطره (توزیع لجستیک) $P(r_{t-1}\|r_t)$
۲۴۸	۰٫۹۸۸۸	۰٫۱۴۴۶	-۵٫۲۶۴۵
۲۴۹	۰٫۹۹۳۹	۰٫۱۴۵۰	۰٫۹۹۶۱
۲۵۰	۰	۰٫۱۴۵۳	۰٫۹۹۸۴

در شکل ۵-۱۷، نمودار سری زمانی T_2 (قاب بالا) و مقادیر احتمال طول گشت $P(r_t|x_{1:t})$ (قاب پایین) به ازای تمام طول گشت‌های مختلف با یک طیف رنگی مشخص شده‌است. طول گشت اصلی که با خط مورب قرمز مشخص شده نشان‌دهنده افزایش طول گشت تا نقطه زمانی ۲۴۹ بوده و در نقطه زمانی ۲۵۰ طول آن به صفر می‌رسد و گشت جدیدی شروع می‌شود.



شکل ۵-۱۷: نمودار سری زمانی T_2 (قاب بالا) و نمودار طول گشت‌ها و مقادیر احتمال آن با یک طیف رنگی (قاب پایین)

بنابراین با توجه به توضیحات ذکر شده کاهش یکباره مقدار احتمال در نقطه ۲۵۰ بدان معنا است که این نقطه با مقدار احتمال نزیک به صفر متعلق به این طول گشت بوده و با احتمال ۱٫۰۰ آغازکننده گشت بعدی است یعنی در این نقطه طول گشت صفر شده و طبق مطالب بیان شده در بخش ۴-۲-۲ نقطه زمانی ۲۵۰ به عنوان نقطه تغییر تشخیص داده می‌شود.

۵-۳-۵ خلاصه نتایج برای سری شبیه سازی شده T_2

پس از اعمال هر یک از روش‌ها روی سری شبیه‌سازی‌شده T_2 که در آن سطح واریانس در نقطه $t = 251$ تغییر یافته، می‌توان خلاصه نتایج بخش‌های ۵-۳-۱ تا ۵-۳-۴ را در جدول ۵-۱۰ ملاحظه نمود.

جدول ۵-۱۰: نقطه تغییر واقعی سری شبیه سازی شده T_2 و مقادیر برآورد آن توسط روش‌های مختلف

روش	نام روش	نقطه تغییر مشخص شده	نقطه تغییر اصلی
۱	روش *CUSUM*	۲۴۵	۲۵۱
۲	روش *SSA*	۲۳۲	
۳	روش *PPM*	-	
۴	روش *BOCPD*	۲۵۰	

با توجه به جدول ملاحظه می‌شود که اگر چه نقاط تغییر تعیین شده توسط سه روش *SSA*، *CUSUM* و *BOCPD*، تفاوت‌هایی با زمان واقعی تغییر دارند اما به هر حال قابلیت تشخیص تغییر در واریانس را داشته‌اند. این در حالی است که روش *PPM* فاقد چنین قابلیتی می‌باشد.

همان طور که مشخص است روش *BOCPD* عملکرد مناسب‌تری نسبت به روش‌های *SSA*، *PPM* و *CUSUM* برای تشخیص تغییر در سطح واریانس سری دارد، زیرا نقطه زمانی ۲۵۰ را که بسیار نزدیک نقطه تغییر واقعی $(t = ۲۵۱)$ است به عنوان نقطه تغییر تشخیص داده است.

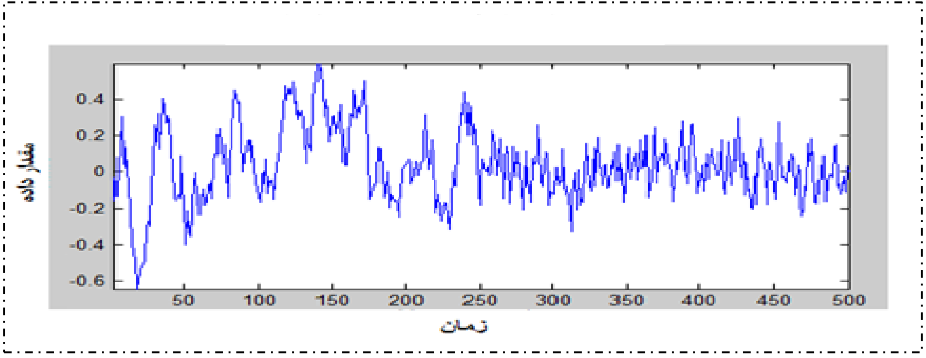
۵-۴ نتایج تشخیص تغییر در خودهمبستگی سری زمانی

سری‌های شبیه‌سازی‌شده در بخش‌های قبل به این صورت در نظر گرفته شدند که در سری T_1، سطح میانگین و در سری T_2، سطح واریانس سری تغییر یافته بود. در این بخش سری شبیه‌سازی‌شده دیگری را در نظر می‌گیریم که تغییر، در خود همبستگی آن رخ داده است. بدین منظور سری T_3 با طول ۵۰۰ را به با استفاده از فرآیند اتورگرسیو $x_{t+1} = \beta x_t + e$ شبیه‌سازی می‌کنیم، که در

آن$e\sim N(۰,۰/1)$ است و پارامتر β در ۲۵۰ نمونه اول برابر β=۰/۹ و در ۲۵۰ نمونه دوم برابر با، β=۰/۵ فرض شده‌است. از آن‌جا که در فرآیند اتورگرسیو داریم

$$\rho_k = \alpha^{|k|} \qquad\qquad k = 0, \pm1, \pm2, \ldots$$

لذا با توجه به سری شبیه‌سازی‌شده می‌توان گفت که در نقطه $t = ۲۵۱$، تغییری در خودهمبستگی سری (تغییر در مقدار β فرآیند اتورگرسیو) رخ داده است. نمودار این سری زمانی شبیه‌سازی‌شده در شکل ۵-۱۸ نمایش داده شده است.

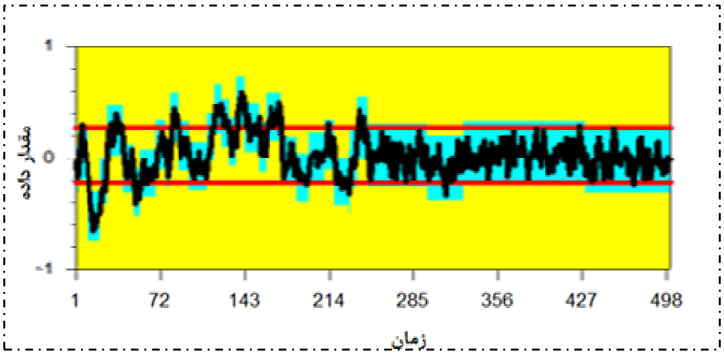


شکل ۵-۱۸: نمودار سری زمانی T_3، که خودهمبستگی آن در نقطه زمانی ۲۵۱ تغییر یافته است

اکنون پس از معرفی سری T_3، نتایج اعمال هر روش روی این سری را در زیر بررسی می‌کنیم.

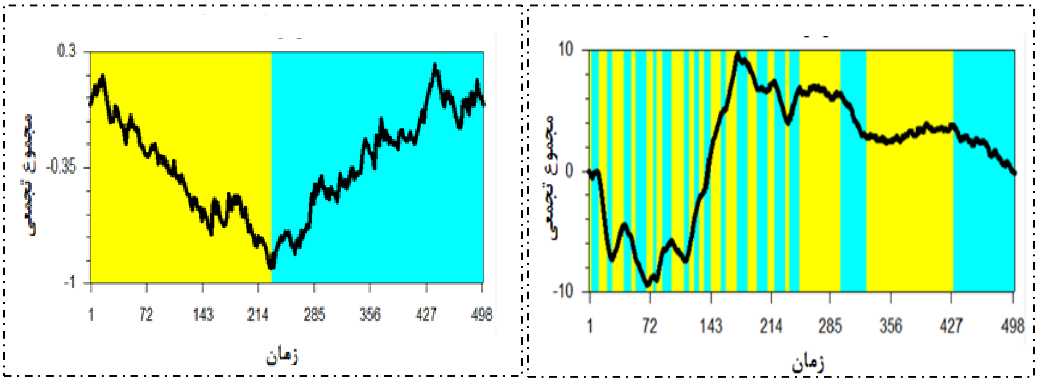
۵-۴-۱ نتایج روش *CUSUM* مبتنی بر نمونه‌های خودگردان

همانند قبل، نمودار کنترل سری T_3 را که در شکل ۵-۱۹نشان داده شده، در نظر بگیرید. ملاحظه می‌شود، علیرغم این که نمودار کنترل زمان‌های بسیاری را به عنوان نقاط تغییر تشخیص داده، اما قادر به تشخیص زمان واقعی تغییر نمی‌باشد.



شکل۵-۱۹: نمودار کنترل سری T_3

در شکل ۵-۲۰ نمودار $CUSUM$ این داده‌ها برای تشخیص تغییر در سطح میانگین (نمودار (الف)) و سطح واریانس (نمودار (ب))، نشان داده شده است. همان طور که مشاهده می‌شود، شیب نمودار (الف) در اطراف بسیاری از نقاط سری T_3 تغییر کرده و این بدان معنا است که در آن نقاط تغییری در میانگین سری رخ داده است. شیب نمودار (ب) تنها در اطراف نقطه زمانی ۲۳۱ تغییر کرده و نبدین ترتیب شان می‌دهد که واریانس سری در این نقطه زمانی تغییر یافته است که در واقع چنین نیست. برای حصول اطمینان از وجود تغییر در این نقاط سطح و فاصله اطمینانی برای این تغییرات محاسبه‌شده و نتایج برای تغییرات میانگین در جدول ۵-۱۲ و برای تغییرات واریانس در جدول ۵-۱۱ آمده است.



(الف) (ب)

شکل۵-۲۰: (الف) نمودار CUSUM سری T_3، برای تشخیص تغییر در میانگین سری و (ب) برای تشخیص تغییر در واریانس سری

جدول ۵-۱۲: نقطه تغییر تعیین شده، سطح اطمینان و فاصله اطمینان برای بررسی تغییر در واریانس سری T_3

شماره داده	فاصله اطمینان	سطح اطمینان	تغییر واریانس از	تغییر واریانس به
۲۳۱	(۳۷۷ ,۱۳۳)	٪۹۱	۰٫۰۷۲۸۴۷	۰٫۰۹۶۴۴

با توجه به جدول ملاحظه می‌شود که واریانس مشاهدات در نقطه زمانی ۲۳۱ از مقدار ۰٫۰۷۲۸۴۷ به مقدار ۰٫۰۹۶۴۴ افزایش یافته و فاصله اطمینان بین نقاط (۱۳۳ ,۳۷۷) با سطح اطمینان ٪۹۱ تعیین شده است. شایان ذکر است که این روش نقاط بسیاری را به عنوان نقاط تغییر در میانگین سری تعیین کرده که نقطه تغییر واقعی نیستند. بنابراین با توجه به نتایج بیان‌شده می‌توان گفت که این روش قادر به تشخیص تغییر در خود همبستگی سری نیست.

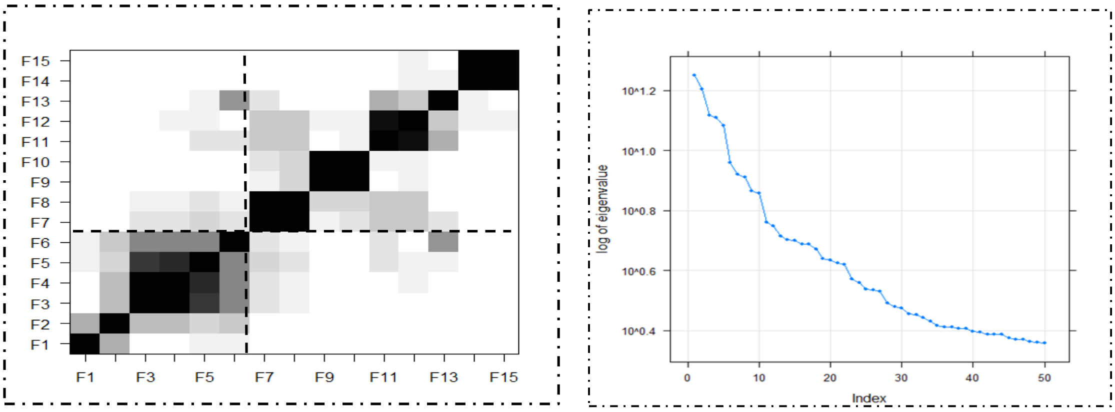
جدول ۵-۱۱: نقطه تغییر تعیین شده، سطح اطمینان و فاصله اطمینان برای بررسی تغییر در میانگین سری T_3

شماره داده	فاصله اطمینان	سطح اطمینان	تغییر میانگین از	تغییر میانگین به
۴	(۱۱ , ۴)	٪۹۲	-۰٫۰۹۹۸۹۷	۰٫۰۴۴۲۳۳
۱۳	(۱۳ , ۱۳)	٪۱۰۰	۰٫۰۴۴۲۳۳	-۰٫۵۲۴۲۶
۲۳	(۲۳ , ۲۲)	٪۹۸	-۰٫۵۲۴۲۶	-۰٫۲۰۰۵۷
۲۹	(۳۰ , ۲۹)	٪۹۸	-۰٫۲۰۰۵۷	۰٫۲۵۱۴۷
۴۲	(۴۲ , ۴۱)	٪۱۰۰	۰٫۲۵۱۴۷	-۰٫۰۹۳۶۸۸
۵۱	(۵۱ , ۴۹)	٪۹۵	-۰٫۰۹۳۶۸۸	-۰٫۳۰۷۵۸
۵۶	(۵۹ , ۵۶)	٪۹۶	-۰٫۳۰۷۵۸	-۰٫۱۲۸۷۹
۶۹	(۶۹ , ۶۸)	٪۹۱	-۰٫۱۲۸۷۹	۰٫۱۲۶۱۹
۷۷	(۷۷ , ۷۵)	٪۹۱	۰٫۱۲۶۱۹	-۰٫۰۴۲۳۳۲
۸۱	(۸۱ , ۸۱)	٪۹۸	-۰٫۰۴۲۳۳۲	۰٫۳۶۳۶۵
۸۸	(۸۸ , ۸۸)	٪۱۰۰	۰٫۳۶۳۶۵	۱۱۴۴۲ر
۹۸	(۱۰۰ , ۹۷)	٪۱۰۰	۰٫۱۱۴۴۲	-۰٫۰۷۲۹۵۱
۱۱۳	(۱۱۳ , ۱۱۱)	٪۹۹	-۰٫۰۷۲۹۵۱	۰٫۱۹۱۱۳
۱۱۹	(۱۱۹ , ۱۱۸)	٪۹۳	۰/۱۹۱۱۳	۰٫۴۴۵۸۵
۱۲۵	(۱۲۶ , ۱۲۵)	٪۹۸	۰٫۴۴۵۸۵	۰٫۳۱۹۱۲
۱۳۱	(۱۳۲ , ۱۳۱)	٪۹۷	۰٫۳۱۹۱۲	۱۵۱۳۹ر
۱۳۷	(۱۳۷ , ۱۳۷)	٪۱۰۰	۰٫۱۵۱۳۹	۰٫۵۰۶۲۷
۱۴۴	(۱۴۵ , ۱۴۴)	٪۱۰۰	۰٫۵۰۶۲۷	۰٫۲۸۷۳
۱۵۶	(۱۵۷ , ۱۵۴)	٪۹۸	۰٫۲۸۷۳	۰٫۱۰۴۳۴

۱۶۲	(۱۶۲ , ۱۶۳)	۹۴٪	۰٫۱۰۴۳۴	۰٫۳۶۸۷
۱۷۵	(۱۷۵ , ۱۷۵)	۱۰۰٪	۰٫۳۶۸۷	-۰٫۰۲۸۷۰۳
۱۸۸	(۱۸۳ , ۱۸۸)	۹۶٪	-۰٫۰۲۸۷۰۳	-۰٫۱۶۳۲۷
۱۹۸	(۱۹۸ , ۱۹۹)	۱۰۰٪	-۰٫۱۶۳۲۷	۰٫۰۲۰۷۳۶
۲۱۱	(۲۰۸ , ۲۱۴)	۹۰٪	۰٫۰۲۰۷۳۶	۰٫۱۲۵۳۳
۲۱۹	(۲۱۹ , ۲۱۹)	۱۰۰٪	۰٫۱۲۵۳۳	-۰/۲۰۲۹۳
۲۳۲	(۲۳۱ , ۲۳۲)	۹۶٪	-۰/۲۰۲۹۳	۰٫۰۱۹۲۸
۲۳۸	(۲۳۷ , ۲۴۱)	۹۹٪	۰٫۰۱۹۲۸	۰٫۲۷۰۶۳
۲۴۹	(۲۴۹ , ۲۶۰)	۱۰۰٪	۰٫۲۷۰۶۳	۰٫۰۲۳۵۸
۲۹۷	(۲۸۴ , ۳۰۲)	۱۰۰٪	۰٫۰۲۳۵۸	-۰٫۰۸۳۸۹۳

۵-۴-۲ نتایج روش تحلیل طیفی منفرد

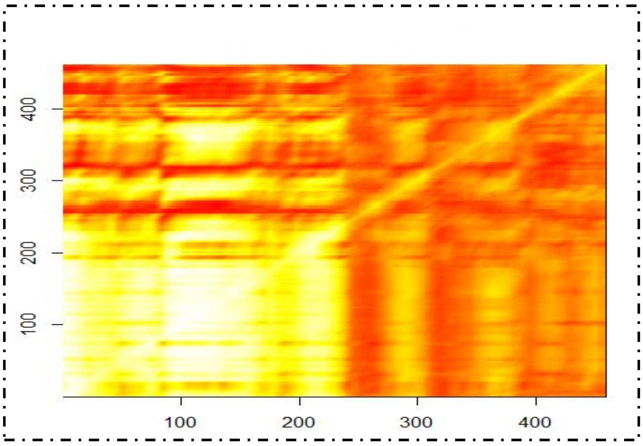
همانند قبل، این روش با انتخاب مقادیر مختلف B و T اجرا شده و پارامتر l نیز با دو روش، l مقدار ویژه بزرگتر و ماتریس W-همبستگی تعیین می‌گردد. نمودار هر دو روش در شکل ۵-۲۱ رسم شده‌اند. نمودار (الف) نشان می‌دهد که ۱۰ مقدار ویژه اول دارای درصد قابل قبولی هستند و با استفاده از روش W-همبستگی و تفکیک‌پذیری که روی نمودار (ب) انجام شده نیز دیده می‌شود که ۷ مولفه اول می‌توانند به‌خوبی بیان‌کننده سری واقعی داده‌ها باشند.



(ب) (الف)

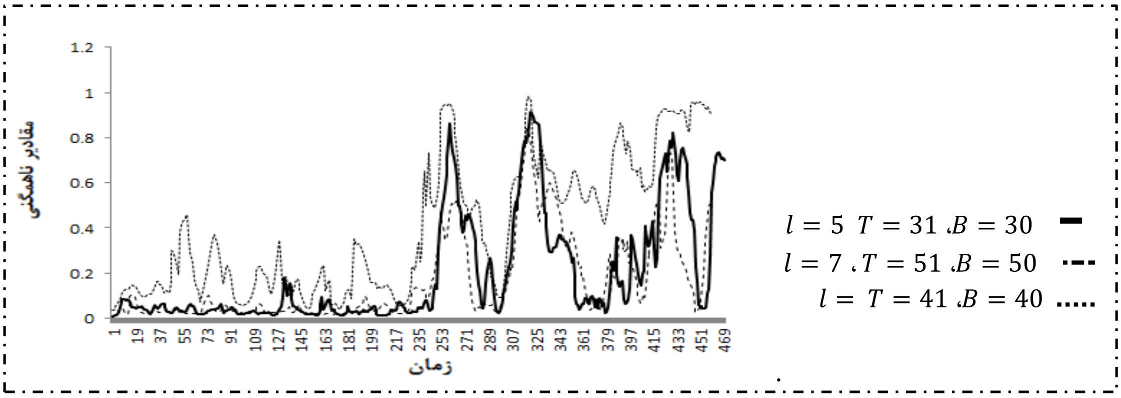
شکل ۵-۲۱: نمودار مقادیر ویژه سری T_3 برای تشخیص تعداد بردارهای مولفه‌های اصلی

پس از انتخاب مقادیر مختلف برای پارامترهای B، T و l ، ماتریس ناهمگنی در بهترین حالت در شکل ۵-۲۲ رسم شده‌است. با توجه به شکل ملاحظه می‌شود که تغییر رنگی در اطراف نقطه زمانی ۲۵۰ رخ داده و این بدان معنا است که در این ناحیه روند ناهمگنی بین مشاهدت تغییر کرده است.



شکل ۵-۲۲: نمایش ماتریس ناهمگنی سری T_3

برای هر نوع تنظیم پارامترها، ماتریس ناهمگنی مشاهدات رسم شده و نقاط تغییر با استفاده از توابع تشخیص، محاسبه شده است. چند نمونه از این توابع برای بهترین انتخاب پارامترها در شکل ۵-۲۳ دیده می‌شود. نقاط تغییر تشخیص داده شده توسط هر یک نیز در جدول (۵-۱۳) آمده‌است.



شکل ۵-۲۳: نمودار توابع تشخیص سطری ماتریس ناهمگنی سری T_3

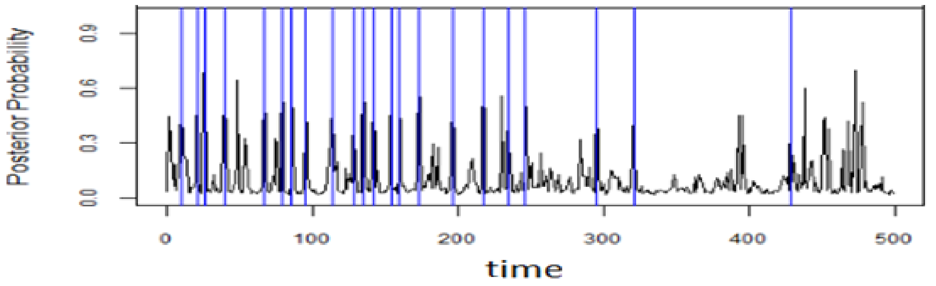
جدول ۵-۱۳: مقدار ناهمگنی در داده‌های سری T_3

	مقدار ناهمگنی به ازای B=30 ,T=31 , l=6	مقدار ناهمگنی به ازای B=40 ,T=41 , l=10	مقدار ناهمگنی به ازای B=50 ,T=51 , l=7
شماره نقطه تغییر	۲۴۶	۲۲۳	۲۴۳
مقدار ناهمگنی	۰٫۰۴۲۶۹	۰٫۰۵۷۶۵	۰٫۰۸۷۷۶

با توجه به نمودار ۵-۲۳ و جدول ۵-۱۳ می‌توان گفت که با انتخاب پارامترهای $B = 30$، $T = 31$ و $l = 6$ ، نقطه زمانی ۲۴۶ با کمترین مقدار ناهمگنی یعنی ۰٫۰۴۲۶۹ به عنوان بهترین برآورد نقطه تغییر واقعی $(t = 251)$ معرفی می‌شود. از این نقطه به بعد روند نمودار ناهمگنی تغییر کرده و این دلیل دیگری است برای این که سری در این نقطه دچار تغییر شده است.

۵-۴-۳ نتایج روش افراز ضربی

در این‌جا تمام شرایط اعمال شده توسط این روش برای سری‌های T_1 و T_2، برای سری T_3 نیز در نظر گرفته شده است. نتایج این روش، برای مقادیر پارامترهای p_0 و w_0 برابر ۰٫۱، ۰٫۲ و ۰٫۳ در جدول ۵-۱۴ آمده‌است. این جدول شامل مقدار احتمال‌های پسین (احتمال نقطه تغییر بودن هر مشاهده)، برای نقاطی است که به عنوان نقاط تغییر، تشخیص داده شده‌اند. برای تشخیص مقدار احتمال پسین در هر نقطه با استفاده از نمونه‌گیری گیبز در روش $MCMC$، ۲۰۰۰۰ نمونه در نظر گرفته شده است. با توجه به جدول مشاهده می‌شود که این روش نقاط بسیاری را به عنوان نقطه تغییر تشخیص می‌دهد که هیچ کدام نقطه تغییر واقعی نیستند. نمودار این احتمال پسین برای تمام مشاهدات در حالت $w_0 = p_0 = 0/2$ در شکل ۵-۲۴ رسم شده‌است. در این نمودار نقاط تغییر با خطوط عمودی آبی رنگ مشخص شده اند.



شکل ۵-۲۴: نمودار احتمال پسین نقطه تغییر بودن هر مشاهده با استفاده از مدل PPM به ازای $w_\circ = p_\circ = 0/2$

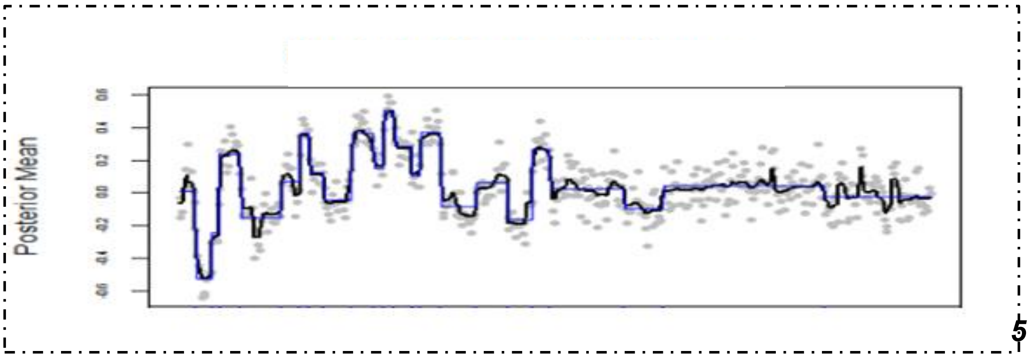
جدول ۵-۱۴: مقدار احتمال پسین برای نقاط تغییر برآورد شده به ازای $w_\circ = p_\circ = 0/1, 0/2, 0/3$

شماره داده	$w_\circ = 0/2$ $p_\circ = 0/2$	$w_\circ = 0/1$ $p_\circ = 0/1$	$w_\circ = 0/3$ $p_\circ = 0/3$
۱۲	۰٫۵۵۶	۰٫۴۳۶	۰٫۵۲۸
۲۲	۰٫۸۴۰	۰٫۸۱۴	۰٫۸۶۲
۲۷	۰٫۳۴۴	۰٫۳۴۶	۰٫۶۳۶
۴۱	۰٫۵۹۴	۰٫۳۳۲	۰٫۷۷۴
۶۸	۰٫۸۳۸	۰٫۸۰۸	۰٫۸۱۲
۸۰	۰٫۸۴۶	۰٫۸۷۰	۰٫۸۷۲
۸۷	۰٫۷۶۰	۰٫۷۱۰	۰٫۷۱۴
۹۶	۰٫۴۶۶	۰٫۴۲۸	۰٫۴۸۶
۱۱۵	۰٫۶۴۸	۰٫۶۴۸	۰٫۵۸۶
۱۳۰	۰٫۴۶۴	۰٫۳۸۴	۰٫۴۷۸
۱۳۶	۰٫۸۵۰	۰٫۸۵۰	۰٫۸۶۴
۱۴۳	۰٫۶۶۴	۰٫۷۴۰	۰٫۵۹۸
۱۵۵	۰٫۸۲۸	۰٫۸۳۴	۰٫۸۴۲
۱۶۱	۰٫۷۸۴	۰٫۸۲۶	۰٫۷۴۸
۱۷۴	۰٫۸۴۲	۰٫۸۶۶	۰٫۸۰۲
۱۹۷	۰٫۵۳۸	۰٫۷۷۸	۰٫۵۹۴
۲۱۸	۰٫۹۲۴	۰٫۹۷۲	۰٫۹۴۸
۲۳۵	۰٫۶۳۴	۰٫۶۵۲	۰٫۷۴۸
۲۴۷	۰٫۶۵۰	۰٫۶۲۴	۰٫۶۶۴
۳۲۲	۰٫۳۰۸	۰٫۳۳۰	۰٫۳۱۶
۳۴۹	۰٫۴۲۰	۰٫۳۶۸	۰٫۳۸۲

همان طور که در بخش ۴-۱-۵ گفته شده، این روش از دیدگاه دوم، تغییر در پارامتر توزیع مشاهدات را نیز برای تشخیص نقاط تغییر بررسی می‌کند. در این‌جا برای زمانی‌هایی که به عنوان نقاط تغییر تعیین شده‌اند، مقدار مورد انتظار میانگین مشاهدات در جدول ۵-۱۵ به ازای $w_{۰} = p_{۰} = ۰/2$ آمده است. با توجه به جدول ملاحظه می‌شود که بر اساس این روش نیز نقاط بسیاری به عنوان نقاط تغییر تشخیص داده می‌شود که هیچ یک نقطه تغییر واقعی $(t = ۲۵۱)$ نیستند. . نمودار مقادیر مورد انتظار برای تمام مشاهدات در حالت $w_{۰} = p_{۰} = ۰/2$ در شکل ۵-۲۵ رسم شده است.

جدول ۵-۱۵: مقدار تغییر در میانگین پسین، برای نقاط تغییر تعیین شده در حالت $w_{۰} = p_{۰} = 0/2$

شماره داده	تغییرمیانگین پسین	میانگین پسین
۱۲	۰٫۰۰۸۲۰۰۷	-۰٫۰۵۲۴۲
۲۲	-۰٫۰۵۲۴۲	-۰٫۲۴۵۳۸
۲۷	-۰٫۲۴۵۳۸	-۰٫۲۳۵۱۸
۴۱	-۰٫۲۳۵۱۸	-۰٫۱۵۰۱۹
۶۸	-۰٫۱۵۰۱۹	۰٫۰۷۰۰۱
۸۰	۰٫۰۷۰۰۱	۰٫۳۶۳۶۵
۸۷	۰٫۳۶۳۶۵	۰٫۱۲۳۹۳
۹۶	۰٫۱۲۳۹۳	-۰٫۰۴۲۸۹۷
۱۱۵	-۰٫۰۴۲۸۹۷	۰٫۳۶۵۷۴
۱۳۰	۰٫۳۶۵۷۴	۰٫۱۵۱۳۸۹
۱۳۶	۰٫۱۵۱۳۸۹	۰٫۵۰۶۲۷
۱۴۳	۰٫۵۰۶۲۷	۰٫۲۸۷۳۰
۱۵۵	۰٫۲۸۷۳۰	۰٫۱۰۴۳۴
۱۶۱	۰٫۱۰۴۳۴	۰٫۳۶۸۷
۱۷۴	۰٫۳۶۸۷	-۰٫۰۸۷۲۱
۱۹۷	-۰٫۰۸۷۲۱	۰٫۰۶۰۵۸
۲۱۸	۰٫۰۶۰۵۸	-۰٫۱۶۶۷۴
۲۳۵	-۰٫۱۶۶۷۴	۰٫۲۶۷۲۶
۲۴۷	۰٫۲۶۷۲۶	۰٫۰۲۴۷۷
۳۲۲	۰٫۰۲۴۷۷	-۰٫۱۰۰۶۳
۲۴۹	-۰٫۱۰۰۶۳	۰٫۰۴۰۳۶



شکل ۵-۲۵: نمودار مقادیر مورد انتظار میانگین برای هر مشاهده با استفاده از مدل PPM به ازای $w_\circ = p_\circ = 0/2$

با توجه به جدول ۵-۱۴ و ۵-۱۵، ملاحظه می‌شود که روش PPM، با استفاده از هر دو دیدگاه نتوانسته تغییر در خودهمبستگی سری را به خوبی تشخیص دهد، زیرا نقاط بسیاری را به عنوان نقطه تغییر معرفی می‌کند در حالی نقطه تغییر واقعی $t = 251$ است.

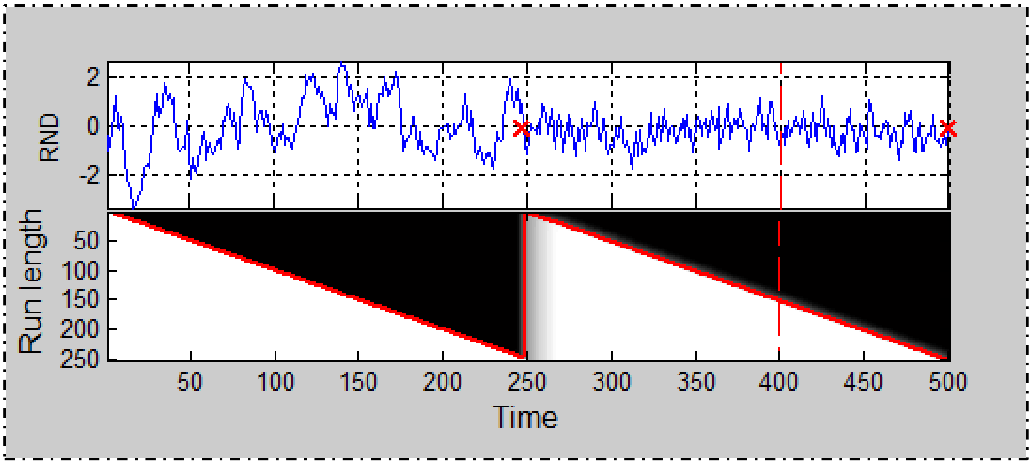
۵-۴-۴ نتایج روش برخط بیزی

در این بخش روند بخش‌های ۵-۲-۴ و ۵-۳-۴ را دنبال می‌کنیم. با توجه به جدول ۵-۱۶ ملاحظه می‌شود که مقدار احتمال طول گشت $P(r_t|x_{1:t})$ از نقطه صفر با احتمال ۱٫۰۰۰ شروع شده و در نقطه زمانی ۲۴۵ به مقدار ۰٫۹۸۵۴ رسیده و در نقطه زمانی ۲۴۶ به بیشترین مقدار خود یعنی ۰٫۹۹۸۲ افزایش پیدا می‌کند. اما مقدار این احتمال در نقطه زمانی ۲۴۷ کاهش چشمگیری داشته و به مقدار صفر می‌رسد. از این نقطه به بعد نیز کاهش ادامه یافته و به مقدار ۳/۳۰۰۴۳e-۰۵ و کمتر نزول می‌کند.

جدول۵-۱۶: احتمال طول گشت $P(r_t|x_{1:t})$، احتمال توزیع پیشگو $P(x_{t+1}|x_{1:t})$، و مقدار تابع مخاطره $P(r_{t-1}|r_t)$

شماره داده	احتمال طول گشت $P(r_t\|x_{1:t})$	احتمال توزیع پیشگو $P(x_{t+1}\|x_{1:t})$	مقدار پارامترهای تابع مخاطره (توزیع لجستیک $P(r_{t-1}\|r_t)$
۲۴۵	۰٫۹۸۵۴	۰٫۰۳۰۴	-۸٫۷۱۰۲
۲۴۶	۰٫۹۹۸۲	۰٫۱۰۰۶	۰٫ ۹۸۴
۲۴۷	۰	۰٫۱۶۳۸	۰٫۹۹۰۵

بنابراین با توجه به کاهش چشم‌گیر مقدار احتمال طول گشت در نقطه ۲۴۷ می‌توان گفت که این نقطه متعلق به این طول گشت نبوده و با احتمال یک، نقطه آغاز یک گشت جدید است. بنابراین نقطه ۲۴۷ به عنوان نقطه تغییر تشخیص داده می‌شود. در شکل ۵-۲۶ نمودار سری زمانی T_3 (قاب بالا) و مقادیر احتمال تمام طول گشت‌های مختلف (قاب پایین) با یک طیف رنگی مشخص شده‌اند. در قاب پایین این نمودار، محتمل‌ترین طول گشت با خط مورب به رنگ قرمز مشخص شده و ملاحظه می‌شود که تغییر در زمانی رخ می‌دهد که گشت جدیدی آغاز شده است.



شکل ۵-۲۶: نمودار سری T_3 (قاب بالا) و نمودار طول گشت‌ها و مقادیر احتمال آن با یک طیف رنگی (قاب پایین)

بر اساس به جدول ۵-۱۶ و شکل ۵-۲۶ می‌توان این طور استنباط نمود که این روش توانسته است نقطه زمانی ۲۴۷ را که نزدیک به نقطه تغییر واقعی $(t = 251)$ است به عنوان نقطه تغییر تشخیص دهد. لذا می‌توان گفت که این روش، وجود تغییر در خود همبستگی سری را به خوبی تشخیص داده است.

۵-۴-۵ خلاصه نتایج برای سری شبیه‌سازی شده T_3

خلاصه نتایج هریک از روش‌ها روی سری T_3 که تغییری در خود همبستگی آن رخ داده، در جدول ۵-۱۷ آورده شده است. با توجه به جدول ملاحظه می‌شود که دو روش CUSUM و *PPM* دارای

عملکرد بسیار ضعیفی بوده و نتوانسته‌اند تغییر را تشخیص دهند. اما دو SSA و $BOCPD$ به خوبی توانستند تغییر در خودهمبستگی سری را تشخیص داده و به ترتیب نقطه زمانی ۲۴۶ و ۲۴۸ را به عنوان نقطه تغییر تعیین نمایند، که البته با نقطه تغییر واقعی $(t = ۲۵۱)$ کمی فاصله دارد. بنابراین روش $BOCPD$ عملکرد مناسب‌تری نسبت به سه روش دیگر برای تشخیص تغییر در خودهمبستگی سری دارد.

جدول ۵-۱۷: نقطه تغییر واقعی سری شبیه‌سازی‌شده T_3 و مقادیر برآورد آن توسط روش‌های مختلف

روش	نام روش	نقطه تغییر مشخص شده	نقطه تغییر اصلی
۱	روش $CUSUM$	۴-۱۳-۲۳-۲۹-۴۲-۵۱-۵۶-۶۹-۷۷-۸۱-۸۸-۹۸-۱۱۳-۱۱۹-۱۲۵-۱۳۱-۱۳۷-۱۴۴-۱۵۶-۱۶۲-۱۷۵-۱۸۸-۱۹۸-۲۱۱-۲۳۲-۲۳۸-۲۴۹-۲۹۷	۲۵۱
۲	روش SSA	۲۴۶	
۳	روش PPM	۱۲-۲۲-۲۷-۴۱-۶۸-۸۰-۸۷-۹۶-۱۱۵-۱۳۰-۱۳۶-۱۴۳-۱۵۵-۱۶۱-۱۷۴-۱۹۷-۲۱۸-۲۳۵-۲۴۷-۲۹۶-۳۲۲-۴۲۹	
۴	روش $BOCPD$	۲۴۷	

۵-۵ مقایسه روش‌ها توسط نتایج سری‌های شبیه‌سازی

با توجه به نتایج اعمال روش‌های بیان‌شده برای تشخیص نقطه تغییر روی سری‌های شبیه‌سازی شده در بخش‌های قبل، به بررسی قابلیت تشخیص تغییر این روش‌ها می‌پردازیم. خلاصه نتایج اعمال هر یک از روش‌های ذکرشده بر روی سری‌های زمانی شبیه‌سازی شده T_1، T_2 و T_3 در جدول ۵-۱۸ آورده شده است.

جدول ۵-۱۸: نقاط تغییر تعیین شده توسط روش‌های ذکرشده روی سری‌های زمانی T_1، T_2، T_3 و T_4

روش	نام روش	سری T_1	سری T_2	سری T_3
۱	روش *CUSUM*	۲۵۱	۲۴۵	۴-۱۳-۲۳-۲۹-۴۲-۵۱-۵۶-۶۹-۷۷-۸۱-۸۸-۹۸-۱۱۳-۱۱۹-۱۲۵-۱۳۱-۱۳۷-۱۴۴-۱۵۶ -۱۶۲-۱۷۵-۱۸۸-۱۹۸-۲۱۱-۲۳۲-۲۳۸-۲۴۹-۲۹۷
۲	روش *SSA*	۲۴۵	۲۳۲	۲۴۶
۳	روش *PPM*	۲۵۰	-	۱۲-۲۲-۲۷-۴۱-۶۸-۸۰-۸۷-۹۶-۱۱۵-۱۳۰-۱۳۶-۱۴۳-۱۵۵-۱۶۱-۱۷۴-۱۹۷ -۲۱۸-۲۳۵-۲۴۷-۲۹۶- ۳۲۲-۴۲۹
۴	روش *BOCPD*	۲۵۰	۲۵۰	۲۴۷

با توجه به جدول می‌توان مقایسه‌ای بین این روش‌ها انجام داد و از بین آن‌ها مناسب‌ترین روش را برای تشخیص نقاط تغییر مجموعه داده واقعی انتخاب کرد. همان طور که در جدول ملاحظه می‌شود، برای سری T_1 که تغییری در سطح میانگین آن رخ داده است، تمام روش‌ها عملکرد مناسبی داشته و توانستند وجود تغییر در میانگین سری را تشخیص دهند. ولی روش *CUSUM* نسبت به سه روش دیگر عملکرد بهتری داشته و توانسته نقطه تغییر واقعی $t = 251$ را به طور دقیق، تعیین کند.

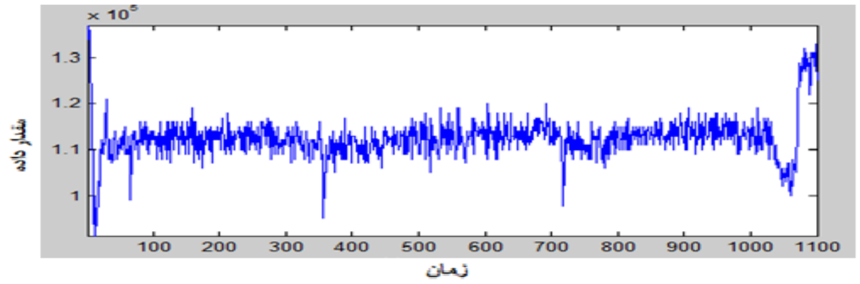
با توجه به ستون دوم جدول، در حالتی که تغییری در سطح واریانس سری رخ داده است (سری T_2) روش‌های *CUSM*، *SSA* و *BOCPD* توانسته‌اند به ترتیب نقاط زمانی، ۲۴۵، ۲۳۲ و ۲۵۰ را به عنوان نقاط تغییر تشخیص دهند، ولی مدل *PPM* فاقد توانایی تشخیص نقطه تغییر است. بنابراین روش *BOCPD*، دارای عملکرد مناسب‌تری نسبت به سایر روش‌ها است. زیرا نقطه زمانی$t = 250$ را که فاصله اندکی با نقطه تغییر واقعی $(t = 251)$ دارد به عنوان نقطه تغییر تشخیص داده است.

در ستون سوم جدول که شامل نتایج مربوط به سری T_3 می‌باشد تغییری در خودهمبستگی سری رخ داده است و ملاحظه می‌شود که دو روش *SSA* و *BOCPD* توانسته‌اند به ترتیب نقاط زمانی ۲۴۶ و ۲۴۷ را به عنوان نقطه تغییر تشخیص دهند که در فاصله کمی از نقطه تغییر واقعی $(t = 251)$ قرار دارند. بنابراین *BOCPD* عملکرد بهتری نسبت به روش *SSA* دارد، زیرا نقطه زمانی ۲۴۷ در فاصله کمتری از نقطه ۲۵۱ واقع می‌شود. دو روش دیگر، نقاط متعددی را به عنوان نقطه تغییر معرفی می‌کنند و از آن‌جا که فقط یک تغییر در سری وجود دارد لذ این دو روش دارای عملکرد نامناسبی هستند.

با توجه به مطالب ذکر شده می‌توان چنین نتیجه گیری کرد، در حالتی که تغییری در سطح میانگین، واریانس و خود همبستگی سری رخ داده است روش *BOCPD* عملکرد مناسب‌تری را نسبت به روش‌های دیگر دارد. بنابراین برای تشخیص نقاط تغییر مجموعه داده واقعی T_4 که نوع تغییر در آن مشخص نیست می‌توان از این روش استفاده کرد. نقاط تغییر تشخیص داده شده توسط این روش بر روی سری T_4 به‌طور مسلم نسبت به روش‌های دیگر محتمل‌ترین نقاط تغییر خواهند بود. بنابراین در بخش بعد این موضوع را مورد بررسی قرار می‌دهیم.

۵-۶ نتایج تشخیص تغییر در سری واقعی

در این بخش، داده‌های واقعی مربوط به اندازه‌گیری واکنش مغناطیسی هسته‌ای در حفاری چاه که مربوط به تعیین نوع لایه‌های پوسته‌ی زمین است، استفاده می‌کنیم. این سری زمانی که آن را سری T_4 می‌نامیم، شامل ۱۱۰۰ مقدار اندازه‌گیری شده است که نمودار زمانی آن در شکل ۵-۲۷ نمایش داده شده است.



شکل ۵-۲۷: نمودار سری زمانی T_4 شامل ۱۱۰۰ مقدار اندازه‌گیری مغناطیسی هسته‌ای در حفاری چاه

با توجه به نتایج سری‌های شبیه‌سازی‌شده بخش قبل ملاحظه گردید که از بین روش‌های ذکرشده روش تشخیص برخط نقطه تغییر بیزی دارای عملکرد مناسب‌تری نسبت به سه روش دیگر است. بنابراین امید است که نقاطی که توسط این روش به عنوان نقطه تغییر معرفی می‌شوند (طبق مطالعه سری‌های شبیه‌سازی‌شده) فاصله بسیار اندکی با نقاط تغییر واقعی داشته باشند.

۵-۶-۱ نتایج روش تشخیص برخط بیزی

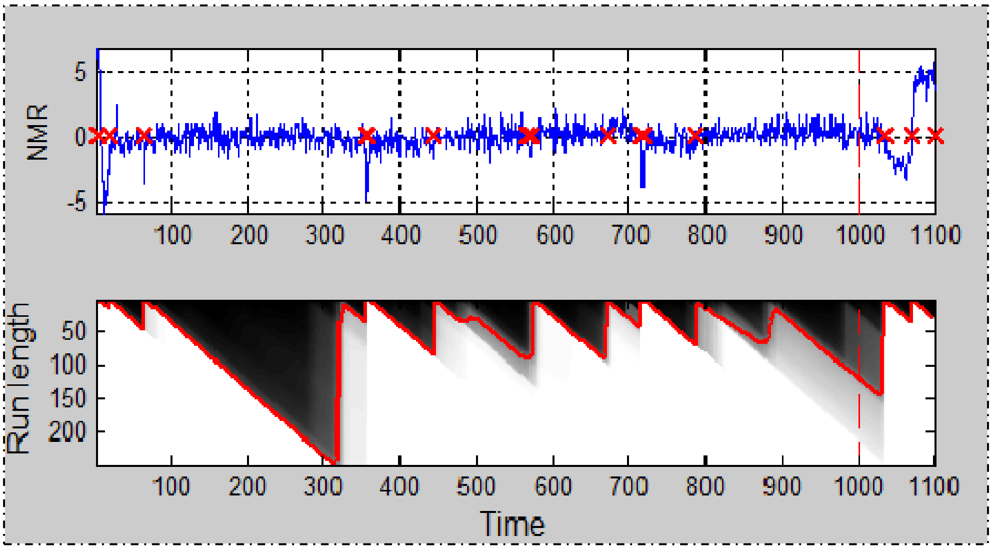
همانند قبل، جدول ۵-۱۹ شامل مقادیر احتمال طول گشت $P(r_t|x_{1:t})$، احتمال توزیع پیشگو $P(x_{t+1}|x_{1:t})$، و تابع مخاطره $P(r_{t-1}|r_t)$ به کار برده شده است. در واقع این جدول تنها شامل نقاطی است که مقدار احتمال طول گشت آن‌ها کاهش چشمگیری یافته و به عنوان نقاط تغییر تعیین شده‌اند.

جدول ۵-۱۹: احتمال طول گشت $P(r_t|x_{1:t})$، احتمال توزیع پیشگو $P(x_{t+1}|x_{1:t})$، و مقدار تابع مخاطره $P(r_{t-1}|r_t)$

شماره داده	احتمال طول گشت $P(r_t\|x_{1:t})$	احتمال توزیع پیشگو $P(x_{t+1}\|x_{1:t})$	مقدار پارامترهای تابع مخاطره (توزیع لجستیک $P(r_{t-1}\|r_t)$
۵	۰٫۰۷۰۱	۰٫۷۳۱۵	-۳٫۹۹۲۰ ۱٫۲۰۱۹ ۱٫۱۰۷۴
۱۹	۰٫۰۶۴	۰٫۵۹۵۸	
۶۵	۰٫۱۱۶	۰٫۵۴۶۲	
۶۶	۰٫۰۱۲	۰٫۵۲۲۵	
۳۵۵	۰٫۱۰۱	۰٫۴۴۶۶	
۳۵۸	۰٫۰۲۳	۰٫۲۵۹۰	

۴۴۵	۰٫۰۱۱	۰٫۱۰۵۴	
۵۶۷	۰٫۰۱۰	۰٫۰۲۰۰	
۵۷۱	۰٫۰۱۰	۰٫۰۶۶۲	
۶۷۱	۰٫۰۰۸	۰٫۰۳۰۷	
۷۱۵	۰٫۰۱۴	۰٫۲۴۳۲	
۷۱۸	۰٫۰۲۶	۰٫۲۰۷۵	
۷۸۷	۰٫۰۰۸	۰٫۰۳۲۴	
۱۰۳۴	۰٫۰۰۹	۰٫۱۱۳۶	
۱۰۷۰	۰٫۰۱۰	۰٫۱۲۵۳	
۱۱۰۰	۱٫۰۰۰	۰٫۲۳۴۸	

کاهش چشم‌گیر مقدار احتمال طول گشت در این نقاط بدان معنا است که این نقاط با مقدار احتمال بسیار بالاتری، نقاط آغازی یک گشت جدید هستند و این نقاط به عنوان نقاط تغییر تعیین می‌شوند. در شکل ۵-۲۸ نمودار سری زمانی T_4 (قاب بالا) و مقادیر احتمال تمام طول گشت‌های مختلف (قاب پایین) با یک طیف رنگی مشخص شده اند. در قاب پایین، محتمل‌ترین طول گشت‌ها با خط مورب قرمز مشخص شده و ملاحظه می‌شود که در نقاط زمانی که گشت جدیدی آغاز شده، تغییری رخ داده است.

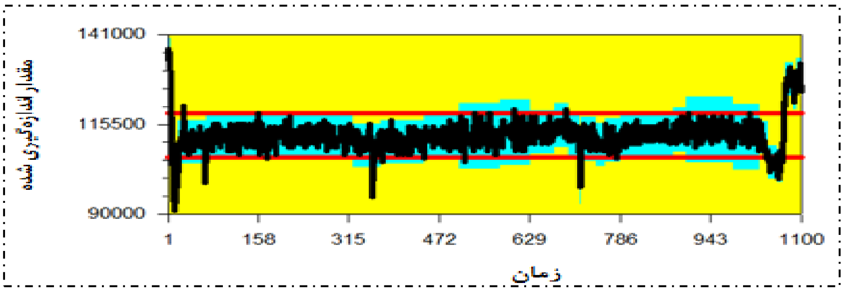


شکل ۵-۲۸: نمودار سری T_4 (قاب بالا) و نمودار طول گشت‌ها و مقادیر احتمال آن با یک طیف رنگی (قاب پایین)

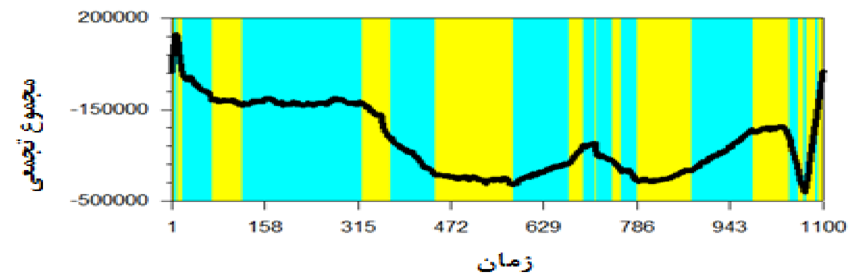
با توجه به نتایج اجرای این روش روی سری‌های شبیه‌سازی‌شده، مشاهده شد که این روش توانسته است تغییر در میانگین، واریانس و خود همبستگی سری را تشخیص دهد. بنابراین نقاط تغییر تشخیص داده شده (جدول ۵-۱۹) توسط این روش تا حدودی القاکننده نقاط تغییر واقعی هستند. در این‌جا این سوال مطرح می‌شود که عملکرد سه روش دیگر ذکر شده بر روی مجموعه داده واقعی T_4 چگونه است. آیا آن‌ها نیز نتایج مشابهی را خواهند داشت یا خیر؟ در پاسخ به این سوال ابتدا به ذکر نتایج سه روش دیگر می‌پردازیم.

۵-۶-۲ نتایج روش $CUSUM$ مبتنی بر نمونه‌های خودگردان

نمودار کنترل داده‌ها (شکل ۵-۲۹) حاکی از آن است که نقاط بسیاری خارج از حدود کنترل قرار گرفته و به عنوان نقاط تغییر تشخیص داده می‌شوند. ولی با توجه به این که این نمودار قادر به تشخیص تغییر در سری‌های شبیه سازی شد نبود، همانند قبل برای بررسی نقاط تغییر، از نمودار $CUSUM$ استفاده می‌کنیم که در شکل ۵-۳۰ برای تشخیص تغییر در سطح میانگین سری و در شکل ۵-۳۱ برای تشخیص تغییر در واریانس سری نشان داده شده است.

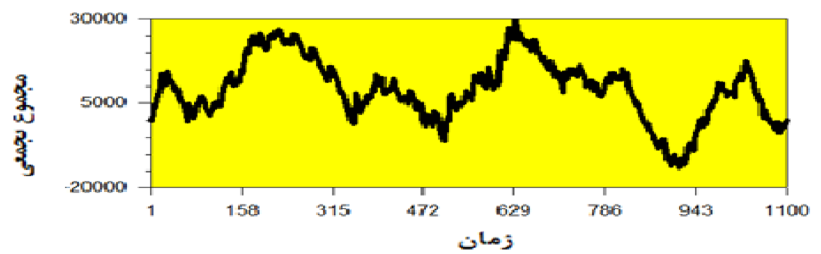


شکل ۵-۲۹: نمودار کنترل سری T_4



شکل ۵-30: نمودار CUSUM سری T_4 برای بررسی تغییر در سطح میانگین

همان طور که ملاحظه می‌شود (شکل ۵-۳۰)، شیب نمودار $CUSUM$ (بررسی تغییر در سطح میانگین) در اطراف بسیاری از نقاط سری T_4 تغییر کرده و این بدان معنا است که در این نقاط، تغییراتی در سطح میانگین سری، روی داده است. اما با توجه به شکل ۵-۳۱، نمودار $CUSUM$ (تغییر در واریانس) نشان می‌دهد که تغییری در واریانش سری رخ نداده‌است. برای نقاطی که نشان‌دهنده تغییر در سطح میانگین هستند سطح و فاصله اطمینان محاسبه شده و در جدول ۵-۲۰ آمده است.



شکل ۵-۳۱: نمودار CUSUM سری T_4 برای بررسی تغییر در واریانس

جدول ۵-۲۰: نقطه تغییر تعیین شده، سطح و فاصله اطمینان نقاط تغییر برای سری T_4

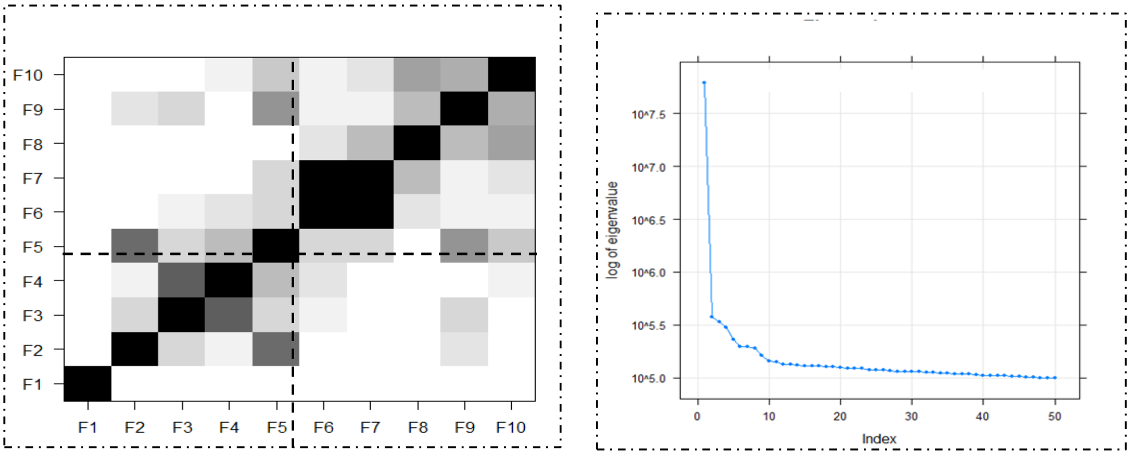
شماره داده	فاصله اطمینان	سطح اطمینان	تغییر میانگین از	تغییر میانگین به
۷	(۷ , ۷)	٪۹۵	۱۳۳۸۳۰	۱۱۱۰۰۰
۱۱	(۱۱ , ۱۳)	٪۹۲	۱۱۱۰۰۰	۹۷۲۸۹
۲۰	(۲۰ , ۲۵)	٪۱۰۰	۹۷۲۸۹	۱۱۰۶۷۰
۶۹	(۴۴ , ۸۶)	٪۹۷	۱۱۰۶۷۰	۱۱۲۱۳۰
۱۲۱	(۱۰۶ , ۲۸۸)	٪۹۳	۱۱۲۱۳۰	۱۱۲۶۲۰
۳۲۳	(۲۸۵ , ۳۲۶)	٪۱۰۰	۱۱۲۶۲۰	۱۰۹۶۰۰
۳۷۱	(۳۴۰ , ۴۱۵)	٪۹۶	۱۰۹۶۰۰	۱۱۰۷۱۰
۴۴۶	(۴۳۶ , ۴۸۵)	٪۱۰۰	۱۱۰۷۱۰	۱۱۲۳۲۰

۵۷۸	(۵۲۰ , ۶۰۵)	۹۵٪	۱۱۲۳۲۰	۱۱۳۳۹۰
۶۷۲	(۶۲۹ , ۶۷۳)	۹۴٪	۱۱۳۳۹۰	۱۱۵۴۰۰
۶۹۷	(۶۹۲ , ۷۰۰)	۹۸٪	۱۱۵۴۰۰	۱۱۳۰۵۰
۷۱۶	(۷۱۵ , ۷۱۶)	۹۴٪	۱۱۳۰۵۰	۹۹۳۳۳
۷۱۹	(۷۱۹ , ۷۲۶)	۹۳٪	۹۹۳۳۳	۱۱۱۵۴۰
۷۴۵	(۷۲۹ , ۷۴۸)	۹۲٪	۱۱۱۵۴۰	۱۰۹۸۷۰
۷۶۰	(۷۵۷ , ۷۸۲)	۹۸٪	۱۰۹۸۷۰	۱۱۱۱۸۰
۷۸۸	(۷۸۰ , ۸۲۱)	۹۸٪	۱۱۱۱۸۰	۱۱۳۰۰۰
۸۷۹	(۸۵۹ , ۹۱۹)	۹۹٪	۱۱۳۰۰۰	۱۱۴۰۸۰
۹۸۲	(۹۶۱ , ۹۹۳)	۱۰۰٪	۱۱۴۰۸۰	۱۱۲۰۲۰
۱۰۴۵	(۱۰۳۴ , ۱۰۴۵)	۹۹٪	۱۱۲۰۲۰	۱۰۴۷۹۰
۱۰۵۹	(۱۰۵۱ , ۱۰۶۷)	۹۵٪	۱۰۴۷۹۰	۱۰۳۸۰۰
۱۰۶۹	(۱۰۶۸ , ۱۰۶۹)	۹۳٪	۱۰۳۸۰۰	۱۱۶۰۰۰
۱۰۷۳	(۱۰۷۳ , ۱۰۷۴)	۹۴٪	۱۱۶۰۰۰	۱۲۸۴۴۰
۱۰۸۹	(۱۰۷۹ , ۱۰۸۹)	۹۵٪	۱۲۸۴۴۰	۱۲۴۶۷۰
۱۰۹۲	(۱۰۹۲ , ۱۰۹۶)	۹۵٪	۱۲۴۶۷۰	۱۲۹۶۷۰

با توجه به جدول ملاحظه می‌شود که این روش برخی از مشاهدات را به عنوان نقاط تغییر تشخیص می‌دهد. ولی سوال این است که نقاط تعیین شده تا چه حد به نقاط تغییر واقعی نزدیک هستند. به این سوال در بخش ۵-۷ پاسخ داده می‌شود.

۵-۶-۳ نتایج روش تحلیل طیفی منفرد

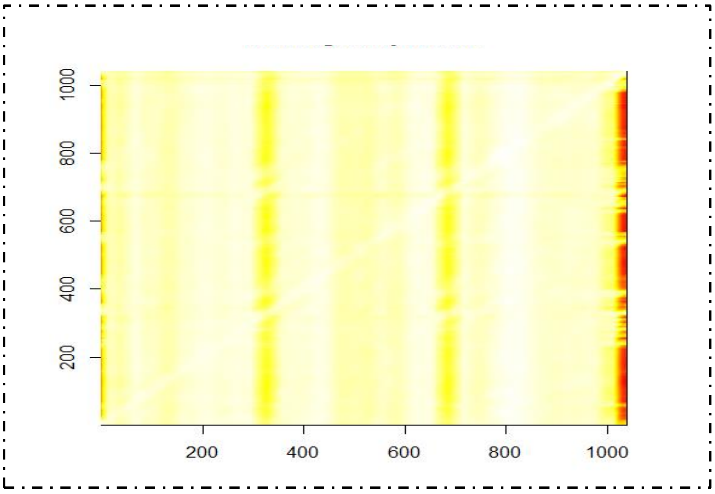
این روش برای مقادیر مختلف B و T اجرا شده همانند قبل، پارامتر l توسط دو روش l مقدار ویژه بزرگتر و ماتریس W-همبستگی تعیین می‌شود، که نمودار آن‌ها در شکل ۵-۳۲ رسم شده است. با توجه به نمودار (الف)، شش مقدار ویژه اول دارای درصد تاثیرپذیری بالایی (حدود۹۰٪)، می‌باشند. این موضوع در نمودار (ب) نیز تایید شده است که در آن با استفاده از روش W-همبستگی و تفکیک این ماتریس، مشخص است که شش مولفه اول به خوبی بیان‌کننده سری واقعی داده‌ها هستند.



(الف) (ب)

شکل ۵-۳۲: نمودار مقادیر ویژه سری T_4 برای تشخیص تعداد بردارهای مولفه‌های اصلی

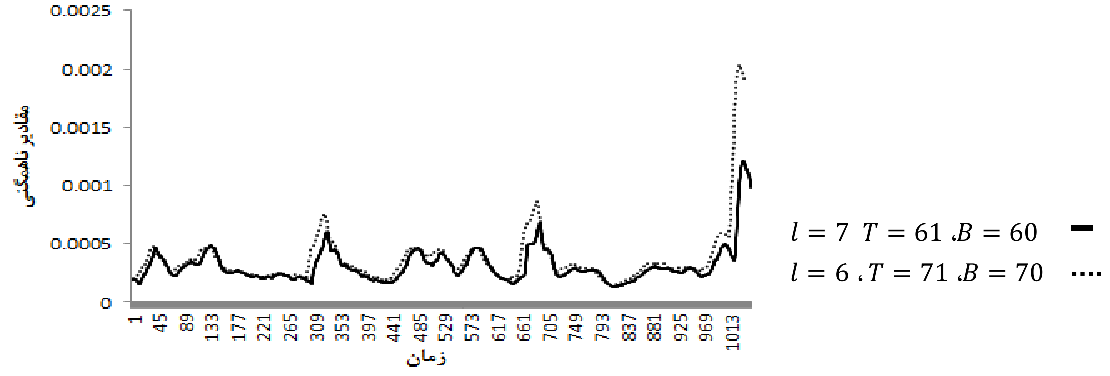
پس از انتخاب مقادیر مختلف برای پارامترهای B، T و l ، ماتریس ناهمگنی، در بهترین حالت انتخاب پارامترها در شکل ۵-۳۳ رسم شده است. مشاهده می‌شود که در اطراف برخی از نقاط سری، تغییر رنگ رخ داده و این بدان معنا است که در این نواحی روند ناهمگنی بین مشاهدت تغییر کرده است.



شکل ۵-۳۳:نمایش ماتریس ناهمگنی سری T_4

به ازای تنظیم‌های متفاوت پارامترها، ماتریس ناهمگنی مشاهدات رسم گردیده و نقاط تغییر با استفاده از توابع تشخیص، محاسبه شده‌اند. چند نمونه از این توابع برای انتخاب بهترین پارامترها در

شکل ۵-۳۴ رسم شده است. نقاط تغییر تعیین شده توسط هر یک از این ترکیب‌ها در جدول (۵-۲۱) آمده است.



شکل ۵-۳۴: نمودار توابع تشخیص ستونی ماتریس ناهمگنی سری T_3

جدول ۵-۲۱: مقدار ناهمگنی در داده‌های سری T_4

شماره نقطه تغییر	مقدار ناهمگنی به ازای B=60 ,T=61 , l=6	شماره نقطه تغییر	مقدار ناهمگنی به ازای B=70 ,T=71 , l=6
۵	۰٫۰۰۰۱۸۷	۵	۰٫۰۰۰۱۱۲
۳۶	۰٫۰۰۰۴۷۶	۱۵	۰٫۰۰۰۱۵
۶۷	۰٫۰۰۰۲۴۴	۷۵	۰٫۰۰۰۲۲۱
۲۹۵	۰٫۰۰۰۲۰۶	۳۰۵	۰٫۰۰۰۱۵۷
۴۲۷	۰٫۰۰۰۱۸۱	۴۳۲	۰٫۰۰۰۱۶۱
۵۴۸	۰٫۰۰۰۲۸۱	۵۵۳	۰٫۰۰۰۲۲۵
۶۴۰	۰٫۰۰۰۱۸۵	۶۴۸	۰٫۰۰۰۱۵۵
۷۲۰	۰٫۰۰۰۲۶	۷۲۶	۰٫۰۰۰۲۱۱
۸۱۳	۰٫۰۰۰۱۴۱	۸۱۸	۰٫۰۰۰۱۲۸
۹۶۴	۰٫۰۰۰۲۶۲	۱۰۲۰	۰٫۰۰۰۲۰۹

با توجه به نمودار ۵-۳۴ و جدول ۵-۲۱ می‌توان گفت که این روش در بهترین حالت انتخاب پارامترها $B = 70$، $T = 71$ و $l = 6$، توانسته نقاط نشان داده شده در جدول را به عنوان نقاط تغییر تشخیص دهد. این نقاط دارای کمترین مقدار ناهمگنی بوده و از این نقاط به بعد روند نمودار ناهمگنی تغییر کرده است، یعنی در این نقاط حتما تغییری در مشاهدات رخ داده است. در واقع نقاط تغییر تشخیص داده شده، متمل‌ترین نقاط تغییر می‌باشند زیرا این نقاط زمانی، کمترین فاصله را با نقاط تغییر تعیین شده توسط روش تشخیص برخط بیزی دارند.

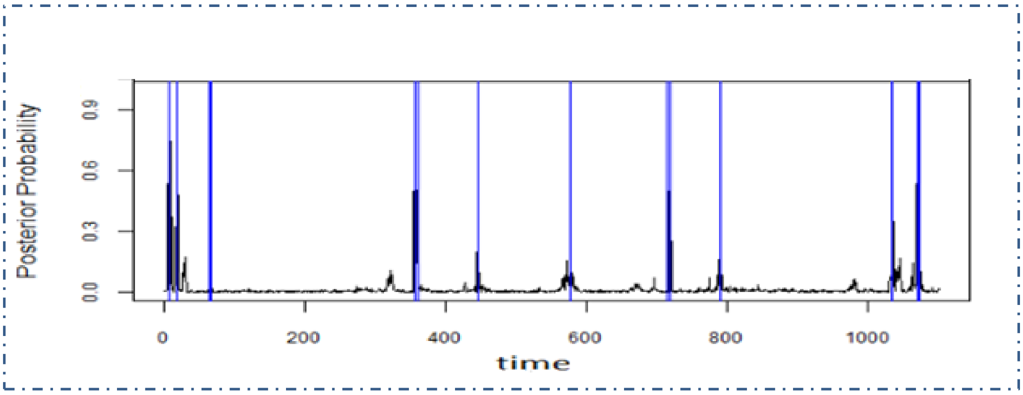
۵-۶-۴ نتایج روش افراز ضربی

همانند قبل، تمام شرایط اعمال شده توسط این روش برای سری‌های T_2, T_1 و T_3، نیز برای سری T_4 در نظر گرفته شده است. مقادیر احتمال پسین برای پارامترهای $p_\circ$ و $w_\circ$ برابر ۰٫۱، ۰٫۲و ۰٫۳ در نقاطی که به عنوان نقاط تغییر تشخیص داده شدند، در جدول ۵-۲۲ آمده است. با توجه به جدول استنباط می‌شود که این روش نقاط بسیاری را به عنوان نقاط تغییر تعیین کرده است. همچنین به ازای $w_\circ = p_\circ = 0/2$ نمودار مقادیر احتمال پسین در شکل ۵-۳۵ نشان داده شده است که در آن خطوط عمودی آبی رنگ مشخص کننده نقاط تغییر هستند.

جدول۵-۲۲: مقدار احتمال پسین برای نقاط تغییر تعیین شده در حالت $w_\circ = p_\circ = 0/3, 0/2, 0/1$

شماره داده	$w_\circ = 0/2$ $p_\circ = 0/2$	$w_\circ = 0/1$ $p_\circ = 0/1$	$w_\circ = 0/3$ $p_\circ = 0/3$
۶	۰٫۸۹۲	۰٫۸۷۸	۰٫۹۰۰
۸	۰٫۹۸۰	۰٫۹۷۴	۰٫۹۸۶
۱۹	۰٫۹۸۶	۰٫۹۴۰	۰٫۹۶۴
۶۴	۰٫۰۰۴	۰٫۰۳۶	۰٫۰۱۲
۶۶	۰٫۹۹۲	۰٫۹۳۲	۰٫۹۴۴
۳۵۵	۱٫۰۰۰	۰٫۵۳۶	۰٫۹۹۴
۳۶۰	۰٫۴۳۴	۰٫۵۲۸	۰٫۵۴۰
۴۴۵	۰٫۳۰۶	۰٫۳۱۸	۰٫۴۲۴
۵۷۷	۰/۰۶۶	۰/۰۳۶	۰/۰۴۶

۷۱۵	۱٫۰۰۰	۱٫۰۰۰	۱٫۰۰۰
۷۱۸	۰٫۵۶۶	۰٫۶۱۰	۰٫۵۰۲
۷۸۹	۰٫۱۳۲	۰٫۰۹۴	۰٫۱۰۲
۱۰۳۴	۰٫۵۶۶	۰٫۲۹۸	۰٫۶۰۶
۱۰۷۰	۰٫۹۶۸	۰٫۹۱۸	۰٫۸۸۲
۱۰۷۲	۰٫۴۶۲	۰٫۵۷۴	۰٫۴۷۶



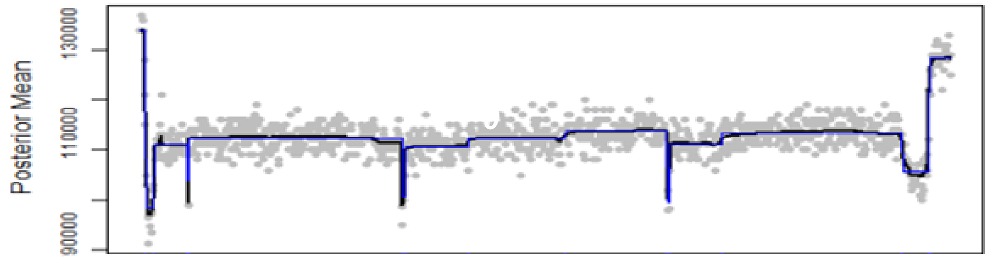
شکل ۵-۳۵: نمودار احتمال پسین نقطه تغییر بودن هر مشاهده با استفاده از مدل PPM به ازای $w_0 = p_0 = 0/2$

نتایج تشخیص نقاط تغییر با استفاده از دیدگاه دوم این روش، که تشخیص تغییر در پارامتر توزیع مشاهدات برای حالت $w_0 = p_0 = 0/2$ می‌باشد، در جدول ۵-۲۳ آمده‌است. با توجه به جدول ۵-۲۲ و ۵-۲۳ ملاحظه می‌شود، همان نقاطی که دیدگاه اول این روش آن‌ها را به عنوان نقاط تغییر معرفی کرده است از دیدگاه دوم نیز به عنوان نقطه تغییر، تشخیص داده شده‌اند. نمودار مقادیر مورد انتظار پارامتر (میانگین) به ازای هر مشاهده در شکل ۵-۳۶ نشان داده شده است.

جدول ۵-۲۳: مقدار تغییر در میانگین پسین، برای نقاط تغییر تعیین شده در حالت $w_0 = p_0 = 0/2$

شماره داده	احتمال تغییرمیانگین پسین	میانگین پسین
۶	۱۳۳۸۳۳٫۳۳	۱۱۸۰۰۰٫۰۰
۸	۱۱۸۰۰۰٫۰۰	۹۸۵۰۹٫۰۹
۱۹	۹۸۵۰۹٫۰۹	۱۱۱۰۲۲٫۲۲
۶۴	۱۱۱۰۲۲٫۲۲	۱۰۴۰۰۰٫۰۰
۶۶	۱۰۴۰۰۰٫۰۰	۱۱۲۳۱۸٫۳۴
۳۵۵	۱۱۲۳۱۸٫۳۴	۱۰۰۷۴۰٫۰۰
۳۶۰	۱۰۰۷۴۰٫۰۰	۱۱۰۵۸۸٫۲۴

۴۴۵	۱۱۰۵۸۸٫۲۴	۱۱۲۳۱۸٫۱۸
۵۷۷	۱۱۲۳۱۸٫۱۸	۱۱۳۷۱۰٫۱۴
۷۱۵	۱۱۳۷۱۰٫۱۴	۹۹۳۳۳٫۳۳
۷۱۸	۹۹۳۳۳٫۳۳	۱۱۱۰۵۶٫۳۴
۷۸۹	۱۱۱۰۵۶٫۳۴	۱۱۳۴۲۰٫۴۱
۱۰۳۴	۱۱۳۴۲۰٫۴۱	۱۰۵۶۶۶٫۶۷
۱۰۷۰	۱۰۵۶۶۶٫۶۷	۱۲۱۵۰۰٫۰۰
۱۰۷۲	۱۲۱۵۰۰٫۰۰	۱۲۸۴۲۸٫۵۷



شکل ۵-۳۶: نمودار مقادیر مورد انتظار میانگین برای هر مشاهده با استفاده از مدل PPM به ازای $w_{0} = p_{0} = 0/2$

با توجه به جدول ملاحظه می‌شود که این روش نیز همانند روش قبل (*SSA*) نقاط بسیاری را به عنوان نقاط تغییر تشخیص می‌دهد و بار دیگر این سوال مطرح است که آیا این روش، توانسته است نقاط تغییر را به درستی تشخیص دهد یا خیر. به این سوال در بخش ۵-۷ پاسخ داده خواهد شد.

۵-۷ خلاصه نتایج برای سری شبیه‌سازی‌شده T_4

پس از اعمال هریک از روش‌ها روی سری T_4، خلاصه‌ای از نتایج در جدول ۵-۲۴ آمده است.

جدول ۵-۲۴: نقاط تغییر برآورد شده سری T_4، توسط روش‌های مختلف

روش	نام روش	نقطه تغییر برآورد شده
۱	روش *CUSUM*	۷-۱۱-۲۰-۶۹-۱۲۱-۳۲۳-۳۷۱-۴۴۶-۵۷۸-۶۷۲-۶۹۷-۷۱۹-۷۴۵-۷۶۰-۷۸۸-۸۷۹-۹۸۲- ۱۰۴۵-۱۰۵۹-۱۰۶۹-۱۰۷۳-۱۰۸۹-۱۰۹۲

۲	روش *SSA*	۱۰۲۰-۸۱۸-۷۲۶-۶۴۸-۵۵۳-۴۳۲-۳۰۵-۷۵-۱۵-۵
۳	روش *PPM*	۱۰۷۲-۱۰۷۰-۱۰۳۴-۷۸۹-۷۱۸-۷۱۵-۵۷۷-۴۴۵-۳۶۰-۳۵۵-۶۶-۶۴-۱۹-۸-۶
۴	روش *BOCPD*	۱۱۰۰-۱۰۷۰-۱۰۳۴-۷۸۷-۷۱۸-۷۱۵-۶۷۱-۵۷۱-۵۶۷-۴۴۵-۳۵۸-۳۵۵-۶۶-۶۵-۱۹-۵

عملکرد مناسب روش *BOCPD* روی سری‌های شبیه‌سازی شده (بخش ۵-۵ را ببینید)، باعث شده که بتوان این روش را برای سری واقعی T_4 نیز ارجح‌تر بدانیم و بتوانیم نتایج به دست آمده از این روش را با سه روش دیگر مقایسه کنیم. همان طور که در جدول ۵-۲۴ ملاحظه می‌شود، نقاط تغییر تشخیص داده شده توسط مدل *PPM*، بسیار مشابه روش *BOCPD* است و این بدان معنا است که مدل *PPM*، نیز عملکرد مناسبی در تشخیص نقاط تغییر سری T_4 داشته است. ولی دو روش دیگر، یعنی روش *CUSUM* و *SSA* نقاط بسیار متفاوتی را تشخیص داده‌اند. بنابراین می‌توان گفت که این دو روش، عملکرد مناسبی نسبت به روش *BOCPD* و مدل *PPM* نداشته‌اند.

از آن‌جا که دو روش *BOCPD* و مدل *PPM* نقاط مشابهی را به عنوان نقاط تغییر تشخیص می‌دهند، این سوال مطرح خواهد شد که نقاط تغییر تشخیص داده شده توسط کدام روش محتمل‌ترین نقاط تغییر هستند. در پاسخ به این سوال می‌توان گفت، از آنجا که مدل *PPM*، از بین سه تغییر بیان‌شده در سری‌های شبیه‌سازی، نتوانسته است تغییر در واریانس سری را تعیین کند ولی روش *BOCPD* توانسته است هر سه تغییر در سری را تشخیص دهد لذا می‌توان چنین نتیجه گرفت که نقاط تغییر تعیین شده توسط روش *BOCPD* به نقاط تغییر واقعی نزدیکتر هستند.

۵-۸ مقایسه روش‌ها بر حسب هزینه محاسبات

در تحلیل‌های آماری علاوه بر دقت نتایج، هزینه‌ی محاسبات نیز دارای اهمیت بسیاری می‌باشد. بنابراین علاوه بر مقایسه دقت نتایج چندین روش می‌توان آن‌ها را از دیدگاه هزینه محاسبات نیز مورد مقایسه قرار داد که آیا با توجه به هزینه صرف شده، دقت نتیجه به دست آمده مقرون به صرفه است

یا خیر؟ برای این مقایسه نیازمند زمان اجرای هر یک از روش‌های ذکر شده بر روی سری‌های شبیه‌سازی‌شده T_1، T_2، T_3 و مجموعه داده واقعی T_4 هستیم که در جدول ۵-۲۵ آورده شده اند.

جدول ۵-۲۵: مدت زمان مورد نیاز برای اجرای روش‌های ذکر شده برای سری‌های T_1، T_2، T_3 و T_4 بر حسب ثانیه (s)

روش	نام روش	سری T_1	سری T_2	سری T_3	سری T_4
۱	روش $CUSUM$	16^s	3^s	26^s	42^s
۲	روش SSA	4^s	4^s	3^s	12^s
۲	روش PPM	183^s	145^s	182^s	4708^s
۴	روش $BOCPD$	98^s	98^s	122^s	861^s

با توجه به جدول ملاحظه می‌شود که زمان اجرای PPM بر روی تمام سری‌ها شبیه‌سازی‌شده نسبت به روش‌های دیگر بیشتر است. با توجه به دقت هر روش در تشخیص نقطه تغییر و مدت زمان اجرای آن، اگر مجموعه داده کوچک باشد، روش CUSUM نیز یک روش مفید و کم‌هزینه برای تشخیص نقطه تغییر در سطح میانگین و واریانس سری می‌باشد و تحلیل به نسبت دقیق‌تری نیز ارایه می‌دهد. اگر مجموعه داده بزرگ باشد، با توجه به مطالب گفته شده در بخش (۵-۷)، دقت این روش در تشخیص نقطه تغییر کاهش می‌یابد و روش BOCPD، تحلیل دقیق‌تری نسبت به روش‌های دیگر ارایه می‌دهد. بنابراین، از بین روش‌های اعمال‌شده بر روی سری T_4، روش $BOCPD$ دارای عملکرد مناسب‌تری نسبت به روش‌های دیگر بوده و مدل PPM نیز نتایج مشابهی با این روش دارد. با توجه به جدول ۵-۲۵ از آن‌جا که این روش دارای زمان اجرای کمتری نسبت به PPM است، لذا اجرای این روش مقرون به صرفه‌تر می‌باشد.

در پایان، می‌توان نتیجه گرفت که اجرای روش $BOCPD$ نسبت به سایر روش‌ها دارای عملکرد مناسب‌تری برای داده‌های واقعی است و مقرون به‌صرفه‌تر می‌باشد.

فصل ششم

نتیجه‌گیری

نتیجه‌گیری

همان طور که در فصل‌های قبل گفته شده، تشخیص نقطه تغییر یکی از یکی از چالش‌برانگیزترین مسایل در آماری است. بنابراین تعیین روش‌های که این نقاط را به طور دقیق تشخیص دهند بسیار حائز اهمیت است. در این پایان نامه چهار روش تشخیص نقطه تغییر روی سه سری شبیه‌سازی‌شده و یک مجموعه داده واقعی مورد مطالعه قرار گرفت.

بنابر نتایج تشخیص و تعیین نقطه تغییر روی سری‌های شبیه‌سازی‌شده و مجموعه داده واقعی مبتنی بر روش‌های نمودار $CUSUM$ مبتنی بر نمونه‌های خودگردان، تحلیل طیفی منفرد SSA، مدل افراز ضربی بری و هارتیگن PPM، تشخیص برخط نقطه تغییر بیزی $(BOCPD)$، می‌توان گفت که روش $BOCPD$ عملکرد مناسب‌تری را نسبت به سه روش دیگر هم از نظر تشخیص و تعیین نقطه تغییری که در کمترین فاصله از نقطه تغییر واقعی قرار دارد و هم از نظر هزینه محاسبات، دارد. با توجه به نتایج می‌توان گفت که هر یک از این روش‌ها دارای نقاط قوت و ضعفی هستند و با توجه به نظر کارشناس مربوطه در مورد دقت تشخیص نقطه تغییر و هزینه، روشی که مناسب‌تر است، انتخاب می‌شود.

پشنهادات

به منظور انجام تحقیقات آتی در زمینه تشخیص نقطه تغییر، می‌توان از راهکارهای پیشنهادی ذیل بهره گرفت:

۱- استفاده از این روش‌ها برای تشخیص نقطه تغییر روی داده‌هایی با توزیع‌های دیگر و مقایسه آن‌ها

۲- استفاده از معیارهای سنجش دقت برای هر روش به منظور سنجش کارایی روش‌ها

۳- مقایسه هریک از این روش‌ها با روش‌های دیگر در رویکردهای مختلف

۴- بررسی تشخیص نقطه تغییر در چند ویژگی هم‌زمان سری زمانی از جمله میانگین، وایانس و خودهمبستگی

فهرست منابع

کمالی، م.، (۱۳۸۹) پایان‌نامه ارشد: "شناسایی و تحلیل نقطه تغییر در فرآیندهای پواسون"، دانشگاه صنعتی شریف

Abdella, M., Marwala, T., (2005) "Treatment of missing data using neural networks and genetic algorithms". ***In: Proceedings of International Joint Conference on Neural Networks.Montreal***, 598-603.

Adams, R.P., and Mackay, D. J.C., (2007) "Bayesian online changepoint detection". ***Technical report,, University of Cambridge, Cambridge, UK,.*** *arXiv:* 0710.3742

Auret, L., Aldrich, C., (2010) " Change point detection in time series data with random forests". **Control Engineering Practice** 18,990-1002.

Barry,D., (1990) " Partition models." Comm. Statist. Theory Method, 19,2745-2756

Barry,D. and Hartigan,J.A. (1992) "product partition models for change point problems", **The Annals of Statistics,**20,260-279

Barry,D., and Hartigan, J.A.,(1993) "ABayesian analysis for change point problems", ***. Journal of the American Statistical Association, ,*** 88:309-319.

Basseville, M., Nikiforov, I.,(1993), "Detection of abrupt changes: theory and application", ***Englewood Cliffs, N.J.: Prentice Hall***

Brodsky B.E., Darkhovsky, B.S., Kaplan A.Ya., Shishkin S.L.,(1999)"A nonparametric method for the segmentation of the EEG" ***Computer Methods and Programs in Biomedicine,*** 60,93-106

Brodsky, B.E., Darkhovsky, B.S., (2000) " Nonparametric Statistical Diagnosis: Problems and Methods", ***Kluwer Academic Publishers, Dordrecht, th Netherlands***

Broomhead, D.S., and King, G.P., (1986) "Extracting Qualitative Dynamics from Experimental Data", **Physica D,** 20,217-236

Broomhead, D.S., Jones, R., and King, G.P.,(1987) "Topological Dimension and Local Coordinates From Time Series Data", **Physica A,** 20,L563-L569.

Bulunga , M.L., (2012), "Change point detection dynamical systems using Auto-Associative neural networks", ***Thesis for the Degree of Master of Science in engineering at Stellenbosch University***

Chen, Y., Hwang, K., and Wei-Shinn, K., (2007) "Distributed change-point detection of DDoS attacks:Experimental results on DETER test bed", ***DETER Community Workshop on Cyber Security Experimentation and Test. Boston: USENIX Association Berkeley***

Cheon, S., and Kim, J., (2009) "Multiple change-point detection of multivariate mean vectors with Bayesian approach", ***Computational Statistics and Data Analysis (Elsevier BV)***54,406-415

Crowley, E.M., (1997) "Product partition models for normal means", ***Journal of the American Statistical Association,*** 92,192-198.

De Oca, V.M., Jeske, D.R., Zhang, Q., Rendon, C., (2010) "Marvasti M. A cusum change-point detection algorithm for non-stationary sequences", ***The Journal of Systems and Software,*** 1288-1297.

Downey, A.B., (2008) "A novel changepoint detection algorithm", ***Applied Microbiology and Biotechnology,*** 1-11.

Frisen, M., (2003) "Statistical surveillance, optimality and methods", ***International Statistical Review ,71, no.*** 2,403-434

Gavit, P., Baddour, Y., Tholmer, R., (2009) "Use of change-point analysis for processmonitoring and control", ***BioPharm International***, 46-54

Goljandina, N., Nekrutkin, V., Zhigljavsky, A., (2000) "Analysis of Time Series Structure: SSA and related techniques", ***London: Chapman and Hall***

Habibi, R.S., Sadooghi-Alvandi, M., and Nematollahi, A.R., (2005) "Change-point detection in general class of distributions", ***Communications in statistics*** 34, no.9,1935-1938

Hajizabeth, E., Ardakani, H. D., Shahrabi, J., (2010), "Application of data mining techniques in stockMarkets", ***A survey. Journal of Economics and International Finance,*** 109-118

Hassani, H., (2007) "Singular spectrum analysis: methodology and comparison", ***Journal of Data Science*** 5,239-257.

Hammersley, J.M., and Handscomb, D. C., (1964) "Monte Carlo Methods", ***London: Methuen***

Hinkley, D., and Schechtman, E., (1987) "Conditional bootstrap methods in the mean-shift model", ***Biometrika,*** 741,85-93.

Kawahara, Y., Yairi, T., Machida, K., (2007) "Change-point detection in time-series data based on subspace identififcation", ***Seventh IEEE International Conference on Data Mining.***

Kawahara, Y., Sugiyama, M., (2009), "Change-point detection in time-series data by direct densityratio Estimation", ***In: Proceedings of*** 2009***SIAM International Conference on Data Mining*** (SDM2009),389-400

Kelsey, A., (2008), " A Novel Approach to Bayesian Online Changepoint Detection", ***Thesis for degree PhD.of Science in Computer at University of Colorado, Boulder***

Khodadadi, A., Asgharian y, M., (2008), " Change-point Problem and Regression: An Annotated Bibliography", ***Hosted by The Berkeley Electronic Press***

Lee, J., Yoo, C., Choi, S.W., Vanrolleghem, P.A., Lee, I., (2004), "Nonlinear process monitoring using kernel principal component analysis", ***Chemical Engineering Science,*** 223-234

Loschi, R.H., Cruz, F.R.B., (2001) "An analysis of the influence of some prior specifications in the identification of change points via product partition model", **Computational Statistics and Data Analysis** 39(4),477-501

Loschi, R. H., Cruz,F. R. B., Iglesias, P. L., & Arellano-Valle, R. B.,(2003) "A Gibbs sampling scheme to the product partition model: An application to change point problems", **Computers and Operations Research**, 30(3),463-482.

Loschi, R.H., Cruz, F.R.B., (2005) "Extension to the product partition model: computingthe probability of a change", **Computational Statistics and Data Analysis,** 48,255-268.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., (1953) "Equations of State Calculations by Faste Computing Machines", **Journal of Chemical Physic**, 21,1087-1091

Moskvina, V., (2000) "Distribution of Random Quadratic Forms Arising in Singular-Spectrum Analysis", **Mathematical Communications**, 5,161-171.

Moskvina, V., (2001) "Application of the singular spectrum analysis for change-point detection in time series", **PhD Thesis, Cardiff University, UK**.

Nazario, D., Ramirez, B., Tep, S., (1997) "Transient detection with an application to a chemical Process", **Computers ind. Eng,** 896-908

Page, E.S., (1954) "Continuous inspection schemes", **Biometrika**, 1,100-115

Page, E.S., (1955) "A test for a change in a parameter occuring at an unknown point" **Biometrika**, 42,523-526

Page, E.S., (1957) "On problem in which a change in a parameter occurs at an unknown point", **Biometrika**, 44,248-252

*Paquet, u.,(2007), "Empirical Bayesian Change Point Detection",***Department of Process Engineerin University of Stellenbosch South Africa**

Pignatiello, J.J., Simpson, J.R.,(2002), "A magnitude-robust control chart for monitoring and estimating step changes for normal process means", **Quality and Reliability Engineering International** 18,1-13.

Pignatiello, J.J., Samuel, T.R., (2001), "Estimation of the change point of a normal process mean in SPC applications", **Journal of Quality Technology** 33,82-95.

Saatci, Y., Turner, R., Rasmussen, C.W.,(2010) " Gaussian Process Change Point Models", **Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel,** 2010

Samuel, T.R., Pignatiello, J.J., Calvin, J.A., (1998), "Identifying the time of a step change in a normal process variance", **Quality Engineering** 10,529-538.

Smadi,M,M., (2006) "Observed Abrupt Changes in Minimum and Maximum Temperatures in Jordan in the 20thCentury", ***American Journal of Environmental Sciences*** *2(3),114-120*

Snelson, E., Ghahramani, Z., (2005) "Compact approximations to Bayesian predictive distributions", ***In ICML '05: Proceedings of the 22nd international conference on Machine learning, New York, NY, USA, ACM Press,*** *840-847*

Staudacher, M., Telser, S., Amann, A., Hinterhuber, H., Ritsch-Marte, M., (2005) "A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep", **Physica A,**582-596.

Takens, F., (1981) "Detecting strange attractors in turbulence", ***Tn Dynamical Systems and Turbulence (D. A. Rand and L. S. Young, editors). Berlin: Springer,*** *366.*

Taylor, W., (2000a) " Change-Point Analyzer 2.0 shareware program", ***Taylor Enterprises, Libertyville, Illinois.*** *Web: http://www.variation.com/cpa*

Taylor, W., (2000b) "A Pattern Test for Distinguishing Between Autoregressive and Mean-Shift Data", ***Taylor Enterprises, Libertyville, Illinois.***

Turner, R., (2010) "Bayesian Change Point Detection for Satellite Fault Prediction", ***University of Cambridge, Cambridge, UK***

Turner, R., and Rasmussen, C. E., (2010), "Supervised Bayesian online change point detection", ***In Press****.*

Vaisman, L.,(2008), "Application of Singular Spectrum-based Change-point Analysis toEMG Event Detection", **Thesis for the degree of Master of Applied Science in Biomedical Engineering, University of Toronto**

Vaisman, Lev., Zariffa, J., Popovic, M.R., (2010) "Application of singular spectrum-based changepoint analysis to EMG-onset detection", ***Journal of Electromyography and Kinesiology**, 750-760*

Vautard, R., Yiou, P., and Ghil, M.,(1992) "Singular-Spectrum Analysis:a Toolkit for Short, Noisy Chaotic Signals", ***Physica D**,58,95-126*

Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S., (2003), "A review of process fault detection and diagnosis Part III: Process history based methods", ***Computers and Chemical Engineering**,327-346*

Wheeler, D., (1993) "Understanding Variation – The Key to Managing Chaos", ***SPC Press, Knoxville, Tennessee**.*

Wilson, R.C.,Nassar, M.R., Gold, J.I., (2010), "Bayesian Online Learning of the Hazard Rate in Change-Point Problems", ***Neural Computation,** 22,2452-2476*

Yao, Y., (1984) "Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches", ***The Annal of Statistics** 12,1434-1447.*

Zhigljavsky, A., Moskvina, V., (2003) "An algorithm based on singular spectrum analysis for changepoint detection", **Communication in Statistics: Simulation and Computation**, 32,319-352

پیوست

روش مونت کارلو

روش مونت کارلو راه حل تقریبی مسایل ریاضی، آمار، فیزیک و... است که برای شبیه سازی کمیت‌های تصادفی برای تقریب(به عبارت بهتر برآورد) کمیتی مجهول به کار می‌رود. در واقع هر روشی که سعی بر حل مسئله با تولید اعداد تصادفی داشته باشد را روش مونت کارلو گویند.

محاسبه انتگرال ها با روش مونت کارلو

فرض کنید g یک تابع حقیقی مقدار باشد که علاقمند به محاسبه انتگرال g روی [a,b] باشیم.

$$\theta = \int_a^b g(x)dx < \infty$$

برای محاسبه این انتگرال می توان از روش‌های زیر استفاده کرد:

۱- روش تحلیلی: با استفاده از قضیه اساسی حساب دیفرانسیل و انتگرال، داریم

$$\theta = \int_a^b g(x)dx = G(b) - G(a)$$

۲-.روش‌های عددی: در این روش عمدتاا تابع g با توابع ساده‌تری مانند توابع خطی تکه‌ای، چند جمله‌ای و ... تقریب شده و مقدار انتگرال برای این توابع محاسبه شده، به عنوان تقریب θ ارائه می‌گردد.

۳- روش مونت کارلو: این روش در مواردی که روش‌های عددی دارای پیچیدگی‌های زیادی باشد و یا پیش فرض‌های روش‌های عددی برقرار نبوده و g چند بعدی باشد، کاربرد دارد.

در این روش هدف محاسبه انتگرال زیر است

$$\theta = \int_a^b g(x)dx$$

برای این منظور توزیع دلخواهی بر [a,b] با چگالی f(x) اختیار می‌کنیم. علاوه بر متغییر تصادفی X ، که بر فاصله [a,b] با چگالی f(x) تعریف شده است احتیاج به متغییر تصادفی

$$h(x) = \frac{g(x)}{f(x)}$$

داریم که در این صورت

$$\theta = E\big(h(x)\big) = \int_a^b \frac{g(x)}{f(x)} f(x)dx$$

حال نمونه تصادفی $X_1, X_2, \dots, X_n$ را از چگالی f استخراج می‌کنیم. آنگاه $h(X_1), h(X_2), \dots, h(X_n)$ نیز یک نمونه تصادفی خواهد بود. بنابراین داریم، $E(\frac{1}{n}\sum_{i=1}^{n} h(X_i)) = \theta$ و طبق قانون قوی اعداد بزرگ

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{p} \theta$$

بنابراین $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$ برآورد معقولی برای θ است که این برآورد معادل تقریب انتگرال مورد نظر است. با این روش بسیاری از انتگرال‌ها را که با روش تحلیلی قابل حل نیستند، می‌توان تقریب زد.

پر واضح است که روش مونت کارلو را می‌توان برای تقریب توابع چند متغیره نیز بکار برد. در واقع مزیت و برتری این روش به روش‌های عددی معمولی در حل مسائل با بعد بیشتر از یک نمایان می‌شود.

زنجیره مارکف

دنباله متغیر تصادفی $\{X^{(t)}: t \geq ۰\}$، که در مجموعه شمارای S قرار می‌گیرد، را « زنجیره‌ی مارکف همگن» با فضای وضعیت S گویند هرگاه در شرط زیر صدق کند

$$\Pr\big(X^{(t+1)} = j \big| X^{(t)} = i, X^{(t-1)} = i_{t-1}, \dots, X^{(1)} = i_1, X^{(۰)} = i_۰\big)$$
$$= \Pr\big(X^{(t+1)} = j \big| X^{(t)} = i\big) = p_{ij}$$

که در آن p_{ij}، احتمال انتقال از وضعیت i به وضعیت j بوده و $P = [p_{ij}]_{i,j\in S}$ ماتریس احتمال انتقال زنجیره $\{X^{(t)}\}$ است. بدیهی است که $\forall i, j \in S:\ p_{ij} \geq ۰\ ;\ \sum_{j\in S} p_{ij} = 1$.

روش مونت کارلو با زنجیره‌های مارکف (MCMC)

برای تقریب کمیت

$$\theta = E(X) = \int_S h(x)f(x)dx$$

به روش مونت کارلو، می‌توان کمیت θ را با استخراج مشاهدات مستقل و هم توزیع، از چگالی $f(x)$ و با استفاده از رابطه $\hat{\theta}_{MC} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ تقریب کرد. اما مشکل اساسی روش مونت کارلو استخراج یک نمونه از چگالی $f(x)$ می‌باشد، که با استفاده از دو روش معکوس تابع توزیع و روش پذیرش–رد انجام می‌پذیرد، اما به علت محدودیت‌های این دو ورش به دنبال روش قدرتمندتری برای استخراج نمونه هستیم. که این روش، روش مونت کارلو با زنجیره‌های مارکف (MCMC)، می‌باشد. این روش یک زنجیره‌ی مارکف تحویل ناپذیر، نامتناوب و زمان برگشت پذیر با توزیع ایستای حدی یکتای $f(x)$ بر S را برای استخراج یک نمونه تقریبی از $f(x)$ استفاده کرده و با استخراج نمونه، روش مونت کارلو را برای تقریب کمیت‌ها به کار می‌برد. اما مسئله اصلی ساختن زنجیره‌ی مارکف تحویل ناپذیر، نامتناوب و زمان برگشت پذیر با چگالی ایستای حدی یکتای $f(x)$، یعنی بخش زنجیره‌ی مارکف MCMC می‌باشد، که برای تولید چنین زنجیره‌ای دو الگوریتم متروپلیس- هستینگ و نمونه گیری گیبز پیشنهاد شده است.

نمونه گیری گیبز

برای تقریب $\theta = E(X) = \int_S h(x)f(x)dx$ به روش مونت کارلو وقتی که با روش‌های معمول نتوان نمونه‌ای از $f(x)$ استخراج کرد، الگوریتم متروپلیس- هستینگ روشی عمومی برای تولید یک زنجیره ارگودیک با چگالی ایستای حدی یکتای $f(x)$ می‌باشد که تحققهای این زنجیره را می‌توان به عنوان نمونه تقریبی از $f(x)$ فرض کرد. اما اجرای این الگوریتم زمانی امکان پذیر است که هسته‌ی احتمال

انتقالات (چگالی پیشنهادی) مناسبی را برای تولید یک مشاهده به شرط داشتن مشاهده قبلی، داشته باشیم. از آنجا که به طور کلی نمی‌توان قاعده‌ای برای انتخاب توزیع پیشنهادی ارائه کرد.

حال الگوریتم نمونه‌گیری گیبز را برای حل این مسئله مطرح می‌کنیم:

فرض کنید $X = (X_1, \dots, X_p) \sim f(x)$، X_{-i} به صورت زیر تعریف می‌شود:

$$\mathbf{X}_{-i} = \left(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p\right)\ ;\ \ i = 1, \dots, p$$

بنابراین $X_i | \mathbf{X}_{-i} = \mathbf{x}_{-i}$ دارای چگالی $f(X_i | x_{-i})$ است. چنانچه از چگالی $f(x_i | \mathbf{x}_{-i})$، $i = 1, \dots, p$ بتوان نمونه تولید کرد، آنگاه با انتخاب وضعیت اولیه‌ی $x^{(0)}$، الگوریتم زیر را می‌توان با داشتن $\mathbf{X}^{(t)}$، $\mathbf{X}^{(t+1)}$ تولید کرد.

۱- ترکیبی از مولفه‌های $\mathbf{x}^{(t)}$ را انتخاب می‌کنیم.

۲- برای تولید X_i^*، $(i = 1, \dots, p)$ از چگالی $f(x_i | \mathbf{x}_{-i}^{(t)})$ استفاده می‌کنیم.

۳- $\mathbf{X}^{(t+1)} = \mathbf{X}^*$ قرار می‌دهیم.

توجه:

- این الگوریتم را گیبز برای تقریب یک انتگرال در فیزیک پیشنهاد کرده است و به همین خاطر به نمونه گیری گیبز یا الگوریتم گیبز مشهور است.

- از آن جا که $\mathbf{X}^{(t+1)}$ تنها به $\mathbf{X}^{(t)}$ وابسته است زنجیره $\{\mathbf{X}^{(t)}\}_{t \geq 0}$، یک زنجیره مارکف می‌باشد.

- در گام دوم به جای اینکه $\mathbf{X}^*$ از $f(\mathbf{x} | \mathbf{x}^{(t)})$ تولید شود، با تولید p مولفه X_i^* بردار $\mathbf{X}^*$ به دست می‌آید. برای اینکه ترتیب تولید مولفه‌ها بر تولید $\mathbf{X}^*$ تاثیر نداشته باشد، در گام اول از بین !p جایگشت ممکن از مولفه‌های $\mathbf{X}^{(t)}$ یکی را انتخاب کرده و گام دوم را براساس این ترتیب انتخاب شده اجرا می‌کنیم. به همین دلیل در برخی از موارد ابتدا یک توزیع احتمال را بر !p جایگشت، مولفه‌های $\mathbf{X}^{(t)}$ در نظر گرفته، و آن را به عنوان ترکیبی از مولفه‌های $\mathbf{X}^{(t)}$ در نظر می‌گیرند. حال برای جایگشت‌های مولفه‌های $\mathbf{X}^{(t)}$ گام دوم را اجرا می‌کنند.

- اجرای گام دوم براساس ترتیب انتخاب شده در گام اول را یک سیکل می‌گوییم.

- برای یک سیکل خاص، در i امین مرحله‌ی سیکل الگوریتم با داشتن $\mathbf{X}^{(t)}$، مقدار پیشنهادی $\mathbf{X}^* = (x_1^{(t)}, \ldots, X_i^*, \ldots, x_p^{(t)})$ را از چگالی پیشنهادی $g_i(\mathbf{x}^*|\mathbf{x}^{(t)})$ تولید می‌کند. که در آن:

$$g_i\left(\boldsymbol{x}^*\middle|\boldsymbol{x}^{(t)}\right) = \begin{cases} f\left(\boldsymbol{x}_i^*\middle|\boldsymbol{x}_{-i}^{(t)}\right) & ; \quad \boldsymbol{x}_{-i}^* = \boldsymbol{x}_{-i}^{(t)} \\ \text{۰} & ; \quad \text{سایر نقاط} \end{cases}$$

در این حالت اگر $R\left(\mathbf{x}^{(t)}, \mathbf{x}^*\right) = \frac{f(x^*)g(x^{(t)}|x^*)}{f\left(x^{(t)}\right)g(x^*|x^{(t)})} \equiv 1$، مقدار پیشنهادی $\mathbf{X}^*$ پذیرفته می‌شود. بنابراین هر سیکل از p، یک الگوریتم متروپلیس- هستینگ تشکیل می‌دهد و بنابراین می‌توان الگوریتم فوق را حالت خاصی از الگوریتم متروپلیس- هستینگ در نظر گرفت.

- می‌دانیم که $f(x_i|\mathbf{x}_{-i}) = c_i f(\mathbf{x})$ است که در آن $c_i^{-1} = \int f(\mathbf{x}_{-i})d\mathbf{x}_{-i}$. چنانچه در این الگوریتم $f(x_i|\mathbf{x}_{-i})$ برای برخی از i ها نرمال نشده باشد[1] یعنی : c_i مجهول باشد، چون هر سیکل در الگوریتم متشکل از p الگوریتم متروپلیس- هستینگ است، مشکلی برای اجرای الگوریتم نخواهیم داشت.

- برای بهبود عملکرد این الگوریتم، در یک سیکل وقتی که مولفه‌ی اول پیشنهادی X_1^* با داشتن $\mathbf{X}^{(t)}$ تولید شد، مولفه‌ی دوم پیشنهادی X_2^* را از $g_i(\mathbf{x}^*|x_1^{(t+1)}, x_2^{(t)}, \ldots, x_p^{(t)})$ به جای $g_i(x^*|x_1^{(t)}, x_2^{(t)}, \ldots, x_p^{(t)})$ تولید کرده و به همین ترتیب برای سایر مولفه‌ها تا کامل شدن سیکل از مقدار جدید $x_i^{(t+1)}$ به جای $x_i^{(t)}$ استفاده می‌کنیم.

- زنجیره‌ی مارکفی که با این الگوریتم تولید می‌شود تحویل ناپذیر، نامتناوب و زمان برگشت پذیر و در نتیجه ارگودیک بوده، بنابراین این الگوریتم نیز یکی از ابزارهای روش MCMC است.

---

[1] تابع $f: \mathfrak{R} \to \mathfrak{R}$ تابع چگالی است اگر ۱) $\forall x : f(x) \geq 0$ ۲) $\int_{\mathfrak{R}} f(x)dx = 1$. اگر f در شرط اول صدق کند و بدانیم $\int_{\mathfrak{R}} f(x)dx < \infty$، آنگاه f را یک چگالی نرمال نشده می‌گوییم

واژه نامه فارسی به انگلیسی

آمار چند متغیره	*Multivariate statistics*
آماره اسمیرنف- کلموگروف	*Kolmogorov smirnov statistics*
الکتیو کاریوگرام	*Electrocardiogram*
اندیس	*Index*
انحرافات با دلیل	*Assignable causes*
باز سازی	*Restructuring*
بازنگری فرآیند	*Process monitoring*
برآورد	*Estimate*
برخط	*Online*
بردار ویژه	*Eigenvector*
بیش برازشی	*Over dispersion*
پارامترهای اضافه	*Hyper parametres*
پسرو	*Backward*
پسین	*Posteriori*
پیشین	*Priori*
پیشرو	*Forward*
تابع خودهمبستگی	*Autocorrelation function*
تابع مرحله ای	*Step function*

تغییرپله‌ای یگانه	*Single point step change*
تجزیه مقدار منفرد	*Singular value decomposition*
تحلیل طیفی منفرد	*Singular spectrum analysis*
تحلیل نوسان خروجی پیش رونده	*Progressive detrended fluctuation analysis*
تشخیص احتمالی	*Probabilistic diagnostics*
تشخیص برخط نقطه تغییر بیزی	*Bayesian Online Change point Detection*
تصادفی	*Random*
تعبیه سازی	*Embedding*
تغییر با روند خطی	*Linear trend change*
تغییر پله ای چندگانه	*Multiple point step change*
تقریب	*Approximate*
توزیع	*Distribution*
توزیع پیشگوی شرطی	*Predictive distribution conditional*
خودگردان ساز	*Bootstrapping*
دنباله تشخیصی	*Diagnostic sequence*
روش دنباله ای	*Sequential method*
زنجیره مارکوف مونت کارلوی	*Marakov chain monte carlo*
سه گانه	*Eigentriple*
سری زمانی	*Time series*
شاخص ناهمگنی	*Heterogeneity index*
شبکه عصبی مصنوعی	*Artificial neural networks*

شوهارت	*Shewhart*
طول گشت	*Run legngth*
فرایند	*Process*
فرآیند کنترل آماری	*Statistical process control*
کم‌برازشی	*Under dispersion*
ماتریس آزمون	*Teat matrix*
ماتریس اصلی	*Base matrix*
ماتریس کواریانس تاخیر	*Lag-covariance matrix*
ماتریس مسیر	*Trajectory matrix*
ماتریس هنکل	*Hankel matrix*
ماکسیمم درستنمایی	*Maximum likelihood*
متعامد	*Orthogonal*
مجموع تجمعی	*Cumulative sum*
مجموع توان دوم فاصله اقلیدسی	*Sum of squared euclidean distances*
مدل افراز ضربی	*Product Partition Model*
مدل فضای حالت	*State space model*
مساله تشخیص	*Disorder problem*
مقدار ویژه	*Eigenvalue*
مولفه اصلی	*Principal component*
میانگین گیری قطری	*Diagonal averaging*
میانگین متحرک اتورگرسیو	*Autoregressive moving average*

میانگین متحرک موزون نمایی	*Exponentially weighted moving average*
نسبت ماکسیمم درستنمایی	*Maximum log likelihood ratio*
نرخ مخاطره	*Hazard rate*
نقطه تغییر	*Change point*
نمونه‌گیری گیبز	*Gibs sampling*
نمونه های خودگردان	*Bootstrap samples*
نوسان زدایی شده	*Detrended fluctuation analysis*
واخط	*Off line*
همگنی تصادفی	*Stochastic homogeneity*
هندسه چند متغیره	*Multivariate geometry*
هنکل سازی	*Hankelization*
هنکل سازی معکوس	*Inverse hankelization*

واژه نامه انگلیسی به فارسی

تقریب	*Approximate*
شبکه عصبی مصنوعی	*Artificial neural networks*
انحرافات با دلیل	*Assignable causes*
میانگین متحرک اتورگرسیو	*Autoregressive moving average*
تابع خودهمبستگی	*Autocorrelation function*
پسرو	*Backward*
ماتریس اصلی	*Base matrix*
تشخیص برخط نقطه تغییر بیزی	*Bayesian Online Change point Detection*
خودگردان ساز	*Bootstrapping*
نمونه های خودگردان	*Bootstrap samples*
نقطه تغییر	*Change point*
مجموع تجمعی	*Cumulative sum*
نوسان زدایی شده	*Detrended fluctuation analysis*
میانگین کیری قطری	*Diagonal averaging*
دنباله تشخیصی	*Diagnostic sequence*
مساله تشخیص	*Disorder problem*
توزیع	*Distribution*
سه گانه	*Eigentriple*
مقدار ویژه	*Eigenvalue*

Eigenvector	بردار ویژه
Electrocardiogram	الکتیو کاریوگرام
Embedding	تعبیه سازی
Estimate	برآورد
Exponentially weighted moving average	میانگین متحرک موزون نمایی
Forward	پیشرو
Gibs sampling	نمونه‌گیری گیبز
Hankelization	هنکل سازی
Hankel matrix	ماتریس هنکل
Hazard rate	نرخ مخاطره
Heterogeneity index	شاخص ناهمگنی
Hyper parametres	پارامترهای اضافه
Index	اندیس
Inverse hankelization	هنکل سازی معکوس
Kolmogorov smirnov statistics	آماره اسمیرنف- کلموگروف
Lag-covariance matrix	ماتریس کواریانس تاخیر
Linear trend change	تغییر با روند خطی
Marakov chain monte carlo	زنجیره مارکوف مونت کارلوی
Maximum likelihood	ماکسیمم درستنمایی
Maximum log likelihood ratio	نسبت ماکسیمم درستنمایی
Multiple point step change	تغییر پله ای چندگانه

Multivariate geometry	هندسه چند متغیره
Multivariate statistics	آمار چند متغیره
Off line	واخط
Online	برخط
Orthogonal	متعامد
Over dispersion	بیش برازشی
Predictive distribution conditional	توزیع پیشگوی شرطی
Priori	پیشین
Principal component	مولفه اصلی
Probabilistic diagnostics	تشخیص احتمالی
Process	فرایند
Process monitoring	بازنگری فرآیند
Product Partition Model	مدل افراز ضربی
Progressive detrended fluctuation analysis	تحلیل نوسان خروجی پیش رونده
Posteriori	پسین
Random	تصادفی
Restructuring	باز سازی
Run legngth	طول گشت
Sequential method	روش دنباله ای
Shewhart	شوهارت
Single point step change	تغییرپله‌ای یگانه

Singular spectrum analysis	تحلیل طیفی منفرد
Singular value decomposition	تجزیه مقدار منفرد
Statistical process control	فرآیند کنترل آماری
State space model	مدل فضای حالت
Step function	تابع مرحله ای
Stochastic homogeneity	همگنی تصادفی
Sum of squared euclidean distances	مجموع توان دوم فاصله اقلیدسی
Teat matrix	ماتریس آزمون
Time series	سری زمانی
Trajectory matrix	ماتریس مسیر
Under dispersion	کم‌برازشی

Abstract:

In a time series can be basically a sequence of observations in order time such as $x_1, \ldots, x_n$, *there many be poimts where in these points, the statistical properties of random variables have been changed. For example, there may be* n_0 *first observation have distribution* F_0 *and other observation have distribuation* F_1. *The point* n_0 *is called change point and finding this point which the time series features is modified, include a wide range of issues in the real world. Finding the change point is one of the most challenging Statistical issues, because the number and location of these points are unknown. For this purpose several methods have been proposed with different capabilities.*

In this thesis, the four methods, cumulative sum (CUSUM) based on bootstrap samples, singular spectrum analysis (SSA), an Bayesian online change point detection (BOCPD) and Bayesian partition production model (PPM) have been studied. In order to verify the accuracy and ability of each of these methods in detecting change points, of three time series simulated and an actual time series are used so that the time series of the simulated are respectively explainer change in surface of the average, change in surface of the variance and change in the autocorrelation.The results of simulated data showed that between the *presented methods, Bayesian online change-point detection method, in comparetion whit other methods, has better performance and is able to properly estimated any changes to mentioned. Choosing this method as the best method, real time series data are studied and the performances are compared with other methods. Since the cost time spent for analyze the data and identify of the change is important, the* execution time *of methods have been compared. The results obtained demonstrate that Bayesian online change-point detection method in large data sets is cost effectiveness.*

Keywords: time series, the change point, cumulative sum (CUSUM), Singular spectrum analysis (SSA), the detection function, matrix heterogeneity, Bayesian online change point detection (BOCPD), product partition model (PPM)

Shahrood University of Technology
Faculty of Mathmatic

M.S Thesis

Change Point Detection in the Time Series Data

By:

Omekolcoom Hematirad

Supervisor:

Dr. D. Shahsavani

Dr. A. Nezakati Reza Zade

Date:

***February** 2013*

179