شناسایی نقطهٔ تغییر در داده‌های پانلی به روش هم‌زمان میانگین متحرک موزون‌نمایی و جمع تجمعی

کریم آتشگر[1*]، ناصر رفیعی[2]

۱. دانشیار مهندسی صنایع دانشگاه صنعتی مالک‌اشتر، تهران

۲. دانشجوی کارشناسی ارشد مهندسی صنایع، دانشگاه صنعتی مالک‌اشتر، تهران

(تاریخ دریافت: ۹۷/۱۰/۰۹، تاریخ دریافت روایت اصلاح‌شده: ۹۸/۰۱/۲۴، تاریخ تصویب: ۹۸/۰۱/۳۱)

چکیده

شناسایی نقطهٔ تغییر در داده‌های پانلی به کارشناسان کمک می‌کند تا دریابند مقاطع داده‌ها در چه زمانی تغییر کرده است. شناسایی این زمان به تحلیل واقعی‌تری از تغییرات در فرایند می‌انجامد. روش‌های مختلفی برای شناسایی نقطهٔ تغییر در داده‌های پانلی معرفی شده است، اما افزایش حساسیت روش‌ها در شناسایی نقطهٔ تغییر داده‌های پانلی، اهمیت بسیاری دارد. در این مقاله، به‌کمک رویکردی ترکیبی روشی با حساسیت بالا برای شناسایی نقطهٔ تغییر در داده‌های پانلی (با ابعاد بزرگ) معرفی شده است. در این مقاله، روشی جدید با رویکرد ترکیبی با نام جمع تجمعی دوبل-میانگین متحرک موزون (double CUSUM-EWMA) که حساسیت بسیاری دارد، برای شناسایی نقطهٔ تغییر داده‌های پانلی پیشنهاد می‌شود. همچنین با استفاده از داده‌های شبیه‌سازی‌شده، فرایندی در ابعاد بزرگ بررسی می‌شود که داده‌های پانلی آن از تغییر پله‌ای تأثیر پذیرفته است. نتایج مقایسهٔ عملکرد روش پیشنهادی با یکی از بهترین روش‌ها در ادبیات نشان‌دهندهٔ بهبود عملکرد روش پیشنهادی این مقاله از روش بررسی‌شده است.

واژه‌های کلیدی: جمع تجمعی دوبل، داده‌های پانلی، میانگین متحرک موزون‌نمایی، نقطهٔ تغییر.

مقدمه

مشاهدات آماری به سه دستهٔ داده‌های سری زمانی، داده‌های مقطعی و داده‌های پانلی تقسیم می‌شوند. داده‌های سری زمانی داده‌هایی هستند که طی یک دورهٔ زمانی از مقادیر مربوط به یک یا چند متغیر جمع‌آوری شده‌اند و به‌صورت $y_t = y_1, y_2, \dots, y_T$ یا $\{y_t; t \in T\}$ نشان داده می‌شوند؛ به‌طوری‌که در آن $t = 1, 2, \dots, T$ بیانگر زمانی است که متغیر تصادفی y در آن اندازه‌گیری شده است. T نیز نشان‌دهندهٔ طول سری زمانی یا تعداد کل مشاهدات است. داده‌های مقطعی داده‌هایی هستند که براساس مقادیر یک یا چند متغیر در یک‌زمان مشخص جمع‌آوری می‌شوند. به عبارت دیگر داده‌های مقطعی به مقادیر مربوط به یک یا چند متغیر در زمان معین برای واحدهای متعدد گفته می‌شود. داده‌های پانلی داده‌هایی هستند که هم‌زمان شامل ویژگی‌های داده‌های مقطعی و سری زمانی هستند. همچنین به مجموعه‌ای از داده‌ها گفته می‌شود که براساس آن‌ها، داده‌های مشاهده‌شده از یک یا چند متغیر برای بسیاری از واحدهای مقطعی (n) که به‌صورت تصادفی انتخاب می‌شوند، در طول یک دورهٔ زمانی مشخص (T) به‌دست آمده باشند. در این صورت ماتریسی را که $n \times T$ دادهٔ آماری دارد، داده‌های پانلی می‌نامند.

تغییرات تصادفی در داده‌های پانلی، بیانگر شرایط طبیعی در سیستم یا فرایند مورد مطالعه است، اما هنگامی که یک سیستم یا فرایند از عوامل غیرطبیعی (غیر تصادفی) تأثیر می‌گیرد، بخشی از داده‌ها، تحت شرایط غیرطبیعی ایجاد می شوند. زمانی که سیستم یا فرایندی از عوامل غیرطبیعی تأثیر می‌گیرد و از آن زمان به بعد داده هایی با تغییرات غیرتصادفی تولید می شود نقطهٔ تغییر نام دارد؛ از این‌رو داده‌هایی که پس از نقطهٔ تغییر ایجاد شده اند، با تغییر در حداقل یکی از پارامترهای میانگین، واریانس یا ساختار وابستگی مواجه خواهند شد.

مدل داده های پانلی در این پژوهش به‌صورت رابطهٔ ۱ تعریف می‌شود:

$$x_{j,t} = \mu_j + \delta_j I\{t > t_0\} + \varepsilon_{j,t}, \quad t = 1, \dots, T; \; j = 1, \dots, n \qquad (1)$$

که در آن، ε_{jt} خطای فرایند در مقطع j ام و زمان t ام است و فرض می‌شود در بعد زمانی ایستا است و $E(\varepsilon_{j,t}) = 0$ است.

* نویسندهٔ مسئول تلفن: ۲۲۹۷۰۳۰۹ Email: atashgar@iust.ac.ir

همچنین $\delta_j \neq 0$ مقدار تغییر در میانگین μ_j از بعد از نقطهٔ تغییر نامعلوم $t_0 \in [1, T)$ است. در مدل فوق μ_j نشان‌دهندهٔ میانگین مقطع j ام در شرایط تحت کنترل است. بدیهی است باید انتظار داشت که در شرایط تحت کنترل، هر مشاهده در هر مقطع به‌صورت تصادفی حول میانگین قرار بگیرد. در مدل داده‌های پانلی ۱، میانگین مقطع j ام از زمان بعد از t_0، از μ_j به $\mu_j + \delta_j$ تغییر می‌کند.

هرچند روش‌های مختلفی برای شناسایی نقطهٔ تغییر در داده‌های پانلی معرفی شده است، اما حساسیت بالا برای روش‌ها در شناسایی نقطهٔ تغییر داده‌های پانلی، اهمیت بسیاری دارد. با توجه به حساسیت روش‌های جمع تجمعی و میانگین متحرک موزون، که برخلاف رویکرد شوهارت اثر داده‌های گذشته را در پایش داده‌های فعلی مدنظر قرار می‌دهد، در این مقاله با رویکردی ترکیبی، نقطهٔ تغییر در داده‌های پانلی (با ابعاد بزرگ) شناسایی شده است.

در بخش بعدی این مقاله، ادبیات مربوط به مسئلهٔ نقطهٔ تغییر در داده های پانلی بررسی می شود. بخش سوم این مقاله به معرفی روش های مختلف شناسایی نقطهٔ تغییر در داده های پانلی با رویکرد جمع تجمعی اختصاص یافته است. در بخش چهارم، آمارهٔ پیشنهادی این مقاله برای شناسایی نقطهٔ تغییر در داده‌های پانلی تشریح شده است. در بخش پنجم، ضمن تجزیه و تحلیل کمی قابلیت روش پیشنهادی، به مقایسهٔ عملکرد روش پیشنهاد این مقاله با دیگر روش های موجود در ادبیات پرداخته شده است. در بخش ششم، روش پیشنهادی در یک نمونهٔ واقعی استفاده شده و درنهایت بخش آخر این مقاله به نتیجه‌گیری اختصاص یافته است.

پیشینهٔ پژوهش

جوزف و ولفسون [۱، ۲] پیشگام پژوهش درمورد مسئلهٔ نقطهٔ تغییر در داده‌های پانلی هستند. آن‌ها با فرض اینکه در n سری داده، نقاط تغییر به طور مستقل و با توزیع یکسان (iid[1]) هستند، مدلی را برای شناسایی نقطهٔ تغییر تصادفی پیشنهاد کردند؛ به طوری‌که هر سری نقطهٔ تغییر خاص خود را دارد. آن‌ها نشان دادند توزیع مشترک نقاط تغییر می‌تواند به طور سازگار برآورد شود. بای [۳] بر وضعیت نقطهٔ تغییر مشترک تمرکز می کند؛ به‌طوری که در پژوهش او تمام سری‌ها، نقطهٔ تغییر مشابهی دارند. وی از روش برآورد کمترین مربعات خطا (LSE[2]) برای برآورد نقطهٔ تغییر در میانگین و از روش حداکثر درست‌نمایی گوسی (QML[3]) برای برآورد نقطهٔ تغییر در میانگین، واریانس یا هر دو به‌صورت هم‌زمان استفاده کرده است. هورواث و هوسکووا [۴] آمارهٔ مربوط به برآوردگر نقطهٔ تغییر در میانگین های داده های پانلی را که بای [۳] با استفاده از روش حداکثر درست‌نمایی گوسی ایجاد کرده است بررسی کردند و توزیع تقریبی آماره را ارائه دادند. تمرکز آن ها در این پژوهش بر آزمونی برای تغییرات احتمالی در موقعیت پارامتر میانگین در داده‌های پانلی است و در آن آزمونی طراحی شده است که به تغییرات نسبتاً بزرگ میانگین در تعداد کمی از واحدهای مقطعی و هم زمان به تغییرات نسبتاً کوچک میانگین در تعداد زیادی از واحدهای مقطعی حساس است. این آزمون از استدلال درست‌نمایی ناشی شده و مبتنی بر تطبیق روش جمع تجمعی (CUSUM[4]) بر داده های پانلی است. لی و همکاران [۵] آمارهٔ هورواث و هوسکووا [۴] را توسعه دادند و آمارهٔ مبتنی بر روش جمع تجمعی برای شناسایی نقطهٔ تغییر مشترک در واریانس را پیشنهاد کردند. دی واچتر و تزاوالیس روش آزمون تشخیص شکست ساختاری برای مدل داده های پانلی خطی پویا با رگرسورهای برون‌زا را توسعه دادند [۶] (منظور از شکست ساختاری تغییر در ساختار یک مدل است). روش ارائه‌شده می تواند تغییر در پارامترهای شیب و تأثیرات ثابت را در خود جای دهد. این کار براساس چارچوب روش تعمیم‌یافتهٔ گشتاورهای (GMM[5]) استفاده‌شدهٔ آرلانو و باند [۷] انجام شده است. در این روش، تفاوت در مقدار توابع هدف GMM با فرض صفر (هیچ تغییری رخ نداده است) و هریک از جایگزین‌ها (تغییر در یکی از زمان‌ها رخ‌داده است) اساسی برای آمارهٔ آزمون خواهد بود. زو و همکاران رویکرد تشخیص نقطهٔ تغییر مبنی بر تابع کاپیولا[6] را در داده‌های پانلی نامتعادل ارائه کردند [۸]. در این رویکرد، کاپیولا برای توصیف ساختار وابستگی داده‌ها استفاده می‌شود و پارامتر کاپیولا نشان‌دهندهٔ درجه وابستگی است. رویکرد این روش در تشخیص نقطهٔ تغییر را می‌توان به‌عنوان فرایند سه‌مرحله‌ای خلاصه کرد. در ابتدا، کاپیولا یا عضو رابط مناسب که می‌تواند ساختار وابستگی

مجموع داده‌ها را توصیف کند، انتخاب می‌شود. سپس این کاپیولا متناسب با زیرمجموعهٔ داده‌ها به‌صورت پویا با اضافه‌کردن داده‌های جدید به داده‌های قبلی در هر زمان به‌روزرسانی می‌شود. درنهایت تغییرات می‌تواند با تجزیه و تحلیل روند پارامترهای متناسب‌شده در هر نقطهٔ زمانی شناسایی شود. چن و هو [۹] برآوردگر نقطهٔ تغییر مشترک را برای میانگین، با پیچیدگی کمتر و دقت بیشتر از روش برآورد کمترین مربعات خطای بای [۳] پیشنهاد کردند که در آن از روش جمع تجمعی برای برآورد نقطهٔ تغییر مشترک استفاده می شود. پستووا و پستا [۱۰] آزمونی را با استفاده از بوت استرپ برای تغییر مشترک در میانگین واحدهای مقطعی پانل در دورهٔ زمانی ثابت، بررسی کردند که می‌تواند نسبتاً کوچک یا متوسط باشد. همچنین پستووا و پستا [۱۱] برآوردگر نقطهٔ تغییر مشترک را برای میانگین واحدهای مقطعی پانل پیشنهاد کردند. به‌طوری که روش آن‌ها برگرفته از روش کمترین مربعات خطا بود. روش پستووا و پستا [۱۱] این ویژگی را دارد که وضعیتی که در آن نقطهٔ تغییر آخرین مشاهده را نشان می‌دهد به این معنی است که هیچ تغییری در میانگین‌های مقاطع رخ نداده است. انوموتو و ناگاتا [۱۲] روش ماهالانوبیس-تاگوچی (MT[۷]) را به کمک استنتاج بیز، برای تشخیص نقطهٔ تغییر در داده‌های پانلی گسترش دادند که BMT[۸] نام دارد. در این روش، برای تشخیص نقطهٔ تغییر، مقاطع در زمانی معین به‌صورت یک بردار از مشاهدات درنظر گرفته می‌شود. فاصلهٔ این بردار با بردار میانگین چند مشاهدهٔ زمانی اخیر، با استفاده از فاصلهٔ ماهالانوبیس محاسبه می‌شود. افزایش ناگهانی فاصله در زمانی خاص نشان می‌دهد ستون مشاهده‌شده از مقاطع در آن زمان در مقایسه با ستون‌های مشاهده‌شده در چند مشاهدهٔ زمانی اخیر، تغییر ناگهانی داشته است، اگر این افزایش ناگهانی از مقدار آستانهٔ فرض‌شده بیشتر باشد، سیگنال خارج از کنترل نقطهٔ تغییر در داده‌های پانلی را نشان می‌دهد.

برخی آزمون‌های تشخیص نقطهٔ تغییر در داده‌های پانلی، براساس روش جمع تجمعی و برای داده‌هایی با ابعاد بزرگ ارائه شده‌اند که توانایی تشخیص چند نقطهٔ تغییر را دارند. در این مورد می‌توان به ژانگ و همکاران [۱۳]، انیکیوا و هارچوی [۱۴] و جیراک [۱۵] اشاره کرد. در روش‌های فوق، مجموع جمع‌های تجمعی در نقطهٔ تغییر برای تمام مقاطع درنظر گرفته می‌شود. براساس مطالعات تجربی چو و فریزلیچ [۱۶] چنین رویکردهایی ممکن است منجر به عملکرد پایین‌تر در تشخیص و تعیین نقاط تغییر در تنظیمات با ابعاد بزرگ شوند؛ از این‌رو این دو پژوهشگر روشی را ارائه کردند که در آن جمع‌های تجمعی در نقاط تغییر به مقاطعی محدود می‌شود که در آن‌ها تغییر رخ داده است. چو [۱۷] با ارائهٔ آمارهٔ جمع تجمعی دوبل (DC[۹]) رویکرد چو و فریزلیچ را به‌طور قابل‌ملاحظه‌ای ارتقا داد.

آمارهٔ جمع تجمعی دوبل با استفاده از ساختار نقطهٔ تغییر مقطعی با بررسی جمع تجمعی از جمع‌های تجمعی مرتب در هر نقطهٔ تغییر حاصل می‌شود؛ یعنی ساختار برآورد نقطهٔ تغییر در هر مقطع، در میان مقاطع نیز برای هر نقطهٔ تغییر احتمالی به‌کار گرفته می‌شود و در این صورت تنها مقاطعی که دچار تغییر شده اند در تشخیص نقطهٔ تغییر مشترک شرکت می کنند؛ بنابراین نه‌تنها این اطلاعات در تفسیر نقاط تغییر شناسایی شده مفید است، بلکه می‌تواند نقش مهمی در کارآمدی جمع‌های تجمعی سری های با ابعاد بزرگ داشته باشد. چو [۱۷] روش خود را هم در توان آزمون و هم در دقت برآورد با روش های ارائه‌شدهٔ ژانگ و همکاران [۱۳]، انیکیوا و هارچوی [۱۴]، جیراک [۱۵] و چو و فریزلیچ [۱۶] مقایسه کرد. در جدول ۱، قابلیت‌های کاربردی روش‌های مختلف شناسایی نقطهٔ تغییر در داده‌های پانلی به تفصیل آمده است.

آمارهٔ جمع تجمعی و آمارهٔ جمع تجمعی دوبل

فرض کنید C_{t_0} عامل جمع تجمعی را نشان می‌دهد؛ به‌طوری‌که $x_{j,t}$ را براساس مدل ۱ در فاصلهٔ زمانی $t \in [1, T]$، به‌عنوان ورودی دریافت می‌کند و $\mathbb{Y}^{j}_{1,t_0,T}$ را برای $t_0 = 1 \ldots T\text{-}1$ به‌صورت زیر بازمی‌گرداند. رابطهٔ ۲ $\mathbb{Y}^{j}_{1,t_0,T}$ را برای $t_0 = 1 \ldots T\text{-}1$، نشان می‌دهد:

$$\mathbb{Y}^{j}_{1,t_0,T} = C_{t_0}\left(\left\{x_{j,t}\right\}_{t=1}^{T}\right) = \left(\frac{t_0\left(T-t_0\right)}{T}\right)^{\gamma}\left(\frac{1}{t_0}\sum_{t=1}^{t_0} x_{j,t} - \frac{1}{\left(T-t_0\right)}\sum_{t=t_0+1}^{T} x_{j,t}\right) \quad (۲)$$

که در آن اندیس j نشان‌دهندهٔ مقطع j ام است و $j \in \{1, \ldots, n\}$، همچنین $\gamma \in (0, 1)$.

جدول ۱. مقایسهٔ روش‌های شناسایی نقطهٔ تغییر در داده‌های پانلی

قابلیت / مرجع	تشخیص نقطهٔ تغییر در میانگین	تشخیص نقطهٔ تغییر در واریانس	تشخیص نقطهٔ تغییر در ساختار وابستگی	تشخیص چند نقطهٔ تغییر	کارایی روش با وجود هم‌بستگی در مشاهدات در طول زمان	کارایی روش با وجود هم‌بستگی در مقاطع در مقایسه با یکدیگر	بررسی کارایی در دقت برآورد	بررسی کارایی در آزمون تشخیص تغییر	روش استفاده‌شده در تشخیص نقطهٔ تغییر	توضیحات
بای [۳]	*	*					*		LSE, QML	برای داده‌هایی با ابعاد کوچک کارایی دارد (برای مثال ۱۰ =T ,۲۰= n)
هوروات و هوسکووا [۴]	*				*			*	CUSUM	برای داده‌هایی با ابعاد بزرگ کارایی دارد (برای مثال ۱۰۰ =T ,۲۰۰= n)
لی و همکاران [۵]		*			*			*	CUSUM	برای داده‌هایی با ابعاد متوسط کارایی دارد (برای مثال ۵۰ =T ,۵۰= n)
دی واچتر و تزاوالیس [۶]			*		*		*	*	GMM	برای داده‌هایی با ابعاد متوسط کارایی دارد (برای مثال ۶ =T ,۱۰۰= n)
زو و همکاران [۸]			*	*	*	*			Copula	اگر تعدادی از داده‌های پانلی در مقاطع و زمان‌های معین نباشند، کارایی خود را از دست نمی‌دهند.
چن و هو [۹]	*				*		*		CUSUM	برای داده‌هایی با ابعاد متوسط کارایی دارد (برای مثال ۱۰ =T ,۱۰۰= n)
پستوها و پستا [۱۱]	*				*		*		LSE	برای داده‌هایی با ابعاد کوچک کارایی دارد (به‌طور مثال ۱۰ =T ,۲۰= n)
انوموتو و ناگاتا [۱۲]	*	*	*	*		*	*	*	BMT	یک نمودار کنترل را طراحی می‌کند که کارایی آن در ابعاد کوچک نشان داده شده است. (برای مثال ۴= n , ۳۰= T)
چو [۱۷]	*			*	*	*	*	*	Double CUSUM	برای داده‌هایی با ابعاد بزرگ کارایی دارد (برای مثال ۱۰۰ =T ,۲۰۰= n)
پژوهش حاضر	*				*	*	*	*	double CUSUM – EWMA	برای داده‌هایی با ابعاد بزرگ، توان تشخیص بیشتری دارد.

چن و هو [۹] برآوردگر جمع تجمعی را براساس رابطهٔ ۲ و با استفاده از جمع‌آوری جمع‌های تجمعی همهٔ مقـاطع، به‌منظور برآورد نقطهٔ تغییر در داده های پانلی تعریف کردند و نشان دادند که مقادیر مختلف γ تا حد معینـی بـر میـزان دقت برآورد تأثیرگذار است.

می‌توان رابطهٔ ۲ را به‌کمک انحراف معیار مشاهدات در طول هر مقطع، به‌صورت رابطهٔ ۳ تعریف کرد:

$$\mathbb{X}^{j}_{1,t_0,T}=C_{t_0}\left(\left\{\sigma_j^{-1}x_{j,t}\right\}_{t=1}^{T}\right)=\frac{1}{\sigma_j}\sqrt{\frac{t_0(T-t_0)}{T}}\left(\frac{1}{t_0}\sum_{t=1}^{t_0}x_{j,t}-\frac{1}{(T-t_0)}\sum_{t=t_0+1}^{T}x_{j,t}\right) \qquad (۳)$$

که در آن σ_j، انحراف معیار مشاهدات در طول مقطـع j ام است که بایـد بـرآورد شـود (در ایـن مقالـه بـرآورد σ_j بـا اسـتفاده از روش ارائـه‌شـده در ملاحظـهٔ (۲، ۳) چـو [۱۷]

به‌دست آمده است). رابطهٔ ۳ از پژوهش سن و سریواستاوا [۱۸] ناشی شده است که این رابطه را با روش حداکثر درست‌نمایی به‌دست آورده‌اند. اگر در رابطهٔ ۳، $\sigma_j = 1$ درنظر گرفته شود، رابطهٔ ۲ به‌دست می‌آید. مطالعات عددی نشان می‌دهد این ساده‌سازی منطقی است [۱۸، ۱۳].

هوروات و هوسکووا [۴] و ژانگ و همکاران [۱۳]، آماره‌هایی را براساس رابطهٔ ۳ و با استفاده از جمع آوری مجذور جمع‌های تجمعی همهٔ مقاطع، برای تشخیص نقطهٔ تغییر داده‌های پانلی تعریف کردند. جیراک [۱۵] آمارهٔ آزمون خود را براساس رابطهٔ ۳ و با استفاده از به‌دست‌آوردن حداکثر نقطه به نقطه روی جمع‌های تجمعی در هر مقطع و هر زمان معرفی کرده است.

روش‌های فوق در استفاده از روش جمع تجمعی، حداکثر نقطه به نقطه یا مجموع (مجذور) جمع‌های تجمعی را درنظر می‌گیرند. مطالعات تجربی چو و فریزلیچ [۱۶] نشان می‌دهد چنین رویکردهایی ممکن است به عملکرد پایین‌تر در تشخیص و تعیین نقاط تغییر با ابعاد بزرگ منجر شوند؛ از این‌رو آن‌ها روشی را ارائه دادند که در آن جمع‌آوری جمع‌های تجمعی در نقاط تغییر به مقاطعی محدود می‌شود که در آن‌ها تغییر رخ داده است.

چو [۱۷] با ارائهٔ آمارهٔ جمع تجمعی دوبل، رویکرد چو و فریزلیچ [۱۶] را به‌شکل قابل‌ملاحظه‌ای ارتقا داد. برای به‌دست‌آوردن آمارهٔ جمع تجمعی دوبل، ابتدا با استفاده از رابطهٔ ۳، مقدار $\mathbb{X}^{j}_{1,t_0,T}$ برای $t_0 = 1 \dots T\text{-}1$ و $j = 1 \dots n$ به‌دست می آید. درواقع $\mathbb{X}^{j}_{1,t_0,T}$ در رابطهٔ ۳، مقدار جمع تجمعی محاسبه‌شده بر $x_{j,t}$ با $t\in[1,\ T]$ را برای $t_0 = 1 \dots T\text{-}1$ نشان می‌دهد. D^{φ}_{m} نیز عامل آمارهٔ جمع تجمعی دوبل است و به‌صورت زیر تعریف می‌شود:

$$D^{\varphi}_{m}\left(\left\{\left|\mathbb{X}^{(j)}_{1,t_0,T}\right|\right\}^{n}_{j=1}\right) = \left\{\frac{m(2n-m)}{2n}\right\}^{\varphi}\left(\frac{1}{m}\sum_{j=1}^{m}\left|\mathbb{X}^{(j)}_{1,t_0,T}\right| - \frac{1}{2n-m}\sum_{j=m+1}^{n}\left|\mathbb{X}^{(j)}_{1,t_0,T}\right|\right) \tag{۴}$$

که در آن برای $m\in\{1,\ \dots,\ n\}$ و $t_0 \in [1, T)$، آمارهٔ جمع تجمعی دوبل، اپراتور D^{φ}_{m} را به‌ترتیب برای مقادیر جمع تجمعی، $\left|\mathbb{X}^{(1)}_{1,t_0,T}\right| \geq \left|\mathbb{X}^{(2)}_{1,t_0,T}\right| \geq \dots \geq \left|\mathbb{X}^{(n)}_{1,t_0,T}\right|$ در هر $t_0 \in [1, T)$، به‌عنوان ورودی دریافت می‌کند. همچنین $\varphi\in[0,\ 1]$ است. سپس برای تشخیص نقطهٔ تغییر در فاصلهٔ زمانی $[1,\ T]$ آمارهٔ جمع تجمعی دوبل به‌صورت زیر تعریف می‌شود:

$$\mathbb{T}^{\varphi}_{1,T} = \max_{t_0\in[1,T)}\max_{1\leq m\leq n} D^{\varphi}_{m}\left(\left\{\left|\mathbb{X}^{(j)}_{1,t_0,T}\right|\right\}^{n}_{j=1}\right) \tag{۵}$$

بنابراین، برآوردگر نقطهٔ تغییر روی بازهٔ زمانی $[1,\ T]$ به‌صورت زیر است:

$$\hat{\eta} = \arg\max_{t_0\in[1,T)}\max_{1\leq m\leq n} D^{\varphi}_{m}\left(\left\{\left|\mathbb{X}^{(j)}_{1,t_0,T}\right|\right\}^{n}_{j=1}\right) \tag{۶}$$

با توجه به خواص ارائه شده از مدل فاکتور پویای عمومی (برای مطالعهٔ مدل فاکتور پویای عمومی (GDFM[۱۰]) به فورنی و همکاران [۱۹] رجوع کنید.) چو [۱۷] روش بوت استرپ را برای انتخاب معیار آزمون پیشنهاد می‌کند که هم‌بستگی مقطع‌ها و درون‌سری‌ها را در داده‌های با ابعاد بزرگ به‌شمار می‌آورد. این معیار آزمون، زمینهٔ تشخیص چند نقطهٔ تغییر در داده هایی با ابعاد بزرگ را با استفاده از الگوریتم دودویی فراهم می‌کند. همچنین الگوریتم تقسیم‌بندی دودویی جمع تجمعی دوبل (DCBS[۱۱]) را برای تقسیم‌بندی داده‌های پانل تشکیل می‌دهد که با معیار آزمون $\pi^{\varphi}_{n,T}$ مجهز شده‌اند؛ بنابراین اگر در هر فاصلهٔ زمانی مانند $\{(s, e);\ 1\leq s < e < T\}$، $\mathbb{T}^{\varphi}_{s,e} > \pi^{\varphi}_{n,T}$ باشد، آنگاه $\hat{\eta}$ برآورد نقطهٔ تغییر است. مقادیر s و e با استفاده از الگوریتم تقسیم بندی دودویی (BS[۱۲]) به شیوهٔ بازگشتی حاصل می شود (برای مطالعهٔ الگوریتم تقسیم بندی دودویی برای تشخیص چند نقطهٔ تغییر به بخش ۳-۲ چو [۱۷] رجوع کنید).

آمارهٔ پیشنهادی جمع تجمعی دوبل-میانگین متحرک موزون

اگر $x_{j,t}$ مقدار مشاهده‌شده براساس مدل ۱ در زمان t و در مقطع j باشد، میانگین متحرک موزون‌نمایی (EWMA[۱۳]) برای مشاهدهٔ زمانی $t\in[1,\ T]$، در مقطع $j\in\{1,\ \dots,\ n\}$، به‌صورت رابطهٔ ۷ تعریف می‌شود:

$$z_{j,t} = \lambda x_{j,t} + (1-\lambda) z_{j,t-1} \tag{۷}$$

در رابطهٔ ۷، λ ضریب هموارسازی و دارای مقدار ثابتی بین صفر و ۱ است. $z_{j,0}$ نیز مقدار میانگین اطلاعات اولیه $\left(\sum_{t=1}^{d} x_{j,t} / d\right)$ است. در داده‌هایی با ابعاد بزرگ هنگامی‌که به‌دنبال نقطهٔ تغییر در بازهٔ زمانی معین $\{1, T\}$ هستیم، فاصلهٔ کوتاه در نزدیک نقطهٔ ابتدایی و انتهایی $\{1, T\}$ نقطهٔ تغییر را شامل نمی‌شوند؛ زیرا این نقاط به شدت از نقاط قبل خود تأثیر می‌پذیرند؛ بنابراین تجزیه و تحلیل نقاط تغییر در آن‌ها باید شامل نقاط زمانی قبل یا بعد از بازهٔ زمانی تعیین شده باشد. چو اندازهٔ این فاصله را با $d = \lceil c \log^2 T \rceil$ تعیین کرده است که در آن $c > 0$ است؛ از این‌رو جست‌وجوی نقاط تغییر تنها در $[1, T] \setminus \mathcal{L}_{1,T}$ صورت می‌گیرد که $\mathcal{L}_{1,T} = [1, 1+d] \cup [T-d, T]$ است؛ بنابراین اگر به جای $z_{j,t-1}$ معادل آن جایگزین شود، رابطهٔ ۸ به‌دست می‌آید:

$$z_{j,t} = \lambda \sum_{l=0}^{t-1} (1-\lambda)^l x_{j,(t-l)} + \frac{1}{d} (1-\lambda)^t \sum_{t=1}^{d} x_{j,t} \qquad (۸)$$

اگر در هر مقطع مشاهدات $x_{j,t}$ روی فاصلهٔ زمانی $t \in [1, T]$، توزیعی نرمال با میانگین μ_j و واریانس σ_j^2 داشته باشند، برای سری‌های زمانی بزرگ، $z_{j,t}$ توزیعی نرمال با میانگین μ_j و واریانس حدی $\left(\frac{\lambda}{2-\lambda}\right) \sigma_j^2$ دارد (برای اثبات به مرجع مونتگومری [۲۰] رجوع کنید). اکنون فرض کنید در مدل داده های پانلی ۱، میانگین مقطع j ام از زمان t_0 به بعد، از μ_j به $\mu_j + \delta_j$ تغییر می‌کند؛ بنابراین اندازهٔ تغییر در میانگین برحسب انحراف معیار، $\frac{\delta_j}{\sigma_j}$ است. از آنجا که $\frac{\delta_j}{\sigma_j} \le \frac{\delta_j}{\sigma_j} \sqrt{\frac{2-\lambda}{\lambda}}$ است، در حالتی که مقدار آستانهٔ آزمون برحسب فرض صفر عدم تغییر در مدل داده های پانلی ۱به‌دست آمده است، کم‌شدن پراکندگی $z_{j,t}$ نسبت به $x_{j,t}$ بر فاصلهٔ زمانی $t \in [1, T]$ این انتظار را به‌وجود می آورد که توان آزمون برای تشخیص نقطهٔ تغییر در میانگین با استفاده از مقادیر $z_{j,t}$ از $x_{j,t}$ بیشتر خواهد شد.

اگر مقادیر متغیر $z_{j,t}$ در رابطهٔ ۸ جایگزین متغیر $x_{j,t}$ در رابطهٔ ۳ شوند، CE_{t_0}، عامل ترکیب جمع تجمعی و میانگین متحرک موزون‌نمایی به‌صورت رابطهٔ ۹ است؛ به‌طوری‌که $z_{j,t}$ بر یک فاصلهٔ زمانی $t \in [1, T]$ به‌عنوان یک ورودی دریافت می شود و $\mathbb{Z}_{1,t_0,T}^{j}$ برای $t_0 = 1 \dots T\text{-}1$، به‌صورت زیر بازمی‌گردد:

$$\mathbb{Z}_{1,t_0,T}^{j} = CE_{t_0}\left(\left\{\sigma_j^{-1} z_{j,t}\right\}_{t=1}^{T}\right) = \frac{1}{\sigma_j}\sqrt{\frac{t_0 (T-t_0)}{T}}\left(\frac{1}{t_0}\sum_{t=1}^{t_0} z_{j,t} - \frac{1}{(T-t_0)}\sum_{t=t_0+1}^{T} z_{j,t}\right) \qquad (۹)$$

سپس برای تعریف آمارهٔ ترکیبی جمع تجمعی دوبل-میانگین متحرک موزون (DCE[۱۴])، عامل D_m^{φ} به‌صورت رابطهٔ ۱۰ تعریف می‌شود:

$$D_m^{\varphi}\left(\left\{\left|\mathbb{Z}_{1,t_0,T}^{(j)}\right|\right\}_{j=1}^{n}\right) = \left\{\frac{m(2n-m)}{2n}\right\}^{\varphi}\left(\frac{1}{m}\sum_{j=1}^{m}\left|\mathbb{Z}_{1,t_0,T}^{(j)}\right| - \frac{1}{2n-m}\sum_{j=m+1}^{n}\left|\mathbb{Z}_{1,t_0,T}^{(j)}\right|\right) \qquad (۱۰)$$

که در آن برای $m \in \{1, \ldots, n\}$ و $t_0 \in [1, T)$، آمارهٔ جمع تجمعی دوبل-میانگین متحرک موزون، اپراتور D_m^{φ} را به‌ترتیب برای مقادیر $\left|\mathbb{Z}_{1,t_0,T}^{(1)}\right| \ge \left|\mathbb{Z}_{1,t_0,T}^{(2)}\right| \ge \ldots \ge \left|\mathbb{Z}_{1,t_0,T}^{(n)}\right|$ در هر $t_0 \in [1, T)$، به عنوان ورودی دریافت می کند. سپس برای تشخیص نقطهٔ تغییر در فاصلهٔ زمانی $[1, T]$، آمارهٔ جمع تجمعی دوبل-میانگین متحرک موزون به صورت زیر تعریف می‌شود:

$$\mathbb{T}_{1,T}^{\varphi} = \max_{t_0 \in [1,T)} \max_{1 \le m \le n} D_m^{\varphi}\left(\left\{\left|\mathbb{Z}_{1,t_0,T}^{(j)}\right|\right\}_{j=1}^{n}\right) \qquad (۱۱)$$

الگوریتم روش پیشنهادی جمع تجمعی دوبل-میانگین متحرک موزون

به طورکلی روش پیشنهادی این مقاله از الگوریتم زیر پیروی می‌کند.

گام اول: اگر $x_{j,t}$ مقدار مشاهده‌شده براساس مدل ۱ در زمان t و در مقطع j باشد، مقادیر میانگین متحرک موزون‌نمایی $z_{j,t}$، براساس رابطهٔ ۸ حاصل می‌شود.

گام دوم: عامل جمع تجمعی را بر مقادیر حاصل از گام اول به ازای هر $t_0 = 1 \dots T\text{-}1$ و $j = 1 \dots n$، اعمال می‌کنیم؛ بنابراین براساس رابطهٔ ۹ مقادیر $\mathbb{Z}_{1,t_0,T}^{j}$ حاصل می‌شود.

گام سوم: برای هر نقطهٔ تغییر احتمالی $t_0 = 1...T\text{-}1$، مقادیر قدر مطلق از $\mathbb{Z}_{1,t_0,T}^{j}$ به ازای $j = 1...n$ به‌صورت نزولی مرتب می‌شود؛ بنابراین مقادیر $\left|\mathbb{Z}_{1,t_0,T}^{(1)}\right| \geq \left|\mathbb{Z}_{1,t_0,T}^{(2)}\right| \geq \ldots \geq \left|\mathbb{Z}_{1,t_0,T}^{(n)}\right|$ در هر $t_0 \in [1,T)$، حاصل خواهد شد.

گام چهارم: عامل جمع تجمعی بر مقادیر حاصل از گام سوم، به ازای هر $m \in \{1, \ldots, n\}$ و $t_0 \in [1,T)$ اعمال می‌شود؛ بنابراین براساس رابطهٔ ۱۰ مقادیر D_m^{φ} حاصل می‌شود.

گام پنجم: برای تشخیص نقطهٔ تغییر در فاصلهٔ زمانی $[1,\ T]$، آمارهٔ جمع تجمعی دوبل-میانگین متحرک موزون، به‌صورت رابطهٔ ۱۱ تعریف می‌شود.

گام ششم: معیار آزمون براساس روش بوت استرپ چو [۱۷] به نام روش GDFM_Boot به‌دست می‌آید. این معیار در کتابخانهٔ نرم‌افزار R، با عنوان پکیج Hdbinseg قابل‌دسترسی است.

گام هفتم: اگر معیار آزمون حاصل از گام ششم را با $\pi_{n,T}^{\varphi}$ نشان دهیم و در فاصلهٔ زمانی $[1,\ T]$، $\mathbb{T}_{1,T}^{\varphi} > \pi_{n,T}^{\varphi}$ باشد، آنگاه $\hat{\eta}$ برآورد نقطهٔ تغییر است و به‌صورت زیر تعریف می‌شود:

$$\hat{\eta} = \arg\max_{t_0 \in [1,T)} \max_{1 \leq m \leq n} D_m^{\varphi}\left(\left\{\left|\mathbb{Z}_{1,t_0,T}^{(j)}\right|\right\}_{j=1}^{n}\right) \qquad (۱۲)$$

مطالعات شبیه‌سازی برای تشخیص نقطهٔ تغییر

داده‌های استفاده‌شده در این مطالعه، براساس داده‌های شبیه‌سازی‌شده چو [۱۷] است که در آن داده‌ها از مدل خودهم‌بستهٔ میانگین متحرک ((ARMA(2, 2) پیروی می‌کنند؛ به‌طوری‌که:

$$\varepsilon_{j,t} = 0.2\varepsilon_{j,t-1} - 0.3\varepsilon_{j,t-2} + u_{j,t} + 0.2u_{j,t-1} \qquad (۱۳)$$

که در آن:

$$u_{j,t} = \sum_{i=0}^{99} \varrho_i v_{j-i,t} \qquad (۱۴)$$

همچنین $v_{j,t} \sim_{iid} N(0,\ \sigma_v^2)$، $\sigma_v = 0.1\varrho^{-1}$، $\varrho_i = \varrho(i+1)^{-1}$ و $\varrho \in \{۰/۲\ ,۰/۵\}$ است. در اینجا، ϱ ضریب تعدیل پراکندگی داده‌هاست.

پانلی با ۲۵۰ مقطع و ۱۰۰ مشاهدهٔ زمانی در هر مقطع را درنظر بگیرید که در آن m مقطع که به صورت تصادفی انتخاب شده‌اند، دارای یک نقطهٔ تغییر در زمان η_1 در میانگین خود هستند. اندازهٔ تغییر برای هر مقطع از تابع یکنواخت (۱/۲۵, ۰/۷۵) U با ضریب δ_1 پیروی می‌کند که در آن $\delta_1 \in \{۰/۰۵\ ,۰/۰۷۵\ ,۰/۱\}$ است. در همهٔ آماره‌ها φ=۰/۵ است. توجه داشته باشید در مشاهداتی با ابعاد بالا که در آن میان داده‌ها هم‌بستگی وجود دارد، برآورد انحراف معیار در هر بار تولید داده با اریبی‌های مختلف همراه خواهد بود؛ بنابراین اگر اندازهٔ تغییر نسبتی از انحراف معیار برآورد شده باشد، نتایج حاصل از هر بار شبیه‌سازی متفاوت خواهد بود. برای کاهش این تفاوت چو [۱۷] به پیروی از جیراک [۱۵] اندازهٔ تغییر را بدون برآورد انحراف معیار تولید می‌کند. درواقع اندازه‌های {۰/۱, ۰/۰۷۵, ۰/۰۵} تقریباً با {σ,۰/۷۵σ, ۰/۵σ} معادل هستند. تعداد تکرار شبیه‌سازی و تعداد تکرار بازنمونه‌گیری بوت استرپ ۱۰۰۰ است. همچنین شاخص اندازه‌گیری دقت برآورد با تعداد تکرارهای شبیه‌سازی برابر است که در شرایط $|\hat{\eta}_1 - \eta_1| < \log T$ صدق می‌کند که در آن η_1 نقطهٔ تغییر صحیح مشترک و $\hat{\eta}_1$ برآورد آن است.

جدول ۲ و ۳ توان آزمون با α=۰/۰۵ و شاخص دقت شناسایی مکان نقطهٔ تغییر، برای آمارهٔ جمع تجمعی دوبل و آمارهٔ پیشنهادی جمع تجمعی دوبل-میانگین متحرک موزون با ۰/۸، ۰/۷۵، ۰/۷، ۰/۶۵، ۰/۶ =λ به‌ترتیب برای ϱ=۰/۲ و ϱ=۰/۵ را نشان می‌دهد. در آمارهٔ جمع تجمعی دوبل-میانگین متحرک موزون، تعداد مشاهدات اولیه d=۵ درنظر گرفته شده است. همچنین، معیار آزمون در هر دو روش یکسان است و براساس روش بوت استرپ چو [۱۷] بر داده‌های پانلی مدل ۱ به‌دست آمده است. باید توجه داشت که مقادیر کوچک λ به کاهش دقت برآورد منجر خواهد شد. شاید این کاهش دقت به این دلیل باشد که مقدار مشاهدهٔ زمانی پس از نقطهٔ تغییر وابستگی بسیاری به مشاهدهٔ زمانی قبل از نقطهٔ تغییر دارد.

جدول ۲. توان آزمون و اندازهٔ شاخص دقت برای روش‌های DC و DCE با λ های مختلف (ϱ=۰/۲)

			Location accuracy (1/1000)						Power Test					
δ_1	m	η_1	DCE λ=۰/۶	DCE λ=۰/۶۵	DCE λ=۰/۷	DCE λ=۰/۷۵	DCE λ=۰/۸	DC	DCE λ=۰/۶	DCE λ=۰/۶۵	DCE λ=۰/۷	DCE λ=۰/۷۵	DCE λ=۰/۸	DC
۰/۰۵	$\sqrt{n}$	۰/۱T	۰	۰	۰	۰	۰	۰	۰/۰۰۱	۰/۰۰۱	۰	۰/۰۰۱	۰/۰۰۱	۰
		۰/۵T	۲	۰	۰	۰	۰	۰	۰/۰۰۶	۰/۰۰۲	۰/۰۰۳	۰/۰۰۳	۰/۰۰۱	۰/۰۰۱
	۰/۴n	۰/۱T	۶	۴	۵	۲	۲	۰	۰/۰۱۴	۰/۰۱۱	۰/۰۱۴	۰/۰۰۵	۰/۰۰۷	۰/۰۰۳
		۰/۵T	۳۵۹	۳۲۷	۲۹۸	۲۸۲	۲۸۶	۲۱۷	۰/۶۶۰	۰/۶۰۰	۰/۵۴۸	۰/۴۸۸	۰/۴۸۴	۰/۳۹۶
۰/۰۷۵	$\sqrt{n}$	۰/۱T	۰	۰	۰	۰	۰	۰	۰	۰/۰۰۱	۰/۰۰۱	۰/۰۰۴	۰/۰۰۱	۰
		۰/۵T	۵	۸	۶	۲	۳	۱	۰/۰۳۳	۰/۰۳۰	۰/۰۱۵	۰/۰۱۳	۰/۰۰۹	۰/۰۰۵
	۰/۴n	۰/۱T	۵۶	۵۲	۴۳	۲۹	۳۳	۲۵	۰/۰۸۰	۰/۰۷۰	۰/۰۶۷	۰/۰۴۰	۰/۰۴۴	۰/۰۳۷
		۰/۵T	۸۱۴	۸۰۴	۷۹۶	۸۱۰	۸۱۶	۷۹۰	۱	۱	۰/۹۹۹	۱	۰/۹۹۹	۰/۹۹۹
۰/۱	$\sqrt{n}$	۰/۱T	۰	۰	۰	۱	۰	۰	۰/۰۰۲	۰/۰۰۷	۰/۰۰۲	۰/۰۰۳	۰/۰۰۲	۰/۰۰۵
		۰/۵T	۲۲	۱۰	۹	۷	۴	۵	۰/۰۵۰	۰/۰۲۹	۰/۰۲۰	۰/۰۱۵	۰/۰۰۷	۰/۰۱۰
	۰/۴n	۰/۱T	۵۱۶	۴۹۵	۴۳۸	۴۱۷	۳۹۶	۳۴۵	۰/۶۲۹	۰/۵۶۶	۰/۵۲۰	۰/۴۹۵	۰/۴۶۰	۰/۳۹۰
		۰/۵T	۹۰۲	۹۱۱	۹۱۰	۹۴۲	۹۳۱	۹۲۹	۱	۱	۱	۱	۱	۱

جدول ۳. توان آزمون و اندازهٔ شاخص دقت برای روش‌های DC و DCE با λ های مختلف (ϱ=۰/۵)

			Location Accuracy (1/1000)						Power Test					
δ_1	m	η_1	DCE λ=۰/۶	DCE λ=۰/۶۵	DCE λ=۰/۷	DCE λ=۰/۷۵	DCE λ=۰/۸	DC	DCE λ=۰/۶	DCE λ=۰/۶۵	DCE λ=۰/۷	DCE λ=۰/۷۵	DCE λ=۰/۸	DC
۰/۰۵	$\sqrt{n}$	۰/۱T	۰	۰	۰	۰	۰	۰	۰/۰۰۲	۰	۰	۰	۰	۰
		۰/۵T	۰	۰	۱	۰	۰	۰	۰/۰۰۵	۰/۰۰۲	۰/۰۰۳	۰/۰۰۲	۰/۰۰۱	۰
	۰/۴n	۰/۱T	۱۹	۱۷	۹	۸	۱۴	۱۳	۰/۰۶۵	۰/۰۵۳	۰/۰۳۵	۰/۰۳۴	۰/۰۳۶	۰/۰۳۵
		۰/۵T	۴۰۶	۳۸۵	۳۲۷	۳۲۲	۳۳۳	۲۷۱	۰/۷۹۸	۰/۷۲۲	۰/۶۳۶	۰/۶۰۴	۰/۶۰۷	۰/۴۹۶
۰/۰۷۵	$\sqrt{n}$	۰/۱T	۰	۰	۰	۰	۰	۰	۰/۰۰۱	۰/۰۰۱	۰	۰	۰/۰۰۱	۰/۰۰۱
		۰/۵T	۳	۳	۰	۱	۲	۲	۰/۰۱۲	۰/۰۰۸	۰/۰۰۳	۰/۰۰۶	۰/۰۰۹	۰/۰۰۴
	۰/۴n	۰/۱T	۶۹	۵۵	۵۲	۳۳	۴۳	۳۶	۰/۰۹۴	۰/۰۷۳	۰/۰۶۸	۰/۰۵۳	۰/۰۶۰	۰/۰۴۲
		۰/۵T	۷۹۴	۸۰۱	۸۱۹	۸۱۷	۸۲۳	۸۱۷	۱	۱	۱	۱	۱	۱
۰/۱	$\sqrt{n}$	۰/۱T	۰	۰	۰	۰	۰	۰	۰/۰۰۳	۰/۰۰۱	۰	۰	۰	۰
		۰/۵T	۳۱	۱۹	۱۴	۴	۹	۱	۰/۰۵۹	۰/۰۴۵	۰/۰۳۰	۰/۰۱۵	۰/۰۱۸	۰/۰۱۰
	۰/۴n	۰/۱T	۷۵۳	۷۷۱	۷۸۹	۷۶۲	۷۶۴	۶۴۹	۰/۹۲۶	۰/۹۲۰	۰/۹۱۵	۰/۸۹۲	۰/۸۹۵	۰/۸۲۹
		۰/۵T	۹۰۳	۹۰۹	۹۰۶	۹۲۶	۹۲۲	۹۳۲	۱	۱	۱	۱	۱	۱

در جدول ۲ و ۳، DC مخفف Double CUSUM و نشان‌دهندهٔ آمارهٔ جمع تجمعی دوبل است. همچنین DCE مخفف Double CUSUM_EWMA و نشان‌دهندهٔ آمارهٔ جمع تجمعی دوبل-میانگین متحرک موزون و λ ضریب هموارسازی به‌کار گرفته‌شده در آماره است. در جدول ۲ و ۳، Power Test نشان‌دهندهٔ توان آزمون در تشخیص وجود نقطهٔ تغییر با حساسیت یک‌هزارم است و Location Accuracy مقدار شاخص دقت برآورد را نسبت به هزار بار تکرار شبیه‌سازی نشان می‌دهد. در این مقاله از نرم‌افزار R ورژن ۳/۵/۱ و پکیج Hdbinseg برای محاسبات استفاده شد.

با توجه به اینکه چو [۱۷] عملکرد DC را با دیگر روش های ادبیات مقایسه کرده و برتری مدل DC را از دیگر مدل ها نشان داده است، این مقاله، مدل پیشنهادی خود را تنها با DC مقایسه می‌کند. نتایج جدول ۲ و ۳ در پنج حالت زیر آمده است:

حالت ۱: هنگامی‌که به‌طور متوسط ۴۰ درصد از مقاطع (۰/۴ n) از تغییر پله‌ای به اندازهٔ ۰/۱ یا ۰/۰۷۵ در زمان ۰/۵T تأثیر می‌گیرند، در حالتی که ضریب تعدیل انحراف معیار داده‌ها ϱ =۰/۲ است، همچنین زمانی که ضریب تعدیل انحراف معیار داده‌ها ϱ=۰/۵ است، توان آزمون برای آماره های مختلف تقریباً صد درصد است. در این موارد توان تشخیص تغییر وجود دارد و شاخص دقت برآورد نقطهٔ تغییر نیز مقادیر نزدیک به هم و استواری را نشان می‌دهد.

حالت ۲: هنگامی‌که به‌طور متوسط ۴۰ درصد از مقاطع (۰/۴ n) تحت تأثیر یک تغییر پله ای به اندازه ۰/۱ یا ۰/۰۷۵ در زمان ۰/۱T قرار می گیرد، در حالتی که ϱ=۰/۲ و ϱ=۰/۵ است، بهبود توان تشخیص نقطهٔ تغییر با استفاده از آمارهٔ پیشنهادی جمع تجمعی دوبل-میانگین متحرک موزون در مقایسه با آمارهٔ جمع تجمعی دوبل مشهود است؛ به‌طوری‌که در λ=۰/۶ بهترین نتیجه حاصل می‌شود.

حالت ۳: هنگامی که به‌طور متوسط ۴۰ درصد از مقاطع (۰/۴ n) از تغییر پله‌ای به‌اندازهٔ ۰/۰۵ در زمان ۰/۵T تأثیر می‌گیرند، بهبود در توان تشخیص نقطهٔ تغییر در هر دو حالت ϱ=۰/۲ و ϱ=۰/۵ با استفاده از آمارهٔ پیشنهادی مشاهده می‌شود. درصورتی‌که نقطهٔ تغییر در زمان ۰/۱T باشد، نتایج با استفاده از آمارهٔ پیشنهادی در λ های کوچک بهبود می‌یابد.

حالت ۴: هنگامی‌که در تعداد کمی از مقاطع ($\sqrt{n}$) تغییر پله‌ای به اندازهٔ ۰/۱ یا ۰/۰۷۵ در زمان ۰/۵T رخ می دهد، توان تشخیص نقطهٔ تغییر با استفاده از آمارهٔ پیشنهادی در λ های کوچک بهبود می‌یابد.

حالت ۵: هنگامی‌که در تعداد کمی از مقاطع ($\sqrt{n}$) تغییر پله‌ای به اندازهٔ ۰/۱، ۰/۰۷۵ یا ۰/۰۵ در زمان ۰/۱T، یا تغییر پله‌ای ۰/۰۵ در زمان ۰/۵T رخ می‌دهد، احتمال تشخیص یک تغییر غیرتصادفی در آماره های مختلف تقریباً صفر است.

موردکاوی

برای ارزیابی روش پیشنهادی این پژوهش، گزارش های پایان هر روز قیمت‌های سهام مدنظر قرار می‌گیرد که چو [۱۷] در فاصلهٔ زمانی ۲۴ فوریه ۲۰۱۵ تا ۲۳ فوریه ۲۰۱۶ برای ارزیابی مدل پیشنهادی خود استفاده کرده است. باید توجه داشت داده‌های فوق به قیمت‌های سهام ۱۰۰ شرکت برتر در بازار بورس ایالات‌متحده آمریکا مربوط است که با شاخص S&P۱۰۰ ارزیابی می‌شوند. فهرست مؤلفه‌های شاخص S&P۱۰۰ برای دورهٔ زمانی مورد نظر از آدرس اینترنتی https://nemozny.github.io/dataset قابل‌دستیابی است. چو [۱۷] در بررسی خود، به آخرین نقطهٔ تغییر توجه کرد. در این پژوهش، با استفاده از آمارهٔ پیشنهادی جمع تجمعی دوبل-میانگین متحرک موزون و با λ=۰/۶ (به‌دلیل داشتن توان بالاتر در ارزیابی شبیه سازی) آخرین نقطهٔ تغییر در قیمت سهام شرکت‌های برتر ایالات‌متحده در بازهٔ تعریف‌شدهٔ بالا شناسایی شد. شکل ۱ نتایج روش پیشنهادی را که در نرم‌افزار R گزارش شده است نشان می‌دهد. همان‌طور که مشاهده می‌شود، آخرین نقطهٔ تغییر، زمان ۲۲۱ و مربوط به ۷ ژانویه ۲۰۱۶ است. چو در گزارش خود، نقطهٔ تغییر را در زمان ۲۲۰ شناسایی کرده که به ۶ ژانویهٔ ۲۰۱۶ مربوط است. در جدول ۴، شرکت‌هایی که بیشترین پرش را در نقطهٔ تغییر دارند، به‌ترتیب آمده‌اند.

```
                [,1]       [,2]       [,3]
level_index    1.000    2.00000    3.00000
child_index    1.000    2.00000    4.00000
s              1.000   18.00000  127.00000
b             17.000  126.00000  221.00000
e            252.000  252.00000  252.00000
test_stat    262.593   34.29913   24.19499

$ecp
[1]  17 126 221

$thr
     99%
17.55767
```

شکل ۱. نتایج اجرای روش پیشنهادی با استفاده از الگوریتم تقسیم‌بندی دودویی در داده‌های مالی

جدول ۴. شرکت‌های دارای نقطهٔ تغییر مشترک در ۷ ژانویهٔ ۲۰۱۶

شماره	نماد شرکت	شماره	نماد شرکت	شماره	نماد شرکت
۱	MS	۱۳	FDX	۲۵	WFC
۲	C	۱۴	PFE	۲۶	BRK.B
۳	AXP	۱۵	WBA	۲۷	VZ
۴	HAL	۱۶	GD	۲۸	CSCO
۵	GS	۱۷	T	۲۹	JPM
۶	GILD	۱۸	CAT	۳۰	AIG
۷	COF	۱۹	BAC	۳۱	BIIB
۸	MET	۲۰	IBM	۳۲	USB
۹	F	۲۱	BK	۳۳	BA
۱۰	NKE	۲۲	MDLZ	۳۴	MA
۱۱	AAPL	۲۳	ABT		
۱۲	COP	۲۴	SO		

نتیجه‌گیری

در این مطالعه، مسئلهٔ تشخیص نقطهٔ تغییر در میانگین داده‌های پانل با استفاده از ترکیب روش میانگین متحرک موزون نمایی با روش جمع تجمعی دوبل، پیشنهاد و تجزیه و تحلیل شد. روش جمع تجمعی دوبل با توجه به دقت بسیاری که در برآورد نقطهٔ تغییر دارد، در تشخیص این نقطه به‌کمک آزمون بوت استرپ توان کمی در تغییرات کوچک دارد؛ بنابراین با هدف بالابردن توان تشخیص تغییر در میانگین داده‌های پانلی با استفاده از این آزمون، پژوهش حاضر آمارهٔ ترکیبی جمع تجمعی دوبل-میانگین متحرک موزون را پیشنهاد می‌دهد. براساس مطالعات شبیه‌سازی، این رویکرد موجب می‌شود دقت برآورد نقطهٔ تغییر کیفیت خود را حفظ کند و توان تشخیص تغییر از دیگر روش‌ها بهبود یابد.

منابع

1. Joseph, L., and Wolfson, D. B., (1992). “Estimation in Multi-Path Change-Point Problems”, *Communications in Statistics Theory and Methods*, Vol. 21, No. 4, PP. 897-913.
2. Joseph, L., and Wolfson, D. B., (1993). “Maximum Likelihood Estimation in the Multi-Path Change Point Problem”, *Annals of the Institute of Statistical Mathematics*, Vol. 45, No. 3, PP. 511-530.
3. Bai, J., (2010). “Common Breaks in Means and Variances for Panel Data”, *Journal of Econometrics*, Vol. 157, No. 1, PP. 78-92.
4. Horváth, L., and Hušková, M., (2012). “Change-Point Detection in Panel Data”, *Time Series Anallysis*, Vol. 33, No. 4, PP. 631-648.
5. Li, F., Tian, Z., Xiao, Y., and Chen, Z., (2015). “Variance Change-Point Detection in Panel Data Models”, *Review of Economics Letters*, Vol. 126, No. 1, PP. 140-143.
6. De Wachter, S., and Tzavalis, E., (2012). “Detection of Structural Breaks in Linear Dynamic Panel Data Models”, *Computational Statistics and Data Analysis*, Vol. 56, No. 11, PP. 3020-3034.
7. Arellano, M., and Bond, S., (1991). “Some Tests of Specification for Panel Data: Monte Carlo Evidence and An Application to Employment Equations”, *Economic Studies*, Vol. 58, No. 2, PP. 277-297.
8. Zhu, X., Li, Y., Liang, C., Chen, J., and Wu, D., (2013). “Copula Based Change Point Detection for Financial Contagion in Chinese Banking”, *Information Technology and Quantitative Management (ITQM)*, Vol. 17, No. 1, PP. 619-626.
9. Chen, Z., and Hu, Y., (2015). “Cumulative Sum Estimator for Change-Point in Panel Data”, *Statistical Papers*, Vol. 58, No. 3, PP. 707-728.
10. Peštová, B., and Pešta, M., (2015). “Testing Structural Changes in Panel Data with Small Fixed Panel Size and Bootstrap”, *Metrika*, Vol. 78, No. 6, PP. 665-689.
11. Peštová, B., and Pešta, M., (2017). “Change Point Estimation in Panel Data Without Boundary Issue”, *Risks*, Vol. 5, No. 1, PP. 7.
12. Enomoto, T., and Nagata, Y., (2016). “Detection of Change Points in Panel Data Based on Bayesian MT Method”, *Total Quality Science*, Vol. 2, No. 1, PP. 36-47.
13. Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z., (2010). “Detecting Simultaneous Change Points in Multiple Sequences”, *Biometrika*, Vol. 97, No. 3, PP. 631-645.
14. Enikeeva, F., and Harchaoui, Z., (2014). *High-Dimensional Change-Point Detection with Sparse Alternatives,* Arxiv Preprint, Arxiv: 1312, 1900.
15. Jirak, M., (2015). “Uniform Change Point Tests in High Dimension”, *Annals of Statistics*, Vol. 43, No. 6, PP. 2451-2483.
16. Cho, H., and Fryzlewicz, P., (2015). “Multiple-Change-Point Detection for High Dimensional Time Series Via Sparsified Binary Segmentation”, *Journal of the Royal Statistical Society: Series B*, Vol. 77, No. 2, PP. 475-507.
17. Cho, H., (2016). “Change-Point Detection in Panel Data Via Double CUSUM Statistic”, *Electronic Journal of Statistics*, Vol. 10, No. 2, PP. 2000-2038.
18. Sen, A., and Srivastava, M. S., (1975). “On Tests for Detecting a Change in Mean”, *Annals of Statistics*, Vol. 3, No. 1, PP. 98-108.
19. Forni, M., Hallin, M., Lippi, M., and Reichlin, L., (2000). “The Generalized Dynamic Factor Model: Identification and Estimation”, *The Review of Economics and Statistics*, Vol. 82, No. 4, 540-554.
20. Montgomery, D. C., (2007). *Introduction to Statistical Quality Control: John Wiley and Sons*.

واژه‌های انگلیسی به ترتیب استفاده در متن

1. Independent and Identically Distributed
2. Least Squares Estimation
3. Quasi-Maximum Likelihood
4. Cumulative Sum
5. Generalized Method of Moments
6. Copula
7. Mahalanobis-Taguchi
8. Bayesian Mahalanobis-Taguchi
9. Double CUSUM
10. Generalized Dynamic-Factor Model

11. Double CUSUM and Binary Segmentation
12. Binary Segmentation
13. Exponentially Weighted Moving Average
14. Double CUSUM_EWMA