بسم الله الرحمن الرحيم

انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990

دانشگاه کردستان
University of Kurdistan

مجموعه مقالات پانزدهمین سمینار

احتمال و فرآیند های تصادفی

- ۸ و ۹ شهریور ۱۴۰۴
- دانشگاه کردستان

مجموعه مقالات پانزدهمین سمینار احتمال و فرایندهای تصادفی

۸ و ۹ شهریور ۱۴۰۴ دانشگاه کردستان

سنندج

شناسنامه کتاب

عنوان: مجموعه مقالات پانزدهمین سمینار احتمال و فرایندهای تصادفی

موضوع: آمار، احتمال، فرآیندهای تصادفی

گردآوری: دکتر سیروس فتحی‌منش، دکتر هادی امامی، دکتر هادی احمدی

تاریخ برگزاری: هشتم و نهم شهریورماه ۱۴۰۴

محل برگزاری: دانشگاه کردستان، سنندج

به‌نام خداوند جان و خرد

پیشگفتار

خداوند متعال را سپاسگزاریم که توفیق برگزاری پانزدهمین سمینار احتمال و فرایندهای تصادفی را در دانشگاه کردستان و با همکاری انجمن آمار ایران، به منظور خدمت به جامعه علمی کشور، نصیب ما نمود. این سمینار با هدف گردهمایی صاحب‌نظران و پژوهشگران حوزه احتمال و فرایندهای تصادفی برگزار شد و فرصت مغتنمی برای تبادل نظر و ارائه تحقیقات و دستاوردهای نوین این عرصه فراهم آورد.

مجموعه حاضر، شامل مقالات فارسی پذیرفته‌شده در این سمینار است که امیدواریم گشاینده افق‌های تازه‌ای پیش‌روی پژوهشگران، به‌ویژه دانشجویان تحصیلات تکمیلی، باشد. در این دوره از سمینار، در مجموع ۹۸ مقاله به زبان‌های فارسی و انگلیسی به دبیرخانه ارسال گردید که پس از طی فرایند داوری، ۷۵ مقاله برای ارائه و چاپ به صورت کامل یا خلاصه پذیرفته شد. در فرایند ارزیابی مقالات، همکاران ارجمندی از دانشگاه‌های سراسر کشور ما را یاری نمودند. بیشترین تعداد مقالات پذیرفته‌شده با ۱۵ مقاله، متعلق به محور سری‌های زمانی و کاربردها بود.

علاوه بر ارائه مقالات، هفت سخنرانی کلیدی توسط اساتید مدعو، شش نشست تخصصی در زمینه‌های احتمال، فرایندهای تصادفی، سری‌های زمانی، نظریه توزیع‌ها، آنالیز تصادفی، ریاضیات مالی، استنباط آماری برای فرایندهای تصادفی و همچنین کاربرد و نقش فرایندهای تصادفی در هوش مصنوعی و علم داده، و یک کارگاه آموزشی در حوزه ریاضیات مالی برگزار گردید. مقالات مندرج در این مجموعه، بر اساس محورهای اصلی سمینار تدوین شده‌اند. این مقالات در پایگاه استنادی علوم جهان اسلام (ISC) و پایگاه سیویلیکا ثبت شده‌اند. امیدواریم این اثر، گامی هرچند کوچک در جهت پیشبرد مرزهای دانش در حوزه تخصصی خود باشد و مورد توجه و استفاده جامعه علمی کشور قرار گیرد. از خوانندگان گرامی دعوت می‌شود تا با مطالعه این مقالات، از یافته‌های نوین پژوهشگران این عرصه آگاه شوند.

برگزاری این رویداد علمی بدون مساعدت و راهنمایی‌های ارزشمند همکاران محترم در دانشگاه‌های مختلف کشور میسر نبود. همکاری این عزیزان در ارسال و داوری مقالات، ارائه راهکارهای سودمند و مشارکت فعال، سهمی بسزا در موفقیت سمینار داشت. همچنین از پشتیبانی و زحمات بی‌دریغ ریاست محترم دانشگاه و رئیس محترم دانشکده علوم پایه دانشگاه کردستان، ریاست محترم انجمن آمار ایران، واحد فناوری اطلاعات دانشگاه کردستان و کلیه

کارکنان و دانشجویانی که در برگزاری هرچه بهتر این سمینار یاری‌گر ما بودند، صمیمانه سپاسگزاریم. در پایان، حمایت‌های مالی سازمان برنامه و بودجه استان کردستان که مشوقی ارزشمند در برگزاری این سمینار بود، را ارج می‌نهیم.

کمیته‌های علمی و اجرایی پانزدهمین سمینار احتمال و فرایندهای تصادفی

فهرست

مجریان و حامیان ۱

اعضای کمیته اجرایی ۲

اعضای کمیته علمی ۳

هیأت داوران ۴

مقالات ارائه شده ۵

مجریان و حامیان :

دانشگاه کردستان

انجمن آمار ایران

سازمان مدیریت و برنامه ریزی

استان کردستان

اعضای کمیته اجرایی سمینار:

- دکتر هادی احمدی
- دکتر هادی امامی (دبیر اجرایی سمینار)
- دکتر حسین بیورانی
- خانم سمیه جعفررمشتی
- دکتر کوروش دادخواه
- دکتر شاهو زارعی
- خانم سمیه زند
- دکتر سیروس فتحی منش
- دکتر اقبال قادری

اعضای کمیته علمی سمینار :

- دکتر محمد امینی (دانشگاه فردوسی مشهد)
- دکتر حسین بیورانی (دانشگاه کردستان)
- دکتر امید خوارزمی(دانشگاه ولی‌عصر رفسنجان)
- دکتر زهرا رضایی قهرودی (دانشگاه تهران)
- دکتر شاهوزارعی (دانشگاه کردستان)
- دکتر احمدرضا سلطانی (دانشگاه شیراز)
- دکتر سیروس فتحی منش (دانشگاه کردستان- دبیر علمی سمینار)
- دکتر علیرضا نعمت اللهی (دانشگاه شیراز)

اسامی داوران :

دکتر بهاره اختری	**دانشگاه اصفهان**
دکتر هادی امامی	**دانشگاه کردستان**
دکتر محمد امینی	**دانشگاه فردوسی**
دکتر ابراهیم امینی سرشت	**دانشگاه بو علی**
دکتر سلمان ایزدخواه	**دانشگاه کردستان**
دکتر محی‌الدین ایزدی	**دانشگاه رازی**
دکتر حسین بیورانی	**دانشگاه کردستان**
دکتر افشین پرورده	**دانشگاه اصفهان**
دکتر عبدالله جلیلیان	**دانشگاه رازی**
دکتر امید خادم نوع	**دانشگاه زنجان**
دکتر مجید خالدی	**دانشگاه تربیت مدرس**
دکتر امید خوارزمی	**دانشگاه ولی عصر(عج) رفسنجان**
دکتر کاووس خورشیدیان	**دانشگاه شیراز**
دکتر زهرا رضایی قهرودی	**دانشگاه تهران**
دکتر شاهو زارعی	**دانشگاه کردستان**
دکتر شیوا زمانی	**دانشگاه صنعتی شریف**
دکتر عرفان صلواتی	**دانشگاه امیرکبیر**
دکتر محمد فتحی	**دانشگاه کردستان**
دکتر سیروس فتحی منش	**دانشگاه کردستان**
دکتر مهدی کلانتری	**دانشگاه پیام نور زنجان**
دکتر مهرناز محمدپور	**دانشگاه مازندران**
دکتر علی محمدیان مصمم	**دانشگاه زنجان**
دکتر رحیم محمودوند	**دانشگاه بوعلی**

مقالات ارائه شده

خوشه‌بندی دوطرفه فازی با استفاده از تابع فاصله اطلاعاتی،

احمدی، ز. و زارعی، ر. ۹

یک روش تنک‌سازی فرایندهای نقطه‌ای مبتنی بر مجموعه‌های تصادفی بولی،

اسدی، ر. و خزایی، م. و گنجعلی، م. ۱۸

آزمون نیکویی برازش تقارن توزیع‌های پیوسته تحت داده‌های از چپ بریده شده و از راست سانسور شده،

اکبری، م ۳۱

یک کلاس از مدل‌بندی سری‌زمانی دوخطی گسسته‌مقدار با کاربردهای آن،

بامدادی، ر. و محمودپور، م. ۳۹

مدل بندی قابلیت اطمینان سیستم های تنش -مقاومت بر اساس رویکرد تابع مولد عام،

تصدیقی، ب. وزارعی، ر. وفتحی، ب. ۴۸

اخلاق در آمار و احتمال،

دولتی، ع. ۵۷

نقش فرآیندهای تصادفی در هوش مصنوعی و علم داده: یک مطالعه موردی،

راستین، ا. ۷۰

فرایند INAR(۱)-PJ: یک جایگزین جدید فرایند INAR (۱) پواسونی،

راسخی، م ۷۸

بررسی انشعاب‌های تصادفی مدل رشد لجستیک جمعیت،

مهموئی، س. و ربیعی مطلق، ا. و محمدی نژاد، ح. م. ۸۸

تشخیص اعداد دست‌نویس با رهیافت یادگیری ماشین،

رضایی طالقانی، ن. و حاج‌رجبی، آ. ۹۶

مقایسه دو رویکرد نیم‌رخی و نامقید در برازش مدل رگرسیون بر اساس نمونه داده‌های ترکیبی با تواتر متفاوت،

رنجبر، ف. و آقامحمدی، ع. و ادیب، م. ۱۰۴

برخی نتایج مجانبی در مورد اختلاف میانگین دو جامعه در داده‌های تابعی جزئی مشاهده‌شده،

سلیمی فر، پ. و حسینی نسب، س. م. و خادم نوع، ا. ۱۱۲

مدل خودبازگشتی صحیح مقدار مرتبه اول فصلی با توزیع حاشیه‌ای دلاپورت،

شالباف، م. و پرهام، غ. ۱۱۹

رویکرد فازی خوشه بندی داده ها بر پایه الگوریتم،

میاهی، م. و طارمی، ب. و میاهی، م. ۱۲۸

بررسی رفتار مجانبی رهیافت PTE در حضور هم‌خطی چندگانه،

طباطبایی شیرازی، س. ا. ح. و عمادی، م. وآرشی، م. و سیف‌اللهی، س. ۱۳۴

چگونه نظریه فرآیندهای تصادفی به مدل گسترش یافته در سنتز مدرن تکامل کمک کرد؟

عریضی، ح. ۱۴۳

بررسی معیار Covar بر اساس رویکرد COUPLE-ARCH و توزیع GED،

علیزاده، ف. و امینی، م. و محتشمی برزادران، غ. ۱۵۳

مروری بر استفاده از خانواده توزیع‌های فاز-نوع در مدل شوک‌های تعمیم یافته،

فتحی‌منش، س. و ایزدی، م. ۱۶۲

بررسی عملکرد مجانبی برآوردگر ترکیبی در مدل‌های خطی با خطا در اندازه‌گیری،

قپانی، ف. ۱۷۴

مقایسه روش‌های کلاسیک و بیزی در برآورد پارامتر قابلیت اعتماد تنش-مقاومت برای توزیع گمپرتز تعمیم‌یافته یکه،

کاراندیش مروستی، ا. و ارمز، ا. و بصیرت، م. ۱۸۲

کاربرد روش استوار مونت‌کارلوی بهینه در استنباط بیزی بدون درست‌نمایی برای داده‌های وابسته،

کریمی، ا. و حسینی، ف. ۱۹۲

مفصل‌های دایره‌ای-خطی: ویژگی‌ها و کاربردها،

کریمی، ن. و دست‌برآورده، ع. ۲۰۰

مدل‌سازی فرایندهای سری‌زمانی شبکه عصبی بیزی برای شناسایی پیش‌زلزله با ناهنجاری‌های یونوسفری زلزله خاش،

کیکاووسی، م. و کریمی، ف. و حسینی، ا. ۲۱۳

پیش‌بینی زمان حوادث ترافیکی با استفاده از شبکه عصبی بقا،

محمدی، ف. و آرشی، م. و حبیبی راد، آ. و محمدزاده، ا. ۲۲۱

مدل‌بندی تصادفات جاده‌ای استان خراسان جنوبی با استفاده از مدل‌های اتورگرسیو صحیح مقدار،

نخعی‌زاده، ز. و جمهوری، س. ۲۳۶

فرایند چوب‌شکنی رگرسیون-بتا با کوواریانس نامانا وابسته به متغیر کمکی،

یارعلی، ا. و ریواز، ف. و جعفری خالدی، م. ۲۴۵

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

خوشه‌بندی دوطرفه فازی با استفاده از تابع فاصله اطلاعاتی

زهرا احمدی [۱] ، دکتر رضا زارعی[۲]

[۱]استادیار گروه آمار، دانشکده علوم ریاضی، دانشگاه گیلان

[۲]فارغ التحصیل کارشناسی ارشد آمارریاضی، گروه آمار، دانشکده علوم ریاضی، دانشگاه گیلان

چکیده: با گسترش روزافزون داده‌ها در حوزه‌های مختلف علمی، نیاز به روش‌های مؤثر برای تحلیل و استخراج الگوهای پنهان در داده‌ها بیش از پیش احساس می‌شود. خوشه‌بندی یکی از روش‌های مهم در داده‌کاوی غیرنظارتی است که با هدف شناسایی ساختارهای پنهان و تقسیم داده‌ها به گروه‌های مشابه به‌کار می‌رود. در این میان، خوشه‌بندی فازی به‌دلیل انعطاف‌پذیری بالا در برخورد با داده‌های نادقیق، جایگاه ویژه‌ای یافته است. در این مقاله ضمن معرفی رویکرد خوشه‌بندی فازی دوطرفه، یک الگوریتم جدید در این حوزه مورد بررسی قرار می‌گیرد. جزئیات محاسباتی این الگوریتم با استفاده از مجموعه داده‌های واقعی در حوزه زیست‌پزشکی تشریح شده و عملکرد آن با چند الگوریتم مشابه مقایسه شده است. نتایج نشان‌دهنده برتری نسبی الگوریتم مذکور از نظر دقت، تفسیرپذیری و معیارهای ارزیابی خوشه‌بندی است.

واژه‌های کلیدی: خوشه‌بندی فازی، خوشه‌بندی دوطرفه، تابع هدف، توابع عضویت

کد موضوع‌بندی ریاضی (۲۰۲۰): 62H30، 03B52، 68T09.

۱ مقدمه

تحلیل و دسته‌بندی داده‌های حجیم و پیچیده یکی از چالش‌های اساسی در علوم داده و آمار است. خوشه‌بندی[1] به عنوان یکی از روش‌های پایه در داده‌کاوی، نقش مؤثری در کشف ساختار درون‌داده‌ها دارد. هدف اصلی خوشه‌بندی، تقسیم داده‌ها به گروه‌هایی است که درون هر گروه، داده‌ها بیشترین شباهت و بین گروه‌ها بیشترین تفاوت را داشته باشند. در روش‌های سنتی خوشه‌بندی یا خوشه‌بندی سخت[2]، هر داده تنها به یک خوشه اختصاص می‌یابد. با این حال، بسیاری از داده‌های دنیای واقعی دارای مرزهای نامشخص یا هم‌پوشانی هستند.

[۱] سخنران، ahmadizahra19.77@gmail.com

[1]Clustering

[2]Hard Clustering

در چنین شرایطی، خوشه‌بندی فازی[3] گزینه‌ای مناسب‌تر است زیرا به هر داده اجازه می‌دهد تا به چند خوشه با درجات عضویت متفاوت تعلق گیرد.

در بسیاری از کاربردها، مانند تحلیل بیان ژن، خوشه‌بندی همزمان نمونه‌ها (ردیف‌ها) و ویژگی‌ها (ستون‌ها) اهمیت فزاینده‌ای پیدا کرده است. این رویکرد که به خوشه‌بندی دوطرفه موسوم است، قادر است زیرگروه‌هایی از نمونه‌ها را همراه با زیرگروه‌هایی از ویژگی‌ها که رفتاری مشابه از خود نشان می‌دهند، شناسایی کند. علی‌رغم پیشرفت‌های قابل توجه در خوشه‌بندی فازی دوطرفه، همچنان چالش‌هایی اساسی در مواجهه با داده‌های بسیار پیچیده و نویزی و به ویژه در استخراج خوشه‌های با مرزهای هم‌پوشان و نتایج با قابلیت تفسیرپذیری بالا وجود دارد. روش‌های موجود اغلب در مدل‌سازی دقیق وابستگی‌های پیچیده میان نمونه‌ها و ویژگی‌ها و همچنین مدیریت بهینه ابهام و عدم قطعیت در داده‌ها با محدودیت‌هایی مواجه‌اند. در این راستا، این مقاله به معرفی و بررسی یک الگوریتم جدید مبتنی بر اطلاعات می‌پردازد. نوآوری اصلی این الگوریتم در استفاده از یک تابع فاصله اطلاعاتی و ترکیب دو عبارت آنتروپی مجزا برای کنترل پراکندگی عضویت‌ها در سطرها و ستون‌ها است که همگرایی پایدارتر، مدیریت بهینه عدم قطعیت و در نتیجه نتایج دقیق‌تر و قابل تفسیرتری ارائه می‌دهد.

باقیمانده این مقاله به شرح زیر سازماندهی شده است: در بخش دوم، مبانی نظری خوشه‌بندی دوطرفه فازی به کوتاهی مرور می‌شود. بخش سوم به معرفی جامع و جزئیات محاسباتی الگوریتم پیشنهادی اختصاص دارد. در بخش چهارم، نتایج آزمایش‌های تجربی روی مجموعه داده‌های واقعی و مقایسه عملکرد الگوریتم تحت بررسی با الگوریتم‌های موجود در این زمینه ارائه می‌شود. در نهایت، بخش پنجم شامل بحث و نتیجه‌گیری کلی از یافته‌های پژوهش است.

۲ خوشه‌بندی دوطرفه

خوشه‌بندی کلاسیک معمولاً تنها بر سطرها (نمونه‌ها) یا ستون‌ها (ویژگی‌ها) انجام می‌شود. اما داده‌های پیچیده‌تر، مانند ماتریس‌های با ابعاد بزرگ و دارای ساختارهای پنهان، نیازمند روشی هستند که بتواند هم‌زمان خوشه‌بندی روی نمونه‌ها و ویژگی‌ها انجام دهد. این روش با عنوان خوشه‌بندی دوطرفه[4] شناخته می‌شود که اولین بار توسط بوریس میرکین مطرح شد. در این روش، هدف شناسایی زیرماتریس‌هایی از داده‌هاست که در آن‌ها گروهی از ویژگی‌ها با گروهی از نمونه‌ها به‌صورت هم‌زمان الگوهای مشترکی دارند. خوشه‌بندی دوطرفه کاربردهای گسترده‌ای در متن‌کاوی، زیست‌فناوری، داده‌های پزشکی و سیستم‌های توصیه‌گر دارد و نسبت به روش‌های سنتی، نتایج دقیق‌تر و تفسیرپذیرتری ارائه می‌دهد. تفاوت این روش با خوشه‌بندی کلاسیک در این است که خوشه‌بندی دوطرفه هم‌زمان روی ردیف‌ها و ستون‌ها کار می‌کند و خروجی آن زیرماتریس‌هایی با الگوهای مشترک است و انعطاف‌پذیری زیادی دارد یعنی هر آیتم می‌تواند در چند خوشه باشد این در حالی است که در خوشه‌بندی کلاسیک خوشه‌بندی فقط روی ردیف یا ستون انجام می‌شود و خروجی آن گروهی از ردیف‌ها یا ستون‌ها است و انعطاف‌پذیری کمی دارد.فرارو.ام.بی. (۲۰۲۴)

فرض کنید X یک ماتریس دادگان $n \times p$ باشد. هدف الگوریتم خوشه‌بندی دوطرفه تقسیم‌بندی همزمان n سطر به k خوشه و p ستون به h خوشه است. k–میانگین دوطرفه توسط ویچی و همکارانش (۲۰۰۱) مطرح شد. نسخه فازی آن توسط فرارو و همکارانش (۲۰۱۵) معرفی شد و سپس در سال ۲۰۲۱ عمیقا مورد بررسی قرار گرفت.

[3] Fuzzy Clustering
[4] Co-Clustering

تابع هدف الگوریتم k–میانگین دوطرفه فازی به صورت زیر مدل‌بندی می‌شود

$$J(\mathbf{X}, U, C) = \sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{f=1}^{h}\sum_{g=1}^{p}(x_i - c_j)^2(u_{ij})^m(v_{gf})^l$$

که در آن

$$\sum_{j=1}^{k} u_{ij} = 1 \quad \sum_{f=1}^{h} v_{gf} = 1,\ u_{ij}, v_{gf} \in [0, 1].$$

همچنین U ماتریسی $n \times k$ حاوی درجات عضویت سطرها و V ماتریسی $p \times h$ حاوی درجات عضویت ستون‌ها می‌باشند. m و l مشخص‌کننده درجات فازی و اعداد حقیقی و بزرگتر از یك هستند.

۳ الگوریتم خوشه‌بندی دوطرفه فازی مبتنی بر اطلاعات

الگوریتم خوشه‌بندی دوطرفه فازی مبتنی بر اطلاعات[5] (به کوتاهی $ibFCC$) یکی از روش‌های نوین در زمینه خوشه‌بندی فازی دوطرفه است که با هدف بهبود دقت خوشه‌بندی و تفسیر بهتر خوشه‌ها طراحی شده است. این الگوریتم از تابع هدفی بهره می‌برد که شامل سه جزء اصلی است: (۱) تابع فاصله اطلاعاتی بین داده‌ها و مراکز خوشه‌ها، (۲) آنتروپی توزیع عضویت نمونه‌ها در خوشه‌ها، و (۳) آنتروپی توزیع عضویت ویژگی‌ها در خوشه‌ها.

در این مدل، دو ماتریس عضویت به نام‌های U و V تعریف می‌شوند که به ترتیب درجات عضویت نمونه‌ها و ویژگی‌ها را در خوشه‌های مربوطه تعیین می‌کنند. پارامترهای Tu و Tv برای تنظیم میزان پراکندگی عضویت‌ها استفاده می‌شوند. مزیت اصلی ibFCC نسبت به روش‌های مشابه در استفاده از فاصله اطلاعاتی به‌جای فاصله اقلیدسی است که باعث می‌شود ارتباط میان ویژگی‌ها و نمونه‌ها بهتر مدل‌سازی شود.

هدف $ibFCC$ به حداقل رساندن تابع هدف معادله (۱.۳) با توجه به محدودیت‌های معادله زیر است

$$J_{ibFCC} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{k} u_{ci}v_{cj}d_{cij} + T_u\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}lnu_{ci} + T_v\sum_{c=1}^{C}\sum_{j=1}^{k} v_{cj}lnv_{cj}. \tag{۱.۳}$$

$$\sum_{c=1}^{C} u_{ci} = 1 \quad \text{و} \quad \sum_{j=1}^{k} v_{cj} = 1 \qquad i = 1, \ldots, n \quad \text{و} \quad c = 1, \ldots, C$$

عبارت اول معادله (۱.۳) درجه تطابق است که باید طی فرایند خوشه‌بندی دوطرفه به حداقل برسد تا افراد و ویژگی‌های بسیار مرتبط را باهم خوشه‌بندی دوطرفه کند. u_{ci} و v_{cj} دو تابع عضویت هستند که به ترتیب عضویت اسناد و ویژگی‌ها را نشان می‌دهند. عبارت دوم و سوم عوامل تنظیم آنتروپی هستند که u_{ci} و v_{cj} را جداگانه ترکیب می‌کنند. T_u و T_v پارامترهای وزنی هستند که درجه فازی را در خوشه‌های نهایی کنترل می‌کنند. معادلات به‌روزرسانی برای عضویت سند و ویژگی که d_{cij} فاصله بین نقطه داده ویژگی و مرکز خوشه ویژگی است، به صورت زیر بدست می‌آید

$$u_{ci} = \frac{\exp\left(-\frac{\sum_{j=1}^{k} v_{cj}d_{cij}}{T_u}\right)}{\sum_{c=1}^{C}\exp\left(-\frac{\sum_{j=1}^{k} v_{cj}d_{cij}}{T_u}\right)} \quad \text{و} \quad v_{cj} = \frac{\exp\left(-\frac{\sum_{i=1}^{n} u_{ci}d_{cij}}{T_v}\right)}{\sum_{j=1}^{k}\exp\left(-\frac{\sum_{i=1}^{n} u_{ci}d_{cij}}{T_v}\right)}. \tag{۲.۳}$$

[5] Information-based Fuzzy Co-Clustering

برای محاسبه اطلاعات فاصله مجموعه دادگان از رابطه زیر استفاده می‌کنیم

$$d_{cij} = \frac{1}{n} x_{ij} log(\frac{x_{ij}}{t_{cij}}) + \frac{|c|}{n} p_{cj} log(\frac{p_{cj}}{t_{cij}}), \tag{۳.۳}$$

که $t_{cij} = \frac{(x_{ij}+|c|*p_{cj})}{(1+|c|)}$، $|c|$ تعداد اسناد در cامین خوشه است. تعریف مقدار $|c|$ در خوشه‌بندی فازی نسبت به خوشه‌بندی سخت کمی پیچیده‌تر است زیرا باید عملیات فازی‌سازی روی ماتریس عضویت انجام شود. پس از فازی‌سازی به راحتی می‌توان مقدار $|c|$ را در خوشه‌بندی سخت بدست آورد.

در الگوریتم $ibFCC$ بدست آوردن p_{cj} به طور صریح دشوار است حتی اگر مقدار p_{cj} از نظر تئوری به صورت u_{ci} و v_{cj} محاسبه شود ممکن است فرایند محاسباتی بالایی داشته باشد. بنابراین یک رویکرد جایگزین انتخاب می‌شود که از روش میانگین وزنی استفاده می‌کند. در خوشه‌بندی فازی مرکز یک خوشه میانگین تمام نقاط است که براساس درجه تعلق آن‌ها به خوشه وزن می‌شود. درنتیجه معادله به‌روزرسانی نرمال شده p_{cj} به صورت زیر بدست می آید

$$p_{cj} = \frac{\sum_{i=1}^{n} u_{ci} x_{ij}}{\sum_{i=1}^{Nn} u_{ci}}. \tag{۴.۳}$$

۱.۳ مراحل اجرای الگوریتم $ibFCC$

مرحله	شرح عملیات
۱	مقداردهی اولیه به پارامترها: تعداد خوشه‌ها C، بیشینه تکرارها $\tau_{\max}$، آستانه همگرایی ε، ضرایب تنظیم آنتروپی T_u و T_v.
۲	تنظیم $\tau = 1$.
۳	مقداردهی اولیه به درجات عضویت u_{ci} و v_{cj} به‌صورت تصادفی.
۴	شروع تکرارها
۵	محاسبه مراکز خوشه ویژگی‌ها با استفاده از رابطه (۴.۳).
۶	محاسبه فاصله اطلاعاتی بین اسناد و ویژگی‌ها با استفاده از رابطه (۳.۳).
۷	به‌روزرسانی درجات عضویت با استفاده از رابطه (۲.۳).
۸	افزایش شمارنده تکرار: $\tau \leftarrow \tau + 1$.
۹	بررسی شرط توقف: $\max \lvert u_{ci}^{(\tau)} - u_{ci}^{(\tau-1)} \rvert \leq \varepsilon$ یا $\tau = \tau_{\max}$.

یکی از تفاوت‌های کلیدی بین الگوریتم $ibFCC$ و الگوریتم‌های فازی دوطرفه کلاسیک، ساختار تابع هدف آن‌هاست. در روش‌های سنتی مانند فازی k–میانگین دوطرفه، تابع هدف بر اساس فاصله اقلیدسی و ترکیب درجه‌های عضویت سطرها و ستون‌ها تعریف می‌شود و تمرکز آن صرفاً بر کمینه‌سازی فاصله بین داده‌ها و مراکز خوشه است. اما این رویکرد در مواجهه با داده‌های پیچیده، نویزی یا دارای هم‌پوشانی بالا، کارایی محدودی دارد. در مقابل، تابع هدف الگوریتم $ibFCC$ با بهره‌گیری از یک تابع فاصله اطلاعاتی (اطلاعات متقابل بین ویژگی و سند) و نیز دو عبارت آنتروپی مجزا برای سطرها و ستون‌ها، رویکردی غنی‌تر و منعطف‌تر ارائه می‌دهد. پارامترهای کنترلی T_u و T_v امکان تنظیم میزان پراکندگی عضویت‌ها را فراهم کرده و از همگرایی سریع و نامناسب جلوگیری می‌کنند. این طراحی باعث می‌شود الگوریتم $ibFCC$ بتواند ساختارهای پنهان و الگوهای معنادارتر را در داده‌ها شناسایی کرده و خوشه‌بندی دقیق‌تری ارائه دهد. به‌طور خلاصه، تفاوت‌های موجود در تابع هدف موجب افزایش تطبیق‌پذیری، دقت، و تفسیرپذیری خروجی الگوریتم $ibFCC$ نسبت به روش‌های سنتی شده‌اند، که در بخش بعدی از طریق ارزیابی عددی در داده‌های واقعی به آن پرداخته می‌شود.

۴ نتایج عددی

برای ارزیابی عملکرد الگوریتم ibFCC، آزمایش‌هایی روی پنج مجموعه داده واقعی در حوزه زیست‌پزشکی انجام شد. ویژگی‌های این مجموعه داده‌ها در جدول زیر آمده است

جدول ۱: جزئیات مجموعه داده‌های مورد بررسی

نام مجموعه داده	تعداد نمونه	تعداد ویژگی	تعداد دسته‌بندی‌ها	توضیحات
Ohsumed	۲۰۰۰۰ (سنداز ۵۰۲۱۶)	۵۰۰	۱۵	۲۳ دسته بیماری‌های موضوعی پزشکی (MeSH).
سرطان ریه (LC)	۲۷	۵۶	۳	برای شناخت انواع پاتولوژیک سرطان ریه.
بافت سینه (BT)	۱۰۶	۹	۶	پیش‌بینی طبقه‌بندی با ادغام کلاس‌های فیبروآدنوم، ماستوپاتی و غده‌ای.
کاردیوتوکوگرافی (CTG)	۲۱۲۶	۲۱	۱۰	۲۱۲۶ کاردیوتوکوگرام جنین، ویژگی‌های تشخیصی اندازه‌گیری شده.
بیان پروتئین موش (MPE)	۱۰۷۶	۶۸	۸	۱۰۷۶ اندازه‌گیری در هر پروتئین، با ۸ دسته از موش‌ها.

هر ۵ الگوریتم به صورت تصادفی مقداردهی و ده بار اجرا شدند تا تاثیر بهینه‌سازی‌های موضعی کاهش یابد. از بین پنج الگوریتم، خوشه‌بندی c–میانگین (به کوتاهی FCM) یک الگوریتم خوشه‌بندی فازی استاندارد است و بقیه، الگوریتم‌های خوشه‌بندی دوطرفه فازی هستند. در ادامه، الگوریتم $ibFCC$ با چهار الگوریتم دیگر شامل

FCM، خوشه‌بندی فازی برای داده‌های چندمتغیره رسته‌ای(به کوتاهی $FCCM$) ، خوشه‌بندی دوطرفه استوار فازی(به کوتاهی $RFCC$) و خوشه‌بندی دوطرفه فازی برای تصاویر(به کوتاهی $FCCI$) مقایسه شد. معیارهای ارزیابی شامل آنتروپی، شاخص F و $Log(CS)$ بودند و نتایج نشان داد که الگوریتم $ibFCC$ در اکثر موارد عملکرد بهتری داشته و به‌ویژه در داده‌های BT و MPE

جدول ۲: مقایسه الگوریتم‌ها بر اساس معیارهای F، آنتروپی و $\log(CS)$

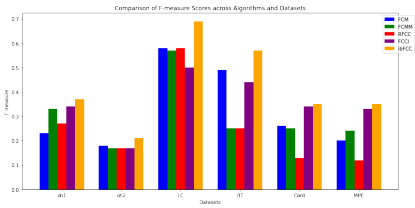
الگوریتم	نویسندگان	داده‌ها	F	آنتروپی	$\log(CS)$
FCM	بزدک (۱۹۷۳)	$Oh۱$	۰/۲۳	۰/۶۶	−۱/۸
		$Oh۲$	۰/۱۸	۰/۹۳	−۱/۶
		LC	۰/۵۸	۰/۵۶	−۱/۵
		BT	۰/۴۹	۰/۵۱	−۱/۹
		$Card$	۰/۲۶	۰/۴۲	−۳/۹
		MPE	۰/۳۵	۰/۲۲	−۵/۱
$FCCM$	او،سی.اچ و همکاران (۲۰۰۱)	$Oh۱$	۰/۳۳	۰/۶۷	−۱/۴
		$Oh۲$	۰/۲۷	۰/۹۶	−۱/۳
		LC	۰/۵۵	۰/۶۶	−۱/۵
		BT	۰/۴۶	۰/۶۰	−۱/۸
		$Card$	۰/۲۵	۰/۳۶	−۲/۱
		MPE	۰/۳۴	۰/۲۶	−۳/۹
$RFCC$	تیجی،دبلیو.سی. و همکاران (۲۰۰۷)	$Oh۱$	۰/۲۷	۰/۶۷	−۱/۷
		$Oh۲$	۰/۲۰	۰/۹۲	−۱/۴
		LC	۰/۵۸	۰/۹۶	−۱/۳
		BT	۰/۵۲	۰/۸۹	−۱/۲
		$Card$	۰/۲۵	۰/۸۵	−۱/۹
		MPE	۰/۳۴	۰/۹۵	−۳/۷
$FCCI$	هانماندلو،ام. و همکاران (۲۰۱۳)	$Oh۱$	۰/۳۴	۰/۶۸	−۱/۳
		$Oh۲$	۰/۲۸	۰/۹۰	−۱/۲
		LC	۰/۵۷	۰/۸۵	−۱/۲
		BT	۰/۴۹	۰/۸۰	−۱/۴
		$Card$	۰/۲۷	۰/۷۸	−۳/۵
		MPE	۰/۳۶	۰/۶۵	−۵/۸
$ibFCC$	لیو و همکاران (۲۰۱۷)	$Oh۱$	۰/۳۵	۰/۷۰	−۱/۳
		$Oh۲$	۰/۲۹	۰/۹۱	−۱/۲
		LC	۰/۵۷	۰/۸۵	−۱/۲
		BT	۰/۴۹	۰/۸۰	−۱/۴
		$Card$	۰/۳۵	۰/۲۲	−۶/۵
		MPE	۰/۳۵	۰/۳۳	−۵/۳

توانسته دقت خوشه‌بندی را به‌طور چشم‌گیری افزایش دهد. بعد از $ibFCC$ الگوریتم $FCCI$ عملکرد بهتری از FCM، $FCCM$ و $RFCC$ دارد.

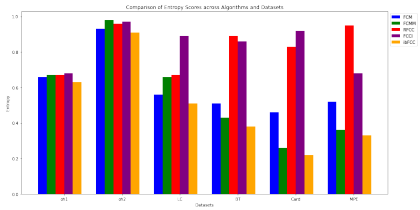
بحث و نتیجه‌گیری

در این مطالعه الگوریتمibFCC به عنوان یک روش جدید برای خوشه‌بندی فازی دوطرفه مورد مطالعه و بررسی قرار گرفت. این الگوریتم با در نظر گرفتن اطلاعات متقابل بین ویژگی‌ها و نمونه‌ها، نسبت به روش‌های موجود که عمدتاً بر فاصله اقلیدسی تمرکز دارند، برتری محسوسی دارد. علاوه بر این، ibFCC با استفاده از دو عبارت آنتروپی مجزا برای سطرها (نمونه‌ها) و ستون‌ها (ویژگی‌ها)، انعطاف‌پذیری بیشتری در مدل‌سازی ساختارهای پنهان داده‌ها ارائه می‌دهد. وجود پارامترهای کنترلی در این الگوریتم، امکان تنظیم میزان پراکندگی عضویت‌ها را فراهم کرده و به همگرایی سریع و جلوگیری از همگرایی نامناسب کمک کرده است.

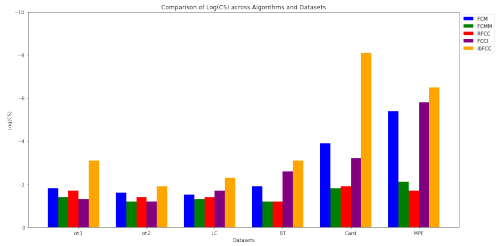
نتایج ارزیابی بر روی پنج مجموعه داده واقعی نشان داد که الگوریتم ibFCC در بیشتر موارد عملکرد بهتری داشته و دقت خوشه‌بندی را به‌طور چشمگیری افزایش داده است. این برتری عملکردی، ibFCC به دلیل توانایی آن در مدل‌سازی بهتر ارتباطات پیچیده میان



(b) معیار F



(a) آنتروپی



Log(CS) (c)

شکل ۱: مقایسه سه معیار F، آنتروپی و $Log(CS)$ برای الگوریتم‌های جدول نتایج روی داده‌ها

ویژگی‌ها و نمونه‌ها، و همچنین انعطاف‌پذیری بالای آن در مواجهه با داده‌های نادقیق و هم‌پوشان است.

در نهایت، با توجه به نتایج به‌دست‌آمده از ارزیابی‌های تجربی، می‌توان نتیجه گرفت که الگوریتم ibFCC به‌عنوان یک ابزار مؤثر و کارآمد برای تحلیل داده‌های پیچیده و دارای ساختار پنهان محسوب می‌شود. قابلیت‌های پیشرفته این الگوریتم در کشف ساختارهای نهفته و ارائه نتایج دقیق‌تر و تفسیرپذیرتر، آن را به گزینه‌ای مناسب برای کاربردهای مختلف در علوم داده، به‌ویژه در حوزه‌هایی مانند زیست‌پزشکی که با داده‌های حجیم و پیچیده سروکار دارند، تبدیل می‌کند.

مراجع

طبیبی،م. و زارعی،ر. (۱۴۰۱)، رویکرد بیزی در خوشه‌بندی تحت شرایط نادقیق، پایان نامه کارشناسی ارشد آمار ریاضی، دانشگاه گیلان.

Bezdek J. C. (1973). Fuzzy Mathematics in Pattern Classification, *John Wiley Sons, Hoboken*, New Jersey.

Ferraro M.B., Vichi M., Grzegorzewski P., Gagolewski M., Hryniewicz O. and Gil M. (2015), Fuzzy double clustering: a robust proposal. , *Advances in Intelligent systems and computing 315. Springer, Cham.* 225-232.

Ferraro M.B. (2024), Fuzzy k-means:history and applications, *Elsevier; 2024: 110-123.*

Hanmandlu M., Verma O.P., Susan S. and Madasu V.K. (2013), Color segmentation by fuzzy co-clustering of chrominance color features, *Neurocomputing 120: 235–249.*

Liu Y., Wu. S., Liu Z. and Chao H. (2017), A fuzzy co-clustering algorithm for biomedical data., *PLoS ONE 12(4): e0176536*

Oh CH., Honda K., and Ichihashi H. (2001), Fuzzy clustering for categorical multivariate data, *2154–2159 vol. 2154.*

Tjhi W.C. and Chen L., (2007), Robust fuzzy Co-clustering algorithm, 1-5.

Vichi M., Borra S., Rocci R., Vichi M. and Schader M., (2001), Doubl kmeans clustering for simultaneous classification of objects and variables, *Advances in classification and data analysis. studies in classification, data analysis, and knowledge organization. Springer, Berlin, Heidelberg*

Fuzzy Co-Clustering using Information Distance

Zahra Ahmadi[1], **Reza Zarei**[2]

[1]Department of Statistics, University of Guilan

[2]Department of Statistics, University of Guilan

Abstract: With the rapid growth of data in various scientific domains, the need for effective methods to analyze and extract hidden patterns has become increasingly important. Among clustering techniques, fuzzy clustering has gained significant attention due to its high flexibility in handling ambiguous and overlapping data. This paper introduces and investigates an advanced algorithm called interval-based bidirectional fuzzy co-clustering (ibFCC) designed for imprecise environments. Using real-world biomedical datasets, the performance of this algorithm is compared with several related methods. The results indicate its relative superiority in terms of accuracy, interpretability, and clustering evaluation metrics.

Keywords: Fuzzy clustering, Co-clustering, Membership functions, Objective function

Mathematics Subject Classification (2020): 62H30,۰۳B۵۲,۶۸T۰۹.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

یک روش جدید تنک‌سازی فرایندهای نقطه‌ای مبتنی بر مجموعه‌های تصادفی بولی

رضا اسدی [۱] ، مجتبی خزائی[۱] ، مجتبی گنجعلی[۱]

[۱]گروه آمار، دانشگاه شهید بهشتی

چکیده: فرایندهای نقطه‌ای فضایی از مدل‌های آماری برای تحلیل الگوهای نقطه‌ای اند. یکی از روش‌های ساخت فرایندهای نقطه‌ای جدید تنک‌سازی فرایندهای نقطه‌ای موجود است. در این مقاله یک الگوریتم تنک‌سازی جدید مبتنی بر یک مجموعه تصادفی بولی ارائه می‌شود. این الگوریتم را می‌توان تعمیمی از روش تنک‌سازی ارائه شده توسط اشتویان در معرفی فرایندهای نقطه‌ای وابریده دانست. بعضی ویژگی‌های فرایند نقطه‌ای تنک‌شده و فرایند نقاط حذف شده حاصل از اعمال این الگوریتم تنک‌سازی بر یک فرایند نقطه‌ای پواسون همگن بررسی شده و بعضی از آماره‌های خلاصه آن‌ها محاسبه شده است. از روش درستنمایی پالم برای برآورد پارامترها استفاده شده است. در نهایت، مدل‌های فوق برای مدل‌بندی دو مجموعه داده واقعی استفاده می‌شود.

واژه‌های کلیدی: فرایند نقطه‌ای پواسون؛ فرایندهای کاکس؛ مجموعه تصادفی بولی؛ تنک‌سازی وابسته؛ درستنمایی پالم.

کد موضوع‌بندی ریاضی (۲۰۲۰): xxAxx، xxBxx، xxCxx.

۱ مقدمه

فرایندهای نقطه‌ای فضایی از مدل‌های آماری برای مدل‌بندی الگوهای نقطه‌ای اند. به عنوان مثال، از فرایندهای نقطه‌ای فضایی برای مدل‌بندی محل انواع درختان جنگلی، محل وقوع زمین لرزه‌ها، لانه پرندگان و ... می‌توان استفاده کرد. رفتار الگوهای نقطه‌ای به سه دسته رفتار خوشه‌ای، رفتار تصادفی محض و رفتار منظم یا دفعی تقسیم‌بندی می‌شود. متناسب با رفتار هر الگوی نقطه‌ای دسته‌ای از مدل‌ها برای مدل‌بندی الگوهای نقطه‌ای ارائه شده است. روش‌های زیادی از جمله اعمال تبدیلات و یا اصلاح فرایندهای نقطه‌ای برای ساخت فرایند نقطه‌ای جدید وجود دارند. چهار رویکرد کلی برای این کار عبارتند از: نگاشت، برهم نهی، خوشه‌بندی و تنک‌سازی.

در رویکرد تنک‌سازی نقاط فرایند اولیه براساس یک الگوریتم حذف می‌شوند. در صورتی که احتمال حذف نقاط فرایند اولیه از هم مستقل باشند، تنک‌سازی مستقل و در صورتی که احتمال حذف نقاط فرایند اولیه به هم وابسته باشند، تنک‌سازی وابسته نامده می‌شوند.

[۱] سخنران، rezaasadi220@gmail.com

در این مقاله ابتدا یک الگوریتم تنک‌سازی معرفی می‌شود که در آن احتمال‌های حذف براساس تحققی از یک مدل بولی مشخص می‌شوند. سپس فرایندهای تنک‌شده و نقاط حذف شده حاصل از اعمال این الگوریتم روی یک فرایند اولیه پواسون همگن بررسی می‌شود. الگوریتم تنک‌سازی ارائه شده در این مقاله را می‌توان تعمیمی از الگوریتم تنک‌سازی ارائه شده توسط اشتویان (۱۹۷۹) و مولر و لاوانسیر (۲۰۱۶) دانست با این تفاوت که در اینجا احتمال حذف درون تحقق مجموعه تصادفی بولی می‌تواند هر مقداری بین صفر و یک باشد در حالی که اشتویان (۱۹۷۹) و مولر و لاوانسیر (۲۰۱۶) تنها احتمال حذف نقاط خارج از مجموعه تصادفی بولی را برابر یک در نظر گرفته‌اند. یک مجموعه تصادفی بولی به صورت زیر تعریف می‌شود.

تعریف ۱.۱. *(اشتویان ، ۱۹۷۹)* یک مجموعه تصادفی بولی از قرار دادن تحققهای یک مجموعه تصادفی بسته، معروف به دانه، روی نقاط یک فرایند نقطه‌ای پواسون، معروف به ریشه، حاصل می‌شود. در صورتی که Z فرایند پواسون مانا روی $\mathbb{R}^2$ با شدت θ و مستقل از M_s ها؛ یعنی مجموعه‌های تصادفی بسته مستقل و هم توزیع با M_o به‌طوری‌که $E\{\sup ||x||, x \in M_o\} < \infty$ باشد، آنگاه مجموعه تصادفی بولی حاصل از دانه بسته تصادفی M_o و فرایند نقطه‌ای Z برابر است با: $\Xi = \bigcup_{s\in Z} (s + M_s)$

الگوریتم‌های تنک‌سازی متنوعی که فرایند تنک‌شده حاصل از اعمال آن‌ها بر فرایند نقطه‌ای فضایی پواسون مورد بررسی قرار گرفته است معرفی شده است. به عنوان مثال فرایندهای هسته سخت مترن، در فرایندهای هسته سخت نقاط فرایند اولیه براساس تابعی از فاصله نقاط از یکدیگر حذف می‌شوند لکن در الگوریتم تنک‌سازی معرفی شده در این مقاله نقاط فرایند اولیه براساس تابعی از فرایند Z که مستقل از فرایند اولیه است حذف می‌شوند.

۲ معرفی الگوریتم تنک‌سازی پیشنهادی و فرایندهای مربوطه

فرایند نقطه‌ای اولیه Y و فرایند نقطه‌ای Z به عنوان فرایندی که بر اساس آن Y تنک می‌شود را در نظر بگیرید. در صورتی که فرایندهای Y و Z بر $\mathbb{R}^2$ تعریف شده و از هم مستقل باشند. حول هر نقطه از فرایند نقطه‌ای فضایی Z دایره‌ای با شعاع r را در نظر گرفته و احتمال حذف نقطه‌ای از فرایند اولیه درون هر دایره را مستقل از سایر دوایر برابر با q در نظر می‌گیریم. به این ترتیب یک الگوریتم تنک‌سازی حاصل می‌شود که در آن:

۱- در ناحیه خارج از اجتماع دوایر احتمال حذف صفر است.

۲- در نواحی درون یک دایره احتمال حذف $q = 1 - p$ است.

۳- در نواحی مشترک بین k دایره احتمال حذف $1 - p^k$ است.

در صورتی که Z فرایند پواسون همگن و همسانگرد با تابع شدت θ باشد و $\Xi = \bigcup_{s\in Z} b(s, r)$ تعریف شود، یک مجموعه تصادفی بولی با دانه‌های دایره‌ای شکل به شعاع r است. الگوریتم تنک‌سازی فوق را می‌توان به صورت یک الگوریتم تنک‌سازی روی تحققهای این مجموعه بولی به صورت زیر تعریف کرد:

۱- در نواحی که خارج از مجموعه تصادفی بولی Ξ است نقاط Y حذف نمی‌شوند.

۲- در نواحی که محل تحقق فقط یک دانه از مجموعه تصادفی بولی Ξ است نقاط Y با احتمال $q = 1 - p$ حذف می‌شوند.

۳- در نواحی که محل اشتراک k دانه از مجموعه تصادفی بولی Ξ است نقاط فرایند اولیه Y با احتمال $q = 1 - p^k$ حذف می‌شوند.

در قضیه زیر توزیع حاشیه‌ای و توام احتمال بقای نقاط محاسبه می‌شود. اثبات قضیه‌های ارائه شده در این مقاله در ضمایم ارائه شده است.

قضیه ۱.۲. در الگوریتم تنک‌سازی پیشنهادی اگر $\Pi(\zeta)$ احتمال بقا در $\zeta \in \mathbb{R}^2$ باشد آن گاه:

(i) اگر $A_\zeta = b(\zeta, r)$ باشد برای هر $\zeta \in \mathbb{R}^2$:

$$P[\Pi(\zeta) = p^k] = P[N(A_\zeta) = k] = \frac{e^{-\theta\pi r^2}(\theta\pi r^2)^k}{k!}, \quad k = 0, 1, 2, \ldots, \tag{۱.۲}$$

(ii) توزیع توام احتمال بقا برای هر $\zeta_1, \zeta_2 \in \mathbb{R}^2$ برابر است با:

$$P[\Pi(\zeta_1) = p^n, \Pi(\zeta_2) = p^m] = \begin{cases} P[\Pi(\zeta_1] = p^n]P[\Pi(\zeta_2) = p^m], & ||\zeta_1 - \zeta_2|| > 2r, \\ C[\Pi(\zeta_1, \zeta_2)], & ||\zeta_1 - \zeta_2|| \leqslant 2r, \end{cases} \tag{۲.۲}$$

که در آن

$$\begin{aligned} C[\Pi(\zeta_1, \zeta_2)] =& e^{-\theta(2\pi r^2 - A(t,r))}[\theta(\pi r^2 - A(t,r))]^{m+n} \\ & \times \sum_{k=0}^{\min(m,n)} \frac{1}{k!(n-k)!(m-k)!}\left[\frac{A(t,r)}{\theta(\pi r^2 - A(t,r))^2}\right]^k, \end{aligned} \tag{۳.۲}$$

و $m, n \in \mathbb{N} \cup \{0\}, r > 0$

$$A(t,r) = 2r^2 Arc\cos(\frac{t}{2r}) - \frac{t}{2}[4r^2 - t^2]^{\frac{1}{2}}, \tag{۴.۲}$$

مساحت ناحیه مشترک بین دو دایره به شعاع r با فاصله t ، ($t < 2r$) است.

در صورتی که فرایند اولیه، Y ، یک فرایند نقطه‌ای پواسون با شدت δ باشد از اعمال الگوریتم فوق بر Y دو فرایند تنک‌شده‌ی Y، X، و نقاط حذف شده از Y، $Y - X$، فرایندهای نقطه‌ای تعریف می‌کنند که به پارامترهای (δ, θ, r, p) بستگی دارد. در ادامه ویژگی‌ها و رفتار این فرایندها بررسی می‌شود.

۳ آماره‌های خلاصه

آماره‌های خلاصه به عنوان ابزاری برای تحلیل اکتشافی داده‌ها، آزمون و بررسی اعتبار مدل توسط آماردان‌ها مورد استفاده قرار می‌گیرند. در تحلیل فرایندهای نقطه‌ای فضایی ویژگی‌های مرتبه اول و دوم متغیر شمارشی $N(B); B \subseteq \mathbb{R}^2$ اطلاعات سودمندی در مورد رفتار فرایند نقطه‌ای به تحلیل‌گر می‌دهند. تابع شدت مرتبه اول و دوم این ویژگی‌ها را تبیین می‌کنند. اگر $\mu(B) = E[N(B)]; B \subseteq \mathbb{R}^2$ ، $\mu(.)$ اندازه شدت نامیده می‌شود. در صورتی که بتوان نوشت $\mu(B) = \int_B \lambda(\zeta)d\zeta; B \subseteq \mathbb{R}^2$ نوشت که در آن $\lambda(.)$ یک تابع نامنفی است، $\lambda(.)$ تابع شدت نامیده می‌شود. اگر $\alpha^{(2)}(C) = E[\sum_{\zeta,\eta \in X}^{\neq} I[(\zeta,\eta) \in C]]; C \subseteq \mathbb{R}^2 \times \mathbb{R}^2$ ، $\alpha^{(2)}(.)$ اندازه شدت فاکتوریلی مرتبه دوم نامیده می‌شود. بعلاوه اگر $\alpha^{(2)}(C) = \int\int I[(\zeta,\eta) \in C]\lambda^{(2)}(\zeta,\eta)d\zeta d\eta$ که در آن $\lambda^{(2)}(.,.)$ یک تابع

نامنفی است به $\lambda^{(۲)}(.,.)$ تابع شدت مرتبه دوم گفته می‌شود. در این‌جا با توجه به این‌که فرایند اولیه‌ی Y توسط یک میدان تصادفی تنک می‌شود لذا هر دو فرایند X و $Y-X$ فرایندهای کاکس هستند (مولر و واگه‌پیترسن ، ۲۰۰۳).

قضیه ۱.۳. برای فرایند تنک‌شده‌ی X، تابع شدت $\lambda(.)$ و تابع شدت مرتبه دوم $\lambda^{(۲)}(.,.)$ برابر است با:

$$(i)\ \lambda(u) = \delta e^{-\theta(۱-p)\pi r^۲}.$$

$$(ii)\ \lambda^{(۲)}(u,v) = \begin{cases} \delta^۲ e^{-۲\pi r^۲\theta(۱-p)+\theta A(t,r)(۱-p)^۲}, & \|t\| \leq ۲r, \\ \delta^۲ e^{-۲(۱-p)\theta\pi r^۲}, & \|t\| > ۲r, \end{cases}$$

که در آن $t = ||u-v||$ است.

تابع همبستگی زوجی یکی دیگر از آماره‌هایی است که ویژگی‌های مرتبه دوم فرایندهای نقطه‌ای را بررسی می‌کند. تابع همبستگی زوجی به صورت $g(\zeta,\eta) = \frac{\lambda^{(۲)}(\zeta,\eta)}{\lambda(\zeta)\lambda(\eta)}$ تعریف می‌شود که در آن برای $a \geqslant ۰$ قرار می‌دهیم $\frac{a}{۰} = ۰$. برای فرایندهای مانا می‌توان نوشت $g(\zeta,\eta) = g(\zeta-\eta)$ و برای فرایندهای مانا و همسانگرد $g(\zeta,\eta) = g(t)$ که در آن $t = ||\zeta-\eta||$. در حالت کلی اگر $t \to \infty$ آن‌گاه $g(t) \to ۱$. ولی برای t های کوچک، $g(t) < ۱$ نشان‌دهنده فرایند با رفتار دفعی، $g(t) = ۱$ نشان‌دهنده فرایند با رفتار تصادفی محض و $g(t) > ۱$ نشان‌دهنده فرایند با رفتار خوشه‌ای است.

قضیه ۲.۳. تابع همبستگی زوجی برای فرایندهای X و $Y-X$ برابر است با:

$$(i)\ g_X(t) = \begin{cases} e^{\theta A(t,r)(۱-p)^۲}, & t \leq ۲r, \\ ۱, & t > ۲r, \end{cases}$$

$$(ii)\ g_{Y-X}(t) = ۱ + \frac{a^۲}{(۱-a)^۲}(g_X(t)-۱),$$

که در آن $t = ||u-v||, a = E[\Pi(o)]$.

(i) نشان می‌دهد که $g_X(t) \geqslant ۱$. و از (ii) نتیجه می‌شود که $g_{Y-X}(t) \geqslant ۱$. نتایج فوق بیانگر این است که هر دو فرایند X و $Y-X$ دارای رفتار خوشه‌ای هستند. تابع شدت پالم یکی دیگر از آماره‌هایی است که در استنباط آماری فرایندهای نقطه‌ای کاربرد زیادی دارد. در صورتی که x نقطه‌ای دلخواه در $\mathbb{R}^۲$ با فاصله t از مبدا، o، باشد. آن‌گاه نرخ تحقق در x به شرط آن‌که o یک نقطه از فرایند باشد برای مجموعه لبگ با اندازه بسیار کوچک dx برابر است با: $\lambda_o(x)dx = P[N(dx) = ۱|N\{o\} = ۱]$ به $\lambda_o(x)$ شدت پالم در نقطه x گفته می‌شود. در صورتی‌که فرایند مانا و همسانگرد باشد تابع شدت پالم صرفا به t وابسته است و می‌توان آن را به صورت $\lambda_o(t)$ نوشت. همچنین برای فرایندهای مانا و همسانگرد $\lambda_o(t) = \lambda g(t)$ (تاناکا و همکاران ، ۲۰۰۸).

قضیه ۳.۳. تابع شدت پالم برای فرایندهای X و $Y-X$ برابر است با:

$$(i)\ \lambda_{o,X}(t) = \begin{cases} \delta e^{-\theta(۱-p)[\pi r^۲ - A(t,r)(۱-p)]}, & t \leq ۲r, \\ \delta e^{-\theta(۱-p)\pi r^۲}, & t > ۲r, \end{cases}$$

$$(ii)\ \lambda_{o,Y-X}(t) = \delta[۱ - e^{-\theta q\pi r^۲}]\{۱ + \frac{a^۲}{(۱-a)^۲}(g_X(t)-۱)\}.$$

۴ برآورد پارامترها و ارزیابی مدل

روش‌های زیادی برای برآورد پارامترهای فرایندهای نقطه‌ای وجود دارد، از جمله حداقل تضادها، ماکسیمم درستنمایی، شبه درستنمایی و درستنمایی پالم. در این مقاله برای برآورد پارامترهای فرایندهای X و $Y-X$ از روش درستنمایی پالم استفاده می‌شود. از روش‌های متداول بررسی نیکویی برازش در فرایندهای نقطه‌ای بررسی پوشش‌ها است. این روش مبتنی بر محاسبه و رسم نمودار چندک‌های آماره خلاصه به ازای t های متفاوت است. از روش بوت استرپ برای محاسبه چندک‌ها می‌توان استفاده کرد. این روش مبتنی‌بر این است که اگر مدل برازش داده شده صحیح باشد بروردگر ناپارامتریک آماره‌های خلاصه الگوی مشاهده شده به ازای t های متفاوت در اغلب اوقات رفتاری مشابه مدل مورد نظر (با پارامترهای برآورد شده) دارد.

۱.۴ درستنمایی پالم

درستنمایی پالم یک روش تقریبی برآورد پارامترهای فرایندهای نقطه‌ای است. این روش مبتنی بر فرایند تفاضل است. برای فرایند نقطه‌ای P که بر پنجره محدب W مشاهده شده است، فرایند تفاضل به صورت $D_{P,W}=\{x-y; x,y\in P_W,\ \ s.t.\ \ x\neq y\}$ تعریف می‌شود. برای فرایند مانا و همسانگرد P، فرایند تفاضل یک فرایند همسانگرد با تابع شدت $N(W)\lambda_o(t)$ است که در آن $\lambda_o(t)$ تابع شدت پالم فرایند P است. همچنین در صورتی که فرض شود فرایند P به خوبی توسط یک فرایند پواسون تقریب زده می‌شود آن‌گاه تابع درستنمایی فرایند برابر است با:

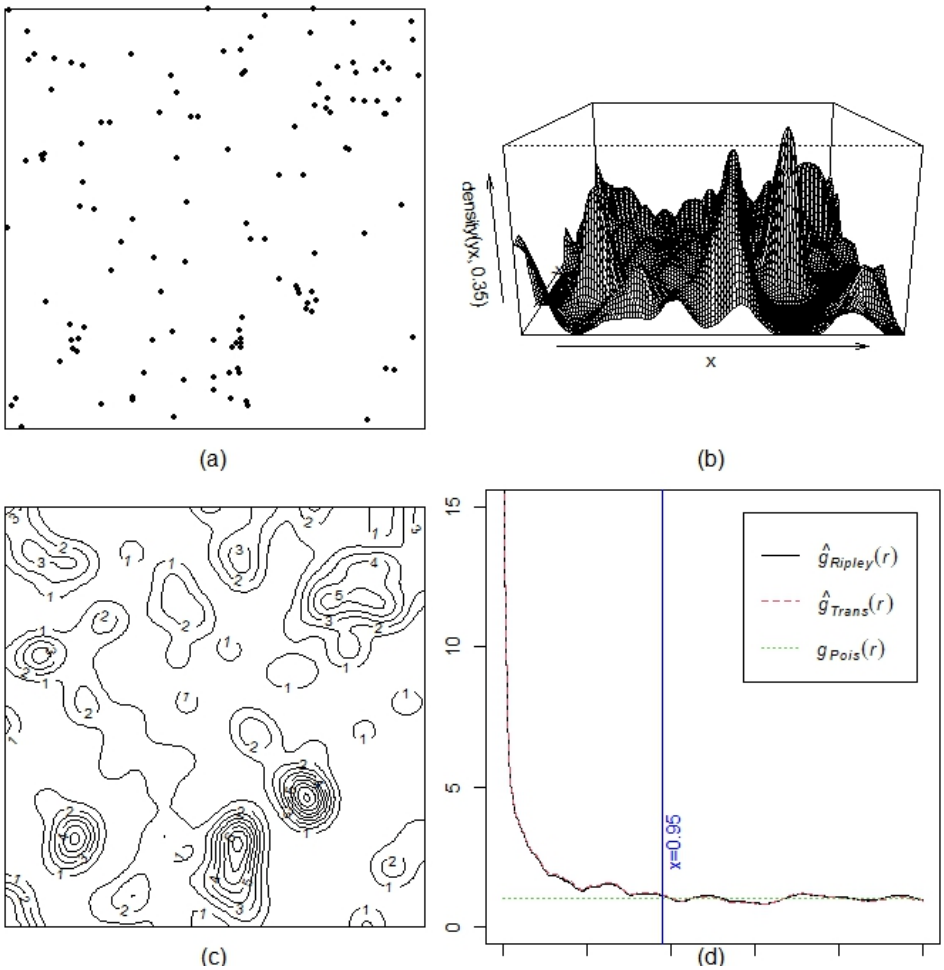
$$\log L(\delta,\theta,p,r,t)=\sum_{\{i,j;i\neq j;t_{ij}<T\}}\log\{N(W)\lambda_o(t_{ij})\}-N(W)\int_{\circ}^{T}\lambda_o(t)۲\pi t dt. \qquad (۱.۴)$$

که در آن t_{ij} فاصله بین نقاط p_i و p_j است و T آستانه‌ای است که فرض می‌شود نقاط با فاصله بیشتر از آن از هم مستقل هستند. به $(\hat{\delta},\hat{\theta},\hat{r},\hat{p})$ که عبارت فوق را ماکسیمم کند برآوردگر ماکسیمم درستنمایی پالم گفته می‌شود. برای آشنایی بیشتر به تاناکا و همکاران (۲۰۰۸) به مراجعه شود.

۵ کاربرد

در این بخش دو مجموعه داده واقعی بررسی می‌شود و فرایندهای X و $Y-X$ به آن‌ها برازش داده می‌شود.

مثال ۱.۵. الگوی نقطه‌ای نهال‌های کاج که در بسته $spatstat$ نرم افزار $\mathbb{R}$ تحت عنوان $finpines$ موجود است، محل قرار گرفتن ۱۲۶ نهال کاج به همراه قطر و ارتفاع آن‌ها در جنگلی در فنلاند بر پنجره $[-۵,۵]\times[-۸,۲]$ را ثبت کرده است. در این‌جا صرفا محل قرار گرفتن نهال‌ها بررسی و مدل‌بندی می‌شود. در شکل ۱ نمودار (a) الگوی نقطه‌ای، (b) نمودار ۳ بعدی تابع شدت، (c) منحنی‌های تراز تابع شدت، (d) نمودار تابع همبستگی زوجی، نهال‌های کاج را نشان می‌دهند. نمودار تابع همبستگی زوجی نشان‌دهنده رفتار خوشه‌ای الگوی نقطه‌ای است همچنین منحنی‌های تراز ونمودار ۳ بعدی تابع شدت نشان‌دهنده این است که الگوی نقطه‌ای تمایل به تجمع در نواحی دایره‌ای شکل دارند. لذا برازش فرایند $Y-X$ به الگوی نقطه‌ای فوق پیشنهاد می‌شود. برای برازش فرایند $Y-X$ به الگوی نقطه‌ای فوق براساس روش درستنمایی پالم پارامترهای مدل فوق برآورد می‌گردد. با توجه به این‌که برآورد ناپارامتریک تابع همبستگی زوجی برای $t\geqslant ۰/۹۵$ کوچکتر از ۱.۱ است. (شکل ۱ مشاهده شود.) لذا T برابر با ۰/۹۵ در نظر گرفته می‌شود. همچنین



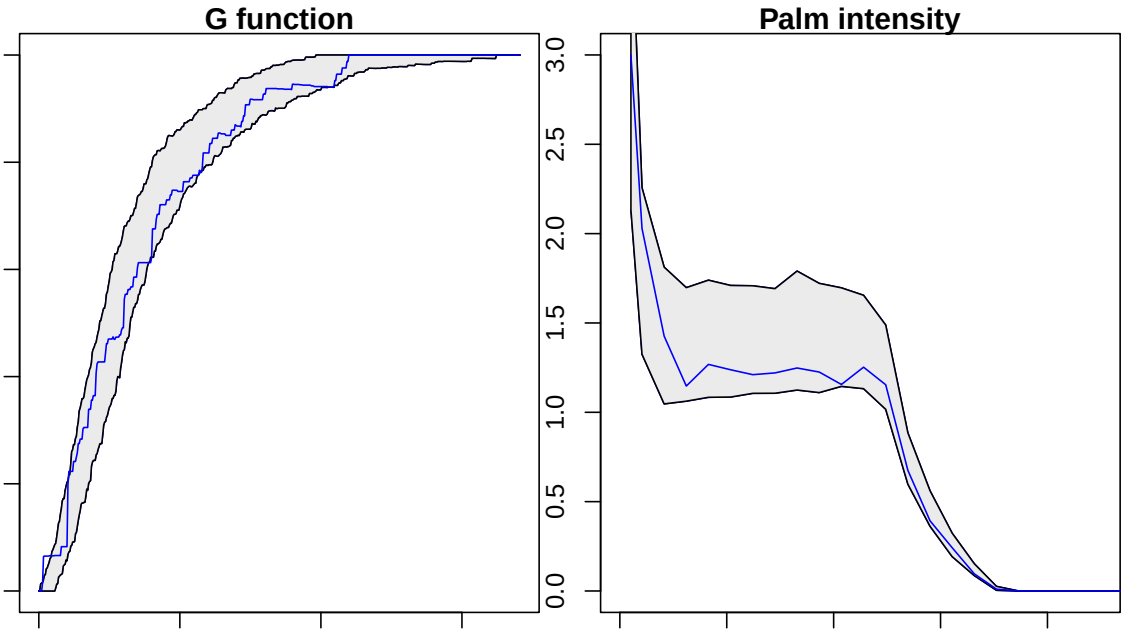
شکل ۱: (a) الگوی نقطه‌ای، (b) نمودار ۳ بعدی تابع شدت، (c) منحنی‌های تراز تابع شدت، (d) تابع همبستگی زوجی، نهال‌های کاج را نشان می‌دهد.

با جایگذاری $\lambda_{o,Y-X}(t)$ در (۱.۴) تابع درستنمایی پالم حاصل می‌شود. برآوردگر ماکسیمم درستنمایی پالم فرایند فوق در جدول ۱ ارائه شده است.

پارامتر	δ	θ	p	r
برآورد	۶۸۳/۵	۷۱۳/۱	۰۴۶/۰	۲۳۸۸/۰

جدول ۱: برآوردهای پارامترهای حاصل از برازش فرایند $Y - X$ به الگوی نهال‌های کاج

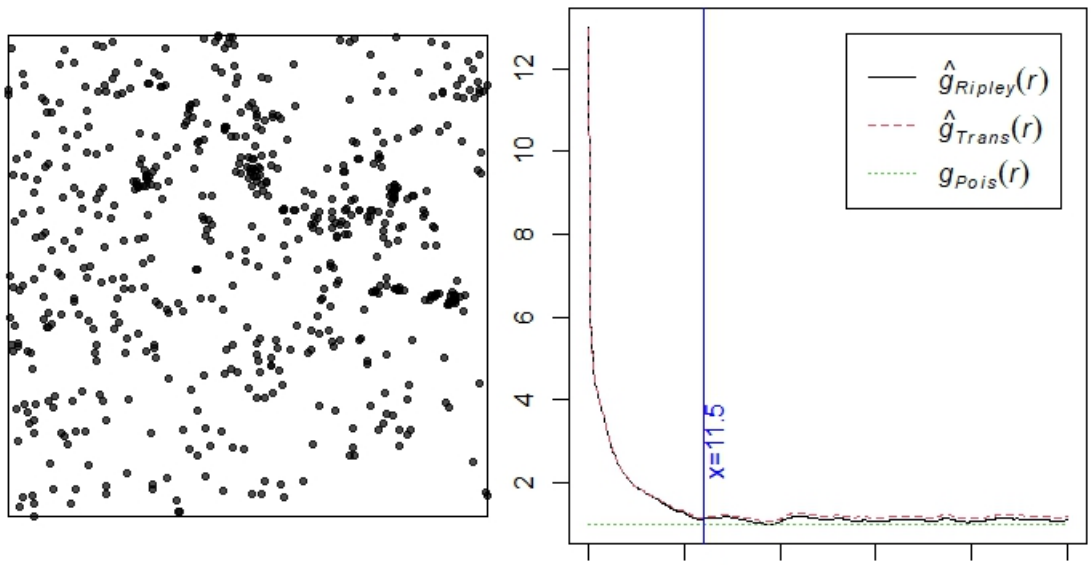
برای بررسی نیکویی برازش مدل برازش داده شده پوشش آماره‌های خلاصه بررسی می‌شود. از آن‌جایی که همه پارامترهای مدل در ساختار تابع شدت پالم وجود دارد پوشش تابع شدت پالم بررسی می‌شود همچنین پوشش تابع G که یکی از آماره‌های خلاصه مبتنی بر فاصله بین نقاط است نیز بررسی می‌شود. در شکل ۲ نمودار پوشش تابع شدت پالم رسم شده است که موید این است که فرایند برازش داده شده به خوبی الگوی نقطه‌ی فوق را مدل‌بندی می‌نماید. همچنین نمودار پوشش تابع G نیز رسم شده است. از آن‌جایی که روش



شکل ۲: نمودار پوشش تابع شدت پالم و تابع G برای مدل برازش داده شده

درستنمایی پالم یک روش تقریبی برآورد پارامترها است و این‌که، در اغلب نقاط تابع G مشاهده شده در فاصله اطمینان ۰/۹۵ قرار گرفته است لذا می‌توان نتیجه گرفت که مدل برازش داده شده با پارامترهای فوق برازش مناسبی به الگوی فوق دارد.

مثال ۲.۵. الگوی نقطه‌ای درختان کاج برگ بلند در ناحیه‌ای در جنوب جورجیا (ایالات متحده امریکا) بر پنجره $[0, 200] \times [0, 200]$ محل قرار گرفتن و قطر ۵۸۴ درخت کاج برگ بلند را نشان می‌دهد. در این‌جا صرفا محل قرار گرفتن درخت‌های فوق بررسی می‌شود. این داده‌ها در پکیج $spatstat$ نرم افزار $\mathbb{R}$ با عنوان $longleaf$ موجود است.



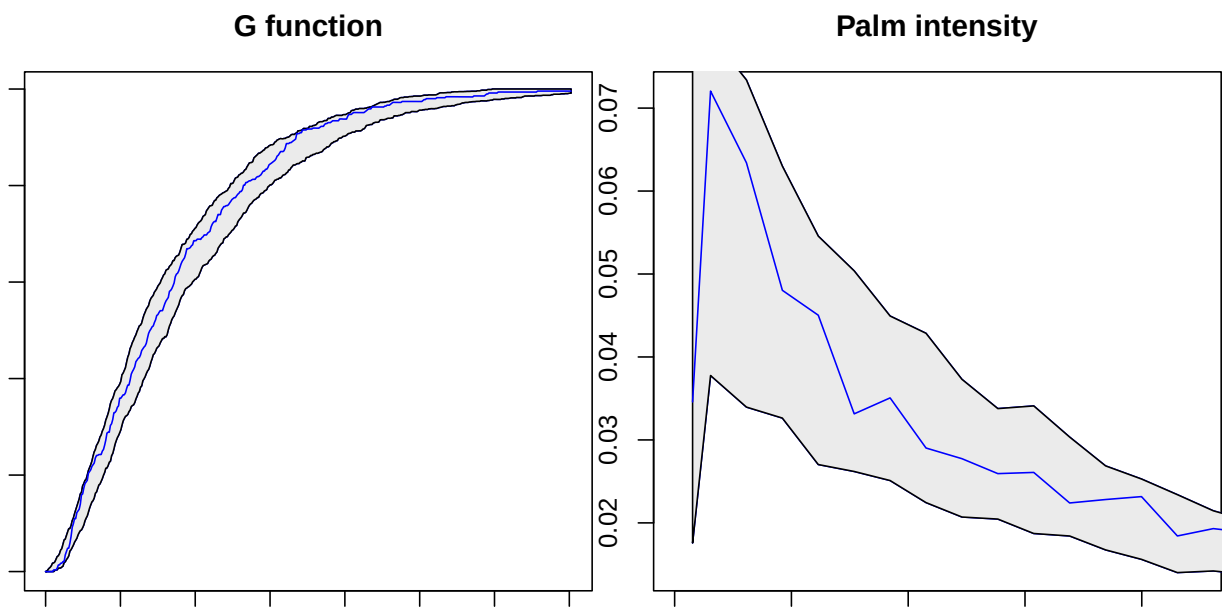
شکل ۳: نمودار الگوی نقطه‌ای محل قرارگرفتن درختان کاج برگ بلند و تابع همبستگی زوجی آن

در شکل ۳ نمودار محل قرار گرفتن درختان کاج و تابع همبستگی زوجی آن‌ها نشان داده شده. در شکل فوق به نظر می‌رسد بعضی نواحی دایره‌ای شکل تنک‌تر هستند. لذا برازش فرایند X به الگوی فوق پیشنهاد می‌شود. برای برازش برآورد پارامترهای مدل توسط روش درستنمایی پالم با توجه به رفتار تابع همبستگی زوجی در شکل ۳ و از آن‌جایی که برای $t \geqslant ۱۱/۵$ تابع همبستگی زوجی در اغلب نقاط کمتر از ۱/۱ است لذا T در رابطه (۱.۴) برابر با ۱۱/۵ در نظر می‌گیریم. با جایگذاری تابع شدت پالم فرایند X، $\lambda_{o,X}(t)$، از ۳.۳ در تابع درستنمایی پالم (۱.۴) برآورد درستنمایی پارامترهای فرایند که در جدول ۲ ارائه شده اند بدست می‌آید.

پارامتر	δ	θ	p	r
برآورد	۵۴۴/۰	۱۰۹/۰	۳۴۹/۰	۵۱۱/۵

جدول ۲: برآوردگر درستنمایی پالم حاصل از برازش فرایند X به الگوی درختان کاج برگ بلند

از طرفی در فرایند X با توجه به رفتار تابع همبستگی زوجی براساس رابطه $\hat{r}_T = \min_t\{g(۲t) < ۱ + \varepsilon \quad \& \quad g(۲t) > ۰\}$ پارامتر r را نیز می‌توان برآورد کرد. برای بهبود و تعدیل برآوردگر پارامتر r، آن را می‌توان به کمک رابطه $\hat{r}_{imp} = a\hat{r}_{Palm} + (۱-a)\hat{r}_T$ برآورد نمود. با در نظر گرفتن $\varepsilon = ۱/$ و $a = ۴/$ آن‌گاه $\hat{r}_{imp} = ۵/۶۵۴$.



شکل ۴: نمودار پوشش برای تابع شدت پالم و تابع G

برای بررسی نیکویی برازش مدل برازش داده شده، پوشش آماره‌های خلاصه بررسی می‌شود. برای این منظور پوشش تابع شدت پالم و تابع G بررسی می‌شود. در شکل ۴ نمودار پوشش برای تابع شدت پالم و تابع G رسم شده است. پوشش آماره‌های خلاصه نشان‌دهنده این است که دلیلی برای عدم برازش مناسب مدل برازش داده شده وجود ندارد و مدل برازش داده شده عملکرد مناسبی دارد.

۶ بحث و نتیجه‌گیری

در این مقاله یک الگوریتم جدید تنک‌سازی ارائه شد. همچنین فرایند تنک‌شده و نقاط تنک شده حاصل از اعمال الگوریتم فوق بر فرایند پواسون همگن بررسی شد و روشی برای برازش مدل‌های فوق ارائه شد. برای مطالعات آتی می‌توان به مطالعه فرایند تنک‌شده حاصل از اعمال الگوریتم فوق بر سایر فرایندهای اولیه اقدام نمود. همچنین پیشنهاد می‌شود سایر روش‌های برآورد پارامترهای مدل بررسی و با روش درستنمایی پالم مقایسه گردد.

مراجع

Stoyan, D. (1979), Interrupted point processes, *Biometrical Journal*, **21(7)**, 607-610.

Lavancier, F. and Møller, J. (2016), Modelling aggregation on the large scale and regularity on the small scale in spatial point pattern datasets, *Scandinavian Journal of Statistics*, **43(2)**, 587-609.

Møller, J., *Statistical inference and simulation for spatial point processes*, CRC Press.

Tanaka, U., Y. Ogata, and D. Stoyan (2008), Parameter estimation and model selection for neyman-scott point processes, *Biometrical Journal: Journal of Mathematical Methods in Bio- sciences* , **50(1)**, 43-57.

Neyman, J. and E. L. Scott (1958), Statistical approach to problems of cosmology, of the Royal Statistical Society: Series B (Methodological), **20(1)**, 1-29.

۷ ضمایم

قضیه ۱.۲ **اثبات.** (i) با استفاده از اینکه Z فرایند نقطه‌ای پواسون است به دست می‌آید. (ii) تساوی به ازای $||\zeta_1 - \zeta_2|| > 2r$ با استفاده از مستقل بودن تعداد نقاط در ناحیه‌های مجزا در فرایند نقطه‌ای پواسون بدست می‌آید. در حالتی که $||\zeta_1 - \zeta_2|| \leqslant 2r$ داریم:

$$\begin{aligned}
P[\Pi(\zeta_1) = p^n, \Pi(\zeta_2) = p^m] &= P[N(A_{\zeta_1}) = n, N(A_{\zeta_2}) = m],\\
&= \sum_{k=0}^{\min(m,n)} P[N(A_{\zeta_1} - A_{\zeta_2}) = n-k, N(A_{\zeta_1} \cap A_{\zeta_2}) = k, N(A_{\zeta_2} - A_{\zeta_1}) = m-k],\\
&= \sum_{k=0}^{\min(m,n)} P[N(A_{\zeta_1} - A_{\zeta_2}) = n-k] P[N(A_{\zeta_1} \cap A_{\zeta_2}) = k] P[N(A_{\zeta_2} - A_{\zeta_1}) = m-k],\\
&= \sum_{k=0}^{\min(m,n)} \frac{e^{-\theta||A_{\zeta_1} - A_{\zeta_2}||}[\theta||A_{\zeta_1} - A_{\zeta_2}||]^{n-k}}{(n-k)!}\\
&\times \frac{e^{-\theta||A_{\zeta_1} \cap A_{\zeta_2}||}[\theta||A_{\zeta_1} \cap A_{\zeta_2}||]^{k}}{k!} \times \frac{e^{-\theta||A_{\zeta_2} - A_{\zeta_1}||}[\theta||A_{\zeta_2} - A_{\zeta_1}||]^{m-k}}{(m-k)!},\\
&= e^{-\theta[2\pi r^2 - A(t,r)]}[\theta(\pi r^2 - A(t,r))]^{m+n} \sum_{k=0}^{\min(m,n)} \frac{1}{k!(n-k)!(m-k)!}\left[\frac{A(t,r)}{\theta(\pi r^2 - A(t,r))^2}\right]^k.
\end{aligned} \tag{1.7}$$

تساوی آخر با کمی محاسبات جبری بدست می‌آید. □

قضیه ۱.۳ **اثبات.** برای (i) با توجه به کاکس بودن فرایند X داریم:

$$\begin{aligned}
\lambda(u) = E[\delta\Pi(u)] = \delta E[\Pi(u)] &= \delta \sum_{k=0}^{\infty} p^k P(\Pi(u) = p^k) = \delta \sum_{k=0}^{\infty} p^k \frac{e^{-\mu(A_u)}\mu(A_u)^k}{k!}\\
&= \delta e^{-\mu(A_u)} \sum_{k=0}^{\infty} \frac{[p\mu(A_u)]^k}{k!} = \delta e^{-\mu(A_u)(1-p)} = \delta e^{-(1-p)\pi\theta r^2}
\end{aligned} \tag{2.7}$$

در تساوی چهارم از رابطه (۱.۲) استفاده شده و $A_u = B(u,r)$. همچنین با توجه به اینکه $Z \sim Poisson(\theta)$ لذا $\mu(A_u) = \theta\pi r^2$ است.

برای (ii) با توجه به کاکس بودن فرایند X داریم:

$$\lambda^{(2)}(u,v) = E[\delta\Pi(u)\delta\Pi(v)] = \delta^2 E[\Pi(u)\Pi(v)]$$

در صورتی‌که $||\zeta_1 - \zeta_2|| \geqslant 2r$ از استقلال ناشی از الگوریتم تنک‌سازی داریم:

$$E[\Pi(\zeta_1)\Pi(\zeta_2)] = E[\Pi(\zeta_1)]E[\Pi(\zeta_2)].$$

و با توجه به قسمت (i) نتیجه حاصل می‌شود.

برای $|\zeta_1 - \zeta_2| < 2r$ با استفاده از رابطه (۲.۲) داریم:

$$
\begin{aligned}
E[\Pi(\zeta_1)\Pi(\zeta_2)] &= \sum_{m,n=0}^{\infty} p^{n+m} P[\Pi(\zeta_1)=p^n, \Pi(\zeta_2)=p^m] \\
&= \sum_{m,n=0}^{\infty} p^{n+m} e^{-\theta[2\pi r^2 - A(t,r)]} [\theta(\pi r^2 - A(t,r))]^{m+n} \sum_{k=0}^{\min(m,n)} \frac{1}{k!(n-k)!(m-k)!} \left[\frac{A(t,r)}{\theta(\pi r^2 - A(t,r))^2}\right]^k \\
&= e^{-\theta[2\pi r^2 - A(t,r)]} \sum_{m,n=0}^{\infty} p^{n+m} [\theta(\pi r^2 - A(t,r))]^{m+n} \sum_{k=0}^{\min(m,n)} \frac{1}{k!(n-k)!(m-k)!} \left[\frac{A(t,r)}{\theta(\pi r^2 - A(t,r))^2}\right]^k \\
&= e^{-\theta[2\pi r^2 - A(t,r)]} \sum_{k=0}^{\infty}\sum_{n=k}^{\infty}\sum_{m=n}^{\infty} \frac{[p\theta(\pi r^2 - A(t,r))]^{m+n}}{k!(n-k)!(m-k)!} \left[\frac{A(t,r)}{\theta(\pi r^2 - A(t,r))^2}\right]^k \\
&+ e^{-\theta[2\pi r^2 - A(t,r)]} \sum_{k=0}^{\infty}\sum_{m=k}^{\infty}\sum_{n=m+1}^{\infty} \frac{[p\theta(\pi r^2 - A(t,r))]^{m+n}}{k!(n-k)!(m-k)!} \left[\frac{A(t,r)}{\theta(\pi r^2 - A(t,r))^2)}\right]^k \\
&= e^{-\theta[2\pi r^2 - A(t,r)]} \sum_{k=0}^{\infty} \frac{1}{k!} [\theta p^2 A(t,r)]^k \sum_{n=0}^{\infty} \frac{[p\theta(\pi r^2 - A(t,r))]^n}{n!} \sum_{m=n}^{\infty} \frac{[p\theta(\pi r^2 - A(t,r))]^m}{m!} \\
&+ e^{-\theta[2\pi r^2 - A(t,r)]} \sum_{k=0}^{\infty} \frac{1}{k!} [\theta p^2 A(t,r)]^k \sum_{m=0}^{\infty} \frac{[p\theta(\pi r^2 - A(t,r))]^m}{m!} \sum_{n=m+1}^{\infty} \frac{[p\theta(\pi r^2 - A(t,r))]^n}{n!} \\
&= e^{-2\pi r^2\theta(1-p)+\theta A(t,r)(1-p)^2} \sum_{n=0}^{\infty} \frac{e^{-p\theta[\pi r^2 - A(t,r)]}[p\theta(\pi r^2 - A(t,r))]^n}{n!} \\
&\times \sum_{m=n}^{\infty} \frac{e^{-p\theta[\pi r^2 - A(t,r)]}[p\theta(\pi r^2 - A(t,r))]^m}{m!} \\
&+ e^{-2\pi r^2\theta(1-p)+\theta A(t,r)(1-p)^2} \sum_{m=0}^{\infty} \frac{e^{-p\theta[\pi r^2 - A(t,r)]}[p\theta(\pi r^2 - A(t,r))]^m}{m!} \\
&\times \sum_{n=m+1}^{\infty} \frac{e^{-p\theta[\pi r^2 - A(t,r)]}[p\theta(\pi r^2 - A(t,r))]^n}{n!},
\end{aligned}
$$

اگر T_1, T_2 متغیرهای تصادفی مستقل و هم توزیع با $pos(p\theta(\pi r^2 - A(t,r)))$ باشند آنگاه :

$$
\begin{aligned}
E[\Pi(\zeta_1)\Pi(\zeta_2)] &= e^{-2\pi r^2\theta(1-p)+\theta A(t,r)(1-p)^2} P(T_1 \leqslant T_2) + e^{-2\pi r^2\theta(1-p)+\theta A(t,r)(1-p)^2} P(T_1 > T_2), \\
&= e^{-2\pi r^2\theta(1-p)+\theta A(t,r)(1-p)^2} [P(T_1 \leqslant T_2) + P(T_1 > T_2)] = e^{-2\pi r^2\theta(1-p)+\theta A(t,r)(1-p)^2},
\end{aligned}
$$

و در نتیجه:

$$
\lambda^{(2)}(u,v) = \begin{cases} \delta^2 e^{-2\pi r^2\theta(1-p)+\theta A(t,r)(1-p)^2}, & ||u-v|| \leq 2r, \\ \delta^2 e^{-2(1-p)\theta\pi r^2}, & ||u-v|| > 2r. \end{cases}
$$

□

قضیه ۲.۳ **اثبات.** قسمت (i) از قضیه ۱.۳ به دست می‌آید. برای (ii) داریم:

$$g_{Y-X}(u,v) = \frac{E\{\delta[۱-\Pi(u)]\delta[۱-\Pi(v)]\}}{E\{\delta[۱-\Pi(u)]\}E\{\delta[۱-\Pi(v)]\}} = \frac{۱-۲E[\Pi(u)]+E[\Pi(u)\Pi(v)]}{\{۱-E[\Pi(u)]\}^{۲}}$$
$$= \frac{۱-۲E[\Pi(u)]+E[\Pi(u)]^{۲}+E[\Pi(u)\Pi(v)]-E[\Pi(u)]^{۲}}{\{۱-E[\Pi(u)]\}^{۲}} = ۱+\frac{E[\Pi(u)\Pi(v)]-E[\Pi(u)]E[\Pi(v)]}{\{۱-E[\Pi(u)]\}^{۲}}$$
$$= ۱+\frac{E[\Pi(u)]^{۲}}{\{۱-E[\Pi(u)]\}^{۲}}[g_X(u,v)-۱] = ۱+\frac{a^{۲}}{(۱-a)^{۲}}[g_X(t)-۱]$$

□

A New Thinning Method of Spatial Point Processes Based on using Boolean Random Set Models

Reza Asadi[1], **Mojtaba Khazaei**[1], **Mojtaba Ganjali**[1]

[1]Department of Statistics, Shahid Beheshti University, Tehran, Iran

Abstract: Spatial point processes are statistical models for the analysis of point patterns. One of the methods of making new point processes is thinning the pre-existing point processes. In this paper, a new dependent thinning algorithm based on a Boolean random set is proposed. This algorithm can be considered as a generalization of the Stoyan thinning algorithm in the interrupted point process. Some properties of the thinned processes and eliminated point processes resulted from the application of the proposed thinning algorithm on the homogeneous Poisson point process, have been studied. Some summary statistics of the both thinned and eliminated point processes are calculated. Palm likelihood method is used to estimate parameters of these models. Finally, the proposed models are used to model two real datasets.

Keywords: Homogeneous Poisson point process; Cox process; Boolean random set; Dependent thinning; Cox process; Palm likelihood.

Mathematics Subject Classification (2020): xxAxx, xxBxx, xxCxx.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

آزمون نیکویی برازش تقارن توزیع‌های پیوسته تحت داده‌های از چپ بریده شده و از راست سانسور شده

معصومه اکبری [۱]

گروه آمار، دانشگاه مازندران، بابلسر، ایران.

چکیده: در این مقاله، ابتدا یک آماره آزمون برای متقارن بودن توزیع‌های پیوسته تحت داده‌های از چپ بریده شده و از راست سانسور شده معرفی می‌شود. نشان داده می‌شود که این آماره آزمون مجانباً نرمال است. در یک مطالعه شبیه سازی گسترده، توان آزمون پیشنهادی مورد بررسی قرار گرفته است. نتایج گویای رضایت بخش بودن عملکرد آزمون است. در نهایت کاربردی بودن آماره آزمون در یک مثال واقعی نشان داده می‌شود.

واژه‌های کلیدی: آزمون نیکویی برازش، توزیع متقارن، داده‌های سانسور شده از راست و بریده شده از چپ.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62N01، 62G10.

---

۱ مقدمه

فرض کنید X_1، X_2،...، X_n یک نمونه تصادفی به حجم n، از یک تابع چگالی ناشناخته f با میانگین متناهی μ باشد. گفته می‌شود چگالی f نسبت به μ متقارن است اگر و فقط اگر

$$H_0 : f(\mu - x) = f(\mu + x), \qquad \forall x \in \mathbb{R}. \tag{۱.۱}$$

بررسی فرض تقارن یکی از موضوعات اساسی در آمار ناپارامتری است و کاربردهای گسترده‌ای در شاخه‌های مختلف علمی دارد. توجه به تقارن توزیع داده‌ها از دیرباز مورد توجه پژوهشگران بوده، چراکه بسیاری از آزمون‌ها و روش‌های آماری کلاسیک بر مبنای این فرض توسعه یافته‌اند. اولین مطالعات در زمینه آزمون تقارن به دهه‌های میانی قرن بیستم بازمی‌گردد، زمانی که فرضیه تقارن به عنوان معیاری برای اعتبارسنجی مدل‌های نرمال در داده‌های تجربی مورد استفاده قرار گرفت. در آن دوران، آزمون‌هایی مبتنی بر مرتب‌سازی داده‌ها، توزیع رتبه‌ها، و مقایسه میانه و میانگین برای آزمون تقارن توسعه یافتند.

---

[۱] سخنران، m.akbari@umz.ac.ir

در سال‌های بعد، با پیچیده‌تر شدن ساختار داده‌ها، روش‌های آماری نیز تکامل یافتند و آزمون‌های متنوع‌تری برای بررسی تقارن توسعه داده شد. به عنوان نمونه، آزمون‌های مبتنی بر آمار ناپارامتری، آزمون‌های استوار، و روش‌هایی مبتنی بر فاصله و تبدیل داده‌ها پیشنهاد شدند. این آزمون‌ها نه تنها در تحلیل داده‌های کلاسیک، بلکه در زمینه‌هایی مانند اقتصاد، ژنتیک، پزشکی و مهندسی نیز به‌کار گرفته شده‌اند.

در مطالعات مالی، فرض تقارن نقش کلیدی در مدل‌سازی بازده دارایی‌ها و قیمت‌گذاری ابزارهای مالی ایفا می‌کند؛ به عنوان نمونه می‌توان به مدل قیمت‌گذاری دارایی‌های سرمایه‌ای شارپ–لینتر و مدل بلک–شولز برای قیمت‌گذاری اختیار معامله اشاره کرد که در آن‌ها تقارن توزیع بازده دارایی‌ها یا خطاها فرض می‌شود. از سوی دیگر، در روش‌های پارامتری و استوار آماری، تقارن اغلب به عنوان فرض پایه‌ای برای خطاها یا متغیرهای پاسخ در مدل‌هایی چون رگرسیون خطی و تحلیل واریانس تلقی می‌شود.

با توجه به اهمیت این موضوع، پژوهش‌های متعددی به آزمون فرضیه H_0 در معادله (۱.۱) پرداخته‌اند. از جمله می‌توان به مطالعات رندلز و همکاران (۱۹۸۰)، مک ویلیامز (۱۹۹۰)، مدرس و گستویرت (۱۹۹۰)،گیبنز و چاکرابرتی (۱۹۹۲)، تاجدین (۱۹۹۴)، باکلیزی (۲۰۰۳، ۲۰۰۷ و ۲۰۰۸)، چنگ و بالاکریشنان (۲۰۰۴)، کورزو و باباتیوا (۲۰۱۳) و شیانگ و همکاران (۲۰۲۱) اشاره کرد که همگی آزمون‌هایی برای داده‌های کامل پیشنهاد داده‌اند.

افزون بر مطالعات کلاسیک، در سال‌های اخیر تلاش‌های متعددی برای توسعه آزمون‌های تقارن با ویژگی‌های آماری بهتر یا سازگار با شرایط خاص داده‌ها صورت گرفته است. به‌عنوان نمونه، ژو و ونگ (۲۰۱۵) آزمونی مبتنی بر نسبت درست‌نمایی تجربی ارائه دادند که ویژگی‌های توان مناسبی دارد. جوریکوا و کالینا (۲۰۱۶) آزمون‌هایی برای تحلیل تقارن در داده‌های اقتصادی ارائه کردند. همچنین گوش و سن (۲۰۱۷) نیز طبقه‌ای از آزمون‌های ناپارامتری مبتنی بر U–آماره‌ها برای بررسی تقارن توسعه دادند. در شرایطی که داده‌ها ناقص باشند، مطالعات محدودی در این حوزه انجام شده‌است. به‌عنوان مثال، شیانگ و همکاران (۲۰۱۹) آزمونی برای تقارن در حضور داده‌های گمشده پیشنهاد کردند و اخیراً کائو و ژو (۲۰۲۱) آزمون‌های ناپارامتری جدیدی برای داده‌های سانسورشده تصادفی توسعه داده‌اند. اکبری و زمانی (۱۴۰۳) آزمونی برای سنجش تقارن در توزیع‌های پیوسته در حضور داده‌های سانسور شده از راست معرفی کرده‌اند.

در این مقاله، به گسترش این موضوع پرداخته و یک آزمون جدید برای بررسی تقارن توزیع پیوسته در شرایط وجود داده‌های بریده شده از چپ و سانسور شده از راست پیشنهاد می‌دهیم. ساختار مقاله به‌صورت زیر است: در بخش ۲، آماره آزمون معرفی می‌شود. از آنجا که توزیع دقیق این آماره به‌راحتی در دسترس نیست، در بخش ۳ با استفاده از روش مونت‌کارلو مقادیر بحرانی و توان آزمون به صورت تقریبی محاسبه می‌شود. نهایتاً، در بخش ۴ کاربرد عملی آزمون با استفاده از یک مجموعه داده واقعی توضیح داده می‌شود.

۲ آماره آزمون

فرض کنید X_1، X_2،...، X_n یک نمونه تصادفی از توزیع پیوسته F با میانه نامعلوم θ باشد. اگر i، j و k سه مقدار مجزا متعلق به مجموعه $\{1, 2, \ldots, n\}$ باشد، در اینصورت مشاهدات سه تایی (X_i, X_j, X_k) تشکیل یک سه تایی راستگرا [1] را می‌دهند، بدین

[1] right triple

معنا که توزیع بنظر می‌رسد که به راست چوله است اگر مشاهده میانی به کوچکترین مشاهده نزیک‌تر باشد. بطور مشخص‌تر فرض کنید

$$\begin{aligned} f^*(X_i, X_j, X_k) &= \frac{1}{3}\{sign(X_i + X_j - 2X_k) \\ &\quad + sign(X_i + X_k - 2X_j) + sign(X_j + X_k - 2X_i)\}, \end{aligned} \tag{1.2}$$

که در آن $sign(u) = -1, 0, 1$ به ترتیب برای مقادیر $u < 0$، $u = 0$ و $u > 0$. با توجه به معادله (۱.۲)، (X_i, X_j, X_k) تشکیل یک سه تایی راست گرا را می‌دهد اگر $f^*(X_i, X_j, X_k) = \frac{1}{3}$ و بطور مشابه تشکیل یک سه تایی چپ گرا را می‌دهد اگر $f^*(X_i, X_j, X_k) = -\frac{1}{3}$ و زمانی که $f^*(X_i, X_j, X_k) = 0$ باشد بیانگر آن است که سه تایی نه راست گرا و نه چپ گراست. قابل توجه است که $f^*(X_i, X_j, X_k)$ فقط می‌تواند مقادیر $-\frac{1}{3}$ ، 0 و $\frac{1}{3}$ را اختیار کند. بر همین اساس رندلز و همکاران (۱۹۸۰) یک آزمون بر اساس U-آماره

$$\hat{\eta} = \frac{1}{\binom{n}{3}} \sum_{i<j<k} f^*(X_i, X_j, X_k), \tag{2.2}$$

پیشنهاد کردند که مقادیر بزرگ $|\hat{\eta}|$ منجر به رد فرضیه تقارن می‌شود. در ادامه نحوه لحاظ کردن مشاهدات از چپ بریده شده و از راست سانسور شده را در آماره آزمون (۲.۲) مورد بررسی قرار می‌دهیم.

فرض کنید که X، C و L به ترتیب با تابع توزیع‌های F، G و H بیانگر متغیرهای تصادفی طول عمر، زمان سانسور و زمان برش از چپ باشند و $\bar{F}(x) = 1 - F(x)$، $\bar{G}(x) = 1 - G(x)$ و $\bar{H}(x) = 1 - H(x)$. تحت سانسور از راست و برش از چپ، n نمونه مستقل و هم توزیع $(T_i\epsilon_i, \delta_i)$ از $(T\epsilon, \delta)$ در نظر گرفته می‌شود که $T = \min(X, C)$، $\delta = I(X < C)$ و $\epsilon = I(T > L)$. به عبارت دیگر، δ نشانه سانسور از راست و ϵ برای مشخص کردن برش استفاده شده است. واضح است که T_i یک مشاهده برای iمین نمونه است اگر $T_i > L_i$ باشد. بر اساس داتا (۲۰۱۰) می‌توان براساس داده‌های از راست سانسور شده و از چپ بریده شده، یک نسخه از آماره (۲.۲) بصورت زیر در نظر گرفت.

$$\hat{\eta}_{LTRC} = \frac{1}{\binom{n}{3}} \sum_{1 \le i<j<k \le n} \frac{f^*(T_i, T_j, T_k)\delta_i\delta_j\delta_k}{\hat{K}_C(T_i)\hat{K}_C(T_j)\hat{K}_C(T_k)},$$

که در آن $\hat{K}_C(\tau)$ برآوردگری از تابع بقا زمان سانسور C تحت برش چپ است و بصورت زیر می‌باشد.

$$\hat{K}_C(\tau) = \prod_{t \le \tau} \left(1 - \frac{dN^C(t)}{Y(t)}\right),$$

که در آن

$$\begin{aligned} dN_i^C(t) &= I(T_i = t, \delta_i = 0), \\ Y(t) &= \sum_{i=1}^{n} Y_i(t), \qquad \text{and} \quad Y_i(t) = I(T_i \ge t \ge L_i). \end{aligned} \tag{3.2}$$

فرض H_0 یعنی متقارن بودن توزیع پیوسته، به ازای مقادیر بزرگ $|\hat{\eta}_{LTRC}|$ رد می‌شود. تحت شرایطی می‌توان نشان داد که آماره آزمون $\hat{\eta}_{LTRC}$ مجانباً نرمال است.

۳ مقادیر بحرانی و توان آزمون

در این بخش، بر اساس شبیه‌سازی مونت کارلو، عملکرد آزمون پیشنهادی مورد بررسی قرار می‌گیرد. همه محاسبات بر اساس نرم افزار R انجام شده است. بدین منظور در ابتدا، مقادیر بحرانی به ازای مقادیر $n = ۲۰, ۵۰, ۷۰$ تقریب زده شده است در حالی که، دو مقدار $p = ۰/۱, ۰/۴$ برای سانسور از راست یعنی $P(X > C) = p$ و $C \sim Exp(\lambda)$ و مستقل از X، در نظر گرفته شده است. علاوه‌براین توزیع متغیر تصادفی برش از چپ بصورت، $L \sim Exp(\mu)$ فرض شده به قسمی که $P(L > X) = ۰/۲$ گردد. نتایج در سطح خطای نوع اول، $\alpha = ۰/۰۵$، بصورت زیر در جدول ۱ گزارش شده است. برای بررسی توان آزمون پیشنهاد شده، توزیع‌های

جدول ۱: مقادیر بحرانی آزمون $\hat{\eta}_{LTRC}$

p	n = ۲۰	۵۰	۷۰
۰/۱	۰/۰۶۸	۰/۰۳۹	۰/۰۳۲
۰/۴	۰/۰۳۵	۰/۰۲۳	۰/۰۱۸

مقابل از خانواده توزیع لامبدای تعمیم‌یافته (GLD[۲]) با تابع چندک زیر در نظر گرفته شده است.

$$F^{-۱}(u) = \lambda_۱ + \frac{u^{\lambda_۳} - (۱-u)^{\lambda_۴}}{\lambda_۲}, \quad ۰ < u < ۱.$$

در جدول ۲، ۹ حالت از GLD به ازای پارامترهای مختلف آن بعنوان ۹ توزیع مقابل نشان داده شده است. جدول ۳ توان آزمون

جدول ۲: ۹ حالت از GLD

$Case$	$\lambda_۱$	$\lambda_۲$	$\lambda_۳$	$\lambda_۴$
۱	۰/۰۰۰۰	۰/۱۹۷۵	۰/۱۳۴۹	۰/۰۱۳۴۹
۲	−۰/۱۱۶۷	−۰/۳۵۱۷	−۰/۱۳۰۰	−۰/۱۶۰۰
۳	۰/۰۰۰۰	−۱/۰۰۰۰	−۰/۱۰۰۰	−۰/۱۸۰۰
۴	۳/۵۸۶۵	۰/۰۴۳۱	۰/۰۲۵۲	۰/۰۹۴۰
۵	۰/۰۰۰۰	−۱/۰۰۰۰	−۰/۰۰۷۵	۰۳۰۰.۰−
۶	۰/۰۰۰۰	۱/۰۰۰۰	۱/۴۰۰۰	۰/۰۲۵۰۰
۷	۰/۰۰۰۰	۱/۰۰۰۰	۰/۰۰۰۱	۰/۱۰۰۰
۸	۰/۰۰۰۰	−۱/۰۰۰۰	−۰/۰۰۱۰	−۰/۱۳۰۰
۹	۰/۰۰۰۰	−۱/۰۰۰۰	−۰/۰۰۰۱	−۰/۱۷۰۰

محاسبه شده $\hat{\eta}_{LTRC}$ را گزارش می‌دهد. همان‌طور که مشاهده می‌شود به ازای افزایش حجم نمونه، توان آزمون‌ها افزایش یافته و توان‌ها زمانی که درجه سانسور کوچک است ($P = ۰/۱$) به نسبت بزرگتر می‌باشد. در مجموع توان آزمون قابل قبول بوده و در اکثر موارد جز در توزیع مقابل حالت دوم، توان خوبی بدست آمده است.

از طرفی حالت اول از توزیع مقابل، همان توزیع نرمال استاندارد است لذا توان آزمون در این مورد، همان خطای نوع اول را نشان می‌دهد که مقادیر گزارش شده حول مقدار ۰/۰۵ همان خطای نوع اول اسمی در نظر گرفته شده است. بنابراین می‌توان گفت برآوردهای خطای نوع اول نیز تحت کنترل و قابل قبول است.

[2] generalized lambda distribution

جدول ۳: توان آزمون

n	$Case۱$	$Case۲$	$Case۳$	$Case۴$	$Case۵$	$Case۶$	$Case۷$	$Case۸$	$Case۹$
				$p = ۰/۱$					
۲۰	۰/۰۵۰۴	۰/۰۷۸	۰/۱۰۶۶	۰/۱۷۰۴	۰/۲۵۶۲	۰/۲۱۶۴	۰/۵۸۳۴	۰/۹۰۰۸	۰/۶۳۴۸
۵۰	−۰/۰۴۴۸	−۰/۰۶۹	−۰/۲۰۰۴	−۰/۴۳۹۸	۰/۶۲۶۸	۰/۵۱۶	۰/۹۶۹۷	۰/۹۹۹۶	۰/۹۸۶
۷۰	۰/۰۶۳۶	۰/۰۸۱	۰/۳۱۱	۰/۵۹۵۸	۰/۷۵۸۸	۰/۶۵۵	۰/۹۸۶۸	۰/۹۹۹۸	۰/۹۹۵
				$p = ۰/۴$					
۲۰	۰/۰۴۹۴	۰/۱۲۷۲	۰/۰۴۴۸	۰/۰۵۷۶	۰/۰۵	۰/۳۳۰۶	۰/۲۲۰۴	۰/۲۴۲۶	۰/۲۴۷۶
۵۰	۰/۰۵۴	۰/۲۰۸۸	۰/۰۲۱۶	۰/۰۵۷۶	۰/۰۳۶۸	۰/۴۶۲۸	۰/۴۵۶۲	۰/۵۰۳۲	۰/۵۲۹۲
۷۰	۰/۰۹۷۶	۰/۳۶۹۴	۰/۰۳۸۶	۰/۱۰۸۲	۰/۰۶۲	۰/۵۹۱۸	۰/۶۷۵۴	۰/۷۳۳۶	۰/۷۶۸۴

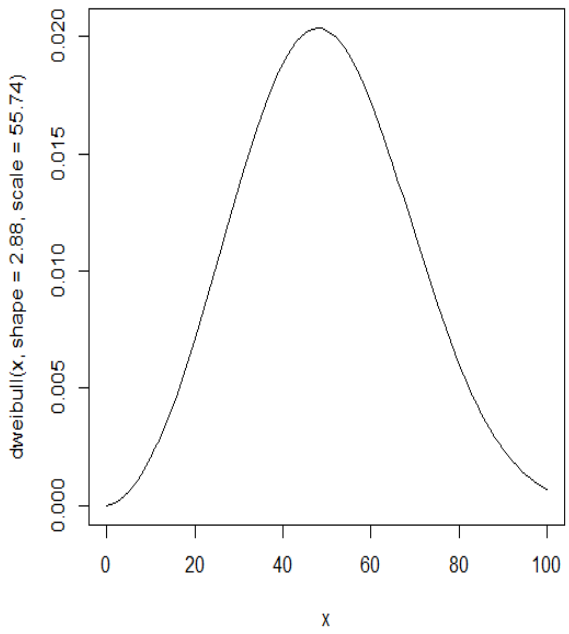
۴ مثال واقعی

برای تشریح نحوه بکار بردن آزمون پیشنهادی، داده‌های از چپ بریده شده و از راست سانسور شده هانگ و همکاران (۲۰۰۹) را بکار می‌بریم که این مجموعه داده شامل طول عمر ۷۱۰ ترانسفورماتور یک شرکت انرژی با ۶۲ شکست است. اگرچه داده‌های اصلی موجود نیست اما یک زیر مجموعه ۲۸۶تایی با ۳۹ شکست در امورا و شیو (۲۰۱۶) آمده است. داده‌ها بین سال‌های ۱۹۸۰ الی ۲۰۰۸ جمع‌آوری شده است. ازاین‌رو ترانسفورماتورهایی که قبل از ۱۹۸۰ نصب شده‌اند بعنوان داده‌های از چپ بریده شده در نظر گرفته می‌شوند. بنابراین داده‌ها از چپ بریده شده و از راست سانسور شده است. نرخ سانسور راست و برش از چپ به ترتیب ۸۶/۴ و ۵۸/۳۹ درصد هستند. هانگ و همکاران (۲۰۰۹) یک توزیع وایبل با پارامتر شکل ۲/۸۸ و پارامتر مقیاس ۵۵/۷۴ به این مجموعه داده‌ها برازش داده‌اند. نمودار چگالی این توزیع در شکل زیر آمده است و به نوعی نشان می‌دهد که توزیع مجموعه داده موردنظر، متقارن است. آماره آزمون پیشنهادی مشاهده شده بصورت $\hat{\eta}_{LTRC} = ۰/۰۶۲۵$ بدست آمده است و برای بدست آوردن p–مقدار یک نمونه به حجم $n = ۵۰$ ازتوزیع وایبل با پارامترهای ذکر شده تولید کرده و $L \sim Exp(\mu)$ و $C \sim Exp(\lambda)$ در نظر می‌گیریم. مقادیر μ و λ را طوری تعیین می‌کنیم که به نرخ‌های به ترتیب ۵۸/۲۹ و ۸۶/۴ منتج شود. بر اساس هر تکرار یک مقدار از آماره $\hat{\eta}_{LTRC}$ بدست خواهد آمد و براساس ۵۰۰۰ تکرار، میانگین تعداد دفعاتی که $\hat{\eta}_{LTRC}$ بیشتر از آماره آزمون مشاهده شده ۰/۰۶۲۵ باشد را به عنوان p–مقدار در نظر خواهیم گرفت. با توجه به مراحل مذکور، p–مقدار بدست آمده برابر با ۰/۷ بدست آمد که حاکی از آن است که توزیع مجموعه داده موردنظر متقارن است.

مراجع

اکبری، م. و زمانی، ز. (۱۴۰۳)، معرفی آزمون جدید برای متقارن بودن یک توزیع در حضور داده‌های سانسور شده، هشتمین همایش ریاضیات و علوم انسانی ریاضیات مالی، ۲۳ و ۲۴ آبان۱۴۰۳، دانشگاه علامه طباطبائی، تهران.

Baklizi A. (2003), A conditional distribution free runs test for symmetry, *Journal of Nonparametric Statistics*, **15**, 713–718.



شکل ۱: چگالی توزیع وایبل

Baklizi A. (2007), Testing symmetry using a trimmed longest run statistics, *Australian and New Zealand Journal of Statistics*, **49**, 339–347.

Baklizi A. (2008), Improving the power of the Hybrid test, *International Journal of Contemporary Mathematical Sciences*, **3**, 497–499.

Cao Y. and Zhu W. (2021). Nonparametric tests for symmetry under random censoring. *Statistics & Probability Letters*, **169**, 108–967.

Cheng W. and Balakrishnan N. (2004), A modified sign test for symmetry, *Communications in Statistics-Simulation and Computation*, **33**, 703–709.

Corzo J. and Babativa G. (2013), A modified runs test for symmetry, *Journal of Statistical Computation and Simulation*, **83**, 984–991.

Datta S., Bandyopadhyay D., and Satten, G. A. (2010), Inverse probability of censoring weighted U-statistics for right-censored data with an application to testing testing hypotheses, *Scandinavian Journal of Statistics* , **37**, 680–700.

Emura T., and Shiu v. (2016), Estimation and model selection for left-truncated and right-censored lifetime

data with application to electric power transformers analysis, Communication in Statisttics: Simulation and Computation, **45**, 171–318.

Gibbons C., and Chakraborti S. (1992), *Nonparametric Statistical Inrefence*, New York: Marcel Dekker.

Ghosh A. K., and Sen A. (2017). A general class of tests for symmetry based on U-statistics. *Journal of Nonparametric Statistics*, **29**, 1–-18.

Hong Y., Meeker W.Q. and McCalley J.D. (2009), Prediction of remaining life of power transformers based on left truncated and right censored lifetime data, Annals of Applied Statistics,**3** 3, 857-–879.

Jurečková J., and Kalina J. (2016). Tests for symmetry of distributions with application in economics. *Statistical Papers*, **57**, 197–-214.

McWillials P. (1990), A distribution-free test for symmetry based on a runs statistic, *Journal of the American Statistical Association*, **85**, 1130–1133.

Modarres R. and Gastwirth J. (1998), Hybrid test for the hypothesis of symmetry, *Journal of Applied Statistics*, **25**, 777-783.

Randles R. H., Fligner M. A., Policello ll G. E., and Wolfe D. A. (1980), An asymptotically distribution-free test for symmetry versus asymmetry, *Journal of the American Statistical Association*, **75**, 168-172.

Tajuddin I. (1994), Distribution-free test for symmetry based on Wilcoxon two-sample test, *Journal of Applies Statistics*, **21**, 409–415.

Xiong J., Zhou X., and Xu J. (2019). A test for symmetry of a distribution with missing data. *Communications in Statistics - Theory and Methods*, **48**, 1682-–1695.

Xiong P., Zhuang W., and Qiu G. (2021), Testing symmetry based on the extropy of record values, *Journal of Nonparametric Statistics*, **33**, 134–155.

Zou C., and Wang Z. (2015). A new test for symmetry based on the empirical likelihood ratio. *Journal of Statistical Computation and Simulation*, **85**, 348–358.

A Symmetry Goodness-of-Fit Test for Continuous Distributions under Left Truncation and Right Censoring

Masoumeh Akbari

Department of Statistics, University of Mazandaran, Babolsar, Iran.

Abstract: In this paper, a test statistic is first introduced for assessing the symmetry of continuous distributions under right-censored and left-truncated data. It is shown that the proposed test statistic is asymptotically normal. In an extensive simulation study, the power of the proposed test is examined. The results indicate that the test performs satisfactorily. Finally, the applicability of the test statistic is demonstrated through a real data example.

Keywords: Goodness-of-fit test, Symmetric distribution, Left-truncated and right-censored data.

Mathematics Subject Classification (2020): 62N01, 62G10.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

یک کلاس از مدل‌بندی سری‌زمانی دوخطی گسسته‌مقدار با کاربردهای آن

رعنا بامدادی [۱] و مهرناز محمدپور

گروه آمار، دانشگاه مازندران

چکیده: در این مطالعه، یک مدل تعدیل‌یافته دوخطی گسسته‌مقدار مرتبه دو برای تحلیل سری‌های زمانی غیرخطی معرفی می‌شود. این مدل از انعطاف‌پذیری بالایی برخوردار است به‌گونه‌ای که ورودی‌های جدید می‌توانند به‌طور مستقل رفتار کنند. مدل‌بندی ارائه‌شده برای فرایندهای بیش‌پراکنده به‌کار می‌رود که بر یک چارچوب غیرخطی تکیه دارند. ویژگی‌های اصلی مدل پیشنهادی به‌طور نظری بررسی شده‌است. همچنین، به‌منظور استفاده عملی از مدل، چند روش برای برآورد پارامترها ارائه و به‌منظور ارزیابی عملکرد این روش‌ها شبیه‌سازی انجام شده‌است. در پایان، برای نشان دادن کاربرد مدل، به تحلیل داده‌ها پرداخته شده است.

واژه‌های کلیدی: پیش‌بینی، مدل تعدیل‌یافته دوخطی گسسته‌مقدار، تقریب نقطه زینی، روش ویتل.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62P10، 62M10، 60G25.

۱ مقدمه

در سال‌های اخیر، مدل‌بندی سری‌های زمانی شمارشی به دلیل کاربردهای وسیع آن مورد توجه قرار گرفته و توسعه‌ی چشمگیری یافته است. این نوع مدل‌بندی در بسیاری از حوزه‌های علمی، از جمله بیمه، اقتصاد، داروسازی، اپیدمیولوژی، سیستم‌های صف، ارتباطات، هواشناسی، مخابرات و غیره کاربردهای گسترده‌ای دارند و ضرورت دارد که برازش مناسب بر روی این داده‌ها صورت گیرد.

برای مدل‌سازی سری‌های زمانی شمارشی، به‌کارگیری عملگرهای رقیق‌ساز [۱] به عنوان روشی مؤثر شناخته شده‌اند. عملگرهای رقیق‌ساز مختلفی توسط ویب (۲۰۰۸) و اسکات و همکاران (۲۰۱۵) معرفی شده‌اند.

جهت مدل‌بندی داده‌هایی با این خصوصیت ژانگ و همکاران (۲۰۱۰) عملگر رقیق‌ساز سری توانی تعمیم‌یافته علامت [۲] به‌صورت زیر

$$\alpha \circledast Y = \text{sign}(\alpha)\text{sign}(Y)\sum_{i=1}^{|Y|} W_i,$$

[۱] سخنران، bamdadirana@gmail.com

[1] Thining operator

[2] Signed Generalized Power Series Thinning Operator

معرفی کرده‌اند، که در آن، دنباله‌ی شمارشی W_i از متغیرهای تصادفی مستقل و هم‌توزیع با توزیع توان تعمیم‌یافته [3] در نظر گرفته می‌شود.

با اینکه در حوزه مدل‌سازی سری‌های زمانی شمارشی پیشرفت‌های زیادی حاصل شده، ولی هنوز مدل‌های غیرخطی در این زمینه کمتر مورد توجه و بررسی قرار گرفته‌اند. در یکی از تلاش‌های برجسته در این حوزه، دوقان و همکاران (۲۰۰۶) مدل‌های دوخطی را با استفاده از عملگر رقیق‌ساز گسترش دادند. یک رویکرد جایگزین برای این مدل‌ها توسط بامدادی و همکاران (۲۰۲۴) ارائه شده است. آن‌ها مدل دوقان و همکاران (۲۰۰۶) را اصلاح کرده و مدل دوخطی را معرفی کردند که انعطاف‌پذیری بیشتری برای ورودی‌های جدید ارائه می‌دهد. این مدل از این مزیت برخوردار است که افراد جدید می‌توانند به‌طور مستقل رفتار کنند. در این مقاله، مدل فوق با مرتبه ۲ برای فرایندهایی که تابع خودهمبستگی جزیی آن در تاخیر ۲ به‌طور معنی‌داری از صفر اختلاف دارد بررسی می‌شود.

۲ معرفی مدل و برخی ویژگی‌های آن

فرآیند $\{X_t\}_{t\in\mathbb{Z}}$، را به‌صورت زیر در نظر بگیرید

$$X_t = a \circledast X_{t-2} + (a \circledast X_{t-2})(b \circledast \varepsilon_{t-1}) + \varepsilon_t, \tag{۱.۲}$$

که در آن $\circledast$ عملگر رقیق‌ساز سری توانی تعمیم‌یافته علامت می‌باشد، $a, b \in [-1, 1]$ و $\{\varepsilon_t\}$ دنباله‌ای از متغیرهای تصادفی گسسته i.i.d. با میانگین μ_ε و واریانس σ_ε^2 هستند که از $\{X_s\}_{s<t}$ مستقل می‌باشند. همچنین، $\{W_i\}$ و $\{\tilde{W}_i\}$ که در عملگرهای $a\circledast$ و $b\circledast$ به‌کار رفته‌اند، دارای توزیع سری توانی با میانگین‌های $|a|$ و $|b|$ و واریانس‌های α و β هستند. این مدل به عنوان مدل تعدیل‌یافته دوخطی گسسته‌مقدار مبتنی بر عملگر رقیق‌ساز از مرتبه دو یا به اختصار MINBL$(2, 0, 2, 1)$ شناخته می‌شود.

گزاره ۱.۲. اگر $(|a| + |ab|\mu_{|\varepsilon|})^2 + |ab|^2\sigma_\varepsilon^2 + |a|^2\beta\mu_{|\varepsilon|} < 1$، آنگاه فرآیند $\{X_t\}_{t\in\mathbb{Z}}$ در *(۱.۲)* ایستا می باشد.

گزاره ۲.۲. فرض کنید $\{X_t\}_{t\in\mathbb{Z}}$ فرآیند $MINBL(2, 0, 2, 1)$ معرفی شده در *(۱.۲)* باشد که شرایط گزاره ۱.۲ برقرار باشد. در این‌صورت، امیدریاضی و تابع خودکواریانس فرآیند به‌صورت زیر می‌باشد

$$E(X_t) = \frac{\mu_\varepsilon}{1 - (a + ab\mu_\varepsilon)}, \tag{۲.۲}$$

و

$$\gamma_X(1) = \frac{ab\sigma_\varepsilon^2 E(X_t)}{1 - (a + ab\mu_\varepsilon)}, \quad \gamma_X(2k) = (a + ab\mu_\varepsilon)^k \gamma_X(0), \quad \gamma_X(2k+1) = (a + ab\mu_\varepsilon)^k \gamma_X(1), \ k \geq 1.$$

گزاره ۳.۲. میانگین شرطی فرآیند $\{X_t\}$ برای $k-$گام جلوتر به‌صورت زیر است

$$\begin{aligned}
E(X_{t+1}|t) &= (a + ab\varepsilon_t)X_{t-1} + \mu_\varepsilon, \\
E(X_{t+2}|t) &= (a + ab\mu_\varepsilon)X_t + \mu_\varepsilon, \\
E(X_{t+k}|t) &= (a + ab\mu_\varepsilon)^{\frac{k}{2}} X_t + \mu_\varepsilon \sum_{i=1}^{\frac{k}{2}} (a + ab\mu_\varepsilon)^{i-1}, \ k = 2h,
\end{aligned} \tag{۳.۲}$$

$$E(X_{t+k}|t) = (a + ab\mu_\varepsilon)^{[\frac{k}{2}]} E(X_{t+1}|t) + \mu_\varepsilon \sum_{i=1}^{[\frac{k}{2}]} (a + ab\mu_\varepsilon)^{i-1}, \ k = 2h+1. \tag{۴.۲}$$

[3] Generalized Power Series

گزاره ۴.۲. واریانس شرطی یک گام به جلو فرآیند $\{X_t\}$ به‌صورت

$$Var(X_{t+1}|t) = a^2\beta|\varepsilon_t|X_{t-1}^2 + (\alpha + 2\alpha b\varepsilon_t + \alpha b^2\varepsilon_t^2 + \alpha\beta|\varepsilon_t|)|X_{t-1}| + \sigma_\varepsilon^2$$

است.

۳ برآوردیابی و شبیه‌سازی

در این بخش، به برآورد پارامترهای مدل MINBL(۲, ۰, ۲, ۱) با استفاده از چهار روش مختلف می‌پردازیم. برای ارزیابی کارایی این روش‌ها، از داده‌های شبیه‌سازی‌شده استفاده شده و نتایج حاصل از هر روش با یکدیگر مقایسه می‌شوند. در روند برآورد، توزیع $\{\varepsilon_t\}$ پواسن با پارامتر μ_ε در نظر گرفته شده است.

۱.۳ روش یول‌واکر

با به‌کارگیری گزاره ۲.۲، امید ریاضی و تابع خودکواریانس مدل تحت فرض توزیع پواسن خطاها به‌صورت زیر بیان می‌شوند

$$\mathrm{E}(X_t) = \frac{\mu_\varepsilon}{1-(a+ab\mu_\varepsilon)},\ \ \gamma_X(1) = \frac{ab\mu_\varepsilon \mathrm{E}(X_t)}{1-(a+ab\mu_\varepsilon)},$$
$$\gamma_X(2k) = (a+ab\mu_\varepsilon)^k\gamma_X(0),\ \ \gamma_X(2k+1) = (a+ab\mu_\varepsilon)^k\gamma_X(1).$$

فرض کنید $A = a + ab\mu_\varepsilon$ و $B = ab\mu_\varepsilon$. با برابر قرار دادن گشتاورهای مرتبه اول و توابع خودکواریانس فرآیند با میانگین نمونه‌ای و تابع خودکواریانس نمونه‌ای برآورد پارامترها به‌صورت

$$\hat{a}_{YW} = \hat{A} - \hat{B},\ \ \hat{b}_{YW} = \frac{\hat{B}}{\hat{a}_{YW}\hat{\mu}_{YW}},\ \ \hat{\mu}_{YW} = \bar{X}(1-\hat{A}),$$

می‌باشد، که در آن $\hat{A} = \frac{\gamma_X(2)}{\gamma_X(0)}$ و $\hat{B} = \frac{\gamma_X(1)(1-\hat{A})}{\bar{X}}$.

۲.۳ روش کمترین مربعات شرطی

برآوردگرهای کمترین مربعات شرطی پارامترهای $\Theta = (a, b, \mu_\varepsilon)$، از طریق مینیمم‌سازی تابع هدف

$$Q(\Theta) = \sum_{t=3}^{n}(X_t - \mathrm{E}(X_t|t-1))^2 = \sum_{t=3}^{n}(X_t - (a+ab\varepsilon_{t-1})X_{t-2} - \mu_\varepsilon)^2,$$

حاصل می‌شود، که در آن $\varepsilon_t = X_t - aX_{t-2} - abX_{t-2}\varepsilon_{t-1}$. مینیمم کردن تابع $Q(\Theta)$ با استفاده از الگوریتم نیوتن–رافسون انجام می‌شود.

۳.۳ روش ویتل

در این بخش، روشی برای برآورد پارامترها در حوزه فرکانس با استفاده از معیار ویتل [4] معرفی می‌شود. در بسیاری از موارد، محاسبه تابع چگالی طیفی، به‌ویژه برای فرآیندهای پیچیده‌تر، ساده‌تر و سریع‌تر از محاسبه تابع احتمال دقیق است. برآوردگرهای ویتل برای پارامترهای

[4] Whittle

$\Theta = (a, b, \mu_\varepsilon)$، با مینیمم‌سازی تابع زیر به دست می‌آیند

$$\hat{l}_n(\Theta) = \frac{1}{n}\sum_{k=1}^{[\frac{n}{2}]}(\log f_X(\omega_k) + \frac{I_n(\omega_k)}{f_X(\omega_k)}),$$

که در آن

$$f_X(\omega) = \frac{1}{2\pi}\frac{(1-(a+ab\mu_\varepsilon))(\gamma_X(0)(1+(a+ab\mu_\varepsilon)) + 2\gamma_X(1)\cos\omega)}{1+(a+ab\mu_\varepsilon)^2 - 2(a+ab\mu_\varepsilon)\cos 2\omega}$$

تابع چگالی طیفی فرآیند MINBL$(2, 0, 2, 1)$ در فرکانس فوریه $\omega_k = \frac{2\pi k}{n}$ و $I_n(.)$ دوره‌نگار فرآیند می‌باشند.

۴.۳ روش درستنمایی ماکزیمم نقطه زینی

روش درستنمایی ماکزیمم به دلیل پیچیدگی‌های مرتبط با عملگر رقیق‌ساز در محاسبه تابع درستنمایی

$$L_n(\Theta) = \sum_{t=1}^{n}\log f_{X_t|t-1}(x_t), \tag{1.3}$$

چالش‌برانگیز است. در ادامه با استفاده از تقریب نقطه زینی به تقریب تابع درستنمایی می‌پردازیم. لگاریتم تابع مولد تجمعی شرطی X_t، را به‌صورت $K_t(u) = \log \mathrm{E}(e^{uX_t}|t-1)$ در نظر بگیرید. تقریب نقطه زینی برای تابع چگالی شرطی (۱.۳) به‌صورت زیر می‌باشد

$$\tilde{f}_{X_t|t}(x_t) = (2\pi K_t''(\tilde{u}_t))^{-\frac{1}{2}}\exp\left\{K_t(\tilde{u}_t) - \tilde{u}_t x_t\right\}, \tag{2.3}$$

به‌طوری که مقدار u که در معادله نقطه زینی $K_t'(u) = x_t$ صدق می‌کند، با $\tilde{u}_t$ نشان داده می‌شود و K_t' و K_t'' به ترتیب مشتقات اول و دوم K_t نسبت به u هستند. محاسبه K_t مدل (۱.۲) در (۲.۳) به راحتی امکان‌پذیر نیست و با بکارگیری بسط تیلور مرتبه اول و دوم $K_t(u)$ در $u = 0$ و اندکی محاسبات می‌توان نشان داد تقریب نقطه زینی تابع چگالی شرطی به‌صورت زیر می‌باشد

$$\tilde{f}_{X_t|t-1}(X_t) = (2\pi\sigma_t^2(\Theta))^{-\frac{1}{2}}\exp\left\{\frac{-(x_t-\mu_t(\Theta))^2}{2\sigma_t^2(\Theta)}\right\}.$$

و لذا برآورد درستنمایی ماکزیمم نقطه‌زینی پارامترها از طریق ماکزیمم کردن تابع زیر بدست می‌آید

$$\tilde{L}_n(\Theta) := \sum_{t=1}^{n}\log\tilde{f}_{X_t|t-1}(x_t) = -\sum_{t=1}^{n}\frac{1}{2}\log(2\pi\sigma_t^2(\Theta)) - \sum_{t=1}^{n}\frac{(x_t-\mu_t(\Theta))^2}{2\sigma_t^2(\Theta)}.$$

که در آن μ_t و$\sigma_t^2(\Theta)$ به ترتیب میانگین شرطی و واریانس شرطی X_t هستند.

۵.۳ شبیه‌سازی

در این بخش، با استفاده از شبیه‌سازی به بررسی عملکرد چهار روش براوردیابی می‌پردازیم. هدف این مطالعه، ارزیابی عملکرد هر یک از این روش‌ها در برآورد پارامترهای مدل MINBL$(2, 0, 2, 1)$ است. مطالعات شبیه‌سازی برای پارامترهای زیر ارائه شده‌اند:

A۱: $(a, b, \mu) = (-0/1, 0/9, 7)$,

A۲: $(a, b, \mu) = (0/2, -0/5, 8)$,

A۳: $(a, b, \mu) = (-0/2, -0/7, 4)$,

A۴: $(a, b, \mu) = (0/3, 0/04, 7)$.

جدول ۱: اریبی و میانگین مربعات خطا برآوردگرهای YW، W، CLS و SPML.

n		YW			W			CLS			SPML		
		$\hat{a}$	$\hat{b}$	$\hat{\mu}$	$\hat{a}$	$\hat{b}$	$\hat{\mu}$	$\hat{a}$	$\hat{b}$	$\hat{\mu}$	$\hat{a}$	$\hat{b}$	$\hat{\mu}$
						$(a, b, \mu) = (-۰/۸, ۰/۹, ۷)$							
۱۰۰	Bias	۰/۳۱۱۴	۰/۷۶۰۷	−۰/۲۹۱۹	−۰/۰۱۰۰	۰/۰۰۰۷	۸/۹۲۱۳e − ۰۵	۰/۰۰۲۴	−۰/۰۰۳۱	−۱/۲۹۰۵e − ۰۷	۰/۰۰۰۱	۰/۰۰۱۳	
	MSE	۰/۱۲۳۹	۰/۵۹۷۸	۰/۱۱۲۶	۰/۰۰۱۰	۴/۸۲۶۲e − ۰۶	۷/۹۵۹۰e − ۰۸	۰/۰۰۰۴	۰/۰۰۰۱	۹/۹۱۶۱e − ۰۵	۱/۲۴۰۲e − ۰۷	۹/۳۰۸۹e − ۰۶	۹/۰۵۵۳e − ۱۴
۳۰۰	Bias	۰/۲۶۵۱	۰/۷۲۹۷	−۰/۳۰۰۹	−۰/۰۱۰۰	۰/۰۰۰۷	۸/۵۹۱۲e − ۰۵	۱/۶۵۴۷e − ۰۵	−۲/۷۴۴۷e − ۰۸	−۱/۱۵۲۵e − ۰۷	−۴/۹۲۷۴e − ۰۵	−۰/۰۰۰۹	
	MSE	۰/۰۹۱۴	۰/۵۶۱۰	۰/۱۱۲۷	۰/۰۰۱۰	۴/۴۷۸۴e − ۰۶	۷/۳۸۰۹e − ۰۸	۱/۳۸۶۶e − ۰۸	۲/۱۴۳۱e − ۰۶	۱/۵۳۸۹e − ۱۴	۲/۱۰۳۲e − ۰۸	۷/۶۵۴۶e − ۰۶	۹/۵۸۰۱e − ۱۴
۵۰۰	Bias	۰/۲۱۴۶	۰/۶۲۴۰	−۰/۳۵۸۳	−۰/۰۰۹۹	۰/۰۰۰۷	۸/۴۶۲۱e − ۰۵	۱/۰۶۸۲e − ۰۵	۰/۰۰۰۳	۳/۶۱۷۰e − ۰۹	۱/۷۹۹۶e − ۰۵	۵/۶۵۳۹e − ۰۵	۲/۸۷۸۱e − ۰۹
	MSE	۰/۰۷۸۱	۰/۴۳۸۳	۰/۱۷۵۷	۰/۰۰۱۰	۴/۳۴۵۶e − ۰۶	۷/۱۶۰۷e − ۰۸	۷/۹۷۲۰e − ۱۰	۱/۷۷۰۲e − ۰۷	۳/۰۷۹۱e − ۱۶	۳/۹۴۳۰e − ۰۹	۳/۷۳۸۴e − ۰۸	۱/۴۷۶۴e − ۱۶
						$(a, b, \mu) = (۰/۲, -۰/۵, ۸)$							
۱۰۰	Bias	−۰/۲۰۹۸	−۰/۱۴۲۶	−۰/۹۴۷۵	۰/۰۱۹۶	−۰/۰۰۷۷	۰/۰۰۰۵	−۰/۰۰۲۷	−۰/۰۰۶۸	۳/۴۷۶۸e − ۰۵	−۰/۰۱۷۱	−۰/۰۲۵۹	−۰/۰۰۰۲
	MSE	۰/۱۰۲۱	۰/۰۵۷۵	۰/۹۶۹۹	۰/۰۰۳۹	۰/۰۰۰۶	۲/۱۵۴۴e − ۰۶	۶/۰۹۸۷e − ۰۵	۰/۰۰۰۷	۱/۲۱۰۲e − ۰۸	۰/۰۰۲۲	۰/۰۰۵۱	۳/۱۲۸۴e − ۰۷
۳۰۰	Bias	−۰/۱۵۷۹	−۰/۱۵۵۲	−۰/۹۴۶۶	۰/۰۱۹۵	−۰/۰۰۷۴	۰/۰۰۰۴	۰/۰۰۱۰	۰/۰۰۲۱	۱/۰۲۴۱e − ۰۷	−۰/۰۰۰۹	−۰/۰۰۲۵	−۵۸۲۰۵e − ۰۷
	MSE	۰/۰۷۵۳	۰/۰۳۲۴	۰/۹۴۴۷	۰/۰۰۳۸	۰/۰۰۰۶	۱/۹۸۱۳e − ۰۶	۶/۹۶۲۰e − ۰۶	۵/۲۲۴۷e − ۰۵	۹/۶۵۹۰e − ۱۴	۶/۷۳۳۶e − ۰۶	۵/۲۴۱۰e − ۰۵	۱/۹۳۵۱e − ۱۲
۵۰۰	Bias	−۰/۱۴۰۱	−۰/۱۴۸۸	−۰/۹۳۲۴	۰/۰۱۹۳	−۰/۰۰۷۴	۰/۰۰۰۴	۱/۹۲۶۸e − ۰۵	۹/۶۰۵۱e − ۰۶	۵/۶۲۸۷e − ۱۰	۴/۶۲۹۳e − ۰۵	۶/۷۵۷۴e − ۰۵	۲/۷۳۶۵e − ۰۸
	MSE	۰/۰۴۰۷	۰/۰۳۳۶	۰/۹۰۲۹	۰/۰۰۳۷	۰/۰۰۰۵	۱/۹۳۷۳e − ۰۶	۲/۶۶۹۴e − ۰۸	۹/۴۳۳۷e − ۰۸	۱/۵۲۳۹e − ۱۷	۱/۲۷۵۳e − ۰۷	۴/۷۹۱۳e − ۰۷	۱/۷۷۵۹e − ۱۴
						$(a, b, \mu) = (-۰/۲, -۰/۷, ۴)$							
۱۰۰	Bias	۰/۱۸۹۰	−۰/۱۰۲۲	۰/۶۷۳۰	−۰/۰۰۸۱	−۰/۰۰۴۰	۰/۰۰۰۷	۰/۰۰۱۱	۰/۰۰۰۶	۰/۰۲۶	−۵/۰۵۰۲e − ۰۵	۰/۰۰۰۳	۱/۴۷۲۹e − ۰۵
	MSE	۰/۰۸۴۱	۰/۰۳۲۶	۰/۵۲۲۰	۰/۰۰۱۲	۰/۰۰۰۳	۸/۸۵۱۹e − ۰۶	۰/۰۰۰۵	۰/۰۰۱۰	۰/۰۰۰۳	۷/۵۶۱۴e − ۰۶	۰/۰۰۰۱	۲/۴۳۴۸e − ۰۸
۳۰۰	Bias	۰/۱۰۵۹	۰/۰۲۰۰	۰/۶۵۱۹	−۰/۰۰۰۲	−۰/۰۰۰۲	۳/۲۷۰۳e − ۰۵	۰/۰۰۰۲	−۰/۰۰۰۷	۰/۰۰۱۰	−۲/۱۳۷۰e − ۰۵	۳/۶۰۵۲e − ۰۵	۸/۳۶۱۳e − ۰۸
	MSE	۰/۰۳۸۳	۰/۰۲۸۶	۰/۴۷۳۷	۰/۰۰۰۲	۵/۰۷۱۳e − ۰۵	۱/۵۸۶۸e − ۰۶	۶/۰۹۵۴e − ۰۶	۶/۸۱۴۰e − ۰۵	۹/۹۳۱۵e − ۰۵	۱/۳۵۹۰e − ۰۶	۱/۹۵۶۹e − ۰۵	۱/۹۹۷۳e − ۱۱
۵۰۰	Bias	۰/۰۸۸۲	۰/۰۳۳۹	۰/۶۲۰۹	۰/۰۰۰۶	۰/۰۰۰۳	−۵/۰۴۶۴e − ۰۵	۰/۰۰۰۲	−۰/۰۰۰۷	۰/۰۰۰۷	−۲/۵۹۴۵e − ۰۵	۸/۰۴۹۴e − ۰۵	۱/۵۳۲۷e − ۰۷
	MSE	۰/۰۲۸۷	۰/۰۲۸۹	۰/۴۱۶۳	۳/۵۰۸۷e − ۰۵	۸/۰۸۶۰e − ۰۶	۲/۵۴۶۷e − ۰۷	۴/۲۶۱۵e − ۰۶	۴/۶۸۰۴e − ۰۵	۵/۱۴۰۱e − ۰۵	۳/۶۱۷۹e − ۰۷	۴/۷۵۶۸e − ۰۶	۹/۳۲۵۶e − ۱۳
						$(a, b, \mu) = (۰/۳, ۰/۰۴, ۷)$							
۱۰۰	Bias	۰/۰۸۲۸	−۰/۰۸۲۶	۰/۱۲۷۵	−۰/۰۰۰۷	−۰/۰۰۱۲	−۶/۵۴۶۲e − ۰۶	−۰/۰۵۶۶	۰/۰۱۸۶	−۰/۰۲۸۳	−۰/۴۵۴۲	−۰/۱۴۳۹	−۰/۰۲۱۴
	MSE	۰/۰۱۲۸	۰/۰۱۸۳	۰/۰۲۵۰	۵/۰۲۷۷e − ۰۶	۱/۳۵۹۶e − ۰۵	۴/۲۸۵۲e − ۱۰	۰/۰۰۶۹	۰/۰۰۰۸	۰/۰۰۳۰	۰/۲۱۹۵	۰/۰۲۱۲	۰/۰۰۴۸
۳۰۰	Bias	۰/۰۱۸۷	−۰/۰۰۸۵	۰/۱۰۴۷	−۰/۰۰۰۵	−۰/۰۰۰۸	−۴/۳۷۷۹e − ۰۶	−۰/۰۲۶۵	۰/۰۱۶۹	−۰/۰۱۱۲	−۰/۴۱۲۲	−۰/۱۳۳۹	−۰/۰۰۵۸
	MSE	۰/۰۰۱۷	۰/۰۰۰۸	۰/۰۱۴۷	۲/۱۹۵۷e − ۰۶	۵/۹۱۵۶e − ۰۶	۱/۹۱۶۶e − ۱۰	۰/۰۰۱۸	۰/۰۰۰۴	۰/۰۰۲۵	۰/۱۷۳۰	۰/۰۱۸۲	۰/۰۰۲۹
۵۰۰	Bias	۰/۰۱۸۴	−۰/۰۱۲۹	۰/۰۸۶۵	−۰/۰۰۰۳	−۰/۰۰۰۵	−۲/۷۴۴۲e − ۰۶	−۰/۰۲۷۰	۰/۰۱۷۲	۰/۰۰۲۵	−۰/۳۸۹۵	−۰/۱۲۷۵	۰/۰۰۶۴
	MSE	۰/۰۰۱۴	۰/۰۰۰۷	۰/۰۱۳۰	۸/۵۷۵۵e − ۰۷	۲/۳۰۸۳e − ۰۶	۷/۵۳۰۵e − ۱۱	۰/۰۰۱۲	۰/۰۰۰۴	۰/۰۱۶	۰/۱۵۷۸	۰/۰۱۶۴	۰/۰۰۳۴

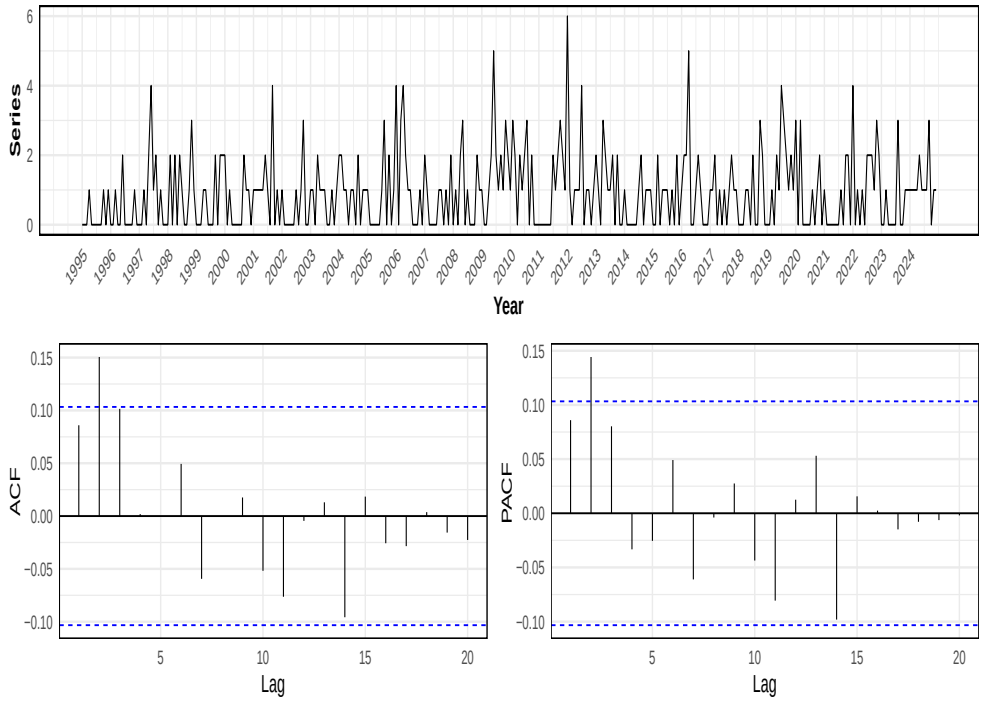
عملگرهای $a\circledast$ و $b\circledast$ عملگر رقیق‌ساز دو جمله‌ای تعمیم‌یافته علامت و سری‌های شمارشی در $a\circledast$ و $b\circledast$ دارای توزیع برنولی به ترتیب با پارامترهای $|a|$ و $|b|$ و همچنین ε_t دارای توزیع پواسن با پارامتر μ می‌باشد. جهت بررسی کارایی چهار روش برآوردیابی، شبیه‌سازی‌هایی با اندازه‌های نمونه‌ای $n = ۱۰۰, ۳۰۰, ۵۰۰$ از مدل با پارامترهای A۱ - A۴ انجام و عملکرد برآوردها با $s = ۱۰۰$ تکرار بررسی شده است. نتایج شبیه‌سازی بر اساس معیارهای Bias و MSE در جدول ۱ ارائه شده است. همان‌طور که در جدول ۱ مشاهده می‌شود، هر چهار روش برآورد از نظر اریبی و میانگین مربعات خطا عملکرد رضایت‌بخشی دارند و برآوردهای ویتل بهترین عملکرد را دارند.

۴ تحلیل داده‌های واقعی

در این بخش، به کاربرد مدل MINBL(۲, ۰, ۲, ۱) پرداخته می‌شود. داده‌ها مربوط به تعداد گزارش‌های ماهانه پلیس در مورد خرید و فروش اموال سرقت‌شده در گانِدا [5]، استرالیا است. این اطلاعات به‌صورت ماهانه از ژانویه ۱۹۹۵ تا دسامبر ۲۰۲۴ جمع‌آوری شده است و داده‌ها از سایت bocsar [6]، گرفته شده است. در نمودار ۱ مسیر نمونه‌ای، ACF و PACF برای داده‌ها نمایش داده شده است. در ادامه،

[5] Gunnedah

[6] http://www.bocsar.nsw.gov.au



شکل ۱: مسیر نمونه‌ای، تابع خودهمبستگی و تابع خودهمبستگی جزئی داده‌های مربوط به خرید و فروش اموال سرقت شده.

ایستایی داده‌ها با استفاده از آزمون دیکی–فولر افزایشی تأیید شده است که $p-$مقدارهای آن برابر با ۰٫۰۱ بوده است. این نتیجه بیان می‌کند که داده‌ها ایستا هستند. همچنین، آزمون غیرخطی کینان که $p-$مقدارهای ۰٫۰۰۰ را نشان می‌دهد، غیرخطی بودن داده‌ها را تأیید می‌کند. از سوی دیگر، داده فوق با شاخص‌های پراکندگی $\hat{I}_X = ۱٫۶۵۰$ بیش‌پراکنده می‌باشند. مدل MINBL(2,0,2,1) به‌عنوان مدلی مناسب برای برازش این مجموعه‌داده‌ها معرفی می‌شود، همان‌طور که در نمودار ۱ مشاهده می‌شود. در ادامه، پارامترهای مدل با در نظر گرفتن توزیع‌های مختلف از جمله پواسن، پواسن–لیندلی، هندسی و هندسی اصلاح‌شده برآورد شده‌اند. جدول ۲ شامل برآوردهای ویتل به همراه مقادیر ریشه میانگین مربعات خطا (RMSE)، میانگین قدرمطلق خطا (MAE) ، میانگین مربعات خطای پیش‌بینی (FMSE) و میانگین قدرمطلق خطای پیش‌بینی (FMAE) هستند. برای محاسبه FMSE، و FMAE، از پیش‌بینی درون‌نمونه‌ای ۱۰–گام به جلو استفاده شده است. نتایج جدول ۲ نشان می‌دهند که مدل PL-MINBL مقادیر کمتری از RMSE، MAE ، FMSE و FMAE را به‌دست می‌آورد، که حاکی از عملکرد قوی این مدل برای داده است.

۱.۴ پیش‌بینی

در این بخش به پیش‌بینی مقادیر آینده می‌پردازیم. همان‌طور که قبلا اشاره شد، پیش‌بینی کلاسیک $k-$گام به جلو از طریق امیدریاضی شرطی (۳.۲) و (۴.۲) به‌صورت زیر محاسبه می‌شود

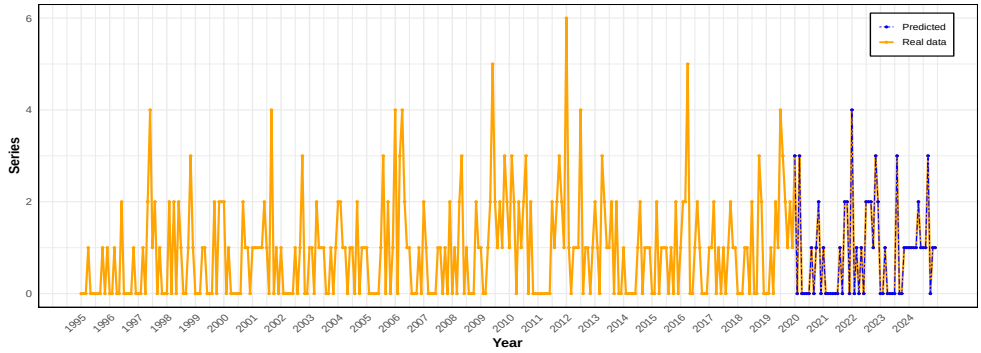
$$\hat{X}_{t+۱} = (a + ab\varepsilon_t)X_{t-۱} + \mu_\varepsilon, \;\; \hat{X}_{t+k} = (a + ab\mu_\varepsilon)X_{t+k-۲} + \mu_\varepsilon.$$

که در عمل پارامترهای a، b و μ_ε با برآوردهای مربوطه جایگزین می‌گردد.

پیش‌بینی بوت‌استرپ

جدول ۲: برآورد پارامترها، RMSE، MAE ، FMSE، و FMAE داده‌های مربوط به اموال سرقت شده.

FMAE	FMSE	MAE	RMSE	برآورد ویتل	مدل
۰٫۵۲۰۲۹۰۷	۰٫۷۰۰۲۷۴۷	۰٫۸۰۳۹۰۱۸	۱٫۰۵۰۷۷۵	$\hat{a} = ۰٫۰۳۱$ $\hat{b} = ۱٫۹۷۲$ $\hat{\mu} = ۰٫۷۶۵$	P-MINBL(۲, ۰, ۲, ۱)
۰٫۵۱۳۱۴۱۴	۰٫۶۸۹۷۵۱	۰٫۷۹۹۵۴۶۶	۱٫۰۴۹۱۸۳	$\hat{a} = ۰٫۰۶۰$ $\hat{b} = ۰٫۳۷۶$ $\hat{\theta} = ۱٫۷۷۶$	PL-MINBL(۲, ۰, ۲, ۱)
۰٫۵۱۶۱۹۴۴	۰٫۶۹۳۳۱۸۲	۰٫۸۰۰۵۲۱	۱٫۰۴۹۳۵۱	$\hat{a} = ۰٫۰۵۱$ $\hat{b} = ۰٫۶۷۲$ $\hat{\mu} = ۰٫۷۶۵$	NG-MINBL(۲, ۰, ۲, ۱)
۰٫۵۳۰۲۱۶۶	۰٫۷۰۵۴۰۵۶	۰٫۸۰۱۶۲۲۹	۱٫۰۵۰۵۶۹	$\hat{a} = ۰٫۰۵۲$ $\hat{b} = ۰٫۶۷۲$ $\hat{p} = ۰٫۵۷۳$	G-MINBL(۲, ۰, ۲, ۱)



شکل ۲: مقادیر واقعی در مقابل مقادیر پیش‌بینی‌شده به روش بوت‌استرپ مربوط به داده‌های اموال سرقت شده در پنج سال گذشته.

در این قسمت از تکنیک بوت‌استرپ اصلاح‌شده برای پیش‌بینی استفاده می‌کنیم. این روش به‌طور خاص برای مدل MINBL(۲, ۰, ۲, ۱) با تنظیمات مشخص اعمال شده و در الگوریتم ۱ به تفصیل شرح داده شده است. برای ارزیابی قابلیت پیش‌بینی این روش در طول زمان، از رویکرد بوت‌استرپ برای پیش‌بینی پنج سال پایانی از داده‌های موجود استفاده شد. نتایج نشان می‌دهند که مقادیر پیش‌بینی‌شده شباهت زیادی به سری‌های اصلی دارند و این موضوع تأیید می‌کند که مدل MINBL(۲, ۰, ۲, ۱) برای پیش‌بینی‌های دقیق در مجموعه داده مؤثر است.

الگوریتم ۱

گام اول: پارامترهای a، b و μ_ε را با استفاده از روش ویتل برآورد کنید.

گام دوم: باقیمانده‌های فرآیند $\hat{\varepsilon}_t = X_t - \hat{a}X_{t-۲} - \hat{a}\hat{b}X_{t-۲}\hat{\varepsilon}_{t-۱}$ را برای $t = ۳, \ldots, n$ تعیین کنید.

گام سوم: توزیع تجربی از باقیمانده‌های تعدیل شده $\tilde{\varepsilon}_t = [\hat{\varepsilon}_t]$ زمانی‌که $\hat{\varepsilon}_t > ۰$، در غیر اینصورت $\tilde{\varepsilon}_t = ۰$.

گام چهارم: داده‌های بوت‌استرپ X_t^b برای $b = ۱, \ldots, B$ به‌صورت زیر شبیه‌سازی شده است.

$$X_t^b = \hat{a} \circledast X_{t-۲}^b + (\hat{a} \circledast X_{t-۲}^b)(\hat{b} \circledast \tilde{\varepsilon}_{t-۱}^b) + \tilde{\varepsilon}_t^b,$$

که در آن $\tilde{\varepsilon}_t^b$ برای $t = ۱, ۲, \ldots, n$ در گام سوم شبیه‌سازی شده است.

گام پنجم: بر اساس داده های شبیه‌سازی شده X_t^b پارامترهای $\hat{a}^b$، $\hat{b}^b$ و $\hat{\mu}_\varepsilon^b$ مشابه گام اول محاسبه می‌شود.

گام ششم: پارامترهای a، b و μ_ε برای B تکرار بر اساس میانگین‌های نمونه‌ای به‌صورت زیر براورد شود:

$$\hat{a}^* = \frac{\sum_{b=۱}^{B} \hat{a}^b}{B}, \ \hat{b}^* = \frac{\sum_{b=۱}^{B} \hat{b}^b}{B}, \ \hat{\mu}_\varepsilon^* = \frac{\sum_{b=۱}^{B} \hat{\mu}_\varepsilon^b}{B}.$$

گام هفتم: پیش‌بینی $k-$گام به جلو بر اساس معادلات بازگشتی زیر محاسبه می‌گردد

$$X_{t+k}^b = \hat{a}^* \circledast X_{t+k-۲}^b + (\hat{a}^* \circledast X_{t+k-۲}^b)(\hat{b}^* \circledast \tilde{\varepsilon}_{t+k-۱}^b) + \tilde{\varepsilon}_{t+k}^b, \ \ k \geq ۱.$$

این طرح با استفاده از اعتبار ویژه پژوهشی دانشگاه مازندران اجرا گردید.

مراجع

Bamdadi, R., Mohammadpour, M., and Ramezani, S. (2024). A bilinear modeling in counts time series with applications, *Communications in Nonlinear Science and Numerical Simulation*, **139**, 108282.

Doukhan, P., Latour, A., and Oraichi, D. (2006). A Simple Integer-Valued Bilinear Time Series Model, *Advances in Applied Probability*, **38**, 559–578.

Scotto, M. G., Weiß, C. H., and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review, *Statistical Modelling*, **15**, 590–618.

Weiß, C. H. (2008). Thinning operations for modeling time series of counts—a survey, *Advances in Statistical Analysis*, **92**, 319–341.

Zhang, H., Wang, D., and Zhu, D. (2010). Inference for INAR(p) processes with signed generalized power series thinning operator, *Journal of Statistical Planning and Inference*, **140**, 667–683.

A class of integer-valued bilinear time series model with applications

Rana Bamdadi and Mehrnaz Mohammadpour

Department of Statistics, University of Mazandaran

Abstract: This study presents a specific class of integer-valued bilinear model of order 2 for nonlinear time series. These models possess the flexibility to incorporate the new individuals into the model in a way that allows them to behave independently. The modeling methodology applies to processes characterized by overdispersion, relying on a non-linear framework. The main characteristics of the proposed model are established. For computational purposes, some approaches for parameter estimation are carried out. To evaluate the performance of them, simulation experiments are conducted. To illustrate the modeling strategy, a social science cases is analyzed.

Keywords: Forecasting, Integer-valued bilinear process, Saddlepoint approximation, Whittle method.

Mathematics Subject Classification (2020): 62P10، 62M10، 60G25.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مدل بندی قابلیت اطمینان سیستم های تنش ـ مقاومت بر اساس رویکرد تابع مولد عام

بهناز تصدیقی، دکتر رضا زارعی، دکتر بهروز فتحی [۱]

[۱]گروه آمار، دانشکده علوم ریاضی، دانشگاه گیلان، رشت، ایران

چکیده: مدل تداخل تنش–مقاومت پیوسته (SSI) تنش و مقاومت را به عنوان متغیرهای تصادفی پیوسته با تابع چگالی احتمال شناخته شده در نظر می‌گیرد. این امر تا حدی منجر به محدودیت در کاربرد آن می‌شود. در این مقاله، تنش و مقاومت به عنوان متغیرهای تصادفی گسسته در نظر گرفته می‌شوند و یک مدل SSI گسسته با استفاده از روش تابع مولد عام ارائه می‌شود. در نهایت، مطالعات موردی اعتبار مدل گسسته را در شرایط مختلف نشان می‌دهد، که در آن تنش و مقاومت را می‌توان با متغیرهای تصادفی پیوسته، متغیرهای تصادفی گسسته یا دو گروه از داده‌های تجربی نشان داد.

واژه‌های کلیدی: قابلیت اطمینان؛ تابع مولد عام؛ مدل تداخل تنش–مقاومت؛ مدل گسسته.

کد موضوع‌بندی ریاضی (۲۰۲۰): 90B25، 60E10، 62N05.

---

۱ مقدمه

مدل تداخل تنش–مقاومت (SSI) به‌طور گسترده‌ای برای طراحی قابلیت اطمینان اجزای مکانیکی استفاده شده‌است. در این مدل، اگر تنش روی یک جزء و مقاومت یک جزء به ترتیب با S_1 و S_2 نشان داده شود، قابلیت اطمینان جزء که با R نشان داده شده‌است، به صورت زیر تعریف می‌شود.

$$R = \Pr(S_2 > S_1) \tag{1.1}$$

معادله (۱.۱) اساسی ترین بیان مدل تداخل تنش–مقاومت است، به این معنی که قابلیت اطمینان جزء به عنوان احتمال بزرگتر بودن مقاومت از تنش در نظر گرفته می‌شود. علاوه‌براین، اگر هر دو تنش–مقاومت به عنوان متغیرهای تصادفی پیوسته و توابع چگالی

[۱] سخنران، behnaztasdighi77@email.com

احتمال آنها که به ترتیب با $f_1(S_1)$ و $f_2(S_2)$ نشان داده می‌شوند، در نظر گرفته شوند، معادله (۱.۱) را می‌توان به صورت فرمول‌های زیر بازنویسی کرد

$$R = \int_{-\infty}^{\infty} f_1(S_1) \left[\int_{S_1}^{\infty} f_2(S_2) dS_2 \right] dS_1 \qquad (۲.۱آ)$$

$$R = \int_{-\infty}^{\infty} f_2(S_2) \left[\int_{-\infty}^{S_2} f_1(S_1) dS_1 \right] dS_2 \qquad (۲.۱ب)$$

برای وضوح، معادله (۲.۱) را می‌توان مدل تداخل تنش–مقاومت پیوسته نامید.

از نظر تئوری، اگر تابع چگالی احتمال تنش و مقاومت موجود باشد، قابلیت اطمینان مؤلفه را می‌توان به‌صورت تحلیلی یا عددی محاسبه کرد. اما در عمل، اغلب توزیع دقیق تنش و مقاومت معلوم نیست و تنها داده‌های آزمایشی محدودی درباره آن‌ها در دسترس است. بنابراین، بررسی روش‌های تقریبی برای محاسبه قابلیت اطمینان مؤلفه‌ها ضروری بوده و پژوهش‌های زیادی در این زمینه انجام شده است.

کاپور (۱) رویکردی برای تعیین مرزهای غیرقابل اطمینان دقیق ابداع کرد و این رویکرد فقط به اطلاعاتی در مورد احتمالات زیربازه‌ای در یک منطقه تداخلی نیاز داشت. پارک و کلارک (۲) برای بهبود دقت محاسبات، فرمول کاپور را در مورد مسئله برنامه‌ریزی درجه دوم اصلاح کردند و راه حلی برای این مشکل ارائه کردند. وانگ و لیو (۳) رویکردی را برای محاسبه عدم اطمینان فازی یک جزء/سیستم ارائه کردند. علاوه بر این، کوتز و همکاران (۴) با خلاصه کردن نتایج تحقیق از رشته‌های مختلف، مدل تنش–مقاومت را تعمیم دادند و روش‌های محاسباتی را بر اساس تخمین حداکثر احتمال ارائه کردند.

در این مقاله، برخلاف مدل‌های پیوسته، تنش و مقاومت به‌صورت متغیر تصادفی گسسته در نظر گرفته شده و توابع جرم احتمال آن‌ها با استفاده از توابع مولد عام (UGF) نمایش داده می‌شوند. با توجه به تعریف قابلیت اطمینان مؤلفه، یک مدل تداخل تنش–مقاومت گسسته ارائه شده که برای محاسبه قابلیت اطمینان در سه حالت زیر کاربرد دارد:

۱. تنش و مقاومت متغیر تصادفی گسسته باشند،

۲. تنش و مقاومت متغیر تصادفی پیوسته باشند،

۳. توزیع تنش و مقاومت در دسترس نباشد اما توزیع فراوانی بر اساس داده‌های آزمایش موجود باشد.

در ادامه در بخش دوم روش UGF مورد بررسی قرار می‌گیرد، در بخش سوم مدل گسسته SSI معرفی می‌شود و در بخش چهارم یک مثال کاربردی ارائه می‌شود.

۲ رویکرد تابع مولد عام در متغیرهای تصادفی گسسته

تعریف ۱.۲. فرض کنید که یک متغیر تصادفی گسسته X دارای یک تابع جرم احتمال باشد که با بردار x متشکل از مقادیر ممکن X و بردار p متشکل از احتمالات مربوطه مشخص می‌شود که می‌توان با عبارات زیر فرمول‌بندی کرد:

$$x = (x_1, x_2, \cdots, x_k)\ ,\ p = (p_1, p_2, \cdots, p_k)\ ,\ p_i = \Pr(X = x_i)\ , \quad i = 1, 2, \cdots, k$$

بر اساس اصل اساسی روش UGF، تابع جرم احتمال متغیر تصادفی گسسته X را می‌توان با یک تابع چند جمله‌ای از متغیر z ، $u_X(z)$ نشان داد که مقادیر ممکن X را به احتمالات مربوطه مرتبط می‌کند.

$$u_X(z) = p_1 z^{x_1} + p_2 z^{x_2} + \cdots + p_k z^{x_k} = \sum_{i=1}^{k} p_i z^{x_i} \tag{1.2}$$

لازم به ذکر است که برای یک متغیر تصادفی گسسته دلخواه، UGF آن منحصراً توسط تابع جرم احتمال آن تعیین می‌شود. این بدان معنی است که یک مکاتبه یک به یک بین تابع جرم احتمال و UGF یک متغیر تصادفی گسسته وجود دارد.

بدون از دست دادن کلیت، n متغیر تصادفی گسسته مستقل $X_1, X_2, \cdots, X_n$ و یک تابع دلخواه $f(X_1, X_2, \cdots, X_n)$ رامی‌توان در نظر گرفت.

تعریف ۲.۲. فرض کنید تعداد مقادیر ممکن هر متغیر تصادفی به ترتیب $k_1, k_2, \cdots, k_n$ باشد. با توجه به معادله (۱.۲)،UGF متغیر تصادفی تکی را می‌توان به صورت زیر به‌دست آورد:

$$\begin{aligned}
u_{X_1}(z) &= \sum_{j_1=1}^{k_1} p_{1j_1} z^{x_{1j_1}}, \\
u_{X_2}(z) &= \sum_{j_2=1}^{k_2} p_{2j_2} z^{x_{2j_2}}, \\
&\vdots \\
u_{X_n}(z) &= \sum_{j_n=1}^{k_n} p_{nj_n} z^{x_{nj_n}}.
\end{aligned}$$

تعریف ۳.۲. برای به‌دست آوردن UGF تابع $f(X_1, X_2, \cdots, X_n)$ ،عملگر ترکیب $\otimes$ به صورت زیر تعریف می‌شود:

$$\begin{aligned}
&\bigotimes \left(\sum_{j_1=1}^{k_1} p_{1j_1} z^{x_{1j_1}}, \sum_{j_2=1}^{k_2} p_{2j_2} z^{x_{2j_2}}, \cdots, \sum_{j_n=1}^{k_n} p_{nj_n} z^{x_{nj_n}} \right) \\
&= \sum_{j_1=1}^{k_1} \sum_{j_2=1}^{k_2} \cdots \sum_{j_n=1}^{k_n} \left(\prod_{i=1}^{n} p_{ij_i} z^{f\left(x_{1j_1}, x_{2j_2}, \cdots, x_{nj_n}\right)} \right).
\end{aligned}$$

در واقع، عملگر ترکیب $\otimes$ یک قانون عملیات را نشان می‌دهد که به شدت به بیان تابع $f(X_1, X_2, \cdots, X_n)$ بستگی دارد. بر اساس این قاعده می‌توان UGF تابع تصادفی $f(X_1, X_2, \cdots, X_n)$ را به‌دست‌آورد که شکلی از تابع چند جمله‌ای نیز دارد و با تابع جرم احتمال تابع $f(X_1, X_2, \cdots, X_n)$ مطابقت دارد.لازم به ذکر است که UGF ها علی‌رغم شباهت به چند جمله‌ای‌ها، چند جمله‌ای‌های منظم نیستند. با این حال، UGF ها خواص اساسی چند جمله‌ای‌های منظم را به ارث می‌برند. به عنوان مثال، در عملیات UGF ، اصطلاحات مشابهی را می‌توان جمع آوری کرد، و قانون جایگزین و قانون انجمنی قابل اجرا هستند:

$$\begin{aligned} u_f(z) &= \bigotimes\left(u_{X_1}(z),\cdots,u_{X_i}(z),u_{X_{i+1}}(z),\cdots,u_{X_n}(z)\right) \\ &= \bigotimes\left(u_{X_1}(z),\cdots,u_{X_{i+1}}(z),u_{X_i}(z),\cdots,u_{X_n}(z)\right), \end{aligned}$$

$$\begin{aligned} u_f(z) &= \bigotimes\left(u_{X_1}(z),\cdots,u_{X_i}(z),u_{X_{i+1}}(z),\cdots,u_{X_n}(z)\right) \\ &= \bigotimes\left(u_{X_1}(z)\right),\cdots,u_{X_i}(z)),\bigotimes\left(u_{X_{i+1}}(z),\cdots,u_{X_n}(z)\right). \end{aligned}$$

هنگامی که UGF تابع تصادفی $f(X_1,X_2,\cdots,X_n)$ به دست آمد، می‌توان آن را به عنوان یک متغیر تصادفی جدید در نظر گرفت و ویژگی‌های آماری آن را تحلیل کرد.

۳ مدل گسسته SSI

تعریف ۱.۳. فرض کنید که تنش‌های وارد بر یک جزء و مقاومت یک جزء دو متغیر تصادفی گسسته مستقل هستند که به ترتیب با S_1 و S_2 نشان داده می‌شوند. اگرتابع جرم احتمال تنش-مقاومت به صورت زیر شناخته شوند:

$$S_1=\left(S_{11},S_{12},\cdots,S_{1k_1}\right),P_1=\left(P_{11},P_{12},\cdots,P_{1k_1}\right)$$

$$S_2=\left(S_{21},S_{22},\cdots,S_{2k_2}\right),P_1=\left(P_{21},P_{22},\cdots,P_{2k_2}\right)$$

که در آن k_1 و k_2 تعداد مقادیر ممکنی هستند که به ترتیب S_1 و S_2 می‌توانند دریافت کنند، سپس، با توجه به معادله (۱.۲)، UGFتنش-مقاومت را می‌توان به صورت زیر به‌دست آورد:

$$u_{S_1}(z)=\sum_{j_1=1}^{k_1}p_{1j_1}z^{S_{1j_1}}\ ,\quad u_{S_2}(z)=\sum_{j_2=1}^{k_2}p_{2j_2}z^{S_{2j_2}}$$

یک تابع $f(S_1,S_2)$ از متغیر تصادفی تنش-مقاومت به‌صورت زیر ساخته می‌شود

$$f(S_1,S_2)=S_2-S_1 \tag{۱.۳}$$

بر اساس روش UGF معرفی شده در بخش ۲، UGF تابع گسسته $f(S_1,S_2)$ را می‌توان به‌صورت زیر به‌دست‌آورد

$$\begin{aligned} u_f(z) &= \bigotimes\left(u_{S_1}(z),u_{S_2}(z)\right)=\bigotimes\left(\sum_{j_1=1}^{k_1}p_{1j_1}z^{S_{1j_1}},\sum_{j_2=1}^{k_2}p_{2j_2}z^{S_{2j_2}}\right) \\ &= \sum_{j_1=1}^{k_1}\sum_{j_2=1}^{k_2}\left(\prod_{i=1}^{2}p_{ij_i}z^{f\left(S_{1j_1},S_{2j_2}\right)}\right)=\sum_{j_1=1}^{k_1}\sum_{j_2=1}^{k_2}\left(\prod_{i=1}^{2}p_{ij_i}z^{\left(S_{2j_2}-S_{1j_1}\right)}\right) \end{aligned} \tag{۲.۳}$$

به عنوان یک حالت قطعی، عملگر $\otimes$ در معادله (۲.۳) به عنوان تفریق توان‌های متناظر در چندجمله‌ای‌های ضرب تعریف می‌شود. درک اینکه شکل نهایی UGF تابع $f(S_1, S_2)$ یک تابع چند جمله‌ای است که شامل $K \leq k_1 \times k_2$ عبارت است (تعداد کل عبارت‌ها می‌تواند کمتر از $k_1 \times k_2$ باشد پس از جمع‌آوری عبارت‌های مشابه) دشوار نیست. بنابراین، معادله (۲.۳) را می توان به صورت زیر بازنویسی کرد

$$u_f(z) = \sum_{j=1}^{k} P_j z^{f_j} \tag{۳.۳}$$

که در آن f_j و $P_j \quad (j = 1, 2, \cdots, k)$ به ترتیب مقادیر ممکن تابع $f(S_1, S_2)$ و احتمالات مربوطه هستند.

همانطور که با معادله (۱.۱) بیان می‌شود، قابلیت اطمینان جزء به عنوان احتمال بزرگتر بودن مقاومت از تنش تعریف می‌شود. با تبدیل معادله (۱.۱) ، می‌توان به‌دست آورد

$$R = \Pr(S_2 - S_1 > 0) \tag{۴.۳}$$

جایگزینی معادله (۱.۳) به معادله (۴.۳) نتیجه می‌دهد

$$R = \Pr(f(S_1, S_2) > 0) \tag{۵.۳}$$

که در آن $f(S_1, S_2)$ اساساً، یک متغیر تصادفی گسسته جدید است و خواص توزیع آن را می‌توان با UGF آن نشان داد.

برای محاسبه احتمال بیان شده توسط معادله (۵.۳) بر اساس UGF تابع $f(S_1, S_2)$ می‌توان یک تابع با ارزش دودویی با دامنه روی مجموعه مقادیر ممکن تابع $f(S_1, S_2)$ تعریف کرد.

$$\alpha(f_i) = \begin{cases} 1, & f_i > 0, \\ 0, & f_i \leq 0. \end{cases}$$

سپس، بر اساس معادله (۵.۳) ، قابلیت اطمینان جزء را می‌توان به صورت زیر محاسبه کرد

$$R = \Pr(f(S_1, S_2) > 0) = \sum_{j=1}^{K} P_j \alpha(f_j) \tag{۶.۳}$$

برای مقایسه، معادله (۶.۳) را می‌توان مدل تداخل تنش–مقاومت گسسته نامید.

۴ مثال کاربردی

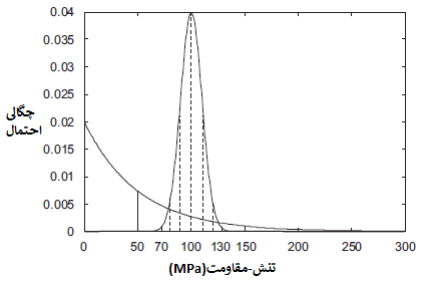
فرض کنید مدل تداخل تنش–مقاومت گسسته ارائه شده در بخش ۳ می‌تواند برای ارزیابی قابلیت اطمینان یک جزء در چندین مورد استفاده شود. بدیهی است که وقتی تنش–مقامت متغیر تصادفی گسسته هستند، قابلیت اطمینان جزء را می‌توان مستقیماً مطابق با معادلات (۲.۳) ، (۳.۳) و (۶.۳) به‌دست آورد. در این بخش دو مورد در نظر گرفته شده است. مورد ۱ نشان می دهد که تنش–مقاومت متغیر تصادفی پیوسته با تابع چگالی احتمال شناخته شده هستند و مورد ۲ نشان می دهد که تنها توزیع فراوانی تنش–مقاومت بر اساس تجزیه و تحلیل آماری داده ها در دسترس است.

مورد ۱. مثال (۳) را در نظر بگیرید: تنش بر روی یک جزء، $S_۱$ ، به صورت نمایی با میانگین $\mu_۱ = ۵۰MPa$ توزیع می‌شود، و مقاومت مولفه، $S_۲$ ، به طور s–نرمال با میانگین $\mu_۲ = ۱۰۰MPa$ و انحراف استاندارد $\sigma_۲ = ۱۰MPa$ توزیع می‌شود. مقدار دقیق قابلیت اطمینان این جزء برابر با ۰٫۸۶۱۹۴ است.

در این روش، برای محاسبه قابلیت اطمینان مؤلفه، متغیر تصادفی پیوسته با تابع چگالی احتمال شناخته‌شده به یک متغیر تصادفی گسسته با تابع جرم احتمال تبدیل می‌شود. در مرحله اول، با توجه به شرایط عملیاتی قطعه، محدوده تقریبی مقادیر ممکن برای تنش و مقاومت تعیین می‌شود که به‌ترتیب با فواصل $\langle S_{۱min}, S_{۱max}\rangle$ و $\langle S_{۲min}, S_{۲max}\rangle$ مشخص می‌شوند. سپس فواصل $\langle S_{۱min}, S_{۱max}\rangle$ و $\langle S_{۲min}, S_{۲max}\rangle$ به ترتیب به زیر بازه های m و n تقسیم می‌شوند ($m = n$ مجاز است) . مقادیر نقطه میانی هر زیر بازه به عنوان مقادیر ممکن متغیر تصادفی و مساحت هر زیر بازه به عنوان احتمال آن در نظر گرفته می‌شوند. با این روش، دو متغیر تصادفی مجزا برای تنش و مقاومت با تابع جرم احتمال مشخص به دست می‌آید. سپس متغیر تصادفی گسسته تنش–مقاومت با استفاده از روش UGF نمایش داده شده و قابلیت اطمینان جزء با مدل تداخل تنش–مقاومت گسسته ارائه شده در بخش ۳ محاسبه می‌شود.

بگذارید دامنه تنش–مقاومت به ترتیب $\langle ۰, ۶\mu_۱\rangle = \langle ۰, ۳۰۰\rangle MPa$ و $\langle \mu_۲ - ۳\sigma_۲, \mu_۲ + ۳\sigma_۲\rangle = \langle ۷۰, ۱۳۰\rangle MPa$ باشد و هر دو بازه را به شش زیر بازه تقسیم کنید. نتیجه تقسیم بندی بازه‌ای در شکل (۱) نشان داده‌شده‌است. در این مورد، مقادیر مساحت زیربازه‌ها با استفاده از *MATLAB* ۱.۰.۷ محاسبه می‌شود. بر اساس مقادیر نقطه میانی و مقادیر مساحت همه زیر بازه‌ها، می‌توان متغیر تصادفی گسسته تنش–مقاومت را به ترتیب به صورت زیر به‌دست‌آورد:

$$
\begin{aligned}
S_۱ &= (۲۵, ۷۵, ۱۲۵, ۱۷۵, ۲۲۵, ۲۷۵)\ ,\ S_۲ = (۷۵, ۸۵, ۹۵, ۱۰۵, ۱۱۵, ۱۲۵) \\
p_۱ &= (۰٫۶۳۲۱, ۰٫۲۳۲۵, ۰٫۰۸۵۵, ۰٫۰۳۱۵, ۰٫۰۱۱۶, ۰٫۰۰۴۳) \\
p_۲ &= (۰٫۰۲۱۴, ۰٫۱۳۵۹, ۰٫۳۴۱۳, ۰٫۱۳۵۹, ۰٫۰۲۱۴)
\end{aligned}
$$



شکل ۱: تابع چگالی احتمال منحنی‌ها و تقسیم بندی‌بازه‌ای

با توجه به روش ارائه‌شده در بخش ۳، قابلیت اطمینان این مولفه به صورت $R = ۰٫۸۶۲۵۰$ به‌دست‌می‌آید. خطای نسبی در مقایسه با مقدار دقیق قابلیت اطمینان ۰٫۰۶۵ است. به منظور نشان دادن تأثیر تقسیم‌بندی بازه‌ای بر دقت محاسبه، می‌توان بازه تنش را به ۱۲ زیربازه تقسیم کرد در حالی که تعداد زیربازه‌های مقاومت را بدون تغییر نگه‌داشت. بنابراین، می‌توان توصیف دیگری از تنش به‌دست‌آورد:

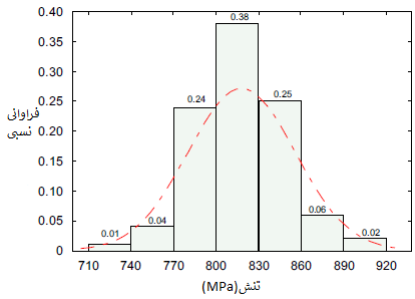
$$S_۱ = (۱۲٫۵۳۷٫۵, ۶۲٫۵, ۸۷٫۵, ۱۱۲٫۵, ۱۳۷٫۵, ۱۶۲٫۵, ۱۸۷٫۵, ۲۱۲٫۵, ۲۳۷٫۵, ۲۶۲٫۵, ۲۸۷٫۵)$$

$$P_۱ = (۰٫۳۹۵۳, ۰٫۲۳۸۷, ۰٫۱۴۴۷, ۰٫۰۸۷۸, ۰٫۰۵۳۳, ۰٫۰۳۲۳, ۰٫۰۱۹۶, ۰٫۰۱۱۹, ۰٫۰۰۷۲,$$
$$۰٫۰۰۴۴, ۰٫۰۰۲۷, ۰٫۰۰۱۶)$$

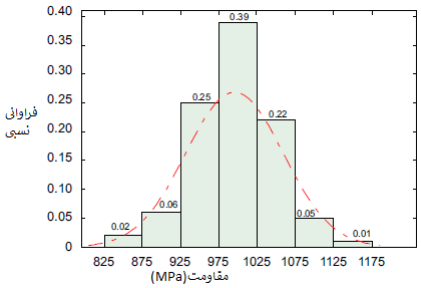
به‌طورمشابه، پایایی این مؤلفه را می‌توان به‌صورت $R = ۰٫۸۶۱۶۳$ محاسبه کرد و خطای نسبی برابر با ۰٫۰۳۶ است. این نتیجه نشان می‌دهد که کاهش طول زیربازه می‌تواند دقت محاسباتی را زمانی که دامنه تنش–مقاومت ثابت است، بهبود بخشد.

مورد ۲. برای محاسبه قابلیت اطمینان یک جزء، اگر مدل تداخل تنش–مقاومت پیوسته استفاده شود، تنش و مقاومت به‌صورت متغیر تصادفی پیوسته در نظر گرفته شده و توزیع آن‌ها با روش‌های برازش توزیع و تخمین پارامتر به‌دست می‌آید. اما در مدل گسسته، نیازی به توجه به توزیع واقعی تنش–مقاومت نیست و می‌توان با استفاده از دو متغیر تصادفی جداگانه اطلاعات آماری را مستقیماً برای محاسبه قابلیت اطمینان به‌کار برد. فرض کنید تعداد داده‌های موجود در هر گروه برابر با ۱۰۰ است و پس از یک پردازش ساده می‌توان داده ها را با هیستوگرام توصیف کرد. همانطور که در شکل های ۲(آ) و ۲(ب) نشان داده شده‌است، فواصل کلاسی داده‌ها و فرکانس‌های نسبی متناظر آن‌ها به‌دست آمده‌است. به‌طور مشابه، مقادیر نقطه میانی هر بازه کلاس به عنوان مقادیر ممکن متغیر تصادفی تنش–مقاومت در نظر گرفته می‌شود و فرکانس‌های نسبی هر بازه کلاس به عنوان احتمالات مربوطه در نظر گرفته می‌شوند. بنابراین، دو متغیر تصادفی جدید تنش–مقاومت با تابع جرم احتمال شناخته شده به شرح زیر به‌دست‌می‌آید:

$$S_۱ = (۷۲۵, ۷۵۵, ۷۸۵, ۸۱۵, ۸۴۵, ۸۷۵, ۹۰۵)\ ,\ S_۲ = (۸۵۰, ۹۰۰, ۹۵۰, ۱۰۰۰, ۱۰۵۰, ۱۱۰۰, ۱۱۵۰)$$
$$p_۱ = (۰٫۰۱, ۰٫۰۴, ۰٫۲۴, ۰٫۳۸, ۰٫۲۵, ۰٫۰۶, ۰٫۰۲)\ ,\ p_۲ = (۰٫۰۲, ۰٫۰۶, ۰٫۲۵, ۰٫۳۹, ۰٫۲۲, ۰٫۰۵, ۰٫۰۱)$$



(آ) هیستوگرام داده‌های تنش



(ب) هیستوگرام داده‌های مقاومت

با استفاده از مدل تداخل تنش–مقاومت گسسته ارائه شده در بخش ۳، قابلیت اطمینان این جزء به صورت $R = ۰٫۹۹۷۲$ به‌دست‌می‌آید.

همانطور که در بخش ۱ ذکر شد، روش ارائه شده توسط (۴) می‌تواند برای حل این مشکل استفاده شود که نوع توزیع احتمال تنش–مقاومت مشخص باشد. فرض کنید هر دو تنش–مقاومت در این حالت متغیر تصادفی نرمال مستقل با پارامترهای ناشناخته هستند. دو گروه از داده‌ها را می‌توان به عنوان مشاهدات تنش–مقاومت در نظر گرفت. با استفاده از روش برآورد حداکثر درست‌نمایی،

پایایی این مؤلفه را می‌توان $R = ۰/۹۹۶۲$ محاسبه کرد. اگر این نتیجه به عنوان مقدار دقیق قابلیت اطمینان در نظر گرفته شود، خطای نسبی حاصل از مدل تداخل تنش–مقاومت گسسته برابر با ۰/۱ است.

بحث و نتیجه‌گیری

در این مقاله، یک رویکرد جدید بر اساس روش تابع مولد عام برای مدل‌سازی قابلیت اطمینان سیستم‌های تنش– مقاومت بر اساس متغیرهای تصادفی گسسته مورد بررسی و مطالعه قرار گرفت که در آن یک مدل تداخل تنش– مقاومت گسسته برای محاسبه قابلیت اطمینان مؤلفه را تحت شرایط مختلف، از جمله زمانی که داده‌ها به صورت پیوسته، گسسته، یا تنها در قالب دو گروه از داده‌های تجربی موجود باشند، داراست. یکی از مهمترین مزایای مدل پیشنهادی، عدم نیاز به اطلاعات قبلی در مورد توزیع دقیق تنش و مقاومت است که به طور قابل توجهی محدودیت‌های موجود در مدل‌های پیوسته SSI را برطرف می‌کند. در بسیاری از کاربردهای عملی، دسترسی به توزیع دقیق داده‌ها دشوار است و تنها داده‌های آزمایشی محدودی در دسترس قرار می‌گیرد. نتایج عددی نشان دهنده برتری روش مورد مطالعه است. برای داده‌های پیوسته که به متغیرهای گسسته تبدیل شدند، مدل دقت بالایی بوده و نتایج نشان داد که کاهش طول زیربازه در فرآیند گسسته‌سازی، دقت محاسبات را به طور محسوسی افزایش می‌دهد. علاوه بر این، مدل SSI گسسته کارایی مناسبی را در شرایطی که تنها دو گروه از داده‌های تنش و مقاومت موجود بود، ارائه نمود.

مراجع

[1] Kapur, K.C., 1975, *Reliability bounds in probabilistic design*, IEEE Transactions on Reliability, 24(3), pp.193-195.

[2] Park, J.W. and Clark, G.M., 1986, *A computational algorithm for reliability bounds in probabilistic design*, IEEE transactions on reliability, 35(1), pp.30-31.

[3] Wang, J.D. and Liu, T.S., 2002, *Fuzzy reliability using a discrete stress-strength interference model*, IEEE transactions on reliability, 45(1), pp.145-149.

[4] Kotz, S., Lumelskii, Y. and Pensky, M., 2003, *The stress–strength model and its generalizations: theory and applications*.

A discrete stress–strength interference model based on universal generating function

Behnaz Tsadighi, Dr.Reza Zarei, Dr.Behroz Fathi [1]

[1]Department of Statistics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran

---

Abstract: Continuous stress–strength interference (SSI) model regards stress and strength as continuous random variables with known probability density function. This, to some extent, results in a limitation of its application. In this paper, stress and strength are treated as discrete random variables, and a discrete SSI model is presented by using the universal generating function (UGF) method. Finally, case studies demonstrate the validity of the discrete model in a variety of circumstances, in which stress and strength can be represented by continuous random variables, discrete random variables, or two groups of experimental data.

Keywords: Reliability; Universal generating function;Stress–strength interference ;Discrete model.

Mathematics Subject Classification (2020): 90B25، 60E10، 62N05.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

اخلاق در آمار و احتمال

علی دولتی [۱]

[۱]**گروه آمار دانشگاه یزد**

چکیده: امروزه با گسترش استفاده از داده‌ها و روش‌های آماری در تصمیم‌سازی‌ها، اقدامهای غیر اخلاقی مانند سوءاستفاده از داده‌ها، ارائه تفسیرهای مغرضانه و نادرست از نتیجه تحلیل‌ها و عدم شفافیت در ارائه گزارش‌ها، به صورت آگاهانه یا ناآگاهانه در انجام کارهای آماری وجود دارد. از این رو، برای تقویت اعتماد عمومی به آمار و حفظ سلامت و اعتبار این حوزه، آموزش و ترویج اصول اخلاقی، امری حیاتی است. هدف این مقاله ارائه الگویی برای گنجاندن اصول اخلاقی در سرفصل‌های درسی آمار است تا دانشجویان علاوه بر یادگیری مفاهیم نظری و کاربردی آمار، به مسئولیت‌پذیری اجتماعی و حرفه‌ای خود نیز آگاهی پیدا کنند.

واژه‌های کلیدی: آموزش آمار، اخلاق در آمار، برنامه‌درسی، مسئولیت‌پذیری حرفه‌ای، حریم خصوصی داده‌ها.

کد موضوع‌بندی ریاضی (۲۰۲۰): ۹۷k۸۰، ۹۷k۹۹.

---

۱ مقدمه

اهمیت ملاحظات اخلاقی در یک‌رشته علمی یا حوزه تخصصی، زمانی برجسته می‌شود که نتایج آن تأثیرات گسترده و قابل‌توجهی بر جهان پیرامون خود داشته باشد. در رشته‌هایی نظیر پزشکی، مهندسی، روانشناسی، علوم‌انسانی و اجتماعی، عموماً این باور وجود دارد که کار و تخصص آنها می‌تواند پیامدهای اخلاقی گسترده و فراگیری برای جامعه داشته باشد و این پیامدها بایستی موردتوجه و ارزیابی قرار گیرند. در این حوزه‌ها سازمان‌ها و نهادهای تخصصی گوناگونی مانند نظام‌مهندسی، نظام‌پزشکی و انجمن‌های مختلفی وجود دارند که با تدوین شیوه‌نامه‌ها و دستورالعمل‌هایی به‌طورجدی بر اهمیت مسئولیت‌های اجتماعی و ملاحظات اخلاقی متخصصان در زمینه‌های کاری خود تأکید می‌ورزند و آن را جزو لاینفک استانداردهای حرفه‌ای خود می‌دانند. این موضوع نشان دهنده این است که در این رشته‌ها، توجه به پیامدهای اخلاقی کارها، بخش مهمی از هویت حرفه‌ای محسوب می‌شود. در ایران نیز مانند سایر کشورهای دنیا در

---

[۱] سخنران، adolati@yazd.ac.ir

زمینه اخلاق علمی و حرفه‌ای، قوانین، دستورالعمل‌ها، شیوه‌نامه‌ها، نظام‌نامه‌ها و راهنماهای مختلفی در حوزه‌هایی مانند وکالت، مهندسی، روان‌شناسی و مشاوره، حسابداری و پزشکی، وجود دارد (۷؛ ۸؛ ۹؛ ۱۰؛ ۱۱؛ ۱۲) در مورد ضرورت آموزش اخلاق در رشته‌های دانشگاهی مقالات زیادی نوشته شده است. به‌عنوان نمونه، مراجع (۱؛ ۲؛ ۳؛ ۴؛ ۵؛ ۶) را ملاحظه کنید. در بسیاری از کشورها، وکلا، پزشکان، زیست شناسان، مهندسان، فیزیک‌دانان، متخصصان علوم کامپیوتر، آماردانان و حسابداران خبره اخلاق تخصصی مرتبط با رشته خود را فرامی‌گیرند، زیرا در مقام متخصص، با تصمیم‌گیری‌های اخلاقی در حرفه خود روبرو خواهند شد. فیسلر و همکاران (؟) در سال ۲۰۲۰ به تحلیل محتوای بیش از صد عنوان درسی با محتوای آموزش اخلاق در رشته‌های مختلف پرداخته‌اند که در برخی از دانشگاه‌های دنیا تدریس می‌شود. این بررسی، طیف وسیعی از رویکردهای موجود در زمینه آموزش اخلاق در رشته‌های مختلف را در طول دهه ها نشان می‌دهد. در ایران نیز در برنامه‌های درسی برخی از رشته‌های مهندسی و علوم پزشکی، دروسی با محتوای اخلاق علمی و حرفه‌ای وجود دارد. رجبعلی پور و همکاران (۵) به تحلیل محتوای دروس ۲۰۵ رشته در حوزه پزشکی و پیراپزشکی پرداخته و دروسی را که محتوای اخلاق حرفه‌ای داشته‌اند، شناسایی کرده‌اند. در حوزه مهندسی نیز دروسی با محتوای اخلاق مهندسی در برنامه درسی برخی از رشته‌های دانشگاهی کشور وجود دارد (۴).

در عصر انفجار داده‌ها و تحولات سریع فناوری در حوزه تحلیل داده‌ها، آمار به عنوان ابزاری حیاتی برای تصمیم‌گیری‌های حساس در حوزه‌های مختلف علمی، پزشکی، اقتصادی و اجتماعی مطرح است. بااین‌حال، قدرت فزاینده روش‌های آماری همراه با مسئولیت اخلاقی روزافزونی است که متخصصان این حوزه باید به آن توجه کنند. اخلاقیات آماری به اصول و دستورالعمل‌هایی اشاره دارد که مسئولیت‌پذیری و رفتار اخلاقی در فعالیت‌های آماری را تنظیم می‌کنند. این اصول تمام مراحل کار آماری، از جمع‌آوری و تحلیل داده‌ها تا تفسیر و گزارش‌دهی را در بر می‌گیرد. رعایت اخلاق در آمار برای حفظ صداقت پژوهش، تقویت اعتماد عمومی و اطمینان از استفاده مسئولانه از داده‌ها در تصمیم‌گیری‌های حیاتی است. موارد متعددی مانند سوءاستفاده از تکنیک‌های آماری، نداشتن تخصص و صلاحیت انجام کار آماری، تحریف عمدی نتایج، نقض حریم خصوصی داده‌ها و تفسیرهای گمراه‌کننده از نتایج، توسط انجام دهندگان کارهای آماری، لزوم توجه جدی‌تر به آموزش اصول اخلاقی در کنار مباحث فنی آمار را آشکار ساخته است. آموزش رسمی دانشجویان رشته آمارعمدتاً بر جنبه‌های محاسباتی و تکنیکی متمرکز بوده و کمتر به ابعاد اخلاقی این حرفه پرداخته است. این در حالی است که یک تصمیم آماری نادرست یا غیراخلاقی می‌تواند پیامدهای گسترده‌ای در سطح جامعه داشته باشد. از خطاهای پزشکی ناشی از تحلیل‌های نادرست گرفته تا سیاست‌گذاری‌های اقتصادی اشتباه بر اساس داده‌های تحریف شده، همگی بر اهمیت رعایت اصول اخلاقی در این حوزه تأکید دارند. نبود آموزش‌های مرتبط با این حوزه،، به‌ویژه زمانی که با سطح آموزش‌های اخلاقی در سایر رشته‌ها مانند پزشکی و مهندسی مقایسه می‌شود، نمود بیشتری پیدا می‌کند. این مقاله باتوجه‌به این ضرورت، به بررسی راهکارهایی برای گنجاندن آموزش اخلاق در برنامه‌های درسی آمار می‌پردازد.

۲ تاریخچه شیوه‌نامه‌های اخلاقی آمار

توجه به ملاحظات اخلاقی در تحقیقات سابقه‌ای طولانی دارد. پس از جنگ جهانی دوم و فجایعی مانند آزمایش‌های نازی‌ها بر روی انسان‌ها و برگزاری دادگاههای نورنبرگ مجموعه‌ای از اصول اخلاقی که به کد نورنبرگ (۱۷) معروف است تدوین شد و به‌عنوان یکی از پایه‌های مهم اخلاق پزشکی و تحقیقات انسانی شناخته می‌شود. مهم‌ترین اصول کد نورنبرگ عبارت‌اند از: رضایت داوطلبانه و مشارکت فرد با آگاهی کامل از خطرات و اهداف تحقیق و ممنوعیت اجبار، فریب یا فشار غیراخلاقی، طراحی آزمایش‌ها بر پایه دانش علمی و

با اهداف مفید برای جامعه، نرساندن رنج یا آسیب جسمی و روانی به افراد، حق انصراف شرکت‌کننده در هر مرحله از تحقیق و نهایتاً برتری قابل‌توجه منافع تحقیق بر خطرات آن. این اصول بعدها در بیانیه هلسینکی (۲۱) و سایر قوانین اخلاق پزشکی جهانی گنجانده شد. آمار به‌عنوان یکی از ابزارهای علمی برای فهم واقعیت‌های اجتماعی، اقتصادی و پزشکی، همواره با داده‌هایی سروکار دارد که گاه به طور مستقیم با زندگی انسان‌ها گره‌خورده‌اند. به همین دلیل، توجه به جنبه‌های اخلاقی در انجام مطالعات آماری، امری حیاتی و اجتناب‌ناپذیر است. سوءاستفاده‌های آماری در پژوهش‌ها (مانند دستکاری داده‌ها یا گزارش انتخابی نتایج) منجر به توجه بیشتر به اخلاق در انتشاریافته‌های علمی بر پایه روشهای آمار شد. اخلاقیات آماری، به عنوان مجموعه اصول راهنمای انجام مسئولانهٔ کارهای آماری، در طول زمان تکامل یافته است. از دهه ۸۰ میلادی، انجمن‌های آماری بین‌المللی دستورالعمل‌هایی برای رفتار حرفه‌ای تدوین کردند. انجمن آمار آمریکا [۱] نقش کلیدی در این تحول ایفا کرده و طی چند دهه، «راهنمای اخلاقی برای انجام کارهای آماری» (۱۴) خود را تدوین و اصلاح کرده است. تاریخچه بیش از سه دهه فعالیت انجمن آمار امریکا برای تدوین این شیوه‌نامه را در مراجع (۱۹؛ ۲۲) ببینید. با گسترش فناورهای نوظهور و جمع‌آوری داده‌های بزرگ و افزایش چالش‌ها و نگرانی‌ها درباره حریم خصوصی و سوءاستفاده از داده‌ها، استانداردهای اخلاقی جامعتری در انجام کارهای آماری لازم گردید و سازمانهای دیگری مانند موسسه بین المللی آمار [۲] فدراسیون انجمنهای ملی آمار اروپا[۳]، انجمن آکچواری (بیمسنجی) امریکا [۴]، انجمن آمار سلطنتی [۵]، اداره آمار بریتانیا [۶]، برای ترویج اخلاق در انجام کارهای آماری دستورالعمل‌های خود را تدوین و منتشر نمودند. کمیسیون آمار سازمان ملل متحد درسال ۱۹۹۴ اصول اساسی آمار رسمی را تصویب کرد. نسخه تجدیدنظر شده این اصول در سال ۲۰۱۳ منتشر شد [۷]. اخیرا انجمن آمار ایران نیز با تشکیل «کمیته اخلاق آماری» و تصویب «شیوه‌نامه اخلاقی برای کارهای آماری» (۱۲) در این زمینه فعالیت‌های خود را آغاز نموده است.

۳ اصول اخلاق آماری

در این بخش به اصول اخلاقی انجام کارهای آماری پرداخته می‌شود. نخست برخی تعریف‌های لازم یادآوری می‌شود. در شیوه‌نامه اخلاقی برای کارهای آماری که توسط انجمن آمار ایران (۱۲) و با بهره‌گیری از شیوه‌نامه‌های معتبر منتشر شده در این زمینه توسط انجمن آمار امریکا و موسسه بین المللی آمار تدوین شده است، «کار آماری»، شامل فعالیت‌هایی مانند نمونه‌گیری، طراحی آزمایش‌ها، خلاصه‌سازی، پردازش، تصویری‌سازی، تحلیل و تفسیر داده‌ها، ایجاد یا بکارگیری مدل یا توسعه و تعمیم الگوریتم‌ها است. «انجام دهندگان کارهای آماری» همه کسانی هستند که بدون توجه به عنوان شغل، حرفه، میزان اطلاعات یا رشته تحصیلی، به این فعالیت‌ها می‌پردازند. کسانی که در یک کار آماری سرمایه‌گذاری می‌کنند، به آن کمک می‌کنند، از آن استفاده می‌کنند، یا به گونه‌ای تحت تأثیر آن قرار می‌گیرند، «ذینفعان» محسوب می‌شوند. سازمان‌ها، مؤسسات، شرکت‌ها، دانشگاه‌ها، نهادهای دولتی یا غیردولتی که به نحوی در انجام، حمایت، تأمین مالی، اجرای پروژه‌های آماری یا بهره‌برداری از نتایج آن نقش دارند، به عنوان «کارفرما، ناظر، حامی مالی، یا استفاده‌کننده از نتایج آماری» شناخته می‌شوند. به فردی که وظیفه نظارت، هدایت و تضمین کیفیت در اجرای کار آماری را بر عهده دارد و ممکن است به عنوان استاد راهنما، مدیر پروژه، سرپرست تحلیل آماری، پژوهشگر ارشد، مربی یا هر عنوان دیگری فعالیت کند و مسئولیت اجرای صحیح، شفافیت

[1] American Statistical Association (ASA)
[2] International Statistical Institute (ISI)
[3] The Federation of European Statistical National Societies (FENStatS)
[4] American Academy of Actuaries
[5] Royal Statistical Society (RSS)
[6] UK Statistics Authority
[7] United Nations Statistical Commission

و رعایت اصول اخلاقی در انجام کار آماری را بر عهده دارد یا در آن شریک است، «راهنمای علمی و فنی در کار آماری» گفته می‌شود. اصول اخلاق در انجام کارهای آماری شامل موارد زیر است(۱۲) :

(اصل الف) **صداقت حرفه‌ای و مسئولیت‌پذیری** : انجام‌دهنده کارهای آماری اخلاق‌مدار، برای ایجاد اعتماد بین خود و ذینفعان، توانمندی‌ها و فعالیت‌های خود را به‌طور صادقانه عرضه می‌کند، با دیگران با احترام برخورد می‌نماید و مسئولیت کارهای خود را می‌پذیرد.

(اصل ب) **درستی و یکپارچگی داده‌ها و روش‌ها:**

انجام دهنده کارهای آماری اخلاق‌مدار در حالی که در صدد کشف و رفع محدودیت‌ها، نقص‌ها یا سوگیری‌های شناخته شده یا مشکوک در داده‌ها یا روش‌ها است، پیامدهای بالقوه آنها را بر تفسیر، نتیجه‌گیری، توصیه‌ها، تصمیم‌گیری‌ها یا سایر نتایج کارهای آماری اطلاع‌رسانی می‌کند.

(اصل پ) **مسئولیت‌های حرفه‌ای:** انجام‌دهندگان کارهای آماری اخلاق‌مدار به اصول علمی و فنی تکیه میکنند و در قبال پیامدهای اجتماعی، اخلاقی و انسانی تصمیمات آماری خود نیز مسئولیت‌پذیر هستند. این مسئولیت‌ها به‌طور مستقیم به نحوه تعامل با ذینفعان، داده‌ها، اعضای تیم‌های چندرشته‌ای و همکاران مربوط می‌شود. با رعایت دقیق وشفاف این مسئولیت‌ها، می‌توان اطمینان حاصل کرد که فرایندهای آماری در راستای منافع عمومی، احترام به حقوق افراد و ارتقای کیفیت تصمیم‌سازیهای مبتنی بر داده‌ها پیش می‌روند.

(اصل پ ۱) **مسئولیت در قبال ذینفعان:** انجام‌دهنده کارهای آماری اخلاق‌مدار، انتظارات ذینفعان از هر پروژه خاص را شناسایی و تبیین می‌کند و به منافع و حقوق آنها احترام می‌گذارد.

(اصل پ ۲) **مسئولیت در قبال داده‌ها و آزمودنیهای تحت تأثیر کارهای آماری:** انجام دهنده‌کارهای آماری اخلاق‌مدار برای تولید و استفاده از داده‌ها، به گونه‌ای عمل می‌کند که موجب آسیب به افراد یا گروه‌ها نشود و همواره به حقوق انسان‌ها، حیوانات و محیط زیست احترام می‌گذارد. در عین حال، امکان سوءاستفاده‌های دیگران از داده‌ها را نیز نادیده نمی‌گیرد. مسئولیت‌پذیر است و به عواقب کارهای آماری خود، به ویژه کارهایی که ممکن است تأثیر مستقیم بر زندگی دیگران ومحیط زیست داشته باشد، توجه دارد.

(اصل پ ۳) **مسئولیت در قبال همکاران و اعضای تیم‌های چندرشته‌ای:** کارهای آماری معمولاً در حوزه‌های مختلف و تیم‌های چندرشته‌ای انجام می‌شود که افراد با تخصص‌ها و دیدگاه‌های متفاوت در آن مشارکت دارند. این تنوع، فارغ از عنوان شغلی یا میزان تحصیلات، می‌تواند منجر به هم‌افزایی و بهبود کیفیت کارها شود، اما در عین حال نیازمند رعایت اصول اخلاقی در تعاملات و همکاری‌ها است. انجام‌دهندگان کارهای آماری باید در محیط‌های گروهی و تیمی به‌طور اخلاقی و حرفه‌ای عمل کنند و همواره اصول احترام متقابل و شفافیت را رعایت نمایند.

(اصل ت) **الزامات و تعهدات اشخاص حقوقی، استادان راهنما، راهبران، سرپرستان و مربیان:** در فرآیند انجام کارهای آماری، علاوه بر فرد انجام‌دهنده، نقش استادان راهنما، راهبران، سرپرستان، مشاوران و سازمانها نیز از اهمیت ویژه‌ای برخوردار است. این افراد و سازمان‌ها باید با رعایت اصول اخلاقی و حرفه‌ای، ترویج‌دهندگان فرهنگ آماری اخلاق‌مدار باشند و از انجام کارهای آماری با کیفیت و متکی به شیوه‌نامه‌های اخلاقی حمایت کنند. در این راستا، اقدامات مشخصی لازم است که هم‌راستایی سازمان‌ها و نهادها با شیوه‌نامه‌های اخلاقی و حرفه‌ای ایجاد شود، برای این که محیطی امن، سازنده و احترام‌آمیز برای تمامی اعضای تیم های آماری فراهم شود.

(اصل ت ۱) **استادان راهنما، راهبران، سرپرستان و مربیان** : استادان راهنما، رهبران، سرپرستان و مربیان، انجام‌دهندگان اصلی، ناظران و مشاوران کار های آماری را به پیروی، ترویج و حمایت ازاصول اخلاقی متعهد می‌کنند و بر انجام کارهای آماری اخلاق‌مدار توسط کارگروه مجری تاکید دارند.

(اصل ت ۲) **سازمان‌ها، موسسات و نهادها:** سازمان‌ها، مؤسسات و نهادهای درگیر در فعالیت‌های آماری (مانند جمع‌آوری، خلاصه‌سازی، پردازش، تحلیل، تفسیر، یا ارائه نتایج آماری با استفاده از انواع داده‌ها)، مسئولیت دارند که کارهای آماری اخلاق‌مدار را ترویج نمایند.

(اصل ث) **تعهدات اخلاقی در مواجهه با تخلفات آماری و حرفه‌ای:** همواره این امکان وجود دارد که انجام‌دهنده کارهای آماری اخلاق‌مدار با شبهاتی در ارتباط با تخلفات احتمالی در زمینه کارهای آماری، شیوه‌های علمی یا عملکرد حرفه‌ای مواجه شود. گاهی اوقات، یک انجام‌دهنده ممکن است فردی را به تخلف متهم کند یا توسط دیگران به تخلفی متهم شود. در برخی مواقع نیز، یک انجام‌دهنده ممکن است در رسیدگی به تخلفات دیگران مشارکت داشته باشد. انجام دهنده کارهای آماری اخلاق‌مدار، تعاریف و شیوه‌های مربوط به اتهام‌های تخلف در حوزه‌های آماری در محیط سازمانی خود را می‌داند، تفاوت رفتار غیراخلاقی با اختلاف نظر و اشتباه صادقانه را تشخیص می‌دهد و قبل از اظهارنظر درباره تخلف دیگران به دنبال کشف حقایق و انگیزه می‌گردد.

هر کدام از اصول بالا، دارای چندین بند است که توضیحات آنها در شیوه‌نامه اخلاقی کارهای آماری انجمن آمار (۱۲) آمده است. هدف اصلی این شیوه‌نامه و شیوه‌نامه‌های مشابه، تبیین استانداردهای اخلاقی و ایجاد فرهنگ پاسخگویی و مسئولیت‌پذیری در میان افرادی است که به نحوی با داده‌ها، روش‌ها و تحلیل‌های آماری و تفسیر نتایج سر و کار دارند. این شیوه‌نامه‌ها به دانشجویان، پژوهشگران و تمامی افرادی که در حوزه‌های مختلف از روش‌های آماری استفاده می‌کنند، کمک می‌کند تا مسئولانه و با رعایت اصول اخلاقی به انجام کارهای آماری بپردازند. پایبندی به اصول اخلاقی، برای تضمین کیفیت، اعتبار علمی و اعتماد عمومی به نتایج کارهای آماری ضروری است. موضوع اخلاق در آمار، جنبه‌های متفاوتی قابل بحث و بررسی است. برای درک بیشتر اهمیت موضوع، می‌توان به کتاب‌هایی که در این زمینه نوشته شده است مراجعه کرد (۱۸؛ ۲۴؛ ۲۶).

۴ اخلاق و آموزش آمار

همان گونه که در دانشگاه طیف وسیعی از مباحث آمار به دانشجویان آموزش داده می‌شود تا برای فرصت‌های تحصیلی و حرفه‌ای مختلف آماده شوند، منطقاً باید طیف وسیعی از موقعیت‌های اخلاقی مرتبط با انجام کارهای آماری در زمینه‌های مختلف نیز آموزش داده شود. ترویج اخلاق در انجام فعالیت‌های مرتبط با آمار، بایستی از طریق گنجاندن آموزش اخلاق در برنامه‌های درسی انجام شود تا بتواند در شکل‌دهی به رفتار حرفه‌ای دانش‌آموختگان موثر باشد. گنجاندن اصول اخلاق در آموزش آمار را می‌توان با استفاده از رویکردهای متنوعی مانند: گنجاندن فعالیت‌ها یا تکلیف‌های خاصی در خلال تدریس برخی از دروس، توزیع مفاهیم اخلاقی در محتوای برخی از دروس، برگزاری کارگاه‌های آموزشی خارج از برنامه درسی رسمی، واگذاری مسئولیت آموزش اخلاق، به دوره‌های ارائه شده توسط متخصصین رشتههای مرتبط و نهایتا گنجاندن درسی مستقل در برنامه درسی انجام داد. در بسیاری از دانشگاه‌ها و مؤسسات آموزشی دنیا دروس یا دوره‌هایی با عنوان «اخلاق علمی» تدریس می‌شود که به مباحث عمومی مرتبط با ملاحظات اخلاقی پژوهش و انتشار نتایج آن اختصاص دارد. از جمله:

- آگاهی در مورد تخلفات علمی مانند سرقت علمی و ادبی و تشریح پیامدهای آن برای اعتبار فرد و جامعه علمی
- آگاهی از تفاوت تخلف علمی و اشتباه صادقانه و نحوه برخورد با تخلفات علمی
- نحوه عملکرد در اصلاح اشتباهات اساسی، پس از انتشار یک پژوهش
- آموزش نحوه ارجاع به مقالات و منابع علمی در قالب استاندارد
- اهمیت صداقت، شفافیت در گزارش و انتشار نتایج پژوهش‌ها و مسئولیت‌پذیری
- اصول اخلاقی در همکاری‌های علمی گروهی
- برخورد صحیح با همکاران در صورت وجود اختلاف نظر علمی
- تقسیم عادلانه کار و رعایت انصاف در نوشتن مقالات مشترک و سهم اعضای گروه
- اصول اخلاقی و قوانین و مقررات مربوط به نوشتن گزارش‌های علمی
- آگاهی از فرآیند بررسی نتایج پژوهش توسط همتایان و نحوه تعامل با داوران.

در برخی از رشته‌ها درکنار آموزش ملاحظات عمومی اخلاق علمی در انتشار نتایج پژوهش‌ها، دوره‌ها یا دروسی با عنوان «اخلاق حرفه‌ای» نیز ارائه می‌شود. در مرجع (۲۰) محتوای ۱۱۵ عنوان درس در زمینه آموزش اخلاق در رشته‌های مختلف، در برخی از دانشگاه‌های دنیا (امریکا، کانادا، اروپا و آسیا) مورد بررسی و تحلیل و دسته‌بندی موضوعی قرار گرفته است. این دسته‌بندی طیف وسیعی از ملاحظات اخلاقی را شامل می‌شود. برخی از این موارد مختص رشته‌های مربوط و برخی نیز عمومی هستند. مسائل اخلاقی مرتبط با هوش مصنوعی و فناوری‌های نوین در اغلب دروس ظاهر می‌شود. از جمله:

- حفظ حریم خصوصی افراد و قوانین مربوط به محرمانگی و امنیت داده‌ها
- ناشناس‌سازی اطلاعات و ردپای افراد در کلان داده‌ها (داده‌های بزرگ) و داده‌های دیجیتال
- توجه به محدودیت‌های مربوط به تلقی داده‌های دیجیتال به عنوان دارایی شرکت‌ها
- رعایت مالکیت معنوی و قوانین کپی رایت و نسخه‌برداری
- دوری از انواع تبعیض و نابرابری اجتماعی و آزار و اذیت افراد
- حمایت از آزادی بیان، حقوق بشر و حفظ ارزش‌های دموکراسی
- مسئولیت در قبال محیط زیست و حیوانات
- مسئولیت در قبال اخبار جعلی، مطالب نادرست و شبه علم.

در زمینه ضرورت و چگونگی گنجاندن آموزش ملاحظات اخلاقی در برنامه درسی رشته آمار و همچنین علوم ریاضی، مقالات زیادی نوشته شده است و نیز در برخی از دانشگاه‌های دنیا، دروسی نیز تعریف شده است. به عنوان نمونه مراجع (۱۳؛ ۱۶؛ ۲۳؛ ۲۷؛ ۲۸؛ ۲۹)

را ملاحظه کنید. در اینجا به محتوای برخی از این دروس، به عنوان نمونه اشاره می‌کنیم. چیدو و مولیر (۱۵) در گزارشی که اخیرا منتشر نموده‌اند، رویکردی برای گنجاندن آموزش اخلاق در علوم ریاضی در برنامه دوره سطح کارشناسی ارائه می‌دهند. این درس برای دانشجویان رشته‌های مهندسی، علوم کامپیوتر، آمار، فیزیک و اقتصاد نیز قابل تدریس است. مواد آموزشی درس به سه نوع تقسیم شده است: تمرین‌ها، پروژه‌ها و مطالب درسی. تمرین‌های انتخابی که برای آموزش ملاحظات اخلاقی، استفاده شده است، در حد آمار مقدماتی و ریاضیات عمومی سال اول و دوم مقطع کارشناسی است. برخی از تمرین‌های اساسی‌تر، به عنوان پروژه‌های مستقل به دانشجویان داده می‌شود تا ملاحظات اخلاقی آنها را به صورت گروهی، به طور عمیق‌تر بررسی کنند. این گونه مسائل، معمولاً به تحقیق و مطالعه بیشتری نیاز دارند و از آنها به عنوان بخشی از ارزیابی دوره استفاده می‌شود. فهرستی از مطالب مربوط به اخلاق در آمار، ریاضی و سایر علوم در این گزارش ارائه شده است که مدرس، با توجه به مخاطبان خود، می‌تواند از آنها استفاده کند. تمرین‌ها به طور موازی هم برای آموزش مطالب نظری و هم برای آگاهی جنبه اخلاقی آنها طراحی شده‌اند و حوزه‌های مختلفی را پوشش می‌دهند. هر تمرین دارای برچسبی است که موضوع رشته مرتبط و یک رکن اخلاقی را توضیح می‌دهد. برچسب‌هایی که برای رشته‌های مرتبط استفاده شده است شامل آمار و احتمال، علوم کامپیوتر، اقتصاد، مهندسی و علوم طبیعی است. برچسب‌های مربوط به حوزه‌های اخلاقی استفاده شده در تمرین‌ها، شامل ۱۰ مورد به شرح زیر است (۱۵):

۱ - توجیه آغاز یک کار: توجیه شما برای انجام یک کار آماری چیست؟ آیا اساساً انجام آن موجه است؟

۲ - بهره‌گیری از تنوع دیدگاه و اجتناب از سوگیری‌ها: آیا از تنوع دیدگاه کافی در بین همکاران و مدیران برخوردارید؟ آیا محدودیت‌ها و سوگیری‌های شناختی خود را درک می‌کنید؟

۳ - مدیریت مسئولانه و اثربخش داده‌ها: آیا منحصراً از مجموعه داده‌های مجاز و به‌دست‌آمده به شیوه‌ای اخلاقی استفاده می‌کنید؟

۴ - پردازش و استنتاج داده‌ها با دقت و صحت: آیا از تخصص کافی برای تحلیل صحیح و اخلاقی داده‌ها و تضمین کیفیت نتایج برخوردارید؟

۵ - مدل‌سازی ریاضی مسئله و پیامدهای واقعی آن: اهداف و محدودیت‌های مدل انتخابی شما چیست؟ پیامدهای واقعی آن برای ذینفعان مختلف کدامند؟

۶ - برقراری ارتباط شفاف و مستندسازی دقیق فرایند کار: چگونه کار خود را به طور شفاف شرح داده و مستند می‌کنید؟ چگونه نتایج را به ذینفعان مربوطه منتقل می‌کنید؟

۷ - قابلیت ابطال و ایجاد حلقه بازخورد مستمر: آیا کار شما قابل ابطال است؟ چگونه تأثیرات گسترده و سازوکارهای بازخورد ایجاد شده را مدیریت می‌کنید؟

۸ - قابلیت تفسیر و ایمنی مدل: آیا خروجی مدل شما قابل تفسیر است؟ آیا سازوکارهای نظارت و نگهداری مناسب و توسعه آن پیش‌بینی شده است؟

۹ - آگاهی از جنبه‌های سیاسی و اجتماعی مرتبط با کار: آیا از جنبه‌های غیر فنی و ماهیت سیاسی و اجتماعی کار خود آگاه هستید؟ برای جلب اعتماد ذینفعان به کار و محصول خود چه تدابیری اندیشیده‌اید؟

۱۰ - تدوین استراتژی‌های واکنش اضطراری مناسب: آیا استراتژی دادن پاسخ غیرفنی مناسب، برای مواجهه با شرایط پیش‌بینی نشده را دارید؟ آیا یک شبکه پشتیبانی متشکل از همکاران آگاه و حامی در اختیار دارید؟

نمونه‌هایی از تمرین‌ها و پروژه‌های این درس به همراه راه حل‌های پیشنهادی، به طور مفصل در گزارش (۱۵) آمده است. در وبگاه انجمن علم داده (۳۰) [۸]، وبگاه مرکز اخلاق مهندسی و علوم (۳۱)[۹]، وبگاه نشریه انجمن اخلاق در برنامه درسی (۳۲) [۱۰] و همچنین وبگاه مرکز ملی آموزش آمار (۳۳) [۱۱] می‌توان مطالب مفیدی در زمینه آموزش اخلاق در آمار و ریاضی جستجو نمود.

۱.۴ مثالی از آموزش ملاحظات اخلاقی در احتمال

شما با دوستتان عباس در تعطیلات به کوه نوردی رفته‌اید. متأسفانه، عباس روی یک سنگ می‌افتد و پای راستش می‌شکند و به یک بیمارستان کوچک شهر نزدیک منتقل می‌شود و مجبور است یک شب دیگر در بیمارستان بماند. برق بیمارستان هر ساعت به صورت مستقل با احتمال p قطع می‌شود.

(الف) فرمولی برای تعداد ساعات مورد انتظار تا قطعی برق بیمارستان استخراج کنید.

(ب) مقدار p باید چقدر کوچک باشد تا تعداد مورد انتظار قطعی برق در یک سال کمتر از ۱ باشد؟

(پ) عباس باید x ساعت دیگر در بیمارستان بماند. اگر برق قطع نشود، او به‌موقع مرخص می‌شود و شما دو نفر هنوز می‌توانید به پرواز برگشتتان برسید. احتمال اینکه برق در $x-۱$ ساعت اول قطع شود چقدر است؟

(ت) عباس از شما می‌پرسد که چه مقدار از p شما را آسوده‌خاطر می‌کند. چگونه تصمیم می‌گیرید؟

پاسخ مسئله

(الف) این مسئله از توزیع هندسی پیروی می‌کند. امیدریاضی متغیر تصادفی X، تعداد ساعات تا اولین قطعی برق برابر است با

$$E(X) = \boxed{\frac{۱}{p}}$$

(ب) می‌خواهیم مقدار p را طوری تعیین کنیم که تعداد قطعی‌های مورد انتظار در یک سال کمتر از ۱ باشد. تعداد ساعات در یک سال $۳۶۵ \times ۲۴ = ۸۷۶۰$ ساعت و تعداد مورد انتظار قطعی‌ها در سال $۸۷۶۰p$ است. شرط مسئله:

$$۸۷۶۰p < ۱ \implies \boxed{p < \frac{۱}{۸۷۶۰} \approx ۰/۰۰۰۱۱۴}$$

(پ) احتمال قطع نشدن برق در یک ساعت $۱-p$، احتمال قطع نشدن برق در $x-۱$ ساعت متوالی $(۱-p)^{x-۱}$ است. بنابراین احتمال حداقل یک قطعی

$$P = \boxed{۱-(۱-p)^{x-۱}}$$

[8] Data Science Association
[9] Ethics Center for Engineering and Science website
[10] Society of Ethics Across the Curriculum website
[11] National Center for Education Statistics (NCES)

(ت) برای انتخاب مقدار مناسب p آیا دانشجو فقط به دوست خود (عباس) فکر می‌کند یا بیماران دیگر را نیز در نظر می‌گیرد

- **دیدگاه فردی:** اگر فقط به پرواز بازگشت فکر کنیم: احتمال قطع برق در x ساعت

$$۱ - (۱ - p)^x \leq ۰/۰۱$$

برای $x = ۲۴$ ساعت، باید $p \approx ۰/۰۰۰۴$ باشد.

- **دیدگاه اجتماعی و ملاحظه اخلاقی:** با در نظر گرفتن ایمنی همه بیماران مقدار p باید بسیار کوچک باشد مثلا:

$$p < ۰/۰۰۰۱.$$

- **نتیجه‌گیری:** مقدار p باید آنقدر کوچک باشد که برای عباس و سایر بیماران ایمنی قابل قبولی فراهم کند. پیشنهاد می‌شود:

$$\boxed{p \leq ۰/۰۰۰۱}$$

این مثال، به موارد ۲ و ۵ و ۶ از ملاحظات اخلاقی ده‌گانه بالا اشاره دارد.

آیا آماردانان بایداخلاق را تدریس کنند؟

استادان و اعضای هیئت علمی خود باید به‌عنوان الگوهای اخلاقی برای دانشجویان خود عمل کنند. آن‌ها باید در انجام تحقیقات علمی و نشر مقالات به اصول اخلاقی پایبند باشند تا دانشجویان از آن‌ها الگو بگیرند. در نهایت، آموزش اخلاق در آمار می‌تواند دانشجویان را قادر سازد تا در مسیر حرفه‌ای خود با مسئولیت‌پذیری، دقت و صداقت عمل کنند و به تقویت اعتبار و سلامت علمی جامعه آماری کمک نمایند. این سؤال پیش می‌آید که آیا آماردانان باید به طور رسمی محتوای درس اخلاق در آمار را آموزش دهند؟ از آنجا که پیامدهای اخلاقی کارهای آماری را تنها متخصصین آموزش‌دیده در آمار می‌توانند درک کنند، خود آماردانان علاقه‌مند بهترین گزینه برای آموزش ملاحظات اخلاقی در آمار هستند. این مسائل را نمی‌توان صرفاً به متخصصان اخلاق و فلاسفه واگذار نمود. برون‌سپاری آن به عنوان یک درس سرویسی یا تدریس مشترک با متخصصین اخلاق، راهکار دیگری است. البته دانشجویان آمار می‌توانند از طریق شرکت در دوره‌های ارائه شده توسط سایر گروه‌های آموزشی مرتبط، در مورد اخلاق حرفه‌ای آموزش ببینند. شرکت در چنین دوره‌هایی، اگرچه تمرکز اصلی بر کاربردهای آماری نیست، اما می‌تواند دانشجویان آمار را با اصول اخلاقی ارزشمندی آشنا کند.

بحث و نتیجه‌گیری

همان گونه که در این مقاله اشاره شد، در ایران در زمینه اخلاق علمی و حرفه‌ای، قوانین، دستورالعمل‌ها، شیوه‌نامه‌ها، نظام‌نامه‌ها و راهنماهای مختلفی در حوزه‌هایی مانند وکالت، مهندسی، روان‌شناسی و مشاوره، حسابداری و پزشکی، وجود دارد. در مورد ضرورت آموزش اخلاق در رشته‌های دانشگاهی نیز مقالاتی نوشته شده است. در برخی از رشته‌های مهندسی و علوم پزشکی دروسی با محتوای اخلاق علمی و حرفه‌ای وجود دارد. در زمینه اخلاق در آمار و آموزش آن، در برنامه‌های درسی دانشگاه‌های

کشور تاکنون کاری انجام نشده است. اخیرا انجمن آمار ایران با تشکیل «کارگروه اخلاق آماری»، «شیوه‌نامه اخلاقی برای کارهای آماری» خود را تدوین و منتشر نموده است. آموزش ملاحظات اخلاقی در برنامه درسی آمار، امری ضروری است که نباید در برابر پیچیدگی‌های فنی و نظری این رشته نادیده گرفته شود. به‌ویژه با توجه به کاربردهای گسترده آمار در حوزه‌های مختلف علمی، مهندسی، پزشکی، اقتصاد و تصمیم‌گیری‌های اجتماعی، توجه به اصول اخلاقی می‌تواند مانع از سوءاستفاده‌ها و خطاهای فاحش شود. جا دارد انجمن آمار ایران، با قرار دادن عنوان «اخلاق در آمار» در محورهای همایش‌های علمی انجمن در جهت ترویج اخلاق در آمار و جلب توجه آماردانان علاقه‌مند کشور به پژوهش در این زمینه، گام بردارد. گام‌های بعدی، می‌تواند تهیه محتوای آموزشی برای افزودن به برنامه درسی رشته آمار باشد. ترویج اصول اخلاقی در جامعه آمار و آموزش آن، به تقویت مسئولیت‌پذیری حرفه‌ای دانش‌آموختگان در انجام کارهای آماری کمک خواهد که به نوبه خود باعث ارتقای کیفیت و اعتبار نتایج حاصل از کارهای آماری آماری در جامعه و صنعت خواهد شد.

قدردانی و تشکر

این مقاله به پیشنهاد کارگروه اخلاق انجمن آمار ایران (متشکل از آقایان دکتر محمد آرشی، دکتر علی دولتی، دکتر علی رجالی، دکتر محمدرضا فقیهی، دکتر رحیم محمودوند و دکتر حمید نیلی ثانی) توسط نویسنده تهیه شده است. از نظرات ارزشمند این عزیزان در تدوین مقاله قدرانی می‌کنم.

مراجع

[۱] امین خندقی، م. و پاک مهر، ح. (۱۳۹۱). آموزش معیارهای اخلاق پژوهش: ضرورتی انکارناپذیر در برنامه‌های درسی آموزش عالی، فصلنامه اخلاق در علوم و فناوری، شماره ۴، صفحات ۱۳-۱.

[۲] بکروی، م. و همکاران (۱۳۹۹). اخلاق حرفه‌ای در علوم کامپیوتر، انتشارات دانشگاه آزاد.

[۳] ثقه‌الاسلامی، ع. (۱۴۰۰). برنامه‌ریزی درسی و تدوین سرفصل‌های آموزشی اخلاق پژوهش، فصلنامه اخلاق در علوم و فناوری، شماره ۳، صفحات ۱۰۱-۹۲.

[۴] خوشدست، ح. و سام، ع. (۱۳۸۸). ارائه الگویی برای آموزش مؤثر اخلاق مهندسی در دوره کارشناسی، فصلنامه آموزش مهندسی ایران، سال ۱۱، شماره ۴۳ ، صفحه ۱۰۸-۹۹.

[۵] رجبعلی‌پور، م.، رستگاری، ف. و قنبری‌زاده، ف. (۱۴۰۱). بررسی دروس مرتبط با اخلاق حرفه‌ای در برنامه درسی رشته‌های مختلف علوم پزشکی درایران، دو فصلنامه آموزش و اخلاق در پرستاری، دوره ۱۱، شماره ۳و۴، صفحات ۴۱-۳۲.

[۶] فراستخواه، م. (۱۳۸۵). اخلاق علمی رمز ارتقای آموزش عالی، فصلنامه اخلاق در علوم و فناوری، شماره ۱، صفحات ۲۷-۱۳.

[۷] منشور اخلاق حرفه‌ای وکالت (۱۳۹۹). هیات مدیره کانون وکلای دادگستری.

[۸] نظامنامه رفتار حرفه‌ای اخلاقی در مهندسی ساختمان (۱۳۹۵). مصوب وزارت راه و شهرسازی.

[۹] نظامنامه اخلاق حرفه‌ای سازمان نظام روان‌شناسی (۱۳۸۶). سازمان نظام روان‌شناسی و مشاوره.

[۱۰] آئین رفتار حرفه‌ای برای حسابداران (۱۳۹۶). جامعه حسابداران رسمی ایران.

[۱۱] قوانین، دستورالعمل‌ها و راهنماهای اخلاق در پژوهش‌های زیست پزشکی (۱۴۰۱). وزارت بهداشت ودرمان و آموزش پزشکی.

[۱۲] شیوه‌نامه اخلاقی برای کارهای آماری (۱۴۰۴)، کارگروه اخلاق انجمن آمار ایران.

[13] Alayont, F. (2022). A case for ethics in the mathematics major curriculum. Journal of Humanistic Mathematics, 12(2), 160-177.

[14] American Statistical Association (ASA). (2018). Ethical Guidelines for Statistical Practice.

[15] Chiodo, M. and Müller, D. (2025). Teaching resources for embedding ethics in mathematics: Exercises, projects, and handouts. arXiv preprint arXiv:2310.08467.

[16] Ernest, P. (2024). Ethics and Mathematics Education. Cham, Switzerland: Springer.

[17] Code, N. (1949). The Nuremberg code. Trials of war criminals before the Nuremberg military tribunals under control council law, 10(1949), 181-2.

[18] Doosti, H. (Ed). (2024). Ethics in Statistics : Opportunities and Challenges , Ethics International Press.

[19] Ellenberg, J. H. (1983). Ethical guidelines for statistical practice: A historical perspective. The American Statistician, 37(1), 1-4.

[20] Fiesler, C., Garrett, N. and Beard, N. (2020). What do we teach when we teach tech ethics? A syllabi analysis, pages 289–295 in Proceedings of the 51st ACM Technical Symposium on Computer Science Education.

[21] Declaration of Helsinki (1964). https://www.britannica.com/topic/Declaration-of-Helsinki.

[22] Hurwitz, S. and Gardenier, J. S. (2012). Ethical guidelines for statistical practice: The first 60 years and beyond. The American Statistician, 66(2), 99-103.

[23] Lesser, L. M. and Nordenhaug, E. (2004). Ethical statistics and statistical ethics: Making an interdisciplinary module. Journal of Statistics Education, 12(3).

[24] Panter, A. T. (Ed.). (2011). Handbook of ethics in quantitative methodology. Taylor Francis.

[25] Royal Statistical Society. 2014. Code of Conduct, RSS, Accessed 04.07.2024. https://rss.org.uk/about/policy-and-guidelines/code-of-conduct/.

[26] Smeyers, P. (2010). Educational research-the ethics and aesthetics of statistics. Springer Science, Business Media.

[27] Tractenberg, R. E. (2016). Why and How the ASA Ethical Guidelines should be integrated into every quantitative course. In Proceedings of the 2016 Joint Statistical Meetings, Chicago, IL, USA (pp. 517-535).

[28] Tractenberg, R. (2023). Ethical Practice of Statistics and Data Science: Ethical Practice of Statistics and Data Science. Ethics International Press.

[29] Vardeman, S. B. and Morris, M. D. (2003). Statistics and ethics: some advice for young statisticians. The American Statistician, 57(1), 21-26.

[30] Data Science Association https://www.datascienceassn.org

[31] Ethics Center for Engineering and Science website, https://onlineethics.org.

[32] Society of Ethics Across the Curriculum website, https://www.seac-online.org/resources/

[33] National Center for Education Statistics (NCES) https://nces.ed.gov/forum/dataethics/

۹

Ethics in Statistics and Probability

Ali Dolati[1]

[1]Yazd University, Department of Statistics

Abstract: Today, with the expansion of the use of statistical data and methods in waste, there are unethical practices such as the misuse of data, the presentation of biased and inaccurate content of the results of analyses, and the lack of transparency in the presentation of reports, in an informed or biased manner. Therefore, it is vital to strengthen public confidence in statistics and to maintain the health and credibility of the field, to educate and promote ethics. The purpose of this paper is to provide a model for the inclusion of ethics in statistics courses, so that students can learn about social and professional development in addition to learning the theoretical and applied concepts of statistics.

Keywords: Statistics education, ethics in statistics, curriculum, professional training, data privacy.

Mathematics Subject Classification (2020): $97k80$, $97k99$.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

نقش فرآیندهای تصادفی در هوش مصنوعی و علم داده: یک مطالعه موردی

اعظم راستین [۱]

دکترای آمار، دانشگاه شهید چمران اهواز

چکیده:

چکیده

در این مقاله، نقش تاثیرگذار فرآیندهای تصادفی در هوش مصنوعی و علم داده با تمرکز ویژه بر یادگیری تقویتی در محیط $Gridworld$ به همراه یک مطالعه‌ی موردی مورد بررسی قرار می‌گیرد. فرآیندهای تصادفی، به‌ویژه مدل‌سازی‌های مارکوفی، چارچوبی ریاضی و مفهومی برای تحلیل و پیش‌بینی رفتار عامل‌های هوشمند در محیط‌های نامعین فراهم می‌آورند. در این راستا، الگوریتم $Q-Learning$ در سه سناریوی مختلف (محیط قطعی، محیط با نویز تصادفی، و محیط با پاداش‌های تصادفی) پیاده‌سازی و تحلیل شده است. نتایج عددی، نمودارهای همگرایی، و مسیرهای یادگیری عامل‌ها نشان می‌دهد که چگونه فرآیندهای تصادفی درک دقیق‌تری از تعامل عامل با محیط فراهم کرده و بر کیفیت یادگیری تأثیر می‌گذارند. این مقاله نشان می‌دهد که ادغام تئوری احتمال و فرآیندهای تصادفی در ساختارهای هوش مصنوعی، مسیر توسعه الگوریتم‌های پایدارتر و تطبیق‌پذیرتر را هموار می‌سازد.

واژه‌های کلیدی: فرآیند تصادفی، یادگیری تقویتی، MDP، $Gridworld$، یادگیری سیاست، ارزش‌گذاری، مدل‌سازی محیط

کد موضوع‌بندی ریاضی (۲۰۲۰): 60J20، 68T05، 68T20.

---

۱ مقدمه

فرآیندهای تصادفی یکی از ابزارهای بنیادی در مدل‌سازی پدیده‌هایی هستند که در آن‌ها عدم قطعیت و تصادف نقش دارند. در دهه‌های اخیر، اهمیت این فرآیندها در حوزه‌هایی مانند هوش مصنوعی و علم داده نیز به طور چشمگیری افزایش یافته است. الگوریتم‌های یادگیری ماشین، تحلیل سری‌های زمانی، یادگیری تقویتی و بسیاری دیگر از شاخه‌های هوش مصنوعی برای مدل‌سازی رفتارهای پیچیده و پیش‌بینی‌های دقیق، به نظریه‌های تصادفی متکی هستند. فرآیندهای تصادفی امکان مدل‌سازی

[۱] سخنران، rastinstat@gmail.com

پویایی‌هایی را فراهم می‌کنند که در آن‌ها آینده سیستم به صورت احتمالی و بر اساس گذشته یا وضعیت فعلی تعیین می‌شود. به ویژه، زنجیره‌های مارکوف، فرآیندهای تصمیم‌گیری مارکوف، و فرآیندهای گاوسی، ابزارهایی کلیدی در توسعه سیستم‌های هوشمند و تحلیل داده‌های پیچیده هستند. در سال‌های اخیر، هوش مصنوعی و علم داده به‌شدت به مفاهیم احتمالاتی و فرآیندهای تصادفی متکی شده‌اند تا بتوانند رفتار پیچیده و غیرقطعی سیستم‌های واقعی را مدل‌سازی کنند. یکی از حوزه‌هایی که به‌طور مستقیم از این ابزارها بهره‌مند است، یادگیری تقویتی است. یادگیری تقویتی یکی از شاخه‌های مهم یادگیری ماشینی است که به‌طور خاص برای تصمیم‌گیری در محیط‌های پویا و تعاملی طراحی شده‌است. در این روش، عامل از طریق تعامل با محیط و بر اساس بازخورد پاداش، سیاستی بهینه برای بیشینه‌سازی پاداش بلندمدت فرا می‌گیرد (سوتون و باروت ، ۲۰۱۸).

در سال‌های اخیر، یادگیری تقویتی به عنوان یکی از زیرشاخه‌های مهم هوش مصنوعی، نقش چشم‌گیری در حل مسائل تصمیم‌گیری در محیط‌های نامطمئن و تصادفی ایفا کرده است. مطالعات متعددی بر اهمیت پایداری الگوریتم‌های یادگیری تقویتی در مواجهه با تغییرات دینامیکی محیط تأکید داشته‌اند (لادوز و همکاران ، ۲۰۲۲). با گسترش کاربردهای یادگیری تقویتی در محیط‌های بزرگ و با فضای کنش گسترده، روش‌های جدیدی برای مقابله با پیچیدگی‌های محاسباتی ارائه شده‌اند. به عنوان نمونه، فوراتی و همکاران (۲۰۲۴) یک نسخه تصادفی از الگوریتم $Q-learning$ را معرفی کردند که امکان یادگیری در محیط‌هایی با فضای کنش بسیار بزرگ را فراهم می‌کند.

فرآیندهای تصادفی نقشی کلیدی در مدل‌سازی عدم قطعیت در یادگیری تقویتی دارند. مطالعات پیشین نشان داده‌اند که مدل‌سازی دقیق نویزها و تصادفی بودن پاداش یا انتقال حالت می‌تواند به طراحی الگوریتم‌های پایدارتر و واقع‌گرایانه‌تر منجر شود (مراجعه شود به ترون و همکاران (۱۹۹۵) و کرنز و سینگز (۲۰۰۲)).

در این مقاله، ما محیط $Gridworld$ را به‌عنوان بستری ساده ولی گویا برای بررسی تأثیر فرآیندهای تصادفی در مدل‌سازی عدم قطعیت انتخاب کرده‌ایم. با استفاده از الگوریتم $Q-learning$ و در سه سناریوی مجزا- محیط قطعی، محیط با نویز در انتقال، و محیط با پاداش تصادفی- روند یادگیری و عملکرد عامل مورد تحلیل قرار گرفته‌اند.

۲ مفاهیم پایه

فرآیند تصادفی به توالی از متغیرهای تصادفی، $\{X_t\}_{t\in T}$، گفته می‌شود که به صورت تابعی از پارامتر t تعریف شده‌اند که در آن T یک مجموعه اندیس گذار بوده و می‌تواند گسسته (مثلاً زمان‌های طبیعی) یا پیوسته (مثلاً بازه‌ای از اعداد حقیقی) باشد. یک مورد خاص از فرآیندهای تصادفی، زنجیره مارکوف است که دارای ویژگی زیر می‌باشد:

$$P(X_{t+1}=x|X_t=x_t,X_{t-1}=x_{t-1},\ldots)=P(X_{t+1}=x|X_t=x_t),t\in T$$

به این خاصیت، ویژگی مارکوفی یا حافظه کوتاه گفته می‌شود. زنجیره‌های مارکوف به طور گسترده‌ای در الگوریتم‌های یادگیری، مدل‌سازی تصمیم‌گیری، و تحلیل سری‌های زمانی به کار می‌روند. فرآیندهای تصادفی مبنای ریاضی مدل‌سازی رفتار سیستم‌هایی هستند که با گذر زمان و تحت تأثیر عدم قطعیت تغییر می‌کنند. در یادگیری تقویتی، به‌ویژه از فرآیندهای تصمیم‌گیری مارکوفی (MDP) استفاده می‌شود (پورتمن ، ۲۰۰۵).

تعریف ۱.۲. فرآیند تصمیم‌گیری مارکوف *(MDP)*: یک MDP که به اختصار آن را با $M=(S,A,P,R,\gamma)$ نشان می‌دهیم،

شامل مجموعه‌ای از حالات ممکن S، اعمال ممکن A، تابع احتمال انتقال $P(s'|s,a)$ (احتمال انتقال از حالت s به حالت s' با انتخاب عمل a) و تابع پاداش $R(s,a)$ (پاداش دریافتی برای اجرای عمل a در حالت s) و ضریب تنزیل آینده $\gamma \in [0,1)$ می‌باشد.

توجه کنید که طبق خاصیت مارکوفی، $P(s_{t+1}|s_t,a_t,s_{t-1},...,s_0) = P(s_{t+1}|s_t,a_t)$. در واقع، این خاصیت تضمین می‌کند که وضعیت آینده فقط به وضعیت و عمل فعلی بستگی دارد، نه به مسیر طی‌شده قبلی.

تعریف ۲.۲. سیاست *(Policy)*: تابعی که مشخص می‌کند عامل در هر حالت چه عملی انجام دهد. سیاست بهینه π^* باعث بیشینه شدن امید ریاضی پاداش تجمعی می‌شود.

تعریف ۳.۲. تابع Q: بیانگر ارزش حالت–عمل تحت یک سیاست خاص است:

$$Q(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t,a_t) \mid s_0 = s, a_0 = a\right]$$

هدف اصلی در MDP، یافتن سیاست بهینه π^* است به‌طوری که تابع ارزش Q بیشینه شود.

یادگیری تقویتی نوعی یادگیری ماشینی است که در آن عامل با تعامل با محیط، سیاست بهینه را از طریق آزمون و خطا می آموزد. در این تحقیق از الگوریتم $Q-Learning$ برای این منظور استفاده شده است. $Q-Learning$ یکی از الگوریتم‌های پایه در یادگیری تقویتی است که بدون مدل محیط و به صورت زیر کار می‌کند:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$$

که در آن α, r, s' و γ به ترتیب بیانگر نرخ یادگیری، پاداش دریافتی در انتقال، حالت بعدی پس از انجام a در s و ضریب تنزیل است. با گذر زمان، مقدارهای Q به سمت مقدارهای بهینه همگرا می‌شوند (واتکینز و دایان ، ۱۹۹۲).

۳ مطالعه موردی: یادگیری تقویتی در محیط $Gridworld$

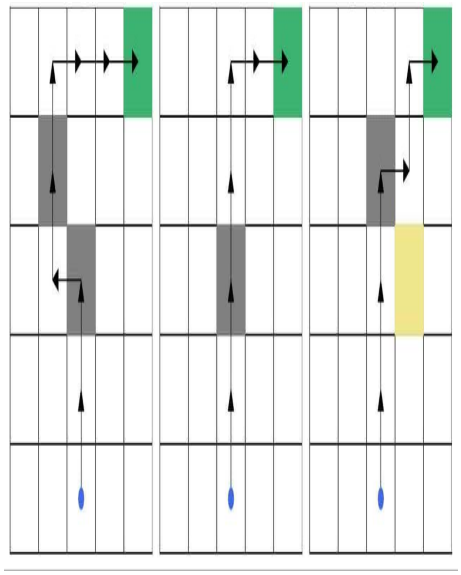
محیط $Gridworld$ یک شبکه دوبعدی است که هر خانه از آن یک حالت را نشان می‌دهد. در این شبکه یک عامل ($Agent$) می‌تواند به چهار جهت اصلی (بالا، پایین، چپ، راست) حرکت کند. بعضی خانه‌ها می‌توانند موانع یا چاله‌هایی باشند که مانع حرکت عامل می‌شوند. هدف عامل، در حالی که از موانع اجتناب و پاداش حداکثری دریافت کند، رسیدن به خانه هدف می‌باشد. در این مطالعه سه نوع محیط تعریف شده است:

۱. **محیط قطعی: انتقال دقیق و بدون نویز به جهت دلخواه**

۲. **محیط نویزدار: احتمال ٪ ۸۰ برای حرکت مورد نظر و ۲۰ ٪ برای انحراف به جهت دیگر**

۳. **محیط با پاداش‌های تصادفی: مقدار پاداش در برخی خانه‌ها به صورت تصادفی تغییر می‌کند**

در شکل ۱ این سه محیط متفاوت نمایش داده شده است. این تصویر سه محیط مختلف را در قالب شبکه ۵×۵ نشان می‌دهد که در هر کدام شرایط متفاوتی حاکم است: در محیط قطعی، حرکت عامل با دقت کامل در جهت اعمال‌شده صورت می‌گیرد و پاداش‌ها ثابت هستند. در محیط نویزی، عامل با احتمال ۰.۸ در جهت اعمال‌شده حرکت می‌کند و با احتمال ۰.۲



شکل ۱: نمایش سه محیط $Gridworld$ با ویژگی‌های مختلف: محیط قطعی، محیط با نویز، محیط با پاداش تصادفی.

به جهت‌های دیگر می‌رود. و در محیط با پاداش تصادفی، مقصد حرکت مشخص است، اما مقدار پاداش از یک توزیع تصادفی (مثلاً یکنواخت یا نرمال) گرفته می‌شود.

در این مطالعه، محیط $Gridworld$ شامل یک شبکه ۵×۵ است که عامل باید از خانه شروع به خانه هدف برسد. عامل در هر مرحله می‌تواند یکی از چهار جهت را انتخاب کند. هدف یافتن سیاست بهینه برای بیشینه کردن پاداش است. پس از اجرای الگوریتم $Q - Learning$، سیاست بهینه استخراج شده نشان‌دهنده بهترین جهت حرکت از هر خانه به سمت هدف است. به عنوان مثال، اگر در خانه‌ای فلش به سمت راست باشد ($\rightarrow$)، به معنی آن است که حرکت به سمت راست بیشترین ارزش را دارد. این سیاست به عامل کمک می‌کند تا با پیمودن مسیر بهینه، در کوتاه‌ترین زمان و بیشترین پاداش به هدف برسد.

پس از آموزش، سیاست بهینه به صورت جهت حرکت از هر خانه به سوی هدف نمایش داده می‌شود. این نتایج نشان می‌دهند که فرآیندهای تصادفی چگونه می‌توانند مسیرهای بهینه را در محیط‌های نامعین بیاموزند.

برای هر یک از سه سناریو، الگوریتم $Q - learning$ به فرم زیر

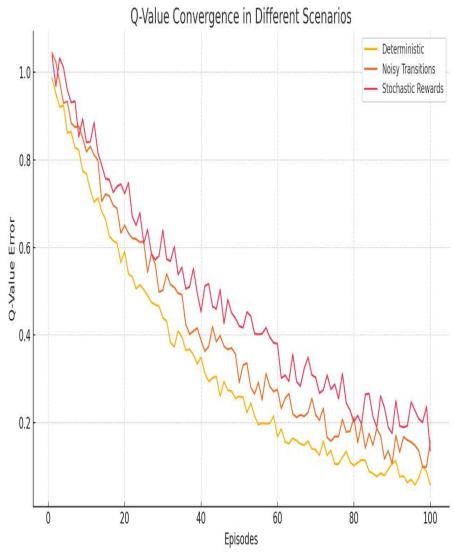
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_a Q(s_{t+۱}, a) - Q(s_t, a_t) \right]$$

در صد اپیزود و با پارامترهای $\alpha = ۰.۱$ و $\gamma = ۰.۹۵$ اجرا شده و نمودار همگرایی و مسیر یادگیری استخراج شده است. در شکل ۲ همگرایی مقدار Q در سه سناریو نمایش داده شده است. همان‌طور که مشاهده می‌شود، مقدار خطای Q در محیط قطعی سریع‌تر کاهش می‌یابد. در محیط‌های نویزی و دارای پاداش تصادفی، نرخ همگرایی کندتر و همراه با نوسان است.

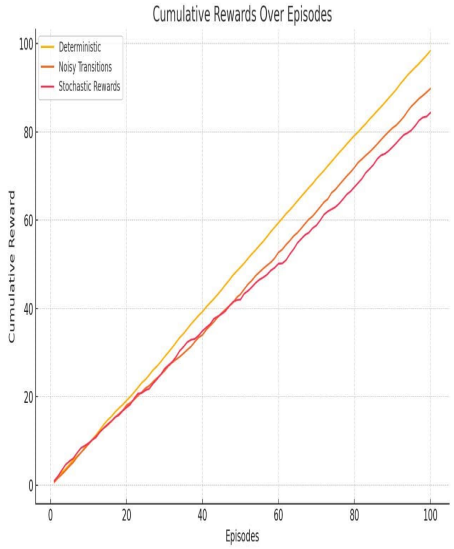
در شکل ۳ پاداش تجمعی در طول اپیزودها برای سه محیط نمایش داده شده است. در محیط قطعی، پاداش تجمعی با شیب مثبت مشخص رشد می‌کند، ولی در محیط‌های تصادفی، پاداش‌ها بی‌ثبات‌ترند.

در شکل ۴ مسیر عامل در شبکه نهایی نمایش داده شده است. عامل در طی فرآیند یادگیری، مسیری بهینه برای رسیدن به هدف پیدا کرده است.

تجزیه‌وتحلیل این نمودارها نشان می‌دهد که تصادفی بودن در پاداش یا انتقال، موجب کاهش سرعت یادگیری و همچنین



شکل ۲: نمودار همگرایی مقدار Q برای سه حالت

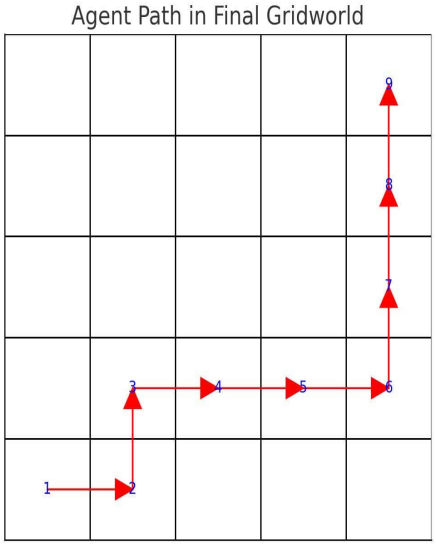


شکل ۳: نمودار پاداش تجمعی در طول اپیزودها برای سه حالت

افزایش نوسان در عملکرد عامل می‌شود. بنابراین، استفاده از تکنیک‌های مقاوم در برابر نویز و سیاست‌های اکتشافی مناسب، اهمیت زیادی دارند.

تحلیل نتایج بخش اصلی این مقاله است که نقش فرآیندهای تصادفی را آشکار می‌کند.

- **محیط قطعی: در این حالت، عامل به سرعت سیاست بهینه را می‌یابد. همگرایی الگوریتم بسیار سریع است. مسیر عامل کاملاً مستقیم و بهینه است. پاداش تجمعی در اپیزودهای پایانی به حداکثر مقدار ممکن می‌رسد.**
- **محیط با نویز حرکتی: عامل مجبور است سیاست محافظه‌کارانه‌تری اتخاذ کند، زیرا هر حرکت ممکن است منجر به حالت ناخواسته شود. این موضوع باعث نوسانات بیشتر در مقادیر Q و تأخیر در همگرایی می‌شود. با این حال،**



شکل ۴: مسیر طی‌شده عامل در شبکه $Gridworld$ نهایی

الگوریتم همچنان قادر به یادگیری سیاست نسبتاً خوب است.

- محیط با پاداش‌های تصادفی: در این حالت عامل باید با تغییرات تصادفی پاداش‌ها سازگار شود. مقدار Q دچار نوسانات دائمی می‌شود. مسیر عامل نیز گاه به گاه تغییر می‌کند، بسته به اینکه کدام مسیر اخیراً پاداش بیشتری داشته است. همگرایی کندتر و نوسانی‌تر از سایر سناریوها است.

بحث و نتیجه‌گیری

مطالعه حاضر نشان داد که فرآیندهای تصادفی نقش کلیدی در تصمیم‌گیری هوشمند دارند. حتی در محیط‌های ساده‌ای مانند $Gridworld$، وجود نویز و تصادفی‌بودن پاداش می‌تواند یادگیری عامل را با چالش روبرو کند. $Q-learning$ با وجود این چالش‌ها، قادر است در بلندمدت سیاست‌های کارآمدی یاد بگیرد. مطالعه موردی ارائه‌شده به‌خوبی نشان می‌دهد که چگونه مفاهیم ریاضی می‌توانند به راه‌حل‌های عملی در مسائل هوش مصنوعی منجر شوند.

برای کارهای آینده، استفاده از الگوریتم‌های پیشرفته‌تر مانند DQN و بررسی اثر حافظه و شبکه‌های عصبی پیشنهاد می‌شود.

مراجع

Bertsekas, D. P. (2017). *Dynamic Programming and Optimal Control*. Athena Scientific.

Durrett, R. (2019). Probability: Theory and Examples. Cambridge University Press.

Kearns, M., and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. Machine Learning, 49(2), 209-232.

Ladosz, P, . Słowik, M., and Dębowski, M. (2022). Exploration in deep reinforcement learning: A survey, Information Fusion, vol. 89, pp. 1–23, 2022.

Fourati, F., Aggarwal, V., and Alouini, M. S. (2024). Stochastic Q-learning for Large Discrete Action Spaces. In Proceedings of the 41st ICML, PMLR 235:13734–13759.

Mnih, V., Kavukcuoglu, K. and Silver, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.

Puterman, M. L. (2005). Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley.

Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

Thrun, S, Burgard, W., and Fox, D. (1995). Probabilistic robotics. MIT press.

Watkins, C. J. C. H., and Dayan, P. (1992). ”Q-learning”. Machine Learning, 8(3-4), 279–292.

The Role of Stochastic Processes in Artificial Intelligence and Data Science: A Case Study

Azam Rastin[1]

[1]**Shahid Chamran university of Ahvaz**

Abstract:

Stochastic processes are a foundational component in modeling uncertainty and dynamics within Artificial Intelligence (AI) and Data Science. This paper presents a theoretical and empirical investigation into the role of stochastic processes in Reinforcement Learning (RL), focusing on a case study using the classical Gridworld environment. We formalize the decision-making process as a Markov Decision Process (MDP), and examine how stochastic transition dynamics influence learning policies and value estimation. Experimental results demonstrate how varying the stochasticity of the environment impacts policy convergence, optimality, and stability. These findings provide critical insights into the importance of accurate modeling of environment dynamics for robust AI systems.

Keywords: Most important keywords of article (3-5 items) should appear here.

Mathematics Subject Classification (2020): 60J20، 68T05، 68T20.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

فرایند INAR(1)-PJ : یک جایگزین جدید فرایند INAR(1) پواسونی

مهدی راسخی [۱]

[۱]گروه آمار، دانشکده علوم ریاضی و آمار، دانشگاه ملایر، ملایر، ایران

چکیده: در این مقاله، یک مدل خودبازگشتی مرتبه اول با مقادیر صحیح نامنفی معرفی می‌شود که نوفه فرآیند از یک توزیع یک پارامتری جدید پیروی می‌کند. برای برآورد پارامترهای این فرآیند، دو روش مختلف برآوردیابی در نظر گرفته شده است: حداکثر درستنمایی شرطی و روش یول–والکر. همچنین در یک مطالعه شبیه‌سازی عملکرد و کارایی این روش‌ها ارزیابی شده است. علاوه بر این، کاربرد فرآیند پیشنهادی از طریق تحلیل یک مجموعه داده واقعی ارائه شده و در این قسمت برتری آن نسبت به مدل INAR(1) پواسونی نشان داده شده است.

واژه‌های کلیدی: توزیع احتمال جوچز، داده‌های شمارشی، بیش‌پراکنش، فرایند خودبازگشتی عدد صحیح مرتبه اول .

کد موضوع‌بندی ریاضی (۲۰۲۰): 62M10، 62F10، 62E15.

۱ مقدمه

داده‌های سری‌های زمانی با مقادیر اعداد صحیح نامنفی در حوزه‌های گوناگونی همچون علوم اجتماعی، اپیدمیولوژی، بیمه، مهندسی صنایع و بسیاری دیگر از زمینه‌ها به طور گسترده‌ای دیده می‌شوند. برای مدل‌سازی بهینه این نوع داده‌ها که دارای وابستگی زمانی نیز هستند، مدل خودبازگشتی عدد صحیح مرتبه اول ((INAR(1)) پیشنهاد شده است که مکنزی (۱۹۸۶) و الاش و الزید (۱۹۸۷) طور مستقل به معرفی آن پرداختند. ویژگی مشترکی که اغلب در این نوع سری‌های زمانی دیده می‌شود، بیش‌پراکنش است؛ یعنی واریانس داده‌ها بیشتر از میانگین آنهاست. با توجه به اینکه توزیع پواسون اغلب برای مدل‌سازی داده‌های شمارشی با بیش‌پراکنش مناسب نیست، برخی از پژوهشگران مدل‌های INAR(1) مبتنی بر توزیع‌های نوفه انعطاف‌پذیرتری را پیشنهاد کرده‌اند. از جمله این مدل‌ها می‌توان به موارد زیر اشاره کرد: مدل INAR(1) گسسته هندسی توسط همکاران و جازی (۲۰۱۲) ، فرایند INAR(1) مبتنی بر خانواده توزیع‌های کاتس که توسط کیم و لی (۲۰۱۷) معرفی شد. در راستای پیشرفت‌های اخیر در مدل‌سازی داده‌های شمارشی، در این مقاله توزیع

[۱] سخنران، m.rasekhi@malayeru.ac.ir

پواسون–جوچز (PJ) معرفی شده است: یک توزیع گسسته جدید با یک پارامتر که از ترکیب توزیع‌های پواسون و جوچز (اکبیری و امبگو (۲۰۲۲)) حاصل شده است. توزیع استفاده شده در ترکیب فوق (توزیع جوچز) در مقایسه با توزیع گاما دارای تعداد پارامتر کمتری است اما از لحاظ انعطاف پذیری در شکل تابع چگالی و میزان شاخص‌های چولگی و کشیدگی همانند توزیع گاما عمل می کند.

توزیع PJ به ویژه در مدل‌بندی داده‌ها با ویژگی بیش‌پراکنش، کارا است و دارای فرم‌های بسته و ساده‌ای برای تابع جرم احتمال و گشتاورهای آن می‌باشد. با استفاده از این توزیع به عنوان مکانیزم نوفه، یک فرآیند خودبازگشتی عدد صحیح مرتبه اول را توسعه داده و بصورت INAR(1)-PJ نامگذاری شده است. فرآیند پیشنهادی روی یک مجموعه داده واقعی اعمال شده است تا کاربردپذیری و انعطاف‌پذیری آن نشان داده شود. همچنین، مطالعات شبیه‌سازی برای ارزیابی دقت و کارایی روش‌های برآوردیابی پارامترها در چارچوب پیشنهادی انجام شده است. مقاله به شرح زیر سازمان‌دهی گردیده است: در بخش ۲ ، توزیع PJ معرفی و ویژگی های آماری آن به طور کامل بررسی شده است. در بخش ۳ ، یک فرایند خودبازگشتی عدد صحیح مبتنی بر نوفه‌های PJ پیشنهاد شده و خواص ساختاری مهم آن ارائه شده است. در بخش ۴ ، دو روش برآوردیابی برای پارامترهای مدل INAR(1)-PJ مورد بحث قرار می‌گیرد: روش حداکثر درستنمایی شرطی و روش یول–والکر. در بخش ۵ ، نتایج یک مطالعه شبیه‌سازی به منظور ارزیابی دقت و کارایی این روش‌های برآوردیابی گزارش شده‌اند. در بخش ۶ ، فرآیند پیشنهادی روی یک مجموعه داده واقعی اعمال شده است تا کاربرد عملی آنها نشان داده شود.

۲ توزیع پواسن–جوچز

تابع جرم احتمال (PMF) توزیع جوچز به صورت زیر تعریف می‌شود:

$$f(x;\theta) = \frac{\theta^{۴}}{\theta^{۳}+\theta^{۲}+۶}(۱+x+x^{۳})e^{-\theta x} \quad x > ۰\,, \tag{۱.۲}$$

که $\theta > ۰$. تابع توزیع تجمعی (CDF) متناظر با توزیع جوچز به صورت زیر بیان می‌شود:

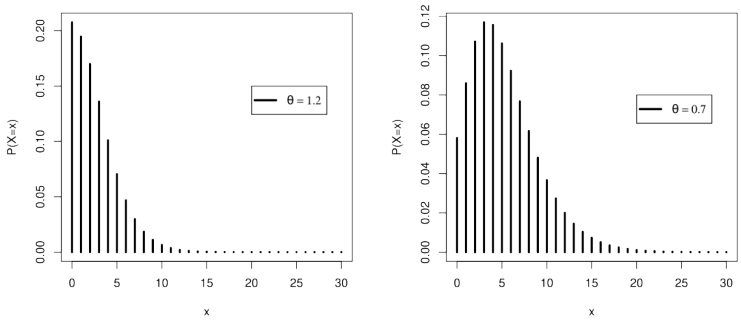
$$F(x;\theta) = ۱ - \left(۱ + \frac{\theta x(\theta^{۲}+\theta^{۲}x^{۲}+۳\theta x+۶)}{\theta^{۳}+\theta^{۲}+۶}\right)e^{-\theta x} \quad x \geq ۰.$$

در ادامه، یک توزیع پواسون ترکیبی جدید یک پارامتری را تعریف می‌کنیم که از ترکیب توزیع پواسون با توزیع جوچز حاصل می‌شود. فرض کنید متغیر تصادفی λ دارای توزیع جوچز با پارامتر θ باشد که در برابری (۱.۲) تعریف شده است. همچنین فرض کنید که توزیع شرطی X به شرط λ ، یک توزیع پواسون با میانگین λ باشد. در این صورت، توزیع حاشیه‌ای X که از حذف λ به دست می‌آید، مربوط به توزیع پواسون–جوچز (PJ) است و دارای تابع جرم احتمال (PMF) به صورت زیر می‌باشد:

$$p(x;\theta) = \left(\frac{\theta^{۴}}{\theta^{۳}+\theta^{۲}+۶}\right)\left(\frac{۱}{(\theta+۱)^{x+۱}} + \frac{(x+۱)}{(\theta+۱)^{x+۲}} + \frac{(x+۳)(x+۲)(x+۱)}{(\theta+۱)^{x+۴}}\right), \quad x = ۰, ۱, ۲, \ldots \tag{۲.۲}$$

فرض کنید X نشان‌دهنده متغیر تصادفی با تابع چگالی (۲.۲) باشد؛ در این صورت می‌نویسیم $X \sim PJ(\theta)$. شکل ۱ انواع مختلفی از شکل‌های تابع جرم احتمال (PMF) توزیع PJ را تحت مقادیر مختلف پارامتر نشان می‌دهد. تابع جرم احتمال این توزیع به ازای مقادیر متفاوت پارامتر، دارای ساختار نزولی و چوله به راست است. دو گشتاور اول توزیع PJ به‌ترتیب به فرم زیر هستند:

$$\begin{aligned} E(X) &= \left(\tfrac{\theta^{۴}}{\theta^{۳}+\theta^{۲}+۶}\right)\left\{\sum_{x=۰}^{\infty}\tfrac{x}{(\theta+۱)^{x+۱}} + \sum_{x=۰}^{\infty}\tfrac{x(x+۱)}{(\theta+۱)^{x+۲}} + \sum_{x=۰}^{\infty}\tfrac{x(x+۳)(x+۲)(x+۱)}{(\theta+۱)^{x+۴}}\right\} \\ &= \left(\tfrac{\theta^{۴}}{\theta^{۳}+\theta^{۲}+۶}\right)\left\{\tfrac{۱}{\theta^{۲}} + \tfrac{۲}{\theta^{۳}} + \tfrac{۲۴}{\theta^{۵}}\right\} = \tfrac{\theta^{۳}+۲\theta^{۲}+۲۴}{\theta(\theta^{۳}+\theta^{۲}+۶)}, \end{aligned} \tag{۳.۲}$$



شکل ۱: نمودار تابع جرم احتمال توزیع PJ تحت مقادیر مختلف پارامتر

$$\begin{aligned} E(X^{۲}) &= \left(\frac{\theta^{۴}}{\theta^{۳}+\theta^{۲}+۶}\right)\left\{\sum_{x=۰}^{\infty}\frac{x^{۲}}{(\theta+۱)^{x+۱}}+\sum_{x=۰}^{\infty}\frac{x^{۲}(x+۱)}{(\theta+۱)^{x+۲}}+\sum_{x=۰}^{\infty}\frac{x^{۲}(x+۳)(x+۲)(x+۱)}{(\theta+۱)^{x+۴}}\right\} \\ &= \left(\frac{\theta^{۴}}{\theta^{۳}+\theta^{۲}+۶}\right)\left\{\frac{\theta+۲}{\theta^{۳}}+\frac{۲(\theta+۳)}{\theta^{۴}}+\frac{۲۴(\theta+۵)}{\theta^{۶}}\right\} = \frac{\theta^{۴}+۴\theta^{۳}+۶\theta^{۲}+۲۴\theta+۱۲۰}{\theta^{۲}(\theta^{۳}+\theta^{۲}+۶)}. \end{aligned} \tag{۴.۲}$$

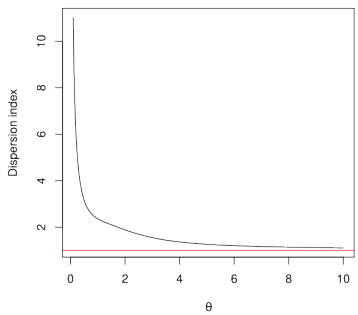
بر اساس روابط (۳.۲) و (۴.۲) واریانس توزیع PJ عبارتست از

$$V(X) = \frac{\theta^{۷}+۴\theta^{۶}+۶\theta^{۵}+۳۲\theta^{۴}+۱۲۰\theta^{۳}+۶۰\theta^{۲}+۱۴۴\theta+۱۴۴}{\theta^{۲}(\theta^{۳}+\theta^{۲}+۶)^{۲}}.$$

همچنین شاخص پراکنش به فرم زیر حاصل می شود.

$$DI(X) = \frac{\theta^{۷}+۴\theta^{۶}+۶\theta^{۵}+۳۲\theta^{۴}+۱۲۰\theta^{۳}+۶۰\theta^{۲}+۱۴۴\theta+۱۴۴}{\theta(\theta^{۳}+\theta^{۲}+۶)(\theta^{۳}+۲\theta^{۲}+۲۴)}.$$

شکل ۲ نشان می‌دهد که برای تمامی مقادیر پارامتر θ ، شاخص پراکنشی بزرگ‌تر از ۱ است. مقدار این شاخص به ازای مقادیر بزرگ پارامتر θ به عدد ۱ نزدیک می‌شود. این رفتار، کارایی این توزیع در مدل‌سازی داده‌های شمارشی با بیش‌پراکنش را تأیید می‌کند.



شکل ۲: نمودار شاخص پراکنش توزیع PJ تحت مقادیر مختلف پارامتر

الگوریتم زیر می‌تواند برای تولید نمونه تصادفی از توزیع PJ با بهره بردن از روش معکوس تابع توزیع و بر اساس توزیع جوجز استفاده شود: (۱) تولید U_i، $i = ۱, ۲, ..., n$ جاییکه $Ui \sim U(۰, ۱)$.

(۲) یافتن ریشه برابری زیر بر اساس λ_i به کمک دستور uniroot در نرم افزار R.

$$\ln\left(1+\frac{\theta\lambda_i(\theta^2+\theta^2{\lambda_i}^2+3\theta\lambda_i+6)}{\theta^3+\theta^2+6}\right)-\theta\lambda_i-\ln(1-U_i)=0$$

(۳) تولید X_i از توزیع $Poisson(\lambda_i)$.

۳ فرایند INAR(1) مبتنی بر نوفه‌های PJ

فرض کنید $\{X_t;t\in\mathbb{Z}\}$ نشان دهنده فرآیند INAR(1) بر اساس تعریف زیر باشد.

$$X_t=p\circ X_{t-1}+\varepsilon_t,\quad t\in\mathbb{Z}, \tag{۱.۳}$$

که در آن $0\leq\alpha<1$ و $\{\varepsilon_t;t\in\mathbb{Z}\}$ دنباله ای از متغیرهای تصادفی هم توزیع و مستقل با مقادیر صحیح نامنفی و میانگین $E(\varepsilon_t)=\mu_\varepsilon$ و واریانس $V(\varepsilon_t)=\sigma_\varepsilon^2$ است. فرض کنید نوفه $\{\varepsilon_t;t\in\mathbb{Z}\}$ مستقل از مشاهدات گذشته $X_{t-k},\ k\geq1$ است. عملگر $\circ$ نشان دهنده عملگر رقیق‌ساز دوجمله ای است که به فرم زیر تعریف می‌شود:

$$p\circ X_{t-1}:=\sum_{j=1}^{X_{t-1}}W_j$$

که در آن $\{W_j;j\geq1\}$ نشان دهنده دنباله ای از متغیرهای تصادفی همتوزیع و مستقل با توزیع برنولی با احتمال موفقیت p است. اگر $0\leq p<1$ این فرآیند ایستا و درغیر اینصورت فرآیند ناایستا است. برای تشخیص ساختار حرکت در فرآیند، احتمال انتقال یک مرحله ای به شرح زیر است:

$$P(X_t=k\,|X_{t-1}=l)=\sum_{i=1}^{\min(k,l)}\binom{l}{i}p^i(1-p)^{l-i}P(\varepsilon_t=k-i),\ \ k,l\geq0$$

که در آن $0<p<1$. برای ایجاد یک فرآیند تصادفی INAR(1) جدید، دنباله متغیرهای تصادفی $\{\varepsilon_t;t\in\mathbb{Z}\}$ که از توزیع PJ پیروی می‌کند را به عنوان متغیر های نوفه فرآیند در برابری (۱.۳) لحاظ می‌شود. درنتیجه احتمال انتقال یک مرحله فرآیند تصادفی جدید INAR(1)-PJ به فرم زیر حاصل می‌شود.

$$\begin{aligned}P(X_t=k\,|X_{t-1}=l)=\left(\tfrac{\theta^4}{\theta^3+\theta^2+6}\right)\sum_{i=1}^{\min(k,l)}\binom{l}{i}p^i(1-p)^{l-i}\Big(\tfrac{1}{(\theta+1)^{k-i+1}}\\+\tfrac{(k-i+1)}{(\theta+1)^{k-i+2}}+\tfrac{(k-i+3)(k-i+2)(k-i+1)}{(\theta+1)^{k-i+4}}\Big).\end{aligned} \tag{۲.۳}$$

از این پس، فرایند توصیف شده در معادله (۲.۳) فرآیند مربوطه را با INAR(1)-PJ نامگذاری میکنیم. بر اساس روش ویب (۲۰۱۸) میانگین، واریانس و شاخص پراکنش فرایند INAR(1)-PJ به‌ترتیب به صورت زیر بیان می‌شوند:

$$\mu_X=\frac{\theta^3+2\theta^2+24}{\theta(\theta^3+\theta^2+6)(1-p)},$$

$$\sigma_X^2=\frac{\left(\begin{array}{l}(p+1)\theta^7+(3p+4)\theta^6+(2p+6)\theta^5+(30p+32)\theta^4\\+(36p+120)\theta^3+60\theta^2+(144p+144)\theta+144\end{array}\right)}{(1-p^2)\theta^2(\theta^3+\theta^2+6)^2},$$

$$DI_X = \frac{\frac{\theta^{۷}+۴\theta^{۶}+۶\theta^{۵}+۳۲\theta^{۴}+۱۲۰\theta^{۳}+۶۰\theta^{۲}+۱۴۴\theta+۱۴۴}{\theta(\theta^{۳}+\theta^{۲}+۶)(\theta^{۳}+۲\theta^{۲}+۲۴)}+p}{۱+p}.$$

تحت فرضیات مدل، بر اساس روش ارائه شده در الزید و الاش (۱۹۸۸) ، میانگین و واریانس شرطی فرآیند به ترتیب به شرح زیر حاصل می‌شود:

$$E(X_t\,|X_{t-۱}) = pX_{t-۱} + \frac{\theta^{۳}+۲\theta^{۲}+۲۴}{\theta(\theta^{۳}+\theta^{۲}+۶)},$$

و

$$V(X_t\,|X_{t-۱}) = p(۱-p)X_{t-۱} + \left(\frac{\left(\begin{array}{l}\theta^{۷}+۴\theta^{۶}+۶\theta^{۵}+۳۲\theta^{۴}\\ +۱۲۰\theta^{۳}+۶۰\theta^{۲}+۱۴۴\theta+۱۴۴\end{array}\right)}{\theta^{۲}(\theta^{۳}+\theta^{۲}+۶)^{۲}}\right).$$

۴ برآوردیابی پارامترهای فرآیند

۱.۴ روش ماکسیمم درستنمایی شرطی

برای انجام روش ماکسیمم درستنمایی شرطی تحت فرآیند INAR(1)-PJ، تابع درستنمایی شرطی زیر تعریف شده است.

$$\begin{aligned}\ell(p,\theta) &= \sum_{t=۲}^{T}\ln\left(P(X_t=k\,|X_{t-۱}=l)\right)\\ &= \sum_{t=۲}^{T}\ln\left(\left(\frac{\theta^{۴}}{\theta^{۳}+\theta^{۲}+۶}\right)\sum_{i=۱}^{\min(X_t,X_{t-۱})}\left(\begin{array}{c}X_{t-۱}\\ i\end{array}\right)p^{i}(۱-p)^{X_{t-۱}-i}\left(\frac{۱}{(\theta+۱)^{X_t-i+۱}}\right.\right.\\ &\qquad\left.\left.+\frac{(X_t-i+۱)}{(\theta+۱)^{X_t-i+۲}}+\frac{(X_t-i+۳)(X_t-i+۲)(X_t-i+۱)}{(\theta+۱)^{X_t-i+۴}}\right)\right).\end{aligned} \tag{۱.۴}$$

برآوردگر ماکسیمم درستنمایی شرطی (CML) فرآیند INAR(1)-PJ ، که با این $\hat{p}_{CLS},\hat{\theta}_{CLS}$ نشان داده می‌شود، با ماکسیمم کردن تابع درستنمایی شرطی برابری (۱.۴) حاصل می‌شود. برآوردیابی پارامترها را به کمک تابع optim در نرم‌افزار R انجام می‌دهیم. برای مدیریت دامنه پارامترهای p و θ از الگوریتم L-BFGS-B استفاده می‌شود. ماتریس اطلاع مشاهده شده در پارامترهای برآورد شده را با استفاده از خروجی ماتریس هِسیان حاصل از بهینه‌سازی محاسبه می‌شود. خروجی حاصل امکان محاسبه خطاهای استاندارد تقریبی برای برآوردگرها را فراهم می‌نماید. تحت شرایط نظم ، برآوردگرهای CML دارای سازگاری، توزیع نرمال مجانبی و کارایی هستند. جو (۱۹۹۷)

۲.۴ روش یول-والکر

روش برآوردیابی یول-والکر براساس برابری توابع خودکواریانس و میانگین فرآیند با حالت نمونه‌ای آنها و برآورد پارامترها با حل دستگاه معادلات ایجاد شده حاصل می‌شود. برای فرآیند INAR(1) تابع خود همبستگی نمونه ای (ACF) عبارتست از $\rho_X(h)=p^h$ که

$h \geq ۱$ و $p \in (۰, ۱)$ عملگر نازل فرآیند است. بنابراین برآورد یول–والکر p عبارتست از

$$\hat{p}_{YW} = \frac{\sum_{t=۲}^{T} (X_t - \bar{X})(X_{t-۱} - \bar{X})}{\sum_{t=۲}^{T} (X_t - \bar{X})^۲}.$$

برآوردگر یول–والکر $\hat{\theta}_{YW}$ از برابر قرار دادن میانگین نظری فرایند INAR(1)-PJ با مقدار نمونه‌ای آن به دست می‌آید. معادله برآوردیابی با قرار دادن تفاضل این دو کمیت برابر با صفر تشکیل شده و سپس $\hat{\theta}_{YW}$ از حل این معادله نسبت به θ حاصل می‌شود.

$$\left((۱ - \hat{p}_{YW})\bar{X}\right)\theta^۴ + \left((۱ - \hat{p}_{YW})\bar{X} - ۱\right)\theta^۳ - ۲\theta^۲ + \left(۶(۱ - \hat{p}_{YW})\bar{X}\right)\theta - ۲۴ = ۰ \qquad (۲.۴)$$

برابری (۲.۴) به صورت عددی به کمک تابع uniroot در نرم افزار R بدست می‌آید.

۵ مطالعات شبیه سازی

یک مطالعه شبیه‌سازی به منظور ارزیابی رفتار برآوردگرهای CML و YW بر مبنای حجم متفاوت نمونه ای برای پارامترهای مدل INAR(1)-PJ انجام شده و این شبیه سازی بر اساس ۱۰۰۰ تکرار برای حجم های نمونه $n = ۱۰۰, ۳۰۰, ۵۰۰$ اجرا شده است. نمونه‌های تصادفی با توزیع PJ با استفاده از الگوریتم معرفی شده در بخش ۱ تولید شده‌اند. یک مجموعه انتخابی از مقادیر پارامتر برای انجام شبیه‌سازی عبارت است از $p = ۰/۵, \theta = ۱$. عملکرد برآوردگرهای CML و YW با استفاده از سه معیار انحراف (Bias) ، میانگین مربعات خطا (MSE) و میانگین خطای نسبی (MRE) ارزیابی می‌شود. این معیارها به صورت زیر تعریف می‌شوند:

$$Bias = \sum_{j=۱}^{N} \frac{\hat{\gamma}_{i,j} - \gamma_i}{N}, \quad MSE = \sum_{j=۱}^{N} \frac{(\hat{\gamma}_{i,j} - \gamma_i)^۲}{N}, \quad MRE = \sum_{j=۱}^{N} \frac{\hat{\gamma}_{i,j}/\gamma_i}{N},$$

برای $i = ۱, ۲$ و $\gamma = (p, \theta)$.

مقادیر محاسبه شده برای Bias و MSE که در جدول ۱ گزارش شده‌اند، به طور کلی نزدیک به صفر هستند و این موضوع میزان اطمینان هر دو روش برآوردیابی را در حجم‌های محدود تأیید می‌کند. با افزایش حجم نمونه، میانگین خطاهای نسبی (MREs) به سمت یک همگرا شده، که نشان از بهبود دقت برآوردیابی با افزایش حجم داده‌ها دارد. این نتایج نشان می‌دهند که روش CML از لحاظ هر دو معیار Bias و MSE در تمامی حجم‌های نمونه‌ای، عملکرد بهتری نسبت به روش YW دارد. علاوه بر این، مقادیر MRE محاسبه شده نیز برتری روش CML را تأیید می‌کنند، زیرا این مقادیر به طور مداوم به مقدار ایده‌آل یک نزدیک‌تر هستند.

۶ کاربرد فرآیند پیشنهادی

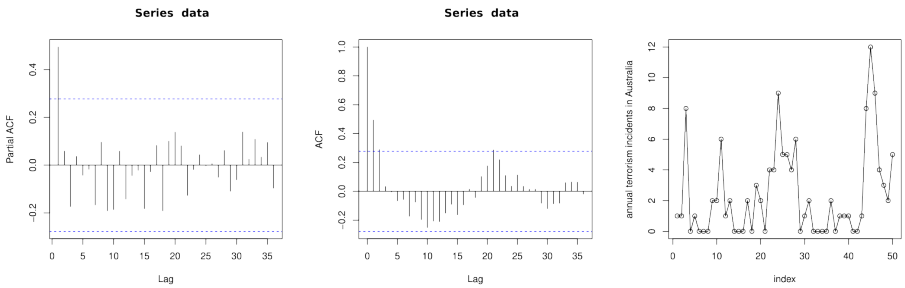
در این بخش، یک مجموعه داده واقعی برای بررسی کاربرد عملی فرایند پیشنهادی INAR(1)-PJ در مقایسه با مدل استاندارد INAR(1) پواسونی (INAR(1)-P) مورد تحلیل قرار می‌گیرد. برای مقایسه مدل‌های رقیب، از معیار AIC و معیار BIC استفاده می‌کنیم، به‌طوری که مقادیر کمتر نشان‌دهنده برازش بهتر مدل به داده‌ها است. ما یک سری زمانی سالانه از وقایع تروریستی در استرالیا که از سال ۱۹۷۰ تا ۲۰۲۰ ادامه داشته است، مورد تحلیل قرار می‌دهیم. داده‌ها از بسته Ecdat در نرم‌افزار R به دست آمده‌اند. آمار توصیفی در این مجموعه داده عبارتست از میانگین ۴/۲، واریانس ۴۸/۸ و شاخص پراکنش ۵۳/۳ ، که به‌خوبی رفتار بیش‌پراکنش را در داده‌های شمارشی نشان

جدول ۱: نتایج شبیه سازی برای فرآیند INAR(1)-PJ : اولین مجموعه انتخابی پارامترها

	$\theta = ۱$		$p = ۰/۵$		
n		CML	YW	CML	YW
۱۰۰	Bias	۰۰۵۸/۰	-۰۳۱۶/۰	-۰۰۳۲/۰	-۰۲۴۶/۰
	MSE	۰۰۹۳/۰	۰۰۸۵/۰	۰۰۲۳/۰	۰۱۸۸/۰
	MRE	۹۷۹۰/۰	۹۱۱۲/۰	۹۶۶۲/۰	۹۴۹۱/۰
۳۰۰	Bias	۰۰۲۴/۰	-۰۱۰۳/۰	-۰۰۱۵/۰	-۰۰۶۹/۰
	MSE	۰۰۳۵/۰	۰۰۲۶/۰	۰۰۰۹/۰	۰۰۶۳/۰
	MRE	۹۹۱۰/۰	۹۶۸۷/۰	۹۸۵۲/۰	۹۸۲۷/۰
۵۰۰	Bias	۰۰۱۹/۰	-۰۰۴۹/۰	-۰۰۰۲/۰	-۰۰۲۷/۰
	MSE	۰۰۱۹/۰	۰۰۱۵/۰	۰۰۵/۰	۰۰۴۰/۰
	MRE	۹۹۰۹/۰	۹۷۹۰/۰	۹۸۸۴/۰	۹۸۶۲/۰

می‌دهند و لزوم استفاده از مدل‌هایی قوی تر از مدل استاندارد INAR(1) پواسونی را مطرح می‌کند. روش ارائه شده در شویر و ویب (۱۹۸۶) یک آزمون برای وجود بیش‌پراکنش در فرایندهای خودبازگشتی با مقادیر عدد صحیح نامنفی ارائه دادند. با اعمال این روش روی مجموعه داده حاضر، آماره آزمونی برابر با ۳۷/۱۱ و مقدار p آزمون کمتر از ۰۰۱/۰ به دست می‌آید که منجر به رد فرضیه صفر هم‌پراکنش می‌شود. این نتیجه برجسته می‌کند که توزیع نوفه مدل باید قادر به مدل‌سازی تغییرات بیشتر از میانگین باشد؛ بنابراین، ممکن است مدل مبتنی بر پواسون برای این داده‌ها مناسب نباشند.

شکل ۳ نمودار سری زمانی تعداد سالانه حملات تروریستی در استرالیا را به همراه توابع خودهمبستگی (ACF) و خودهمبستگی جزئی (PACF) برآورد شده را نشان می‌دهد. تابع PACF پس از گام اول قطع می‌شود، که نشان می‌دهد ساختار خودبازگشتی مرتبه اول به خوبی وابستگی دنباله‌ای داده‌ها را مدل‌سازی می‌کند. بنابراین، مدلی از نوع AR(1) برای مدل‌سازی این مجموعه داده مناسب است.



شکل ۳: نمودار سری زمانی، تابع خودهمبستگی (ACF) و تابع خودهمبستگی جزئی (PACF) برای تعداد سالانه حملات تروریستی در استرالیا

در جدول ۲ ، پارامترهای برآوردشده مدل‌ها، خطاهای استاندارد مرتبط به آنها و معیارهای انتخاب مدل ارائه شده است. مدل INAR(1)-PJ از نظر هر دو معیار AIC و BIC مقادیر پایین‌تری را نسبت به مدل INAR(1)-P ارائه می‌دهد. این نتایج حاکی از عملکرد بهتر مدل پیشنهادی در توصیف داده‌های مربوط به حملات تروریستی سالانه در استرالیا دارد.

جدول ۲: برآورد ماکسیمم درستنمایی شرطی پارامترهای مدل، خطای استاندارد، AIC، BIC و آماره های مرتبط با تعداد سالانه حملات تروریستی در استرالیا

مدل	پارامتر	برآورد	خطای استاندارد	AIC	BIC	μ_X	$\sigma_X^۲$	DI_X
(۱INAR-PJ(	p	۲۳۷/۰	۰۱۱۰/۰	۸۴۹/۲۰۷	۶۳۳/۲۱۱	۳۶/۲	۴۵/۴	۸۸/۱
	θ	۵۱۴/۱	۰۳۱/۰					
(۱INAR-P(	p	۳۳۱/۰	۰۰۵/۰	۳۳۹/۲۳۹	۱۲۳/۲۴۳	۴۶/۲	۴۶/۲	۱
	λ	۶۵۰/۱	۰۵۱/۰					
تجربی						۴/۲	۴۸/۸	۵۳/۳

۷ نتیجه‌گیری

در این مقاله، یک مدل جدید خودبازگشتی عدد صحیح نامنفی مرتبه اول با نوفه‌های PJ با نام (INAR(1)-PJ)، برای مدل‌سازی سری‌های زمانی شمارشی با بیش‌پراکنش پیشنهاد شده است. به منظور برآورد پارامترهای مدل، دو روش مختلف مورد بررسی قرار گرفته و عملکرد آنها از طریق شبیه‌سازی مقایسه شده است. مدل پیشنهادی روی یک مجموعه داده واقعی اعمال شده و نتایج آن با مدل استاندارد INAR(1) پواسونی مقایسه گردیده است. یافته‌های تجربی نشان می‌دهد که مدل INAR(1)-PJ در مواجهه با بیش‌پراکنش عملکرد بهتری دارد. بنابراین مدل پیشنهادی برای مدلبندی داده‌هایی با چولگی مثبت و شاخص پراکنش بزرگتر از ۱ گزینه مناسب‌تری محسوب می‌شود.

مراجع

Al-Osh, M and Alzaid, A (1987). First-order integer-valued autoregressive (INAR (1)) process. *J Time Ser Anal*,**8(3)**: 261-275.

Alzaid, A, Al-Osh, M (1988). First-order integer-valued autoregressive (INAR (1)) process: distributional and regression properties. *Stat Neerland*,**42(1)**: 53-61.

Echebiri, U.V., Mbegbu, J.I. (2022). Juchez Probability Distribution: Properties and Applications. *Asian Journal of Probability and Statistics*,**20(2)**: 56-71.

Jazi, M., Jones, G. and Lai, C.D. (2012). Integer valued AR (1) with geometric innovations. *J. Iran. Stat. Soc. (JIRSS)*,**11**: 173-190.

Joe, H (1997). Multivariate models and multivariate dependence concepts. *Chapman and Hall, London*.

Kim, H. and Lee, S. (2017). On first-order integer-valued autoregressive process with Katz family innovations. *J Stat Comput Simul*,**87**: 546-562.

McKenzie, E (1986). Autoregressive moving-average processes with negative binomial and geometric marginal distrbutions.*Adv Appl Probabx*, **18**: 679-705.

Schweer, S, WeiB, CH (2014). Compound Poisson INAR (1) processes: stochastic properties and testing for overdispersion. *Comput Stat Data Anal*,**77**: 267-284.

WeiB, C.H.(2018). An introduction to discrete-valued time series. *Wiley*.

The INAR(1)-PJ Process: A New Alternative to the Poisson INAR(1) Process

Mahdi Rasekhi[1]

[1]Department of Statistics, Faculty of Mathematical Sciences and Statistics,

Malayer University, Malayer, Iran

Abstract: In this paper, we introduce a first-order non-negative integer-valued autoregressive model, in which the innovation process follows a newly proposed one-parameter distribution. To estimate the unknown parameters of the model, two estimation approaches are considered: conditional maximum likelihood and Yule-Walker methods. A simulation study is conducted to evaluate the performance and efficiency of these estimation techniques. Furthermore, the practical applicability of the proposed model is demonstrated through the analysis of a real-world data sets, showing that it outperforms the traditional Poisson INAR(1) model.

Keywords: Juchez probability distribution, count data, over-dispersion, INAR(1) process.

Mathematics Subject Classification (2020): 62M10، 62F10، 62E15.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

بررسی انشعاب‌های تصادفی مدل رشد لجستیک جمعیت

سمیرا مهموئی [۱] ، امید ربیعی مطلق [۲] ، حاجی محمد محمدی‌نژاد [۳]
[۱,۲,۳] دانشکده علوم ریاضی و آمار دانشگاه بیرجند

چکیده: تحلیل رفتار جمعیت و فرآیندهای زیستی از مسائل کلیدی در علوم ریاضی و زیست‌شناسی به‌شمار می‌آید. در این زمینه، مدل رشد لجستیک به‌عنوان ابزاری مؤثر برای بررسی رفتار جمعیت‌ها و فرآیندهای بیولوژیکی مورد استفاده قرار می‌گیرد. با توجه به افزایش جمعیت و چالش‌های محیطی، مدیریت پایدار منابع امری ضروری است. این مقاله به بررسی انشعاب‌های معادله رشد لجستیک تحت تأثیر حرکت وینر و عوامل تصادفی می‌پردازد. نتایج به‌دست آمده نشان‌دهنده تغییرات ناگهانی در امید ریاضی و وقوع P-انشعاب در رفتار سیستم است. هدف ارائه چارچوبی جامع برای درک بهتر دینامیک‌های سیستم‌های جمعیتی در شرایط ناپایدار است.

واژه‌های کلیدی: مدل رشد لجستیک جمعیت، انشعاب تصادفی، معادلات تصادفی، تابع چگالی احتمال.
کد موضوع‌بندی ریاضی (۲۰۲۰): 37G10، 34F05، 37N25.

۱ مقدمه

مدل رشد لجستیک[۱] در ساختارهای ریاضی متنوعی مانند مدل‌سازی رشد جمعیت باکتری‌ها، پیاده‌سازی شبکه‌های عصبی مصنوعی، تخمین پویایی $COVID-۱۹$ برای پیشگیری از شیوع، تحلیل مدل‌های رشد جمعیت تومورها به منظور درک رفتار فرآیند آن‌ها و ... کاربرد دارد(Tsagkis, 2023; Postnikov, 2020; Mansour, 2022; Koyama, 2021). با توجه به روند افزایشی جمعیت جهان، نیاز برای مدیریت رشد پایدار و اطمینان از اینکه رشد جمعیت در محدوده منابع موجود باقی می‌ماند، ضروری است. همانطور که می‌دانیم، چالش‌هایی مانند تخریب محیط زیست با افزایش جمعیت تشدید می‌شود. اندازه جمعیت به‌طور قابل توجهی بر توسعه اجتماعی، اقتصادی، سیاسی و محیط زیستی تأثیر می‌گذارد.

[۱] سخنران، s.mahmoiy@birjand.ac.ir

[1] Logistic Growth Model

با توجه به تأثیر عوامل متعددی بر نرخ رشد جمعیت، ورهولست [۲] این عوامل را در قالب مدل لجستیک فرمول‌بندی کرده است که نشان‌دهنده رشد محدود شونده‌ی یک جمعیت زیستی است. این معادله غالباً به عنوان نمونه استاندارد در نمایش رشد جمعیت مورد استفاده قرار می‌گیرد، که در آن نرخ تولد مستقیماً با اندازه‌ی کنونی جمعیت و منابع موجود مرتبط است.

در ادامه، از توسیع مدل ورهولست به شکل زیر استفاده می‌شود:

$$\begin{aligned} dx &= \Big[x(1-\frac{x}{y})-ax^2\Big]\,dt, \\ dy &= \Big[-h\frac{x(y-x)}{y(c+x)^2}-bxy+\alpha y\Big]\,dt, \end{aligned} \tag{1.1}$$

$$x(0)=x_0,\quad y(0)=y_0,$$

در این مدل، $x(t)$ نشان‌دهنده‌ی جمعیتی است در لحظه t که محیط زیست خودش را تخریب می‌کند و در عین حال رقابت درون‌گروهی هم دارد، $y(t)$ نقش ظرفیت برد سیستم را ایفا می‌کند که نشان‌دهنده‌ی حداکثر جمعیتی است که محیط قادر است در زمان t پشتیبانی کند. برخلاف مدل کلاسیک ورهولست که ظرفیت برد مقداری ثابت است، در اینجا $y(t)$ به‌صورت تابعی پویا مدل شده که تحت تأثیر چندین عامل زیست‌محیطی تغییر می‌کند. عبارت $-bxy$ در معادله‌ی dy نشان می‌دهد که افزایش جمعیت x منجر به تخریب سریع‌تر محیط از طریق استفاده بی‌رویه از منابع یا آلودگی آن می‌شود؛ بنابراین، افزایش پارامتر b باعث کاهش سریع‌تر $y(t)$ می‌شود. در مقابل، عبارت $+\alpha y$ بیانگر بازسازی طبیعی ظرفیت محیط است که به‌صورت تدریجی موجب بهبود شرایط زیستی می‌شود. افزایش α رشد سریع‌تر $y(t)$ را در پی دارد و بنابراین به پایداری بیشتر ظرفیت برد منجر می‌شود. افزون‌بر این، عبارت غیرخطی $-\frac{hx(y-x)}{y(c+x)^2}$ نوعی اصطکاک پویا میان جمعیت و ظرفیت محیط را مدل‌سازی می‌کند. پارامترهای a,b,c,h,α همگی مثبت بوده و دینامیک کلی سیستم زیستی را تعیین می‌کنند.

در سیستم‌های دینامیکی غیرتصادفی، انشعاب زمانی رخ می‌دهد که تغییر در پارامترهای سیستم باعث تغییر ناگهانی در رفتار کیفی جواب‌های آن شود. برای مثال، از دست رفتن پایداری یک نقطه تعادل و یا ظهور چرخه‌های حدی یا تغییر در تعداد نقاط تعادل نمونه‌های از وقوع یک انشعاب در دستگاه‌های معادلات دیفرانسیل غیرتصادفی هستند. در محیط‌های تصادفی، این انشعابات تحت تأثیر نویز قرارمی‌گیرند، که می‌تواند رفتار انشعاب را تغییر دهد یا پدیده‌های جدیدی ایجاد کند که در سیستم‌های قطعی مشاهده نمی‌شوند. نظریه انشعاب تصادفی [۳] تغییرات کیفی را در رفتار مجانبی سیستم‌های دینامیکی تصادفی هنگامی که پارامترهای سیستم تغییر می‌کنند، بررسی می‌کند. انشعاب‌های تصادفی در دو قالب کلّی دسته بندی می‌شوند:

۱- (Dـانشعاب[۴] یا انشعاب دینامیکی[۵]): این نوع انشعاب وابسته به تغییر رفتار کیفی جواب‌ها است. در صورت نبودن فاکتور تصادف در معادله، این انشعاب همان انشعاب معمولی در معادلات دیفرانسیل است.

۲- (Pـانشعاب[۶] یا انشعاب پدیده‌ای[۷]): این انشعاب وابسته به تغییر ماهیت تصادفی جواب‌ها است (به عنوان مثال وقوع یک انشعاب یا تغییر کیفی در تابع چگالی احتمال جواب‌ها، یک Pـانشعاب محسوب می‌شود).

[۲] مدل رشد لجستیک اولین بار توسط پیر فرانسوا ورهولست $Pierre-FranoisVerhulst$ در سال ۱۸۳۸ معرفی شد.

[3] stochastic bifurcation

[4] D-bifurcation

[5] dynamical bifurcation

[6] P-bifurcation

[7] phenomenological bifurcation

تحلیل P- انشعاب‌ها معمولاً با مطالعه‌ی معادله معروف فوکر–پلانک[8] انجام می‌گیرد که توزیع احتمال مسیرهای سیستم را توصیف می‌کند؛ در این‌جا ابتدا به معرفی این معادله می‌پردازیم تا بستر لازم برای مطالعه‌ی پدیده‌ی انشعاب در سیستم‌های تصادفی فراهم شود.

فرض‌کنید $f : \mathbb{R}^n \times [t_\circ, T] \to \mathbb{R}^n$ و $g : \mathbb{R}^n \times [t_\circ, T] \to \mathbb{R}^{n\times d}$ توابعی هموار و $\{W(t)\}_{t\geqslant \circ}$ یک حرکت براونی d–بعدی باشد. معادله دیفرانسیل تصادفی زیر را درنظربگیرید:

$$dx(t) \quad = \quad f(x(t),t)\,dt + g(x(t),t)\,dW(t),$$

فرض‌کنیم $p \equiv p(x,t)$ تابع چگالی احتمال جواب $x(t)$ باشد، دراین‌صورت می‌توان نشان داد که p در معادله فوکر–پلانک متناظر زیر صدق می‌کند (برای ملاحظه جزئیات، مرجع (Arnold , 1974, p. 44) را ملاحظه کنید):

$$\partial_t p = -\sum_i \partial_i \left[f_i(x(t),t)p\right] + \frac{۱}{۲}\sum_{i,j}\partial_i\partial_j\left\{\left[g(x(t),t)g^{tr}(x(t),t)\right]_{ij} p\right\},$$

این معادله تکامل زمانی توزیع احتمال جواب‌های تصادفی سیستم را توصیف می‌کند. مطالعه‌ی دقیق این معادله نه تنها امکان پیش‌بینی رفتار سیستم را در آینده فراهم می‌کند، بلکه به ما اجازه می‌دهد تا رفتار سیستم در حضور نویز را تحلیل کنیم.

۲ بررسی انشعاب‌های تصادفی مدل رشد لجستیک جمعیت

با توجه به تأثیر عوامل تصادفی و غیرقابل پیش‌بینی بر رفتار جمعیت‌ها، بررسی این پدیده در مدل‌های ریاضی اهمیت زیادی دارد. در این بخش، انشعاب‌های ماهیتی معادله رشد لجستیک جمعیت در حضور حرکت وینر مورد مطالعه قرار می‌گیرد تا اثرات نویز تصادفی بر پایداری و رفتار بلندمدت سیستم تحلیل و مدل‌سازی شود. هدف اصلی این بخش، ارائه چارچوبی جامع برای درک عمیق‌تر رفتار سیستم‌های پویا در محیط‌های تصادفی و تحلیل تأثیر تغییرات پارامتری بر ویژگی‌های دینامیکی آن‌ها است. مدل رشد لجستیک جمعیت ۱.۱ را در نظر بگیرید. پس از افزودن فاکتور تصادف به سیستم داریم:

$$\begin{aligned} dx &= \overbrace{\left[x(۱-\frac{x}{y}) - ax^۲\right]}^{=L_۱(x,y)} dt + \sigma x\, dw_۱(t) \\ dy &= \underbrace{\left[-h\frac{x(y-x)}{y(c+x)^۲} - bxy + \alpha y\right]}_{=L_۲(x,y)} dt + \sigma y\, dw_۲(t), \end{aligned} \tag{۱.۲}$$

$$x(\circ) = x_\circ, \quad y(\circ) = y_\circ,$$

که در آن $dw_۱(t)$ تغییرات تصادفی جمعیت و $dw_۲(t)$ تغییرات تصادفی محیطی را مدل‌سازی می‌کنند. نقطه $E = (x^*, y^*)$ نقطه ثابت دستگاه غیرتصادفی است که در آن $y^* = x^*/(۱ - ax^*) > \circ$ و $x^* > \circ$ جواب معادله زیر است:

$$U(x) = x^۲\left(a^۲(-h) - \alpha + ۲bc\right) + x\left(ah + bc^۲ - ۲\alpha c\right) + bx^۳ - \alpha c^۲.$$

با استفاده از مقادیر پارامترهای جدول ۱، نقطه تعادل به صورت تقریبی برابر با $E = (x^*, y^*) = (۰/۰۸۲۵, ۰/۰۸۹۹)$ می‌باشد.

[8] Fokker–Planck equation

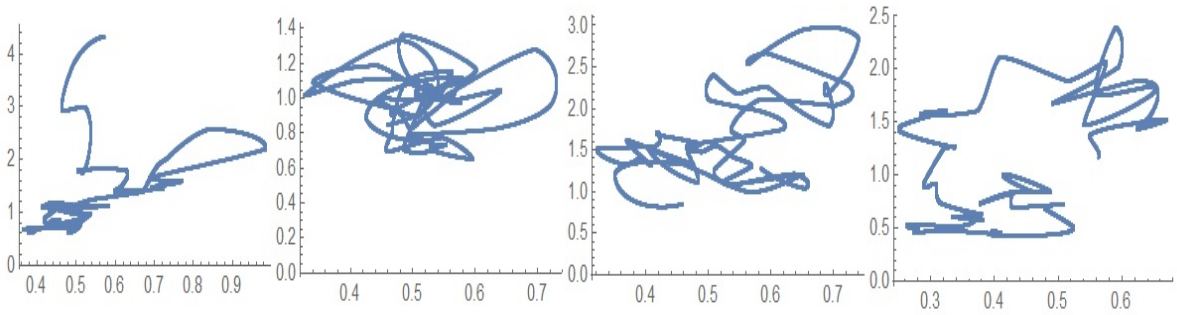
ماتریس قسمت خطی سیستم به صورت زیر است:

$$\begin{bmatrix} ۱-\dfrac{۲x(۱+y)}{y} & \dfrac{x^۲}{y^۲} \\ -\dfrac{۲}{(۲+x)^۳}-y+\dfrac{x(۴+y)}{(۲+x)^۳y} & ۰/۱+x\left(-۱-\dfrac{x}{(۲+x)^۲y^۲}\right) \end{bmatrix}$$

محاسبه‌ی مقادیر ویژه این ماتریس در نقطه‌ی E منجر به دو مقدار $\lambda_۱ = -۱/۰۹۵, \lambda_۲ = -۰/۸۱۲$ می‌شود. از آن‌جا که هر دو مقدار ویژه دارای قسمت حقیقی منفی و غیرصفر هستند، نتیجه می‌گیریم که نقطه‌ی E یک نقطه‌ی تعادل هذلولوی است. این موضوع، نشانگر ثبات و پایداری موضعی نقطه‌ی E است. بنابراین، در همسایگی این نقطه هیچ‌گونه D-انشعاب (که معمولاً در نقاط ناپایدار رخ می‌دهد) روی نمی‌دهد. شکل ۱ جواب‌های سیستم (۱.۲) را در اطراف E به ازای مقادیر پارامتر جدول ۱ نشان می‌دهد.

σ	T	$y(۰)$	$x(۰)$	b	c	h	α	a
۰/۳	۱۰	۰/۰۸۹۹	۰/۰۸۲۵	۱	۲	۱	۰/۵	۱

جدول ۱: مقادیر پارامتر سیستم (۱.۲)



شکل ۱: جواب‌های تصادفی معادله (۱.۲)

به منظور بررسی P-انشعاب دستگاه ۱.۲، فرض‌کنید $P \equiv P(t,x,y)$ تابع چگالی احتمال جواب این سیستم باشد، معادله فوکر-پلانک وابسته به این معادله را به شکل زیر در نظر بگیرید:

$$\begin{aligned}\frac{\partial P}{\partial t} &= -\frac{\partial}{\partial x}[L_۱(x,y)P]-\frac{\partial}{\partial y}[L_۲(x,y)P]+\frac{\sigma^۲}{۲}\frac{\partial^۲}{\partial x^۲}[x^۲P]+\frac{\sigma^۲}{۲}\frac{\partial^۲}{\partial y^۲}[y^۲P] \\ &= -P\frac{\partial L_۱(x,y)}{\partial x}-L_۱(x,y)\frac{\partial P}{\partial x}-P\frac{\partial L_۲(x,y)}{\partial y}-L_۲(x,y)\frac{\partial P}{\partial y} \\ &\quad +\frac{\sigma^۲}{۲}\left(۲P+۴x\frac{\partial P}{\partial x}+x^۲\frac{\partial^۲P}{\partial x^۲}+۲P+۴y\frac{\partial P}{\partial y}+y^۲\frac{\partial^۲P}{\partial y^۲}\right)\end{aligned}$$

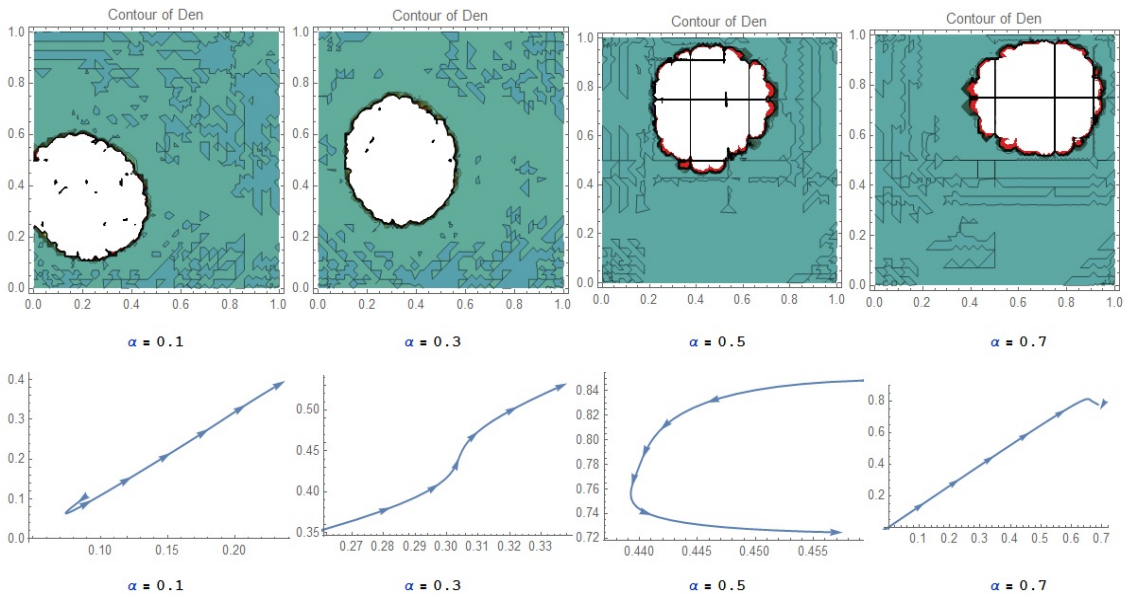
$$\begin{aligned}\frac{\partial P}{\partial t} &= \frac{\partial P}{\partial x}\Big[۲\sigma^۲x-L_۱(x,y)\Big]+\frac{\partial P}{\partial y}\Big[۲\sigma^۲y-L_۲(x,y)\Big] \\ &\quad +P\Big[۲\sigma^۲-\frac{\partial L_۱(x,y)}{\partial x}-\frac{\partial L_۲(x,y)}{\partial y}\Big]+\frac{\partial^۲P}{\partial x^۲}[\frac{\sigma^۲x^۲}{۲}]+\frac{\partial^۲P}{\partial y^۲}[\frac{\sigma^۲y^۲}{۲}]\end{aligned}$$

با توجه به شرط اولیه زیر،

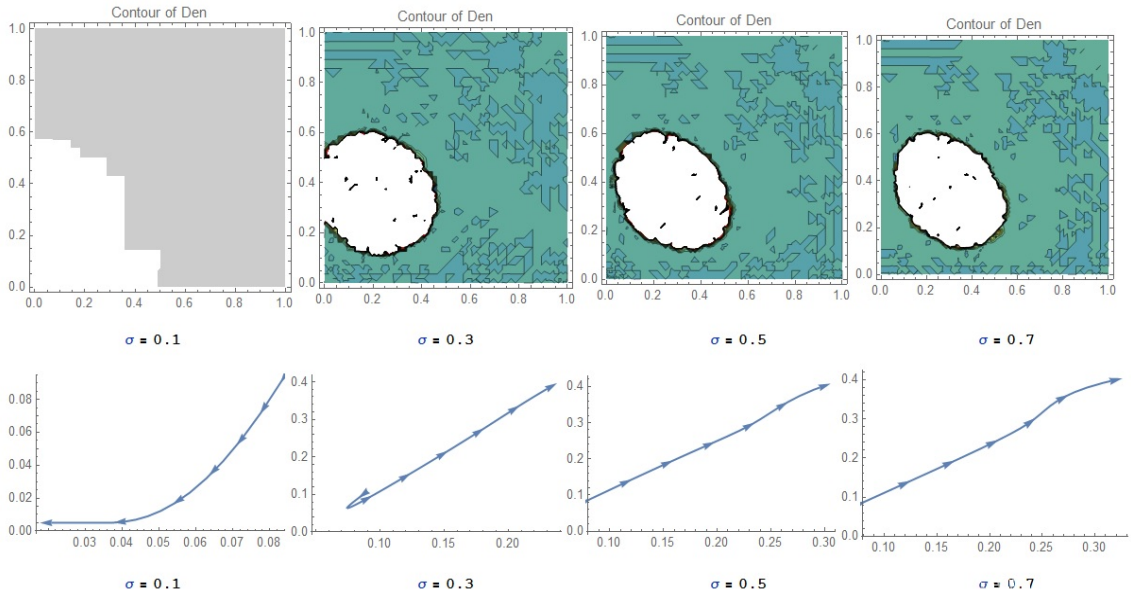
$$P(\circ, x, y) = \Delta(x, y, \epsilon) = \frac{۱}{\pi\epsilon} \exp\left(-\frac{(x - x_\circ)^۲ + (y - y_\circ)^۲}{\epsilon}\right), \qquad \circ < \epsilon \ll ۱$$

$$\lim_{\epsilon \to \circ} \Delta(x, y, \epsilon) = \delta(x, y)$$

که در آن $\delta(x, y)$ تابع دلتای دیراک است. نمودارهای ۲ و ۳ تغییرات جواب معادله فوکر–پلانک در صفحه را برحسب تغییرات به‌ترتیب α و σ نشان می‌دهند و نمودارهای قسمت پایین هر کدام از شکل‌ها امید ریاضی جواب دستگاه (۱.۲) را نشان می‌دهند. همان‌طورکه ملاحظه می‌شود نمودار امید ریاضی دچار تغییرات ناگهانی شده که تغییر رفتار سیستم (۱.۲) را نشان می‌دهد و بنابراین این سیستم دچار P–انشعاب شده است.



شکل ۲: برای $\sigma = \circ/۳; \epsilon = \circ/\circ\circ۲; a = ۱; b = ۱; c = ۲; h = ۱$



شکل ۳: برای $\epsilon = ۰/۰۰۲; a = ۱; b = ۱; c = ۲; h = ۱; \alpha = ۰/۱$

بحث و نتیجه‌گیری

در این مطالعه، با تکیه بر مدل رشد لجستیک و توسعه آن در بستر سیستم‌های تصادفی، به بررسی پایداری دینامیکی جمعیت در حضور نویزهای تصادفی پرداخته شد. تحلیل معادله رشد لجستیک با در نظر گرفتن حرکت وینر نشان داد که اعمال اختلالات تصادفی می‌تواند منجر به بروز رفتارهای پیچیده و انشعابات تصادفی در سیستم شود. نتایج حاصل از محاسبه مقادیر ویژه در نقطه‌ی تعادل E نشان داد که این نقطه هذلولوی بوده و در آن D–انشعاب رخ نمی‌دهد، در حالی‌که بررسیP–انشعاب در چارچوب معادله فوکر–پلانک، تصویر روشن‌تری از ناپایداری‌های ممکن در سیستم ارائه می‌دهد و نشان می‌دهد که حتی تغییرات کوچک در پارامترهای محیطی می‌توانند منجر به تحولات بنیادی در پاسخ سیستم شوند.

این نتایج اهمیت در نظر گرفتن نویزهای تصادفی در مدل‌سازی پدیده‌های واقعی، به‌ویژه در حوزه‌هایی مانند رشد جمعیت، انتشار بیماری‌ها و گسترش تومورها را برجسته می‌سازد. چارچوب ارائه‌شده در این پژوهش می‌تواند به‌عنوان مبنایی برای تحلیل پایداری و پاسخ سیستم‌های زیستی و اجتماعی در برابر ناپایداری‌های محیطی مورد استفاده قرار گیرد. در مجموع، بررسی مدل رشد لجستیک در بستر تصادفی نه تنها درک عمیق‌تری از رفتار بلندمدت سیستم‌های پویا فراهم می‌سازد، بلکه راهگشای توسعه سیاست‌ها و راهکارهای مبتنی بر داده برای مدیریت پایداری در مواجهه با عدم‌قطعیت‌های دنیای واقعی خواهد بود.

مراجع

Arnold L. (1974), *Stochastic Differential Equations: Theory and Applications*, Wiley Interscience.

Tsagkis P., Bakogiannis E. and Nikitas A. (2023), Analysing urban growth using machine learning and

open data: An artificial neural network modelled case study of five Greek cities, *Sustainable Cities and Society,***89**, 104337.

Postnikov E. B. (2020), Estimation of COVID-19 dynamics “on a back-of-envelope”: Does the simplest SIR model provide quantitative parameters and predictions?, *Chaos, Solitons and Fractals,***135**, 109841.

Mansour M. B. A. and Abobakr A. H. (2022), Stochastic differential equation models for tumor population growth, *Chaos, Solitons and Fractals,***164**, 112738.

Koyama K., Hiura S., Abe H. and Koseki S. (2021), Application of growth rate from kinetic model to calculate stochastic growth of a bacteria population at low contamination level, *Journal of Theoretical Biology,***525**, 110758.

Investigation of Stochastic Bifurcations in the Logistic Growth Model of Populations

Samira Mahmooee[1], **Omid RabieiMotlagh**[2],**Hajimohammad Mohammadinejad**[3]

[1,2,3] Dept. of Applied Math., University of Birjand, Birjand, Iran

---

Abstract: The logistic growth model is employed as a key tool in analyzing the behavior of populations and biological processes. Given the increasing population and environmental challenges, sustainable resource management is essential. This paper investigates the bifurcations of the logistic growth equation under the influence of Wiener motion and stochastic factors. The results obtained reveal sudden changes in the expected value and the occurrence of P-bifurcation in the system's behavior. The objective is to provide a comprehensive framework for a better understanding of the dynamics of population systems under unstable conditions.

Keywords: Logistic growth model, Stochastic bifurcation, Stochastic equations, Probability density function.

Mathematics Subject Classification (2020): 37G10، 34F05، 37N25.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

تشخیص اعداد دست‌نویس با رهیافت یادگیری ماشین

نوشین رضائی طالقانی [۱] ، آرزو حاج رجبی

گروه آمار، دانشگاه بین المللی امام خمینی (ره)

چکیده:

در سال‌های اخیر با گسترش داده‌های تصویری و افزایش نیاز به پردازش و تحلیل خودکار آن‌ها، مسئله‌ی تشخیص اعداد دست‌نویس به عنوان یکی از مسائل مهم و کاربردی در حوزه یادگیری ماشین و بینایی ماشین مطرح شده است. در این پژوهش مسئله‌ی تشخیص اعداد دست‌نویس با استفاده ازالگوریتم‌های یادگیری ماشین مانند جنگل تصادفی و درخت تصمیم بررسی و پیاده‌سازی گردیده است و برای این منظور از پایگاه‌های داده فارسی Hoda بهره گرفته شده است. عملکرد این مدل‌ها با معیارهای آماری مناسب ارزیابی شده و نتایج، حاکی از کارایی بالای جنگل تصادفی در مقایسه با درخت تصمیم می‌باشد.

واژه‌های کلیدی: یادگیری ماشین، تشخیص اعداد دست‌نویس، جنگل تصادفی ، درخت تصمیم.

۱ مقدمه

در سال‌های اخیر با گسترش فناوری‌های مبتنی بر هوش مصنوعی، مسائلی نظیر شناسایی و دسته‌بندی الگوهای نوشتاری در تصاویر به‌ویژه تشخیص اعداد دست‌نویس، توجه بسیاری از پژوهشگران را به خود جلب کرده است. تشخیص اعداد دست‌نویس یکی از مسائل مهم در حوزه شناسایی الگوها و پردازش تصویر است که در سامانه‌های بانکی، مراکز پستی، آرشیو اسناد و سایر نهادهای اداری کاربرد فراوان دارد. به دلیل تنوع در سبک نگارش افراد و ناهماهنگی‌های مختلف در تصاویر، این مسئله همواره به‌عنوان یکی از چالش‌های اصلی در این حوزه مطرح بوده است. در زمینه بازشناسی اعداد و حروف دست‌نویس پژوهش‌های بسیاری انجام شده است، به عنوان مثال با به‌کارگیری شبکه عصبی و سه مرحله پیش‌پردازش، استخراج ویژگی و دسته‌بندی، بر روی پایگاه Hoda با دقت ۹۵/۳۷٪ بازشناسی شده است (اسدی و دیگران ، ۱۳۹۵). زربافی و همکاران (۱۴۰۲)با استفاده از شبکه های عصبی کانولوشنی و مدل ترکیبی شامل حافظه کوتاه مدت و بلند مدت به اعتبار سنجی ۹۸/۰۳ روی پایگاه داده‌ی اعداد فارسی دست یافتند.

همچنین در پژوهشی ، عملکرد سه الگوریتم ،SVM درخت تصمیم و شبکه عصبی در تشخیص اعداد دست‌نویس فارسی بر روی پایگاه داده Hoda مقایسه شده است .(آقایی و اخوت ، ۱۴۰۳)

[۱] سخنران، rezaeinrt@gmail.com

در این مقاله برای بازشناسی اعداد دست‌نویس از پایگاه داده فارسی Hoda و از الگوریتم درخت تصمیم و جنگل تصادفی استفاده شده است و با استفاده از معیارهای ارزیابی نشان داده شده که الگوریتم جنگل تصادفی دارای دقت بیشتری نسبت به درخت تصمیم است. در بخش دوم به یادگیری ماشین ومفاهیم مقدماتی پرداخته شده است. الگوریتم جنگل تصادفی و درخت تصمیم در بخش سوم بررسی شده است. در بخش چهارم به مطالعه‌ی روش‌های پیشنهادی بر روی پایگاه داده Hoda پرداخته شده است.

۲ یادگیری ماشین

یادگیری ماشین شاخه‌ای از هوش مصنوعی است که به ماشین‌ها امکان می‌دهد از داده‌ها، الگو بیاموزند و پیش‌بینی کنند. این حوزه در مسائل گوناگون از جمله طبقه‌بندی تصاویر و پردازش متن، کاربرد گسترده‌ای دارد. یکی از مسائل مهم، شناسایی و طبقه‌بندی اعداد دست‌نویس است که به دلیل تنوع در سبک نگارش، چالشی پیچیده به شمار می‌رود. روش‌های یادگیری ماشین به چهار نوع اصلی نظارت‌شده،[1] بدون نظارت،[2] نیمه‌نظارت‌شده[3] و تقویتی[4] تقسیم می‌شوند. در این پژوهش، تمرکز بر روش‌های نظارت‌شده است که این نوع یادگیری یکی از شاخه‌های اصلی یادگیری ماشین است که در آن مدل با استفاده از مجموعه داده‌های برچسب‌دار آموزش می‌بیند تا بتواند پیش‌بینی‌های دقیقی برای داده‌های جدید ارائه دهد. در این روش، داده‌ها شامل جفت‌های ورودی–خروجی هستند که ورودی‌ها (ویژگی‌ها) و خروجی‌های مورد انتظار (برچسب‌ها) را مشخص می‌کنند. هدف، یادگیری یک تابع نگاشت است که بتواند ورودی‌ها را به برچسب‌های صحیح مرتبط سازد. این رویکرد به‌ویژه در مسائل طبقه‌بندی، مانند تشخیص اعداد دست‌نویس، و رگرسیون کاربرد گسترده‌ای دارد. یکی از نکات کلیدی در یادگیری ماشین، انتخاب ویژگی‌های مناسب از داده‌هاست. ویژگی‌هایی مانند شدت پیکسل‌ها در تصاویر یا الگوهای هندسی می‌توانند به شناسایی بهتر اعداد کمک کنند. پیش‌پردازش داده‌ها مانند نرمال‌سازی یا کاهش نویز دقت مدل‌ها را بهبود می‌بخشد. برای مثال در مسائل طبقه‌بندی تصاویر، تبدیل داده‌ها به مقیاس‌های استاندارد یا استفاده از روش‌های کاهش

جدول ۱: دسته‌بندی الگوریتم‌های یادگیری ماشین

ردیف	نوع یادگیری	الگوریتم‌های پرکاربرد
۱	یادگیری نظارت‌شده	رگرسیون خطی، رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، نزدیک ترین همسایه
۲	یادگیری بدون نظارت	خوشه‌بندی، تحلیل مؤلفه‌های اصلی
۳	یادگیری نیمه‌نظارت‌شده	خودآموز، ماشین بردار پشتیبان نیمه‌نظارتی
۴	یادگیری تقویتی	روش ترکیبی شبکه عمیق و ماشین بردار پشتیبان

ابعاد مانند تحلیل مؤلفه‌های اصلی، سرعت و کارایی مدل‌ها را افزایش می‌دهد. در جدول ۱ به انواع یادگیری ماشین اشاره شده است که نتایج نشان می دهد که انتخاب درست الگوریتم نقش مهمی را در برازش مدل مناسب به داده‌ها ایفا می‌کند.

[1] Supervised
[2] Unsupervised
[3] Semi-Supervised
[4] reinforcement

۳ جنگل تصادفی و درخت تصمیم

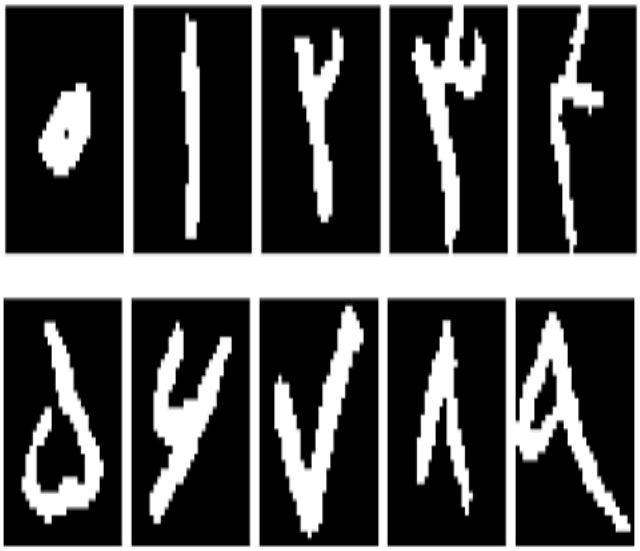
درخت تصمیم یک الگوریتم یادگیری ماشین است که برای شناسایی دست‌نوشته‌های اعداد به کار می‌رود. این روش داده‌های ورودی، مانند ویژگی‌های پیکسلی یا الگوهای هندسی تصاویر دست‌نوشته را در یک ساختار درختی سازمان‌دهی می‌کند. هر گره در درخت یک ویژگی خاص (مثلاً شدت رنگ پیکسل در یک ناحیه) را بررسی می‌کند. و شاخه‌ها، داده‌ها را بر اساس معیارهای تصمیم به زیرمجموعه‌های کوچک‌تر تقسیم می‌کنند تا در نهایت به یک برگ برسند که نشان‌دهنده یک عدد خاص (مثلاً ۰ تا ۹) است. این روش به دلیل سادگی و قابلیت تفسیر آسان، برای مسائل شناسایی اعداد مناسب است. با این حال، اگر داده‌ها دارای خطا باشند یا تنوع زیادی در سبک‌های دست‌خط وجود داشته باشد، درخت تصمیم ممکن است بیش از حد به داده‌های آموزشی وابسته شود و بیش‌برازش بوجود آید و عملکردش روی داده‌های جدید کاهش یابد. جنگل تصادفی یک الگوریتم پیشرفته‌تر است که با ترکیب چندین درخت تصمیم، دقت شناسایی دست‌نوشته‌های اعداد را افزایش می‌دهد. این روش با انتخاب تصادفی زیرمجموعه‌هایی از داده‌ها (مانند تصاویر دست‌نوشته) و ویژگی‌ها (مثلاً بخشی از پیکسل‌ها)، تعداد زیادی درخت تصمیم مستقل می‌سازد و سپس، برای پیش‌بینی نهایی نتایج همه درخت‌ها را ترکیب می‌کند. این رویکرد باعث می‌شود جنگل تصادفی در برابر خطا و تغییرات در دست‌خط‌ها مقاوم‌تر باشد و معمولاً دقت بالاتری نسبت به یک درخت تصمیم تنها ارائه دهد. با این حال، به دلیل استفاده از چندین درخت، محاسبات آن سنگین‌تر است و تفسیر مدل به اندازه یک درخت تصمیم ساده نیست. برای جزییات بیشتر به (Geron ، ۲۰۱۹) مراجعه شود.

۴ تحلیل داده‌ی واقعی

در این پژوهش از مجموعه داده Hoda استفاده شده است که اولین مجموعه بزرگ اعداد دست نویس فارسی است. این مجموعه شامل ۱۰۲۳۵۳ تصاویر خاکستری از اعداد ۰ تا ۹ است که در شکل ۱ قابل مشاهده است. این مجموعه داده‌ها باتوجه به حجم کم، ساختار ساده و برچسب‌گذاری دقیق، گزینه‌ای ایده‌آل برای آزمایش سریع و مقایسه عملکرد الگوریتم‌ها محسوب می‌شوند. جدول ۲، تعداد نمونه‌ها در هر کلاس مشخص شده است.

جدول ۲: تعداد نمونه‌ها در هر کلاس

کلاس	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
تعداد نمونه	۱۰۰۷۰	۱۰۳۳۰	۹۹۲۳	۱۰۳۳۴	۱۰۳۳۳	۱۰۱۱۰	۱۰۲۵۴	۱۰۳۶۳	۱۰۲۶۴	۱۰۳۷۱



شکل ۱: .تصویر اعداد ۰ تا ۹

در گام اول تصاویر ۳۲×۳۲ پیکسل از مجموعه داده Hoda به آرایه‌های یک‌بعدی ۱۰۲۴ تایی تبدیل شده است، که هر عنصر نشان‌دهنده شدت رنگ خاکستری یک پیکسل (بین ۰ تا ۲۵۵) می باشد، این مقادیر بر ۲۵۵ تقسیم شده تا در بازه [۰،۱]نرمال‌سازی [۵] شوند و برای مدل‌های یادگیری ماشین مناسب‌تر گردند. سپس ۷۰ درصد داده ها را به مجموعه‌های آموزشی و ۳۰ درصد را به مجموعه آزمایشی تقسیم کرده و پس از آن مجموعه آموزشی را به مدل‌های یادگیری نظارت‌شده مانند درخت تصمیم و یا جنگل تصادفی می دهیم. مدل با داده‌های آموزشی و برچسب‌های مربوطه (اعداد ۰ تا ۹) آموزش می‌بیند، در مدل درخت تصمیم با استفاده از شاخص جینی [۶] به‌عنوان معیار پیش‌فرض تقسیم‌بندی، گره‌ها و شاخه‌ها تشکیل شده است. شاخص جینی میزان ناخالصی گره‌ها را با فرمول(۱.۴)

$$\text{Gini} = ۱ - \sum_{i=۱}^{K} p_i^۲ \tag{۱.۴}$$

محاسبه می‌کند، که در آن p_i احتمال حضور در کلاس i است. مدل جنگل تصادفی از مجموعه‌ای از درخت‌های تصمیم مستقل تشکیل شده که هر یک با شاخص جینی آموزش دیده و پیش‌بینی‌های آن‌ها از طریق رأی‌گیری اکثریت ترکیب می شوند. برای ارزیابی عملکرد مدل‌ها از معیارهای زیر استفاده شده است :

۱. **دقت**[۷] : این معیار نسبت نمونه‌های درست پیش‌بینی‌شده به کل نمونه‌ها را نشان می‌دهد.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- TP[۸] : نمونه‌های مثبت که به‌درستی مثبت پیش‌بینی شده‌اند (مثبت درست).
- TN[۹] : نمونه‌های منفی که به‌درستی منفی پیش‌بینی شده‌اند (منفی درست).

[5] Normalization
[6] Gini Impurity
[7] Accuracy
[8] True Positive
[9] True Negative

- FP[10] : نمونه‌های منفی که به‌اشتباه مثبت پیش‌بینی شده‌اند (مثبت نادرست).
- FN[11] : نمونه‌های مثبت که به‌اشتباه منفی پیش‌بینی شده‌اند (منفی نادرست).

۲. **صحت**[12] : این معیار نسبت پیش‌بینی‌های درست مثبت به کل پیش‌بینی‌های مثبت را نشان می‌دهد و برای کاهش خطاهای مثبت نادرست (FP) اهمیت دارد.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

۳. **بازخوانی یا حساسیت**[13] : این معیار نسبت نمونه‌های مثبت درست شناسایی‌شده به کل نمونه‌های مثبت واقعی را نشان می‌دهد و برای کاهش خطاهای منفی نادرست (FN) حیاتی است.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

۴. **امتیاز F۱**[14] : میانگین هارمونیک صحت و حساسیت است که برای داده‌های نامتوازن مناسب بوده و تعادل بین این دو معیار را برقرار می‌کند.

$$\text{F۱ Score} = ۲ \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

معیارها به‌صورت درصدی در بازه [۰٪, ۱۰۰٪] گزارش شده‌اند، که ۱۰۰٪ بهترین و ۰٪ بدترین عملکرد را نشان می‌دهد. نتایج ارزیابی مدل‌های درخت تصمیم درجدول۳ و جنگل تصادفی در جدول ۴ ارائه شده است.

ماتریس‌های درهم‌ریختگی ارائه‌شده، عملکرد دو مدل یادگیری ماشین، یعنی جنگل تصادفی و درخت تصمیم، را بر روی داده‌های آزمایشی برای تشخیص اعداد دست‌نویس (۰ تا ۹) نشان می‌دهند. ماتریس درهم‌ریختگی جنگل تصادفی نشان‌دهنده دقت بسیار بالای این مدل است. مقادیر قطری (پیش‌بینی‌های درست) برای هر کلاس (اعداد ۰ تا ۹) نشان می‌دهند که اکثر نمونه‌ها به‌درستی طبقه‌بندی شده‌اند. به‌عنوان مثال، برای عدد ۰، مدل در ۲۹۶۴ مورد از مجموع نمونه‌ها پیش‌بینی درست داشته و تنها در ۴ مورد خطا کرده است (۱ و ماتریس درهم‌ریختگی درخت تصمیم عملکرد ضعیف‌تری را نسبت به جنگل تصادفی نشان می‌دهد. به طور کلی، جنگل تصادفی به دلیل استفاده از چندین درخت تصمیم و کاهش واریانس، عملکرد بهتری نسبت به درخت تصمیم در تشخیص اعداد دست‌نویس دارد. تعداد بالای پیش‌بینی‌های درست و خطاهای کمتر در ماتریس درهم‌ریختگی جنگل تصادفی، برتری این مدل را تأیید می‌کند.

ماتریس درهم‌ریختگی درخت تصمیم (دو کلاس اول)

$$\begin{pmatrix} ۲۹۰۸ & ۳ & ۱ & ۵ & ۱۵ & ۱۹ & ۴ & ۸ & ۵ & ۰ \\ ۱ & ۳۰۰۵ & ۱۰ & ۱ & ۷ & ۳ & ۱۲ & ۲ & ۳ & ۲۲ \end{pmatrix}$$

ماتریس درهم‌ریختگی جنگل تصادفی (دو کلاس اول)

$$\begin{pmatrix} ۲۹۶۴ & ۰ & ۰ & ۰ & ۱ & ۳ & ۰ & ۰ & ۰ & ۰ \\ ۲ & ۳۰۵۴ & ۱ & ۰ & ۳ & ۰ & ۳ & ۰ & ۰ & ۳ \end{pmatrix}$$

[10] False Positive
[11] False Negative
[12] Precision
[13] Recall
[14] Score-۱F

جدول ۳: معیارهای ارزیابی درخت تصمیم

کلاس	صحت	بازخوانی یا حساسیت	امتیاز F۱
۰	%۹۷	%۹۸	%۹۸
۱	%۹۶	%۹۸	%۹۷
۲	%۸۹	%۹۲	%۹۱
۳	%۹۲	%۹۱	%۹۱
۴	%۹۱	%۸۹	%۹۰
۵	%۹۵	%۹۲	%۹۴
۶	%۹۱	%۹۰	%۹۱
۷	%۹۵	%۹۶	%۹۶
۸	%۹۷	%۹۷	%۹۷
۹	%۹۴	%۹۳	%۹۴
دقت : %۹۴			

جدول ۴: معیارهای ارزیابی جنگل تصادفی

کلاس	صحت	بازخوانی یا حساسیت	امتیاز F۱
۰	%۹۹	%۱۰۰	%۱۰۰
۱	%۹۹	%۱۰۰	%۹۹
۲	%۹۷	%۹۸	%۹۸
۳	%۹۸	%۹۷	%۹۷
۴	%۹۷	%۹۷	%۹۷
۵	%۹۹	%۹۹	%۹۹
۶	%۹۹	%۹۷	%۹۸
۷	%۹۹	%۹۹	%۹۹
۸	%۱۰۰	%۱۰۰	%۱۰۰
۹	%۹۸	%۹۹	%۹۸
دقت : %۹۸			

بحث و نتیجه‌گیری

این پژوهش به بررسی تشخیص اعداد دست‌نویس با استفاده از روش‌های یادگیری ماشین پرداخت. با بهره‌گیری از مجموعه داده Hoda و اعمال پیش‌پردازش‌های لازم، الگوریتم‌های جنگل تصادفی و درخت تصمیم آزمایش شدند. بوسیله ماتریس‌های درهم‌ریختگی و معیارهای ارزیابی نشان داده شد که جنگل تصادفی خطاهای کمتر و دقت پیش‌بینی بهتری نسبت به درخت تصمیم دارد. این برتری به دلیل ترکیب چندین درخت تصمیم و کاهش واریانس در جنگل تصادفی است. در نهایت، این پژوهش نشان داد که انتخاب الگوریتم مناسب نقش کلیدی در بهبود دقت مدل‌های یادگیری ماشین دارد.

مراجع

اسدی، فرشید و صیدی پیری، رسول و نوری، زینب و لطفی پور، امین (۱۳۹۵)، تشخیص حروف دست‌نویس فارسی با استفاده از حسگر شتاب‌سنج و الگوریتم‌های یادگیری ماشین، چهارمین کنفرانس بین‌المللی مهندسی برق و کامپیوتر، تهران، ایران.

زربافی، م. و همکاران (۱۴۰۲)، تشخیص اعداد گفتاری با استفاده از شبکه‌های یادگیری عمیق، مجله علمی-پژوهشی فناوری اطلاعات و ارتباطات، ۱۰، ۴۵-۵۶.

آقایی، امیرعلی و اخوت، راضیه سادات (۱۴۰۳)، مقایسه عملکرد ماشین بردار پشتیبان، درخت تصمیم و شبکه عصبی در تشخیص اعداد دست‌نویس فارسی، نوزدهمین کنفرانس سیستم‌های هوشمند ایران، سیرجان، ایران.

Geron, A. (2019), *Hands on Machine Learning with Scikit Learn, Keras and TensorFlow*, Second Edition, O'Reilly Media, Inc.

Handwritten Digit Recognition Using a Machine Learning Approach

Arezo Hajrajabi[1], **Nooshin Rezaei Taleghani**[2]

[1]Department of Statistics, Imam Khomeini International University, Qazvin

[2]Department of Statistics, Imam Khomeini International University, Qazvin

---

Abstract:

In recent years, with the expansion of visual data and the growing need for its automatic processing and analysis, the issue of handwritten digit recognition has emerged as a significant and practical problem in the fields of machine learning and computer vision. In this research, the problem of handwritten digit recognition has been investigated and implemented using machine learning algorithms such as Random Forest and Decision Tree. For this purpose, the Persian Hoda dataset has been utilized. The performance of these models has been evaluated using appropriate statistical metrics, and the results indicate the superior efficiency of Random Forest compared to Decision Tree.

Keywords: Machine Learning, Handwritten Digit Recognition, Random Forest, Decision Tree.

Mathematics Subject Classification (2020): 68T07, 68T10, 62H30.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مقایسه دو رویکرد نیم‌رخی و نامقید در برازش مدل رگرسیون بر اساس نمونه داده‌های ترکیبی با تواتر متفاوت

فرزین رنجبر[۱]، علی آقامحمدی[۲] ، مجید ادیب[۳]

[۲]گروه آمار، دانشگاه زنجان

[۳]گروه ریاضی، دانشگاه زنجان

چکیده: این مقاله به مقایسه روش نیم‌رخی و نامقید در برآورد پارامترهای مدل رگرسیون بر اساس نمونه داده‌های ترکیبی با تواتر متفاوت (MIDAS) می‌پردازد. روش نیم‌رخی با ترکیب حداقل مربعات معمولی (OLS) برای برآورد ضرایب رگرسیونی و بهینه‌سازی پارامترهای وزن‌دهی چندجمله‌ای، کارایی محاسباتی بهتر و خطای پیش‌بینی کمتری نسبت به روش براورد پارامتر نامقید نشان می‌دهد.

واژه‌های کلیدی: مدل رگرسیونی نمونه داده‌های ترکیبی با تواتر متفاوت، روش برآورد پارامتر نامقید، روش برآورد پارامتر نیم‌رخی.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62M10، 91B84.

---

۱ مقدمه

در اقتصادسنجی مدرن، تحلیل داده‌های سری زمانی با تواتر متفاوت به یکی از چالش‌های اساسی محققان تبدیل شده است. مدل‌های رگرسیونی بر اساس نمونه داده‌های ترکیبی با تواتر متفاوت (MIDAS) [۲] که توسط گیلز و همکاران (۲۰۰۴) و گیلز و همکاران (۲۰۰۶) معرفی شدند، ابزاری قدرتمند برای مواجهه با این چالش محسوب می‌شوند. این مدل‌ها قابلیت ترکیب اطلاعات متغیرهای اقتصادی با تواترهای متفاوت نظیر داده‌های روزانه، ماهانه و فصلی را فراهم می‌آورند. اهمیت این تحقیق در چندین جنبه نمایان می‌شود. نخست، مدل‌های MIDAS سنتی نیازمند استفاده از روش‌های حداقل مربعات غیرخطی هستند که از لحاظ محاسباتی پیچیده و گاهی دارای مشکلات همگرایی می‌باشند آندرو و همکاران (۲۰۱۰).

در ادامه برای کاهش پیچیدگی محاسباتی و کاهش تعداد پارامترها، روش استنباطی نیم‌رخی ارائه شده است که ابتدا به مفهوم آن می‌پردازیم.

---

[۱] Farzin.r@znu.ac.ir

[۲] Mixed Data Sampling

مفهوم تابع درستنمایی نیم‌رخی [۱] ریشه‌ای عمیق در ادبیات آماری دارد و توسط محققانی چون پاتیفلد (۱۹۷۷) توسعه یافته است. فرض کنید متغیرهای تصادفی y_t مستقل و هم‌توزیع با تابع چگالی احتمال $f(y;\delta,\theta)$ باشند و هدف برآورد پارامترهای δ,θ باشد. با یک نمونه به اندازه T لگاریتم تابع درستنمایی به صورت زیر تعریف می‌شود.

$$\mathcal{L}(\delta,\theta)=\sum_{t=1}^{T}\log f(y_t;\delta,\theta)$$

در بسیاری مواقع، بیشینه‌سازی این تابع نسبت به تمامی پارامترها دشوار است. اما اگر θ را ثابت در نظر گرفته شود، یعنی $\theta=\bar{\theta}$، بیشینه‌سازی نسبت به δ برای θ ثابت ساده‌تر است. در چنین مواردی می‌توان لگاریتم تابع درستنمایی به شکل $L_{T,\delta}(\theta)$ بازنویسی کرد که به آن لگاریتم تابع درستنمایی نیم‌رخی گفته می‌شود و نسبت به δ برای یک θ معین بهینه می‌شود. بنابرین مسئله بهینه‌سازی به $(\hat{\delta},\hat{\theta})=\arg\max_{\theta}L_{T,\delta}(\theta)$ تبدیل می‌شود. این توابع عمدتاً زمانی مورد استفاده قرار می‌گیرند که δ پارامتری با بعد کم و θ پارامتری با بعد بالا یا مزاحم محسوب شود. مورفی (۲۰۰۰) توجیه کلی برای استفاده از تابع درستنمایی نیم‌رخی را به عنوان یک رهیافت استنباطی ارائه دادند.

در این پژوهش، رویکرد تابع درستنمایی نیم‌رخی برای مدل‌های رگرسیون MIDAS ارائه می‌شود که در آن θ نشان‌دهنده پارامترهای وزن چندجمله‌ای MIDAS و δ شامل پارامترهای عرض از مبدأ و شیب مدل رگرسیون است. برای مقدار ثابت $\theta=\bar{\theta}$، رویکرد نیم‌رخی مسئله برآورد را به مجموعه‌ای از برآوردگرهای حداقل مربعات معمولی (OLS) مدل‌های رگرسیون خطی نسبتاً ساده $\hat{\delta}(\bar{\theta})$ تقلیل می‌دهد. از آنجا که پارامترهای چندجمله‌ای MIDAS یعنی θ دارای ابعاد پایین هستند - گاهی اوقات حتی تک‌بعدی - انتخاب $\bar{\theta}$ در یک شبکه نقاط، نسبتاً آسان است. که این امر بهینه‌سازی را از نظر محاسباتی ساده می‌نماید.

روش دیگر در تحلیل مدل های رگرسیون MIDAS ، روش MIDAS نامقید (U-MIDAS) [۲] است که توسط فورنی و همکاران (۲۰۱۵) پیشنهاد شده است. U-MIDAS یک روش برآورد برای مدل‌های MIDAS است که برخلاف روش استاندارد MIDAS که از توابع پارامتریک (مانند چندجمله‌ای بتاو ...) برای وزن‌دهی به وقفه داده‌های با تواتر بالا استفاده می‌کند. U-MIDAS هیچ محدودیت پارامتریکی بر روی وزن‌های وقفه‌ها اعمال نمی‌کند. در عوض، این روش هر وزن وقفه را به صورت جداگانه و بدون ساختار با استفاده از حداقل مربعات معمولی (OLS) برآورد می‌کند. این روش زمانی جذاب است که اختلاف تواترها کم باشد. مثلاً در داده‌های فصلی/ماهانه (که هر فصل شامل ۳ ماه است)، U-MIDAS می‌تواند ۳، ۶ یا ۹ وقفه ماهانه را مستقیماً در مدل وارد کند و تمام پارامترها را با OLS برآورد کند. این رویکرد برای مواردی که ساختار وزن‌ها نامشخص است یا تفاوت تواترها کم است، مناسب می‌باشد. اما اگر اختلاف تواترها بزرگ باشد (مثلاً داده‌های روزانه در مدل‌های فصلی که هر فصل شامل ۹۰ روز است) U-MIDAS نیاز به تخمین تعداد زیادی پارامتر (مثلاً وزن ۹۰ وقفه) دارد که منجر به پیچیدگی محاسباتی بالا و مشکل بیش برازش می‌شود. نتیجه قابل توجه و غیرمنتظره این است که در مواردی که U-MIDAS جذاب به نظر می‌رسد، رویکرد جدید نیم‌رخی برای مدل رگرسیون MIDAS از نظر محاسباتی سریع‌تر از روش OLS است فورنی و همکاران (۲۰۱۵). دلیل این امر آن است که U-MIDAS مثلاً در حالت فصلی/ماهانه یک مدل رگرسیونی با ۳، ۶ یا ۹ وقفه از داده‌های ماهانه و داده‌های فصلی با وقفه است که بر روی داده‌های فصلی جاری یا آینده تصویر می‌شوند، در حالی که روش برآورد نیم‌رخی تنها شامل یک متغیر توضیحی (با نادیده گرفتن عرض از مبدأ در هر دو حالت) است که این متغیر توضیحی با اعمال چندجمله‌ای MIDAS (با فرض $\bar{\theta}$) بر داده‌های با تواتر بالا ساخته می‌شود. روش برآورد نیم‌رخی از نظر محاسباتی کارآمد، از لحاظ آماری جذاب و پیچیدگی محاسباتی روش کمترین مربعات غیرخطی را برای برآورد پارامترهای مدل MIDAS را ندارد.

[۱] Profiling

[۲] Unrestricted Mixed Data Sampling

در مجموعه مدل‌های MIDAS ، مطالعات متنوعی انجام شده است. کلمنتز و همکاران (۲۰۰۹) بر جنبه‌های پیش‌بینی این مدل‌ها تمرکز کردند، در حالی که بررسی‌های جامع‌تری توسط آرمستو و همکاران (۲۰۱۰) و آندرو و همکاران (۲۰۱۰) انجام شده است. گیلز و همکاران (۲۰۱۱) نخستین بار ایده نیم‌رخی را در رگرسیون چندکی MIDAS به کار گرفتند. هدف اصلی این مطالعه، ارائه روشی نوین برای برازش مدل‌های MIDAS است که مزایای محاسباتی OLS را با دقت و انعطاف‌پذیری مدل‌های پارامتری ترکیب می‌کند. نوآوری اصلی در استفاده از رویکرد نیم‌رخی برای برآورد پارامترهای چندجمله‌ای بتا نهفته است، که امکان کاهش پیچیدگی محاسباتی را فراهم می‌آورد.

یافته‌های ما نشان می‌دهد که مدل MIDAS با رویکرد نیم‌رخی، علی‌رغم دارا بودن ضریب تعیین (R^2) کمتر در نمونه، از قدرت پیش‌بینی بالاتری در خارج از نمونه برخوردار است. میانگین مربعات خطای پیش‌بینی خارج از نمونه در روش نیم‌رخی به طور قابل توجهی پایین‌تر است. این یافته نشان‌دهنده این است که U-MIDAS مشکل بیش برازشی نیز دارد. در بخش دوم مدل رگرسیون MIDAS به اختصار توضیح داده شده و رویکردهای نیم‌رخی و نامقید در این مدل‌ها بررسی می‌شوند. بخش سوم شامل شبیه‌سازی برای مقایسه این رویکرد و تحلیل نتایج حاصل از آن است.

۲ مدل رگرسیون MIDAS

مدل‌های MIDAS معمولا بر ترکیبی از دو تواتر متفاوت، به ترتیب تواتر بالا و پایین، تمرکز دارند. فرض کنید $t = 1, \ldots, T$ نشان‌دهنده واحد زمانی با تواتر پایین و m تعداد دفعاتی است که تواتر بالا در همان واحد زمانی ظاهر می‌شود. به عنوان مثال، برای رشد تولید ناخالص داخلی فصلی و شاخص‌های ماهانه به عنوان متغیرهای توضیحی، $m = 3$ خواهد بود. متغیر با فرکانس پایین با y_t^L و متغیر با فرکانس بالا نیز با $x_{t-j/m}^H$ نمایش داده می‌شود که در آن $t - j/m$ بیانگر دوره زمانی jام (گذشته) با تواتر بالاتر است و $j = 0, 1, \ldots$. برای یک ترکیب فصلی/ماهانه، x_t^H، $x_{t-1/3}^H$، $x_{t-2/3}^H$ به ترتیب نشان‌دهنده آخرین، دومین و اولین ماه فصل t هستند.

مدل‌های رگرسیونی MIDAS در اصل نوعی مدل‌های رگرسیونی هستند که داده‌هایی با تواترهای متفاوت (مثلاً ماهانه و فصلی) را به طور همزمان در یک مدل ترکیب می‌کنند. تأثیر متغیر با تواتر بالا با استفاده از یک تابع وزن‌دهی مدل می‌شود که به‌صورت خودکار وزن‌های تمام وقفه‌های گذشته را محاسبه و توزیع می‌کند. این کار باعث می‌شود تعداد پارامترهای مدل به حداقل برسد و مشکلات مربوط به انتخاب تعداد وقفه‌های مناسب نیز رفع گردد.

مدل اساسی MIDAS با یک متغیر توضیحی با تواتر بالا برای پیشبینی h دوره بعد در تواتر پایین، در حالتی که داده‌های با تواتر بالا تا x_t^H در دسترس هستند، به صورت زیر

$$y_{t+h}^L = a_h + b_h C(L^{1/m}; \theta_h) x_t^H + \varepsilon_{t+h}^L \tag{۱.۲}$$

است، که در آن $C(L^{1/m}; \theta) = \sum_{j=0}^{j_{\max}-1} c(j;\theta) L^{j/m}$ و $C(1;\theta) = \sum_{j=0}^{j_{\max}-1} c(j;\theta) = 1$ است. مؤلفه j^{max} حداکثر تعداد وقفه‌هایی است که از متغیر با تواتر بالا (مثلاً داده‌های روزانه) در مدل استفاده می‌شود.این مدل اغلب DL-MIDAS نامیده می‌شود که DL به معنای وقفه توزیع‌شده است. پارامتری‌سازی ضرایب وقفه‌دار $c(j;\theta)$ به شیوه‌ای صرفه‌جویانه، یکی از ویژگی‌های کلیدی مدل‌های MIDAS است.

هرچند محدود کردن فضای پارامتری θ به یک بعد اجباری نیست، اما این کار محاسبات را تسهیل می‌کند. در این راستا، چندجمله‌ای‌های

بتا که ابتدا توسط گیلز و همکاران (۲۰۰۶) پیشنهاد شدند، گزینه مناسبی هستند. این توابع مبتنی بر توزیع بتا بوده و تنها به دو پارامتر نیاز دارند، یعنی:

$$c(j;\theta_1,\theta_2)=\frac{f(\frac{j}{j_{\max}},\theta_1;\theta_2)}{\sum_{j=0}^{j_{\max}-1}f(\frac{j}{j_{\max}},\theta_1;\theta_2)}, \tag{۲.۲}$$

که در آن

$$f(x,a,b)=\frac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}, \tag{۳.۲}$$

و $\Gamma(a)=\int_0^\infty e^{-x}x^{a-1}$ است. با ثابت نگه‌داشتن یکی از پارامترها (مثلاً $\theta_1=1$) ، می‌توان مدل را به یک پارامتر (θ_2) تقلیل داد که برای داده‌های اقتصادی مطلوب است.

۱.۲ مدل اتورگرسیو MIDAS

روش ارائه‌شده در این مقاله به مدل DL-MIDAS در معادله (۱.۲) محدود نمی‌شود. به عنوان مثال، افزودن متغیرهای وابسته با وقفه (یعنی متغیرهای کم تواتر) منجر به دسته‌ای از مدل‌های ADL-MIDAS [1] می‌شود. برای ساده‌سازی نمادگذاری، در ادامه مقاله (بدون از دست دادن کلیت مسئله) فرض می‌کنیم $h=1$. با این فرض یک مدل اتورگرسیو از مرتبه p را به صورت

$$y_{t+1}^L=a+\sum_{j=1}^{p}\rho_j y_{t-j+1}^L+bC(L^{1/m};\theta)x_t^H+\varepsilon_{t+1}^L \tag{۴.۲}$$

بازنویسی می‌کنیم. حال می‌توان روش نیم‌رخی را با ثابت درنظر گرفتن پارامترهای چند جمله‌ای MIDAS برای این مدل نیز بکار گرفت.

۲.۲ مدل MIDAS نامقید

مدل U-MIDAS که توسط فورنی و همکاران (۲۰۱۵) معرفی شد. برای توصیف این مدل فرض کنید m کوچک باشد، (مثلاً برابر ۳ برای ترکیب داده‌های فصلی/ماهانه) برخلاف مدل استاندارد MIDAS از هیچ تابع وزن‌دهی از پیش تعیین‌شده‌ای در این مدل استفاده نمی‌شود. برای هر وقفه از متغیر با تواتر بالا، یک ضریب جداگانه در نظر گرفته می‌شود. یعنی به جای برآورد $bC(L^{1/m};\theta)$ در معادله ۴.۲ ضرایب متغیر با تواتر بالا به صورت جداگانه برآورد می‌شود. فرم کلی این مدل به صورت

$$y_{t+1}=a+\sum_{j=1}^{p}\rho_j y_{t-j+1}^L+\sum_{j=0}^{J_{\max}-1}c_j x_{t-j/m}^H+\varepsilon_{t+1}^L \tag{۵.۲}$$

است. این بدان معناست که علاوه بر پارامترهای a و ρ_j، ما $j_{max}-2$ پارامتر اضافی را تخمین می‌زنیم. وقتی $m=3$ و $j_{max}-1$ کوچک باشد (مثلاً حداکثر وقفه سالانه برابر ۴ باشد) و حجم نمونه به اندازه کافی بزرگ باشد که جمله خطا ϵ_{t+1}^L را ناهمبسته سازد، پارامترهای مدل U-MIDAS را می‌توان با حداقل مربعات معمولی (OLS) تخمین زد.

[1] Autoregressive Distributed Lag MIDAS

۳ شبیه‌سازی

در این بخش با استفاده از یک مطالعه شبیه‌سازی به بررسی کارایی هریک از رویکردهای ارائه شده می‌پردازیم. فرض کنید مقدار $m = ۳$ برای ترکیب داده‌های فصلی/ماهانه باشد. مدل رگرسیون MIDAS را به صورت

$$y^L_{t+۱} = \sum_{j=۰}^{j^{\max}-۱} c(j,\theta) x^H_{t-\frac{j}{m}} + \epsilon^L_{t+۱}, \quad t = ۱, \ldots, T \tag{۱.۳}$$

در نظر می‌گیریم. در این مدل به دلیل اینکه ضرایب عرض از مبدأ و اتورگرسیو مربوط به متغیر کم تواتر در تحلیل ضروری نیستند، بنابراین کنار گذاشته می‌شوند. مدل U-MIDAS فرض می‌کند که ضرایب بدون هیچ محدودیتی و با استفاده از روش حداقل مربعات معمولی (OLS) برآورد می‌شوند؛ در حالی که مدل MIDAS با چندجمله‌ای بتا، وزن‌ها را به‌صورت بسیار مقید و پارامتریک به شکل $c(j,\theta) \propto (۱ - \frac{j}{j_{\max}})^{\theta-۱}$ مدل‌سازی می‌کند. روش نیم‌رخی فرض می‌کند پارامتر θ ثابت است و با روش حداقل مربعات معمولی، پارامترهای دیگر (مثل شیب) را برآورد کرده سپس مقدار θ را طوری تنظیم می‌کند که مدل بهترین برازش را داشته باشد.

در شبیه‌سازی مونت‌کارلو، داده‌های ماهانه از یک مدل $AR(۱)$ با خودهمبستگی ۰/۸ و داده‌های فصلی از مدل MIDAS با چندجمله‌ای بتا تولید شده‌اند. مقادیر خودهمبستگی‌های ۰/۱, ۰/۹ را نیز آزمایش شد و نتایج مشابهی حاصل گردید. مقادیر $T = \{۵۰, ۲۰۰, ۲۰۰۰\}$ (اندازه نمونه)، $j^{max} = \{۶, ۹, ۱۲\}$ (حداکثر وقفه) و $\theta = \{۲, ۱۰\}$ (پارامترهای توزیع بتا) را در نظر می‌گیریم که ترکیبات آنها منجر به ۱۸ سناریوی مختلف می‌شود. شبیه‌سازی را به تعداد ۱۰۰۰ مرتبه تکرار کرده و در هر سناریو و تکرار، داده‌های ماهانه/فصلی را شبیه‌سازی می‌کنیم سپس مدل رگرسیون MIDAS را با استفاده از رویکرد نیم‌رخی و مدل U-MIDAS را بر داده‌ها برازش می‌دهیم. مدل‌ها را با استفاده از معیارهای ارزیابی به شرح ذیل ارزیابی می‌کنیم:

$$Bias^۲ = \sum_{j=۰}^{j_{\max}-۱} (\hat{c}(j,\theta) - c(j,\theta))^۲,$$

$$R^۲ = ۱ - \frac{\sum_{t=۱}^{T} (\hat{y}^L_t - y^L_t)^۲}{\sum_{t=۱}^{T} y^۲_t},$$

$$MSE = \frac{۱}{r} \sum_{t=T+۱}^{T+r} (\hat{y}^L_t - y^L_t)^۲,$$

که در آن $\hat{c}(j,\theta)$ وزن‌های برآورد شده و $\hat{y}^L_t$ مقدار برازش شده است.

نتایج شبیه‌سازی در جدول ۱ ارائه شده است که در آن نتایج ۱۸ سناریو شبیه‌سازی برای سه شاخص عنوان شده، گزاش شده است. با توجه به جدول ۱ مشخص است که برآوردهای نیم‌رخی در تمامی سناریوها مربع اریبی کمتری نسبت به U-MIDAS دارد. به عنوان مثال، در سناریوی $T = ۵۰$، $j^{max} = ۶$ و $\theta = ۲$، مربع اریبی روش نیم‌رخی ۰.۰۸۷ است، در حالی که این مقدار برای U-MIDAS به ۰.۳۹۱ می‌رسد. این نشان می‌دهد که برآورد نیم‌رخی به طور قابل توجهی دقیق‌تر از U-MIDAS است.

در بیشتر سناریوها $R^۲$ درون‌نمونه‌ای روش نیم‌رخی کمتر از U-MIDAS است. با این حال، قدرت پیش‌بینی خارج از نمونه‌ای آن بالاتر است، به طوری که MSE پیش‌بینی آن به طور قابل توجهی کمتر از U-MIDAS است. به عنوان مثال، در سناریوی $T = ۵۰$، $j_{max} = ۱۲$ و $\theta = ۲$، MSE پیش‌بینی روش نیم‌رخی ۰.۹۸۹ است، در حالی که این مقدار برای U-MIDAS برابر ۱.۳۲۲ (حدود ۲۶٪ بیشتر) است. این نتایج نشان می‌دهد که U-MIDAS از مشکل برازش بیش از حد رنج می‌برد. پارامترهای غیرمقید در U-MIDAS می‌توانند با $R^۲$ درون‌نمونه‌ای بالا تنظیم شوند، اما نمی‌توانند در پیش‌بینی برون‌نمونه‌ای از یک مدل کم پارامتر پیشی

جدول ۱: مقایسه عملکرد پروفایلی و U-MIDAS در ۱۸ سناریوی شبیه‌سازی‌شده

T	$j_{\max}$	θ	$Bias^2$		R^2		MSE Forecast		FLOPS (10^3)	
			پروفایلی	U-MIDAS	پروفایلی	U-MIDAS	پروفایلی	U-MIDAS	پروفایلی	U-MIDAS
۵۰	۶	۲	۰.۰۸۷	۰.۳۹۱	۰.۶۵۸	۰.۶۹۰	۱.۰۴۱	۱.۱۴۸	۵.۸	۷.۶
۵۰	۶	۱۰	۰.۱۱۹	۰.۴۰۶	۰.۷۴۲	۰.۷۶۵	۱.۰۲۵	۱.۱۲۲	۵.۸	۷.۶
۵۰	۹	۲	۰.۰۶۶	۰.۴۹۷	۰.۶۱۰	۰.۶۸۷	۱.۰۳۴	۱.۱۹۷	۹.۱	۱۲.۱
۵۰	۹	۱۰	۰.۱۲۷	۰.۵۰۵	۰.۷۲۱	۰.۷۶۲	۱.۰۶۴	۱.۲۴۷	۹.۱	۱۲.۱
۵۰	۱۲	۲	۰.۰۵۵	۰.۶۱۹	۰.۵۷۳	۰.۶۶۹	۰.۹۸۹	۱.۳۲۲	۱۲.۳	۱۹.۸
۵۰	۱۲	۱۰	۰.۱۱۴	۰.۶۳۱	۰.۶۹۳	۰.۷۵۵	۱.۰۳۶	۱.۳۲۵	۱۲.۳	۱۹.۸
۲۰۰	۶	۲	۰.۰۴۱	۰.۱۸۱	۰.۶۶۵	۰.۶۷۷	۱.۰۰۳	۱.۰۴۷	۲۹.۴	۲۰.۲
۲۰۰	۶	۱۰	۰.۰۶۵	۰.۱۹۰	۰.۷۳۱	۰.۷۴۰	۱.۰۲۴	۱.۰۶۷	۲۹.۴	۲۰.۲
۲۰۰	۹	۲	۰.۰۲۸	۰.۲۳۴	۰.۶۲۷	۰.۶۶۴	۱.۰۱۷	۱.۰۸۹	۳۷.۱	۴۱.۴
۲۰۰	۹	۱۰	۰.۰۵۹	۰.۲۴۱	۰.۷۰۱	۰.۷۳۲	۱.۰۲۹	۱.۱۱۱	۳۷.۱	۴۱.۴
۲۰۰	۱۲	۲	۰.۰۲۲	۰.۳۱۰	۰.۵۹۸	۰.۶۴۸	۰.۹۷۱	۱.۱۲۸	۴۴.۹	۶۹.۲
۲۰۰	۱۲	۱۰	۰.۰۵۲	۰.۳۱۷	۰.۶۹۲	۰.۷۲۵	۱.۰۱۵	۱.۱۳۰	۴۴.۹	۶۹.۲
۲۰۰۰	۶	۲	۰.۰۱۳	۰.۰۶۲	۰.۶۷۰	۰.۶۷۳	۰.۹۸۰	۰.۹۸۱	۲۹۵.۰	۲۰۴.۹
۲۰۰۰	۶	۱۰	۰.۰۳۲	۰.۰۶۴	۰.۷۳۳	۰.۷۳۵	۱.۰۰۰	۱.۰۰۲	۲۹۵.۰	۲۰۴.۹
۲۰۰۰	۹	۲	۰.۰۱۰	۰.۰۸۲	۰.۶۳۶	۰.۶۴۰	۰.۹۸۷	۰.۹۹۱	۳۶۶.۲	۴۱۶.۷
۲۰۰۰	۹	۱۰	۰.۰۳۷	۰.۰۸۵	۰.۷۱۰	۰.۷۱۴	۱.۰۰۶	۱.۰۱۰	۳۶۶.۲	۴۱۶.۷
۲۰۰۰	۱۲	۲	۰.۰۰۸	۰.۱۰۲	۰.۶۰۷	۰.۶۱۶	۱.۰۱۲	۱.۰۲۰	۴۷۳.۴	۷۰۰.۸
۲۰۰۰	۱۲	۱۰	۰.۰۳۶	۰.۱۰۵	۰.۶۹۲	۰.۶۹۹	۰.۹۹۹	۱.۰۰۴	۴۳۷.۴	۷۰۰.۸

بگیرند.

با افزایش اندازه نمونه (T) و ثابت نگه داشتن j_{max}، هر دو روش OLS و نیم‌رخی سازگار هستند. مربع اریبی هر دو روش با افزایش اندازه نمونه کاهش می‌یابد. برای مثال، مربع اریبی روش نیم‌رخی (U-MIDAS) از ۰.۰۸۷ (۰.۳۹۱) به ۰.۰۴۱ (۰.۱۸۱) و سپس به ۰.۰۱۳ (۰.۰۶۲) کاهش می‌یابد، همچنان که T از ۵۰ به ۲۰۰ و سپس به ۲۰۰۰ افزایش می‌یابد. نتایج شبیه‌سازی تأیید می‌کند که روش نیم‌رخی کاراتر از برآوردگر OLS است.

تعداد وقفه‌ها نیز عامل کلیدی در عملکرد پیش‌بینی است. نتایج شبیه‌سازی نشان می‌دهد که U-MIDAS در سناریوهای با j^{max} بزرگ و T کوچک، عملکرد ضعیف‌تری نسبت به روش نیم‌رخی دارد. در حالت $T = ۵۰$، $j^{max} = ۱۲$ و $\theta = ۲$، MSE روش نیم‌رخی ۰.۹۸۹ است، در حالی که MSE روش U-MIDAS برابر ۱.۳۲۲ است (افزایش ۲۶%).

مقایسه تعداد عملیات محاسباتی مورد نیاز (FLOPS) نشان می‌دهد که روش پروفایلی به طور ملموسی ساده‌تر و سریع‌تر از U-MIDAS است. با رشد اندازه نمونه و تعداد وقفه‌ها، اختلاف بار محاسباتی میان دو روش افزایش می‌یابد. این مزیت محاسباتی روش نیم‌رخی را

برای کاربردهای عملی با حجم داده بزرگ و یا مدل‌های با وقفه زیاد، گزینه‌ای ایده‌آل‌تر می‌سازد.

بحث و نتیجه‌گیری

شبیه‌سازی‌های انجام‌شده نشان داد که روش نیم‌رخی در تمامی سناریوها از نظر اریبی کمتر، قدرت پیش‌بینی بالاتر و کارایی محاسباتی برتری دارد. اگرچه U-MIDAS ممکن است درون نمونه R^2 بالاتری ارائه کند اما مشکل بیش‌برازش و افزایش MSE خارج نمونه را دارد. بنابراین برای کاربردهای عملی MIDAS به ویژه در موارد نیازمند پیش‌بینی دقیق و صرفه‌جویی محاسباتی، روش نیم‌رخی توصیه می‌گردد.

مراجع

Andreou, E., Ghysels, E., Kourtellos, A., (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics* 158(2), 246–261.

Armesto, M.T., Engemann, K.M., Owyang, M.T., (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review* 92(6), 521–536.

Clements, M.P., Galvão, A.B., (2009). Forecasting US output growth using leading indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics* 24(7), 1187–1206.

Foroni, C., Marcellino, M., Schumacher, C., (2015).Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A* 178(1), 57–82.

Ghysels, E., Santa-Clara, P., Valkanov, R., (2004). The MIDAS touch: Mixed data sampling regression models. *UNC Working Paper*.

Ghysels, E., Santa-Clara, P., Valkanov, R., (2006a). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics* 131(1-2), 59–95.

Ghysels, E., Qian, H., Sargent, T.J., (2011). MIDAS regressions with MIDAS shocks. *UNC Working Paper*.

Murphy, S.A., van der Vaart, A.W., (2000). On profile likelihood. *Journal of the American Statistical Association* 95(450), 449–465.

Patefield, W.M., (1977). On the maximized likelihood function. *Sankhyā: The Indian Journal of Statistics, Series B* 39(1), 92–96.

A Comparison of the Profiling and Unrestricted Approaches in Fitting Regression Models Based on Mixed-Frequency Data Sampling

Farzin ranjbar, Ali Aghamohammadi[1], Majid Adib[2]

[1]Department of Statistics, University of Zanjan

[2]Department of Mathematics, University of Zanjan

Abstract: This article compares the profiling and unrestricted estimation methods for parameter estimation in regression models based on mixed-frequency data (MIDAS). The profile method, which combines ordinary least squares (OLS) for estimating regression coefficients with the optimization of polynomial weighting parameters, demonstrates superior computational efficiency and lower prediction error compared to the unrestricted parameter estimation approach.

Keywords: Mixed data sampling regression, Unrestricted estimation method, Profiling estimation method.

Mathematics Subject Classification (2020): 62M10, 91B84.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

برخی نتایج مجانبی در مورد اختلاف میانگین دو جامعه در داده‌های تابعی جزئی مشاهده‌شده

پویا سلیمی فر [۱] ، سید محمد ابراهیم حسینی نسب[۲] ، امید خادم نوع [۳]

[۱] دانشگاه شهید بهشتی

[۲] دانشگاه شهید بهشتی [۳] دانشگاه زنجان

چکیده: امروزه ثبت اطلاعات از مشاهداتی مانند $X_1, X_2, \ldots, X_n$ که به‌صورت تابعی پیوسته از یک پدیده باشند، مرسوم است. مشاهدات تابعی اندازه‌گیری‌هایی از پدیده‌ای است که نسبت به یک متغیر مانند زمان، بسامد و . . . پیوسته است. ماهیت پیوسته این داده‌ها منبع غنی از اطلاعات است و امکان مدلسازی با دقت بالا را فراهم می‌کند، در برخی شرایط و مطالعات تجربی عواملی مانند محدودیت زمانی در مطالعه، ابزارهای اندازه‌گیری ناکارامد یا عوامل ناشناخته سبب می‌شوند ثبت داده‌ها در سراسر دامنه‌ای معین در جریان مطالعه به‌صورت کامل انجام نگیرد، در این وضعیت داده‌های تابعی به‌صورت ناکامل یا جزئی مشاهده‌شده [۱] در دسترس هستند و روش‌های برآوردیابی معمول که برای داده‌های تابعی کامل ارائه شده است، قابل استفاده نیست. یک رویکرد این است که با در نظر گرفتن مقدار تابع در زیر مجموعه‌های مشاهده‌شده از کل دامنه به تحلیل داده‌ها پرداخته می‌شود. در این مقاله، برخی خواص مجانبی از اختلاف میانگین دو جامعه در داده‌های تابعی جزئی مشاهده‌شده ارائه می‌شود.

واژه‌های کلیدی: داده‌های تابعی جزئی مشاهده‌شده، همگرایی در احتمال، فرآیند گاوسی، عملگر قطری .

---

۱ مقدمه

داده‌های تابعی جزئی مشاهده شده شامل تحقق‌هایی از توابع تصادفی هستند که در کل دامنه مشاهده نشده و هر تابع در نمونه ممکن است در زیر مجموعه‌های متفاوتی از دامنه مشاهده شود و هیچ اطلاعی درباره مقادیر تابع در سایر نقاط دامنه در دسترس نباشد. با در نظر گرفتن $\tau = [0, 1]$ برای i امین متغیر تابعی $X_i \in L^2[0, 1]$ زیر مجموعه‌ای مانند $O_i \subseteq [0, 1]$ وجود دارد به طوری که $X_i(t)$ برای $t \in O_i$ مشاهده شده است و برای $t \in \tau \setminus O_i$ مشاهده نشده است، به طور کلی الگوی گمشدگی در مطالعات تابعی

---

[۱] سخنران، p_salimifar@sbu.ac.ir

[1] Partially observed

نیز می‌تواند در چهار حالت نمایان شود. اولین حالت که به آن اشاره می‌شود، الگوی گمشدگی کاملا تصادفی [۲] است که در آن گمشدگی بخش‌هایی از تابع، مستقل از تابع (به این معنا که در آن ساختار گمشدگی نه به بخش‌های مشاهده شده و نیز به بخش‌های گمشده از تابع وابسته نیست) است. دومین حالت، الگوی گمشدگی تصادفی نام دارد که در آن گمشدگی تنها با بخش مشاهده شده از تابع مرتبط است و سومین الگوی گمشدگی، گمشدگی غیر تصادفی نام دارد که به هر دو بخش مشاهده شده و گمشده از تابع وابسته خواهد بود. الگوی چهارم که الگوی گمشدگی سیستماتیک [۳] است از سه وضعیت موجود متفاوت است و در آن اغلب یک عامل شناخته شده در جریان مطالعه سبب گمشدگی بخشی از تابع در دامنه می‌شود، همچنین یادآور می‌شویم که تاکنون الگوی گمشدگی تصادفی و غیر تصادفی به دلیل فرض‌های بسیار محدود کننده بر الگوی داده‌ها در پژوهش‌های مربوط به داده‌های تابعی مورد مطالعه قرار نگرفته‌اند. مسائل کاربردی مرتبط با این حالت باعث ایجاد یک مجموعه کارهای تحقیقاتی با جنبه‌های مختلف در این مسئله شده است، باگنی (۲۰۱۲)، به آزمون مشخصه‌سازی برای داده‌های تابعی ناکامل (که به ارزیابی اعتبار فرضیات احتمالی تشکیل دهنده یک مدل آماری اشاره دارد) پرداخت، کراووس (۲۰۱۵)، داده‌های پایش فشار خون سرپایی و نمایه ضربان قلب شامل الگو تغییرات ضربان قلب در حالات مختلف فعالیت، استراحت و یا شرایط خاص پزشکی را مورد مطالعه قرار داد و با استفاده از قسمت مشاهده شده از منحنی ضربان قلب برای قسمت گمشده از آن یک پیشگو ارائه کرد، سپس برای قسمت گمشده از منحنی یک باند پیشگویی ساخت. دلایگل و هال (۲۰۱۶)، داده‌های تابعی ناکامل را با استفاده از بخش‌هایی از یک مدل زنجیر مارکوف تقریب زدند. گرومنکو و همکاران (۲۰۱۷)، در یک مطالعه به بررسی روند سرمایش در لایه یونوسفر با استفاده از رگرسیون تابعی با منحنی‌های ناکامل پرداختند. شیبانی و همکاران (۲۰۱۸)، به طبقه‌بندی با متغیرهای کمکی تابعی ناقص پرداختند. اِستِفنسکی و همکاران (۲۰۱۸)، به تفکیک داده‌های تابعی جزئی مشاهده شده بر اساس تحلیل مولفه‌های اصلی پرداختند. دیسکری و همکاران (۲۰۱۹)، به بازسازی کوواریانس از مشاهدات تابعی ناکامل پرداختند. لیبل و رامسدر (۲۰۱۹)، در یک مطالعه که هدف آن، برآورد منحنی میانگین قیمت مصرف انرژی در بازار ذخیره کنترلی برق در آلمان بود، تمرکز کردند و با توجه به این که در این بازار مزایده‌ای، منحنی‌های قیمت تنها به صورت جزئی در انتها دامنه مشاهده می‌شدند مکانیسم گمشدگی به‌صورت سیتماتیک، وابسته به استراتژی های معاملاتی (عامل شناخته شده) بود و این عامل سبب نقض فرض گمشدگی کاملاً تصادفی می‌شد و نتایج ناسازگاری را در برآوردگر میانگین ایجاد می‌کرد از این رو آنها از طریق قضیه اساسی حساب دیفرانسیل و انتگرال و مشتق پذیری داده‌های تابعی برآوردهای سازگاری برای تابع میانگین و عملگر کوواریانس در این وضعیت ارائه دادند. یان کیو و همکاران (۲۰۲۱)، به مطالعه‌ی آزمون فرض دو نمونه‌ای برای میانگین پرداختند. نایپ و لیبل (۲۰۲۰)، یک روش بازسازی برای قسمت گمشده از منحنی‌های جزئی مشاهده شده ارائه دادند. از دیگر مطالعات می‌توان به کتب منتشر شده توسط روبین (۱۹۷۶)، هی و همکاران (۲۰۲۲)، اشاره کرد.

با توجه به این که استنباط در مورد میانگین داده‌های تابعی جزئی مشاهده‌شده از اهمیت بالایی برخوردار است و در بسیاری از مطالعات شامل آزمون فرض مورد استفاده قرار می‌گیرد در این پژوهش نیز به مطالعه‌ی برخی خواص مجانبی در مورد اختلاف میانگین یک متغیر تابعی در دو جامعه پرداخته می‌شود. نتایج این پژوهش می‌تواند در آزمون فرض برای برابری میانگین دو جامعه و در حالت تعمیم یافته آن به چند جامعه مورد استفاده قرار گیرد.

[2] Missing-Completely-At-Random

[3] Systematic Missingness

۲ برخی نتایج حدی

در مطالعه‌ی داده‌های تابعی جزئی باید دو عامل مهم در نظر گرفته شود، نخست این که وزن های مربوط به شانس گمشدگی در هر گروه متفاوت است و باید در ساختار اختلاف میانگین‌ها اعمال شود، دوم این که ممکن است در اکثر مطالعات واقعی برای هر گروه تابع کوواریانس متمایزی وجود داشته باشد که هر دو این عوامل باید در ساختار مسئله اعمال شود. اکنون به تعریف برخی نماد مورد استفاده در مقاله می‌پردازیم:

$O_{ij}(t)$ معرف تابع نشانگر در لحظه t برای گروه j است. از این رو $N_j(t) = \sum_{i=1}^{n} O_{ij}(t)$ مجموع مشاهدات در دسترس برای $j = 1, 2$ است که در آن اندیس j نشانگر تعداد گروه‌ها است و برای هر گروه مشاهدات تابعی،می‌توانند متفاوت باشند. اگر $N_j(t) = 0$، آنگاه $J_j(t) = I_{[N_j(t))>0]} = 0$، به این معنا است که مقدار مشاهده شده‌ای از تابع برای گروه j ام در لحظه‌ی t وجود ندارد.

$$\begin{cases} \hat{\rho}(s,t) = \dfrac{I(s,t)}{M(s,t)} \displaystyle\sum_{i=1}^{n} U_i(s,t)\{X_i(s) - \hat{\mu}_{st}(s)\}\{X_i(t) - \hat{\mu}_{st}(t)\}, \\ I(s,t) = I_{[M(s,t)>0]}, \\ \hat{\mu}_{st}(s) = \dfrac{I_{[M(s,t)>0]}}{M(s,t)} \displaystyle\sum_{i=1}^{n} U_i(s,t) X_i(s). \end{cases}$$

$U_i(s,t) \equiv O_i(s)O_i(t)$ معرف تابع نشانگر است و به این معناست که در دو لحظه‌ی معین s و t مقدار تابع به‌صورت توأم در دسترس باشد ، از این رو $M(s,t) \equiv \sum_{i=1}^{n} U_i(s,t)$ معرف مجموع مشاهدات توام در هر گروه است و در صورتی که $I_{[M(s,t)>0]} = 0$ (به این معنا که هیچ مقادیری از تابع در نقاط جفتی مشاهده نشود) ، برآوردگر $\hat{\mu}_{st} = 0$ در نظر گرفته می‌شود، در ادامه برآورد میانگین گروهی و میانگین کل به‌صورت زیر است:

$$\sum_{j=1}^{2} \hat{\omega}_j(t) = 1, \quad N(t) = \sum_{i=1}^{n} O_i(t), \quad \hat{\mu}(t) = \sum_{j=1}^{2} \hat{\omega}_j(t)\hat{\mu}_j(t).$$

که در آن برآورد میانگین کل به صورت ترکیب خطی از برآورد میانگین‌های گروهی است و هر گروه با ضریب تاثیر $\hat{\omega}_i(t)$ در میانگین کل حضور دارند بعلاوه میانگین برای هر گروه و تابع وزن برای j–امین گروه به صورت زیر است:

$$\hat{\mu}_j(t) = J_j(t)N_j(t)^{-1} \sum_{i=1}^{n_j} O_{ji}(t)X_{ji}(t), \quad j \in \{1, 2\}, \quad \hat{\omega}_j(t) = \frac{\dfrac{N_j(t)}{\hat{r}_j^2}}{\displaystyle\sum_{k=1}^{K} \dfrac{N_k(t)}{\hat{r}_k^2}}.$$

بعلاوه احتمال این که در لحظه‌ی t مقدار تابع مشاهده شود و نیز احتمال این که در دو لحظه معین به‌صورت توام مقادیر تابع مشاهده شود به‌صورت زیر تعریف می‌شوند:

$$\pi_i(t) \equiv EO_i(t) = \Pr\{O_i(t) = 1\}, \quad \nu_i(s,t) = E[U_i(s,t)] = P[O_i(t) = 1, O_i(s) = 1].$$

و ضریب استاندارد ساز در $\hat{\omega}_j(t)$ نیز به‌صورت $\hat{r}_j = \int_0^1 \hat{\rho}_j(t,t)\, dt$ تعریف می‌شود.

نخست تعریف می‌کنیم:

$$SSR_n(t) = \sum_{i=1}^{2} n_i \left(\bar{y}_{i\cdot}(t) - \bar{y}_{\cdot\cdot}(t)\right)^2,$$

که همان تغییرات بین گروهی است که در تحلیل واریانس مورد استفاده قرار می‌گیرد. مشابه تعریف فوق مولفه‌ی تصادفی $Z_{n_j}(t)$ نیز مشابه جملات مجموع بالا و با تعدیل آن توسط دو عامل ذکر شده برای داده‌های تابعی جزئی مشاهده‌شده به‌صورت زیر ارائه می‌شود:

$$Z_{n_j}(t) = \hat{r}_j^{-1} N_j(t)^{1/2} \left[\hat{\mu}_j(t) - \hat{\mu}(t)\right].$$

اکنون برای سهولت در اعمال دو مولفه‌ی مهم یعنی، ساختار گمشدگی نامتقارن و تابع کوواریانس متمایز و دوری از پیچیدگی های غیر ضروری با تعریف یک عملگر قطری می‌توان $Z_{n_j}(t)$ را به‌صورت ضرب نقطه‌ای از عناصر $f : \mathcal{L}^2 \to \mathcal{L}^2$ ایجاد کرد که در آن برآورد عملگر قطری یعنی $\hat{\psi}_{jl}$ بر روی هر عضو دلخواه f_l متعلق به فضای L^2 به صورت زیر عمل می‌کند:

$$(\hat{\psi}_{jl} f_l)(t) = \hat{r}_j^{-1} \left\{\delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) J_l(t) N_l(t)^{-1/2}\right\} f_l(t)$$

بعلاوه تعریف می‌کنیم:

$$H_j(t) = N_j(t)^{1/2} [\hat{\mu}_j(t) - \mu(t)], \quad j = 1, 2$$

از این رو می‌توان $Z_{n_j}(t)$ را برای هر جامعه به‌صورت زیر بازنویسی کرد،

$$Z_{n_j}(t) = \hat{\psi}_{j1} H_1(t) + \hat{\psi}_{j2} H_2(t), \quad j = 1, 2 \tag{1.2}$$

اکنون عناصر $Z_{n_j}(t)$ به صورت زیرتعریف می‌شوند:

$$Z_{n_j}(t) = \hat{r}_j^{-1} N_j(t)^{1/2} \left[\hat{\mu}_j(t) - \hat{\mu}(t)\right] = \sum_{i=1}^{2} \hat{\psi}_{ji} H_i(t),$$

از این رو $Z_{n_j}(t)$ با اعمال عملگر قطری $\hat{\psi}_{jl}$ روی $H_l(t)$ به صورت زیر نمایش داده می‌شود:

$$(\hat{\psi}_{jl} H_l)(t) = \hat{r}_j^{-1} \left\{\delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) J_l(t) N_l(t)^{-1/2}\right\} N_l(t)^{1/2} [\hat{\mu}_l(t) - \mu(t)], \quad l = 1, 2$$

برای حالتی که $J_j(t) = \mathbb{I}_{\{N_j(t) > 0\}} = 1$، به این معنا که در هر گروه مشاهده داشته باشیم، $Z_{n_j}(t)$ به‌صورت زیر ساده می‌شود:

$$Z_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{\delta_{jl} N_l(t)^{1/2} - N_j(t)^{1/2} \hat{\omega}_l(t)\right\} [\hat{\mu}_l(t) - \mu(t)],$$

اما با استفاده از تعریف دلتای کرونکر، داریم:

$$\sum_{k=1}^{2} \hat{r}_j^{-1} \delta_{jl} N_l(t)^{1/2} [\hat{\mu}_l(t) - \mu(t)] = \hat{r}_j^{-1} N_j(t)^{1/2} [\hat{\mu}_j(t) - \mu(t)],$$

در ادامه مولفه‌ی $Z_{n_j}(t)$ به صورت مجموع دو مولفه‌ی تصادفی تجزیه می‌شود و به همگرایی هر دو جز به صورت کلی اشاره می‌شود برای این منظور دو بخش تصادفی را بر اساس انتقال میانگین‌ها بازنویسی کرده و به همگرایی دو جز تصادفی به صورت کلی اشاره می‌کنیم.

با فرض این که میانگین‌های هر گروه به‌صورت $\mu_j(t) = \mu(t) + h_j(t)$ نمایش داده شوند، $Z_{n_j}(t)$ بر اساس این انتقال به اندازه‌ی هر تابع حقیقی مقداری مانند $h_j(t)$ به‌صورت زیر بازنویسی می‌شود:

$$Z_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{ \delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) N_l(t)^{-1/2} \right\} N_l(t)^{1/2} \left[h_l(t) + (\hat{\mu}_l(t) - \mu_l(t)) \right]. \tag{2.2}$$

همگرایی مولفه‌های تجزیه شده $Z_{n_j}(t)$

با توجه به تعریف $Z_{n_j}(t)$ در رابطه (۲.۲) می‌توان نوشت: $Z_{n_j}(t) = \Delta_{n_j}(t) + S_{n_j}(t)$ که در آن

$$\Delta_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{ \delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) N_l(t)^{-1/2} \right\} N_l(t)^{1/2} h_l(t), \tag{3.2}$$

$$S_{n_j}(t) = \sum_{l=1}^{2} \hat{r}_j^{-1} \left\{ \delta_{jl} - N_j(t)^{1/2} \hat{\omega}_l(t) N_l(t)^{-1/2} \right\} N_l(t)^{1/2} \left[\hat{\mu}_l(t) - \mu_l(t) \right], \tag{4.2}$$

قضیه ۱.۲. با قرار دادن $h_j(t) = \frac{d_j(t)}{n_j^{1/2}}, \quad j = 1, 2$ که در آن $d_j(t)$ هر تابع حقیقی مقدار است، می‌توان نشان داد که همگرایی در توزیع برای $\Delta_{n_j}(t)$ و $S_{n_j}(t)$ به صورت زیر نتیجه می شود:

$$\Delta_{n_j}(t) = \hat{r}_j^{-1} N_j(t)^{1/2} \left\{ n_j^{-1/2} d_j(t) - \sum_{l=1}^{2} \hat{\omega}_l(t) n_l^{-1/2} d_l(t) \right\} \xrightarrow{d} r_j^{-1} \pi_j(t)^{1/2} \left\{ d_j(t) - a_j^{1/2} \sum_{l=1}^{2} \omega_l(t) a_l^{-1/2} d_l(t) \right\}, \tag{5.2}$$

$$S_{n_j}(t) \xrightarrow{d} r_j^{-1} \left[S_j^{\infty}(t) - \sum_{l=1}^{2} \omega_l(t) \left[\frac{\pi_j(t) a_j}{\pi_l(t) a_l} \right]^{1/2} S_l^{\infty}(t) \right], \tag{6.2}$$

که در آن

$$a_j = \lim_{n \to \infty} \frac{n_j}{n}.$$

نسبت حدی اندازه نمونه هر گروه به اندازه نمونه کل گروه‌ها است، از این رو میانگین مجانبی فرآیند گاوسی با انتقال میانگین همان توزیع حدی بخش غیر تصادفی در (۵.۲) است و تابع کوواریانس آن مرتبط با بخش تصادفی در (۶.۲) است. که در آن $S_j^{\infty}(t) \stackrel{d}{=} \mathcal{GP}(0, \mathcal{K}_j)$ به صورت فرآیند گاوسی با میانگین صفر و هسته عملگر کوواریانس زیر است:

$$\kappa_j(s,t) = \left[\pi_j(t) \right]^{-1/2} \left[\pi_j(s) \right]^{-1/2} \nu_j(s,t) \rho_j(s,t).$$

در ادامه هسته مربوط به توزیع مجانبی عملگر $Z_{n_j}(t)$ به‌صورت زیر ارائه می‌شود:

$$
\begin{aligned}
\nu_{jk}^{\infty}(s,t) &= \operatorname{cov}\left[Z_j^{\infty}(t), Z_k^{\infty}(s)\right] \\
&= \sum_{l=۱}^{۲} \psi_{jl}(s)\kappa_l(s,t)\psi_{kl}^{*}(t) \\
&= \sum_{l=۱}^{۲} r_j^{-۱}\left\{\delta_{jl} - \pi_j(s)^{۱/۲} a_j^{۱/۲}\omega_l(s)\pi_l(s)^{-۱/۲} a_l^{-۱/۲}\right\} \\
&\quad \times \kappa_l(s,t)\left\{\delta_{kl} - \pi_k(t)^{۱/۲} a_k^{۱/۲}\omega_l(t)\pi_l(t)^{-۱/۲} a_l^{-۱/۲}\right\} r_k^{-۱},
\end{aligned}
$$

که در آن $\|\hat{\psi} - \psi\|_{\infty} \xrightarrow{p} ۰$ ، $\psi_{kl}^{*}(t)$ به‌صورت عملگر الحاقی تعریف می‌شود، از این رو توزیع مجانبی $Z_{n_j}(t)$ به صورت زیر است:

$$
Z_{n_j}(t) \xrightarrow{d} \mathcal{GP}\left(\Delta_j^{\infty}, \nu_{jj}^{\infty}\right),
$$

که یک فرآیند گاوسی با میانگین مجانبی زیر است:

$$
\Delta_j^{\infty}(t) = r_j^{-۱}\pi_j(t)^{۱/۲}\left\{d_j(t) - a_j^{۱/۲}\sum_{l=۱}^{۲}\omega_l(t)a_l^{-۱/۲}d_l(t)\right\}.
$$

قضیه (۱.۲) به‌عنوان یک نتیجه مهم در تقریب توزیع آماره‌ی آزمون مورد استفاده قرار می گیرد.

مراجع

[1] Bugni, F. A., Smith, J., and Lee, K. (2012). Specification test for missing functional data, *Econometric Theory*, 28(5), 959-1002.

[2] Gromenko, O., Kokoszka, P., and Sojka, J. (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves, *Ann. Appl. Stat*, 11(2): 898-918.

[3] Liebl, D., Rameseder, S. (2019). Partially observed functional data: The case of systematically missing parts, *Computational Statistics Data Analysis*, 131, 104-115.

[4] Stefanucci, M., Sangalli, L. M., and Brutti, P. (2018). PCA-based discrimination of partially observed functional data, with an application to AneuRisk65 data set, *Statistica Neerlandica*, 72(3), 246-264.

[5] Mojirsheibani, M., Shaw, C. (2018). Classification with incomplete functional covariates. *Probability Letters*, 139, 40-46.

[6] Delaigle, A., Hall, P. (2016). Approximating fragmented functional data by segments of Markov chains, *Biometrika*,103(4), 779-799.

[7] Qiu, Z., Chen, J., and Zhang, J. T. (2021). Two-sample tests for multivariate functional data with applications, *Computational Statistics Data Analysis* , 157, 107160.

[8] Kraus, D. (2015). Components and completion of partially observed functional data, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4), 777-801.

[9] Kneip, A., Liebl, D. (2020). On the optimal reconstruction of partially observed functional data, *The Annals of Statistics*, 1692-1717.

[10] Descary, M. H., Panaretos, V. M. (2019). Recovering covariance from functional fragments, *Biometrika* , 106(1), 145-160.

[11] *Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.*

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مدل خودبازگشتی صحیح مقدار مرتبه اول فصلی با توزیع حاشیه‌ای دلاپورت

مریم شالباف [۱] ، غلامعلی پرهام [۲]

[۱]گروه آمار و ریاضی، واحد شوشتر، دانشگاه آزاد اسلامی، شوشتر، ایران

[۲]گروه آمار، دانشکده علوم ریاضی و کامپیوتر، دانشگاه شهید چمران اهواز

چکیده: در سری‌های زمانی شمارشی اغلب پدیده بیش پراکنشی مشاهده می‌شود. در این مقاله، یک فرایند خودبازگشتی صحیح مقدار مرتبه اول با ساختار فصلی معرفی می‌شود. توزیع حاشیه‌ای مدل توزیع دلاپورت می‌باشد. بر این اساس توزیع نوآوری‌ها پیچشی از توزیع پواسون با تعداد α تا توزیع هندسی تعدیل‌یافته در صفر خواهد شد. برخی از خواص مدل نمایش داده می‌شود. برای برآورد پارامترها از روش‌های یول–والکر، حداقل مربعات شرطی و حداکثر درستنمایی شرطی استفاده می‌شود. با انجام یک شبیه‌سازی عملکرد این برآوردگرها مورد ارزیابی قرار می‌گیرد. در نهایت این مدل به یک مجموعه داده واقعی برازش داده می‌شود.

واژه‌های کلیدی: توزیع دلاپورت، مدل خودبازگشتی صحیح مقدار مرتبه اول فصلی، α تا توزیع هندسی تعدیل‌یافته، عملگر رقیق‌کننده دوجمله‌ای

---

۱ مقدمه

داده‌های سری زمانی با ویژگی‌های فصلی را می‌توان در زمینه‌های مختلفی مانند علوم آماری، بهداشت، اقتصاد، محیط زیست و غیره یافت. آن‌ها عمدتاً یک الگوی فصلی با دوره‌هایی را نشان می‌دهند که پس از یک بازه زمانی منظم تکرار می‌شوند. کوچکترین دوره زمانی برای این رویداد، دوره فصلی نامیده می‌شود. عوامل متعددی مانند آب‌وهوا و ویژگی‌های ذاتی می‌توانند باعث ایجاد ساختارهای فصلی شوند. مدل‌های سری زمانی گسسته مقدار (INAR) معمولاً براساس عملگر رقیق‌کننده دوجمله‌ای که توسط استیوتل و ون هارن (۱۹۷۹) به‌صورت $\rho \circ X = \sum_{i=1}^{X} Y_i$ تعریف شده است، ساخته می‌شوند، به‌طوری که $\{Y_t\}_{t\in\mathbb{Z}}$ یک دنباله از متغیرهای تصادفی مستقل و هم‌توزیع برنولی با احتمال موفقیت $\rho \in [0,1]$ می‌باشد و X یک متغیر تصادفی صحیح مقدار غیرمنفی و مستقل از Y می‌باشد. ملکنزی (۱۹۸۵) و آلاوش و آلزاید (۱۹۸۷) مدل INAR(1) را به‌صورت زیر معرفی کردند:

$$X_t = \rho \circ X_{t-1} + \varepsilon_t \qquad (1.1)$$

---

[۱] سخنران، Maryamshal@iau.ac.ir

که در آن $0 \leq \rho < 1$ و ε_t دنباله‌ای از متغیرهای تصادفی با مقادیر صحیح غیرمنفی ناهمبسته با میانگین μ_ε و واریانس متناهی σ_ε^2 است. اگر $\varphi_X(s)$ و $\varphi_\varepsilon(s)$ به ترتیب نشان‌دهنده تابع مولد احتمال $\{X_t\}_{t\in\mathbb{Z}}$ و $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ باشند، آنگاه توزیع حاشیه‌ای $\{X_t\}_{t\in\mathbb{Z}}$ را می‌توان به‌صورت $\varphi_X(s) = \varphi_X(1-\rho+\rho s)\varphi_\varepsilon(s)$ نوشت.

متداول‌ترین فرایند برای مدل‌بندی این نوع از داده‌ها، مدل پواسون خودبازگشتی گسسته مقدار مرتبه اول ((PINAR(1)) است بیش پراکنشی (یعنی حالتی که واریانس از میانگین بزرگتر است) پدیده‌ای است که معمولا در سری‌های شمارشی مشاهده می‌شود. چون توزیع پواسون هم پراکنش است (یعنی میانگین با واریانس برابر است) فرایند PINAR برای مدل‌سازی داده‌های با بیش پراکنشی مناسب نیست. یک رویکرد معمول برای برخورد با بیش پراکنشی در داده‌های شمارشی استفاده از توزیع‌های پواسون آمیخته است که توسط بارتو–سوآزا (۲۰۱۹) بررسی شده است. توزیع‌های پواسون آمیخته بوسیله معرفی یک اثر تصادفی پنهان روی میانگین توزیع پواسون به‌دست می‌آیند. مدل خود بازگشتی صحیح مقدار با توزیع حاشیه‌ای دلاپوریت توسط شالباف و همکاران (۲۰۲۲) بررسی شده است. در سال‌های اخیر بر روی مدل‌های خود بازگشتی صحیح مقدار مطالعات زیادی صورت گرفته است ولی فرایندهای فصلی با توزیع حاشیه‌ای گسسته به‌ندرت بررسی و معرفی شده‌اند. بورگینکون و همکاران (۲۰۱۶) فصلی بودن را براساس مدل PINAR(1) بررسی کرده‌اند. فرایند $INAR(1)_s$ فصلی مرتبه اول با دوره فصلی s ساختار ساده‌ای دارد و برای برآورد کردن پارامترهای کمتری دارد اما برای مدل‌سازی سری‌های زمانی فصلی با بیش پراکنشی مناسب نیست. تیان و همکاران (۲۰۱۸) یک فرایند INAR(1) فصلی را براساس عملگر رقیق‌کننده دوجمله‌ای منفی و با توزیع حاشیه‌ای هندسی معرفی کردند. در این مقاله ما در جهت کمک به تحلیل سری‌های زمانی گسسته مقدار، یک فرایند خودبازگشتی صحیح مقدار مرتبه اول فصلی با توزیع حاشیه‌ای دلاپورت $(DELINAR(1)_s)$ را معرفی می‌کنیم. این مدل برای مدل‌سازی سری‌های زمانی فصلی با مقدار صحیح غیرمنفی و با بیش پراکنشی مناسب است. بر این اساس مدل در بخش ۲ معرفی شده است. همچنین برخی از ویژگی‌های مدل بیان شده است. در بخش ۳، روش‌های برآورد پارامترهای مدل مورد بحث قرار می‌گیرد. در بخش ۴ برخی از نتایج پیش‌بینی ارائه شده است. بخش ۵ برخی از نتایج شبیه‌سازی را برای روش‌های برآورد نشان می‌دهد. در بخش ۶، مدل بر روی یک مجموعه داده واقعی اعمال می‌شود. در نهایت در بخش ۷ نتایج بیان شده است.

۲ فرایند INAR(1) فصلی با توزیع حاشیه‌ای دلاپورت

در این بخش، ویژگی‌های ساختاری این فرایند، مانند توزیع‌های حاشیه‌ای و نوآوری، میانگین و واریانس این توزیع‌ها، تابع خودکوواریانس، امید شرطی و واریانس شرطی متغیر تصادفی حاشیه‌ای و احتمالات انتقال را بررسی می‌کنیم.

۱.۲ توزیع دلاپورت

متغیر تصادفی X دارای توزیع دلاپورت با پارامترهای λ، α و β است و آن را با نماد $DEL(\lambda,\alpha,\beta)$ نشان می‌دهند اگر تابع مولد احتمال آن به‌صورت $G_X(S) = e^{-\lambda(1-s)}\left(\frac{1}{1+\beta(1-s)}\right)^\alpha$ باشد، به‌طوری که $|s| \leq 1$، $\lambda > 0$ و $\alpha,\beta > 0$ (دلاپورت، ۱۹۵۹). با دوبار مشتق گرفتن از تابع مولد احتمال توزیع دلاپورت می‌توان نشان داد که:

$$\mu = E(X_t) = \lambda + \alpha\beta \quad , \quad \sigma^2 = Var(X_t) = \lambda + \alpha\beta(1+\beta)$$

با توجه به اینکه واریانس بزرگ‌تر از میانگین است، می‌توان گفت این توزیع بیش پراکنشی دارد.

۲.۲ فرایند خودبازگشتی صحیح مقدار مرتبه اول فصلی با توزیع حاشیه‌ای دلاپورت $DELINAR(1)_s$

فرض کنید N_0 نشان‌دهنده مجموعه مقادیر صحیح غیرمنفی باشد. همچنین فرض کنید X یک متغیر تصادفی با مقادیر صحیح غیرمنفی و $\rho \in [0, 1]$ باشد. با استفاده از عملگر رقیق‌کننده دوجمله‌ای، مدل دلاپوریت فصلی مرتبه اول به‌صورت

$$X_t = \rho \circ X_{t-s} + \varepsilon_t \tag{1.2}$$

تعریف می‌شود، به‌طوری که $\rho \in [0, 1]$ است و $\{\varepsilon_t\}$ دنباله نوآوری از متغیرهای تصادفی با مقادیر صحیح غیرمنفی و مستقل از مقادیر گذشته $\{X_t\}$ می‌باشند و $s \in \mathbb{N}$ دوره فصلی را نشان می‌دهد. متغیرهای برنولی در عملگرهای رقیق‌کننده مستقل از هم و مستقل از نوآوری $\{\varepsilon_t\}$ می‌باشند. این مدل را با نماد $DELINAR(1)_s$ نشان می‌دهیم.

قضیه ۱.۲. متغیر تصادفی ε_t را می‌توان به صورت $\varepsilon_t = Y_1 + Y_2$ نشان داد که در آن Y_1 دارای توزیع پواسون با پارامتر $\lambda(1-\rho)$ و Y_2 پیچش α تا توزیع هندسی تعدیل‌یافته در صفر است. بنابراین امید ریاضی و واریانس ε_t به‌صورت $\mu_\varepsilon := E(\varepsilon_t) = (1-\rho)\mu_X$ و $\sigma_\varepsilon^2 := Var(\varepsilon_t) = (1-\rho^2)\sigma_X^2 - \rho(1-\rho)\mu_X$ می‌باشند. همچنین احتمال شرطی X_t بشرط X_{t-1} به‌صورت زیر به‌دست می‌آید.

$$p_{ij} = P(X_t = j | X_{t-1} = i) = P(\rho \circ X_{t-1} + \varepsilon_t = x_t | X_{t-1} = x_{t-1}).$$

اکنون فرض کنید

$$B^{\rho}_{X_{t-1}} := \rho \circ X_{t-1} | X_{t-1} \sim Binomial(X_{t-1}, \rho)$$

همچنین ε_t مستقل از X_{t-1} می‌باشد و دارای ساختار پیچشی از دو متغیر تصادفی $P_{(\lambda,\rho)} \sim Po(\lambda(1-\rho))$ و پیچش α تا توزیع هندسی تعدیل‌یافته در صفر با پارامترهای $p = \frac{1}{1+\beta}$ و p_0 می‌باشد، که $p_0 = \frac{(1+\rho\beta)}{1+\beta}$ میزان جرم احتمال در صفر می‌باشد. بنابراین احتمال شرطی را می‌توان به‌صورت زیر نوشت:

$$\begin{aligned} P(X_t = x_t | X_{t-1} = x_{t-1}) &= P(B^{\rho}_{X_{t-1}} + P_{(\lambda,\rho)} + \alpha FZMG(p, p_0) = x_t) \\ &= \sum_{s=0}^{x_t} P(\alpha FZMG = s) \sum_{d=0}^{\min\{x_t - s, x_{t-1}\}} P(B^{\rho}_{X_{t-1}} = d) P(P_{(\lambda,\rho)} = x_t - s - d), \end{aligned} \tag{2.2}$$

۳ برآورد پارامترها

فرض کنید $X_1, X_2, \ldots, X_n$ یک نمونه از فرایند $DELINAR(1)_s$ باشند. در این فرایند ما ۴ پارامتر داریم، فرض می‌کنیم پارامتر α معلوم است پس ۳ پارامتر دیگر باید برآورد شوند. در ادامه روش‌های یول والکر، حداقل مربعات شرطی و حداکثر درستنمایی شرطی و برخی از نتایج تحلیلی و مجانبی برآوردگرها را معرفی و توضیح می‌دهیم.

۱.۳ برآورد یول والکر

فرض کنید $\gamma(k)=\sum_{t=2}^{n-k}(X_t-\bar{X})(X_{t+k}-\bar{X})$ و $0\leq k<n$ تابع خودکواریانس نمونه‌ای باشد. چون $\rho=\frac{\gamma(1)}{\gamma(0)}$ و $E(X_t)=\lambda+\alpha\beta$ و $Var(X_t)=\lambda+\alpha\beta(1+\beta)$، برآوردهای یول والکر پارامترها به‌صورت زیر می‌باشد:

$$\hat{\rho}_{YW}=\frac{\sum_{t=s+1}^{n}(X_t-\bar{X})(X_{t-s}-\bar{X})}{\sum_{t=1}^{n}(X_t-\bar{X})^2}\quad,\quad \hat{\beta}_{YW}=\sqrt{\frac{S^2-\bar{X}}{\alpha}}\quad,\quad \hat{\lambda}_{YW}=\bar{X}-\alpha\sqrt{\frac{S^2-\bar{X}}{\alpha}},$$

به‌طوری که $\bar{X}$ و S^2 بترتیب نشان‌دهنده میانگین نمونه‌ای و واریانس نمونه‌ای می‌باشد.

۲.۳ برآورد حداقل مربعات شرطی

روش حداقل مربعات شرطی را در دو مرحله اجرا می‌کنیم (کلیمکو و نلست، ۱۹۸۷). در گام اول با نوشتن معادله حداقل مربعات شرطی برای پارامترهای ρ و μ و به کمک مشتق گیری جزئی نسبت به پارامترهای ρ و μ از معادله و کمینه کردن آن داریم:

$$\hat{\mu}_{cls}=\frac{\sum_{t=s+1}^{n}X_t-\hat{\rho}_{cls}\sum_{t=s+1}^{n}X_{t-s}}{(n-s)(1-\hat{\rho}_{cls})},\quad \hat{\rho}_{cls}=\frac{(n-s)\sum_{t=s+1}^{n}X_tX_{t-s}-\sum_{t=s+1}^{n}X_t\sum_{t=s+1}^{n}X_{t-s}}{(n-s)\sum_{t=s+1}^{n}X_{t-s}^2-(\sum_{t=s+1}^{n}X_{t-s})^2}.$$

در گام دوم، با نوشتن معادله حداقل مربعات شرطی برای پارامتر σ^2 و کمینه کردن آن نسبت به σ^2 داریم:

$$\hat{\sigma}_{cls}^2=\frac{\sum_{t=s+1}^{n}(X_t-\hat{\rho}_{cls}X_{t-s}-(1-\hat{\rho}_{cls})\hat{\mu}_{cls})^2-\hat{\rho}_{cls}(1-\hat{\rho}_{cls})\sum_{t=s+1}^{n}(X_{t-s}-\hat{\mu}_{cls})}{(n-s)(1-\hat{\rho}_{cls}^2)}$$

سرانجام با حل دستگاه و فرض معلوم بودن α، برآورد کمترین مربعات شرطی برای پارامترهای λ و β به‌صورت زیر به‌دست می‌آید:

$$\hat{\lambda}_{cls}=\hat{\mu}_{cls}-\alpha\hat{\beta}_{cls}\qquad,\qquad \hat{\beta}_{cls}=\sqrt{\frac{\hat{\sigma}_{cls}^2-\hat{\mu}_{cls}}{\alpha}}.$$

۳.۳ برآورد حداکثر درستنمایی شرطی

برآوردگرهای حداکثر درستنمایی شرطی با بیشینه کردن تابع لگاریتم درستنمایی به‌دست می‌آیند. تابع لگاریتم درستنمایی شرطی به‌صورت $CL(\rho,\lambda,\beta)=\sum_{t=s+1}^{n}\log P(X_t=j|X_{t-s}=i)$ می‌باشد، به‌طوری که $P(X_t=j|X_{t-s}=i)$ در (۲.۲) تعریف شده است. مقادیر برآورد حداکثر درستنمایی شرطی با بیشینه کردن تابع فوق بدست می‌آید. چون $CL(\rho,\lambda,\beta)$ یک تابع غیرخطی است، برآوردهای حداکثر درستنمایی شرطی باید به کمک روش‌های عددی محاسبه شوند.

۴ پیش‌بینی

در اینجا هدف، محاسبه پیش‌بینی مقدار $X_{n+h}(h\in\mathbb{N})$ براساس اطلاعات تا زمان n ام می‌باشد. توزیع X_{n+h} براساس تعریف مدل $DELINAR(1)_s$ را می‌توان به‌صورت زیر بیان کرد:

$$X_{n+h}\overset{d}{=}\rho^q\circ X_{n+h-qs}+\sum_{j=0}^{q-1}\rho^j\circ\varepsilon_{n+h-js}$$

به‌طوری که $q=\lceil h/s\rceil$ و $\lceil x\rceil=\min\{n\in\mathbb{N}|x\leq n\}$. در واقع $\lceil x\rceil$ نشان‌دهنده کوچک‌ترین عدد طبیعی بزرگ‌تر (یا مساوی) x می‌باشد. از عبارت بالا مشاهده می‌شود که توزیع پیش‌بینی h گام جلوتر شکل بسیار پیچیده‌ای دارد و بنابراین ما پیش‌بینی h گام

جلوتر را به کمک امید شرطی زیر بدست می‌آوریم:

$$\hat{X}_{n+h}|X_n = E(X_{n+h}|X_n) = \rho^q(X_{n+h-qs} - \mu_x) + \mu_x.$$

به‌طوری که $\mu_x = \lambda + \alpha\beta$.

۵ شبیه‌سازی

در این قسمت عملکرد برآوردهای یول والکر، حداقل مربعات شرطی و حداکثر درستنمایی شرطی برای پارامترهای مدل برای اندازه‌های نمونه مختلف با دوره فصلی ۱۲ در جداول ۱ تا ۳ ارائه شده است. اریبی‌ها و میانگین توان دوم خطاها ۱۰۰۰ بار تکرار و محاسبه شده‌اند. این جداول نشان می‌دهد که اریبی و خطای استاندارد برآورد پارامترها با افزایش حجم نمونه برای همه موارد کاهش می‌یابد. همان‌طور که از جداول مشاهده می‌شود، روش حداقل مربعات شرطی و روش یول والکر، رفتارهای مشابهی را نشان می‌دهند. با توجه به جداول، برآوردگرهای حداکثر درستنمایی شرطی مقادیر کم‌تری از اریبی (مقادیر قدرمطلق) و میانگین مربعات خطا در مقایسه با برآوردگرهای یول والکر و حداقل مربعات شرطی را نشان می‌دهند.

جدول ۱: میزان اریبی و میانگین مربعات خطا (در داخل پرانتز) برآوردگرها برای $(\rho, \beta, \lambda) = (۰/۲, ۳, ۱/۵)$

$\hat{\lambda}_{CML}$	$\hat{\lambda}_{cls}$	$\hat{\lambda}_{YW}$	$\hat{\beta}_{CML}$	$\hat{\beta}_{cls}$	$\hat{\beta}_{YW}$	$\hat{\rho}_{CML}$	$\hat{\rho}_{cls}$	$\hat{\rho}_{YW}$	n
۰/۰۱۷۶	۰/۰۴۹۲	۰/۰۰۷۵	−۰/۰۱۵۹	−۰/۰۳۲۰	−۰/۰۱۰۶	۰/۰۰۰۸	−۰/۰۰۲۳	−۰/۰۱۴۲	۲۰۰
(۰/۱۷۰۸)	(۰/۲۸۰۴)	(۰/۲۶۸۵)	(۰/۰۷۶۶)	(۰/۱۰۱۳)	(۰/۰۹۵۶)	(۰/۰۰۲۷)	(۰/۰۰۵۰)	(۰/۰۰۴۷)	
۰/۰۰۶۰	۰/۰۱۵۵	−۰/۰۰۰۳	−۰/۰۰۸۲	−۰/۰۱۳۱	−۰/۰۰۵۱	۰/۰۰۰۰	−۰/۰۰۱۳	−۰/۰۰۷۴	۴۰۰
(۰/۰۸۷۰)	(۰/۱۴۲۴)	(۰/۱۳۴۴)	(۰/۰۳۷۴)	(۰/۰۴۹۸)	(۰/۰۴۷۴)	(۰/۰۰۱۳)	(۰/۰۰۲۶)	(۰/۰۰۲۵)	
۰/۰۱۱۹	۰/۰۱۹۶	۰/۰۰۹۵	−۰/۰۰۶۷	−۰/۰۱۰۵	−۰/۰۰۵۷	−۰/۰۰۱۶	−۰/۰۰۲۹	−۰/۰۰۵۹	۸۰۰
(۰/۰۴۶۵)	(۰/۰۷۷۴)	(۰/۰۷۵۹)	(۰/۰۱۸۹)	(۰/۰۲۶۸)	(۰/۰۲۶۲)	(۰/۰۰۰۷)	(۰/۰۰۱۳)	(۰/۰۰۱۳)	

جدول ۲: میزان اریبی و میانگین مربعات خطا (در داخل پرانتز) برآوردگرها برای $(\rho, \beta, \lambda) = (۰/۵, ۲, ۱)$

$\hat{\lambda}_{CML}$	$\hat{\lambda}_{cls}$	$\hat{\lambda}_{YW}$	$\hat{\beta}_{CML}$	$\hat{\beta}_{cls}$	$\hat{\beta}_{YW}$	$\hat{\rho}_{CML}$	$\hat{\rho}_{cls}$	$\hat{\rho}_{YW}$	n
۰/۰۲۷۲	۰/۰۹۵۳	۰/۰۴۶۰	−۰/۰۱۱۸	−۰/۰۴۶۱	−۰/۰۲۳۵	−۰/۰۰۲۹	−۰/۰۰۸۶	−۰/۰۳۸۵	۲۰۰
(۰/۱۵۵۵)	(۰/۲۷۷۹)	(۰/۲۴۵۹)	(۰/۰۶۶۰)	(۰/۰۹۰۸)	(۰/۰۸۰۸)	(۰/۰۰۱۹)	(۰/۰۰۴۴)	(۰/۰۰۵۴)	
۰/۰۲۵۴	۰/۰۶۴۵	۰/۰۴۵۳	−۰/۰۱۶۳	−۰/۰۳۶۷	−۰/۰۲۷۸	−۰/۰۰۳۴	−۰/۰۰۸۱	−۰/۰۲۳۱	۴۰۰
(۰/۰۷۲۵)	(۰/۱۳۵۳)	(۰/۱۲۷۸)	(۰/۰۳۱۴)	(۰/۰۴۵۱)	(۰/۰۴۲۶)	(۰/۰۰۰۸)	(۰/۰۰۲۲)	(۰/۰۰۲۶)	
۰/۰۰۱۷	۰/۰۲۳۳	۰/۰۱۴۲	−۰/۰۰۶۸	−۰/۰۱۷۷	−۰/۰۱۳۰	−۰/۰۰۱۵	−۰/۰۰۳۷	−۰/۰۱۰۹	۸۰۰
(۰/۰۳۳۲)	(۰/۰۶۰۸)	(۰/۰۵۹۶)	(۰/۰۱۴۷)	(۰/۰۲۱۴)	(۰/۰۲۰۹)	(۰/۰۰۰۴)	(۰/۰۰۱۱)	(۰/۰۰۱۲)	

جدول ۳: میزان اریبی و میانگین مربعات خطا (در داخل پرانتز) برآوردگرها برای $(\rho, \beta, \lambda) = (۰/۸, ۱/۵, ۴)$

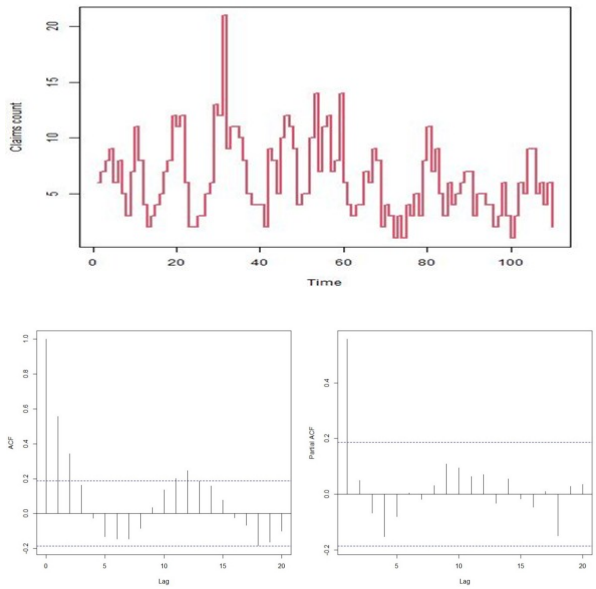
$\hat{\lambda}_{CML}$	$\hat{\lambda}_{cls}$	$\hat{\lambda}_{YW}$	$\hat{\beta}_{CML}$	$\hat{\beta}_{cls}$	$\hat{\beta}_{YW}$	$\hat{\rho}_{CML}$	$\hat{\rho}_{cls}$	$\hat{\rho}_{YW}$	n
۰/۰۴۶۱	۰/۲۶۰۲	۰/۱۵۳۲	−۰/۰۱۵۹	−۰/۱۲۴۱	−۰/۰۶۵۴	−۰/۰۰۳۰	−۰/۰۱۰۳	−۰/۰۵۸۵	۲۰۰
(۰/۴۴۸۸)	(۱/۲۰۰۵)	(۰/۸۱۵۱)	(۰/۱۰۸۸)	(۰/۲۶۷۶)	(۰/۱۹۳۶)	(۰/۰۰۰۵)	(۰/۰۰۲۲)	(۰/۰۰۵۷)	
۰/۰۱۸۰	۰/۱۷۶۸	۰/۱۱۲۰	−۰/۰۱۱۱	−۰/۰۸۸۸	−۰/۰۵۹۴	−۰/۰۰۲۱	−۰/۰۰۸۱	−۰/۰۳۲۶	۴۰۰
(۰/۲۴۱۶)	(۰/۵۶۴۵)	(۰/۴۹۴۰)	(۰/۰۵۵۵)	(۰/۱۳۳۴)	(۰/۱۱۶۵)	(۰/۰۰۰۲)	(۰/۰۰۱۰)	(۰/۰۰۲۱)	
۰/۰۳۱۸	۰/۱۱۳۰	۰/۰۹۰۷	−۰/۰۱۰۰	−۰/۰۵۰۲	−۰/۰۴۰۲	−۰/۰۰۰۹	−۰/۰۰۴۶	−۰/۰۱۶۹	۸۰۰
(۰/۱۱۳۰)	(۰/۲۶۷۳)	(۰/۲۴۸۶)	(۰/۰۲۶۹)	(۰/۰۶۷۶)	(۰/۰۶۲۵)	(۰/۰۰۰۱)	(۰/۰۰۰۵)	(۰/۰۰۰۸)	

۶ داده واقعی

در این بخش برای مدل‌سازی و پیش‌بینی و تجزیه و تحلیل از داده‌های سری زمانی شمارشی که از انجمن جبران خسارت کارگران (WCB) بریتیش کلمبیا و کانادا جمع‌آوری شده است، استفاده می‌کنیم. داده‌ها شامل ۱۲۰ مشاهده است که از ژانویه ۱۹۸۵ شروع شده و در دسامبر ۱۹۹۴ به پایان می‌رسد. میانگین نمونه ۶/۱۳ و واریانس نمونه ۱۱/۸۰ است. از این‌رو، به‌نظر می‌رسد داده‌ها بیش‌پراکنش داشته می‌باشند.

شکل ۱ نمودار سری زمانی داده‌ها، تابع خودهمبستگی (ACF) و تابع خودهمبستگی جزئی (PACF) را نشان می‌دهد. نمودار سری زمانی نشان می‌دهد که سری در میانگین ایستاست. الگوی کاهش هندسی با دوره فصلی ۱۲ را می‌توان در نمودار ACF مشاهده کرد و نشان می‌دهد که یک سری زمانی با رفتارهای همبستگی سریالی وجود دارد. فصلی بودن و بیش پراکنشی این مجموعه داده‌ها آشکار است، این ویژگی‌ها باعث شد که ما مدل $DELINAR(۱)_{۱۲}$ را به‌عنوان مدل برازش شده انتخاب کنیم. مجموعه داده‌ها به دو بخش تقسیم می‌شود. ۱۱۰ مشاهده اول برای مدل‌سازی سری استفاده می‌شوند و ۱۰ مشاهده باقیمانده برای اهداف پیش‌بینی در نظر گرفته شده است. همچنین برای مقایسه مدل‌های $INAR(۱)_{۱۲}$، $DELINAR(۱)$ و $PINAR(۱)$ را نیز در نظر می‌گیریم.

همان‌طور که قبلا اشاره شد پارامتر α را ثابت و معلوم در نظر گرفتیم و با توجه به این‌که α مقادیر صحیح غیرمنفی را می‌گیرد، بنابراین مدل را با مقادیر ۱، ۲ و ۳ برای α با رویکرد ماکزیمم درستنمایی شرطی برازش کردیم تا بهترین را با توجه به معیار اطلاعاتی آکاییک (AIC) و معیار اطلاعات بیزی (BIC) انتخاب کنیم. با توجه به تجزیه و تحلیل حساسیت پارامتر α، به نظر می‌رسد که معیار اطلاع آکاییک با افزایش مقدار α کمی افزایش می‌یابد. بنابراین ما برای مدل $\alpha = ۱$ را در نظر می‌گیریم.



شکل ۱: نمودار سری زمانی و تابع خودهمبستگی و تابع خودهمبستگی جزیی داده‌های واقعی از سال ۱۹۸۵ تا ۱۹۹۴

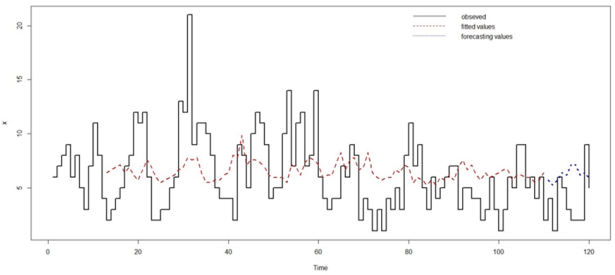
جدول ۴ برآوردهای حداقل مربعات و حداکثر درستنمایی شرطی (با میانگین مربعات خطا در پرانتز) را برای چهار مدل برازش شده و معیار اطلاع آکاییک (AIC) و معیار اطلاع بیزی (BIC) را نشان می‌دهد. همان‌طور که می‌بینیم، AIC و BIC مربوط به مدل $DELINAR(۱)_{۱۲}$ در مقایسه با مدل‌های دیگر کمترین است. نتایج بدست آمده در جدول ۴ نشان می‌دهد که مدل خودبازگشتی با

مقادیر صحیح غیرمنفی مرتبه اول دلاپوریت با ساختار فصلی بهتر از سایر مدل‌های خودبازگشتی با مقادیر صحیح غیرمنفی مرتبه اول فصلی و غیرفصلی می‌باشد. جهت بررسی توانایی مدل در پیش‌بینی، از معیار ریشه میانگین مربع خطا (RMSE) استفاده شده که برای همه مدل ها محاسبه شده است و در جدول ۴ نوشته شده است. مقادیر جدول ۴ نشان می‌دهد که ریشه میانگین مربع خطا در هر دو مورد مدل دلاپورت فصلی و غیرفصلی خیلی مشابه است ولی در بین همه مدل ها مدل پواسون غیرفصلی کمترین مقدار را دارد.

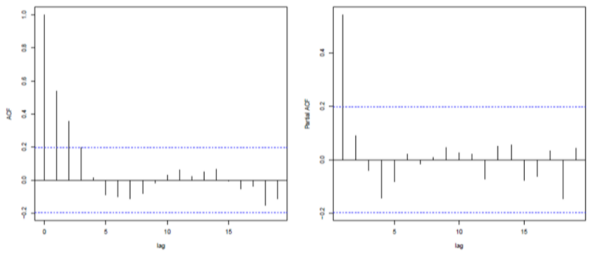
در شکل ۲ مقادیر مشاهده شده، مقادیر برازش و مقادیر پیش‌بینی به ترتیب با خطوط مشکی، قرمز و آبی نشان داده شده است. شکل ۳ نیز نشان دهنده تابع خودهمبستگی و تابع خودهمبستگی جزیی باقیمانده‌هاست و همان‌طور که مشاهده می‌شود همچنان همبستگی سریالی در باقیمانده‌ها دیده می‌شود.

جدول ۴: براورد پارامترها (میانگین مربعات خطا در پرانتز) و مقادیر AIC و BIC

مدل	پارامتر	برآورد CML	برآورد CLS	AIC	BIC	RMSE
$PINAR(۱)$	ρ	۰٫۵۷۰۵(۰٫۰۰۰۵)	۰٫۵۶۵۱(۰٫۰۰۷۵)	۵۴۵٫۸۰	۵۵۱٫۲۰	۳٫۱۰۱
	λ	(۰٫۰۱۸۸)۲٫۷۰۵۹	(۰٫۳۱۹۹)۲٫۷۳۵۶			
$INAR(۱)_{۱۲}$	ρ	(۰٫۰۰۰۴)۰٫۲۴۵۹	(۰٫۰۰۹۸)۰٫۲۶۶۷	۵۳۲٫۰۹	۵۳۷٫۴۹	۳٫۳۷۸
	λ	(۰٫۰۴۱۱)۴٫۶۷۶۹	(۰٫۴۲۰۱)۴٫۵۳۸۹			
$DELINAR(۱)$	ρ	(۰٫۰۰۱۶)۰٫۴۴۳۶	(۰٫۰۰۷۰)۰٫۵۶۵۱	۵۲۷٫۰۸	۵۳۵٫۱۸	۳٫۷۷۹
	β	(۰٫۰۰۷۷)۱٫۲۷۶۸	(۰٫۱۰۰۰)۱٫۶۸۳۱			
	λ	(۰٫۱۱۴۸)۳٫۶۹۴۵	(۰٫۳۲۲۳)۲٫۹۲۵۳			
$DELINAR(۱)_{۱۲}$	ρ	(۰٫۰۱۰۴)۰٫۲۲۸۵	(۰٫۰۱۰۰)۰٫۲۶۶۷	۵۰۰٫۲۱	۵۰۸٫۳۱	۳٫۷۸۵
	β	(۰٫۲۰۳۲)۲٫۷۳۴۵	(۰٫۳۱۹۷)۲٫۵۰۸۹			
	λ	(۰٫۴۸۱۶)۳٫۵۵۷۵	(۰٫۲۹۱۱)۳٫۶۸۱۱			



شکل ۲: نمودار سری زمانی مقادیر مشاهده شده، برازش شده و پیش‌بینی مربوط به داده‌های واقعی از سال ۱۹۸۵ تا ۱۹۹۴



شکل ۳: تابع خودهمبستگی و تابع خودهمبستگی جزیی باقیمانده‌ها

۷ نتیجه‌گیری

در این مقاله یک فرایند خودبازگشتی مرتبه اول فصلی با توزیع حاشیه‌ای دلاپورت معرفی گردید که توزیع نوآوری مدل به‌صورت پیچشی از توزیع پواسون و α تا توزیع هندسی تعدیل‌یافته در صفر می‌باشد. در ادامه ویژگی‌های اصلی مدل مانند میانگین واریانس و تابع خودهمبستگی مورد بررسی قرارداده شد. برای برآورد پارامترها از روش‌های یول–والکر، حداقل مربعات شرطی و حداکثر درستنمایی شرطی استفاده شده است. همچنین به پیش‌بینی مدل پرداخته شده است. با انجام یک شبیه‌سازی در حجم نمونه‌های متناهی عملکرد این برآوردگرها مورد ارزیابی قرار گرفت. در نهایت این مدل به یک مجموعه داده واقعی، برازش داده شد. برای مقایسه مدل‌های $PINAR(۱)$، $INAR(۱)_{۱۲}$ و $DELINAR(۱)$ نیز در نظر گرفته شد و نتیجه نشان داد که براساس معیار AIC و BIC، مدل پیشنهادی ما در مقایسه با سایر مدل‌های INAR بهتر است. به منظور تحقیقات بیشتر، می‌توان مدل را به مرتبه‌های بیشتر گسترش داد. همچنین، طبق ACF باقیمانده‌ها که در شکل ۳ نشان داده شده است، همچنان همبستگی سریالی در باقیمانده‌ها وجود داشت بنابراین می‌توان یک مدل جدید معرفی کرد که بتواند همبستگی فصلی و سریالی را کاهش دهد.

مراجع

Al-Osh MA, Alzaid AA. (1987) First-order interger-valued autoregressive (INAR(1)) process, *J Time Ser Anal.*, **8**(3), 261–275.

Barreto-Souza W. (2019), Mixed Poisson INAR(1) processes, *Stat Pap.*, **60**(3), 2119-2139.

Bourguignon M, Vasconcellos KL, Reisen VA, Ispány M (2016), A Poisson INAR(1) process with a seasonal structure, *J Stat Comput Simul*, **86**(2), 373–387.

Delaporte P. (1959), Quelques problémes de statistiquemathématique posés par Ìassurance automobile et le bonus non sinistre, *Bulletin Trimestriel de l'lnstitut des Actuaires Franą uis.*, **227**, 87–102.

Klimko, L.A. and Nelson, P.I., (1978), On conditional least squares estimation for stochastic processes, *The Annals of statistics*, 629-642.

McKenzie E. (1985), Some simple models for discrete variate time series, *J Am Water Resour Assoc.*, **21**(4): 645–650.

Steutel FW, Van Harn K. (1979), Discrete analogues of self-decomposability and stability, *Ann Probab*; **7**(5), 893–899.

Shalbaf, M., Parham, G. and Chinipardaz, R., (2022), Binomial Thinning Integer-Valued AR (1) with Poisson-α Fold Zero Modified Geometric Innovations, *Journal of Sciences, Islamic Republic of Iran*, **33**(1), 55-63.

Tian S, Wang D, Cui S. (2020), A seasonal geometric INAR process based on negative binomial thinning operator, *Statistical Papers*, **61**(6), 2561–2581.

A seasonal Integer-Valued AR(1) model with Delaporte marginal distribution

Maryam Shalbaf[1], **Gholam Ali Parham**[2]

[1]Department of Statistics and Mathematics, Sho.C., Islamic Azad University, Shoushtar, Iran

[2]Department of Statistics, Faculty of Mathematical Sciences and Computer, Shahid Chamran University of Ahvaz

---

Abstract: Real-count data time series often show the phenomenon of the overdispersion. In this paper, we introduce the first-order integer-valued autoregressive process with a seasonal structure. The univariate marginal distribution is derived from the Delaporte distribution and the innovations are convolution of Poisson with α-fold zero modified geometric distribution, based on binomial thinning operator. Some properties of the model are derived. The methods of Yule-Walker, conditional least squares and conditional maximum likelihood are used for estimating the parameters. The Monte Carlo experiment is conducted to evaluate the performances of these estimators in finite samples. At the end this model is illustrated using a real data set.

Keywords: Delaporte distribution, seasonal INAR(1) model, α-fold zero modified geometric, binomial thining.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

رویکرد فازی خوشه بندی داده ها بر پایه الگوریتم FCM

مولود میاهی [۱] ، بهرام طارمی [۲] ، مینا میاهی [۳]

[۱] گروه علوم پایه، دانشکده علوم پزشکی ساوه

[۲] بخش آمار، دانشکده علوم، دانشگاه شیراز

[۳] گروه ریاضی، مدعو دانشکده علوم پزشکی ساوه

چکیده: با افزایش حجم داده‌ها، تحلیل و استخراج الگوها به چالشی مهم تبدیل شده است. خوشه‌بندی یکی از روش‌های موثر برای گروه‌بندی داده‌های مشابه است. در روش‌های سنتی، هر داده تنها به یک خوشه تعلق دارد، اما در واقعیت برخی داده‌ها در مرز خوشه‌ها قرار می‌گیرند و نمی‌توان آن‌ها را به‌طور قطعی دسته‌بندی کرد. خوشه‌بندی فازی با امکان تعلق هم زمان یک داده به چند خوشه، رویکردی منعطف تر ارائه می‌دهد. این روش به دلیل مدل‌سازی بهتر، کاهش خطا و انعطاف‌پذیری مورد توجه محققان قرارگرفته است. این مقاله، به معرفی افراز فازی، خوشه بندی فازی و الگوریتم Fuzzy C-Means (FCM) می پردازد.

واژه‌های کلیدی: افراز فازی، خوشه‌بندی فازی، الگوریتم FCM

کد موضوع‌بندی ریاضی (۲۰۲۰): ۶۲M۱۰، ۹۱B۸۴.

---

۱ مقدمه

خوشه‌بندی یک تکنیک اکتشافی است که با تقسیم اشیاء همگن به گروه‌های مجزا، هدف آن ساده‌سازی مجموعه داده‌های پیچیده است. الگوریتم‌های خوشه‌بندی متعددی از دیدگاه‌های مختلف توسعه داده شده‌اند که روش‌های پارتیشن‌بندی، روش‌های مبتنی بر چگالی، روش‌های مبتنی بر مدل و روش‌های سلسله مراتبی پرکاربردترین آنها در عمل هستند. (هو و همکاران ، ۲۰۲۵) برخی از این الگوریتم‌ها معمولاً بر بهینه‌سازی یک تابع هدف خاص تمرکز دارند. با این حال، کمبود قابل توجهی در تفسیرپذیری و توضیح‌پذیری این نتایج خوشه‌بندی وجود دارد که چالش‌های تصمیم‌گیری را برای کاربران هنگام استفاده از این الگوریتم‌ها ایجاد می‌کند. (هو و همکاران ،

---

[۱] M.Miahi@savehums.ac.ir

[۲] سخنران، Tarami@shirazu.ac.ir

۲۰۲۵) یکی از این روش های چالش برانگیز روش متداول الگوریتم C-Means است که خوشه‌ها را بر پایه کمینه‌سازی فاصله تا مراکز خوشه تعیین می‌کند. این روش به شکل و پراکندگی داده‌ها حساسیت ندارد و همچنین در مواجهه با داده‌هایی که فاصله‌ای مشابه از چند مرکز دارند، دارای عملکرد نامطلوبی است. (عسگریان و همکاران ، ۱۳۸۶) همانطور که مشاهده می شود اصول مدل بندی کلاسیک در عدم قطعیت نمی تواند کارائی داشته باشد. سوالی در اینجا پیش می آید که با عدم قطعیت چه باید کرد؟ پاسخ این سوال را منطق فازی می دهد. (میاهی و نایینی ، ۱۳۹۱) بنابراین برای رفع این محدودیت‌ها، الگوریتمی مانند Fuzzy C-Means (FCM) توسعه داده شده است که امکان شناسایی خوشه‌های فازی با اشکال متنوع را فراهم می‌کنند. (عسگریان و همکاران ، ۱۳۸۶)

۲ دست آورد پژوهش

تعریف ۱.۲. افراز c تایی فازی :فرض کنید $X = \{x_۱, x_۲, ..., x_n\}$ مجموعه‌ای از داده‌ها باشد و V_{cn} مجموعه تمام ماتریس های حقیقی $c \times n$ که $۲ \leq c < n$ است، باشد و ماتریس $\tilde{U} = [u_{ik}] \in Vcn$ یک افراز c تایی فازی نامیده می‌شود اگر شرایط زیر را داشته باشد:

۱. $۱ \leq k \leq n \quad ۱ \leq i \leq c \quad ۰ \leq u_{ik} \leq ۱$

۲. $۱ \leq k \leq n \quad \sum_{i=۱}^{c} u_{ik} = ۱$

۳. $۱ \leq i \leq c \quad ۰ < \sum_{j=۱}^{n} u_{ik} < n$

مجموعه تمام ماتریس‌هایی که شرایط فوق را برآورده می‌سازند، مجموعه افرازهای فازی c تایی (M_{fc}) نامیده می‌شود. شرط اول نشان میدهد که در مقایسه با افراز قطعی، در افراز فازی داده‌ها می‌توانند با درجات عضویت مختلف به چندین خوشه تعلق داشته باشند. شرط دوم تضمین می‌کند که مجموع درجات عضویت یک داده در تمام خوشه‌ها برابر یک باشد، و شرط سوم بیان می‌کند که هیچ خوشه‌ای نباید کاملاً خالی یا شامل تمام داده‌ها باشد.

تعریف ۲.۲. تابع هدف : در افراز فازی c تایی، با استفاده از معیار پراکندگی و فاصله مینکوفسکی، تابع هدف بصورت زیر تعریف می‌شود:

$$minz(\tilde{U}, V) = \sum_{i=۱}^{c} \sum_{k=۱}^{n} (u_{ik})^m \|x_k - v_i\|^۲. \tag{۱.۲}$$

بطوریکه

$$v_i = \frac{۱}{\sum_{k=۱}^{n} u_{ik}} \sum_{k=۱}^{n} (u_{ik})^m . x_k \qquad i = ۱, ۲, ..., c. \tag{۲.۲}$$

است. که در آن $m > ۱$ ،v_i مرکز خوشه i ام و میانگین وزن دار x_k هاست که وزن هر x_k ، توان mام درجه عضویتش می باشد. هرچه درجه عضویت بالاتر باشد تاثیر بیشتری در تعیین v_i دارند و همچنین با افزایش m نیز این تاثیر افزایش می یابد.

تعریف ۳.۲. نرم عمومی : با فرض اینکه G ماتریس متقارن و مثبت $p \times p$ باشد، نرم عمومی را بصورت زیر تعریف می کنیم

$$\|x_k - v_i\|_G^۲ = (x_k - v_i)' . G . (x_k - v_i). \tag{۳.۲}$$

بادرنظر گرفتن نرم عمومی مسئله افراز فازی بصورت زیر نوشته می شود :

$$minz(\tilde{U}, V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\|x_k - v_i\|_G^2. \tag{۴.۲}$$

مزیت مسئله بالا این است که با کمک از معادلات دیفرانسیل می توان شرایط لازم را برای نقطه بهینه موضعی تعیین کرد. درواقع حل آن با درنظر گرفتن شرط مجموع درجات عضویت برابر با یک و مرکز خوشه بدست آمده از فرمول (۲.۲)، شرط

$$\mu_{ik} = \frac{(\frac{1}{\|x_k - v_i\|_G^2})^{\frac{1}{m-1}}}{\sum_{j=1}^{c}(\frac{1}{\|x_k - v_{ji}\|_G^2})^{\frac{1}{m-1}}}. \qquad i = 1, 2, ..., c \qquad k = 1, 2, ..., n \tag{۵.۲}$$

و با مشتق گیری از تابع هدف نسبت به v_i (به ازای $\tilde{U}$ ثابت) و نسبت به μ_{ik} (به ازای v ثابت) صورت می پذیرد.

در واقعیت معادلات بالا به دلیل پیچیدگی و غیرخطی بودن به‌صورت تحلیلی قابل حل نمی باشند. بنابراین با بهره گیری از الگوریتم‌های تکراری می توان مینیمم تابع هدف را بطور تقریبی بدست آورد. الگوریتمی که برای این منظور در این مقاله استفاده می شود الگوریتم FCM است.

تعریف ۴.۲. الگوریتم *FCM* : یک الگوریتم با یک روند تکرار بسیار ساده برای رسیدن به یک بهینه موضعی همگرا در حل معادله (۴.۲) است. برای طراحی این الگوریتم پارامترهای تعداد خوشه ها *(c)*، توان وزنی که فازی بودن نتایج خوشه ها را کنترل کند*(m)*، ماتریس متقارن و مثبت *($G_{P\times P}$)* ، مقدار دهی اولیه ماتریس *($\tilde{U}^{(0)}$)* ،معیار اختتام الگوریتم *(Δ)* نیاز است. این الگوریتم برای یافتن خوشه‌های نسبتا کروی مناسب است. با انتخاب یک مقدار ثابت برای c $(2 \leq c < n)$ و رعایت شرط $\sum_{i=1}^{c} u_{ik} = 1$ شروع و گام‌های زیر را سپری می‌شود:

۱. مقدار m و ماتریس متقارن مثبت $G_{p\times p}$ را انتخاب نموده سپس با قرار دادن $L = 0$ ماتریس $\tilde{U}^{(L)} \in M_{fc}$ مقدار دهی اولیه می شود.

۲. مراکز c خوشه فازی اولیه $v_i^{(L)}$ را با کمک از ماتریس $\tilde{U}^{(L)}$ و رابطه v_i *(فرمول (۲.۲))* بدست می آید.

۳. ماتریس درجات عضویت را به روز رسانی میشود یعنی ماتریس جدیدی به نام $\tilde{U}^{(L+1)}$ را یافته و برای یافتن هر یک از عناصر این ماتریس جدید از شرط زیر استفاده می گردد :

اگر به ازای هر i، $x_k \neq v_i^{(L)}$ شود از μ_{ik} *(فرمول (۵.۲))* کمک گرفته شود در غیر اینصورت از رابطه زیر استفاده گردد:

$$\mu_{jk}^{(L+1)} = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}$$

۴. محاسبه Δ و بررسی شرط توقف در این گام صورت می پذیرد. اگر $\Delta = \|\tilde{U}^{(L+1)} - \tilde{U}^{(L)}\|_G$ باشد و مقدار آن از ε بیشتر باشد آنگاه بجای L، $L + 1$ قرار داده و به گام دوم رفته و الگوریتم تکرار می گردد، درغیراینصورت الگوریتم متوقف می‌شود.

۳ دستورات در R

قبل از اجرای الگوریتم، داده‌ها را باید وارد R شده و در یک متغیر به نام x ذخیره گردد، سپس مراحل زیر انجام شود :

۱. **نصب و بارگذاری بسته:** ابتدا بسته‌ی ppclust نصب و بارگذاری می‌شود تا بتوان از الگوریتم FCM استفاده کرد.

```
install.packages("ppclust")
library(ppclust)
```

۲. **اجرای الگوریتم FCM :** با دستور مربوطه الگوریتم بر اساس تعداد خوشه‌ها و ضریب فازی روی داده‌ها اجرا می‌شود.

```
fcmresult <- fcm(x, centers = number of cluster, m )
```

۳. **مشاهده مراکز خوشه‌ها:** دستور دیگری مختصات مراکز خوشه‌ها را نشان می‌دهد.

```
fcmresul$centers
```

۴. **مشاهده ماتریس درجه تعلق:** براساس این ماتریس میتوان میزان تعلق هر داده به هر خوشه را نمایش داد.

```
fcmresult$u
```

۵. **تولید برچسب خوشه‌ها:** با استفاده از بیشترین درجه تعلق، برای هر داده یک برچسب خوشه مشخص می‌شود.

بحث و نتیجه‌گیری

تحلیل خوشه اي مبني بر افراز مجموعه اي از نقاط داده در تعدادي از خوشه ها (زیر گروه ها) استوار گردیده است، جايي که اشياء داخل يك خوشه(زیر گروه) يك درجه معين از شبيه بودن را نشان مي دهند. خوشه بندي متداول هر نقطه (بردار ويژگي) را به يكي و فقط يكي از خوشه ها با درجه عضويت يك اختصاص مي دهد. فرضيات به خوبي مرز بين خوشه ها را تعيين مي كنند ولی مرز اغلب توصيفي از داده حقيقي را منعكس نمي كند. بنابراين اگر بين زير گروه ها ارتباط فازي باشد، جايي که توصيف گوناگوني از شباهت اشياء در يك خوشه خاص نياز باشد، مشكلات فراواني ايجاد مي شود. در این شرایط خوشه‌بندی فازی می‌تواند راهکار مناسبی ارائه دهد. با این حال، چالش‌هایی مانند تعیین تعداد بهینه خوشه‌ها، موقعیت اولیه مراکز، و تأثیر شکل و چگالی داده‌ها در نتایج نهایی، از مهم‌ترین مسائلی هستند که در پیاده‌سازی موفق این روش باید مورد توجه قرار گیرند.

مراجع

عسگریان، ا.، معین‌زاده، ح.، سریانی، م.، و حبیبی، ج. (۱۳۸۶)، رویکرد جدید برای خوشه‌بندی فازی بوسیله الگوریتم ژنتیک، سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران.

میاهی، م. و نایینی، ح. (۱۳۹۱)، خوشه‌بندی سری زمانی فازی بر اساس تابع خودهمبستگی، اولین همایش بین‌المللی اقتصادسنجی، روش‌ها و کاربردها.

Hu L., Jiang M., Liu X., and He Z. (2025), *Significance-based decision tree for interpretable categorical data clustering*, Inform. Sci., 690, 121588.

Hu L., Jiang M., Dong J., Liu X., and He Z. (2025), *Interpretable categorical data clustering via hypothesis testing*, Pattern Recogn., 111364.

A Fuzzy Approach to Data Clustering Based on the FCM Algorithm

Moloud Miahi[1], **Bahram Tarami**[2], **Mina Miahi**[3]

[1]Department of Basic Sciences, Faculty of Medical Sciences, Saveh, Iran

[2]Department of Statistics, Faculty of Basic Sciences, Shiraz University, Shiraz, Iran

[3]Department of Mathematics, Adjunct Lecturer, Saveh University of Medical Sciences, Saveh, Iran

---

Abstract: With the growing volume of data, analyzing and extracting meaningful patterns has become a major challenge. Clustering is a widely used technique for grouping similar data points. In conventional clustering methods, each data point is assigned to a single cluster; however, in practice, some data points lie on the boundaries between clusters and cannot be classified definitively. Fuzzy clustering, which allows a data point to belong to multiple clusters simultaneously, provides a more flexible and realistic approach. This method has gained considerable attention from researchers due to its superior modeling capabilities, error reduction, and adaptability. This paper presents an overview of fuzzy partitioning, fuzzy clustering, and the Fuzzy C-Means (FCM) algorithm.

Keywords: Fuzzy partitioning, Fuzzy clustering, Fuzzy C-Means (FCM) algorithm

Mathematics Subject Classification (2020): $62M10$، $91B84$.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

بررسی رفتار مجانبی رهیافت PTE در حضور همخطی چندگانه

سید امیرحسین طباطبائی شیرازی [۱]، مهدی عمادی، محمد آرشی، سولماز سیف‌اللهی

گروه آمار، دانشگاه فردوسی مشهد

چکیده: رگرسیون خطرات متناسب کاکس یکی از مدل‌های پرکاربرد در تحلیل داده‌های بقا با متغیرهای کمکی است. با این حال، در حضور هم‌خطی چندگانه، کارایی این مدل کاهش می‌یابد و برآوردهای حاصل از روش استاندارد درستنمایی جزئی ممکن است غیرقابل اعتماد شوند. در این مقاله، برآوردگر پیش‌آزمون را بر اساس برآوردگر لیو بسط می‌دهیم تا دقت برآورد ضرایب افزایش یابد. ویژگی‌های نظری حدی این برآوردگر را تعیین و عملکرد آن از طریق شبیه‌سازی‌های گسترده مونت‌کارلو ارزیابی می‌گردد. همچنین، نحوه کاربرد این استراتژی برآورد بر داده‌های بقا مورد بررسی قرار می‌گیرد. نتایج نشان‌دهنده بهبودهای چشمگیر بوده و مزایای عملی این برآوردگر را برای پژوهشگران به‌خوبی آشکار می‌سازد.

واژه‌های کلیدی: مدل خطرات متناسب کاکس، برآوردگر لیو، همخطی، برآوردگر پیش‌آزمون.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62N02، 62F30.

---

۱ مقدمه

در عصر حاضر، تحلیل داده‌های پیچیده و حجیم، به‌ویژه در حوزه‌هایی مانند پزشکی، زیست‌فناوری و مهندسی، به یکی از چالش‌های اساسی پژوهشگران تبدیل شده است. در این میان، مدل‌های رگرسیونی ابزارهای مهمی برای تفسیر روابط میان متغیرها به‌شمار می‌روند. یکی از پرکاربردترین این مدل‌ها در تحلیل داده‌های بقا، مدل خطرات متناسب کاکس (CPHM) است که با امکان بررسی هم‌زمان تأثیر چندین عامل بر نرخ وقوع یک رویداد، ابزاری قدرتمند در مطالعات طولی به‌حساب می‌آید (کاکس ، ۱۹۷۲).

مدل کاکس با فرض ثبات ضرایب در طول زمان و بهره‌گیری از روش درست‌نمایی جزئی برای برآورد پارامترها، ساختاری انعطاف‌پذیر برای تحلیل زمان وقوع رویدادهای حیاتی فراهم می‌آورد. با این حال، در عمل با چالش‌هایی روبه‌روست؛ یکی از مهم‌ترین آن‌ها، هم‌خطی چندگانه یا همبستگی شدید میان متغیرهای مستقل است. این پدیده می‌تواند منجر به ناپایداری ضرایب، افزایش واریانس برآوردها و

---

[۱] سخنران، tabatabaeishirazi.se@um.ac.ir

کاهش دقت و قابلیت تفسیر نتایج شود. راهکارهایی نظیر حذف یا گروه‌بندی متغیرهای همبسته، اگرچه مرسوم‌اند، اما ممکن است موجب افت کارایی مدل و بروز اریبی در نتایج گردند. از این رو، توجه به این چالش و ارائه راهکارهای مؤثر برای مقابله با آن، ضرورتی اجتناب‌ناپذیر است. در راستای مقابله با هم‌خطی در CPHM، برآوردگرهای متعددی پیشنهاد شده‌اند؛ از جمله برآوردگر ریج، برآوردگر نوع–لیو و برآوردگر کیبریا–لکمان. در این میان، برآوردگر لیو به‌سبب ساختار ساده، برخورداری از تنها یک پارامتر تنظیم و سهولت در تعیین آن، مورد توجه ویژه قرار گرفته است. مطالعات متعدد نشان داده‌اند که این برآوردگر، به‌ویژه در شرایط وجود هم‌خطی، از نظر معیار میانگین مربعات خطا، عملکرد بهتری نسبت به برآوردگر حداکثر درست‌نمایی جزئی (MPLE) دارد.

در برخی از تحلیل‌های رگرسیونی، اطلاعات پیشین درباره پارامترهای مدل ممکن است در دسترس باشد. به عنوان مثال، نتایج تحلیل‌های مقدماتی می‌توانند نشان دهند که برخی از متغیرها از لحاظ آماری معنادار نیستند و در نتیجه باید از مدل حذف شوند. این قبیل اطلاعات را می‌توان به صورت قیود خطی از نوع $\boldsymbol{C\beta} = \boldsymbol{b}$ در ساختار مدل اعمال کرد. برآوردگری که تحت این قیود به‌دست می‌آید، برآوردگر مقید نامیده می‌شود. با این وجود، در بسیاری از موارد، نسبت به صحت این قیود تردید وجود دارد. در چنین شرایطی، استفاده از برآوردگر پیش‌آزمون پیشنهاد می‌شود. این برآوردگر، ترکیبی از برآوردگر معمول و برآوردگر مقید است که تصمیم‌گیری درباره استفاده از هر یک، بر اساس نتایج آزمون فرض آماری انجام می‌گیرد. مطالعات نشان داده‌اند که در صورت اعتبار اطلاعات پیشین، برآوردگر پیش‌آزمون می‌تواند نسبت به برآوردگرهای کلاسیک عملکرد بهتری از لحاظ میانگین مربعات خطا داشته باشد.

هدف این مقاله، معرفی برآوردگر پیش‌آزمون در (CPHM) با استفاده از برآوردگر لیو و بررسی ویژگی‌های مجانبی آن است. در ادامه، با بهره‌گیری از شبیه‌سازی مونت‌کارلو، عملکرد این برآوردگر با برآوردگر لیو در شرایط مختلف، مقایسه خواهد شد و در نهایت، کارایی آن در تحلیل داده‌های مربوط به سرطان ریه مورد ارزیابی قرار خواهد گرفت.

۲ مدل خطرات متناسب کاکس

مدل CPHM که نخستین‌بار توسط کاکس (۱۹۷۲) معرفی شد، به‌عنوان یکی از ارکان اصلی در تحلیل بقا شناخته می‌شود. این مدل، به‌ویژه به دلیل توانایی‌اش در برقراری ارتباط میان زمان وقوع یک رویداد و یک یا چند متغیر کمکی، بدون نیاز به فرض شکل مشخصی برای تابع خطر پایه، از محبوبیت فراوانی برخوردار است. فرض کنید $\tau_i = \min(t_i, c_i)$ که در آن، t_i زمان بقا در صورت مشاهده رویداد و c_i زمان سانسور در صورت عدم مشاهده رویداد است. همچنین، بردار متغیرهای کمکی را به صورت $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ در نظر بگیرید. در این صورت، مدل CPHM به صورت $h(t_i \mid \mathbf{x}_i) = h_0(t_i)\exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ تعریف می‌شود که در آن، $h_0(t)$ تابع خطر پایه و $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ بردار ضرایب مجهول مدل می‌باشد. برآوردگر حداکثر درست‌نمایی جزئی (MPLE) برای برآورد $\boldsymbol{\beta}$ به صورت زیر محاسبه شده است:

$$\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} = (\hat{\boldsymbol{Z}}^\top \hat{\boldsymbol{Z}})^{-1} \hat{\boldsymbol{Z}} \hat{\boldsymbol{V}}, \tag{۱.۲}$$

بطوریکه $\hat{\boldsymbol{Z}}$ ماتریسی به ابعاد $\sum(m_i + 1) \times p$ است، m_i اندازه مجموعه‌ی ریسک آزمودنی i است و $\hat{\boldsymbol{V}}$ برداری با ابعاد $p \times 1$ می باشدکه در$\hat{\boldsymbol{\beta}}_{PMPLE}$ محاسبه می‌شود تحت شرایط نظم، نشان داده شده که برآوردگرهای MPLE دارای توزیع مجانبی نرمال به صورت زیر می‌باشد: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{\beta}) \sim N_p(0, \boldsymbol{F}^{-1})$ که در آن، $\boldsymbol{F}^{-1} = \hat{\boldsymbol{Z}}^\top \hat{\boldsymbol{Z}}$ ماتریس اطلاع فیشر است (ماگنوس ، ۱۹۹۷).

فرض کنید اطلاعات پیشین درباره ضرایب مدل را بتوان در قالب q قید خطی مستقل بیان کرد. برای بررسی صحت این قیود در

مدل، می‌توان آن‌ها را بر اساس فرضیه‌های آماری $H_0 : \boldsymbol{C\beta} = \boldsymbol{b}\ vs\ H_1 : \boldsymbol{C\beta} \neq \boldsymbol{b}$ آزمود. در این رابطه، $\boldsymbol{C}$ و $\boldsymbol{b}$ به ترتیب ماتریس و برداری معلوم با ابعاد $q \times p$ و $q \times 1$ می‌باشند. بر اساس توزیع مجانبی MPLE، آماره‌ای که برای آزمون فرض صفر H_0 به کار می‌رود، به صورت زیر تعریف می‌شود:

$$T_n = \left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}\right)^{\top}\left(\boldsymbol{C}\boldsymbol{I}^{-1}\boldsymbol{C}^{\top}\right)^{-1}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}\right). \tag{۲.۲}$$

با افزایش اندازه نمونه (n)، آماره آزمون T_n تحت H_0 دارای توزیع مجانبی کای‌دو با q درجه آزادی است. در صورتی که H_0 برقرار باشد، می‌توان برای برآورد ضرایب مدل از برآوردگر درست‌نمایی جزئی مقید (RMPLE) استفاده کرد که به صورت زیر تعریف می‌شود:

$$\hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}} = \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{A}\left(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}\right), \tag{۳.۲}$$

که در آن $\boldsymbol{A} = \boldsymbol{F}^{-1}\boldsymbol{C}^{\top}\left(\boldsymbol{C}\boldsymbol{F}^{-1}\boldsymbol{C}^{\top}\right)^{-1}$.

۳ برآوردگرهای پیشنهادی در مدل CPHM

برآوردگر لیو یکی از روش‌های شناخته‌شده برای برآورد ضرایب در حضور همخطی در مدل‌های رگرسیونی است. در CPHM، این برآوردگر توسط احمد و همکاران (۲۰۲۳)به صورت $\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} = \left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + I_p\right)^{-1}\left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + dI_p\right)\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}$ تعریف شده است که در آن $d \in [0, 1]$ پارامتر لیو می‌باشد. برای سادگی می‌توان برآوردگر لیو را به فرم $\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} = \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}$ نوشت که در آن $\boldsymbol{B} = \left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + I_p\right)^{-1}\left(\hat{\boldsymbol{Z}}^{\top}\hat{\boldsymbol{Z}} + dI_p\right)$ می‌باشد. برای برآورد پارامتر لیو از برآوردگری زیر در این مقاله استفاده می‌کنیم:

$$\hat{d} = 1 - \frac{\sum_{j=1}^{p} 1/\lambda_j(\lambda_j + 1)}{\sum_{j=1}^{p} \hat{\boldsymbol{\beta}}^{2}_{\mathrm{MPLE(j)}}/(\lambda_j + 1)^2}. \tag{۱.۳}$$

با الهام گرفتن از کیبریا و صالح (۲۰۰۴) می‌توان فرم برآوردگر مقید لیو (RCLE) به صورت، $\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}} = \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}}$ تعریف کرد.

برآوردگری که هدف اصلی این مقاله می‌باشد، برآوردگر پیش آزمون (PTE) در مدل CPHM می‌باشد. این برآوردگر که از انواع برآوردگرهای انقباضی می‌باشد، بر اساس صحت و یا نادرستی فرض مطرح شده در مدل، بین دو برآوردگر مقید و نامقید تصمیم‌گیری می‌کند. با استفاده از برآوردگر MPLE و RMPLE برآوردگر پیش آزمون به فرم زیر می‌توان تعریف کرد

$$\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} = \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - (\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}})I_{\left(Tn < \chi^{2}_{(q,\alpha)}\right)}, \tag{۲.۳}$$

و بنا به رابطه ای که بین $\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}$ و $\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}$ به ترتیب با $\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}$ و $\hat{\boldsymbol{\beta}}_{\mathrm{RMPLE}}$ دارند، می‌توان برآوردگر پیش آزمون بر اساس برآوردگر لیو را به صورت $\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}} = \boldsymbol{B}\hat{\boldsymbol{\beta}}_{\mathrm{PTE}}$ تعریف کرد. یا به عبارتی دیگر،

$$\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}} = \hat{\boldsymbol{\beta}}_{\mathrm{RCLE}} - (\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) I_{\left(Tn < \chi^{2}_{(q,\alpha)}\right)}, \tag{۳.۳}$$

فرم برآوردگر پیش‌آزمون لیو حکایت از این دارد که اگر H_{0} پذیرفته شود، برآوردگر پیش‌آزمون برابر RCLE و در غیر این صورت، برابر برآوردگر CLE خواهد بود. این برآوردگر توسط ماگنوس (۱۹۹۷) معرفی شد و به شدت در مدل‌های رگرسیونی مورد توجه قرار گرفت.

۱.۳ ویژگی برآوردگرهای پیشنهادی

در این بخش به بررسی ویژگی برآوردگرهای معرفی شده به خصوص CPTE می‌پردازیم. در ابتدا آماره آزمون T_n در رابطه (۲.۲) را تحت $H_{\nu} : \boldsymbol{C\beta} = \boldsymbol{b} + \boldsymbol{\nu}, \boldsymbol{\nu} \in \mathbb{R}^{+q}$ در نظر می‌گیریم. در این حالت، زمانیکه $n \to \infty$، خواهیم داشت $T_n \to \infty$، به‌گونه‌ای که $\lim_{n\to\infty} P_{\boldsymbol{\nu}}(T_n \geq x) = 1$. به بیان دیگر، T_n یک آزمون سازگار است. در نتیجه، تحت فرض‌های جایگزین ثابت، زمانی که $n \to \infty$ داریم

$$\begin{aligned}
&n(\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} - \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}) \left(\frac{1}{n}\boldsymbol{I}\right) (\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} - \hat{\boldsymbol{\beta}}_{\mathrm{MPLE}}) \\
&= n(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b})^{\top} (\frac{1}{n}\boldsymbol{C}\boldsymbol{I}^{-1}\boldsymbol{C}^{\top})^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{b}) I_{\left(Tn < \chi^{2}_{(q,\alpha)}\right)} \\
&= T_n I_{\left(Tn < \chi^{2}_{(q,\alpha)}\right)} \leq \chi^{2}_{(q,\alpha)} I_{\left(Tn < \chi^{2}_{(q,\alpha)}\right)} \to 0.
\end{aligned}$$

به عبارتی دیگر $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{PTE}} - \boldsymbol{\beta}) \overset{D}{\equiv} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{MPLE}} - \boldsymbol{\beta})$ و در نتیجه $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}} - \boldsymbol{\beta}) \overset{D}{\equiv} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \boldsymbol{\beta})$. بنابراین، به‌منظور تمایز مناسب میان توزیع‌های مجانبی برآوردگرها، فرض‌های جایگزین محلی $H_{0(n)}$ را به صورت زیر در نظر می‌گیریم:

$$H_{0(n)} : \boldsymbol{C\beta} = \boldsymbol{b} + \frac{\boldsymbol{\vartheta}}{\sqrt{n}}; \qquad \boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_q)^{\top} \neq 0 \tag{۴.۳}$$

در این صورت، وقتی $n \to \infty$، T_n دارای توزیع غیر مرکزی کای‌دو با q درجه آزادی و پارامتر غیرمرکزی $\Delta = \boldsymbol{\vartheta}^{\top}(\boldsymbol{C}\boldsymbol{I}^{-1}\boldsymbol{C}^{\top})^{-1}\boldsymbol{\vartheta}$ خواهد بود. حال می‌توان تحت $H_{0(n)}$ و زمانی که $n \to \infty$ اثبات کرد که

$$\begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \boldsymbol{\beta}) \\ \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}} - \boldsymbol{\beta}) \\ \sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}} - \hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) \end{pmatrix} \sim N_{3p}\left(\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \boldsymbol{b}_3 \end{pmatrix}, \begin{pmatrix} \boldsymbol{B}\boldsymbol{I}^{-1}\boldsymbol{B}^{\top} & \boldsymbol{B}(\boldsymbol{I}^{-1} - \boldsymbol{J})\boldsymbol{B}^{\top} & \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{\top} \\ \boldsymbol{B}(\boldsymbol{I}^{-1} - \boldsymbol{J})\boldsymbol{B}^{\top} & \boldsymbol{B}(\boldsymbol{I}^{-1} - \boldsymbol{J})\boldsymbol{B}^{\top} & 0 \\ \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{\top} & 0 & \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^{\top} \end{pmatrix} \right) \tag{۵.۳}$$

که در آن $\boldsymbol{\gamma} = \boldsymbol{B\vartheta}$، $\boldsymbol{J} = \boldsymbol{B}\boldsymbol{C}\boldsymbol{I}^{-1}$، $\boldsymbol{b}_2 = \boldsymbol{b}_1 - \boldsymbol{B\gamma}$ و $\boldsymbol{b}_3 = \boldsymbol{B\gamma}$ (کیبریا و صالح، ۲۰۰۴).

حال اگر اریبی مجانبی را به صورت زیر تعریف کنیم

$$\mathbb{B}(\hat{\boldsymbol{\beta}}) = \lim_{n\to\infty} \mathbb{E}\big(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big), \tag{۶.۳}$$

در این صورت با استفاده از رابطه (۵.۳) ، برای برآوردگرهای بر اساس برآوردگر لیو خواهیم داشت:

$$\mathbb{B}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}) = \boldsymbol{b}_1, \qquad \mathbb{B}(\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) = \boldsymbol{b}_2, \qquad \mathbb{B}(\hat{\boldsymbol{\beta}}_{\mathrm{CLPTE}}) = \boldsymbol{b}_1 - \boldsymbol{b}_3 \Upsilon_q(\Delta; \chi^2_{(q,\alpha)}).$$

که در آن $\Upsilon_q(\Delta; z)$ نمایانگر تابع توزیع تجمعی در نقطه z برای توزیع غیر مرکزی کای‌دو با q درجه آزادی و پارامتر غیر مرکزی Δ می‌باشد. همچنین، اگر واریانس برآوردگر را به فرم

$$\mathbb{V}(\hat{\boldsymbol{\beta}}) = \lim_{n\to\infty} \mathbb{E}\big(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\big).$$

تعریف کنیم، در این صورت داریم:

$$\begin{aligned}
&\mathbb{V}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}) = \boldsymbol{B}\boldsymbol{I}^{-1}\boldsymbol{B}^\top + \boldsymbol{b}_1\boldsymbol{b}_1^\top,\\
&\mathbb{V}(\hat{\boldsymbol{\beta}}_{\mathrm{RCLE}}) = \boldsymbol{B}(\boldsymbol{I}^{-1} - \boldsymbol{J})\boldsymbol{B}^\top + \boldsymbol{b}_2\boldsymbol{b}_2^\top,\\
&\mathbb{V}(\hat{\boldsymbol{\beta}}_{CLPTE}) = \mathbb{V}(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}) - \Big[2(\boldsymbol{b}_1 - \boldsymbol{b}_3)\boldsymbol{b}_3^\top + \boldsymbol{B}\boldsymbol{J}\boldsymbol{B}^\top\Big]\Upsilon_{q+2}(\Delta; \chi^2_{(q,\alpha)}) - \boldsymbol{b}_3\boldsymbol{b}_3^\top\Upsilon_{q+4}(\Delta; \chi^2_{(q,\alpha)}),
\end{aligned}$$

۴ شبیه‌سازی

در این بخش، با استفاده از شبیه‌سازی مونت کارلو به ارزیابی عملکرد برآوردگرهای پیشنهادی خواهیم پرداخت. برای بررسی عملکرد برآوردگرها در حضور همخطی در مدل، متغیرهای کمکی را مطابق مدل زیر تولید می‌کنیم:

$$x_{ij} = \sqrt{1-\rho^2}Z_{ij} + \rho Z_{i(p+1)}; \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p, \tag{۱.۴}$$

که در آن Z_{ij} نمونه‌های تصادفی از توزیع نرمال استاندارد می‌باشد. پارامتر ρ شدت همخطی بین متغیرهای کمکی را کنترل می‌کند و مقدار آن را $p = \{0.5, 0.7, 0.95\}$ در نظر گرفته می‌گیریم. همچنین $n = \{50, 100, 200\}$، $p = \{6, 12\}$ و احتمال سانسور را $cp = \{0.2, 0.3\}$ و زمان‌های بقاء پایه را توزیع نمایی با پارامتر نرخ ۰/۵ فرض می‌کنیم. مقادیر واقعی ضرایب را به صورت $\boldsymbol{\beta} = (0.5, 1, -0.6, 0.8, 0, 0, \ldots, 0)^\top$ تعریف می‌کنیم بدین معنا که، $p-4$ ضریب آخر بردار $\boldsymbol{\beta}$ برابر صفرند و فرض صفر مدل را بر اساس این پارامترهای صفر در مدل در نظر می‌گیریم. در نهایت، داده‌ها برای CPHM مطابق روش توضیح داده شده توسط استین (۲۰۱۲) تولید می‌کنیم. علاوه بر این، برای بررسی اثر انحراف از مقدار واقعی ضرایب واقعی، بردار $\boldsymbol{b}$ در فرض صفر را به صورت $\boldsymbol{b} = (\delta, 0^\top_{p-5})^\top$ تعریف می‌کنیم که در آن $\delta \in [0, 4]$. این رابطه نشان می‌دهد که $\delta = 0$ متناظر با مقادیر واقعی ضرایب است و مقادیر دیگر نشان‌دهنده انحراف از مقادیر واقعی می‌باشند. برای ارزیابی عملکرد برآوردگرها، تولید داده‌ها را ۱۰۰۰ بار تکرار کرده و کارایی نسبی برآوردگرهای پیشنهادی را که به صورت زیر تعریف می‌شود، محاسبه می‌کنیم:

$$\mathrm{RE}(\hat{\boldsymbol{\beta}}) = \frac{\sum_{m=1}^{1000} \left(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}^{(m)} - \boldsymbol{\beta}\right)^{\top} \left(\hat{\boldsymbol{\beta}}_{\mathrm{CLE}}^{(m)} - \boldsymbol{\beta}\right)}{\sum_{m=1}^{1000} \left(\hat{\boldsymbol{\beta}}_{*}^{(m)} - \boldsymbol{\beta}\right)^{\top} \left(\hat{\boldsymbol{\beta}}_{*}^{(m)} - \boldsymbol{\beta}\right)}, \tag{۲.۴}$$

که در آن $\hat{\boldsymbol{\beta}}_{*}^{(m)}$ نشان‌دهنده هر یک از برآوردگرهای پیشنهادی در تکرار mام است. از نتایج شبیه‌سازی که در جدول ۱ گزارش شده است. با توجه به نتایج، زمانی که مقدار $\delta = 0$ باشد، برآوردگرهای پیشنهادی عملکرد بهتری نسبت به سایرین دارند. اما با فاصله گرفتن δ از مقدار واقعی، کارایی آن‌ها کاهش یافته و به‌ویژه برآوردگر لیو مقید به‌شدت افت عملکرد دارد، در حالی‌که کارایی برآوردگر پیش‌آزمون به مقدار یک میل می‌کند. همچنین، با افزایش اندازه نمونه، کارایی نسبی برآوردگرها کاهش می‌یابد، اما افزایش تعداد متغیرها موجب بهبود کارایی نسبی آن‌ها می‌شود. افزایش ضریب همبستگی بین متغیرها نیز منجر به ارتقای کارایی نسبی می‌گردد. در نهایت، با افزایش احتمال سانسور، هرچند مقادیر کارایی نسبی اندکی افزایش می‌یابد، اما روند کلی نتایج تغییر چندانی نمی‌کند.

جدول ۱: کارایی برآوردگرهای پیشنهادی نسبت به برآوردگر لیو در مدل CPHM.

		$cp = 0.2$						$cp = 0.3$					
		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.95$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.95$	
n	δ	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE	RCLE	CLPTE
$p = 6$													
50	0	1.625	1.371	1.686	1.415	1.610	1.380	1.634	1.386	1.685	1.419	1.599	1.394
	1	0.225	0.969	0.281	0.941	0.743	0.811	0.251	0.944	0.314	0.889	0.798	0.846
	2	0.064	1.000	0.082	1.000	0.309	0.872	0.072	1.000	0.094	1.000	0.350	0.835
	3	0.029	1.000	0.038	1.000	0.159	0.982	0.033	1.000	0.044	1.000	0.184	0.970
	4	0.017	1.000	0.022	1.000	0.095	1.000	0.019	1.000	0.025	1.000	0.111	0.997
100	0	1.526	1.372	1.592	1.423	1.638	1.472	1.510	1.377	1.572	1.423	1.615	1.460
	1	0.108	1.000	0.135	1.000	0.447	0.772	0.124	1.000	0.154	1.000	0.490	0.757
	2	0.029	1.000	0.036	1.000	0.147	0.993	0.033	1.000	0.042	1.000	0.167	0.987
	3	0.013	1.000	0.016	1.000	0.070	1.000	0.015	1.000	0.019	1.000	0.080	1.000
	4	0.007	1.000	0.009	1.000	0.040	1.000	0.009	1.000	0.018	1.000	0.047	1.000
200	0	1.493	1.346	1.576	1.386	1.689	1.460	1.482	1.337	1.563	1.373	1.667	1.444
	1	0.052	1.000	0.065	1.000	0.251	0.925	0.058	1.000	0.073	1.000	0.276	0.898
	2	0.013	1.000	0.017	1.000	0.072	1.000	0.015	1.000	0.019	1.000	0.081	1.000
	3	0.006	1.000	0.008	1.000	0.033	1.000	0.007	1.000	0.009	1.000	0.037	1.000
	4	0.003	1.000	0.004	1.000	0.019	1.000	0.004	1.000	0.005	1.000	0.021	1.000
$p = 12$													
50	0	4.504	2.525	4.657	2.544	3.689	2.287	4.400	2.512	4.552	2.642	3.580	2.294
	1	0.639	0.972	0.788	0.965	1.751	1.254	0.705	0.967	0.868	0.971	1.831	1.298
	2	0.185	1.000	0.236	1.000	0.754	0.936	0.208	1.000	0.267	0.999	0.829	0.937
	3	0.085	1.000	0.110	1.000	0.394	0.967	0.096	1.000	0.125	0.999	0.442	0.965
	4	0.048	1.000	0.110	1.000	0.238	1.000	0.055	1.000	0.072	1.000	0.270	0.996
100	0	3.813	2.491	4.117	2.614	4.038	2.666	3.824	2.452	4.103	2.554	3.948	2.567
	1	0.261	1.000	0.335	0.995	1.102	0.972	0.296	1.000	0.379	0.992	1.199	0.999
	2	0.069	1.000	0.090	1.000	0.363	0.976	0.079	1.000	0.103	1.000	0.410	0.958
	3	0.031	1.000	0.041	1.000	0.173	1.000	0.036	1.000	0.047	1.000	0.197	1.000
	4	0.018	1.000	0.023	1.000	0.100	1.000	0.020	1.000	0.027	1.000	0.114	1.000
200	0	3.326	2.463	3.552	2.660	3.924	2.785	3.179	2.462	3.464	2.593	3.815	2.789
	1	0.110	1.000	0.143	1.000	0.570	0.902	0.125	1.000	0.164	1.000	0.635	0.878
	2	0.028	1.000	0.037	1.000	0.164	1.000	0.032	1.000	0.043	1.000	0.109	1.000
	3	0.013	1.000	0.017	1.000	0.075	1.000	0.015	1.000	0.019	1.000	0.086	1.000
	4	0.007	1.000	0.009	1.000	0.043	1.000	0.008	1.000	0.011	1.000	0.049	1.000

۵ تحلیل داده‌های بقا

در این بخش، برای ارزیابی عملکرد برآوردگرهای پیشنهادی، داده‌های مربوط به بقای بیماران مبتلا به سرطان پیشرفته ریه از گروه درمانی سرطان North Central تحلیل خواهیم کرد. این داده‌ها، موجود در بسته آماری survival نرم‌افزار R با عنوان cancer است که شامل اطلاعاتی درباره‌ی توانایی بیماران در انجام فعالیت‌های روزمره است که با استفاده از نمرات عملکردی ارزیابی شده‌اند. پس از حذف رکوردهای دارای مقادیر گمشده، مجموعه داده نهایی شامل ۱۶۷ بیمار می‌باشد. متغیرهای موجود در این مجموعه داده شامل زمان بقا بر حسب روز، وضعیت بیمار، سن، جنسیت، نمره عملکردی کارنوفکسی بر اساس ارزیابی پزشک (ph.karno)، نمره عملکردی ECOG بر اساس نظر پزشک (ph.ecog) (۰ بدون علامت، ۱ علامت‌دار اما کاملاً قادر به حرکت، ۲ بستری کمتر از ٪۵۰از روز، ۳ بستری بیش از ٪۵۰ از روز ولی نه کاملاً بستری، ۴ کاملاً بستری) و کاهش وزن در شش ماه گذشته می‌باشد.

نتایج آزمون نسبت درست‌نمایی و آزمون والد برای برازشCPHM بر این داده‌ها، حاکی از برازش مناسب مدل با p–مقدار به ترتیب برابر با ۹×۱۰^{-۵} و $۱۰^{-۴}$ می‌باشند. بررسی همبستگی میان متغیرهای کمکی، همبستگی قوی به میزان ۰٫۸۱۱ بین متغیرهای ph.karno و ph.ecog را نشان می‌دهد. همچنین عدد شرطی محاسبه‌شده برابر با ۲۲۳٫۹۴ بوده که وجود همخطی را تایید می‌کند.
با توجه به نبود اطلاعات پیشین درباره ضرایب مدل، آزمون ضرایب انجام شد که نشان داد متغیرهای سن و کاهش وزن از نظر آماری معنادار نیستند و ضرایب آن‌ها صفر در نظر گرفته شد. برای ارزیابی عملکرد برآوردگرهای پیشنهادی مبتنی بر لیو، از روش بوت‌استرپ با اندازه نمونه ۵۰ و ۱۰۰۰ تکرار استفاده شد. نتایج جدول ۲ نشان می‌دهد که این برآوردگرها دارای انحراف معیار کمتر و کارایی نسبی بالاتر از یک هستند، که به معنای برتری آن‌ها نسبت به برآوردگر لیو است؛ یافته‌ای که با نتایج شبیه‌سازی نیز هم‌خوانی دارد.

جدول ۲: برآورد، انحراف معیار sd و کارایی نسبی برآوردگرها برای داده‌های سرطان ریه.

CLPTE		RCLE		CLE		
sd	برآورد	sd	برآورد	sd	برآورد	متغیرها
۰٫۲۷۴	−۰٫۵۱۲	۰٫۲۶۴	−۰٫۴۹۵	۰٫۲۸۴	−۰٫۵۵۱	جنسیت
۰٫۰۲۳	۰٫۰۱۰	۰٫۰۱۰	۰٫۰۲۲	۰٫۰۲۴	۰٫۰۱۱	ph.ecog
۰٫۰۱۰	۰٫۳۵۲	۰٫۰۰۱	۰٫۳۴۵	۰٫۰۱۳	۰٫۳۵۶	ph.karno
۱٫۲۱۷		۱٫۲۳۲		-		RE

بحث و نتیجه‌گیری

در این مقاله به بررسی کارایی برآوردگر که براساس برآوردگر لیو در مدل کاکس با وجود همخطی بین متغیرهای کمکی پرداخته می‌شود. ضمن معرفی برآوردگر، ویژگی‌های مجانبی آن مانند اریبی و واریانس استخراج کرده و با انجام مطالعات شبیه‌سازی گسترده، عملکرد آن‌ها را در شرایط مختلف ارزیابی نمودیم. نتایج نشان داد که این برآوردگر پیشنهادی به طور پیوسته کارایی نسبی بالاتری نسبت به برآوردگر اصلی لیو کسب کرده به‌ویژه در شرایطی که همخطی چندگانه شدید باشد، وجود دارد. در نهایت، مزایای عملی رویکرد پیشنهادی با استفاده از مجموعه داده سرطان ریه NCCTG نشان داده شد. این یافته‌ها بر ارزش استفاده از برآوردگر پیش‌آزمون مبتنی بر لیو در ارائه استنباط‌های قابل اطمینان‌تر در تحلیل بقاء تأکید دارند.

بیانیه تأمین مالی

تحقیق محمد آرشی تحت حمایت مادی بنیاد ملی علم ایران INSF برگرفته شده از طرح شماره ۴۰۱۵۳۲۰ انجام شده است.

مراجع

Ahmad S., Aslam M., Ahmad S. (2023), *Extending the Liu estimator for the Cox proportional hazards regression model with multicollinearity* Commun. Stat. Simul. Comput., **53**(12), 5828–5841.

Austin PC. (2012), *Generating survival times to simulate Cox proportional hazards models with time-varying covariates*,**31**(29),:3946–3958.

Cox DR. (1972), *Regression models and life-tables (with discussion)*, J. R. Stat. Soc., B: Methodol. ;**34**(1):187–220.

Kibria B.M.G. and Saleh A. (2004), *Performance of positive rule estimator in the ill-conditioned Gaussian regression model*, Calcutta Statistical Association Bulletin, **55**, 209–240.

Magnus J.R. (1997), *The traditional pretest estimator*, Theory of Probability and Its Applications, **44**(2), 293–308.

Invastigation in PTE Asymptotic Behavior in the Present of Multicollinearity

Seyed Amirhossein Tabatabaei Shirazi, Mahdi Emadi, Mohammad Arashi, Solmaz Seifollahi

Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran

---

Abstract: The Cox proportional hazards regression model is widely used for analyzing survival data with covariates. However, its performance can deteriorate in the presence of multicollinearity, leading to unreliable estimates when using the standard partial likelihood approach. In this paper, we extend the pretest estimator based on the Liu method to improve the accuracy of coefficient estimation. We derive the asymptotic theoretical properties of the proposed estimator and evaluate its performance through extensive Monte Carlo simulations. We also demonstrate how this estimation strategy can be applied to real survival data. The results show significant improvements, highlighting the practical advantages of this estimator for researchers working in survival analysis.

Keywords: Cox proportional hazards model, Liu estimator, Multicollinearity, Pretest estimator.

Mathematics Subject Classification (2020): 62N02, 62F30.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

چگونه نظریه فرآیندهای تصادفی به مدل گسترش یافته در سنتز مدرن تکامل کمک کرد؟

دکتر حمیدرضا عریضی [۱]

[۱] **گروه روانشناسی، دانشگاه اصفهان**

چکیده: شاید مهم‌ترین نظریه‌ای که همهٔ مفاهیم انسانی و اجتماعی را تغییر داد، ظهور تکامل داروین و ژنتیک مندلی باشد که در فاصلهٔ شش سال در قرن نوزدهم ارائه شد و تفکر بشر را در مورد انسان، محیط او، همهٔ حیوانات و حتی گیاهان تغییر داد. جولیان هاکسلی در سال ۱۹۴۲ این دو نظریه را در یک الگوی ترکیبی با هم پیوند داد که نام آن را سنتز مدرن نهاد. این الگوی ترکیبی کمک کرد که این نظریه پا برجا باقی بماند. در مورد جهش تکامل دو دیدگاه لامارک و داروین (دانشمندان فرانسوی و انگلیسی) وجود دارد که با هم متفاوت است. در این مقاله، به کمک فرآیندهای تصادفی انواع جهش‌ها و طبقه‌بندی آن‌ها را تشریح می‌کنیم. این طبقه‌بندی به نظریهٔ سنتز مدرن کمک کرده است تا هم‌چنان پیوند تکامل و ژنتیک استوار بماند.

واژه‌های کلیدی: فرآیندهای تصادفی – تکامل داروین – ژنتیک مندلی – جهش – طبقه‌بندی جهش‌ها.

کد موضوع‌بندی ریاضی (۲۰۲۰): 60J70.

---

۱ کاربرد فرآیندهای تصادفی در مدل گسترش یافته سنتز مدرن تکامل

شاید فرآیندهای تصادفی در هیچ علمی به اندازه زیست شناسی (تکامل) کاربرد نداشته باشد. فرایندهای تصادفی در سه زمینه ژنوتیپ ، فنوتیپ و برازش نقش اصلی را در تبیین آن داشته است. (سارکار، ۱۹۹۲).

---

[۱] سخنران، Dr.oreyzi@edu.ui.ac.ir

۱.۱ ارتباط نظریه فرآیندهای تصادفی به مدل گسترش یافته در سنتز مدرن تکامل

داروین در کتاب دوران ساز خود منشاء انواع که در ۱۸۵۹ منتشر شد، رویکرد تکاملی خود را با سه اصل مهم انتخاب طبیعی ، انتقال صفات و فرآیندهای تصادفی شرح داد. این در هم تنیدگی با نظریه احتمال و فرآیندهای تصادفی سبب شد که آماردانان برجسته‌ای همچون فیشر (۱۹۳۰) و رایت (۱۹۳۱) به دفاع از نظریه تکامل برخیزند. این تلاش‌ها موجب شده است که احتمال و فرآیندهای تصادفی در ردیف محورهای بسیار مهمی در دفاع از نظریه تکامل قرار گیرند (سارکار، ۱۹۹۲). یکی از این تلاش‌ها در پوشش شرط‌بندی بیولوژیکی است. پوشش شرط‌بندی بیولوژیکی زمانی اتفاق می‌افتد که برازش اورگانیسم‌ها در شرایط معمولی کاهش یافته، هرچند در تبادل در شرایط پر فشار بایستی برازش خود را افزایش دهند. این مفهوم زمینه را برای بحث در مورد فرآیندتصادفی در آثار فیشر و رایت گشود (کورنینگ و کافمن، ۲۰۲۳).

موقعیتی را در نظر بگیرید که در هر سال افراد دو راهبرد تولید مثل با احتمالات a و $1-a$ را بکار ببرند. واریانس باروری (لمان و بالو، ۲۰۰۷) را با تابع F نشان می‌دهیم. راهبر تولید مثل اول به کیفیت سال بستگی دارد. در سال‌هایی که کیفیت باروری بالاست، انحراف معیار باروروی $F(1(\sigma))$ و در سال‌هایی که کیفیت باروری پایین است، انحراف معیار باروری $F(1+\sigma)$ است. راهبرد تولید مثل دوم در هر سال میزان باروری به میزان ثابتی به اندازه $F(1(c))$ کاهش می‌یابد (مقدار ثابت $c =$ و واریانس $= \sigma^2$). حال یک جهش[1] با بسامد P با راهبرد متفاوت $a + da$ را در نظر بگیرید. می‌توان تغییر در بسامد جهش در سال‌های خوب و بد را طبق فرمول (۱.۱) به‌دست آورد.

$$\begin{aligned} \Delta P_{good} &= S_{good} P(1-P) + O(da)^2 \\ \Delta P_{bad} &= S_{bad} P(1-P) + O(da)^2 \end{aligned} \tag{1.1}$$

که در آن برازش خوب و بد (w_{good} و w_{bad}) برای da (راهبرد متفاوت) در هرنوع سال با فرمول‌های (۲.۱) تعیین می‌شود.

$$\begin{aligned} w_{good} &= 1 + S_{good} \\ w_{bad} &= 1 + S_{bad}. \end{aligned} \tag{2.1}$$

در مدل رایت و فیشر وقتی طبقات افراد در تکامل در نظر گرفته می‌شوند که بر مبنای برازش یافتن با وضعیت جهش است. O توزیع افراد در طول این طبقه‌ها[2]، d مدل زمان گسسته[3] و S ضریب انتخاب است(لمان و بالو، ۲۰۰۷).

لمان و بالو (۲۰۰۷) برای میانگین استراتژی دو راهبرد خوب و بد مقدار زیر را بدست آورده‌اند: (σ^2 واریانس و مقدار ثابت c در بالا تعریف شده‌اند.)

$$da \frac{c(1-c(1-a)) - a\sigma^2}{(1-c(1-a))^2 - a^2\sigma^2}.$$

درک داروین از نظریه احتمال، متناسب با دانش زمان خودش قرن ۱۹ و پیر لاپلاس بود و هنوز اصول مهم کلموگروف پدیدار نشده بود. در فلسفه علم، پوپر بیش از همه به نقش فرآیندهای تصادفی و نظریه احتمال در نظریه تکامل داروین پرداخته است و این ارجاع به نظریه احتمال و فرآیندهای تصادفی را علاوه بر مشاهدات داروین که در کتاب منشاء انواع به آن اشاره کرده بود را ملاک علمی بودن

[1] Mutation bias
[2] Occupancy process
[3] Discrete time

نظریه تکامل دانسته است. اما داروین از طریق فرانسیس گالتون (۱۸۲۲-۱۹۱۱) که آماردان بود توانست اهمیت احتمال را برای بیان نظریه تکامل پیدا کند (گالتون، ۱۸۸۹). اساس استدلال داروین بر چهار گزاره متکی است:

۱- در هر گونه خاص میان ویژگی تک تک موجودات تنوع هست.

۲- ویژگی‌های والدین احتمالاً به فرزندانشان منتقل می‌شود.

۳- گونه‌های جانداران به طور غریزی توانایی آن را دارند که با آهنگی هندسی تکثیر شوند.

۴- به طور معمول منابع محیط برای چنین آهنگ رشدی بسنده نیست.

گزاره دوم داروین اشاره می‌کند که ویژگی‌های والدین احتمالاً به فرزندانشان منتقل می‌شود. او این آگاهی را به صورتی ناهشیار دارد. تنها شش سال بعد مندل در ژنتیک مندلی ساز و کار آن را در می‌یابد. کتاب داروین در سال انتشار به پرفروش‌ترین کتاب در انگلستان تبدیل شد و با شور فراوان مورد استقبال قرار گرفت[۴]. بیشتر افرادی که کتاب او را می‌خواندند مشاهدات او که در جزایر آمریکای لاتین (و به خصوص گالاپاگوس) رخ داده بود را جالب می‌یافتند اما به مفاهیم او در نظریه احتمالات و فرآیندهای تصادفی کمتر توجه می‌کردند. اما همین اشارات متخصصان علم آمار را از همان آغاز تا امروز برمی‌انگیخت تا مانع آن شوند که بر تابوت این نظریه[۵] آخرین میخ‌ها زده شود. امروزه غیر از شواهد زیست‌شناسی، از شبیه‌سازی آزمایش رایانه‌ای برای بررسی فرآیندهای تصادفی و احتمالات (گزاره ۲) استفاده می‌شود (مود و اسلیمن، ۲۰۱۲،تژادا-لاپورتا و همکاران، ۲۰۲۵).

در برابر این نظریه به خصوص مخالفت داروین با غایت نگری[۶] و به خصوص رویکرد افلاطونی طراحی هوشمند[۷] مقاومت جامعه مذهبی در پذیرش این نظریه را برمی‌انگیخت تا اینکه یک کشیش به نام گرگور مندل[۸] از کلیسایی در شهر برنو (اینک دومین شهر بزرگ چک) مبانی استدلالی برای دو گزاره ۱ و ۲ فراهم می‌ساخت که با پرورش لوبیا در باغ حیاط کلیسا اولین اشاره‌ها به ژن را انجام داد. یکی از ایده‌های مهم تمایز مندل بین ژن‌های غالب و مغلوب بود که چرا با وجود آنکه گاهی برخی ویژگی‌های پدربزرگ و مادربزرگ‌ها که در فرزندان بلافصل آن‌ها تجلی پیدا نکرده است در نوه‌های آن‌ها ظاهر می‌شود.

نظریه مندل نظریه بسیار مهمتری نسبت به نظریه داروین بود و هنگامی که او در ۶۱ سالگی در ۱۸۸۴ از دنیا رفت فقط تعداد اندکی که بولتن خبری کلیسا در ۱۸۶۵ را خوانده بودند از کار او آگاهی داشتند. در آن زمان نظریه داروین با ژنتیک مندلی متضاد در نظر گرفته می‌شدند. اینکه گزاره ۲ داروین به خوبی با کشف مندل تبین می‌شود کمتر مورد توجه بود. برای این تضاد بین درک داروینی و مندلی به کتاب پیتر بولر (۱۹۸۹) نگاه کنید که اینک خوشبختانه به فارسی توسط محی‌الدین غفرانی ترجمه شده است. با این حال توضیح این که چرا بین ژن‌های موجود درون جمعیت یک گونه واحد تفاوت و تنوع وجود دارد نه توسط داروین و نه توسط مندل قابل تشخیص نبود و همچنان به صورت یک معما باقی مانده بود. این متخصصان احتمال و فرآیندهای تصادفی بودند که علاوه بر اینکه پاسخی برای این معما یافتند، آن را در زمینه‌های گوناگون توسعه دادند (سود و میستلی، ۲۰۲۲). باید توجه کرد که مفهوم تصادف در زیست‌شناسی تکامل کاملاً خاص و مربوط به تطابق و سازگاری است (گارسیاپینو، ۲۰۲۴). به طور کلی مستقل از ژنوتیپ و نیز محیط که تفاوت‌های فردی در آن رخ می‌دهد، می‌توان نظریه تکامل را در قالب فرآیندهای تصادفی مطرح کرد (لنرماند و همکاران ، ۲۰۰۹) به همین دلیل فرآیندهای تصادفی در نظریه مدرن تکامل جایگاه بسیار اساسی دارد.

[۴] این کتاب چهار دهه پیش توسط نورالدین فرهیخته در ایران منتشر شد.

[۵] بسیاری از دانشمندان زمان داروین آن را نظریه‌ای می‌پنداشتند که در حال مرگ است و تقریباً تا زمان سنتز مدرن درست می‌گفتند.

[۶] Teleological

[۷] Intelligent design

[۸] George Mandle

به طور کلی فرآیندهای تصادفی به سه صورت در تبیین نظریه تکامل به کار رفته است:

۱- فرآیندهای تصادفی در جهش و تغییر [۹]

۲- فرآیندهای تصادفی در چرخه حیات [۱۰]

۳- فرآیندهای تصادفی در تغییرات محیطی [۱۱]

این سه مدل تکاملی به کمک هم می‌توانند تکامل را توضیح دهند. امروزه بیان نظریه تکامل بدون کمک فرآیندهای تصادفی بسیار دشوار است که خود کاربرد مهم این فرآیندهارا نشان می‌دهد. قبل از بیان این سه نوع فرآیند تصادفی دو مدل سنتز مدرن و گسترش سنتز مدرن که در واقع به کمک مفهوم فرآیندهای تصادفی ایجاد شده‌اند را توضیح می‌دهیم. در سنتز مدرن مفهوم رانش ژنتیکی [۱۲] که در کارهای فیشر و رایت مورد استفاده قرار گرفته بود برای ترکیب و سنتز مفهوم تکامل در نظریه داروین و ژن در نظریه مندل به کار رفت. رانش ژنتیکی هر گونه تغییری است که در جمعیت پیش آید به نحوی که فراوانی آلل‌های [۱۳] خاصی در جمعیت فزونی یابد. از این ترکیب داروین‌گرایی نوین [۱۴] پدید آمد که در آن با استناد به دانسته‌های فعلی ما درباره ژن‌ها و کروموزوم‌ها به توضیح منشأ و علل تنوع ژنتیکی می‌پردازد که این خود ماده خام انتخاب طبیعی است. سنتز مدرن بعداً با مشمول کردن فرآیندهای تصادفی در تغییرات محیطی شکل گسترش یافته سنتز مدرن را پدید آورد. این شکل گسترش یافته کمک کرد تا هر دو نظریه داروین و مندل در کنار هم رشد کنند. بنابراین فرآیندهای تصادفی زبان پیوند بین دو نظریه است. در جهش و تغییر مفهوم فرآیندهای تصادفی به این صورت است که جهش‌ها از تأثیرات فنوتیپیک [۱۵] خود به طور مستقل رخ می‌دهند و معطوف به آن نیستند. معنی این که آن‌ها تصادفی هستند این است که اعم از این که جهش برای ارگانیسم مفید یا مضر باشد، رخ می‌دهند. این مسئله که بسیاری از جهش‌ها برای ارگانیسم مخرب هستند امروزه امر شناخته شده‌ای است (نگاه کنید به آیرواکر و کیتلی (۲۰۰۷)). این موقعیت را می‌توان در مدل‌های هندسی تطابق و سازگاری [۱۶] که در آن جهش‌ها اثرات فنوتیپیک نااریب دارند یافت (اور،۲۰۰۵، مارتین و لئورماند، ۲۰۰۶). با وجود آن که میانگین اثرات برازشی جهش‌ها مخرب‌تر می‌شوند جمعیت به حالت بهینه فنوتیپی نزدیک‌تر می‌شود. یامپولسکی و استولتفوتز (۲۰۰۱) به حالتی در تکامل پرداخته‌اند که حجم جمعیت N و میزان جهش چندان بزرگ نباشد (محدود باشد). در این حالت جهش‌ها به سرعت تغییر می‌کنند و ترتیب ظهور جهش‌ها هم بسیار تغییر می‌کند. یامپولسکی و استولتفوتز به کمک فرآیندهای تصادفی، این دو عامل متغیر را سازگار با انتخاب طبیعی مدلسازی کرده‌اند. در ادامه مقاله یامپولسکی و استولتفوتز (۲۰۰۱)، دو نظریه احتمال تکامل موازی [۱۷] (اور، ۲۰۰۵) و تکامل با برازش تصادفی (جانسون و بارتون، ۲۰۰۲) به رابطه جهش‌ها و انتخاب طبیعی پرداخته‌اند.

دومین کاربرد فرآیندهای تصادفی در تکامل مربوط به چرخه حیات است. مدل‌های تکاملی به طور ضمنی به چرخه حیات افراد مربوط می‌شوند که در آن رویدادهایی مثل تولد، تولید مثل و مرگ وجود دارد. مقیاس معمول مشاهده برای زیست‌شناسانی که جمعیت‌ها را مطالعه می‌کنند رویدادهای تصادفی مستقل برای هر فرد است. مثلاً بقای هر فرد از زمان t به زمان $t+1$ را می‌توان به کمک احتمال

[۹] Stochasticity of Mutation and variation
[۱۰] Stochasticity of life histories
[۱۱] Stochasticity of environmental change
[۱۲] Genetic drift
[۱۳] Allele، هر نسخه از یک ژن، در جانداران از هر ژن دو آلل وجود دارد که از پدر و مادر به ارث می‌رسد.
[۱۴] Neo-Darwinism
[۱۵] Phenotypic effects
[۱۶] Geometrical models of adaptation
[۱۷] Parallel evolution

نشان داد (کورنینگ و کافمن، ۲۰۲۳). در عین حال تعداد فرزندان یک فرد و نیز تعداد گامت‌های یک ژنوتیپ را می‌توان با بازترکیب و جداسازی از توزیع احتمال (توزیع چندجمله‌ای و توزیع پواسون) بیرون کشید. این نوع فرآیندهای تصادفی می‌تواند منجر به تغییر تصادفی در اندازه جمعیت شود (فرآیندهای تصادفی جمعیت شناختی). در مدل‌های ژنتیک جمعیتی اثر این فرآیندهای تصادفی در چرخه حیات واریانس تغییر در بسامدهای ژنوتیپ از یک نسل به نسل دیگر را نشان می‌دهد که بزرگی اندازه آن به تعداد افراد بستگی دارد که به آن رانش ژنتیکی گویند. در این مورد دو مدل رایت– فیشر و مدل موران پرکاربردترین مدل‌ها هستند. در مدل استاندارد رایت–فیشر نسل‌های ناهمپوش و گسسته ارائه می‌شود که در هر نسل افراد فرزندانی تولید می‌کنند و می‌میرند. برای پدیدآیی نسل بعدی، تعداد (مفروض) افراد جایگزین والدین (مستقل از همدیگر) می‌شوند. که متناظر با نمونه‌گیری چندجمله‌ای[۱۸] از ژنوتیپ‌هاست. در مدل موران (۱۹۵۸) جمعیت در زمان‌های گسسته $t = ۱, ۲, ۳, \ldots$ نمایش داده می‌شود. فردی برای تولید مثل و فرد دیگری برای برای مرگ گزینش می‌شود. بنابراین این مدل جزء نمونه‌های مدل زاد و مرگ[۱۹] است که در فرآیندهای تصادفی بسیار مطالعه شده است. در این‌جا نسل‌ها همپوشانی دارند و میانگین طول زندگی یک فرد در یک جمعیت به اندازه $۲N$ شامل $۲N$ مرحله (گام) زمانی است. می‌توان نشان داد که اگر p بسامد نوع A در یک زمان مفروض باشد واریانس بسامد $۲N$ مرحله یا گام بعدی pq/N است. بنابراین در مدل موران اندازه جمعیت مؤثر $N_e = N/۲$ است. در واقع بسیاری از نتایجی که در مدل فیشر– رایت بدست آمده است را می‌توان در مدل موران با جایگزینی $N/۲$ به جای N بدست آورد (تژادا–لاپورتا و همکاران، ۲۰۲۵).

سومین فرآیند تصادفی مربوط به تغییرات محیطی است و در حالی که در فرآیندهای تصادفی چرخه حیات افراد مستقل از یکدیگر می‌باشند در تغییرات محیطی تأثیر محیط بر همه افراد در یک جمعیت وجود دارد. هرچند محیط ثابت نیست اما می‌توان در یک زمان آن را دارای ماهیت تناوبی (مثلاً تغییر فصل) دانست. گاهی عوامل محیطی (مثل آب و هوا) کمتر پیش‌بینی پذیر هستند، هرچند ممکن است در بدو امر فرآیندهای تصادفی غیر زیستی (مثل بروز آتشفشان) به ذهن آیند اما فرآیندهای تصادفی زیستی بسیار بیشتر هستند (در این زمینه می‌توان به مقاله لی اسمیت و همکاران (۲۰۲۵) که در آن کاربرد فرآیندهای زیستی در تغییرات محیطی دیده می‌شود مراجعه کرد). در مورد تغییرات محیطی مقیاس زمان[۲۰] بسیار مهم است. در این حالت نوسانات محیطی نسبت به زمان نسلی[۲۱] بسیار سریع‌تر است. اگر نوسانات فرآیند تصادفی در محاسبه میانگین افراد برای اثر فرآیندهای تصادفی بر ارگانیسم بسیار کوچک باشد، بهتر است نوسانات فرآیند تصادفی در تغییر محیطی در مقیاس‌های زمانی بزرگ مورد مطالعه قرار گیرد (سود و میستلی، ۲۰۲۲).

فرآیندهای تصادفی در فرآیند تکامل در چهار موقعیت نقش اساسی دارند که به ترتیب عبارتند از :

الف) ناسازگاری

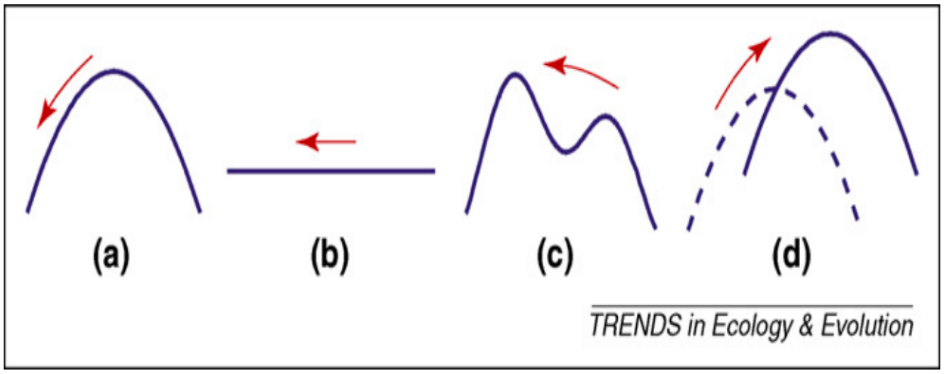
ب) سرشت تصادفی در جهش‌های خنثی

ج) انقلاب‌هایی تکاملی

د) شکل‌دهی به انتخاب طبیعی

نقش فرآیندهای تصادفی در این چهار موقعیت اساسی در شکل ۱ ترسیم شده است. نخست اینکه انتخاب تصادفی همواره در جهت سازگاری و مثبت نبوده و همه شکل‌های گفته شده در فرآیندهای تصادفی حدودی برای سازگاری تعیین می‌کنند(به کسرگلو و همکاران

[۱۸] Multinomial sampling
[۱۹] Birth death models
[۲۰] Timescale
[۲۱] Generation time



شکل ۱: فرآیندهای تصادفی در فرآیند تکامل، (a) ناسازگاری، (b) سرشت تصادفی در جهش‌های خنثی، (c) انقلاب‌های تکاملی، (d) شکل‌دهی به انتخاب طبیعی.

(۲۰۲۳) نگاه کنید)، به همین دلیل در ترجمه فارسی گاهی به جای تکامل از فرگشت استفاده می‌کنند تا بر این بازگشت (عقب‌گردها) تاکید کنند. به دلیل پیشینه سازگاری یک جهش خنثی یا تغییر تصادفی در محیط ممکن است به جمعیت‌ها آسیب رسانیده و به ناسازگاری بیانجامد.

دو عامل تعیین کننده در رانش ژنتیکی μ (میزان جهش) و S (شدت انتخاب) است و تعادل بین آن‌ها با $\frac{\mu}{S}$ بدست می‌آید اما همواره چنین نیست و در موارد زیادی رانش ژنتیکی به شکل‌های دیگری بدست می‌آید (لنرماند و همکاران ، ۲۰۰۹). در مورد دوم یعنی سرشت تصادفی در جهش‌های خنثی باید توجه کرد که گزینش محدودیت‌های فنوتیپیکی قوی اعمال می‌شوند، مثلاً می‌توان پروتئین‌هایی که در مقیاس زمانی طولانی در ارگانیسم‌های مجزا حفظ می‌شود را مورد ملاحظه قرار داد. عدم تغییر در این مقیاس زمانی با یک انتخاب گزینش قطع می‌شود. این مربوط به نظریه‌ای است که توسط استیو می گولد به نام تعادل منقطع مشهور شده است (دوران، ۲۰۲۴).

در یک حالت جهش های خنثی ممکن است نقشی مهم در تکامل بازی کنند.طول مقیاس زمانی در جهش‌های خنثی بی اهمیت است، در این حالت سرشت یک جهش را می‌توان کاملاً تابع فرآیندهای تصادفی دانست. به عبارت دیگر این جهش‌ها تنها در برخی محیط‌ها و پیشینه ژنتیکی حتمی می‌باشند و بنابراین نوعی تعادل منقطع ایجاد می‌کنند که تغییر در طول زمانی طولانی وجود ندارد اما در یک انتخاب گزینش این تعادل منقطع می‌شود. این مطالعات با واژه تعادل منقطع نام گذاری شده است که با الهام از توماس کوهن (۱۴۰۳) فلسفه علم به این‌ها اطلاق می‌شود. شرح کاملی از این مورد را در گاوریلتس (۲۰۰۳) ببینید. در مورد انقلاب‌های تکاملی منظور آن است که گاهی انتخاب طبیعی یک جمعیت را به یک اوج سازگاری موضعی می‌رساند اما از پیشرفت آن به قله‌های دیگر حتی و اگر این اوج سازگاری موضعی پائین‌تر باشد جلوگیری می‌کند. دلیل آن ایجاد اغتشاش در بسامدهای آلل‌ها می‌باشد. که رانش از یک قله به قله دیگری را پدید می‌آورد و به صورت استعاری آن را انقلاب می‌نامند. این وضعیت به‌ویژه موقعی که حجم جمعیت کوچک باشد رخ می‌دهد.

۲.۱ سوگیری جهش در یک گیاه

مفهوم فرآیندهای تصادفی در نظریه داروین اهمیت بسیار دارد، با این حال در ۲۰۲۲ یک گیاه این مفهوم را زیر سوال برد. این گیاه یک مثال نقض برای مفهوم فرآیندهای تصادفی در تکامل انواع است. مونرو و همکاران (۲۰۲۲) در شماره ۶۰۵ مجله نیچر[۲۲] در مورد

[۲۲] Nature

گیاهی به نام اربیدوپسیس تالیانا [۲۳] مقاله‌ای به چاپ رسانیدند. این گیاه اینک در پژوهش‌های گیاه‌شناسی، زیست شناسی مولکولی و ژنتیک ارزش بالایی یافته‌است چون بر خلاف طرفداران نظریه تکامل می‌توان به صورت مداخله‌ای جهش‌های غیرتصادفی (طراحی هوشمند) پدید آورد. تا به حال تعداد بسیار زیادی از این نوع جهش‌های غیر تصادفی در آن ایجاد کرده‌اند. این گیاه دارای ژنوم کوچکی به تعداد ۱۳۵ میلیون جفت باز است. اینک درباره این گیاه مقاله‌های زیادی چاپ شده است که حداقل در یک نمونه فرآیندهای تصادفی در تکامل را زیر سؤال می‌برد. با این حال فرآیندهای تصادفی یک پارادایم بسیار قوی در نظریه تکامل است که با یک نمونه، حداقل بنابر فلسفه توماس کوهن، زوال نخواهد یافت.

بحث و نتیجه‌گیری

فرآیندهای تصادفی امروزه در مطالعه نظریه تکامل کاربردهای بسیاری یافته است. نظریه‌های جدید به خصوص بعد از سنتز گسترش یافته تکامل همه به کمک فرآیندهای تصادفی توسعه یافته‌اند با این حال ارزشیابی این که چگونه شکل‌های مختلف فرآیندهای تصادفی بر تکامل صفت‌های خاصی تأثیر می‌گذارد، دشوار است.

از بین سه شکل فرآیند تصادفی نوع فرآیندهای تصادفی در جهش و تغییر که رانش ژنتیکی را توجیه می‌کند بیشترین کاربرد را دارد. این نظریه‌ها امروزه به این توافق رسیده‌اند که رانش ژنتیکی تطابق و سازگاری در انواع را محدود می‌کند. از زمان فیشر و رایت که دو آماردان برجسته انگلیسی و آمریکایی و هر دو مدافع نظریه داروین بودند و همه تلاش خود را به کار بردند تا آن را بر مبنای فرآیندهای تصادفی بیان کنند، تا کنون گرایش به سمت نظریه لامارک (که داروین با این شکل از بیان تکامل کاملاً مخالف بود) همه از فرآیندهای تصادفی برای توجیه نظریه تکامل استفاده می‌کنند. به عنوان مثال اپی ژنتیک که گرایش امروزه در تکامل است (نزدیک‌تر به نظریه لامارک است) در مقاله بیور و همکاران (۲۰۲۰) به نقش فرآیندهای تصادفی در تفاوت‌های اپی ژنتیکی پرداخته است.

مراجع

بولر، ج. پ. (۱۹۸۹)،انقلاب مندل، ترجمه محی‌الدین زعفرانی (۱۳۷۲)، شرکت انتشارات علمی و فرهنگی، تهران.

کوهن، ت. (۱۴۰۳)، ساختار انقلاب‌های علمی، ترجمه سعید زیبا کلام، انتشارات سمت، تهران.

Biwer, C., Kawam, B., Chapelle, V., and Silvestre, F. (2020), The role of stochasticity in the origin of epigenetic variation in animal populations,*Integrative and Comparative Biology*, **60(6)**, 1544-1557.

Corning, P. A. , Kaufman, S. A. (2023), *Evolution on Purpose*, The Mit Press, Cambridge,MA01239.

Duran, N. (2024), The many ways toward punctuated evolution, *Palaeontology*, **67(5)**, 12731-12736.

Eyre-Walker, A., and Keightley, P. D. (2007), The distribution of fitness effects of new mutations, *Nature Reviews Genetics*, **8(8)**, 612-618.

Fisher, R. A. (1930),*The genetical theory of natural selection*, The Clarendon Press, Oxford, England.

[۲۳] Arabidopsis thaliana

Galton, F. (1889), *Natural inheritance*, Macmillan, London.

García-Pintos, L. P. (2024). Limits on the evolutionary rates of biological traits. *Scientific Reports*, **14(1)**, 11314.

Gavrilets S. (2003), Models of Speciation: What Have We Learned in 40 Years?, *Evolution; International Journal of Organic Evolution*,**57**, 2197–2215.

Johnson, T., and Barton, N. H. (2002), The effect of deleterious alleles on adaptation in asexual populations.*Genetics*, **162(1)**, 395-411.

Keseroglu, K., Zinani, O. Q., Keskin, S., Seawall, H., Alpay, E. E., and Özbudak, E. M. (2023), Stochastic gene expression and environmental stressors trigger variable somite segmentation phenotypes, *Nature communications*, **14(1)** , 6497-6501.

Lea-Smith, D. J., Hassard, F., Coulon, F., Partridge, N., Horsfall, L., Parker, K. D. and Krasnogor, N. (2025), Engineering biology applications for environmental solutions: Potential and challenges, *Nature Communications*, **16(1)**, 3538-3547.

Lehmann, L., and Balloux, F. (2007), Natural selection on fecundity variance in subdivided populations: kin selection meets bet hedging.,*Genetics*, **176(1)**, 361-377.

Lenormand, T. ,Roze, D. ,Rousset. (2009) Stochasticity in Evolution, *Trends in Ecology and Evolution*, **24(3)** ,157-165.

Monroe, J.G. ,Srikant, T., Weigle, D. (2022), Mutation Bias Reflects Natural Selection in Arabidopsis Thaliana, *Nature*, **602** ,101-105.

Mode, C. J., and Sleeman, C. K. (2012), *Stochastic processes in genetics and evolution: computer experiments in the quantification of mutation and selection*, World Scientific,Singapore (SG).

Moran P. (1958), Random Processes in Genetics,*Proc. Camb. Philos. Soc.*, **54**, 60–71.

Orr, H. A. (2005), The probability of parallel evolution, *Evolution* , **59(1)**, 216-220.

Sarkar, S. (1992), The founders of Evolutionary Genetics,*Boston studies in the Philosophy and History of Science* , **142**,89-116.

Sood, V., and Misteli, T. (2022), The stochastic nature of genome organization and function.*Current opinion in genetics and development* , **72**, 45-52.

Tejada-Lapuerta, A., Bertin, P., Bauer, S., Aliee, H., Bengio, Y. and Theis, F. J. (2025), Causal machine learning for single-cell genomics, *Nature Genetics*, **20**, 1–12.

Wright, S. (1931), Evolution in Mendelian populations, *Genetics*, **16(2)**, 97-159.

Yampolsky, L. Y., and Stoltzfus, A. (2001), Bias in the introduction of variation as an orienting factor in evolution, *Evolution and development*, **3(2)**, 73-83.

How did the theory of stochastic processes contribute to the model developed in the modern synthesis of evolution?

Dr. Hamidreza Oreizi[1]

[1]Department of Psychology, University of Isfahan

Abstract: Perhaps the most important theory that changed all human and social concepts was the emergence of Darwinian evolution and Mendelian genetics, which were presented within six years of each other in the 19th century and changed human thinking about humans, their environment, all animals and even plants. In 1942, Julian Huxley linked these two theories in a combined model that he called the modern synthesis. This combined model helped this theory to survive. In this article, we explain mutation, which is viewed differently in the views of Darwin and Lamarck (English and French scientists), using random processes, and we explain the types of mutations and their classification using random processes. This classification has helped the modern synthesis theory to continue to maintain the link between evolution and genetics.

Keywords: Stochastic processes – Darwinian evolution – Mendelian genetics – Mutation – Classification of mutations.

Mathematics Subject Classification (2020): 60J70.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران
B۰۶۶A"۰۶۶"۹F۰۶"۸F۰۶"۷F۰۶"۶F۰۶"۵F۰۶"۴F۰۶"۳F۰۶"۲F۰۶"۱F۰۶"۰F۰۶"

برآورد معیار $CoVaR$ بر اساس رویکرد $Copula - GARCH$ و توزیع GED

فاطمه علیزاده [۱] محمد امینی [۲] ، غلامرضا محتشمی برزادران [۳] ،

گروه آمار، دانشگاه فردوسی مشهد

چکیده: ریسک یکی از مفاهیم پایه ای در بازارهای مالی می‌باشد و اندازه گیری و تحلیل آن بسیار اهمیت دارد. در این مقاله یکی از مهم ترین روش های اندازه گیری ریسک یعنی ارزش در معرض خطر شرطی را در حالت دو متغیره با استفاده از دو سری زمانی بازدهی مطالعه می‌کنیم در ادامه ساختار وابستگی این بازدهی‌ها را بر اساس تابع مفصل مورد بررسی قرار داده و در سری زمانی برای پیش بینی واریانس از مدل‌های گارچ با توزیع حاشیه‌ای GED استفاده می‌کنیم و در پایان روش مورد نظر در داده واقعی به کار گرفته می‌شود.

واژه‌های کلیدی: ریسک سیستمی، ارزش در معرض خطرشرطی، مدل گارچ، تابع مفصل، توزیع GED .

کد موضوع‌بندی ریاضی (۲۰۲۰): 62P05، 62M10، 62H05، 91B84.

---

۱ مقدمه

با گذشت زمان به دلیل گسترش حوادث نامطلوب مختلف در جهان که بخشی از آن از افزایش فعالیت‌های اقتصادی، اجتماعی، سیاسی و ... نشئت می‌گیرد، نااطمینانی درباره آینده بیشتر شده است. ریسک یکی از مفاهیم پایه‌ای در بازار های مالی است که تعریف واحدی از آن وجود ندارد. اقتصاددانان، دانشمندان علوم رفتاری، نظریه پردازان ریسک و آمارشناسان هر یک تصور خاصی از ریسک دارند، البته به طور سنتی ریسک به عنوان عدم اطمینان تعریف شده است و بر اساس این مفهوم، در این‌جا ریسک به عنوان عدم اطمینان مرتبط با وقوع یک خسارت و یا زیان تعریف می‌شود.

در سال‌های اخیر با بررسی علل نوسانات مالی مشخص می‌شود که عوامل مختلفی در ایجاد نوسانات در بازارهای مالی (بیمه‌ها، بانک‌ها و ...) مؤثر بوده‌اند از جمله بحران نفتی سال ۱۹۷۳، بحران‌های مالی سال‌های ۱۹۹۷-۱۹۹۸ در جنوب شرق آسیا و رویدادهای طبیعی نظیر زلزله و سونامی. بروز چنین حوادثی و به تبع آن نوسانات بازارهای مالی موجب ایجاد اختلال در فعالیت‌های سازمان‌های مالی، تجاری و حتی تولیدی می‌شود. از این رو اندازه‌گیری و ارزیابی ریسک اهمیت پیدا می‌کند و از جمله مهم ترین کاربردهای آن می‌توان

---

[۱] فاطمه علیزاده، fateme.alizade301073@gmail.com

به قیمت‌گذاری سهام‌ها، تعیین حق بیمه متناسب با ریسک و تعیین سرمایه اقتصادی [۱] شرکت که پشتوانه‌ی فعالیت‌های مالی است و خسارت‌های ناگهانی بالاتر از حد انتظار را پوشش می‌دهد، اشاره کرد.

تاکنون روش‌های متعددی برای اندازه‌گیری ریسک معرفی شده است که هر یک از این روش‌ها مزایا و معایب خاص خود را دارد. یکی از مهم ترین معیارهای اندازه گیری ریسک ارزش در معرض خطر [۲] است. این شیوه اندازه گیری را ابتدا تیل گلدیمن [۳] در سال ۱۹۸۰ ارائه کرد و در اواخر دهه ۱۹۸۰ توسط موسسه جی–پی مورگان [۴] گسترش یافت. این شاخص حداکثر زیان ممکن را در یک افق زمانی مشخص با توجه به یک سطح احتمال معین α بیان می‌کند و با نماد VaR_α نشان داده می‌شود، به عنوان مثال ارزش در معرض خطر با سطح احتمال ۰/۰۵ برای یک بازه زمانی یک روزه گویای اینست که حداکثر زیان تحمیلی طی روز بعدی با احتمال ۵ درصد از مقدار ارزش در معرض خطر بالاتر می‌رود. این معیار را می‌توان با استفاده از سری بازدهی محاسبه کرد. فرض کنید $\{p_t\}$ سری زمانی بهای نوعی دارایی مالی باشد، بازدهی در زمان t ام به صورت $X_t = log(p_t) - log(p_{t-1})$ تعریف می‌شود.

تعریف ۱.۱. اگر $\{X_t\}$ یک سری بازدهی با تابع توزیع F_{X_t} باشد، آن‌گاه

$$VaR_\alpha(X_t) = F_{X_t}^{-1}(\alpha) = inf\{x; F_{X_t}(x) \geq \alpha\} \ ; \ \alpha \in (0, 1).$$

رویدادهایی که در نهادهای مالی رخ می‌دهد ممکن است برسیستم مالی سایر نهادها و یا حتی بر کل سیستم اقتصاد اثر بگذارد در این صورت نمی‌توان تاثیر متقابل بین ریسک‌ها را نادیده گرفت به همین دلیل از زمان شروع بحران‌های مالی و اقتصادی در سال ۲۰۰۷ ریسک سیستمی [۵] یعنی ریسکی که بحران در یک نهاد، بر سایر نهادها یا کل سیستم مالی تاثیر می‌گذارد مطرح شد. با مروری بر بحران های مالی در جهان ملاحظه می‌شود مهمترین خطری که ثبات مالی را تهدید می‌کند ریسک سیستمی ناشی از بحران‌های سیستم بانکی است. این بحران‌های بانکی اخیر در جهان، هزینه‌های زیادی برای آن کشورها بهمراه داشته و نگرانی‌هایی را برای نظام مالی آنها به وجود آورده است. لذا به مطالعه ریسک سیستمی در بانک ها توجه زیادی شده و با معیارهای مختلفی ریسک سیستمی در بانک ها و مؤسسات مالی جهان ارزیابی گردیده است. یکی از روش‌های اندازه گیری ریسک سیستمی، ارزش در معرض خطر شرطی [۶] است که اولین بار توسط آدرین و برونرمیر (۲۰۱۱) [۷] معرفی شد و پس از آن جیراردی و ارگان (۲۰۱۳) [۸] تعریف دیگری از آن ارایه دادند. این معیار ارزش در معرض خطر یک سازمان با بازدهی Y را در صورتی که سازمانی دیگر با بازدهی X در بحران قرار داشته باشد بدین معنی که بازدهی آن حداکثر برابر ارزش در معرض خطرش در سطح احتمال α باشد، در سطح احتمال β را محاسبه می‌کند، یعنی

$$CoVaR_{\alpha,\beta}(Y|X) = VaR_\beta(Y \mid X \leq VaR_\alpha(X)); \ \alpha, \beta \in (0, 1), \tag{۱.۱}$$

که به آن ارزش در معرض خطر شرطی ($CoVaR$) گویند. در صورتی که بازدهی های و وابسته باشند می‌توان از توابع مفصل برای بررسی ساختار وابستگی بین بازدهی‌ها استفاده کرد. از جمله مطالعاتی که در زمینه برآورد این معیارها با استفاده از تابع مفصل انجام شده است می‌توان به موارد زیر اشاره کرد.

[1] Economic Capital
[2] Value at Risk
[3] Till Guldimann
[4] J.P. Morgan
[5] Systemic Risk
[6] Conditional Value-at-risk
[7] Adrin and Brunnermeier
[8] Girardi and Ergun

ماینیک و اسکنینگ (۲۰۱۴) معیار $CoVaR$ را بر اساس توابع مفصل گاوسی و استودنت، ربوردو و یوگولینی (۲۰۱۵) در مفصل‌های تغییر شکل یافته گامبل و برناردی و همکاران (۲۰۱۷) نیز برای توابع مفصل کران‌های فرشه، مارشال اولکین، FGM و برخی مفصل‌های ارشمیدسی یعنی گامبل، فرانک، جو و کلیتون تحلیل کرده اند. کاریمالیس و نومیکس (۲۰۱۸)، $CoVaR$ را با توجه به نتایج برناردی و همکاران (۲۰۱۷) و مدل‌های سری زمانی گارچ بررسی و سپس ریسک سیستمی در سیستم بانکی اروپا را با توابع مفصل فرانک و جو محاسبه کرده اند. جی و همکاران (۲۰۱۹) $CoVaR$ را با استفاده از مفصل‌های زمان متغیر برای قیمت نفت و دلار، هوانگ و همکاران (۲۰۲۴) به روش شبیه‌سازی مونت کارلو و مفصل، کیلمن و همکاران (۲۰۲۲) و وانگ و همکاران (۲۰۲۵) با مفصل‌های واین طالعه کرده اند.

۱.۱ تابع مفصل

در سال‌های اخیراستفاده از توابع مفصل [9] به عنوان ابزاری مفید برای مدل‌سازی ساختار وابستگی بین دو یا چند متغیر گسترش یافته است و کاربرد بسیاری در علوم بیمه و اقتصاد دارد. مفصل‌ها توابعی هستند که بین توابع توزیع چند متغیره و توزیع حاشیه ای آن ها ارتباط برقرار می‌کنند. از طرفی مفصل خود تابع توزیعی است که توزیع حاشیه‌ای آن از توزیع یکنواخت استاندارد پیروی می‌کند.

قضیه ۲.۱ که با نام قضیه اسکلار معروف است رابطه بین توزیع‌های حاشیه ای، تابع مفصل و توزیع توام دو متغیر تصادفی وابسته را بیان می‌کند.

قضیه ۲.۱. اگر (X,Y) یک بردار تصادفی با تابع توزیع توام $F(.,.)$ و توابع توزیع حاشیه ای F_X و F_Y باشند آنگاه مفصل $C : I^2 \longrightarrow I$ وجود دارد به طوریکه:

$$F(x,y) = C(F_X(x), F_Y(y)) \ ; \ \forall x, y \in R.$$

اگر F_X و F_Y پیوسته باشد آنگاه C یکتاست در غیر این صورت به صورت یکتا روی $Range\, F_X \times Range\, F_Y$ تعیین می شود و با داشتن توزیع های حاشیه‌ای و توام برای یافتن مفصل متناظر قرار می‌دهیم: $u = F_X(x)$ ، $v = F_Y(y)$ و داریم :

$$C(u,v) = F(F_X^{-1}(u), F_Y^{-1}(v)) \ ; \ \forall u, v \in [0, 1].$$

خانواده‌ای از مفصل‌ها به نام مفصل‌های ارشمیدسی، وجود دارند که با استفاده از یک تابع به نام تابع مولد $\phi : [0, 1] \to [0, \infty)$ که در آن $\phi(1) = 0$ و $\phi(0) = \infty$ می‌باشد به شکل زیر ساخته می‌شوند:

$$C(u,v) = \phi^{-1}(\phi(u) + \phi(v)) \tag{۲.۱}$$

۲.۱ مدل بندی تغییر پذیری

مدل‌بندی تغییرپذیری که اغلب به صورت واریانس تعریف می‌شود برای پیش بینی نوسانات آینده یکی از مباحث مهم در مطالعات مالی می‌باشد. اغلب مشاهده می‌شود که بازدهی روزانه سهام بعد از یک دوره تغییرات شدید بها، واریانس شرطی بزرگتری نسبت به یک دوره نسبتا پایدار دارند. مدل‌های $ARCH$ و $GARCH$ [10] برای فرایند واریانس شرطی که با آن بتوانیم تغییر پذیری مقادیر آینده را بر

[9] Copula Function

[10] Generalized Auto regressive Conditional Heteroscedasticity

اساس داده های جاری و گذشته پیشگویی کنیم، در مدل ARCH(q) فرض می‌شود که سری بازدهی $\{X_t\}$ به صورت

$$X_t = \mu + a_t \ , \ \ a_t = \sigma^2_{t|t-1}\varepsilon_t, \tag{۳.۱}$$

$$\sigma^2_{t|t-1} = \omega + \alpha_1 a^2_{t-1} + \alpha_2 a^2_{t-2} ... + \alpha_q a^2_{t-q}. \tag{۴.۱}$$

تولید می‌شود که در آن پارامترهای α_i و ω مجهولند، $\{\varepsilon_t\}$ دنباله ای از متغیرهای تصادفی مستقل و هم توزیع با میانگین صفر و واریانس واحد است و ε_t برای $j = 1, 2, ...$ مستقل از a_{t-j} است.

در رویکردی دیگر علاوه بر q تاخیر از توان دوم a_t ، p تاخیر از واریانس شرطی را هم وارد مدل می‌کند. یعنی،

$$\sigma^2_{t|t-1} = \omega + \alpha_1 a^2_{t-1} + ... + \alpha_q a^2_{t-q} + \beta_1 \sigma^2_{t-1|t-2} + ... + \beta_p \sigma^2_{t-p|t-p-1}. \tag{۵.۱}$$

۲ ارزش در معرض خطر شرطی

ماینیک و اسکنیگ (۲۰۱۴) برای دو متغیر بازدهی X و Y با توابع توزیع حاشیه‌ای $F_X(x)$ و $F_Y(y)$ ساختار وابستگی تابع مفصل $C(u, v)$ و توزیع توأم $F(x, y)$ ارزش در معرض خطر شرطی به ازای هر $\alpha, \beta \in [0, 1]$ به شرح زیر به دست آوردند:

$$CoVaR_{1-\alpha,1-\beta}(Y_t \mid X_t) = F_{Y_t}^{-1}(F^{-1}_{V|U\leq\alpha}(\beta)) = VaR_{F^{-1}_{V|U\leq\alpha}(\beta)}(Y_t) \tag{۱.۲}$$

که در آن (U, V) دارای توزیع توأم $C(u, v)$ با توابع توزیع حاشیه‌ای یکنواخت استاندارد می‌باشند و

$$F_{V|U\leq\alpha}(v) = \frac{C(\alpha, v)}{\alpha}. \tag{۲.۲}$$

علاوه بر این برناردی (۲۰۱۷) نشان داد با در نظر گرفتن $w^* = F^{-1}_{V|U\leq\alpha}(\beta)$ ، و استفاده از (۲.۲) ، w^* از طریق حل معادله زیر به دست می‌آید:

$$C(\alpha, w^*) = \alpha\beta. \tag{۳.۲}$$

سپس w^* را در مفصل‌های کران‌های فرچت، مارشال اولکین، FGM و مفصل های ارشمیدسی فرانک، کلیتون و گامبل محاسبه کرده است.

حال در حالت خاص اگر $C(u, v)$ یک مفصل ارشمیدسی با تابع مولد $\varphi(t)$ باشد با محاسبات جبری ساده داریم:

$$w^* = \varphi^{-1}(\varphi(\alpha\beta) - \varphi(\alpha)), \tag{۴.۲}$$

در نتیجه

$$CoVaR_{1-\alpha,1-\beta}(Y_t \mid X_t) = VaR_{w^*}(Y_t). \tag{۵.۲}$$

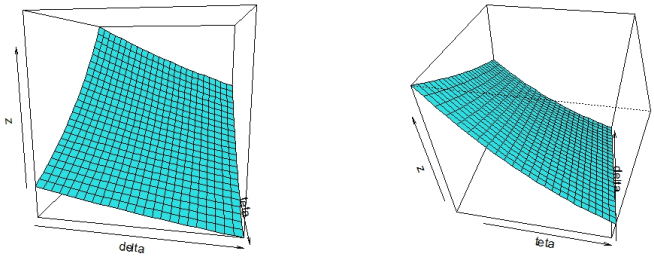
:

حال اگر X و Y دارای ساختار وابستگی مفصل $BB1$ با تابع مولد $\varphi(t) = (t^{-\theta} - 1)^\delta$ باشد، آنگاه معکوس تابع مولد به صورت $\varphi^{-1}(s) = (s^{\frac{1}{\delta}} + 1)^{-\frac{1}{\theta}}$ و با استفاده از رابطه (۴.۲) داریم:

$$w^* = [((\alpha\beta)^{-\theta} - 1)^\delta - (\alpha^{-\theta} - 1)^\delta)^{\frac{1}{\delta}} + 1]^{-\frac{1}{\theta}}$$

(۶.۲)

در شکل ۱ تغییرات w^* براساس تغییرات θ و δ برای $\alpha = \beta = ۰/۰۵$ رسم شده است که در آن مشاهده می‌شود w^* با ثابت نگاه داشتن یکی از پارامترهای θ و δ نسبت به پارامتر دیگر نزولی می‌باشد و از آنجایی که معکوس تابع توزیع تابعی صعودی نسبت به سطح احتمال (w^*) است از این رو با افزایش هر یک از مقادیر θ و δ مقدار $CoVaR$ کاهش پیدا می‌کند. اکنون فرض کنید سری زمانی



شکل ۱: نمودار w^* براساس تغییرات θ و δ مفصل$BB۱$

بازدهی به صورت $Y_t = \mu_t + \sigma_{t|t-۱}\varepsilon_t$ باشد. ابتدا مناسب‌ترین مدل سری زمانی بازدهی Y_t را به دست آورده، پارامترهای مدل سری زمانی، پارامتر توزیع خطای تعمیم یافته و میانگین (μ_t) و انحراف معیار ($\sigma_{t|t-۱}$) زمان آتی را برآورد کرده و برای براورد ارزش در معرض خطر شرطی در سطح احتمال w^* که به روش زیر بدست می‌آید عمل می‌کنیم:

$$P(Y_t < CoVaR) = P(\mu + \sigma_{t|t-۱}\varepsilon_t < CoVaR)) = w^* \Rightarrow$$
$$P(\varepsilon_t < \frac{CoVaR - \mu}{\sigma_{t|t-۱}}) = w^* \Rightarrow$$
$$CoVaR = \mu + \sigma_{t|t-۱}F^{-۱}_{\varepsilon_t}(w^*).$$

که در آن وقتی خطاها از توزیع GED تبعیت کنند تابع چگالی آن به صورت $f_{\varepsilon t}(x) = \frac{\beta}{۲\Gamma(\frac{۱}{\beta})}e^{-|x|^\beta}$ است. برای $w^* < \frac{۱}{۲}$ با تغییر متغیرهای $y = (-x)^\beta$ و $z = x^\beta$ داریم:

$$F_{\varepsilon_t}(\eta) = \frac{\beta}{۲\Gamma(\frac{۱}{\beta})}(\int_۰^\infty \frac{۱}{\beta}y^{\frac{۱}{\beta}-۱}e^{-y}dy + \int_۰^{\eta^\beta} \frac{۱}{\beta}z^{\frac{۱}{\beta}-۱}e^{-z}dz),$$

و با تعاریف تابع گاما و گامای ناقص پایینی می‌توان نوشت:

$$F_{\varepsilon_t}(\eta) = \frac{۱}{۲\Gamma(\frac{۱}{\beta})}(۲\Gamma(\frac{۱}{\beta}) - \Gamma(\frac{۱}{\beta}, \eta^\beta)). \tag{۷.۲}$$

و $CoVaR$ از طریق حل معادله زیر بدست می‌آید.

$$۱ - \frac{\Gamma(\frac{۱}{\beta}, CoVaR^\beta)}{۲\Gamma(\frac{۱}{\beta})} = w^*. \tag{۸.۲}$$

و برای $w^* > \frac{۱}{۲}$ با استفاده از رابطه

$$F_{\varepsilon_t}(\eta) = \frac{\Gamma(\frac{۱}{\beta}, (-\eta)^\beta))}{۲\Gamma(\frac{۱}{\beta})}. \tag{۹.۲}$$

از طریق حل معادله

$$\frac{\Gamma(\frac{۱}{\beta},(-CoVaR)^{\beta})}{۲\Gamma(\frac{۱}{\beta})}=w^{*}, \qquad (۱۰.۲)$$

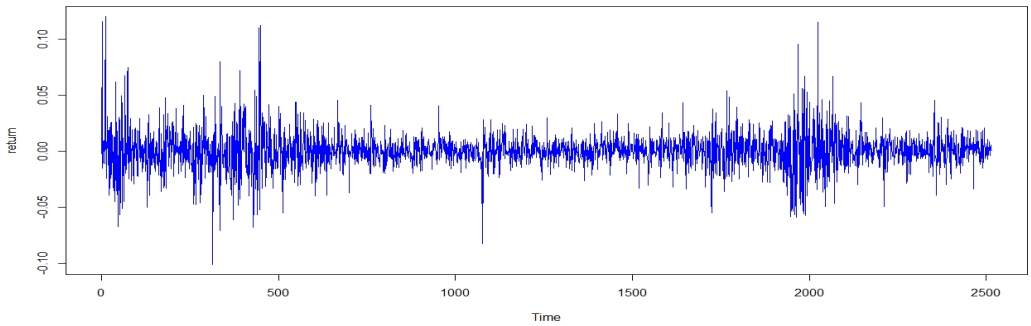
برای حل این معادلات نیازمند روش‌های عددی می‌باشیم.

۳ کاربرد در داده واقعی

در این بخش داده‌های بازدهی روزانه شرکت‌های IBM و SP از سال ۲۰۰۲ تا ۲۰۱۰ مورد بررسی قرار گرفته است.

ابتدا برای محاسبه مقدار ساختار وابستگی بین بازدهی‌ها را مورد بررسی قرار می‌دهیم. با مقایسه تاو کندال، وابستگی دمی بالا، پایین و معیارهای AIC و BIC توابع مفصل مختلف برازش داده شده مفصل $BB۱$ مفصل مناسبی برای بیان ساختار وابستگی می‌باشد، پی–مقدار آزمون نیکویی برازش کرامر ون مایسز و کولموگروف اسمیرنوف بیشتر از پنج صدم بنابراین فرض مناسب بودن مفصل $BB۱$ را رد نمی‌کنیم.

برآورد پارامترهای این مفصل به روش ماکزیمم درستنمایی $\theta=۵/۷۶$ و $\delta=۰/۳۲$ بدست آمد بنابراین مقدار w^{*} برای $\alpha=\beta=$ ۰/۰۵ بر اساس رابطه ۶.۲ تقریبا ۰/۰۷۲ محاسبه شد .

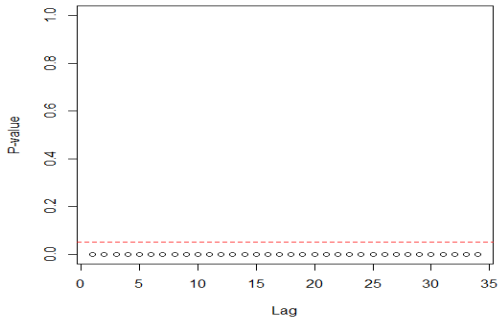


شکل ۲: بازدهی روزانه سهام IBM

شکل ۲ از سری بازدهی IBM نشان می‌دهد که در برخی از دوره ها بسیار ناپایدار شدند به عبارت دیگر واریانس شرطی سری زمانی در طی زمان تغییر می‌کند، به همین منظور در ادامه به بررسی مدل‌های آرچ و گارچ می‌پردازیم.

در مورد داده های بازدهی سهام IBM برای بررسی وجود اثرهای ARCH با رسم پی–مقدارهای آزمون مک‌لیود–لی در شکل ۳ مشاهده می‌شود که همگی در سطح ۵ درصد معنادارند و به طور رسمی شواهدی قوی به نفع ARCH در داده ها را نشان می‌دهد.

با بررسی سری توان دوم و قدرمطلق داده‌های زیان سهام IBM ، نمودارهای خودهمبستگی، (acf)، خودهمبستگی جزیی، $(pacf)$ و خصوصا خودهمبستگی گسترده $(eacf)$ یک مدل $ARMA(۱,۱)$ را گزارش می‌کند، بنابراین می‌توان یک مدل $GARCH(۱,۱)$ را برای داده های اصلی در نظر گرفت.



شکل ۳: پی–مقدارهای آزمون مک‌لیود–لی برای زیان روزانه سهام IBM

اگر چه نشانه ای از یک مدل $GARCH(۲,۲)$ نیز وجود دارد، بنابراین با استفاده از روش‌های مشخص سازی مثل AIC و BIC $GARCH(۱,۱)$ را انتخاب می‌کنیم. پیش از آن که مدل برازنده شده‌ای را بپذیریم کار اساسی این است که وارسی کنیم آیا مدل به درستی مشخص شده یا نه. اگر مدل به درستی مشخص شده باشد مانده های استاندارد شده یک سری نوفه سفیدند. در این مورد در نمودارهای ACF و $PACF$ مانده های استانداردشده و توان دومشان هیچ کدام از مقادیر خارج از بازه نیستند که این پیشنهاد می‌کند سری باقی مانده ها نوفه سفید است، علاوه بر این نتایج آزمون های کلی نگر هم حاکی از مدل نوفه سفید برای باقیمانده های استاندارد شده است. برای باقی‌مانده ها توزیع‌های خطای تعمیم یافته، نرمال و استودنت مورد بررسی قرار گرفت که توزیع خطای تعمیم یافته با مقایسه معیارهای AIC و MSE عملکرد بهتری داشت.

پس از برآورد ضرایب مدل سری زمانی مقدار $CoVaR$ شرکت IBM در صورت بحران در شرکت SP و محاسبه مقدار w^* برای $\alpha = \beta =$ ۰٫۰۵ که بر اساس رابطه ۶.۲ تقریبا ۰٫۰۷۲ محاسبه شد بر اساس مدل سری زمانی $GARCH(۱,۱)$ و توزیع حاشیه‌ای GED با پارامتر ۳٫۱۷۴ مقدار $CoVaR$ برابر با ۲٫۶۴۱− بدست آمد.

۴ نتیجه‌گیری

در این مقاله یکی از روش‌های مهم بررسی ریسک سیستمی یعنی ارزش در معرض خطر شرطی بر اساس تابع مفصل ارشمیدسی $BB۱$ و مدل‌های سری زمانی گارچ در صورتی که ماندهای آن از توزیع GED پیروی کند، مورد مطالعه قرار گرفت و در پایان این معیار برای شرکت IMB در صورتی که شرکت SP در شرایط بحرانی قرار گیرد محاسبه شد که در آن توابع مفصل مختلف برای بررسی ساختار وابستگی بازدهی مدل‌های سری زمانی گارچ برای پیش بینی نوسان روز آینده شرکت IMBبررسی شد، نتایج حاکی از آن بود که مفصل $BB۱$ و سری زمانی $GARCH(۱,۱)-$ بهترین مدل‌بندی را ارایه می‌دهند.

مراجع

[1] Adrian, T., Brunnermeier, M.K., (2016). CoVaR, *Amer. Econ. Rev.*, **106**, 1705–1741.

[2] Bernardi, M., Durante, F., Jaworski, P., (2017). CoVaR of families of copulas. *Statist.Probab. Lett.* **120**, 8–17.

[3] Girardi, G., Ergün, A.T., (2013). Systemic risk measurement: multivariate GARCH estimation of CoVaR. *J. Bank. Finance*, **37**, 3169–3180.

[4] Huang, W., Lin, N., Hong, L. J. (2024). Monte Carlo Estimation of CoVaR, *Operations Research,* **72(6)**, 2337-2357.

[5] Ji, Q., Bouri, E., Roubaud, D., Shahzad, S. J. H. (2018). Risk spillover between energy and agricultural commodity markets: A dependence-switching CoVaR-copula model, *Energy Economics*, **75**, 14-27.

[6] Karimalis, E. N., Nomikos, N. K., (2018). Measuring systemic risk in the European banking sector: a copula CoVaR approach, *The European Journal of Finance*, **24(11)**, 944-975.
MReboredo, Juan C., Ugolini, A., (2015). Systemic risk in European sovereign debt markets: A CoVaR-copula approach,em *Journal of International Money and Finance*, **51**, 214–244.

[7] Nelsen, R.B. (2006). *An Introduction to Copulas*, vol. 139. Springer Science & Business Media.

[8] Wang, J., Yan, X., Cao, Y., Wang, X. (2025), Multi-scale dependence and risk contagion among international financial markets based on VMD-Vine copula-CoVaR, *Applied Economics*, **57(6)**, 658-677.

Estimating CoVaR Based on Copula-GARCH method and GED Distribution

Fatemeh Alizadeh[1]**,Mohammad Amini** [3]**, Gholam Reza Mohtashami Borzadaran**[2]

Department of Statistics, Ferdowsi Univercity of Mashhad

Abstract: Risk is one of the basic concepts in financial markets and its measurement and analysis is very important. In this article, we study one of the most important methods of measuring risk, namely conditional value at risk, in the bivariate case using two time series of returns. Next, we examine the dependence structure of these returns based on the detailed function and use GARCH models with marginal distributions (GED) in the time series to predict the variance. Finally, the method is applied to real data.

Keywords: Sistemic Risk,Conditional Value at Risk, GARCH Model, Copula Function, GED Distribution.

Mathematics Subject Classification (2020): 62P05, 62M10, 62H05, 91B84.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مروری بر استفاده از خانواده توزیع‌های فاز–نوع در مدل شوک‌های تعمیم یافته

سیروس فتحی منش [۱] ، محی‌الدین ایزدی[۲]

[۱] گروه آمار– دانشگاه کردستان

[۲]گروه آمار– دانشگاه رازی

چکیده: استفاده از خانواده توزیع‌های فاز–نوع که بر اساس فرایندهای مارکف زمان گسسته و پیوسته تعریف می‌شوند در مدل‌بندی متغیرهای تصادفی نامنفی گسسته و پیوسته رو به گسترش است. در دهه اخیر، استفاده از این خانواده توزیع‌ها در برخی مدل‌های شوک مورد توجه پژوهشگران زمینه قابلیت اعتماد قرارگرفته است. در این مقاله هدف ما مروری بر نحوه استفاده از خانواده توزیع‌های فاز–نوع در سه مدل شوک کرانگین تعمیم‌یافته، تجمعی تعمیم‌یافته و دلتای تعمیم یافته است. استفاده از این خانواده از توزیع‌ها در مدل شوک‌های ذکرشده توجیه می‌شود و سپس توزیع تعداد شوک‌ها و طول عمر سیستم بر اساس روابط ماتریسی موجود در خانواده توزیع‌های فاز–نوعی بیان می‌شوند.

کلمات کلیدی : خانواده توزیع‌های فاز–نوع، فرایندهای مارکف، مدل شوک تجمعی تعمیم یافته، مدل شوک دلتای تعمیم یافته، مدل شوک کرانگین تعمیم یافته.

---

۱ مقدمه

در مدل‌های آماری استفاده از خانواده توزیع‌های منعطف که با تغییر پارامتر، دامنه وسیعی از توزیع‌ها را شامل شود بسیار حائز اهمیت است، زیرا اشتباه در انتخاب توزیع برای برازش به یک مجموعه داده متناسب با وسعت دامنه آن خانواده، کاهش می‌یابد. خانواده توزیع‌های فاز–نوعی [۱] یک خانواده گسترده از توزیع‌های آماری است که برای هر دو حالت گسسته و پیوسته قابل تعریف می‌باشند. این خانواده به‌طور گسترده‌ای در مدل‌بندی متغیرهای تصادفی نامنفی در زمینه‌های مختلفی مانند قابلیت اعتماد، نظریه صف، ریسک‌های بیمه‌ای و مالی، زیست شناسی و سایر علوم مورد استفاده قرار می‌گیرد. ایده اولیه این توزیع در ارلنگ (۱۹۰۹) مطرح گردید و سپس

---

[۱] سخنران، s.fathimanesh@uok.ac.ir

[1]phase-type distribution

نیوتس [۲] در سال ۱۹۷۵ آن را در یک چارچوب ریاضی مدون بسط داد و به عنوان یک خانواده از توزیع‌ها معرفی نمود (نیوتس ، ۱۹۷۵).

خاصیت بسته بودن این خانواده از توزیع‌ها نسبت به مجموع متناهی، توزیع‌های آمیخته و سیستمهای منسجم قابلیت اعتمادی در اسف و لویکسون (۱۹۸۲) مورد بررسی قرار گرفت. در دهه ۹۰ میلادی و بعد از آن، پژوهشگران مختلفی به تطبیق این توزیع‌ها بر پدیده‌های تصادفی مختلف و همچنین به کنکاش بیشتری از خواص این خانواده پرداختند. در سیندی (۱۹۹۰) ویژگی‌ها و نحوه مشخص‌سازی این توزیع‌ها مورد مطالعه قرارگرفت. همچنین، نحوه محاسبه براورد درستنمایی ماکسیمم پارامترهای این خانواده از توزیع‌ها در بابیو و کومنی (۱۹۹۲) شرح داده شد. در آلن (۱۹۹۵) و فدی (۱۹۹۵) از توزیع‌های فاز–نوع به ترتیب برای مدل‌بندی داده‌های تحلیل بقا و قابلیت اعتماد استفاده شد. در لی (۲۰۰۳) و کای و لی (۲۰۰۵) به کاربرد حالت چند متغیره این توزیع‌ها در مدلهای شوک و نظریه ریسک پرداخته شد. همچنین، بلاد (۲۰۰۵) یک مقاله مروری در زمینه مدل‌بندی ریسک با استفاده از توزیع‌های فاز–نوع نوشت. از جمله مقالات پیشرو در زمینه کاربرد این توزیع‌ها در مدل‌های شوک می‌توان به منتورو و همکاران (۲۰۰۹) و سگویا و لابیو (۲۰۱۳) اشاره نمود. در چند سال اخیر نیز موجی جدید از توجه به توزیع‌های فاز–نوع در انواع مدل شوک ایجاد شده است. از جمله این مقالات می‌توان به ایرلماز (۲۰۱۷) ، ژائو و همکاران (۲۰۱۸) ، ازکوت و ایرلماز (۲۰۱۹) ، وانگ و همکاران (۲۰۲۲) ، ، ایرلماز و انلو (۲۰۲۳) ، اوزکوت و همکاران (۲۰۲۴) ، دانگ و همکاران (۲۰۲۵) و منش و همکاران (۲۰۲۵) اشاره نمود. در این مقاله، هدف ما مروری بر استفاده از خانواده توزیع‌های فاز–نوع در چند مدل شوک جدید و مختلف معرفی شده در چند سال اخیر است. برای این منظور به‌طور ویژه، سه مدل شوک: کرانگین تعمیم یافته، [۳] دلتای تعمیم یافته، [۴] و تجمعی تعمیم یافته [۵] را مورد بررسی قرار می‌دهیم و ویژگی‌های توزیعی و قابلیت اعتمادی طول عمر سیستم‌های تحت این گونه شوک‌ها را بر مبنای توزیع‌های فاز–نوع بیان می‌کنیم. در ادامه این مقاله نخست، به معرفی اجمالی خانواده توزیع‌های فاز–نوع گسسته و پیوسته می‌پردازیم. سپس مدل شوک‌های پایه‌ای را بصورت مختصر توضیح داده و در نهایت نحوه استفاده از خانواده توزیع‌های فاز–نوع را برای مدل شوک‌های بیان شده در بالا را شرح می‌دهیم.

۲ توزیع‌های فاز–نوع

توزیع‌های فاز–نوع به هر دو صورت گسسته و پیوسته و بر مبنای فرآیندهای مارکف گسسته و پیوسته قابل تعریف هستند. در این بخش نخست، حالت گسسته و سپس حالت پیوسته را بیان می‌کنیم.

تعریف ۱.۲. فرایند مارکف زمان–گسسته $\{X_n, n = ۰, ۱, ۲, \ldots\}$ با فضای حالت $L = \{۱, ۲, \ldots, m, m+۱\}$، که در آن m حالت اول گذرا و حالت $(m+۱)$ام جاذب است، ماتریس احتمال انتقال

$$\mathbf{P} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ ۰ & ۱ \end{pmatrix}$$

که در آن $\mathbf{T}$ یک زیر ماتریس $m \times m$ مربوط به احتمال انتقال حالت‌های گذرا است و توزیع اولیه $\pi_۰ = (\boldsymbol{\alpha}', \alpha_{m+۱})'$ را در نظر

[2] Neuts
[3] generalized extreme shock model
[4] generalized δ shock model
[5] generalized cumulative shock model

بگیرید. اگر متغیر تصادفی X را زمان تا جذب فرایند تعریف کنیم در اینصورت X دارای توزیع فاز-نوع گسسته با پارامترهای $\boldsymbol{\alpha}$ و $\mathbf{T}$ است و آن را با نماد $X \sim PH_d(\boldsymbol{\alpha}, \mathbf{T})$ نشان می‌دهیم.

برای این توزیع می‌توان روابط زیر را بدست آورد.

- تابع جرم احتمال:

$$p_X(k) = \boldsymbol{\alpha}'\mathbf{T}^{k-1}\mathbf{t}, \qquad k = 1, 2, 3, \ldots.$$

- تابع توزیع تجمعی:

$$F_X(k) = 1 - \boldsymbol{\alpha}'\mathbf{T}^{k}\mathbf{1}, \qquad k = 0, 1, 2, 3, \ldots.$$

- تابع مولد احتمال:

$$H_X(z) = E(z^X) = \alpha_{m+1} + z\boldsymbol{\alpha}'(\mathbf{I} - z\mathbf{T})^{-1}\mathbf{t},$$

-

$$\frac{d^k H(z)}{dz^k}|_{z=1} = E(X(X-1)\ldots,(X-(k-1)) = k!\boldsymbol{\alpha}'(\mathbf{I} - z\mathbf{T})^{-k}\mathbf{T}^k\mathbf{1}.$$

تعریف ۲.۲. فرایند مارکف زمان-پیوسته $\{X_t, t \geq 0\}$ با فضای حالت $L = \{1, 2, \ldots, m, m+1\}$ که در آن m حالت اول گذرا و حالت $m+1$ام جاذب است، ماتریس پارامترهای بینهایت کوچک [6]

$$\mathbf{Q} = \begin{pmatrix} \mathbf{T}_{m\times m} & \mathbf{t}_{m\times 1} \\ 0_{1\times m} & 0 \end{pmatrix}$$

که در آن $\mathbf{T}_{m\times m}$ یک زیر ماتریس $m \times m$ مربوط به پارامترهای بی‌نهایت کوچک حالت‌های گذرا است و توزیع اولیه $\pi_0 = (\boldsymbol{\alpha}', \alpha_{m+1})'$ را در نظر بگیرید. اگر متغیر تصادفی پیوسته X را زمان تا جذب فرایند تعریف کنیم در اینصورت X دارای توزیع فاز-نوع پیوسته با پارامترهای $\boldsymbol{\alpha}$ و $\mathbf{T}$ است و آن را با نماد $X \sim PH_c(\boldsymbol{\alpha}, \mathbf{T})$ نشان می‌دهیم.

برای حالت پیوسته نیز می‌توان روابط زیر را بدست آورد.

- تابع چگالی احتمال:

$$f_X(x) = \boldsymbol{\alpha}' \exp(\mathbf{T}x)\mathbf{t}, \qquad x \geq 0$$

که در آن

$$\exp(\mathbf{T}x) = \sum_{i=0}^{\infty} \frac{(\mathbf{T}x)^i}{i!}$$

به ماتریس نمایی معروف است و $\mathbf{T}^0 = \mathbf{I}$ تعریف می‌شود.

- تابع توزیع تجمعی:

$$F_X(x) = 1 - \boldsymbol{\alpha}' \exp(\mathbf{T}x)\, \mathbf{1}, \qquad x \geq 0$$

که در آن $\mathbf{1}' = (1, \ldots, 1)$.

[6] Infinitesimal

- تابع تبدیل لاپلاس:

$$L_X(s) = E(\exp(-sX)) = \alpha_{m+1} + \boldsymbol{\alpha}'(s\mathbf{I} - z\mathbf{T})^{-1}\mathbf{t},$$

این خانواده از توزیع‌ها دارای ویژگی‌های بستاری جالبی در هر دو حالت گسسته و پیوسته می‌باشند که در پایین به بیان مختصری از آنها می‌پردازیم.

- نسبت به ترکیبات خطی بسته است. به بیان دیگر اگر $X_1, \ldots, X_n$ متغیرهای تصادفی مستقل که همگی عضو خانواده توزیع‌های فاز-نوع گسسته (پیوسته) باشند، آنگاه برای مقادیر ثابت و نامنفی $a_1, \ldots, a_n$ $\sum_{i=1}^{n} a_i X_i$ نیز دارای توزیع فاز-نوع گسسته (پیوسته) است.

- نسبت به آمیختگی متناهی بسته است. به بیان دیگر اگر $X_1, \ldots, X_n$ متغیرهای تصادفی مستقل که همگی عضو خانواده توزیع‌های فاز-نوع گسسته (پیوسته) باشند، آنگاه برای متغیرهای تصادفی $I_i \sim Ber(p_i)$ با شرط $\sum_{i=1}^{n} I_i = 1$ نتیجه می‌شود که توزیع آمیخته متناهی آنها یعنی $\sum_{i=1}^{n} I_i X_i$ نیز دارای توزیع‌های فاز-نوع گسسته (پیوسته) است.

- نسبت به مجموع تصادفی بسته است. به بیان دیگر اگر $X_1, \ldots, X_n, \ldots$ دنباله‌ای از متغیرهای تصادفی مستقل و هم‌توزیع $(i.i.d)$ [7] فاز-نوع پیوسته باشند و N دارای توزیع فاز-نوع گسسته باشد، آنگاه $\sum_{i=1}^{N} X_i$ دارای توزیع فاز-نوع پیوسته است.

- نسبت به آماره‌های مرتب بسته است. به بیان دیگر اگر $X_1, \ldots, X_n$ متغیرهای تصادفی مستقل که همگی عضو خانواده توزیع‌های فاز-نوع گسسته (پیوسته) باشند، آنگاه آماره‌های مرتب متناظر با آنها، یعنی $X_{(1)}, \ldots, X_{(n)}$ نیز دارای توزیع فاز-نوع گسسته (پیوسته) می‌باشند.

- خانواده توزیع‌های فاز-نوع در بین خانواده توزیع‌های نامنفی چگال است یعنی برای هر متغیر تصادفی نامنفی دلخواه می‌توان دنباله‌ای از متغیرهای تصادفی عضو خانواده توزیع‌های فاز-نوع را به گونه‌ای یافت که به متغیر تصادفی مورد نظر در توزیع همگرا باشد.

- ممکن است برای یک متغیر تصادفی عضو خانواده توزیع‌های فاز-نوعی بیش از یک نمایش از لحاظ شیوه تعریف پارامترها وجود داشته باشد و لزوما یکتا نمی‌باشند اما در هر صورت بدیهی است که ویژگی‌های توزیعی در نهایت یکسان خواهند بود.

برای توضیحات بیشتر در مورد این خانواده از توزیع‌ها و اثبات روابط بالا می‌توان به بلاد و نیلسون (۲۰۱۷) و هی (۲۰۱۴) مراجعه نمود.

۳ مدل‌های شوک

یکی از مدل‌های رایج در قابلیت اعتماد در مدل‌بندی طول عمر سیستم‌ها یا قطعاتی با عامل خرابی خارجی (خارج از سیستم) مدل‌های شوک می‌باشند. در این مدل‌ها، خرابی سیستم بر اساس الگوی معینی از نحوه وارد شدن شوک‌ها از لحاظ شدت و فاصله زمانی بین

[7] independent and identical distribution

شوک‌ها تعریف می‌شود. شوک‌های خارجی، بسته به سیستم مورد مطالعه، می‌توانند نوسانات الکتریکی و دمایی، ضربات و تکانه‌ها، خرابی بخشی از سیستم و غیره باشد. همچنین، این شوک‌ها در مسائل مالی و بیمه نیز بصورت اتفاقات نامطلوب قابل تعبیر است. به عنوان مثال، در یک مدل ریسک، هر ادعای خسارت می‌تواند به عنوان یک شوک تلقی گردد. در قابلیت اعتماد، انواع مختلفی از مدل‌های شوک تعریف شده‌اند و از آنها جهت مدل‌بندی طول عمر سیستم‌های تصادفی مختلف به منظور مطالعه شاخص‌های قابلیت اعتمادی و یا سیاست نگهداری بهینه استفاده شده است.

فرض کنید یک سیستم از لحظه صفر شروع به کار می‌کند وشوک‌ها نیز طبق یک فرایند تصادفی معین، به این سیستم وارد می‌شوند. فصله زمانی بین شوک $(i-1)$ام و iام را با X_i، شدت شوک iام را با Y_i، تعداد شوک‌های وارد شده به سیستم تا لحظه خرابی را با N، مقدار ثابت و معین حد آستانه را با d و طول عمر سیستم را با $T=\sum_{i=1}^{N} X_i$ نشان می‌دهیم. مدل‌های شوک مختلفی بر اساس نحوه تعریف N قابل تعریف هستند که از بین آنها سه مدل شوک زیر پایه‌ای و بنیادی هستند که بقیه مدل شوک‌ها نیز بصورت تعمیمی از این مدل شوک‌ها تعریف می‌شوند.

- **مدل شوک کرانگین :** در این مدل تا زمانیکه شدت شوک‌های وارده از یک حد آستانه مانند d بیشتر نشود سیستم فعال است و در نتیجه $N=\min\{n, Y_n > d\}$ (گات و سلر ، ۱۹۹۹).

- **مدل شوک تجمعی :** در این مدل تا زمانیکه مجموع شدت شوک‌های وارده از یک حد آستانه مانند d بیشتر نشود سیستم فعال است و در نتیجه $N=\min\{n, Y_1+\ldots, Y_n > d\}$ (گات ، ۱۹۹۰).

- **مدل شوک دلتا :** در این مدل، بدون توجه به میزان شدت شوک‌ها، اگر فاصله زمانی بین دو شوک پیاپی از حد آستانه δ کمتر باشد سیستم فعال است و در نتیجه $N=min\{n, X_1>\delta, \ldots, X_{n-1}>\delta, X_n \leq \delta\}$ (لی و ژائو ، ۲۰۰۷).

نویسندگان بسیاری تعمیم‌هایی از مدل‌های بالا و همچنین مدل‌های شوک آمیخته را که ترکیبی از دو یا چند مدل شوک های قبلی هستند را ارائه کرده‌اند. برای توضیحات بیشتر به عنوان نمونه می‌توان به گات (۲۰۰۱) و مالور و همکاران (۲۰۰۶) مراجعه کرد.

۴ مدل‌های شوک بر اساس خانواده توزیع‌های فاز-نوع

با توجه به نحوه تعریف N در بسیاری از موارد و تحت شرایط نه‌چندان محدود کننده، می‌توان از توزیع‌های فاز-نوع گسسته برای مدل‌بندی N و همچنین توزیع‌های فاز-نوع پیوسته برای مدل‌بندی T استفاده نمود. در این بخش، نحوه استفاده از توزیع‌های فازنوع را برای چند مدل شوک تعمیم یافته مورد بحث قرار می‌دهیم.

۱.۴ مدل شوک کرانگین تعمیم یافته

این مدل شوک بوسیله بوزبولت و ایرلماز (۲۰۲۰) و برخی خواص آن نیز بوسیله منش و همکاران (۲۰۲۳) مورد بررسی قرار گرفت. در این مدل مانند مدل شوک کرانگین معمولی، اگر شوک وارد شده به سیستم از حد آستانه d عبور کند سیستم از کار می‌افتد با این تفاوت که درحالت تعمیم یافته به جای یک منبع شوک، m منبع شوک مستقل و جود دارند. اگر $Y_{i,k}$ برای $k=1,\ldots,m$ نشان‌دهنده میزان شدت iامین شوک وارد به سیستم از منبع شوک kام باشد و d_k حد آستانه آن باشد در اینصورت $P_k=P(Y_i^k>d_k)$ احتمال خرابی

سیستم توسط این شوک است. همچنین احتمال وارد شدن شوک از طریق منبع kام برابر π_k است. تحت این مفروضات و نمادگذاری‌ها، دو نوع مدل کرانگین تعمیم یافته در بوزبولت و ایرلماز (۲۰۲۰) تعریف شده است

- **مدل ۱:** در این مدل، در طول فعال بودن سیستم، شوک‌ها می‌توانند از منابع مختلف به سیستم وارد شوند.
- **مدل ۲:** در این مدل، اولین شوک از هر منبعی وارد شود در ادامه بقیه شوک‌ها نیز تا زمان از کار افتادن سیستم فقط از آن منبع وارد می‌شوند.

فرض کنید تعداد شوک‌های وارد شده به سیستم تا زمان خرابی مطابق این دو مدل به ترتیب N_1 و N_2 و طول عمر سیستم تحت این دو مدل به ترتیب $T_1 = \sum_{i=1}^{N_1} X_i$ و $T_2 = \sum_{i=1}^{N_2} X_i$ باشند. در اینصورت در مدل ۱، N_1 زمان جذب یک فرآیند مارکف گسسته با فضای وضعیت $L = \{0, 1\}$ است که در آن ۱ فعال بودن سیستم (وضعیت گذرا) و ۰ از کار افتادن سیستم (وضعیت جاذب) را نشان می‌دهد. بنابراین،

$$N_1 \sim Ph_d(1, 1 - \sum_{i=1}^{m} p_i\pi_i) \sim Geo(1 - \sum_{i=1}^{m} p_i\pi_i).$$

در مدل ۲، N_2 زمان جذب یک فرآیند مارکف گسسته با فضای وضعیت $L = \{1, 2, \ldots, m, m+1\}$ است. وضعیت i نشان دهنده این است که شوک وارد شده از منبع i باعث از کار افتادن سیستم نشده است و وضعیت $m+1$ از کار افتادن سیستم را نشان می‌دهد. در این حالت، توزیع اولیه فرایند $\mathbf{a} = (\pi_1, \ldots, \pi_m)'$ و زیر ماتریس احتمال انتقال حالت‌های گذرا به ترتیب بصورت

$$\mathbf{Q} = \begin{pmatrix} 1-p_1 & 0 & \ldots & 0 \\ 0 & 1-p_2 & \ldots & 0 \\ \vdots & & & \\ 0 & 0 & \ldots & 1-p_m \end{pmatrix}$$

می‌باشند. در نتیجه $N_2 \sim PH_d(\mathbf{a}, \mathbf{Q})$. اگر X_iها را نیز متغیرهای تصادفی $(i.i.d)$ از توزیع $PH_c(\boldsymbol{\alpha}, \mathbf{A})$ باشند آنگاه T_1 و T_2 نیز عضو خانواده توزیع‌های فازنوع بصورت زیر هستند.

$$\begin{aligned} T_1 &\sim PH_c(\boldsymbol{\alpha}, \mathbf{A} \otimes \mathbf{I} + (-\mathbf{A}\mathbf{1}'\boldsymbol{\alpha}) \otimes (1 - \sum_{i=1}^{n} \pi_i p_i)) \\ T_2 &\sim PH_c(\boldsymbol{\alpha} \otimes \mathbf{a}, \mathbf{A} \otimes \mathbf{I} + (-\mathbf{A}\mathbf{1}'\boldsymbol{\alpha}) \otimes \mathbf{Q}). \end{aligned}$$

در این روابط، نماد $\otimes$ نشان دهنده ضرب کرونکر [8] بین دو ماتریس است. برای توضیحات بیشتر در مورد این نوع ضرب ماتریسی می‌توان به گراهام (۲۰۱۸) مراجعه کرد. حال، با توجه به روابط موجود در توزیع‌های فاز–نوعی می‌توان ویژگی‌های قابلیت اعتمادی مختلف مانند تابع قابلیت اعتماد، تابع نرخ خطر و میانگین طول عمر باقی مانده را برای T_1 و T_2 و بصورت روابط ماتریسی یافت.

۲.۴ مدل شوک تجمعی تعمیم یافته

در مدل شوک تجمعی تعمیم یافته همانند مدل شوک تجمعی، اگر مجموع شوک‌های وارد شده به سیستم از یک حد آستانه عبور کند سیستم از کار می‌افتد اما با این تفاوت که در حالت تعمیم یافته به جای یک منبع شوک، m منبع شوک مستقل وجود دارد که هر شوک

[8] kronecker product

با احتمال π_i از منبع iام صادر می‌شود (گانگ و همکاران ، ۲۰۲۰). اگر برای $i = ۱, \ldots, m$ شدت شوک وارد شده از منبع iام را با $Y^{(i)}$ نشان دهیم و شوک وارد شده به سیستم را با Z نشان دهیم در اینصورت

$$P(Z > t) = \sum_{i=۱}^{m} \pi_i P(Y^{(i)} > t).$$

در این مدل

$$N = \min\{n, Z_۱ + \ldots + Z_n > d\}, \qquad T = \sum_{i=۱}^{N} X_i.$$

حال اگر فرض کنیم $Y^{(i)} \sim PH_c(\boldsymbol{\alpha_i}, \mathbf{A}_i)$ در اینصورت از بسته بودن خانواده توزیع‌های فازنوعی نسبت به توزیع‌های آمیخته نتیجه می‌شود که

$$Z = \sum_{i=۱}^{m} I_i Y^{(i)} \sim PH_c(\boldsymbol{\alpha}, \mathbf{A}), \quad \boldsymbol{\alpha} = (\pi_۱ \boldsymbol{\alpha}_۱, \ldots \pi_m \boldsymbol{\alpha_m}), \quad \mathbf{A} = diag\{\mathbf{A}_۱, \ldots, \mathbf{A}_m\}.$$

همچنین با توجه به بسته بودن این خانواده نسبت به مجموع، داریم

$$\sum_{i=۱}^{n} Z_i \sim PH_c(\boldsymbol{\gamma}, \boldsymbol{\Gamma})$$

که در آن

$$\boldsymbol{\gamma} = (\boldsymbol{\alpha}', c\boldsymbol{\alpha}', \ldots, c^{n-۱}\boldsymbol{\alpha}')', \qquad c = ۱ - \boldsymbol{\alpha}'\mathbf{۱} \qquad \mathbf{A}^\circ = -\mathbf{A}\mathbf{۱},$$

و

$$\boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{A} & \mathbf{A}^\circ\boldsymbol{\alpha}' & \ldots & c^{n-۲}\mathbf{A}^\circ\boldsymbol{\alpha}' \\ \circ & \mathbf{A} & \ldots & c^{n-۳}\mathbf{A}^\circ\boldsymbol{\alpha}' \\ \vdots & \vdots & \vdots & \vdots \\ \circ & \circ & \ldots & \mathbf{A} \end{pmatrix}.$$

اکنون، با داشتن توزیع $\sum_{i=۱}^{n} Z_i$ می‌توانیم توزیع N را از رابطه زیر بیابیم.

$$\begin{aligned} P(N = n) &= P(\sum_{i=۱}^{n} Z_i > d, \sum_{i=۱}^{n-۱} Z_i \leq d) \\ &= P(\sum_{i=۱}^{n} Z_i > d) - P(\sum_{i=۱}^{n-۱} Z_i > d). \end{aligned}$$

با در نظر گرفتن فرض استقلال فاصله زمانی بین شوک‌ها و شدت آنها، تابع قابلیت اعتماد سیستم نیز بصورت یک توزیع آمیخته به شکل زیر قابل محاسبه است.

$$P(T > t) = P(\sum_{i=۱}^{N} X_i > t) = \sum_{n=۱}^{\infty} P(\sum_{i=۱}^{n} X_i > t) p(N = n).$$

علاوه براین، اگر فرض کنیم فاصله زمانی بین ورود شوک‌ها iid ازتوزیع فاز–نوعی هستند آنگاه $\sum_{i=۱}^{n} X_i$ نیز عضو این خانواده باقی می‌مانند و توزیع T قابل بیان بصورت آمیخته‌ای از توزیع‌های فاز–نوعی هستند.

۳.۴ مدل شوک دلتای تعمیم یافته

مدل شوک دلتای تعمیم یافته بوسیله ایرلماز و انلو (۲۰۲۳) و به عنوان تعمیمی از مدل شوک دلتا ارائه گردید. بر اساس این مدل، سیستم در لحظه صفر فعال است و فاصله بین زمان‌های ورود شوک متغیرهای تصادفی iid می‌باشند. این سیستم تا زمانیکه m شوک پیاپی در یک فاصله زمانی کمتر از δ به سیستم وارد نشود، سیستم فعال است. اگر تعداد شوک‌های وارد شده به این سیستم تا لحظه خرابی را با $N_{\delta,m}$ نشان دهیم در اینصورت برای $n \geq m$ داریم

$$\{N_{\delta,m} = n\} = \{X_1 + \ldots X_m \geq \delta, X_2 + \ldots X_{m+1} \geq \delta, \ldots, X_{n-m} + \ldots X_{n-1} \geq \delta, X_{n-m+1} + \ldots X_n < \delta\}$$

و طول عمر سیستم نیز بصورت $T = \sum_{i=1}^{N_{\delta,m}} X_i$ خواهد بود. اگر $m = 1$ باشد این مدل، به مدل شوک δ معمولی تبدیل می‌شود. به جهت وابستگی مجموع زمان‌های تعریف شده در تعریف $N_{\delta,m}$، محاسبه دقیق توزیع آن مشکل است ایرلماز و انلو (۲۰۲۳) با تعریف $p_1 = P(X_1 + \ldots X_m \geq \delta, X_2 + \ldots X_{m+1} \geq \delta)$ و $p_2 = P(X_1 + \ldots X_m \geq \delta)$ آن را با متغیر تصادفی $N^*_{\delta,m}$ بصورت زیر تقریب زدند.

$$P(N^*_{\delta,m} = n) = \begin{cases} 0 & n < m \\ 1 - p_2, & n = m \\ p_2 (\frac{p_1}{p_2})^{n-m-1} (1 - \frac{p_1}{p_2})), & n > m. \end{cases}$$

تحت فرض فاز–نوعی بودن توزیع X_iها و با فرض $P(X_i > 0) = 1$ توزیع تقریبی $N_{\delta,m}$ برای هر دو حالت گسسته و پیوسته بودن X_iها قابل محاسبه است.

حالت گسسته

در این حالت، فرض کنید $X \sim PH_d(\mathbf{a}, \mathbf{Q}_{k\times k})$ باشد در اینصورت با توجه به خواص توزیع‌های فاز–نوعی $\sum_{i=1}^{m} X_i \sim PH_d(\boldsymbol{\pi}_m, \mathbf{Q}_m)$ که در آن $\boldsymbol{\pi}_m = (1, 0, \ldots, 0)'$ یک بردار mk بعدی، $\mathbf{Q}_1 = \mathbf{Q}$ و

$$\mathbf{Q}_m = \begin{pmatrix} \mathbf{Q}_{m-1} & (\mathbf{I}_{(m-1)k} - \mathbf{Q}_{m-1})\mathbf{1}\mathbf{a}' \\ 0 & \mathbf{Q} \end{pmatrix}.$$

با در نظر گرفتن این مفروضات p_1 و p_2 از روابط زیر قابل محاسبه هستند.

$$\begin{aligned} p_1 &= \boldsymbol{\pi}'_{m-1}\mathbf{Q}^{\delta-2}_{m-1}\mathbf{1} + \sum_{s=1}^{\delta-2} \mathbf{a}'(\mathbf{Q}^{\delta-s-1}\mathbf{1})^2 \boldsymbol{\pi}'_{m-1}\mathbf{Q}^{s-1}_{m-1}(\mathbf{I}_{(m-1)k} - \mathbf{Q}_{m-1})\mathbf{1} \\ p_2 &= \boldsymbol{\pi}'_m \mathbf{Q}^{\delta-1}_{m-1}\mathbf{1}. \end{aligned}$$

حالت پیوسته

اگر فاصله زمانی بین ورود شوک‌ها متغیرهای تصادفی پیوسته باشند مقادیر p_1 و p_2 در حالت کلی بصورت زیر قابل محاسبه هستند.

$$p_1 = \bar{F}_{m-1}(\delta) + \int_0^\delta (\bar{F}(\delta - s))^2 f_{m-1}(s)ds \tag{۱.۴}$$

$$p_2 = \bar{F}_{m-1}(\delta) + \int_0^\delta \bar{F}(\delta - s) f_{m-1}(s)ds \tag{۲.۴}$$

در روابط (۱.۴) و (۲.۴) عبارت $\bar{F}_{m-1}$ و f_{m-1} به ترتیب تابع قابلیت اعتماد و تابع چگالی احتمال $X_1+\ldots+X_{m-1}$ می‌باشند. اگر فرض کنیم $X \sim PH_c(\boldsymbol{\alpha}, \mathbf{A})$، در اینصورت با توجه به خواص توزیع‌های فاز–نوعی روابط (۱.۴) و (۲.۴) بصورت زیر قابل محاسبه هستند.

$$
\begin{aligned}
p_1 &= \boldsymbol{\alpha}' \exp(\mathbf{A}s)\mathbf{1} + \int_0^{\delta} (\boldsymbol{\alpha}' \exp(\mathbf{A}(\delta - s))\mathbf{1})^2 \boldsymbol{\alpha}' \exp(\mathbf{A}s)\mathbf{a}^0 \, ds \\
p_2 &= (\boldsymbol{\alpha}', c\boldsymbol{\alpha}') \exp(\begin{pmatrix} \mathbf{A} & \mathbf{A}^0\boldsymbol{\alpha}' \\ 0 & \mathbf{A} \end{pmatrix} \delta)\mathbf{1}.
\end{aligned}
$$

مراجع

AALEN, O.O. (1995), Phase type distributions in survival analysis, *Scand. J. Statist* , **22**, 447-463.

Assaf, D. and Levikson, B. (1982), Closure of phase-type distributions under operations arising in reliability theory, *The Annals of Probability* , **10**, 265-269.

Bladt, M. (2005), A review on phase-type distributions and their use in risk theory, *ASTIN Bulletin: The Journal of the IAA*, **35(1)**, 145-161.

Bladt, M., & Nielsen, B. F. (2017), *Matrix-exponential distributions in applied probability (Vol. 81, pp. 1-736)*, New York: Springer.

Bobbio A, Cumani A (1992), ML estimation of the parameters of a PH distribution in triangular canonical form, in: Balbo G, Serazzi G (eds) Computer performance evaluation, Elsevier, Amsterdam, 173-206.

Bozbulut, A.A. & Eryilmaz, S. (2020), extreme shock models and their applications, *Communications in Statistics – Simulation and Computation*, **49(1)** 110-120.

Cai, J., and Li, H. (2005), Multivariate risk model of phase type, with applications to shock models., *Insurance: Mathematics and Economics*, **36(2)**, 137-152.

Dong, Q., Bai, M., Yan, Z., & Wu, B. (2025), An Optimal load adjustment policy for multi-state k-out-of-n balanced systems with self-healing mechanisms, *Reliability Engineering & System Safety*, 111091.

ERLANG, A. (1909), Sandsynlighedsregning og telefonsamtaler, *Nyt tidsskrift for Matematik* , **20**, 33-39.

Eryilmaz, S. (2017), Computing optimal replacement time and mean residual life in reliability shock models, *Computers and industrial engineering*, **103**,40-45.

Eryilmaz, S., & Unlu, K. D. (2023), A new generalized δ-shock model and its application to 1-out-of-(m+1): G cold standby system, *Reliability Engineering & System Safety*, **234**,109203.

Faddy MJ. (1995), Phase-type distributions for failure times, *Math Comput Model* , **22**, 63-70. Phase-type distributions for failure times.

Gong, M., Eryilmaz, S., & Xie, M. (2020), Reliability assessment of system under a generalized cumulative shock model, *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, **234(1)**, 129-137.

Graham, A. (2018), *Kronecker products and matrix calculus with applications*, Courier Dover Publications.

Gut, A. (1990), Cumulative shock models, *Advances in Applied Probability*, **22 (2)** 504-507.

Gut, A. (2001), Mixed shock models, *Bernoulli*, **7(3)** 541-555.

Gut, A., & Hu¨ sler, J. R. (1999), Extreme shock models, *Extremes*, **2** 295-307.

He, Q. M. (2014), *Fundamentals of matrix-analytic methods (Vol. 365)*, New York: Springer.

Li, H., (2003), Association of multivariate phase-type distributions, with applications to shock models., *Statist. Probab. Lett* , **64**, 281-392.

Li, Z. and Zhao, P. (2007), Reliability Analysis on the δ-Shock Model of Complex Systems, *IEEE Transactions on Reliability*, **56(2)** 340-348.

Mallor, F., Omey, E. and Santos, J. (2006), Asymptotic results for a run and cumulative mixed shock model, *Journal of Mathematical Sciences*, **138(1)** 5410-5414.

Manesh, S. F., Izadi, M., & Khaledi, B. E. (2023), On stochastic ordering among extreme shock models, *Probability in the Engineering and Informational Sciences*, **37(4)** 961-972.

Manesh, S. F., Izadi, M., & Khaledi, B. E. (2025), A new mixed generalized δ-shock model, *statistics & probability letters*, submitted.

Montoro-Cazorla, D., Pérez-Ocón, R., and Segovia, M. C. (2009), Shock and wear models under policy N using phase-type distributions, *Applied mathematical modelling*, **33(1)**,543-554.

Neuts, M.F.,. (1975), Probability distributions of phase type, In: Liver Amicorum Prof. Emeritus H. Florin, University of Louvain, Belgium, **20**, 173-206.

O'CINNEIDE, C.A. (1990), Characterization of phase-type distributions, *Comm. Statist Stochastic Models* , **6**, 1-57.

Ozkut, M., and Eryilmaz, S. (2019), Reliability analysis under Marshall–Olkin run shock model, *Journal of Computational and Applied Mathematics*, **349**,52-59.

Ozkut, M., Kan, C., & Franko, C. (2024), Analyzing the multi-state system under a run shock mode, *Probability in the Engineering and Informational Sciences*, **38(4)**,619-931.

Segovia, M. C., and Labeau, P. E. (2013), Reliability of a multi-state system subject to shocks using phase-type distributions, *Applied mathematical modelling*, **37(7)**,488-554.

Wang, X., Zhao, X., Wu, C., & Wang, S. (2022), Mixed shock model for multi-state weighted k-out-of-n: F systems with degraded resistance against shocks, *Reliability Engineering & System Safety*, **217**,108098.

Zhao, X., Guo, X., and Wang, X. (2018), Reliability and maintenance policies for a two-stage shock model with self-healing mechanism, *Reliability Engineering and System Safety*, **172**,185-194.

A review of the use of the family of phase-type distributions in the generalized shock model.

Sirous Fathi Manesh [1], **Muhyiddin Izadi** [2]

[1] Department of Statistics, University of Kurdistan, Sanandaj, Iran

[2]Department of Statistics, Razi University, Kermanshah, Iran

---

Abstract: The utilization of phase-type (PH) distributions - defined through discrete-time and continuous-time Markov processes - is increasingly prevalent in modelling non-negative discrete and continuous random variables. In recent years, these distributions have gained considerable attention among reliability researchers for their application in various shock models. This paper presents a comprehensive review of PH distribution applications in three principal shock models: the generalized shock model, generalized cumulative shock model, and generalized delta shock model. We provide theoretical justification for employing PH distributions in these shock models, followed by derivation of both shock count distribution and system lifetime distribution using the inherent matrix-based relationships characteristic of PH distributions.

Keywords: Generalized cumulative shock model, Generalized delta shock model, Generalized extreme shock model, Markov Processes, Phase-type distribution family.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

بررسی عملکرد مجانبی برآوردگر ترکیبی در مدل‌های خطی با خطا در اندازه‌گیری

فاطمه قپانی [۱] ،

گروه آمار و ریاضی ،واحد شوشتر، دانشگاه آزاد اسلامی، شوشتر، ایران

چکیده: در این مقاله یک برآوردگر ترکیبی برای بهبود دقت برآوردها در مدل‌های با خطی در اندازه‌گیری خطی تحت محدودیت‌های خطی تصادفی معرفی و ویژگی‌های مجانبی برآوردگر معرفی شده تعیین می‌شود. عملکرد برآوردگر معرفی شده با استفاده از معیار ماتریس میانگین مربعات خطا نسبت به بعضی برآوردگرهای موجود مورد بررسی قرار می‌گیرد. در پایان عملکرد برآوردگر معرفی شده با استفاده از یک مطالعه‌ی شبیه سازی ارزیابی می‌شود.

واژه‌های کلیدی:برآورردگر ترکیبی، مدل خطی با خطا در اندازه‌گیری،محدودیتهای خطی تصادفی، هم‌خطی .

کد موضوع‌بندی ریاضی (۲۰۲۰): 62J05، 62J07.

---

۱ مقدمه

وجود وابستگی خطی بین ستون های ماتریس متغیرهای توضیحی باعث کاهش دقت برآورد و حساسیت برآورد نسبت به تغییر کوچکی در مجموعه داده‌ها می‌شود. یک روش برخورد با مسئله هم‌خطی به دست آوردن برآورد پارامترهای مدل با اضافه کردن اطلاعات پیشین به صورت محدودیت‌های خطی دقیق یا تصادفی روی پارامترهای نامعلوم مدل است رائو و همکاران (۲۰۰۸) . در این مقاله یک برآوردگر ترکیبی با استفاده از روش کبریا –لکمن و در مدل‌های خطی با خطا در اندازه‌گیری تحت محدودیت های خطی تصادفی معرفی می‌شود. هدف اصلی برآورد ترکیبی، ترکیب برآوردگرهای مختلف برای کاهش انحراف و خطا در برآورد ضرایب مدل است. یک مدل با خطا در اندازه گیری را به‌صورت

$$y = Z\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$$

$$X = Z + U, U \sim MN(0, I_n \otimes \Sigma)$$

---

[۱] سخنران، Fatemeh.Ghapani@iau.ac.ir

در این مدل $y = (y_1, y_2, ..., y_n)'$ بردار $n \times 1$ از مشاهدات،Z ماتریس $n \times p$ از متغیرهای تو ضیحی ،β بردار $p \times 1$ از پارامترهای ناشناخته مدل و ε بردار $n \times 1$ از خطای تصادفی مدل است.X ماتریس مشاهده شده Z با خطای اندازه گیری U است. فرض کنید ε و U از هم مستقل و U یک ماتریس با عناصر معلوم است.

فرض کنید بردار پارامتر β تحت محدودیت‌های خطی تصادفی به‌صورت زیر باشد:

$$r = R\beta + e, e \sim N(0, \sigma^2 W)$$

در این رابطه r یک بردار معلوم $m \times 1$ از مشاهدات، R یک ماتریس معلوم $m \times p$ پرتبه سطری $m < p$ و e بردار $m \times 1$ از خطای تصادفی است.برآوردگر آمیخته در مدل‌های با خطا در اندازه‌گیری توسط قپانی و همکاران (۲۰۱۵) به‌صورت زیر معرفی شد:

$$\hat{\beta}_{ME} = S_r^{-1}(X'y + R'W^{-1}r),$$

در این رابطه $S_r = S + R'W^{-1}R$.

۲ برآوردگر ترکیبی

برای تعیین برآوردگر جدید در مدل‌های با خطا در اندازه‌گیری براساس کبریا و لکمن (۲۰۲۰)رابطه‌ی زیر را در نظر می‌گیریم:

$$\Phi_1 = (y - X\beta)'(y - X\beta) - n\beta'\Sigma\beta + k\left[(\beta + \hat{\beta})'(\beta + \hat{\beta}) - c\right]$$

در این رابطه c ضریب ثابت و k ضریب لاکرانژ در نظر گرفته می‌شود. با مشتق گرفتن از Φ_1 برآوردگر $\hat{\beta}_{KL}$ به‌صورت زیر تعیین می‌شود:

$$\hat{\beta}_{KL} = S_k^{-1}(X'y - k\hat{\beta})$$

به‌منظور در نظر گرفتن محدودیت‌های تصادفی در برآوردگر $\hat{\beta}_{KL}$ رابطه‌ی زیر را در نظر می‌گیریم:

$$\Phi_2 = (y - X\beta)'(y - X\beta) - n\beta'\Sigma\beta + k\left[(\beta + \hat{\beta}_{ME})'(\beta + \hat{\beta}_{ME}) - c\right] + (r - R\beta)'W^{-1}(r - R\beta)$$

با مشتق گرفتن از Φ_2 نسبت به β برآوردگر ترکیبی به‌صورت زیر معرفی می‌شود:

$$\hat{\beta}_{MKL} = (S + R'W^{-1}R + kI_p)^{-1}(X'y - k\hat{\beta}_{ME} + R'W^{-1}r)$$

۳ خواص مجانبی

در این بخش خواص مجانبی برآوردگر $\hat{\beta}_{MKL}$ مورد بررسی قرار می‌گیرد. فرض می‌کنیم وقتی n به بینهایت میل می‌کند حدهای $n^{-1}(Z'Z + R'W^{-1}R)$، $n^{-1}(Z'Z + R'W^{-1}R + kI_p)$ و $n^{-1}(Z'Z + R'W^{-1}R - kI_p)$ وجود دارند.

قضیه ۱.۳. تحت فرض‌های بیان شده $\hat{\beta}_{MKL}$ به طور مجانبی دارای توزیع نرمال با میانگین و کواریانس ماتریس به‌صورت زیر است

$$E(\hat{\beta}_{MKL}) = G_{rk}\beta$$

$$AVar(\hat{\beta}_{MKL}) = A_r^{-1}(G_{rk}BG_{rk} + \sigma^2 G_{rk}A_rG_{rk})A_r^{-1}$$

که در آن $B = (n\sigma^2 + \beta'Z'Z\beta)\Sigma$، $G_{rk} = (Z'Z + R'W^{-1}R + kI_p)^{-1}(Z'Z + R'W^{-1}R - kI_p)$ و $A_r = Z'Z + R'W^{-1}R$.

اثبات. از آنجا که $E(X'X) = Z'Z + n\Sigma$ فانگ و همکاران (۲۰۰۳)،بنابراین داریم:

$$n^{-1}(X'X + R'W^{-1}R) = n^{-1}(Z'Z + R'W^{-1}R) + \Sigma + O_p(n^{-\frac{1}{2}})$$

همچنین داریم

$$\begin{aligned}\sqrt{n}\hat{\beta}_{MKL} &= \left[n^{-1}(Z'Z + R'W^{-1}R) + O_p(n^{-\frac{1}{2}})\right]^{-1} n^{-\frac{1}{2}}\left[G_{rk}(X'y + R'W^{-1}r) + O_p(n^{\frac{1}{2}})\right]\\ &= \left[I_p + O_p(n^{-\frac{1}{2}})\right]^{-1} C^{-1}n^{-\frac{1}{2}}\left[G_{rk}(X'y + R'W^{-1}r) + O_p(n^{\frac{1}{2}})\right]\end{aligned}$$

زمانی که حد $C = n^{-1}(Z'Z + R'W^{-1}R)$ وجود دارد رابطه‌ی زیر نتیجه می‌شود:

$$\sqrt{n}\hat{\beta}_{MKL} = C^{-1}\xi + O_p(n^{-\frac{1}{2}}),$$

که $\xi = n^{-\frac{1}{2}}\left[G_{rk}(X'y + R'W^{-1}r)\right]$ به‌طور مجانبی نرمال است (فانگ و همکاران (۲۰۰۳)).

از آنجا که $E\left[G_{rk}(X'y + R'W^{-1}r)\right] = G_{rk}A_r\beta$، نتیجه گرفته می‌شود که $E(\xi) = n^{-\frac{1}{2}}G_{rk}A_r\beta$. همچنین

$$\sqrt{n}(\hat{\beta}_{MKL} - G_{rk}\beta) = C^{-1}\left[\xi - E(\xi)\right] + O_p(n^{-\frac{1}{2}})$$

بنابراین $\sqrt{n}(\hat{\beta}_{MKL} - G_{rk}\beta)$ دارای توزیع مجانبی نرمال با میانگین صفر است. همچنین داریم. $AVar(\sqrt{n}\hat{\beta}_{MKL}) = C^{-1}Var(\xi)C^{-1}$ برای محاسبه واریانسξ روش زیر را بکار می‌بریم:

$$\begin{aligned}Var(\xi) &= E_{y_r}\left[Var(\xi\,|y_r)\right] + Var_{y_r}\left[E(\xi\,|y_r)\right]\\ &= n^{-1}E_{y_r}(G_{rk}y'y\Sigma G_{rk}) + n^{-1}Var_{y_r}\left[G_{rk}(Z'y + R'W^{-1}r)\right],\end{aligned}$$

در این رابطه E_{y_r} و Var_{y_r} امید و واریانس نسبت به بردار $y_r' = (y', r')$ می باشند. از آنجا که $E(y'y) = n\sigma^2 + \beta'Z'Z\beta$ و $Var_{y_r}\left[G_{rk}(Z'y + R'W^{-1}r)\right] = \sigma^2 G_{rk}A_rG_{rk}$ می‌توان نوشت, $Var(\xi) = n^{-1}(G_{rk}BG_{rk} + \sigma^2 G_{rk}A_rG_{rk})$ بنابراین خواهیم نوشت:

$$AVar(\hat{\beta}_{MKL}) = A_r^{-1}(G_{rk}BG_{rk} + \sigma^2 G_{rk}A_rG_{rk})A_r^{-1}.$$

□

۴ ماتریس میانگین مربعات خطا

برای بررسی کارائی برآوردگر معرفی شده نسبت به برآوردگرهای موجود دیگر از معیار ماتریس میانگین مربعات خطا استفاده می‌شود. ماتریس میانگین مربعات خطای خطای برآوردگر $\hat{\beta}$ از پارامتر β به‌صورت زیر تعریف می‌شود:

$$MSEM(\hat{\beta}) = E(\hat{\beta}-\beta)(\hat{\beta}-\beta)' = Var(\hat{\beta}) + Bias(\hat{\beta})Bias(\hat{\beta})'$$

که $Bias(\hat{\beta}) = E(\hat{\beta}) - \beta$ مقدار اریبی برآوردگر $\hat{\beta}$ است. مقدار میانگین مربعات خطا معیار دیگری برای بررسی کارائی یک برآوردگر است که به صورت زیر تعریف می‌شود:

$$MSE(\hat{\beta}) = tr\left[MSEM(\hat{\beta})\right] = tr\left[Var(\hat{\beta})\right] + Bias(\hat{\beta})'Bias(\hat{\beta})$$

ماتریس میانگین مربعات برآوردگر های ذکر شده به طور مجانبی به صورت زیر نوشته می‌شود:

$$AMSEM(\hat{\beta}) = A^{-1}(B+\sigma^2 A)A^{-1}$$

$$AMSEM(\hat{\beta}_{KL}) = A^{-1}(G_k B G_k + \sigma^2 G_k A G_k)A^{-1} + b_1 b_1'$$

$$AMSEM(\hat{\beta}_{ME}) = A_r^{-1}(B+\sigma^2 A_r)A_r^{-1}$$

$$AMSEM(\hat{\beta}_{MKL}) = A_r^{-1}(G_{rk} B G_{rk} + \sigma^2 G_{rk} A_r G_{rk})A_r^{-1} + b_2 b_2'$$

که $b_1 = -2kA_k^{-1}\beta$ ، $A_k = Z'Z + kI_p$ و $b_2 = -2kA_{rk}^{-1}\beta$.

۱.۴ مقایسه ماتریس میانگین مربعات خطای برآوردگرهای $\hat{\beta}_{KL}$ و $\hat{\beta}$

به‌منظور مقایسه کارائی دو برآوردگر $\hat{\beta}_{KL}$ و $\hat{\beta}$ ماتریس زیر را در نظر می‌گیریم:

$$\begin{aligned}\Delta_1 &= AMSEM(\hat{\beta}) - AMSEM(\hat{\beta}_{KL}) = D_1 - b_1 b_1' \\ &= 2kA_k^{-1}(A^{-1}B + BA^{-1} + 2\sigma^2 I_p)A_k^{-1} - b_1 b_1',\end{aligned}$$

از آنجا که $D_1 = 2kA_k^{-1}(A^{-1}B + BA^{-1} + 2\sigma^2 I_p)A_k^{-1}$ یک ماتریس معین مثبت است . بنابراین Δ_1 معین مثبت است هرگاه $b_1' D_1^{-1} b_1 \leq 1$.

۲.۴ مقایسه ماتریس میانگین مربعات خطا برآوردگرهای $\hat{\beta}_{MKL}$ و $\hat{\beta}_{ME}$

به منظور مقایسه ماتریس میانگین مربعات خطای دو برآوردگر $\hat{\beta}_{MKL}$ و $\hat{\beta}_{ME}$ماتریس زیر را در نظر می گیریم:

$$\begin{aligned}\Delta_2 &= AMSEM(\hat{\beta}_{ME}) - AMSEM(\hat{\beta}_{MKL}) = D_2 - b_2 b_2' \\ &= 2kA_{rk}^{-1}(A_r^{-1}B + BA_r^{-1} + 2\sigma^2 I_p)A_{rk}^{-1} - b_2 b_2',\end{aligned}$$

از آنجا که $D_2 = 2kA_{rk}^{-1}(A_r^{-1}B + BA_r^{-1} + 2\sigma^2 I_p)A_{rk}^{-1}$ یک ماتریس معین مثبت است بنابراین Δ_2معین مثبت است هرگاه $b_2' D_2^{-1} b_2 \leq 1$.

۳.۴ مقایسه ماتریس میانگین مربعات برآوردگرهای $\hat{\beta}_{MKL}$ و $\hat{\beta}_{KL}$

به منظور بررسی دو برآوردگر در معیار ماتریس میانگین مربعات خطا ماتریس زیر در نظر گرفته می‌شود:

$$\Delta_3 = AMSEM(\hat{\beta}_{KL}) - AMSEM(\hat{\beta}_{MKL}) = D_3 + b_1 b_1' - b_2 b_2',$$

در این رابطه $D_3 = AVar(\hat{\beta}_{KL}) - AVar(\hat{\beta}_{MKL})$. واضح است که $AVar(\hat{\beta}_{KL}) > 0$ و $AVar(\hat{\beta}_{MKL}) > 0$. بنابراین هرگاه داشته باشیم $\lambda_{\max}\left[AVar(\hat{\beta}_{MKL})AVar(\hat{\beta}_{KL})^{-1}\right] < 1$ آنگاه $D_3 > 0$ بنابراین $\Delta_3 \geq 0$ اگر وفقط اگر داشته باشیم $b_3'(D_3 + b_1 b_1')^{-1} b_3 \leq 1$.

۵ برآورد پارامتر k

در این بخش پارامتر اریب k را طوری براورد می کنیم که ماتریس Δ_1 یک ماتریس معین مثبت شود. ماتریس Δ_1را می توان به صورت زیر نوشت:

$$\Delta_1 = 2kA_k^{-1}(A^{-1}B + BA^{-1} + 2\sigma^2 I_p - 2k\beta'\beta)A_k^{-1},$$

از آنجا که $A_k^{-1} > 0$، $A^{-1}B > 0$ و $BA^{-1} > 0$. بنابراین Δ_1 یک ماتریس معین مثبت است هرگاه $2\sigma^2 I_p - 2k\beta'\beta$ نیمه معین مثبت باشد. بنابراین یک برآورد مناسب برای پارامتر اریب k به صورت $\hat{k} < \frac{\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ تعیین می‌شود.

۶ مطالعه‌ی شبیه‌سازی

به‌منظور ارزیابی نتایج تئوری یک مطالعه‌ی شبیه‌سازی انجام شده است. برای در نظر گرفتن درجات مختلف هم‌خطی با توجه به مقاله مک‌دونالد و گالارنو (۱۹۷۵) متغیرهای توضیحی به‌صورت زیر تولید می‌شوند:

$$z_{il} = (1 - \rho^2)^{\frac{1}{2}} w_{il} + \rho w_{i,p+1}, i = 1, ..., n, l = 1, ..., p,$$

که در آن w_{ij} متغیرهای شبه تصادفی مستقل نرمال استاندارد و ρ^2 همبستگی بین دو متغیر توضیحی را نشان می‌دهد. داده‌ها همجنین استاندارد می‌شوند تا $Z'Z$ به فرم ماتریس همبستگی شود. j- امین مجموعه از داده‌های شبیه‌سازی از مدل با خطا در اندازه‌گیری تحت محدودیت‌های خطی تصادفی به‌صورت زیر تولید می‌شود:

$$\begin{aligned} y_j &= Z\beta + \varepsilon_j, \\ X_j &= Z + U_j, j = 1, ..., 1000, \\ r_j &= R\beta + e_j, \end{aligned}$$

در این رابطه $y_j = (y_{1j}, ..., y_{nj})'$، $Z = (z^{(1)}, z^{(1)}, z^{(3)})$، به طوری که $z^{(l)} = (z_{1l}, ..., z_{nl})'$، $l = 1, 2, 3$ و $\varepsilon_j \sim N(0, \sigma^2 I_n)$ تولید می شود. همجنین R یک ماتریس معلوم و عناصر ان از توزیع نرمال استاندارد تولید شده اند و $r_j = (r_{1j}, ..., r_{mj})'$ و $e_j \sim N(0, \sigma^2 I_m)$. در این مطالعه $p = 3$، $\sigma^2 = 0.5$، $m = 1$ و $\Sigma = diag(0.05, 0.05, 0.05)$

در نظر گرفته شده است. برای انجام شبیه‌سازی از نرم افزار R استفاده شده است . برا ی هر ترکیب از پارامترها ۱۰۰۰ تکرار انجام می شود. عملکرد برآوردگرها با استفاده از مقدار میانگین مربعات خطا از رابطه ی زیر ارزیابی می‌شود:

$$\mathrm{MSE}(\tilde{\beta}) = \frac{۱}{۱۰۰۰} \sum_{j=۱}^{۱۰۰۰} \sum_{l=۱}^{۳} (\tilde{\beta}_{lj} - \beta_l)^۲$$

نتایج شبیه‌سازی در جداول ۱ و ۲ ارائه شده است.

جدول ۱: برآورد مقدار میانگین مربعات خطا براوردگرهای مختلف با $n = ۵۰$

$\hat{\beta}_{MKL}$	$\hat{\beta}_{ME}$	$\hat{\beta}_{KL}$	$\hat{\beta}$	ρ
۰/۰۵۷۱	/۰۶۱۱.۰	۰/۰۵۷۵	۰/۰۶۱۵	۰/۸۰
۰/۱۱۲۲	۰/۱۲۵۱	۰/۱۱۳۵	۰/۱۲۶۷	۰/۹۰
۰/۲۴۹۱	۰/۲۹۵۱	۰/۲۵۴۴	۰/۳۰۲۳	۰/۹۵

جدول ۲: برآورد مقدار میانگین مربعات خطا براوردگرهای مختلف با $n = ۱۰۰$

$\hat{\beta}_{MKL}$	$\hat{\beta}_{ME}$	$\hat{\beta}_{KL}$	$\hat{\beta}$	ρ
۰/۰۴۵۵	/۰۴۷۷.۰	۰/۰۴۵۷	۰/۰۴۷۹	۰/۸۰
۰/۱۰۵۸	۰/۱۱۴۷	۰/۱۰۶۷	۰/۱۱۵۸	۰/۹۰
۰/۲۳۷۰	۰/۲۶۶۹	۰/۲۴۰۵	۰/۲۷۱۳	۰/۹۵

نتایج شبیه‌سازی ارائه شده در جداول ۱ و ۲ بیانگر آن است که با افزایش تعداد نمونه مقدار میانگین مربعات خطای برآوردگرها کاهش می‌یابد. با افزایش سطح هم‌خطی مقدار میاگین مربعات خطای همه برآوردگرها افزایش پیدا می‌کند و در تمام موارد مقدار میانگین مربعات خطای برآوردگر $\hat{\beta}_{MKL}$ نسبت به سایر برآوردگرهای ذکر شده کمتر است. بنابراین در حضور هم‌خطی استفاده از برآوردگر جدید پیشنهاد می‌شود.

بحث و نتیجه‌گیری

در این مقاله یک برآوردگر ترکیبی در مدل‌های خطی با خطا در اندازهگیری تحت محدودیتهای خطی تصادفی معرفی گردید. ویژگی‌های مجانبی برآوردگر معرفی شده بررسی و با استفاده از معیار ماتریس میانگین توان‌های دوم خطا کارایی برآوردگرها مورد ارزیابی قرار گرفت. با استفاده از مطالعه‌ی شبیه‌سازی نشان داده شد که برای مقدار مشخصی از پارامتر اریب برآوردگر معرفی شده کارائی بیشتری نسبت به برآوردگرمحدود شده دارد.

مراجع

Fung, W. K., Zhong, X. P. and Wei, B. C. (2003). On estimation and influence diagnostics in linear mixed measurement error models. *American Journal of Mathematical and Management Sciences,* **23**, 37–59.

Ghapani, F., A. R. Rasekh, and B. Babadi. (2015). Detection of outliers and influential observations in linear ridge measurement error models with stochastic linear restrictions. *Journal of Sciences Islamic Republic of Iran*. **26**(4), 355–66.

Kibria, B. M. G. and Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model. *Simulations and applications. Hindawi Scientifica*. https://doi.org/10.1155/2020/9758378.

McDonald, C., Galarneau, D. A. (1975). A Monte Carlo Evaluation of some Ridge-Type Estimators. *Journal of the American Statistical Association*. **70**, 407-416.

Rao, C. R., Toutenburg, H., Shalabh and Heumann, C. (2008), *Linear Models and Generslizations*, Springer. Berlin.

Investigating the asymptotic performance of the hybrid estimator in the linear measurement error models

Fatemeh Ghapani

Department of Mathematics and Statistics, Sho.C., Islamic Azad University, Shoushtar, Iran

---

Abstract: In this paper, a hybrid estimator is introduced for improving the accuracy of estimators in linear measurement error models with stochastic linear restrictions,and the asymptotic properties of the introduced estimator are investigated.The performance of the introduced estimator is examined using the mean square error matrix criterion compared to some existing estimators.Finally, the performance of the introduced estimator is evaluated using a simulation study.

.

Keywords: Hybrid estimator, Linear measurement error model,Stochastic linear restrictions, Multicollinearitys.

Mathematics Subject Classification (2020): 62J05, 62J07.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مقایسه روش‌های کلاسیک و بیزی در برآورد پارامتر قابلیت اعتماد تنش-مقاومت برای توزیع گمپرتز تعمیم‌یافته یکه

اعظم کاراندیش مروستی [۱]، احسان ارمز [۲]، مریم بصیرت [۳]

[۱] گروه ریاضی و آمار، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

[۲] گروه ریاضی و آمار، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

[۳] گروه ریاضی و آمار، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

چکیده: در این مقاله، برآورد پارامتر قابلیت اعتماد تنش-مقاومت $R = P(Y < X)$ مورد بررسی قرار گرفته است؛ حالتی که در آن X (مقاومت) و Y (تنش) متغییرهای تصادفی مستقلی هستند که از توزیع گمپرتز تعمیم‌یافته یکه (UGG) پیروی می‌کنند. به‌منظور برآورد پارامتر R، از سه رویکرد اصلی شامل روش ماکسیمم درستنمایی (MLE)، روش خودگردان (Bootstrap)، و روش بیزی (Bayesian) با بهره‌گیری از الگوریتم‌های گیبز و متروپلیس-هستینگس استفاده شد. مطالعه حاضر علاوه‌بر شبیه‌سازی داده‌ها، از داده‌های واقعی بارندگی دو منطقه اقلیمی ایران (یاسوج و ساری) طی سال‌های ۱۳۹۰ تا ۱۴۰۲ بهره برد تا اعتبار برآوردها در شرایط واقعی سنجیده شود. نتایج نشان داد که روش بیزی، به‌ویژه با در نظر گرفتن وزن‌دهی متفاوت داده‌ها (نسبت‌های $m = ۰/۳$ و $n = ۰/۷$)، عملکرد پایدارتر و فاصله‌های اعتباری دقیق‌تری نسبت به سایر روش‌ها ارائه می‌دهد. همچنین روش خودگردان توانست در داده‌های محدود، پوشش مناسبی برای برآورد R فراهم کند، اگرچه میانگین برآورد آن نسبت به MLE پایین‌تر بود. در مجموع، برآوردگرهای بیزی در شرایط داده‌های نامتقارن یا کوچک، دقت بالاتری نشان دادند، که این امر در مقایسه با مطالعات پیشین نیز تأیید می‌شود. یافته‌های تحقیق نشان می‌دهند که ترکیب توزیع‌های انعطاف‌پذیر مانند گمپرتز تعمیم‌یافته یکه با رویکردهای بیزی می‌تواند چارچوب قابل اعتمادتری برای تحلیل قابلیت اطمینان سامانه‌ها در حوزه‌های مهندسی و اقلیم‌شناسی فراهم آورد.

واژه‌های کلیدی: توزیع گمپرتز تعمیم‌یافته یکه، تنش-مقاومت، ماکسیمم درستنمایی ، روش بیزی.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62Mxx

---

[۱] سخنران، azam.karandish@iau.ac.ir

۱ مقدمه

توزیع های آماری مهم ترین ابزار در مدل بندی داده های واقعی هستند. با توجه به رشد روز افزون آمار در تمامی زمینه‌های علمی، معرفی توزیع های جدید با انعطاف بیشتر به منظور تبیین و تفسیر بهتر پدیده ها، یک ضرورت انکار ناپذیر است. در دهه های اخیر تلاش های زیادی برای تحقق این موضوع انجام شده است. بیشتر روش هایی که قبل ازسال ۱۹۸۰ میلادی برای ساخت توزیع ها استفاده شده است مبتنی بر معادلات دیفرانسیل بود که به پیدایش دستگاه پیرسنی و معادلات بور انجامید. بسیاری از توزیع های کلاسیک در این قالب می گنجد. بعد از سال ۱۹۸۰ میلادی تا ۲۰۱۴ میلادی بیشتر روش ها مبتنی بر افزودن پارامترها به توزیع های کلاسیک به منظور افزایش انعطاف پذیری، انجام شده است. امروزه استفاده از تجهیزات اندازه گیری مدرن مانند حسگرها در چندین شاخه از علوم کاربردی ازجمله قابلیت اطمینان، مهندسی کیفیت و علوم زیست پزشکی و اجتماعی به سرعت در حال رشد است. این موضوع منجر به ایجاد مجموعه داده‌های جدید با رفتارهای مختلف در چولگی و کشیدگی می شود که نمی توانند به خوبی توسط توزیع های آماری کلاسیک مانند توزیع های نرمال و وایبول توصیف شوند. بر این اساس، خواسته های جدیدی برای توسعه توزیع های انعطاف پذیرتر برای پوشش این پدیده ها به وجود آمده است. در بیشتر موقعیت ها، توزیع های جدید با گسترش توزیع های شناخته شده موجود ساخته می شوند. یکی از توزیع های معروف در قابلیت اطمینان، زیست شناسی و کهولت شناسی توزیع گمپرتز است که گمپرتز (۱۸۲۵) آن را معرفی کرد. ویلکنز (۲۰۰۱) نشان داد که توزیع گمپرتز روابط دقیق یا حدی با توزیع هایی مانند وایبول، لوژستیک تعمیم یافته، نمایی، نمایی دوگانه و گامبل دارد. قابلیت اعتماد تنش–مقاومت که به‌صورت $R = P(Y < X)$ تعریف می‌شود، یکی از معیارهای مهم در تحلیل قابلیت سامانه‌هاست. در این مقاله، دو متغیر تصادفی X و Y به‌ترتیب معرف مقاومت و تنش هستند و فرض می‌شود که هر دو از توزیع گمپرتز تعمیم‌یافته یکه (UGG) تبعیت می‌کنند. تابع چگالی احتمال این توزیع در رابطه (۱.۱) ارائه شده است:

$$f(x) = \theta\lambda x^{-(c+1)} e^{-\frac{\lambda}{c}(x^{-c}-1)}[1 - e^{-\frac{\lambda}{c}(x^{-c}-1)}]^{\theta-1}; \qquad 0 < x < 1,\ \theta > 0,\ \lambda > 0, c > 0 \qquad (1.1)$$

لذا مقدار تحلیلی R به صورت زیر به‌دست می‌آید:

$$R = P(Y < X) = \int_0^1 P(Y < X | X = x) f_X(x) dx = \frac{\theta_2}{\theta_1 + \theta_2}.$$

بنابراین قابلیت اعتماد تنش–مقاومت فقط تابعی از θ_1 و θ_2 است، که صرفاً به پارامترهای شکل توزیع وابسته است.

۲ روش‌های برآورد

۱.۲ روش ماکسیمم درستنمایی (MLE)

تابع لگاریتم درستنمایی داده‌ها از معادله (۱.۲) استخراج شده است:

$$\begin{aligned} l(\theta_1,\theta_2,\lambda,c) &= \log L(\theta_1,\theta_2,\lambda,c) \\ &= m\log(\theta_1)+n\log(\theta_2)+(m+n)\log\lambda-(c+1)\left[\sum_{i=1}^{m}\log x_i+\sum_{i=1}^{n}\log y_i\right] \\ &\quad -\frac{\lambda}{c}\left[\sum_{i=1}^{m}(x_i^{-c}-1)+\sum_{i=1}^{n}(y_i^{-c}-1)\right]+(\theta_1-1)\sum_{i=1}^{m}\log\left(1-e^{-\frac{\lambda}{c}(x_i^{-c}-1)}\right) \\ &\quad +(\theta_2-1)\sum_{i=1}^{n}\log\left(1-e^{-\frac{\lambda}{c}(y_i^{-c}-1)}\right). \end{aligned} \tag{۱.۲}$$

با محاسبه مشتق‌های جزئی نسبت به پارامترها، معادلات به‌دست‌آمده برای θ_1، θ_2، λ و c حل می‌شوند و سپس مقدار $\hat{R}$ از فرمول زیر محاسبه می‌شود:

$$\begin{aligned} \frac{\partial}{\partial\theta_1}l(\theta_1,\theta_2,\lambda,c) &= \frac{m}{\theta_1}+\sum_{i=1}^{m}\ln\big(C(x_i)\big), \\ \frac{\partial}{\partial\theta_2}l(\theta_1,\theta_2,\lambda,c) &= \frac{n}{\theta_2}+\sum_{i=1}^{n}\ln\big(C(y_i)\big), \\ \frac{\partial}{\partial\lambda}l(\theta_1,\theta_2,\lambda,c) &= \frac{m+n}{\lambda}-\Big(\sum_{i=1}^{m}A(x_i)+\sum_{i=1}^{n}A(y_i)\Big) \\ &\quad +(\theta_1-1)\Big(\sum_{i=1}^{m}\frac{A(x_i)e^{-\lambda A(x_i)}}{C(x_i)}\Big)+(\theta_2-1)\Big(\sum_{i=1}^{n}\frac{A(y_i)e^{-\lambda A(y_i)}}{C(y_i)}\Big), \\ \frac{\partial}{\partial c}l(\theta_1,\theta_2,\lambda,c) &= -\Big(\sum_{i=1}^{m}\ln(x_i)+\sum_{i=1}^{n}\ln(y_i)\Big) \\ &\quad +\frac{\lambda}{c}\Big(\sum_{i=1}^{m}\big(A(x_i)-B(x_i)\big)+\sum_{i=1}^{n}\big(A(y_i)-B(y_i)\big)\Big) \\ &\quad -(\theta_1-1)\frac{\lambda}{c}\sum_{i=1}^{m}\frac{A(x_i)+B(x_i)}{C(x_i)}e^{-\lambda A(x_i)} \\ &\quad -(\theta_2-1)\frac{\lambda}{c}\sum_{i=1}^{n}\frac{A(y_i)+B(y_i)}{C(y_i)}e^{-\lambda A(y_i)}, \end{aligned}$$

که در آن $A(x_i)=(x_i^{-(c)}-1)/c$، $B(x_i)=x_i^{-c}\ln(x_i)$ و $C(x_i)=1-e^{-\lambda A(x_i)}$ تعریف شده‌اند. بنابراین برآوردهای ماکسیمم درستنمایی $\hat{\theta}_1$، $\hat{\theta}_2$، $\hat{\lambda}$ و $\hat{c}$ را با قرار دادن مشتق جزئی برابر با صفر و حل همزمان این معادلات می‌توان محاسبه کرد. از این معادلات بلافاصله نتیجه می‌شود که برآوردگر ماکسیمم درستنمایی θ_1 و θ_2 به صورت زیر است:

$$\begin{aligned} \hat{\theta}_1 &= -\frac{m}{\sum_{i=1}^{m}\ln\left(1-e^{-\frac{\hat{\lambda}}{\hat{c}}(x_i^{-\hat{c}}-1)}\right)}, \\ \hat{\theta}_2 &= -\frac{n}{\sum_{i=1}^{n}\ln\left(1-e^{-\frac{\hat{\lambda}}{\hat{c}}(y_i^{-\hat{c}}-1)}\right)}. \end{aligned}$$

شایان ذکر است که برآوردهای λ و c را می‌توان به صورت عددی بدست آورد. با استفاده از ویژگی پایایی برآوردگرهای ماکسیمم درستنمایی، داریم:

$$\hat{R} = \frac{\hat{\theta}_۲}{\hat{\theta}_۱ + \hat{\theta}_۲}.$$

۲.۲ روش خودگردان (Bootstrap)

دو نوع فاصله اطمینان به روش خودگردان بررسی شد: فاصله اطمینان خودگردان پایه و فاصله اطمینان خودگردان صدکی

فاصله‌های اطمینان خودگردان پایه

یک فاصله‌ی اطمینان خودگردان پایه $۱۰۰(۱-\gamma)\%$ برای R به‌صورت زیر می باشد.

$$\left(۲\hat{R} - \hat{R}^*_{((B+۱)(۱-\frac{\gamma}{۲}))}, ۲\hat{R} - \hat{R}^*_{((B+۱)(\frac{\gamma}{۲}))}\right)$$

که در آن $\hat{R}$، برآوردگر ماکسیمم درستنمایی برای R است و $\hat{R}^*_{((B+۱)\eta)}$، $(B+۱)\eta$-امین مقدار مرتب شده از برآوردهای خودگردان $\{\hat{R}^*_s, s = ۱, ۲, \ldots, B\}$ است.

فاصله‌های اطمینان خودگردان صدکی

یک فاصله‌ی اطمینان خودگردان صدکی $۱۰۰(۱-\gamma)\%$ برای R به‌صورت زیر است.

$$\left(\hat{R}^*_{((B+۱)(\frac{\gamma}{۲}))}, \hat{R}^*_{((B+۱)(۱-\frac{\gamma}{۲}))}\right).$$

۳.۲ روش بیزی

با فرض توزیع‌های پیشین گاما برای پارامترها، توزیع پسین پارامترها طبق روابط زیر حاصل می‌گردد.

$$\begin{aligned}
\pi(\theta_۱, \theta_۲, \lambda, c|\boldsymbol{x}, \boldsymbol{y}) =& \frac{L(\boldsymbol{x}, \boldsymbol{y}|\theta_۱, \theta_۲, \lambda, c)\pi(\theta_۱, \theta_۲, \lambda, c)}{\int_{\theta_۱}\int_{\theta_۲}\int_{\lambda}\int_{c} L(\boldsymbol{x}, \boldsymbol{y}|\theta_۱, \theta_۲, \lambda, c)\pi(\theta_۱)\pi(\theta_۲)\pi(\lambda)\pi(c)dcd\lambda d\theta_۲ d\theta_۱} \\
&\propto L(\boldsymbol{x}, \boldsymbol{y}|\theta_۱, \theta_۲, \lambda, c)\pi(\theta_۱)\pi(\theta_۲)\pi(\lambda)\pi(c) \\
&\propto \left(\prod_{i=۱}^{m} x_i\right)^{-(c+۱)} e^{-\frac{\lambda}{c}\sum_{i=۱}^{m}(x_i^{-c}-۱)} \times \prod_{i=۱}^{m}\left(۱ - e^{-\frac{\lambda}{c}(x_i^{-c}-۱)}\right)^{\theta_۱ - ۱} \\
&\times \left(\prod_{i=۱}^{n} y_i\right)^{-(c+۱)} e^{-\frac{\lambda}{c}\sum_{i=۱}^{n}(y_i^{-c}-۱)} \times \prod_{i=۱}^{n}\left(۱ - e^{-\frac{\lambda}{c}(y_i^{-c}-۱)}\right)^{\theta_۲ - ۱} \\
&\times \theta_۱^{m+a_۱-۱} e^{-\theta_۱ b_۱} \times \theta_۲^{n+a_۲-۱} e^{-\theta_۲ b_۲} \times \lambda^{m+n+d-۱} e^{-\lambda e} \times c^{f-۱} e^{-cg}
\end{aligned}$$

توابع چگالی پسین θ_1، θ_2، λ و c عبارتند از:

$$\pi(\theta_1|\lambda,c,\boldsymbol{x}) \propto \theta_1^{m+a_1-1}e^{-\theta_1 b_1}\prod_{i=1}^{m}\left[1-e^{-\frac{\lambda}{c}(x_i^{-c}-1)}\right]^{\theta_1-1},$$

$$\pi(\theta_2|\lambda,c,\boldsymbol{y}) \propto \theta_2^{n+a_2-1}e^{-\theta_2 b_2}\prod_{i=1}^{n}\left[1-e^{-\frac{\lambda}{c}(y_i^{-c}-1)}\right]^{\theta_2-1},$$

$$\pi(\lambda|\theta_1,\theta_2,c,\boldsymbol{x},\boldsymbol{y}) \propto e^{-\frac{\lambda}{c}[\sum_{i=1}^{m}(x_i^{-c}-1)+\sum_{i=1}^{n}(y_i^{-c}-1)]}\lambda^{m+n+d-1}e^{-\lambda e}$$
$$\times\prod_{i=1}^{m}\left(1-e^{-\frac{\lambda}{c}(x_i^{-c}-1)}\right)^{\theta_1-1}\prod_{i=1}^{n}\left(1-e^{-\frac{\lambda}{c}(y_i^{-c}-1)}\right)^{\theta_2-1},$$

$$\pi(c|\lambda,\theta_1,\theta_2,c,\underline{x},\underline{y}) \propto \left(\prod_{i=1}^{m}x_i\prod_{i=1}^{n}y_i\right)^{-(c+1)} e^{-\frac{\lambda}{c}[\sum_{i=1}^{m}(x_i^{-c}-1)+\sum_{i=1}^{n}(y_i^{-c}-1)]}$$
$$e^{-cg}c^{f-1}\prod_{i=1}^{m}\left(1-e^{-\frac{\lambda}{c}(x_i^{-c}-1)}\right)^{\theta_1-1}\prod_{i=1}^{n}\left(1-e^{-\frac{\lambda}{c}(y_i^{-c}-1)}\right)^{\theta_2-1}.$$

از الگوریتم گیبز و متروپلیس–هستینگس برای نمونه‌گیری از توزیع‌های پسین استفاده شد و برآورد بیزی R بر اساس تابع زیان مربعات میانگین و برای تابع زیان لینکس به صورت زیر محاسبه گردید.

تابع زیان مربع خطا باشد، برآورد R عبارت است از:

$$\hat{R}_{BS} = \frac{1}{B-M}\sum_{i=M+1}^{N} R^{(i)}.$$

تابع زیان لینکس باشد برابر است با

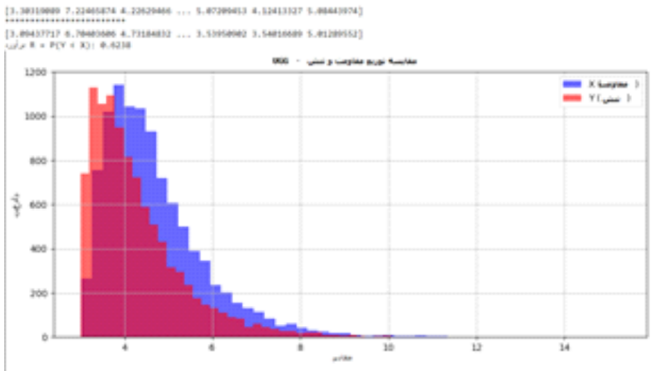
$$\hat{R}_{BL} = -\frac{1}{\phi}\log\left(\frac{1}{B-M}\sum_{i=1+M}^{N} e^{-\phi R^{(i)}}\right)$$

که در آن M دوره سوخت (یعنی تعداد تکرارها قبل از رسیدن به توزیع مانا) است.

۳ تحلیل شبیه‌سازی

با استفاده از داده‌های شبیه‌سازی شده برای مقادیر مختلف θ_1 و θ_2، مقادیر مختلفی از R تولید و با روش‌های بیان شده در بخش دوم برآورد شدند. با استفاده از برنامه پایتون تعداد ۱۰ هزار عدد تصادفی با توزیع گمپرتز تعمیم‌یافته یکه تولید و متناسب با داده های مذکور مقدار قابلیت اعتماد برآورد شده است. نمودار مقایسه‌ای توزیع مشاهدات در مقاومت و تنش در نمودار زیر برای داده های شبیه سازی شده ارائه شده است.

برآورد $R = P(Y < X)$: ۰٫۶۲۳۸



شکل ۱: مقایسه توزیع تنش–مقاومت در داده های شبیه سازی شده

نتایج تحلیل مشاهدات مذکور نشان داد در برآورد ماکسیمم درستنمایی مقدار θ_1، ۸٫۶۹۶ و مقدار θ_2، ۵٫۹۱۴ و مقدار تنش مقاومت برابر ۰٫۴۰۴۸ $(R = ۰٫۴۰۴۸)$ می‌باشد.

ضمناً در روش خودگردان مقدار تنش–مقاومت برابر با ۰٫۶۳۴۸ $(R = ۰٫۶۳۴۸)$ با فاصلهٔ اطمینان ۹۵ درصدی (۰٫۶۶۵۰ و ۰٫۶۰۶۰) به‌دست آمده است. برآورد بیزی با میانگین توزیع پسین ۰٫۴۰۵۰ و تابع لینکس ۰٫۷۰۱۱ که فاصلهٔ اطمینان ۹۵ درصدی (۰٫۴۲۶۳ و ۰٫۳۸۵۰) حاصل شده است.

البته متناسب با پژوهش کاراندیش مروستی و همکاران (۲۰۲۴) می‌توان نتایج بارندگی دو استان کهگیلویه و بویراحمد (یاسوج) و مازندران (ساری) را از سال ۱۳۹۰ تا ۱۴۰۲ با یکدیگر مقایسه و به ترتیب به عنوان تنش–مقاومت در نظر گرفت.

جدول ۱: بارندگی استان‌های کهگیلویه و بویراحمد (یاسوج) و مازندران (ساری) در سال‌های ۱۳۹۰ تا ۱۴۰۲

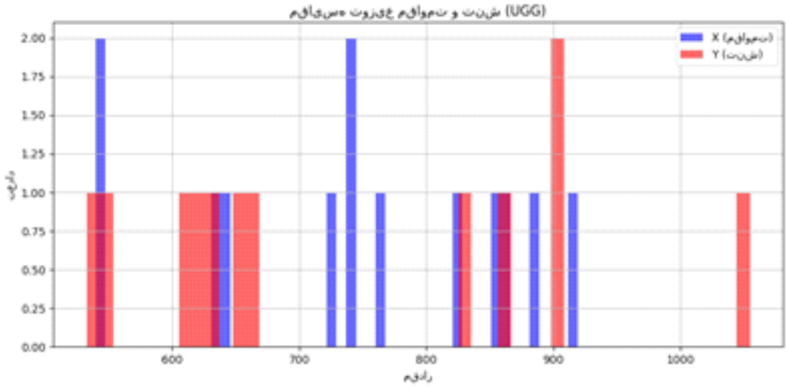
	سال ۱۳۹۰	سال ۱۳۹۱	سال ۱۳۹۲	سال ۱۳۹۳	سال ۱۳۹۴	سال ۱۳۹۵	سال ۱۳۹۶	سال ۱۳۹۷	سال ۱۳۹۸	سال ۱۳۹۹	سال ۱۴۰۰	سال ۱۴۰۱	سال ۱۴۰۲
ساری	۹۱۹	۸۵۵	۷۴۱	۶۴۱	۷۲۵	۴٫۸۲۴	۹٫۵۳۹	۹٫۸۸۷	۷٫۸۶۱	۶۳۱	۴٫۷۴۴	۲٫۵۴۵	۴٫۷۶۶
یاسوج	۶۶۱	۸۲۶	۶۴۸	۶۲۴	۶۱۱	۵٫۵۳۲	۲٫۳۵۶	۷٫۱۰۵۴	۲٫۸۶۵	۲٫۹۰۷	۹٫۶۳۱	۹٫۹۰۱	۹٫۵۴۴

هر چند هر یک از استان ها در برخی از سال ها نسبت به دیگری برتری داشته است؛ برای داده‌های فوق مقدار R به عنوان تنش–مقاومت یک می‌شود.

مهم‌ترین نتایج در جداول فوق عبارتنداز:

- افزایش حجم نمونه موجب کاهش خطای میانگین مربعات (MSE) می‌شود.
- روش بیزی با تابع زیان لینکس کمترین اریبی و خطا را ارائه داد.
- روش صدکی خودگردان و روش گیبز، فاصله‌های اطمینان پوشاننده دقیق‌تری نسبت به MLE فراهم کردند.

تفاوت برآوردهای R در روش کلاسیک و بیزی نمایانگر حساسیت روش‌ها به انتخاب پیشین‌ها و نوع داده‌هاست. نتایج حاصل از روش‌های مختلف در ادامه به صورت خلاصه شده است.



شکل ۲: مقایسه توزیع تنش–مقاومت در داده‌های واقعی شهرهای یاسوج و ساری

جدول ۲: نتایج برآورد برای داده های واقعی

روش	برآورد θ_1 (یاسوج)	برآورد θ_2 (ساری)	مقدار تحلیلی R
روش MLE	۱٫۶۴	۱٫۳۱	۰٫۶۳۶
روش خودگردان	میانگین R	۰٫۴۸۷	
	برآورد فاصله اطمینان ۹۵٪	(۰٫۳۴۶, ۰٫۶۱۵)	
روش بیزی	میانگین پسین R (MSE)	۰٫۷۰۰۸	
	تابع زیان لینکس ($a = ۲$)	۰٫۷۰۰۸	
	برآورد فاصله اطمینان ۹۵٪	(۰٫۶۹۴, ۰٫۷۰۶)	

بحث و نتیجه‌گیری

نتایج این مطالعه نشان داد که در برآورد پارامتر قابلیت اعتماد تنش–مقاومت $R = p(Y < X)$، عملکرد روش‌های مختلف به‌طور معناداری به ساختار داده‌ها و روش مورد استفاده بستگی دارد. در داده‌های واقعی مورد بررسی، روش ماکسیمم درستنمایی (MLE) مقدار بالاتری از R (حدود ۰٫۶۳۶) را نسبت به روش‌های خودگردان و بیزی ارائه داد، که می‌تواند ناشی از تأکید این روش بر ساختار مدل و نادیده‌گیری پراکندگی‌های جزئی در داده‌های محدود باشد. در مقابل، روش خودگردان (Bootstrap) مقدار میانگین پایین‌تری برای R (حدود ۰٫۴۸۷) با فاصله اطمینان گسترده‌تری ارائه کرد، که نشان‌دهنده‌ی حساسیت این روش به پراکندگی‌های نمونه‌ای و قدرت پوشش خوب آن برای داده‌های کوچک است. روش بیزی (Bayesian) با استفاده از توزیع‌های پیشین نمایی ساده و الگوریتم متروپلیس، برآوردی میانی (۰٫۷۰۰۸) با فاصله اعتباری متوازن ارائه داد که با نتایج مطالعات مشابه مانند تاراج و کوتز (۲۰۰۷) [1] و ویو و همکاران (۲۰۱۴)[2] در شرایط داده‌های کوچک و ناهمگن هم‌راستاست. این یافته‌ها بر اهمیت انعطاف‌پذیری روش‌های بیزی به‌ویژه در شرایطی با حجم داده محدود یا توزیع‌های نامتقارن تأکید می‌کند. به‌کارگیری توزیع‌های پیشین مناسب، به‌ویژه پیشین‌های آگاهانه یا بر اساس دانش زمینه‌ای، موجب کاهش اریبی و افزایش دقت برآورد می‌شود. همچنین، تطابق نتایج این تحقیق با مطالعاتی نظیر احمدی

[1] Nadarajah & Kotz
[2] Wu et al.

و زارعی (۲۰۲۰) و کاراندیش مروستی و همکاران (۲۰۲۴) که عملکرد روش‌های بیزی را در تحلیل قابلیت اعتماد برای توزیع‌های تعمیم‌یافته تأیید می‌کنند، بر اعتبار یافته‌های حاضر می‌افزاید.

مراجع

Ahmadi, J., and Zarei, H. (2020), Bayesian inference for the stress-strength reliability parameter based on generalized distributions. *Journal of Statistical Research of Iran (JSRI)*, **17(1)**, 45–62.

Chen, M. H., and Shao, Q. M. (1999), Monte Carlo estimation of bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, **8(1)**, 69–92.

Davison, A. C.,and D. V. Hinkley. (1997), Bootstrap methodsandtheirapplication. *California: Cambridge University Press*.

Downton F. (1973), The estimation of $Pr(Y < X)$ in the normal case. *Technometrics*, **15 (3)**, 551–8. doi: 10.1080/00401706.1973.10489081.

Harris. (1970), The estimation of reliability from stress-strength relationships. *Technometrics*, **12 (1)**, 49–54. doi: 10.1080/00401706.1970.10488633.

Gompertz, B. (1825), On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Esq. FRS & c. Philosophical Transactions of the Royal Society of London*, **115**, 513–583. In a letter to Francis Baily, 1825.

Jha, M. K., S. Dey, and Y. M.Tripathi. (2020), Reliability estimation in a multicomponent stress-strength basedonunit-Gompertzdistribution. *International Journal of Quality & Reliability Management*, **37 (3)**, 428–50. doi: 10.1108/IJQRM-04-2019-0136.

Karandish Marvasti, A., E. Ormoz, and M. Basirat. (2022), Unit generalized Gompertz distribution and its application. *Andishe Ye Amari*, **53 (1)**, 81-92. (In Persian).

Karandish Marvasti, A., E. Ormoz, and M. Basirat. (2024), Estimation of stress-strength reliability based on unit generalized Gompertz distribution. *Communications in Statistics - Theory and Methods*. Doi: 10.1080/03610926.2024.2428988.

Nadarajah, S., and Kotz, S. (2007), Reliability modeling using the generalized Gompertz distribution. *IEEE Transactions on Reliability*, **56(4)**, 570–575.

Willekens, F. (2001), Gompertz in context: the Gompertz and related distributions. *In Forecasting Mortality in Developed Countries, European studies of population*, ed.E.Tabeau,A.vandenBerg Jeths, C. Heathcote, vol. 9, 105–26. Dordrecht: Springer. doi: 10.1007/0-306-47562-6-5.

Wu, J., Zhang, Z., and Zheng, G. (2014), Bayesian estimation of stress-strength reliability using progressive type-II censored data from the generalized distributions. *Computational Statistics & Data Analysis*, **71**, 128–139.

Comparison of Classical and Bayesian Methods in Estimation of Stress-Strength Reliability Parameter for the Unit Generalized Gompertz Distribution

Azam Karandishmarvasti[1], **Ehsan Ormoz**[2], **Maryam Basirat**[3]

[1]Department of Mathematics and statiistics, Ma.C., Islamic Azad University, Mashhad, Iran

[2]Department of Mathematics and statiistics, Ma.C., Islamic Azad University, Mashhad, Iran

[3]Department of Mathematics and statiistics, Ma.C., Islamic Azad University, Mashhad, Iran

---

Abstract: In this paper, the estimation of the stress-resistance reliability parameter $R = P(Y < X)$ is investigated; A case where (X resistance) and (Y tension) are independent random variables that follow the unit generalized Gompertz (UGG) distribution. In order to estimate the parameter R, three main approaches were used, including the maximum likelihood method (MLE), the bootstrap method, and the Bayesian method using the Gibbs and Metropolis-Hastings algorithms. In addition to data simulation, the present study used real rainfall data from two climatic regions of Iran (Yasuj and Sari) during the years 2011 to 2023 to measure the validity of the estimates in real conditions. The results showed that the Bayesian method, especially considering different data weighting (ratio $m = 0.3$ and $n = 0.7$), provides more stable performance and more accurate confidence intervals than other methods. The bootstrap method was also able to provide adequate coverage for estimating R in limited data. Although its average estimate was lower than that of MLE. Overall, Bayesian estimators showed higher accuracy in conditions of asymmetric or small data, which is also confirmed by comparison with previous studies. The research findings indicate that combining flexible distributions such as the generalized Gompertz unit with Bayesian approaches can provide a more reliable framework for analyzing the reliability of systems in the fields of engineering and climatology.

Keywords: Unit generalized Gompertz distribution; stress-strength; maximum likelihood estimation; bayesestimation.

Mathematics Subject Classification (2020): 62Mxx.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

کاربرد روش استوار مونت‌کارلوی بهینه در استنباط بیزی بدون درست‌نمایی برای داده‌های وابسته

امید کریمی [۱] ، فاطمه حسینی
گروه آمار دانشگاه سمنان

چکیده: اغلب موارد در تحلیل فرایندهای سری‌زمانی و فضایی با مدل‌های آماری پیچیده‌ای مواجه می‌شویم که تابع درست‌نمایی آن‌ها براحتی قابل محاسبه نیستند. به همین دلیل، بسیاری از روش‌های استنباط بدون درست‌نمایی با محدودیت‌هایی در دقت و کارایی مواجه هستند. در این مقاله، یک روش استنباط بدون درست‌نمایی با استفاده از روش استوار مونت‌کارلوی بهینه را برای تحلیل داده‌های وابسته فضایی ارائه می‌کنیم. این رهیافت یک چارچوب نوین و کارآمد برای استنباط بدون درست‌نمایی است که نمونه‌های وزنی دقیقی از توزیع پسین ارائه می‌دهد. در نهایت، روش پیشنهادی روی داده‌های شبیه‌سازی شده پیاده‌سازی و ارزیابی می‌شود.

واژه‌های کلیدی: استنباط بیزی، داده‌های وابسته، نمونه‌های مونت کارلویی، استنباط بدون درست‌نمایی.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62F15، 62H11، 62M05.

۱ مقدمه

مدل‌های مبتنی بر شبیه‌ساز به دلیل انعطاف‌پذیری بالایی که در مدل‌سازی ارائه می‌دهند، بسیار جذاب هستند. در واقع، هر مکانیزم تولید داده که بتوان آن را به صورت مجموعه‌ای محدود از گام‌های الگوریتمی تعریف کرد، می‌تواند به عنوان یک مدل مبتنی بر شبیه‌ساز برنامه‌نویسی شود. از این رو، این مدل‌ها اغلب برای شبیه‌سازی پدیده‌های فیزیکی در علوم طبیعی، مانند ژنتیک، اپیدمیولوژی، یا علوم اعصاب، به کار می‌روند (گاتمان و کورندر، ۲۰۱۶). در این مدل‌ها، تولید نمونه‌ها با استفاده از روش‌های مونت کارلویی امکان پذیر است، اما محاسبه تابع درست‌نمایی غیرممکن است. ناتوانی در محاسبه تابع درست‌نمایی، استنباط بدون درست‌نمایی (LFI)[1] ، یعنی تقریب توزیع پسین را به چالشی بزرگ تبدیل می‌کند. روش مونت‌کارلوی بهینه (OMC)[2] که توسط میدز و ولینگ (۲۰۱۵) پیشنهاد شد،

[۱] سخنران، omid.karimi@semnan.ac.ir

[1] Likelihood-Free Inference
[2] Optimization Monte Carlo

یک روش نوین برای LFI است. ایده اصلی این روش، تبدیل مکانیزم تصادفی تولید داده به مجموعه‌ای از فرآیندهای بهینه‌سازی است. گولمیس و همکاران (۲۰۱۶) برخی محدودیت‌های کلیدی OMC را شناسایی و نسخه بهبودیافته‌ای به نام مونت‌کارلوی بهینه استوار (ROMC)[۳] را پیشنهاد کردند. استواری در این روش به معنای کاهش اتکای صرف به اطلاعات موضعی (نقطه‌ای) در فضای پارامتر و استفاده از نواحی پذیرش برای تخمین دقیق‌تر توزیع پسین است. برخلاف روش OMC که تنها بر مقدار بهینه تکیه دارد، ROMC از مجموعه‌ای از نقاط در ناحیه پذیرش استفاده می‌کند و در نتیجه دقت بالاتری دارد. از منظر نظری، روش ROMC در چارچوب کلی فرآیندهای تصادفی تحلیل می‌شود. در واقع، ساختار شبیه‌ساز به‌عنوان نگاشتی از متغیرهای تصادفی کمکی به داده‌ها عمل می‌کند و الگوریتم ROMC با استفاده از نواحی پذیرش در فضای پارامتر، به تخمین توزیع پسین در شرایطی می‌پردازد که تابع درست‌نمایی غیرقابل محاسبه است. از این حیث، ROMC را می‌توان رویکردی مبتنی بر تحلیل هندسی نواحی با احتمال بالا در فرآیندهای تصادفی شرطی دانست

در این مقاله، نحوه پیاده‌سازی ROMC را برای تقریب بهینه توزیع پسین روی داده‌های وابسته ارائه می‌کنیم. ROMC یک چارچوب استنباط بدون درست‌نمایی است که مجموعه‌ای از گام‌های الگوریتمی را برای تقریب توزیع پسین تعریف می‌کند، بدون آنکه الگوریتم خاصی را برای هر گام الزامی کند. بنابراین، ROMC را می توان به عنوان یک روش پایه استفاده کرد. در نهایت با یک مطالعه شبیه‌سازی روش پیشنهادی را ارزیابی و تحلیل می‌کنیم.

۲ مدل‌های مبتنی بر شبیه‌ساز

در این بخش، ابتدا مقدمه‌ای کوتاه درباره مدل‌های مبتنی بر شبیه‌ساز ارائه می‌دهیم، سپس به روش OMC و نسخه استوار آن ROMC می‌پردازیم. مدل مبتنی بر شبیه‌ساز، مکانیزمی تصادفی و پارامتری برای تولید داده است که امکان نمونه‌گیری از داده‌ها را فراهم می‌کند، اما محاسبه تابع درست‌نمایی در آن امکان‌پذیر نیست. در واقع یک مدل مبتنی بر شبیه‌ساز، مجموعه‌ای پارامتری از توابع چگالی احتمال $\{p(\mathbf{y} \mid \boldsymbol{\theta})\}_{\boldsymbol{\theta}}$ بر فضای پارامتر $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^D$ را ارائه می‌دهد که شکل بسته ندارد یا از نظر محاسباتی غیرقابل محاسبه است. در چنین مواردی، تنها دسترسی به شبیه‌ساز $m_r(\boldsymbol{\theta})$ ممکن است، یعنی یک مکانیزم مبتنی بر الگوریتم برنامه‌نویسی که به صورت تصادفی داده‌های $\mathbf{y}$ را از مجموعه‌ای از پارامترها $\boldsymbol{\theta}$ تولید می‌کند. فرآیند نمونه‌گیری از شبیه‌ساز را با $m_r(\boldsymbol{\theta}) \to \mathbf{y}$ نشان می‌دهیم. همان‌طور که میدز و ولینگ (۲۰۱۵) نشان داده‌اند، می‌توان تصادفی بودن شبیه‌ساز را با معرفی مجموعه‌ای از متغیرهای تصادفی $\mathbf{u} \sim p(\mathbf{u})$ جدا کرد. بنابراین، برای یک جفت مشخص $(\boldsymbol{\theta}, \mathbf{u})$، شبیه‌ساز به یک نگاشت قطعی g تبدیل می‌شود، به طوری که $\mathbf{y} = g(\boldsymbol{\theta}, \mathbf{u})$ است. فرض کنید $\mathbf{y}_\circ$ داده‌ی مشاهده‌شده باشد، مشکل اصلی ناتوانی در محاسبه تابع درست‌نمایی $L(\boldsymbol{\theta}) = p(\mathbf{y}_\circ \mid \boldsymbol{\theta})$ است. در واقع درست‌نمایی، احتمال تولید داده $\mathbf{y}$ مشابه داده مشاهده‌شده $\mathbf{y}_\circ$ با استفاده از پارامترهای $\boldsymbol{\theta}$ است. بنابراین، تابع درست‌نمایی $L(\boldsymbol{\theta})$ به‌صورت

$$L(\boldsymbol{\theta}) = \lim_{\epsilon \to \circ} c_\epsilon \int_{\mathbf{y} \in B_{d,\epsilon}(\mathbf{y}_\circ)} p(\mathbf{y} \mid \boldsymbol{\theta}) d\mathbf{y} = \lim_{\epsilon \to \circ} c_\epsilon \Pr(g(\boldsymbol{\theta}, \mathbf{u}) \in B_{d,\epsilon}(\mathbf{y}_\circ) \mid \boldsymbol{\theta})$$

می‌توان نوشت. که در آن c_ϵ یک ضریب تناسب وابسته به ϵ است و $B_{d,\epsilon}(\mathbf{y}_\circ)$ ناحیه‌ای به شعاع ϵ اطراف $\mathbf{y}_\circ$ است که با یک تابع فاصله d به‌صورت $B_{d,\epsilon}(\mathbf{y}_\circ) := \{\mathbf{y} : d(\mathbf{y}, \mathbf{y}_\circ) \leq \epsilon\}$ تعریف می‌شود.

در مواردی که $\mathbf{y}$ بعد بالا باشد از آماره خلاصه T استفاده و سپس فاصله d اعمال می‌شود. در این حالت، $B_{d,\epsilon}(\mathbf{y}_\circ) := \{\mathbf{y} :$

[3] Robust OMC

$d(T(\mathbf{y}), T(\mathbf{y}_\circ)) \leq \epsilon\}$ خواهد شد. محاسبه $\Pr\left(g(\boldsymbol{\theta}, \mathbf{u}) \in B_{d,\epsilon}(\mathbf{y}_\circ) \mid \boldsymbol{\theta}\right)$ به عنوان کسری از نمونه‌هایی که در ناحیه ϵ اطراف $\mathbf{y}_\circ$ قرار دارند، در حالت حدی $\epsilon \to \circ$ از نظر محاسباتی غیرممکن است. بنابراین، محدودیت به $\epsilon > \circ$ کاهش می‌یابد که منجر به درست‌نمایی تقریبی $L_{d,\epsilon}(\boldsymbol{\theta}) = \Pr(\mathbf{y} \in B_{d,\epsilon}(\mathbf{y}_\circ) \mid \boldsymbol{\theta})$ می‌شود. در نتیجه، توزیع پسین تقریبی را می‌توان به‌صورت $p_{d,\epsilon}(\boldsymbol{\theta} \mid \mathbf{y}_\circ) \propto L_{d,\epsilon}(\boldsymbol{\theta})p(\boldsymbol{\theta})$ نوشت. این روش تنها راه‌حل برای مواجهه با مشکل محاسباتی تابع درست‌نمایی نیست، بلکه رویکردهای دیگر شامل مدل‌سازی رابطه (تصادفی) بین $\boldsymbol{\theta}$ و $\mathbf{y}$، معکوس آن، یا تعریف استنباط بدون درست‌نمایی به عنوان یک مسئله براورد نسبت هستند (توماس و همکاران، ۲۰۲۲؛ هرمانس و همکاران، ۲۰۲۰).

۱.۲ مونت‌کارلوی بهینه (OMC)

روش OMC برای اولین بار توسط میدز و ولینگ (۲۰۱۵) ارایه شده است. برای این منظور ابتدا تابع نشانگر $\mu_{B_{d,\epsilon}(\mathbf{y})}(\mathbf{x})$ را به‌صورت $\mu_{B_{d,\epsilon}(\mathbf{y})}(\mathbf{x}) = \begin{cases} ۱ & \text{اگر } \mathbf{x} \in B_{d,\epsilon}(\mathbf{y}) \\ \circ & \text{در غیر این صورت} \end{cases}$ تعریف می‌کنیم، پس، تابع درست‌نمایی تقریبی $L_{d,\epsilon}(\boldsymbol{\theta})$ را می‌توان به‌صورت زیر نوشت:

$$L_{d,\epsilon}(\boldsymbol{\theta}) = \Pr(\mathbf{y} \in B_{d,\epsilon}(\mathbf{y}_\circ) \mid \boldsymbol{\theta}) = \int_{\mathbf{u}} \mu_{B_{d,\epsilon}(\mathbf{y}_\circ)}(g(\boldsymbol{\theta}, \mathbf{u}))d\mathbf{u}.$$

در آن صورت با استفاده از نمونه‌های مونت‌کارلویی برگرفته از شبیه‌ساز به‌صورت زیر تقریب زده می‌شود:

$$L_{d,\epsilon}(\boldsymbol{\theta}) \approx \frac{۱}{N}\sum_{i=۱}^{N} \mu_{B_{d,\epsilon}(\mathbf{y}_\circ)}(g(\boldsymbol{\theta}, \mathbf{u}_i)), \quad \mathbf{u}_i \sim p(\mathbf{u}). \tag{۱.۲}$$

در معادله (۱.۲)، برای هر $\mathbf{u}_i$، ناحیه‌ای به‌صورت $C_\epsilon^i = \{\boldsymbol{\theta} : g(\boldsymbol{\theta}, \mathbf{u}_i) \in B_{d,\epsilon}(\mathbf{y}_\circ)\}$ در فضای پارامتر $\boldsymbol{\theta}$ وجود دارد که تابع نشانگر مقدار یک را برمی‌گرداند. بنابراین، می‌توان درست‌نمایی تقریبی و توزیع پسین را به صورت زیر نوشت:

$$L_{d,\epsilon}(\boldsymbol{\theta}) \approx \frac{۱}{N}\sum_{i=۱}^{N} \mu_{C_\epsilon^i}(\boldsymbol{\theta}), \qquad p_{d,\epsilon}(\boldsymbol{\theta} \mid \mathbf{y}_\circ) \propto p(\boldsymbol{\theta}) \sum_{i=۱}^{N} \mu_{C_\epsilon^i}(\boldsymbol{\theta}). \tag{۲.۲}$$

در واقع، تقریب درست‌نمایی و توزیع پسین به توصیف مجموعه‌های C_ϵ^i تبدیل می‌شود. در روش OMC ، فرض می‌کند که فاصله d، فاصله اقلیدسی $\|\cdot\|_۲$ بین آمار خلاصه T داده‌های مشاهده‌شده و تولیدشده است و نواحی C_ϵ^i می‌توانند به خوبی با بیضی‌های بسیار کوچک تقریبی شوند. این فرضیات منجر به تقریب توزیع پسین با استفاده از نمونه‌های وزنی $\boldsymbol{\theta}_i^*$ می‌شوند که کمترین فاصله بین داده‌های مشاهده‌شده و شبیه‌سازی‌شده را برای هر تحقق $\mathbf{u}_i \sim p(\mathbf{u})$ به‌صورت $\boldsymbol{\theta}_i^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\Phi(\mathbf{y}_\circ) - \Phi(g(\boldsymbol{\theta}, \mathbf{u}_i))\|_۲$ به‌دست می‌آورند.

۲.۲ روش استوار مونت‌کارلوی بهینه (ROMC)

گولمیس و همکاران (۲۰۱۶) نشان دادند که فرض بیضی‌های بسیار کوچک می‌تواند به توزیع‌های پسین بیش از حد مطمئن منجر شود. به طور شهودی، این مشکل به این دلیل رخ می‌دهد که وزن‌ها در OMC تنها بر اساس اطلاعات محلی در $\boldsymbol{\theta}_i^*$ محاسبه می‌شوند و استفاده صرف از اطلاعات محلی می‌تواند گمراه‌کننده باشد. برای مثال، اگر انحنای $\|T(\mathbf{y}_\circ) - T(g(\boldsymbol{\theta}, \mathbf{u}_i))\|_۲$ در $\boldsymbol{\theta}_i^*$ تقریباً صاف باشد، ممکن است به اشتباه نشان دهد که C_ϵ^i بسیار بزرگ‌تر از اندازه واقعی آن است. روش استوار OMC پیشنهادی این مشکل را برطرف می‌کند. ROMC نواحی پذیرش C_ϵ^i را با جعبه‌های مرزی D-بعدی $\hat{C}_\epsilon^i$ تقریبی می‌کند. یک توزیع یکنواخت $q_i(\boldsymbol{\theta})$ روی هر

جعبه مرزی تعریف می‌شود و به عنوان توزیع پیشنهادی برای تولید نمونه‌های پسین $\boldsymbol{\theta}_{ij} \sim q_i$ عمل می‌کند. نمونه‌ها وزن به‌صورت $w_{ij} = \mu_{C_\epsilon^i}(\boldsymbol{\theta}_{ij})\frac{p(\boldsymbol{\theta}_{ij})}{q_i(\boldsymbol{\theta}_{ij})}$ دریافت می‌کنند که استفاده از توزیع‌های پیشنهادی $q_i(\boldsymbol{\theta})$ به جای پیشین $p(\boldsymbol{\theta})$ را تعدیل می‌کند. با توجه به نمونه‌های وزنی، $\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y}_\circ)}[h(\boldsymbol{\theta})]$ از یک تابع $h(\boldsymbol{\theta})$ می‌تواند به‌صورت $\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{y}_\circ)}[h(\boldsymbol{\theta})] \approx \frac{\sum_{ij} w_{ij} h(\boldsymbol{\theta}_{ij})}{\sum_{ij} w_{ij}}$ تقریبی شود. برای محاسبه وزن w_{ij} باید بررسی کنیم که آیا نمونه‌های $\boldsymbol{\theta}_{ij}$، که از جعبه‌های مرزی گرفته شده‌اند، در ناحیه پذیرش C_ϵ^i قرار دارند. این بررسی به عنوان یک مکانیزم ایمنی عمل می‌کند که نادرستی‌های احتمالی در ساخت $\hat{C}_\epsilon^i$ را اصلاح می‌کند. اما این بررسی نیازمند ارزیابی تابع فاصله $d_i(\boldsymbol{\theta}_{ij})$ است که در مدل‌های پیچیده ممکن است پرهزینه باشد. به‌طور کلی، ROMC به دو بخش آموزش و استنباط تقسیم می‌شود. بخش آموزش شامل گام‌هایی برای تخمین نواحی پیشنهادی است و بخش استنباط نمونه‌های وزنی را محاسبه می‌کند. در واقع ابتدا، نواحی پیشنهادی $\hat{C}_\epsilon^i$ تعیین و برآورد می‌شوند و سپس توزیع پسین تقریب زده می‌شود. در ادامه یک مطالعه شبیه‌سازی برای پیاده‌سازی روش پیشنهادی روی داده‌های وابسته ارائه می‌شود.

۳ مطالعه شبیه‌سازی

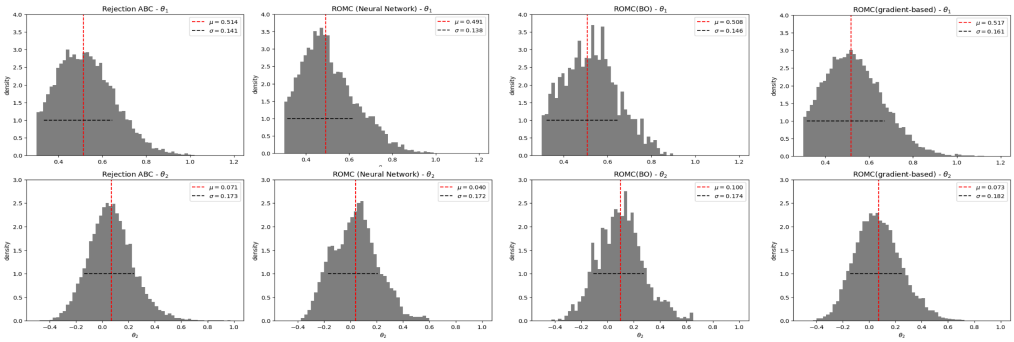
در این بخش، روش پیشنهادی را ابتدا روی مدل میانگین متحرک مرتبه دوم (MA2)، که یکی از مدل‌های فرایندهای سری زمانی است، پیاده‌سازی و ارزیابی می‌کنیم. سپس روی یک مدل رگرسیون فضایی ساده اجرا و تحلیل می‌شود. استنباط با سه رویکرد مختلف روش ROMC انجام می‌شود: (۱) با بهینه‌ساز مبتنی بر گرادیان، (۲) با بهینه‌سازی بیزی و (۳) با برازش یک شبکه عصبی به عنوان مدل جایگزین. مورد آخر نشان‌دهنده قابلیت توسعه روش پیشنهادی است، که در آن بخشی از ROMC با یک روش مشخص توسط کاربر جایگزین می‌شود.

۱.۳ مدل میانگین متحرک مرتبه دوم

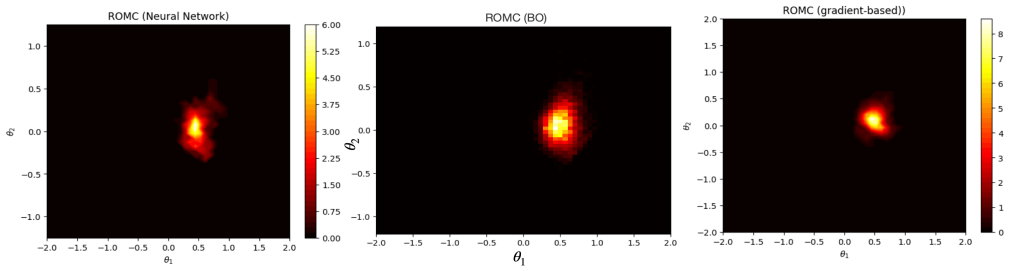
مدل MA2 به‌صورت $y_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}, \quad t = 1, \ldots, K$ تعریف می‌شود، که در آن $\theta_1, \theta_2 \in \mathbb{R}$ و $w_t \sim \mathcal{N}(0, 1)$، است. متغیر تصادفی w_k نوفه سفید است و دو پارامتر مورد نظر، θ_1 و θ_2، وابستگی به مشاهدات قبلی را مدل می‌کنند. پارامتر K تعداد مشاهدات متوالی است که $K = 100$ در نظر گرفته شده است. برای اطمینان از قابل‌شناسایی بودن از پیشین پیشنهادی مارین و همکاران (۲۰۱۲) به‌صورت $p(\boldsymbol{\theta}) = p(\theta_1)p(\theta_2 \mid \theta_1) = \mathcal{U}(\theta_1; -2, 2)\mathcal{U}(\theta_2; \theta_1 - 1, \theta_1 + 1)$ استفاده می‌کنیم. بردار مشاهده $\mathbf{y}_\circ = (y_1, \ldots, y_{100})$ با $\boldsymbol{\theta}^* = (0.6, 0.2)$ تولید شده است. به دلیل بالا بودن بُعد خروجی $\mathbf{y}$، از آماره خلاصه استفاده می‌کنیم. با توجه به اینکه بردار خروجی یک تحقق سری‌زمانی را نشان می‌دهد، از اتوکواریانس‌ها با تأخیر ۱ و ۲ به‌صورت زیر استفاده شده است:

$$
\begin{aligned}
&s_1(\mathbf{y}) = \frac{1}{T-1}\sum_{t=2}^{T} y_t y_{t-1}, \quad s_2(\mathbf{y}) = \frac{1}{T-2}\sum_{t=3}^{T} y_t y_{t-2}, \\
&s(\mathbf{y}) = (s_1(\mathbf{y}), s_2(\mathbf{y})), \quad d = \|s(\mathbf{y}) - s(\mathbf{y}_\circ)\|_2^2,
\end{aligned}
\tag{۱.۳}
$$

فاصله بین مشاهده و خروجی شبیه‌ساز با فاصله اقلیدسی، طبق معادله ۱.۳، اندازه‌گیری می‌شود. همچنین برای شبکه عصبی، از کلاس `MLPRegressor` از بسته `scikit-learn` استفاده می‌کنیم. که در آن $\hat{d}_i$ جایگزین تابع فاصله واقعی d_i در نواحی پیشنهادی می‌شود. بنابراین، تمام اقدامات استنباط، یعنی نمونه‌گیری، محاسبه امید، و ارزیابی پسین، بر اساس $\hat{d}_i$ انجام می‌شود. در این‌جا از یک



شکل ۱: هیستوگرام‌های توزیع‌های پسین حاشیه‌ای برای هر روش



شکل ۲: توزیع‌های پسین حاشیه‌ای پارامترهای مدل فضایی برای سه نسخه ROMC

شبکه عصبی با دو لایه مخفی، هر یک با ۱۰ نورون، استفاده می‌کنیم و آن را با ۵۰۰ نمونه از هر ناحیه پیشنهادی آموزش می‌دهیم. برای مقایسه نتایج استنباط ROMC با الگوریتم سنتی بیز تقریبی ABC، نیز در تحلیل آورده شده است. در جدول ۱، میانگین و انحراف معیار نمونه‌های پسینی پارمترها برای هر روش ارائه شده است. مشاهده می‌شود که نمونه‌های به‌دست‌آمده از توزیع پسین تقریبی در تمام روش‌ها معادل و نزدیک به مقدار واقعی هستند. در شکل ۱، هیستوگرام‌های توزیع‌های پسین حاشیه‌ای پارامترها برای همه روش‌ها

جدول ۱: مقایسه نتایج پسین برای پارامترهای مدل $MA(۲)$ با استفاده از نسخه‌های مختلف روش ROMC و الگوریتم ABC

	$\hat{\mu}_{\theta_۱}$	$\hat{\sigma}_{\theta_۱}$	$\hat{\mu}_{\theta_۲}$	$\hat{\sigma}_{\theta_۲}$
نمونه‌گیری رد ABC	۰/۵۱۶	۰/۱۴۲	۰/۰۷۰	۰/۱۷۲
ROMC (گرادیانی)	۰/۵۰۱	۰/۱۴۲	۰/۰۳۳	۰/۱۶۹
ROMC (بیزی)	۰/۵۱۳	۰/۱۶۹	۰/۰۹۰	۰/۱۷۴
ROMC (شبکه عصبی)	۰/۴۹۱	۰/۱۳۸	۰/۰۴۰	۰/۱۷۲

رسم شده‌اند که نشان می‌دهد توزیع پارامترها در همه روش‌ها نسبتا مشابه به هم هستند. در شکل ۲، توزیع پسین برای سه نسخه مختلف روش ROMC ارائه شده است. نتایج نشان می‌دهند که تمام نسخه‌های ROMC نتایج نسبتا مناسب و یکسانی ارائه می‌دهند که با الگوریتم نمونه‌گیری رد ABC هم‌خوانی دارند. با موازی کردن این روش می‌توان سرعت محاسبات را تا ۵ برابر افزایش داد.

۲.۳ مدل رگرسیون فضایی

در این بخش، روش پیشنهادی را روی مدل رگرسیون فضایی در یک مشبکه منظم ۱۰ × ۱۰ با مختصات $s_i = (i, j)$ و $i, j =$ ۱,...,۱۰ پیاده‌سازی و ارزیابی می‌کنیم. مشابه بخش قبل استنباط با سه رویکرد مختلف روش ROMC انجام می‌شود. مدل

رگرسیون فضایی به‌صورت $y(s_i) = \beta_0 + \beta_1 x(s_i) + w(s_i) + \varepsilon_i$ تعریف شده است، که در آن $i = 1, \ldots, 100$، $x(s_i)$ یک متغیر توضیحی در مکان s_i است که از توزیع یکنواخت $\mathcal{U}(0, 10)$ تولید شده است، $\beta = (\beta_0, \beta_1)$ بردار ضرایب رگرسیون، $w(s_i)$ فرآیند گاوسی با کوواریانس نمایی $\mathrm{Cov}(w(s_i), w(s_j)) = \sigma^2 \exp(-\phi\|s_i - s_j\|)$ با $\sigma^2 = 1$ و $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$ نوفه سفید است. فاصله $\|s_i - s_j\|$ فاصله اقلیدسی بین مکان‌های s_i و s_j روی مشبکه منظم است. پارامترهای مورد نظر برای استنباط $\boldsymbol{\theta} = (\beta_0, \beta_1, \phi, \tau^2)$ هستند. از پیشین‌های معمول زیر استفاده می‌کنیم:

$$p(\boldsymbol{\theta}) = p(\beta_0)p(\beta_1)p(\phi)p(\tau^2) = \mathcal{N}(\beta_0; 0, 10)\mathcal{N}(\beta_1; 0, 10)\mathcal{U}(\phi; 0/1, 5)\mathcal{IG}(\tau^2; 2, 1). \tag{۲.۳}$$

که در آن $\mathcal{IG}$ توزیع گامای معکوس است. بردار مشاهده $\mathbf{y}_0 = (y(s_1), \ldots, y(s_{100}))$ با $\boldsymbol{\theta}^* = (1, 0/5, 1, 0/1)$ تولید شده است. انتخاب آماره‌های خلاصه $T(y)$ برای داده‌های فضایی نقشی کلیدی در عملکرد الگوریتم ROMC ایفا می‌کند. در این مطالعه، از میانگین کلی مشاهدات و واریوگرام تجربی به‌صورت زیر استفاده شده است تا به ترتیب ویژگی‌های سطح کلی و ساختار همبستگی فضایی داده‌ها منعکس گردد. این آماره‌ها نقش تعیین‌کننده‌ای در تعریف تابع فاصله و شکل‌گیری نواحی پذیرش دارند:

$$T_1(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y(s_i), \quad T_2(\mathbf{y}) = \frac{1}{|\mathcal{N}|}\sum_{(i,j)\in\mathcal{N}} (y(s_i) - y(s_j))^2,$$
$$T(\mathbf{y}) = (T_1(\mathbf{y}), T_2(\mathbf{y})), \quad d = \|T(\mathbf{y}) - T(\mathbf{y}_0)\|_2^2,$$

که در آن $\mathcal{N}$ مجموعه جفت‌های مکان‌هایی است که فاصله اقلیدسی آن‌ها کمتر از ۲ واحد است. در واقع از همسایگی‌های نزدیک مشابه استفاده کرده‌ایم. در این‌جا از یک شبکه عصبی با دو لایه مخفی، هر یک با ۱۲ نورون، استفاده می‌کنیم و آن را با ۵۰۰ نمونه از هر ناحیه پیشنهادی آموزش می‌دهیم. برای مقایسه نتایج استنباط ROMC با الگوریتم سنتی نمونه‌گیری رد ABC، نیز در این تحلیل آورده شده است. در جدول ۲، میانگین و انحراف معیار نمونه‌های پسینی پارامترهای β_0، β_1، ϕ، و τ^2 برای همه روش‌ها ارائه شده است. مشاهده می‌شود که نمونه‌های به‌دست‌آمده از توزیع پسین تقریبی در تمام روش‌ها معادل و تقریبا به مقدار واقعی نزدیک هستند. نتایج

جدول ۲: خلاصه نتایج توزیع‌های پسین حاشیه‌ای پارامترهای مدل فضایی برای سه نسخه ROMC و نمونه‌گیری رد ABC

	$\hat{\mu}_{\beta_0}$	$\hat{\sigma}_{\beta_0}$	$\hat{\mu}_{\beta_1}$	$\hat{\sigma}_{\beta_1}$	$\hat{\mu}_{\phi}$	$\hat{\sigma}_{\phi}$	$\hat{\mu}_{\tau^2}$	$\hat{\sigma}_{\tau^2}$
نمونه‌گیری رد ABC	۰/۹۹۲	۰/۱۴۸	۰/۴۹۵	۰/۰۸۵	۱/۰۱۰	۰/۲۱۲	۰/۰۹۸	۰/۰۲۵
ROMC (مبتنی بر گرادیان)	۰/۹۹۸	۰/۱۴۵	۰/۴۹۹	۰/۰۸۳	۱/۰۰۷	۰/۲۰۸	۰/۰۹۹	۰/۰۲۴
ROMC (بهینه‌سازی بیزی)	۰/۹۸۹	۰/۱۵۰	۰/۴۹۴	۰/۰۸۷	۱/۰۱۳	۰/۲۱۵	۰/۰۹۷	۰/۰۲۶
ROMC (شبکه عصبی)	۰/۹۸۷	۰/۱۴۳	۰/۴۹۰	۰/۰۸۲	۱/۰۰۵	۰/۲۱۰	۰/۰۹۶	۰/۰۲۳

نشان می‌دهند که تمام نسخه‌های ROMC نتایج منسجمی ارائه می‌دهند که با الگوریتم نمونه‌گیری رد ABC هم‌خوانی دارند. با وجود قابلیت بالای موازی‌سازی در روش ROMC، این روش همچنان دارای محدودیت‌هایی در کاربردهای واقعی است. حل تعداد زیادی مسئله بهینه‌سازی برای تعیین نقاط θ^* و ساخت نواحی پذیرش، زمان‌بر بوده و نیازمند منابع محاسباتی قابل توجهی است، به‌ویژه در مدل‌هایی با ابعاد بالا یا شبیه‌سازهای پیچیده. همچنین در محیط‌هایی با محدودیت در دسترسی به پردازنده‌های چندهسته‌ای، بهره‌گیری کامل از ظرفیت موازی‌سازی با چالش‌هایی مواجه خواهد بود.

۴ بحث و نتیجه‌گیری

در این مقاله، روش استنباط بدون درست‌نمایی ROMC برای تحلیل داده‌های وابسته، از جمله سری‌های زمانی و داده‌های فضایی، ارائه شد. این روش با استفاده از سه رویکرد مختلف پیاده‌سازی و ارزیابی شد. نتایج مطالعه شبیه‌سازی روی مدل میانگین متحرک مرتبه دوم و مدل رگرسیون فضایی نشان داد که هر سه نسخه ROMC نمونه‌های پسینی با میانگین و انحراف معیار نزدیک به مقادیر واقعی تولید می‌کنند و با الگوریتم نمونه‌گیری رد ABC نتایج مشابهی دارند. همچنین، موازی‌سازی روش ROMC می‌تواند سرعت محاسبات را تا ۵ برابر افزایش دهد، که نشان‌دهنده کارایی بالای این روش در مسائل پیچیده است.

جنبه اصلی این روش این است که ROMC یک چارچوب انعطاف‌پذیر و قابل گسترش ارائه می‌دهد که به کاربران اجازه می‌دهد تا بخش‌های مختلف روش، مانند تابع فاصله یا الگوریتم بهینه‌سازی، را با رویکردهای دلخواه جایگزین کنند و این امکان استفاده گسترده‌تر و توسعه بیشتر توسط پژوهشگران را فراهم می‌کند. با این حال، چالش‌هایی برای بهبود کارایی ،ROMC به‌ویژه در مسائل با ابعاد بالا، همچنان وجود دارد. مهم‌ترین چالش، امکان اجرای ROMC در محیط‌های توزیع‌شده، مانند خوشه‌های محاسباتی، است که می‌تواند زمان محاسبات را به‌طور قابل توجهی کاهش دهد. علاوه بر این، تنظیم تعداد نورون‌ها و لایه‌های شبکه عصبی برای مسائل پیچیده‌تر یا بهینه‌سازی انتخاب آماره‌های خلاصه برای داده‌های فضایی می‌تواند دقت تقریب توزیع پسین را بهبود بخشد. در آینده، تمرکز بر توسعه الگوریتم‌های موازی و بهینه‌سازی‌های عددی می‌تواند قابلیت‌های ROMC را برای کاربردهای گسترده‌تر تقویت کند.

مراجع

Gutmann, M. U. and Corander, J. (2016), *Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models*, *Journal of Machine Learning Research*, **17**, 1–47

Gkolemis, V., Gutmann, M. U. and Pesonen, H. (2024), *An Extendable Python Implementation of Robust Optimization Monte Carlo*, *Journal of Statistical Software*, **110**, 1–26.

Hermans, J., Begy, V. and Louppe, G. (2020), *Likelihood-Free MCMC with Amortized Approximate Ratio Estimators*, *International Conference on Machine Learning*, 4239–4248, PMLR.

Marin, J. M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012), *Approximate Bayesian Computational Methods*, *Statistics and Computing*, **22**, 1167–1180.

Meeds, T. and Welling, M. (2015), *Optimization Monte Carlo: Efficient and Embarrassingly Parallel Likelihood-Free Inference*, *Advances in Neural Information Processing Systems*, **28**.

Thomas, O., Dutta, R., Corander, J., Kaski, S. and Gutmann, M. U. (2022), *Likelihood-Free Inference by Ratio Estimation*, *Bayesian Analysis*, **17**, 1–31.

Robust Optimal Monte Carlo Method

Omid Karimi, Fatemeh Hosseini

Department of Statistics, Semnan University, Semnan, Iran

Abstract: In most cases in the analysis of time-series and spatial processes, we encounter complex statistical models whose likelihood functions are not easily computable. For this reason, many likelihood-free inference methods face limitations in terms of accuracy and efficiency. In this paper, we present a likelihood-free inference method using an optimal robust Monte Carlo approach for analyzing spatially dependent data. This approach provides a novel and efficient framework for likelihood-free inference, offering accurate weighted samples from the posterior distribution. Finally, the proposed method is implemented and evaluated on simulated data.

Keywords: Bayesian inference, dependent data, Monte Carlo methods, likelihood-free inference.

Mathematics Subject Classification (2020): 65C05, 90C26, 62H11.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مفصل‌های دایره‌ای-خطی: ویژگی‌ها و کاربردها

نجیب‌الله کریمی [۱] ، علی دست‌برآورده[۲]

[۱]دانشجوی دکتری آمار، دانشگاه یزد

[۲]عضو هیأت علمی گروه آمار، دانشگاه یزد

چکیده: این مقاله مطالعه مروری از تحقیقات جدید در زمینه مفصل‌های دایره‌ای-خطی است. برای تحلیل ساختار وابستگی داده‌های دایره‌ای و خطی در حالت‌های مختلف، چند مفصل دایره‌ای-خطی نیز مورد مطالعه قرارگرفته‌اند. علاوه‌بر این، به‌معرفی و بکارگیری بسته cylcop در نرم‌افزار R نیز پرداخته شده است. در نهایت، از طریق شبیه‌سازی، مفصل‌های ارائه شده مورد بررسی قرارگرفته و پارامترهای آنها برآورد شده‌اند. نتایج نشان می‌دهند که مفصل‌های با بخش‌های ویژه برای داده‌های حرکتی مناسب هستند.

واژه‌های کلیدی: داده‌های حرکتی، داده‌های دایره‌ای-خطی، مفصل با بخش‌های ویژه، مفصل دایره‌ای-خطی.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62H11، 62G30، 62F99.

---

۱ مقدمه

در بسیاری از حوزه‌های علمی مانند هواشناسی، زیست‌شناسی، علوم زمین و روان‌شناسی با داده‌هایی مواجه هستیم که ماهیتی زاویه‌ای دارند. این داده‌ها که به داده‌های دایره‌ای معروف هستند، بر روی محیط یک دایره واحد تعریف می‌شوند و ویژگی اصلی آن‌ها تناوب و برابری در انتها و ابتدای مقیاس اندازه‌گیری است. نمونه‌هایی از داده‌های دایره‌ای عبارت‌اند از: جهت باد، چرخه‌های زمان در طی شبانه‌روز، زاویه‌های فاز در پردازش سیگنال و جهت حرکت حیوانات. برای مثال، زوایای ۰ درجه و ۳۶۰ درجه در یک فضای دایره‌ای معادل هستند. به همین دلیل، روش‌های آماری متداول که بر پایه مفروضات داده‌های خطی بنا شده‌اند، برای تحلیل این نوع داده‌ها مناسب نیستند (ماردیا و جاپ، ۲۰۰۰).

در بسیاری از مطالعات، متغیر دایره‌ای با یک یا چند متغیر خطی همراه است. برای مثال، در بررسی رفتار حیوانات، ممکن است جهت حرکت (دایره‌ای) با سرعت حرکت (خطی) ترکیب شوند یا در تحلیل باد، سرعت و جهت به‌طور توأم بررسی شوند. به داده‌هایی‌که

---

[۱] سخنران، najibullahkarimi110@gmail.com

از رفتار حیوانات یا وزیدن باد به‌دست می‌آیند، داده‌های حرکتی نیز می‌گویند. در رفتار حیوانات، داده‌های مربوط به جهت حرکت را معمولاً تحت عنوان زاویه‌های چرخش و داده‌های مربوط به سرعت حرکت را تحت عنوان طول‌های گام بررسی می‌کنند (هودل و فبریگ، ۲۰۲۱).

در چنین مواردی، درک دقیق وابستگی بین متغیرهای دایره‌ای و خطی برای تفسیر درست پدیده‌ها ضروری است.

یکی از روش‌های منعطف و قدرتمند برای مدل‌سازی وابستگی بین متغیرها، استفاده از مفصل‌ها است. مفصل‌ها این امکان را فراهم می‌کنند که ساختار وابستگی میان متغیرها جدا از توزیع‌های حاشیه‌ای مدل‌سازی شود (نلسون، ۲۰۰۶). با این وجود، توسعه مفصل‌هایی‌که به‌طور خاص برای مدل‌سازی وابستگی بین متغیرهای دایره‌ای و خطی طراحی شده باشند، نسبتاً جدید و در حال گسترش است. این مفصل‌ها، قادرند با حفظ ماهیت تناوبی متغیرهای دایره‌ای و خواص متغیر خطی، ساختار وابستگی را به دقت مدل کنند (دورنت و سمپی، ۲۰۱۵). از سوی دیگر، پژوهش‌هایی مانند نگار و همکاران (۲۰۲۳) با تمرکز بر کاربردهای عملی مدل‌سازی وابستگی دایره‌ای–خطی در داده‌های زیستی و پزشکی، مسیرهای جدیدی را در این داده‌ها گشوده‌اند.

این مقاله، علاوه بر مقدمه، سه بخش دیگر دارد. در بخش دوم به مفاهیم پایه مرتبط با متغیرهای دایره‌ای و مفصل‌های خطی پرداخته شده و محدودیت‌های این مفصل‌ها در مدل‌سازی وابستگی بین داده‌های دایره‌ای–خطی، بیان شده است. در بخش سوم، مفصل‌های دایره‌ای–خطی معرفی شده و ویژگی‌های آن‌ها مورد بررسی قرار گرفته است. همچنین، به کاربرد و بکارگیری بسته cylcop نیز اشاره شده است. در بخش چهارم، از طریق شبیه‌سازی، کاربردهای مفصل‌های دایره‌ای–خطی در مدل‌سازی وابستگی بین متغیرهای دایره‌ای و خطی، بیان شده است.

۲ مفاهیم پایه

۱.۲ متغیرهای دایره‌ای

متغیرهای دایره‌ای به متغیرهای اطلاق می‌شوند که مقادیر آنها در فضای دایره‌ای تعریف شده و خاصیت تناوبی دارند. این متغیرها که معمولاً با $\theta \in [0, 2\pi)$ نمایش داده می‌شوند، در تحلیل‌های جهت‌دار و زاویه‌ای کاربرد فراوانی دارند. از جمله مهم‌ترین کاربردهای عملی می‌توان به جهت باد در هواشناسی، زوایای حرکت در ناوبری و زمان وقوع پدیده‌های تناوبی در زیست‌شناسی اشاره کرد. مدل‌سازی توزیع‌های احتمالی متغیرهای دایره‌ای موسوم به آمار جهتی است (ماردیا و جاپ، ۲۰۰۰). تابع چگالی احتمال یک متغیر دایره‌ای، Θ، تابعی با دوره تناوب 2π است که در رابطه زیر صدق می‌کند:

$$f_\Theta(\theta) = f_\Theta(\theta + 2\pi), \quad \theta \in [0, 2\pi),\ k \in \mathbb{Z}. \tag{۱.۲}$$

مدل‌های توزیع خطی ممکن است برای متغیرهای دایره‌ای نامعتبر باشند. بنابراین، استفاده از مدل‌های توزیع خاص برای متغیرهای دایره‌ای ضروری است. به‌عنوان مثال می‌توان به توزیع فون‌میزس، توزیع کوشی پوشانده‌شده و توزیع نرمال پوشانده‌شده اشاره کرد. در اینجا فقط به توزیع فون‌میزس که به عنوان معادل دایره‌ای توزیع نرمال شناخته می‌شود، اشاره می‌کنیم. تابع چگالی این توزیع به‌صورت زیر تعریف می‌شود:

$$f(\theta; \mu, k) = \frac{e^{k\cos(\theta-\mu)}}{2\pi I_0(k)}, \tag{۲.۲}$$

که در آن $\mu \in [0, 2\pi)$ پارامتر مکان (میانگین جهت)، $k \geq 0$ پارامتر تمرکز و $I_0(k)$ تابع بسل اصلاح‌شده از نوع اول و مرتبه صفر است (ابراویتز و استیگون، ۱۹۶۵). این توزیع در مدل‌سازی بسیاری از پدیده‌های طبیعی با ماهیت تناوبی کاربرد گسترده‌ای دارد.

۲.۲ مفصل و ساختار وابستگی

نظریه مفصل بر پایه قضیه اسکلار (اسکلار، ۱۹۵۹) بنا شده است و روش منعطف‌تر و مناسب‌تر برای ساخت توزیع توأم چندمتغیره (که در اینجا ما دومتغیره آن را بررسی می‌کنیم) فراهم می‌آورد. قضیه اسکلار بیان می‌کند که تابع توزیع توأم چندمتغیره را می‌توان به‌صورت ترکیبی از یک تابع مفصل که ساختار وابستگی بین متغیرها را توصیف می‌کند و توابع توزیع تجمعی حاشیه‌ای برای تمام متغیرهای حاشیه‌ای نوشت. در مفصل‌ها، ساختار وابستگی میان متغیرها جدا از توزیع هر متغیر در نظر گرفته می‌شود. بنابراین، مدل مناسب برای توزیع حاشیه‌ای هر متغیر را می‌توان به‌طور مستقل و آزادانه انتخاب کرد. فرض کنید $F_{X_1,X_2}(x_1, x_2)$ و $f_{X_1,X_2}(x_1, x_2)$ به‌ترتیب تابع توزیع تجمعی توأم و تابع چگالی احتمال توأم برای متغیرهای تصادفی (X_1, X_2) باشند. در این‌صورت، برای هر $(x_1, x_2) \in \mathbb{R}^2$، با استفاده از یک تابع مفصل C داریم (وانگ و همکاران، ۲۰۲۱):

$$F_{X_1,X_2}(x_1, x_2) = C(u_1, u_2), \tag{۳.۲}$$

$$f_{X_1,X_2}(x_1, x_2) = c(u_1, u_2) f_{X_1}(x_1) f_{X_2}(x_2), \tag{۴.۲}$$

که در آن $u_i = F_{X_i}(x_i)$، $i = 1, 2$، مقدار تابع توزیع حاشیه‌ای X_i در نقطه x_i، $f_{X_i}(x_i)$ مقدار تابع چگالی متناظر با آن و c تابع چگالی مفصل C هستند. تابع چگالی c به‌صورت زیر محاسبه می‌گردد:

$$c(u_1, u_2) = \frac{\partial^2 C(u_1, u_2)}{\partial u_1 \partial u_2}. \tag{۵.۲}$$

تعریف ۱.۲. مفصل تجربی: با استفاده از یک نمونه شامل n مشاهده‌ی دوبُعدی مستقل و هم‌توزیع به‌صورت (x_i, y_i) که از یک توزیع توأم استخراج شده‌اند، می‌توان مفصل تجربی را به‌صورت زیر محاسبه کرد (جینیست و فوری، ۲۰۰۷):

$$C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} I\Big(\frac{r_i}{n+1} \leq u, \frac{s_i}{n+1} \leq v\Big), \tag{۶.۲}$$

که در آن $I(\cdot)$ تابع نشانگر، r_i و s_i به‌ترتیب رتبه مشاهدات x_i و y_i هستند. هنگامی‌که یکی از متغیرهای تصادفی ماهیت دایره‌ای داشته باشد، می‌توان همان رویه را دنبال کرد، اما برای تعیین رتبه‌های متغیر دایره‌ای، نیاز به تعریف یک مرجع خواهیم داشت. در این مقاله، مقدار مرجع را $-\pi$ در نظر می‌گیریم.

۳.۲ محدودیت‌های مفصل‌های کلاسیک برای داده‌های دایره‌ای-خطی

هنگامی‌که یکی از متغیرها دایره‌ای و دیگری خطی باشد، مفصل‌های کلاسیک با چالش‌های اساسی مواجه می‌شوند. مشکل اصلی ناشی از ناسازگاری ذاتی بین ساختار تناوبی متغیرهای دایره‌ای و مفروضات اساسی مفصل‌های کلاسیک است (جانسون و وهرلی، ۱۹۷۸). این ناسازگاری می‌تواند منجر به برآوردهای نادرست و تفسیرهای اشتباه از روابط بین متغیرها شود. علاوه‌بر این، پیچیدگی‌های محاسباتی ناشی از نیاز به تبدیل‌های خاص برای متغیرهای دایره‌ای، استفاده از روش‌های استاندارد را با دشواری مواجه می‌سازد. در مفصل‌های کلاسیک، فرض بر این است که داده‌ها از یک ترتیب ویژه پیروی می‌کنند، در حالی‌که داده‌های دایره‌ای دارای ماهیت تناوبی‌اند و آغاز و

پایان مشخص ندارند. همچنین، استفاده از توزیع‌های حاشیه‌ای نامناسب برای داده‌های دایره‌ای در چارچوب مفصل‌های کلاسیک می‌تواند به نتایج آماری گمراه‌کننده منجر شود. برای رفع این چالش‌ها، مفصل‌های سازگار با داده‌های دایره‌ای-خطی ضروری به‌نظر می‌رسد.

۳ مفصل دایره‌ای-خطی

۱.۳ معرفی و ویژگی‌ها

مفصل‌های دایره‌ای-خطی از دهه ۱۹۷۰ در ادبیات آماری از جمله در آثار جانسون و وهرلی (۱۹۷۸، ۱۹۷۷) مطرح بوده‌اند. هرچند در آن زمان هنوز واژه «مفصل» به‌کار نمی‌رفت. از آن زمان تا کنون، این حوزه موضوع پژوهش‌های فعال بوده است. فرض کنید Θ یک متغیر تصادفی دایره‌ای پیوسته و X یک متغیر تصادفی خطی پیوسته باشند. در حالی‌که X روی کل خط حقیقی $(x \in \mathbb{R})$ و متغیر زاویه‌ای Θ روی دایره واحد $(\theta \in S)$ تعریف می‌شوند، در اینجا ما آن را به‌صورت بازه‌ای از طول 2π، یعنی $[a, a+2\pi)$ در نظر می‌گیریم.

هر تابع چگالی احتمال دایره‌ای-خطی پیوسته روی زیرمجموعه‌ای از فضای استوانه‌ای $(S \times \mathbb{R})$ تعریف شده است. بنابراین، دامنه مفصل دایره‌ای-خطی نیز به جای مربع واحد $([0,1]^2)$، سطح یک استوانه با ارتفاع واحد و محیط واحد خواهد بود و با این فرض، در ادامه u را به‌عنوان بعد دایره‌ای و v را به‌عنوان بُعد خطی در نظر می‌گیریم. برای اختصار، مفصل‌هایی‌که چگالی آن‌ها دارای ویژگی زیر باشند را «دوره‌ای[1] در جهت u» یا به اختصار «دوره‌ای» می‌نامیم (ویژگی اصلی یک مفصل دایره‌ای-خطی):

$$c(0, v) = c(1, v), \ \forall v \in [0, 1]. \tag{۱.۳}$$

بررسی رفتار حیوانات یکی از واضح‌ترین کاربردهای مفصل‌های دایره‌ای-خطی است. البته نتایج آن را می‌توان برای بسیاری از داده‌های دایره‌ای-خطی دیگر نیز تعمیم داد. در رفتار حیوانات، برای نمایش زاویه‌های چرخش معمولاً بازه $\Theta \in [-\pi, \pi)$ را در نظر می‌گیرند. در این بازه، زوایای منفی به چرخش به چپ و زوایای مثبت به چرخش به راست اشاره دارند. تکیه‌گاه متغیر تصادفی X (طول گام)، به اعداد حقیقی نامنفی محدود شده است. زاویه‌های چرخش تا حدودی با سایر متغیرهای دایره‌ای متفاوت‌اند، زیرا منطقی است فرض کنیم که یک حیوان تمایل ذاتی به چرخش به چپ یا راست ندارد. این فرض را می‌توان با استفاده از یک تابع چگالی احتمال حاشیه‌ای که حول $\theta = 0$ متقارن است، به همراه مفصلی‌که نه تنها چگالی آن دوره‌ای است (به رابطه ۱.۳ توجه کنید)، بلکه در بُعد u نیز حول $u = 0/5$ متقارن است، محقق ساخت (هودل و فبریگ، ۲۰۲۱). منظور ما از مفصل متقارن در u یا به‌طور خلاصه مفصل متقارن، مفصلی است که تابع چگالی آن در رابطه زیر صدق می‌کند:

$$c(u, v) = c(1-u, v) \qquad \forall u, v \in [0, 1]. \tag{۲.۳}$$

هرچند انواع دیگر تقارن نیز برای مفصل‌ها قابل تعریف است، اما در اینجا به همین تقارن بسنده می‌کنیم. (به نلسون (۲۰۰۶) مراجعه شود). هر مفصل متقارن، الزاماً دوره‌ای نیز است.

[1] Periodic

۲.۳ چند مفصل دایره‌ای-خطی

در این بخش، چند مفصل دایره‌ای-خطی که ویژگی‌های مورد نیاز در رابطه (۱.۳) را دارند، معرفی می‌شوند. این مفصل‌ها برای مدل‌سازی وابستگی بین متغیرهای دایره‌ای و خطی طراحی شده‌اند.

- **مفصل جانسون-وهرلی:** مفصل جانسون- وهرلی به‌صورت زیر تعریف می‌شود (وانگ و همکاران، ۲۰۲۱):

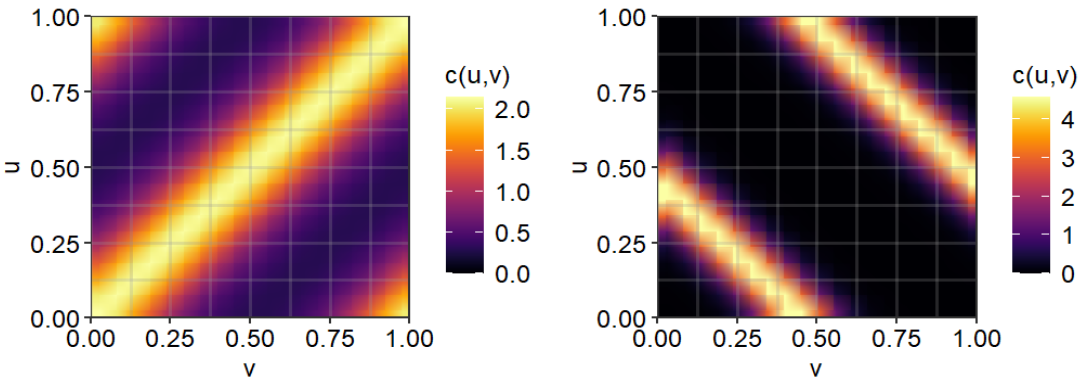
$$c(u,v) = 2\pi g(\eta), \quad \eta = \begin{cases} 2\pi(u-v), & u \geq v \\ 2\pi(u-v)+2\pi, & u < v, \end{cases} \tag{۳.۳}$$

$$C(u,v) = \int_0^u \mathrm{d}w_1 \int_0^v 2\pi g[2\pi(w_1-w_2)]\mathrm{d}w_2 = \int_0^u [-G[2\pi(w_1-w_2)]|_0^v]\mathrm{d}w_1, \tag{۴.۳}$$

که در آن $\eta \in [0, 2\pi]$ یک متغیر دایره‌ای، $g(\cdot)$ تابع چگالی احتمال η و $G(\cdot)$ تابع توزیع متناظر با $g(\cdot)$ هستند. در رابطه (۳.۳)، اگر به جای $g(\cdot)$ توزیع فون‌میزس را قرار دهیم، به‌مفصل به‌دست آمده مفصل فون‌میزس استوانه‌ای می‌گویند. این مفصل به‌صورت زیر است:

$$C(u,v) = uv + \frac{1}{2\pi^2 I_0(k)} \sum_{j=1}^{\infty} I_j(k)\Big(\frac{\cos[2\pi u - 2\pi v - \mu]}{j^2} + \frac{-\cos[j(-2\pi v - \mu)] + \cos[-j\mu]}{j^2}\Big), \tag{۵.۳}$$

که در آن $I_j(\cdot)$ تابع بسل اصلاح‌شده‌ی مرتبه j است (ابرامویتز و استیگون، ۱۹۶۵). پارامتر μ تابع چگالی احتمال را در جهت v جابجا می‌کند و k تمرکز را تعیین می‌نماید. تابع چگالی این مفصل در دو حالت مختلف ($\mu = 0, k = 1$ و $\mu = \pi, k = 4$) در شکل ۱ نشان داده شده است. شکل‌های سمت چپ و راست به‌ترتیب نشان می‌دهند که مفصل همبستگی مثبت دارد یا منفی. برای $k = 0$ مفصل استقلال را داریم و برای $k = \infty$ با شرط همبستگی مثبت به کران بالایی (M) و با شرط همبستگی منفی به کران پایینی (W) فرشه-هافدینگ میل می‌کند. در حالی‌که مفصل فون‌میزس استوانه‌ای یک مفصل دایره‌ای-خطی و دارای ویژگی تناوبی است، اما متقارن نیست؛ موضوعی‌که کارایی آن را برای مدل‌سازی داده‌های حرکتی محدود می‌کند. در ادامه، مفصل‌هایی معرفی می‌شوند که چگالی‌های آن‌ها هم تناوبی و هم متقارن در بُعد u $(u = 0/5)$ هستند.



شکل ۱: چگالی مفصل فون‌میزس استوانه‌ای در دو حالت (چپ: $\mu = 0, k = 1$؛ راست: $\mu = \pi, k = 4$).

- **مفصل چندجمله‌ای مرتبه دو**[2]: مفصل چندجمله‌ای مرتبه دو به‌صورت زیر تعریف می‌شود:

$$c(u,v) = ۱ + ۲\pi\alpha\cos(۲\pi u)(۱ - ۲v), \tag{۶.۳}$$

$$C(u,v) = uv + \alpha\sin(۲\pi u)v(۱ - v), \tag{۷.۳}$$

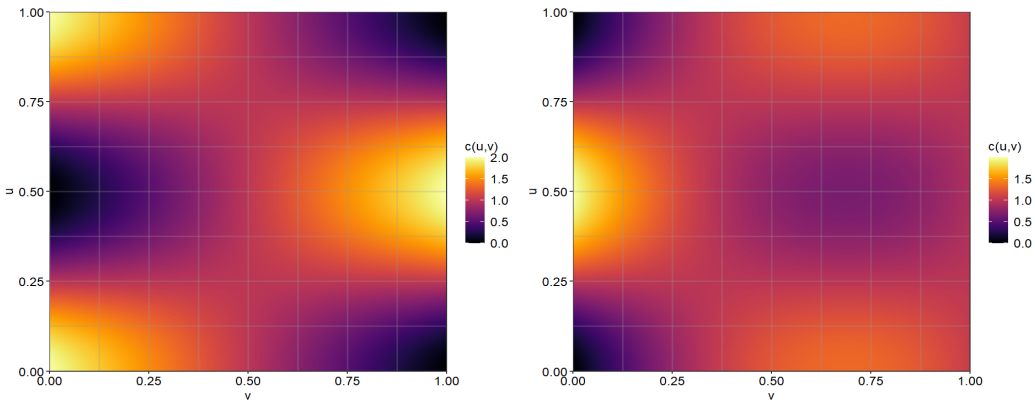
که در آن α پارامتر مفصل است و $|\alpha| \leq (۲\pi)^{-۱}$.

- **مفصل چندجمله‌ای مرتبه سه**[3]: به‌طور مشابه، مفصل چندجمله‌ای مرتبه سه به‌صورت زیر تعریف می‌گردد:

$$c(u,v) = ۱ + ۲\pi\alpha\cos(۲\pi u)(۱ - ۴v + ۳v^۲) + ۲\pi\beta\cos(۲\pi u)(۲v - ۳v^۲), \tag{۸.۳}$$

$$C(u,v) = v(۱ - v)[\alpha\sin(۲\pi u)(۱ - v) - \beta\sin(۲\pi u)v], \tag{۹.۳}$$

که در آن $|\alpha| \leq (۲\pi)^{-۱}$ و $|\beta| \leq (۲\pi)^{-۱}$. برای آسانی کار، از مفصل چندجمله‌ای مرتبه دو به‌عنوان مفصل مرتبه دو و از مفصل چندجمله‌ای مرتبه سه به‌عنوان مفصل مرتبه سه یاد می‌کنیم. آشکار است که مفصل‌های دایره‌ای–خطی مرتبه دو و مرتبه سه در بُعد u دایره‌ای متناوب و متقارن و در بُعد v خطی هستند. مفصل‌های مرتبه دو و مرتبه سه در مقاله هودل و فبریگ (۲۰۲۱) با جزئیات ارائه شده‌اند. همچنین، هودل (۲۰۲۲) بسته cylcop را در نرم‌افزار R برای مفصل‌های دایره‌ای–خطی با تقارن زاویه‌ای، ارائه کرد. توابع چگالی این دو مفصل در شکل ۲ رسم شده است. در این مفصل‌ها برای $\alpha = ۰$ (و $\beta = ۰$) مفصل استقلال به‌دست می‌آید، اما کران‌های فرشه–هافدینگ را به‌دست آورده نمی‌توانیم. برای $\alpha > ۰$ (و $\beta < ۰$)، مفصل همبستگی مثبت را بین u و v نشان می‌دهد هرگاه $u \in [۰, ۰/۵]$ و همبستگی منفی را نشان می‌دهد هرگاه $u \in [۰/۵, ۱]$. برای حالت مقابل یعنی $\alpha < ۰$ (و $\beta > ۰$) نیز این جهت‌های همبستگی عکس حالت قبل هستند.



شکل ۲: چگالی‌های مفصل دایره‌ای–خطی با بخش‌های ویژه نسبت به v. چپ: چگالی مفصل دایره‌ای–خطی مرتبه دو ($\alpha = \frac{۱}{۲\pi}$)، راست: چگالی مفصل دایره‌ای–خطی مرتبه سه ($\alpha = -\frac{۱}{۲\pi}, \beta = ۰/۰۱$).

[2] Copula with Quadratic Sections

[3] Copula with Cubic Sections

۳.۳ برآورد پارامترها

با در اختیار داشتن مجموعه‌ای شامل n مشاهده‌ی مستقل و هم‌توزیع $(i.i.d)$ از یک جفت متغیر دایره‌ای و خطی (θ_i, x_i)، یک خانواده مفصل مفروض $C(u, v; \boldsymbol{\alpha})$ و توزیع‌های حاشیه‌ای مفروض $F_\Theta(\theta; \boldsymbol{\beta})$ و $F_X(x; \boldsymbol{\gamma})$، به‌طور نظری می‌توانیم بردار پارامترهای $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ را از طریق ماکسیمم‌سازی لگاریتم تابع درستنمایی توأم برآورد کنیم. با این حال، این رویکرد مستلزم بهینه‌سازی هم‌زمان تعداد زیادی از پارامترها است و مشخص‌سازی نادرست توزیع‌های حاشیه‌ای می‌تواند منجر به برآوردهای اریب برای پارامترهای مفصل شود و برعکس. اگر توزیع‌های حاشیه‌ای معلوم باشند، لگاریتم تابع درستنمایی به‌صورت زیر خواهد بود:

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \log \left[c\big(F_\Theta(\theta_i), F_X(x_i); \boldsymbol{\alpha}\big) \right]. \tag{۱۰.۳}$$

این رویکرد به دو شیوه جایگزین برای برآورد پارامترهای مفصل منتهی می‌شود که در آنها برآوردهایی از F_Θ و F_X در رابطه (۱۰.۳) جای‌گذاری می‌شوند. این برآوردها، یعنی $\hat{F}_\Theta(\theta)$ و $\hat{F}_X(x)$ می‌توانند به‌صورت پارامتری با استفاده از روش ماکسیمم درستنمایی (MLE)[۴] و ناپارامتری به‌صورت شبه‌مشاهده، یعنی نمونه‌گیری از مفصل تجربی حاصل شوند (هودل و فبریگ، ۲۰۲۱). برآورد پارامترهای مفصل با استفاده از حاشیه‌های ناپارامتری برآوردشده، به‌عنوان برآورد ماکسیمم شبه‌درستنمایی (MPLE)[۵] شناخته می‌شود. برای بسیاری از مفصل‌های خطی–خطی، برآورد پارامترها را می‌توان به‌صورت تحلیلی از روی برآورد یک معیار وابستگی مانند ضریب اسپیرمن یا ضریب کندال به‌دست آورد؛ اما برای بیش‌تر مفصل‌های معرفی‌شده در این مقاله، راه‌حل تحلیلی در دسترس نیست و بنابراین نیاز به بهینه‌سازی عددی داریم.

برای سایر مفصل‌های دایره‌ای–خطی، می‌توان از جستجوی شبکه‌ای[۶] بهره گرفت تا پارامترهایی را یافت که اختلاف یک معیار وابستگی بین مفصل تجربی و مفصل پارامتری را مینیمم کند (هودل و فبریگ، ۲۰۲۱). در نهایت، انتخاب مدل را می‌توان بر پایه معیار اطلاع آکاییک (AIC)[۷] انجام داد؛ ولی در زمینه MPLE مورد مناقشه نیز بوده است؛ یعنی تعدادی از پژوهش‌گران باورمند هستند انتخاب مدلی‌که پارامترهای آن به روش MPLE برآورد شده‌اند، بر مبنای معیار اطلاع آکاییک درست نیست (گرونبیرگ و هجورت، ۲۰۱۴).

۴.۳ معیارهای همبستگی

معیارهای پارامتری همبستگی دایره‌ای–خطی مستلزم آن هستند که متغیر خطی دارای توزیع نرمال باشد (جانسون و وهرلی، ۱۹۷۷؛ ماردیا، ۱۹۷۶) و بنابراین، به‌طور کلی برای تمام داده‌های دایره‌ای–خطی قابل استفاده نیستند. از این‌رو، هودل و فبریگ (۲۰۲۱) از یک برآورد ناپارامتری برای ضریب همبستگی دایره‌ای–خطی، با نماد D، استفاده کردند که برای نخستین‌بار توسط ماردیا (۱۹۷۶) معرفی شد و در مطالعاتی چون سلو و همکاران (۱۹۸۸) و اخیراً توسط تاو (۲۰۱۵) به کار رفته است. این ضریب با رتبه‌بندی داده‌ها برآورد می‌شود و به‌صورت زیر است:

$$D = k\Big(\big(\sum_i s_i \cos \frac{2\pi r_i}{n}\big)^2 + \big(\sum_i s_i \sin \frac{2\pi r_i}{n}\big)^2 \Big), \tag{۱۱.۳}$$

[4] Maximum Likelihood Estimator
[5] Maximum Pseudo-likelihood Estimator
[6] Grid-Search
[7] Akaike Information Criterion

که در آن k یک ثابت نرمال‌سازی، r_i رتبه زاویه‌ای (دایره‌ای) مشاهده i (با نقطه مرجع $-\pi$) و s_i رتبه خطی همان مشاهده هستند. با این‌حال، ضریب D فقط می‌تواند مقادیری بین ۰ و ۱ را بگیرد و بنابراین نمی‌تواند همبستگی مثبت یا منفی را از هم تمایز دهد. همچنین، لیگوی و همکاران (۲۰۱۹) یک معیار همبستگی را بر مبنای اطلاع متقابل معیار واگرایی کولبک–لیبلر بین توزیع توأم و حاصل‌ضرب توزیع‌های حاشیه‌ای پیاده‌سازی کرده‌اند.

$$\begin{aligned} I(\Theta, X) &= \int_{-\pi}^{\pi}\int_{0}^{\infty} f_{\Theta,X}(\theta,x)\log\Big(\frac{f_{\Theta,X}(\theta,x)}{f_{\Theta}(\theta)f_X(x)}\Big)\mathrm{d}\theta\mathrm{d}x \\ &= \int_{-\pi}^{\pi}\int_{0}^{\infty} c\Big(F_{\Theta}(\theta),F_X(x)\Big)f_{\Theta}(x)f_X(x)\log\Big(c\big(F_{\Theta}(\theta),F_X(x)\big)\Big)\mathrm{d}\theta\mathrm{d}x \\ &= \int_{0}^{1}\int_{0}^{1} c(u,v)\log\big(c(u,v)\big)\mathrm{d}u\mathrm{d}v. \end{aligned} \tag{۱۲.۳}$$

برابری آخر با تغییر متغیر $u = F_{\Theta}(\theta)$ و $v = F_X(x)$ به‌دست آمده است. از این عبارت، بلافاصله مشخص می‌شود که اطلاع متقابل برابر منفی آنتروپی مرتبط با تابع چگالی مفصل $c(u,v)$ است (ما و سون، ۲۰۱۱). از آن‌جایی‌که محاسبه انتگرال در خط آخر رابطه (۱۲.۳) دشوار است، نمی‌توانیم پارامترهای خانواده مفصل انتخاب‌شده را به‌طور مستقیم تعیین کنیم. با این‌حال، نشان داده شده است که اطلاع متقابل نسبت به پارامتر وابستگی در دامنه وسیعی از مفصل‌های دومتغیره، رفتار یکنوا دارد (تنزیر و ایلیدن، ۲۰۱۶). بنابراین، انتظار می‌رود که چارچوب بهینه‌سازی تقریبی توضیح داده‌شده در بالا، در استفاده از اطلاع متقابل به عنوان معیار وابستگی، به خوبی عمل کند.

می‌توان برآوردی از اطلاع متقابل را با دسته‌بندی داده‌ها (مشاهدات متغیر تصادفی دایره‌ای و خطی) و گرفتن میانگین اطلاع متقابل نقطه‌ای–که براساس فراوانی نقاط داده در هر دسته محاسبه می‌شود– بر روی تمام ترکیب‌های دسته‌ها به‌دست آورد. هم‌چنین توجه داشته باشید که اطلاع متقابل داده‌ها برابر با اطلاع متقابل مفصل تجربی است. در صورت نیاز، می‌توان اطلاع متقابل را با استفاده از آنتروپی H به بازه‌ی $[0,1]$ طوری نرمال‌سازی کرد که مقدار صفر مفصل استقلال و مقدار یک مفصل بالایی فرشه–هافدینگ را معرفی کنند (تنزیر و ایلیدن، ۲۰۱۶):

$$I_{norm}(u,v) = \frac{I(u,v)}{\sqrt{H(u)H(v)}}. \tag{۱۳.۳}$$

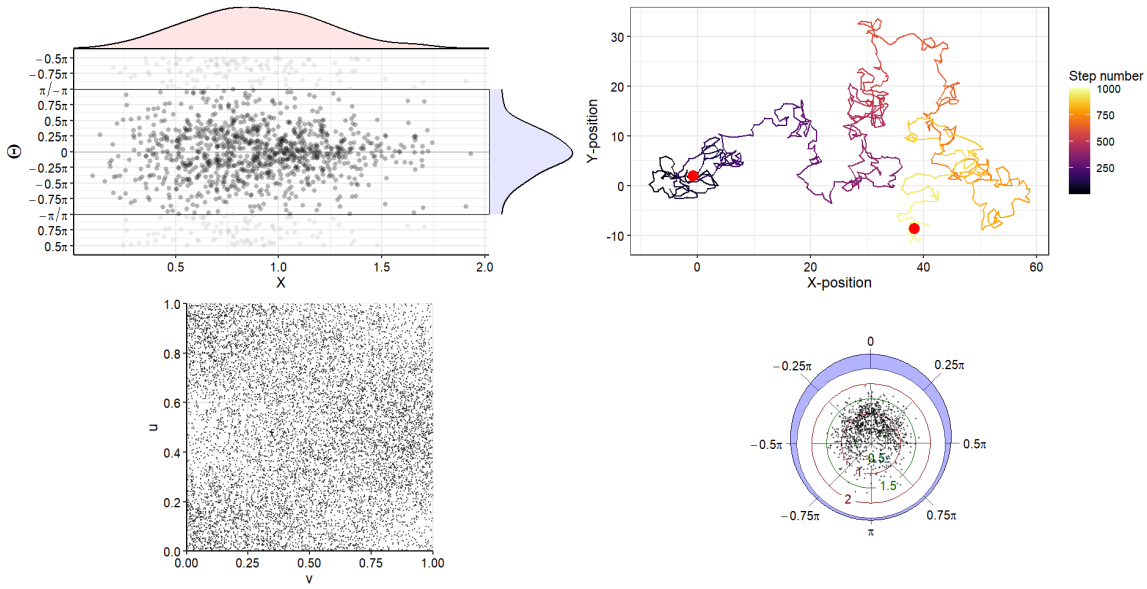
در نهایت، اگر بپذیریم که مفصل واقعی نسبت به متغیر u متقارن است، می‌توانیم مفصل تجربی را نیز متقارن‌سازی کنیم (به این صورت که مقادیر u بزرگ‌تر از ۰/۵ را با $1-u$ جایگزین کنیم).

۴ کاربردها

همان‌طور که قبلاً اشاره شد، مفصل دایره‌ای–خطی فون‌میزس متقارن نیست و استفاده‌ی آن برای داده‌های حرکتی مانند داده‌های باد یا حرکت حیوانات که معمولاً متقارن فرض می‌شوند، برآوردهای مناسب به‌دست نمی‌دهد. اما مفصل‌های مرتبه دو و مرتبه سه این مشکل را برطرف می‌کنند. در این مقاله از مفصل مرتبه دو با $\alpha = 0/08$، تعداد ۱۰۰۰ داده تولید شده که توزیع حاشیه‌ای دایره‌ای (زاویه‌های چرخش) آن را توزیع فون‌میزس با پارامترهای $(\mu,k) = (0,1)$ و توزیع حاشیه‌ای خطی (طول‌های گام) آن را توزیع وایبل با پارامترهای $(a,b) = (3,1)$ تشکیل می‌دهند. چون مفصل‌های معرفی‌شده بیش‌تر به موضوعات حرکتی می‌پردازند، داده‌های تولیدشده از آن‌ها را داده‌های مسیر[8] نیز می‌گویند. نمودارهای این داده‌ها در شکل ۳ آورده شده است. در این شکل، نمودار بالا (سمت چپ)، نمودار پراکنش

[8] Trajectory Data

زاویه‌های چرخش و طول‌های گام به‌ترتیب با برآوردهای چگالی ناپارامتری به رنگ‌های آبی و سرخ را نشان می‌دهد؛ نمودار بالا (سمت راست)، نمودار مسیر است که از دنباله زاویه‌های چرخش و طول‌های گام مشتق شده است؛ نمودار پایین (سمت چپ)، نمودار پراکنش مفصل به‌کاربرده شده را نشان می‌دهد و نمودار پایین (سمت راست)، نمودار پراکنش دایره‌ای زاویه‌های چرخش و طول‌های گام است.



شکل ۳: نمایش داده‌های دایره‌ای-خطی که با استفاده از مفصل مرتبه دو با حاشیه‌های فون‌میزس و وایبل شبیه‌سازی شده است.

ضریب همبستگی بین متغیرهای دایره‌ای و خطی در دو حالت بررسی شده است. مقدار ضریب همبستگی رابطه (۱۱.۳) $D =$ ۰٫۰۴۳۶ و مقدار ضریب همبستگی اطلاع متقابل $I_{norm}(u,v) =$ ۰٫۰۳۲۱ هستند. هردو معیار، همبستگی بین زاویه‌های چرخش و طول‌های گام را کم نشان می‌دهند.

در ادامه، از مفصل مرتبه سه با پارامترهای $(\alpha,\beta) = (۰٫۰۸, -\frac{۱}{۲\pi})$ و توزیع‌های حاشیه‌ای مفصل قبلی داده تولید کردیم. تعداد داده‌های تولیدشده را در اینجا نیز ۱۰۰۰ در نظر گرفتیم و از رسم نمودارهای آن خودداری کردیم. ضریب همبستگی رابطه (۱۱.۳) و اطلاع متقابل به‌ترتیب به‌صورت $D =$ ۰٫۰۹۵۵ و $I_{norm}(u,v) =$ ۰٫۰۴۲۲ به‌دست آمدند. بر مبنای این مفصل نیز ضریب همبستگی‌ها کوچک هستند. در بخش بعدی، پارامترهای مفصل‌ها به‌روش‌های ماکسیمم درستنمایی و ماکسیمم شبه‌درستنمایی برآورد شده و اریبی و جذر میانگین مربعات خطای آن‌ها (RMSE)[9] برای نمونه‌های مختلف در جدول ۱ آورده شده است. هردو روش برآوردهای نزدیک به پارامترهای اصلی را به‌دست می‌دهند. در جدول مشاهده می‌شود که با افزایش اندازه نمونه، اریبی و جذر میانگین مربعات خطا کاهش پیدا می‌کنند و هردو خطای کمی دارند. علاوه‌براین، هردو روش برآوردیابی برای هردو مفصل با پارامترهای مختلف، نتایج قابل قبولی ارائه می‌دهند و بنابراین، استفاده‌ی آن‌ها برای داده‌های واقعی مناسب به‌نظر می‌رسد.

[9] Root-Mean-Square Error

جدول ۱: بررسی عملکرد روش‌های MLE و MPLE برای برآورد پارامترهای مفصل‌ها

مفصل‌های ویژه	روش برآورد	اندازه نمونه (n)	$a = -۰/۱۵۹۲$		$b = ۰/۱۵۹۲$		$a = ۰/۰۸$		$b = -۰/۱۵۹۲$	
			اریبی	RMSE	اریبی	RMSE	اریبی	RMSE	اریبی	RMSE
مرتبه سه	MLE	۱۰۰	۰/۰۱۸۴	۰/۰۳۶۸	−۰/۰۲۰۵	۰/۰۳۷۵	۰/۰۰۱۹	۰/۰۵۶۶	۰/۰۲۲۶	۰/۰۴۲۶
		۵۰۰	۰/۰۰۶۳	۰/۰۱۲۷	−۰/۰۰۶۸	۰/۰۱۳۴	۰/۰۰۱۲	۰/۰۲۶۳	۰/۰۰۷۹	۰/۰۱۵۲
		۱۰۰۰	۰/۰۰۴۴	۰/۰۰۸۸	−۰/۰۰۴۴	۰/۰۰۸۷	۰/۰۰۰۴	۰/۰۱۷۷	۰/۰۰۵۹	۰/۰۱۱۱
	MPLE	۱۰۰	۰/۰۱۶۰	۰/۰۳۳۵	−۰/۰۱۸۳	۰/۰۳۵۳	۰/۰۰۶۲	۰/۰۵۴۸	۰/۰۲۱۶	۰/۰۴۰۷
		۵۰۰	۰/۰۰۶۱	۰/۰۱۲۹	−۰/۰۰۶۶	۰/۰۱۳۱	۰/۰۰۱۳	۰/۰۲۵۵	۰/۰۰۷۷	۰/۰۱۴۹
		۱۰۰۰	۰/۰۰۴۶	۰/۰۰۹۰	−۰/۰۰۴۶	۰/۰۰۸۹	۰/۰۰۱۸	۰/۰۱۸۵	۰/۰۰۵۹	۰/۰۱۰۷
مرتبه دو	MLE	۱۰۰	۰/۰۱۳۷	۰/۰۲۴۴			−۰/۰۰۳۰	۰/۰۳۷۱		
		۵۰۰	۰/۰۰۴۷	۰/۰۰۸۸			۰/۰۰۰۰	۰/۰۱۶۴		
		۱۰۰۰	۰/۰۰۳۱	۰/۰۰۵۸			−۰/۰۰۰۶	۰/۰۱۱۴		
	MPLE	۱۰۰	۰/۰۱۲۷	۰/۰۲۳۶			−۰/۰۰۱۴	۰/۰۳۹۰		
		۵۰۰	۰/۰۰۴۷	۰/۰۰۹۰			۰/۰۰۰۴	۰/۰۱۶۲		
		۱۰۰۰	۰/۰۰۳۲	۰/۰۰۶۰			−۰/۰۰۰۲	۰/۰۱۲۰		

بحث و نتیجه‌گیری

در این مقاله، ضمن معرفی داده‌های دایره‌ای و مفصل‌ها، مفصل‌های دایره‌ای–خطی و ویژگی‌های آن‌ها مورد بررسی قرار گرفت و به چند مفصل مهم در این حوزه پرداخته شد. نتایج شبیه‌سازی نشان دادند که مفصل‌های با بخش‌های ویژه از جمله مفصل مرتبه دو و مفصل مرتبه سه، می‌توانند برازش‌های مناسبی برای داده‌های حرکتی ارائه دهند و ساختار وابستگی در این نوع داده‌ها را به‌طور مناسب مدل می‌کنند. داده‌های حرکتی شامل داده‌های دایره‌ای و خطی می‌شوند و ویژگی‌های تناوب و تقارن را هم‌زمان دارا می‌باشند. نظر به این ویژگی‌ها، هر مفصل دایره‌ای–خطی نمی‌تواند وابستگی بین آن‌ها را به‌درستی مدل کند.

قدردانی و تشکر

نویسندگان مقاله از پیشنهادات داوران گرامی که موجب بهبود و ارائه بهتر آن شده‌اند، کمال سپاس و امتنان دارند.

مراجع

Abramowitz, M. and Stegun, I. A. (1965), *Handbook of Mathematical Functions*, New York, Dover.

Durante, F. and Sempi, C. (2015), *Principles of Copula Theory*, Second Edition, Chapman and Hall/CRC.

Genest, C. and Favre, A. C. (2007), Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask, *Journal of Hydrologic Engineering,* **24**(4), 347–368.

Gronneberg, S. and Hjort, N. L. (2014), The Copula Information Criteria, *Scandinavian Journal of Statistics,* **41**(2), 436–459.

Hodel, F.H. and Fiebreg, J. R. (2021), Cylcop: A Package for Circular-Linear Copulae with Angular Symmetry, *BioRxiv (2021)*: 2021-07.

Hodel, F. H. (2022), *Package 'cylcop'*, URL: https://cran.r-project.org/package=Cylcops.

Johnson, R. A and Wehrly, T. E. (1977), Measures and Models for Angular Correlation and Angular-Linear Correlation, *Journal of the Royal Statistical Society, Series B (Methodological),* **39**(2), 222–229.

Johnson, R. A and Wehrly, T. E. (1978), Some Angular-Linear Distributions and Related Regression Models, *Journal of the American Statistical Association,* **73**(363), 602–606.

Leguey, I., Larranaga, P., Bielza, C. and Kato, K. (2019), A Circular-Linear Dependence Measure under Johnson-Wehrly Distributions and its Application in Bayesian Networks, *Information Sciences,* **486**, 240–253.

Ma, J. and Sun, Z. (2011), Mutual Information is Copula Entropy, *Tsinghua Science and Technology,* **16**(1), 51–54.

Mardia, K. V. (1976), Linear-Circular Correlation Coefficient and Rhythmometry, *Biometrika,* 403–405.

Mardia, K. V. and Jupp, P. E. (2000), *Directional Statistics*, Second Edition, New York, John Wiley and Sons, Ltd, Chichester.

Nagar, P., Bekker, A., Arashi, M., Kat C. J. and Barnard, A. C. (2024), A Dependent Circular-Linear Model For Multivariate Biomechanical Data: ILIZARV RING FIXATOR STUDY, *Statistical Methods in Medical Research*, **33**(9), 1660-1672.

Nelsen, R. B. (2006), *An Introduction to Copulas*, Second Edition, New York, Springer.

Sklar, M. (1959), Functions de repartition a n dimensions et leurs marges, *Annales de I'ISUP,* **8**(3), 229–231.

Solow, A. R., Bullister, J. L. and Nevison, C. (1988), An Application of Circular-Linear Correlation Analysis to the Relationship between Freon Concentration and Wind Direction in Woods Hole, Massachusetts, *Environmental Monitoring and Assessment,* **10**(3), 219–228.

Tenzer, Y. and Elidan, G. (2016), On the Monotonicity of the Copula Entropy, *arXiv preprint arXiv*: 1611.06714.

Tu, R. (2015), A Study of the Parametric and Nonparametric Linear-Circular Correlation Coefficient, *California Polytechnic State University, San Luis Obispo,* 1–24, URL: http://digitalcommons.calpoly.edu/statsp/51/.

Wang, Z. W., Zhang, W. M., Zhang, Y. F. and Liu, Z. (2021), Circular-Linear-Linear Probabilistic Model Based on Vine Copulas: An Application to the Joint Distribution of Wind Direction, Wind Speed, and Air Temperature, *Journal of Wind Engineering and Industrial Aerodynamics,* **215**, 104704.

Circular-Linear Copulas: Properties and Applications

Najibullah Karimi[1], **Ali Dastbaravarde**[2]

[1]PhD Student of Statistics, Yazd University

[2]Department of Statistics, Faculty of Mathematics, Yazd University, 89195-741, Yazd, Iran

---

Abstract: This review paper presents recent studies in the field of circular-linear cupolas. To analyze the dependence structure of circular and linear data, various forms of circular-linear copulas are examined. The use and implementation of the cylcop package in R software have been discussed as well. Furthermore, through simulation, copula parameters are estimated. The results indicate that the copulas with specific sections are suitable for modeling movement data.

Keywords: Movement Data, Circular-Linear Data, Copula with Special Sections, Circular-Linear Copula.

Mathematics Subject Classification (2020): 62H11, 62G30, 62F99.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مدل‌سازی فرایندهای سری‌زمانی شبکه عصبی بیزی برای شناسایی پیش‌زلزله با ناهنجاری‌های یونوسفری زلزله خاش

محدثه کیکاوسی [۱]، فاطمه حسینی، امید کریمی

گروه آمار، دانشگاه سمنان

چکیده: فرایندهای سری‌زمانی حافظه بلند و کوتاه مدت برای مدل‌کردن داده‌های وابسته به گذشته به کار می‌روند که قادرند اطلاعات مربوط به زمان‌های گذشته را برای مدت طولانی حفظ کنند. با بهره‌گیری از شبکه‌های عصبی و روش‌های بیزی می‌توان روند زمانی ناهنجاری‌های یونوسفری را شناسایی و تحلیل کرد. تحلیل ناهنجاری‌های یونوسفری به شناسایی وقوع زمین‌لرزه‌ها با استفاده از پس‌لرزه‌ها کمک می‌کند. در این مقاله، مدل‌های حافظه بلندمدت با روش‌های کلاسیک و بیزی برای تحلیل و پیش‌بینی روندهای زمانی ناهنجاری‌های مربوط به زلزله خاش سال ۱۳۹۲ به کار گرفته می‌شوند. نتایج حاصل نشان می‌دهند که مدل پیشنهادی توانایی مناسبی در شناسایی الگوهای پنهان یونوسفری پیش از زمین‌لرزه دارند و می‌توان از آن‌ها به‌عنوان ابزاری کارآمد در تحلیل پیش‌نشانگرهای لرزه‌ای بهره گرفت.

واژه‌های کلیدی: فرآیندهای زمانی، شبکه‌های عصبی بیزی، حافظه طولانی کوتاه مدت، زمین‌لزره خاش.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62J12، 62H11، 65K10.

۱ مقدمه

پیش‌بینی زلزله همواره یکی از چالش‌های مهم در علوم زمین بوده است. در سال‌های اخیر، استفاده از تحلیل تغییرات محتوای الکترونی کل[۱] (TEC) به عنوان یکی از روش‌های نوین برای شناسایی پیش‌نشانگرهای زلزله مورد توجه قرار گرفته است. پژوهش‌های متعددی نشان داده‌اند که بلایای طبیعی می‌توانند از طریق سازوکارها و شدت‌های مختلف، لایه یونوسفر[۲] را دچار اختلال کنند (چن و همکاران ، ۲۰۲۰؛ فری‌شَه و همکاران ، ۲۰۲۴). با توجه به پیچیدگی رفتار زمین، استفاده از شبکه‌های عصبی با مدل‌سازی عدم قطعیت می‌تواند بینش عمیق‌تری در تحلیل این ناهنجاری‌ها ارائه دهد. در این مقاله، با استفاده از داده‌های TEC استخراج‌شده از فایل‌های یونکس[۳] و

[۱] سخنران، mohadeseh-keykavosi@semnan.ac.ir

[1] Total Electron Content
[2] Ionosphere
[3] IONEX

به‌کارگیری مدل‌های حافظه کوتاه و بلند مدت[4] (LSTM) و LSTM بیزی[5] (سقیب و همکاران ، ۲۰۲۴) به تحلیل ناهنجاری‌های یونوسفری پیش از وقوع زلزله شهر خاش در استان سیستان و بلوچستان پرداخته‌ایم. این زلزله ۲۷ فروردین ۱۳۹۲ با بزرگای حدود ۷/۵ رخ داده است. یکی از اهداف مهم، بررسی پیش‌بینی روند زمانی داده‌های TEC در بازه زمانی پیش از زلزله و ارزیابی هم‌زمانی ناهنجاری‌های شناسایی‌شده با زمان وقوع این زلزله می‌باشد.

۲ مدل‌های LSTM کلاسیک و بیزی

برای مدل‌کردن روند زمانی داده‌های TEC، از دو مدل شبکه عصبی بازگشتی LSTM کلاسیک و بیزی برای شناسایی ناهنجاری‌های احتمالی پیش از زلزله استفاده می‌شود. طبق آگاروال (۲۰۱۸) مدل LSTM کلاسیک یکی از انواع شبکه‌های عصبی بازگشتی است که برای مدل‌سازی داده‌های سری زمانی و وابسته به گذشته طراحی شده است. این مدل با استفاده از ساختار حافظه‌ای ویژه، قادر است اطلاعات مربوط به زمان‌های گذشته را برای مدت طولانی‌تری حفظ کند و از آن‌ها در یادگیری وابستگی‌های زمانی پیچیده بهره گیرد. در مدل‌های LSTM ، چهار مؤلفه‌ی وضعیت سلول و دروازه ورودی[6] و دروازه فراموشی[7] و دروازه خروجی[8] وجود دارد که نقش کلیدی در پردازش و حفظ اطلاعات ایفا می‌کنند. هر یک از این اجزا وظیفه‌ای خاص را بر عهده دارند و با همکاری یکدیگر، توانایی مدل در یادگیری و حفظ اطلاعات بلندمدت را بهبود می‌بخشند. برای درک بهتر عملکرد این مدل، فرض کنید در زمان t، ورودی X_t، بردار وزن‌ها $W_t = [W_f, W_i, W_c, W_o]$ و بردار اریبی‌ها $b_t = [b_f, b_i, b_c, b_o]$ وجود دارد. خروجی مدل در این زمان نیز O_t خواهد بود. دروازهٔ فراموشی به صورت $f_t = \sigma(W_f \cdot [O_{t-1}, X_t] + b_f)$ است که در آن، f_t خروجی دروازهٔ فراموشی در زمان t است. تابع σ نشان‌دهندهٔ تابع فعال‌سازی سیگموید[9] است که خروجی را در محدودهٔ بین ۰ تا ۱ قرار می‌دهد. O_{t-1} خروجی مدل در زمان قبلی $(t-1)$ است. همچنین، W_f وزن‌های مرتبط با دروازهٔ فراموشی و b_f اریبی این دروازه است. دروازهٔ فراموشی با استفاده از این مکانیزم، تصمیم می‌گیرد که چه میزان از اطلاعات قبلی در وضعیت سلول حفظ شود. اگر f_t بزرگتر از ۰/۵ باشد، اطلاعات قبلی در حالت سلولی حفظ می‌شود؛ در غیر این صورت، اطلاعات فراموش می‌شود. این فرایند به مدل LSTM اجازه می‌دهد تا به صورت انتخابی و هوشمندانه، اطلاعات مهم را نگه داشته و اطلاعات غیرضروری را حذف کند، که این امر در پردازش توالی‌های زمانی و یادگیری الگوها بسیار حیاتی است.

دروازه ورودی در مدل LSTM به‌صورت $i_t = \sigma(W_i \cdot [O_{t-1}, X_t] + b_i)$ محاسبه می‌شود. که در آن، i_t خروجی دروازه ورودی در زمان t است و نشان می‌دهد چه اطلاعاتی از ورودی جدید باید وارد وضعیت سلول شود. وضعیت سلول پیشنهادی در مدل LSTM به صورت $\tilde{C}_t = \tanh(W_c \cdot [O_{t-1}, X_t] + b_c)$ تعریف می شود که در آن، $\tilde{C}_t$ مقدار پیشنهادی برای به‌روزرسانی وضعیت سلول در زمان t است. تابع tanh به عنوان تابع فعال‌سازی استفاده می‌شود و خروجی آن در محدوده‌ی بین ۱- و ۱+ قرار می‌گیرد. W_c وزن‌های مرتبط با وضعیت سلول پیشنهادی و b_c اریبی آن است. سپس وضعیت سلول به صورت $C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$ به‌روزرسانی می‌شود. که در آن، C_t وضعیت سلول در زمان t است که با ترکیب اطلاعات قبلی و اطلاعات جدید به‌دست می‌آید. دروازه خروجی مدل LSTM به صورت $h_t = \sigma(W_o \cdot [O_{t-1}, X_t] + b_o)$ به دست می‌آید که در آن، h_t سیگنال خروجی دروازه خروجی

[4] Long Short-Term Memory
[5] Bayesian LSTM
[6] Input Gate
[7] Forget Gate
[8] Output Gate
[9] Sigmoid

در زمان t است که مقداری بین ۰ و ۱ اختیار می‌کند. پس از محاسبه‌ی h_t، خروجی نهایی مدل در زمان t (O_t) با استفاده از وضعیت سلول به‌روزرسانی‌شده (C_t) و سیگنال خروجی دروازه خروجی به صورت $O_t = h_t \cdot \tanh(C_t)$ به دست می‌آید. در این مقاله، مدل LSTM پایه‌ای به‌کار گرفته شده است که شامل یک لایه LSTM با تعداد مشخصی واحد حافظه و یک لایه خروجی چگالی است. این مدل به‌صورت کلاسیک و قطعی آموزش داده می‌شود؛ به این معنا که در مرحله‌ی آموزش، از تکنیک دراپ آوت[۱۰] برای جلوگیری از بیش‌برازش استفاده می‌شود، اما در مرحله آزمون، دراپ آوت غیرفعال می‌گردد. در نتیجه، مدل تنها یک مقدار مشخص و قطعی برای TEC در هر زمان پیش‌بینی می‌کند. این مدل توانایی مناسبی در یادگیری الگوهای زمانی TEC دارد، اما به‌دلیل ماهیت قطعی خود، امکان سنجش عدم قطعیت پیش‌بینی‌ها را فراهم نمی‌کند. به همین دلیل، در ادامه از رهیافت بیزی برای بهبود مدل LSTM بهره گرفته می‌شود تا علاوه بر پیش‌بینی، ارزیابی میزان اطمینان مدل نیز امکان‌پذیر باشد.

در ادامه مؤلفه‌های، W_t و b_t را با استفاده از رهیافت بیزی بهینه می‌کنیم. در مدل‌های LSTM ، وزن‌ها (W_t) و اریبی‌ها (b_t) با استفاده از توزیع‌های پیشین مدل‌سازی می‌شوند. این رهیافت عدم قطعیت‌ها را در پارامترها در نظر می‌گیرد و برازش‌ها را بهینه می‌کند. وزن‌ها و اریبی‌ها دارای توزیع نرمال به صورت $W_t \sim \mathcal{N}(\mu_0, \sigma^2)$ و $b_t \sim \mathcal{N}(\mu_0, \sigma^2)$ است که در آن، μ_0 میانگین اولیه‌ی پیشین است که انتظار اولیه برای مقدار وزن‌ها و اریبی‌ها را نشان می‌دهد و σ^2 واریانس پیشین است که میزان عدم قطعیت در پارامترها را مشخص می‌کند. در مدل LSTM ، تنظیم اریبی برای بهینه‌سازی عملکرد مدل بسیار مهم است. همچنین، برای بهبود توانایی تعمیم‌پذیری مدل و جلوگیری از بیش‌برازش، از تکنیک دراپ آوت استفاده می‌شود. برای افزایش اعتمادپذیری مدل و لحاظ‌کردن عدم قطعیت در پیش‌بینی‌ها، از ساختار بیزی در قالب شبکه عصبی LSTM استفاده می‌شود. در این ساختار، با بهره‌گیری از روش‌های مونت کارلویی[۱۱] ، لایه‌های دراپ آوت نه‌تنها در زمان آموزش بلکه در مرحله‌ی آزمون نیز فعال باقی می‌مانند. نمونه‌های تصادفی مونت کارلویی امکان برآورد توزیع پسین را فراهم می‌سازد. با نمونه‌گیری چندباره از مدل، می‌توان میانگین پیش‌بینی‌ها و انحراف معیار آن‌ها را محاسبه کرد. تابع زیان در مدل LSTM به‌صورت $\mathcal{O}_t^{\text{dropout}} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i, O_t) + \lambda \sum_{j=1}^{L}\left(\|W_t\|^2 + \|b_t\|^2\right)$ تعریف می‌شود، که شامل دو بخش است. بخش اول، میانگین خطای پیش‌بینی را محاسبه می‌کند، که با استفاده از تابع $\mathcal{L}$ اختلاف بین مقادیر واقعی (y_i) و پیش‌بینی‌های مدل (O_t) را اندازه‌گیری می‌کند. بخش دوم، یک عبارت تنظیم‌گر[۱۲] است که با مجموع مربعات وزن‌ها (W) و اریبی‌ها (b) و ضرب آن در پارامتر λ از بیش‌برازش جلوگیری می‌کند.

طبق هستی و همکاران (۲۰۰۹)، برآوردگر ماکسیمم درست‌نمایی برای وزن‌ها (W_t) با بهینه کردن تابع زیان به‌دست می‌آید. این رویکرد تعادلی مناسب بین دقت پیش‌بینی و جلوگیری از بیش‌برازش برقرار می‌کند و به بهبود عملکرد و تعمیم‌پذیری مدل کمک می‌نماید.

در رهیافت بیزی امکان به دست آوردن توزیع احتمالی مشاهدات پیش‌بینی‌شده را فراهم می‌کند. تابع درست‌نمایی به‌صورت زیر تعریف می‌شود:

$$p(y \mid X, W, \sigma^2) \propto \frac{1}{\sqrt[n/2]{\sigma^2}} e^{-\frac{1}{2\sigma^2}\left((y-XW)^T(y-XW)\right)}.$$

این فرمول احتمال اینکه بردار خروجی y از داده‌های ورودی X و پارامترهای مدل W تولید شده باشد را نشان می‌دهد. که در آن، $(y - XW)^T(y - XW)$ مجموع مربعات خطاها (SSE)[۱۳] است که اختلاف بین مقادیر واقعی y و مقادیر پیش‌بینی‌شده XW را اندازه‌گیری می‌کند. در مدل بیزی، از توزیع پیشین مزدوج برای پارامترها استفاده می‌شود. توزیع پیشین به‌صورت $p(W, \sigma^2) =$

[10] Dropout
[11] Monte Carlo Dropout
[12] Regularization
[13] sum squares errors

$p(\sigma^2)\,p(W \mid \sigma^2)$ می‌باشد. درآن صورت توزیع پسین به‌صورت $p(W, \sigma^2 \mid y, X) \propto p(y \mid X, W, \sigma^2)\,p(W \mid \sigma^2)\,p(\sigma^2)$ تعریف می‌شود. توزیع پسین پارامترهای مدل به‌صورت

$$\begin{aligned} p(W, \sigma^2 \mid y, X) \propto & \frac{1}{\sqrt[n/2]{\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-XW)^T(y-XW)} \\ & \times (\sigma^2)^{-(n-v)/2} e^{-\frac{1}{2\sigma^2}(W-\mu_0)^T\Lambda_0(W-\mu_0)} \\ & \times (\sigma^2)^{-(\alpha_0-1)} e^{-\frac{b_0}{\sigma^2}} \end{aligned}$$

به دست می‌آید. که در آن $(\sigma^2)^{-(n-v)/2} e^{-\frac{1}{2\sigma^2}(W-\mu_0)^T\Lambda_0(W-\mu_0)}$ توزیع پیشین وزن‌ها را با میانگین μ_0 و ماتریس دقت Λ_0 مشخص می‌کند و $(\sigma^2)^{-(\alpha_0-1)} e^{-\frac{b_0}{\sigma^2}}$ توزیع پیشین واریانس را با پارامترهای α_0 و β_0 مدل‌سازی می‌کند. ماتریس ریج $\Lambda_0 = cI$ نقشی کلیدی در تنظیم مدل و حل مشکل چندخطی‌بودن دارد. با اضافه کردن این ماتریس به X^TX، ماتریس معکوس‌پذیر شده و از بیش‌برازش جلوگیری می‌کند. پارامتر c شدت تنظیم را کنترل می‌کند؛ هرچه c بزرگتر باشد، تأثیر پیشین قوی‌تر است. میانگین پسین μ_n با فرمول $\mu_n = (X^TX + \Lambda_0)^{-1}(X^Ty\hat{w} + \Lambda_0\mu_0)$ محاسبه می‌شود که ترکیبی از اطلاعات داده‌ها و باورهای پیشین است. به‌روزرسانی پارامترهای بیزی شامل محاسبه $\Lambda_n = X^TX + \Lambda_0$ و $a_n = a_0 + \frac{n}{2}$ است. همچنین، W_t با فرمول $W_t = W_0 + \frac{1}{2}(y^Ty + \mu_0^T\Lambda_0\mu_0 - \mu_n^T\Lambda_n\mu_n)$ به‌روزرسانی می‌شود. بنابراین توزیع پیشگویی نهایی به‌صورت

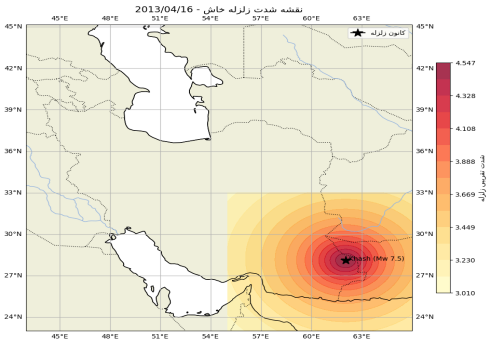
$$p(y \mid m) = \frac{p(m) \times p(y \mid X, W, \sigma, m)}{p(W, \sigma \mid y, X, m)}$$

محاسبه می‌شود. این فرمول بر اساس قانون بیز، توزیع پیشگویی $p(y \mid m)$ را به عنوان نسبت درست‌نمایی $p(y \mid X, W, \sigma, m)$ و توزیع پیشین $p(m)$ به توزیع پسین پارامترها $p(W, \sigma \mid y, X, m)$ محاسبه می‌کند. پیش‌بینی نهایی دارای توزیع نرمال است که نه تنها یک مقدار نقطه‌ای، بلکه یک بازه اطمینان برای مقدار خروجی ارائه می‌دهد. این بازه اطمینان، عدم قطعیت‌های موجود در برآورد پارامترها و نویز مدل را در نظر می‌گیرد و به کاربر امکان می‌دهد تا درک بهتری از محدوده احتمالی مقادیر پیش‌بینی‌شده داشته باشد. در نتیجه، LSTM بیزی نه‌تنها پیش‌بینی عددی ارائه می‌دهد، بلکه بازه‌ی اطمینان برای آن پیش‌بینی را نیز مشخص می‌کند. این ویژگی برای تحلیل پیش‌نشانگرهای زلزله که با عدم قطعیت‌های طبیعی همراه‌اند، اهمیت ویژه‌ای دارد.

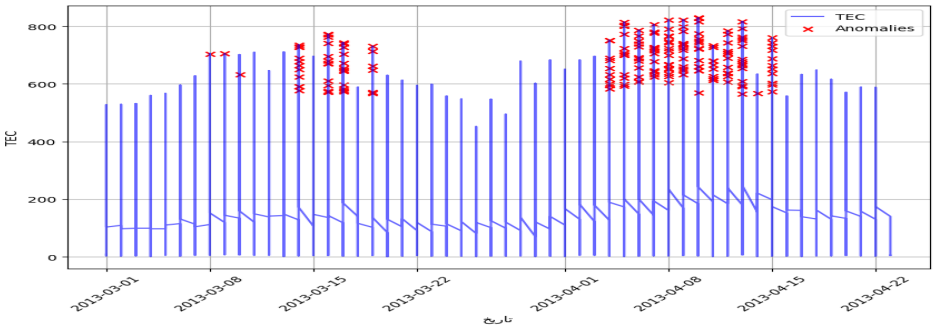
۳ تحلیل داده‌های زلزله خاش

برای مطالعه موردی، زلزله خاش در تاریخ ۲۷ فروردین ۱۳۹۲ انتخاب شد. پس از استخراج داده‌ها، به‌منظور تحلیل دقیق‌تر و شناسایی ناهنجاری‌ها، شاخص نمره‌گذاری استاندارد[14] به صورت $Z_i = \frac{x_i - \mu}{\sigma}$ برای هر روز محاسبه شد، که در آن x_i مقدار TEC در زمان مشخص، μمیانگین و σ انحراف معیار مقادیر TEC در یک بازه‌ی زمانی مرجع است. شکل ۱ موقعیت و شدت زلزله خاش را روی نقشه کشور ایران نشان می‌دهد، این زلزله با بزرگای ۷/۵ در عمق حدود ۹۲ کیلومتری زمین به وقوع پیوست. دلیل انتخاب این زلزله، قدرت بالا، گستره تأثیر قابل توجه و ثبت دقیق آن در منابع بین‌المللی بوده است که امکان بررسی ناهنجاری‌های یونوسفری پیش از رخداد را با دقت و وضوح بیشتری فراهم می‌سازد. شکل ۲ روند تغییرات TEC را در بازه زمانی بین ابتدای مارس تا اواخر آوریل ۲۰۱۳ برای ایستگاه خاش نمایش می‌دهد. در این بازه زمانی، دو دوره‌ی پرتراکم از ناهنجاری‌ها به وضوح قابل مشاهده هستند: دوره اول: از حدود ۸ مارس تا ۱۶ مارس، که در آن تعداد قابل توجهی از مقادیر TEC از روند معمول خود فراتر رفته‌اند. دوره دوم: از حدود ۵

[14] Z-score



شکل ۱: نقشه شدت زلزله خاش



شکل ۲: نمودار TEC با نقاط ناهنجاری: مقادیر TEC با استفاده از خطوط نمایش داده شده‌اند و نقاط ناهنجار استاندارد شده با علامت ستاره مشخص شده‌اند.

آوریل تا ۱۵ آوریل، که در این بازه نیز مقادیر TEC دارای پراکندگی زیاد و رفتار غیرعادی بوده‌اند. این رفتار ناهنجار ممکن است با پیش‌نشانگرهای ژئوفیزیکی یا ژئومغناطیسی زلزله‌ها مرتبط باشد و می‌تواند به عنوان هشدار اولیه مورد استفاده قرار می‌گیرد.

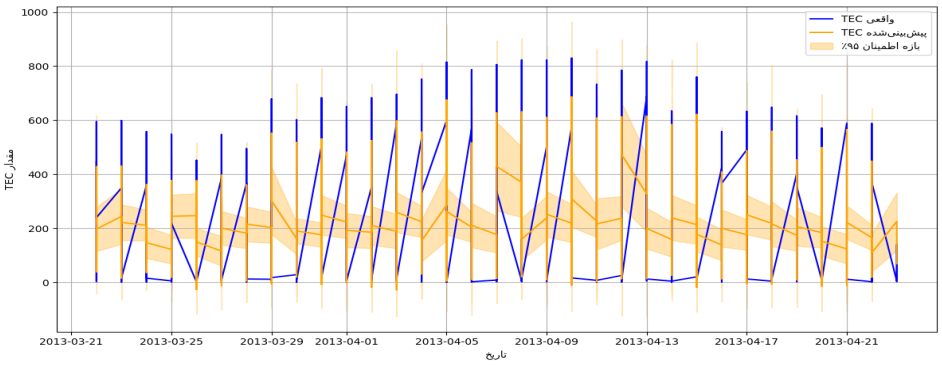
در ادامه، مدل‌های LSTM کلاسیک و بیزی روی داده‌ها اجرا و مورد ارزیابی قرار گرفت. هدف اصلی، بررسی دقت پیش‌بینی این دو مدل، ارزیابی ناهنجاری‌های احتمالی در داده‌ها پیش از رخداد زلزله و تحلیل عدم قطعیت در پیش‌بینی‌ها است. برای این منظور داده‌ها به دو بخش آموزش[۱۵] (۸۰٪) و آزمون[۱۶] (۲۰٪) تفکیک شدند. هر دو مدل روی بازه زمانی پیش از زلزله آموزش داده شدند و عملکرد آن‌ها در بازه نزدیک به وقوع زلزله (تقریباً ۴۶ روز قبل تا ۷ روز بعد از زلزله) مورد ارزیابی قرار گرفتند. معیارهای میانگین قدرمطلق خطا (MAE)[۱۷] و ریشه میانگین مربعات خطا (RMSE)[۱۸] و نرمال میانگین قدرمطلق خطا (NMAE)[۱۹] و نرمال ریشه میانگین مربعات خطا (NRMSE)[۲۰] برای ارزیابی دقت این مدل‌های به‌صورت $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$، $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$، $\mathrm{NMAE} = \frac{\mathrm{MAE}}{\max(y_i)-\min(y_i)}$، $\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\max(y_i)-\min(y_i)}$ محاسبه شده‌اند، که در آن y_i داده‌های واقعی TEC و $\hat{y}_i$ مقدار پیشگویی آن است. نتایج ارزیابی مدل‌های LSTM کلاسیک و بیزی روی داده‌های خاش در جدول ۱ خلاصه شده است. نتایج جدول ۱ نشان می‌دهد که مدل LSTM قادر است تغییرات TEC را با خطای نسبتاً کنترل‌شده‌ای

[15] Train
[16] Test
[17] Mean Absolute Error
[18] Root Mean Squared Error
[19] Normal Mean Absolute Error
[20] Normal Root Mean Squared Error

جدول ۱: خلاصه نتایج دو مدل LSTM کلاسیک و بیزی

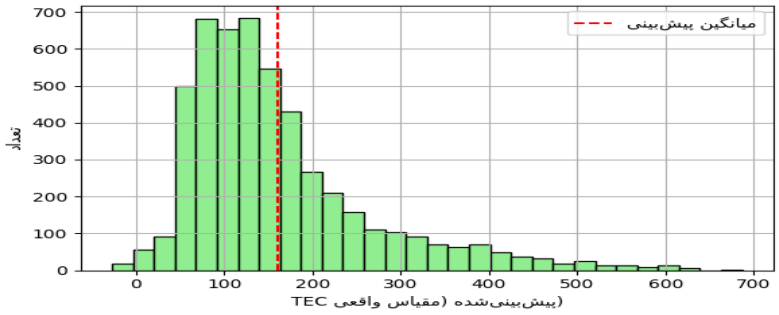
معیار	LSTM کلاسیک	LSTM بیزی
MAE	۱۳۵/۷۰	۱۳۵/۱۶
RMSE	۱۸۷/۵۰	۱۷۵/۱۷
NMAE	۰/۱۶۳۷	۰/۱۶۳۰
NRMSE	۰/۲۲۶۲	۰/۲۱۱۳

پیش‌بینی نماید، به‌ویژه آن‌که مقدار NMAE کمتر از ۰/۲ بوده که نشان‌دهنده پیش‌بینی قابل قبول در مقیاس تغییرات طبیعی داده‌هاست. همان‌گونه که مشاهده می‌شود، هرچند دقت کلی مدل LSTM بیزی نسبت به LSTM تفاوت زیادی ندارد، اما مقدار RMSE کاهش یافته است که بیانگر کاهش اثر نمونه‌های دارای خطای بالا در مدل بیزی است. همه تحلیل داده‌ها با پایتون انجام شده است. در شکل ۳، علاوه بر مقایسه مقدار واقعی با مقدار پیش‌بینی‌شده، بازه اطمینان ۹۵٪ نیز نمایش داده شده است.



شکل ۳: نمودار پیش‌بینی سری زمانی TEC با مدل $LSTM$ بیزی

مدل LSTM بیزی نه تنها پیش‌بینی را انجام می‌دهد، بلکه عدم قطعیت آن را نیز نشان می‌دهد. عرض بازه‌های اطمینان در حوالی روزهای پیش از زلزله افزایش می‌یابد، که می‌تواند نشانه‌ای از افزایش عدم اطمینان مدل و احتمال بروز ناهنجاری باشد. این ویژگی روش‌های بیزی برای تحلیل پیش‌نشانگرهای زلزله بسیار کاربردی است، زیرا به جای پیش‌بینی قطعی، بازه‌ای از پیش‌بینی‌ها ارائه می‌دهد. شکل ۴ توزیع میانگین پیش‌بینی‌ها توسط مدل بیزی را نمایش می‌دهد.توزیع به‌صورت نامتقارن و حول مقدار میانگین



شکل ۴: نمودار هیستوگرام پیشبینی مدل $LSTM$ بیزی

متمرکز است، اما نسبت به هیستوگرام، LSTM پراکندگی کمتری دارد که باعث دقت پیشگویی در مدل بیزی است. این نامتقارنی می‌تواند بازتاب‌دهنده ناپایداری یا تنش در لایه یونوسفر باشد. مدل LSTM بیزی دقت نسبتا بالاتری در پیش‌بینی دارد و مهم‌تر از آن، با ارائه بازه اطمینان، قدرت تحلیل ناهنجاری‌های ساختاری در داده را فراهم می‌آورد. به‌ویژه در نواحی زمانی پیش از زلزله، افزایش عرض بازه اطمینان می‌تواند به عنوان شاخصی برای شناسایی ناهنجاری در لایه یونوسفر در نظر گرفته شود.

بحث و نتیجه‌گیری

در این مقاله با استفاده از مدل سری‌زمانی LSTM کلاسیک و بیزی با شاخص نمره‌گذاری استاندارد، ناهنجاری‌های احتمالی در داده‌های TEC استخراج‌شده از فایل‌های یونکس شناسایی گردید. این ناهنجاری‌ها عمدتاً به‌صورت خوشه‌ای و در بازه‌های زمانی مشخصی پیش از وقوع برخی زلزله‌ها ظاهر شدند که می‌توانند به‌عنوان نشانه‌هایی از پیش‌نشانگرهای یونوسفری تفسیر شوند. در ادامه، به‌منظور بررسی و مدل‌کردن رفتار زمانی این ناهنجاری‌ها، از دو مدل یادگیری عمیق شامل LSTM کلاسیک و LSTM بیزی با مرزهای عدم‌قطعیت استفاده شد. نتایج این مدل‌ها نشان داد که مدل بیزی علاوه بر ارائه دقت قابل‌قبول، قابلیت نمایش عدم‌قطعیت پیش‌بینی‌ها را نیز داراست که در تحلیل داده‌های حساس و پرنوسان مانند TEC بسیار ارزشمند است. با توجه به این یافته‌ها، پیشنهاد می‌شود در مطالعات آتی مدل‌های پیشرفته‌تر بیزی، گراف‌پایه به‌کار گرفته شود تا همبستگی دقیق‌تری بین ناهنجاری‌های TEC مشخص گردد و گامی مؤثر در راستای توسعه سامانه‌های هشدار زلزله برداشته شود.

مراجع

Chen J., Zhang X., Ren X., Zhang J., Freeshah M., Zhao Z. (2020), Ionospheric disturbances detected during a typhoon based on GNSS phase observations: A case study for typhoon Mangkhut over Hong Kong, *Adv. Space Res.*, **66**, 1743–1753.

Freeshah M., Adil M.A., Sentürk E., Zhang X., Ren X., Liu H., Osama N. (2024a), A cyclone formation, eastward plume drag, ionhydration process, and the consequent ionospheric changes following the 2022 Hunga Tonga-Hunga Ha'apai volcanic eruption, *Advances in Space Research*, **73** (5), 2457–2470.

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Saqib, M., Sentürk, E., Adil, M.A., Freeshah, M. (2024). Seismo-ionospheric precursory detection using hybrid Bayesian-LSTM network model with uncertainty-boundaries and anomaly-intensity. *Advances in Space Research*, 74, 1828–1842.

Aggarwal, C. C. (2018). Neural networks and deep learning: A textbook. Springer International Publishing. https://doi.org/10.1007/978-3-319-94463-0

Bayesian Neural Network Temporal Processes Modeling for Earthquake Precursors Identification Using Ionospheric Anomalies of Khash Earthquake”

Mohadeseh Keykavosi, Fatemeh Hoseini, Omid Karimi

Department of statistics, Semnan University

Abstract: Long-short term memory time series processes are employed to model past-dependent data, capable of retaining information from past times for extended periods. By leveraging neural networks and Bayesian methods, the temporal trends of ionospheric anomalies can be identified and analyzed. The analysis of ionospheric anomalies aids in detecting earthquake occurrences using aftershocks. In this paper, long-term memory models, both classical and Bayesian, are utilized to analyze and predict the temporal trends of anomalies related to the 2013 Khash earthquake. The results demonstrate that the proposed model has a suitable capability in identifying hidden ionospheric patterns before earthquakes and can be used as an effective tool in analyzing seismic precursors.

Keywords: Temporal Processes, Bayesian Neural Networks, Long- Short term Memory, Khash Earthquake.

Mathematics Subject Classification (2020): 62J12، 62H11، 65K10.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

پیش‌بینی زمان حوادث ترافیکی با استفاده از شبکه عصبی بقا

فاطمه محمدی [۱] ، محمد آرشی[۱] آرزو حبیبی‌راد [۱] ابوالفضل محمدزاده مقدم [۲]

[۱]گروه آمار دانشگاه فردوسی مشهد

[۲]گروه عمران دانشگاه فردوسی مشهد

چکیده: حوادث ترافیکی یکی از چالش‌های اصلی در مدیریت ترافیک شهری هستند که می‌توانند تاثیرات منفی بر ایمنی و کارایی سیستم‌های حمل و نقل داشته باشند. آژانس‌های حمل و نقل برای مدیریت حوادث ترافیکی و به کارگیری استراتژی‌های مناسب نیاز به پیش‌بینی مدت زمان حادثه دارند. داده‌های زمان حوادث ترافیکی داده‌های زمان تا رویداد هستند. یکی از ویژگی‌های این داده‌ها وابستگی به مدت زمان است. یکی از مدل‌های رایج برای در نظر گرفتن این ویژگی، مدل‌های نرخ خطر مبنا است. عملکرد مدل‌های نرخ خطر مبنا در پیش‌بینی زمان حوادث ترافیکی، با توجه به فرضیاتی که دارند ممکن است دارای محدودیت‌هایی باشد. از جمله این فرضیات می‌توان به فرض خطرات متناسب، در نظر گرفتن توزیع خاص برای داده‌ها، اثر ثابت با زمان عوامل تاثیرگذار و رابطه خطی بین متغیرهای مستقل و وابسته اشاره کرد و ممکن است برخی از این فرضیات در داده‌های حوادث ترافیکی وجود نداشته باشد. یکی دیگر از روش‌های رایج برای مدل‌بندی زمان حوادث ترافیکی شبکه‌های عصبی مصنوعی می‌باشد. نقطه قوت این مدل‌ها این است که برخی از فرضیات مشکل‌ساز مدل‌های نرخ خطر مبنا را ندارند. اگر هدف این مدل‌ها پیش‌بینی مدت زمان پاکسازی حادثه باشد، نمی‌توانند وابستگی به مدت زمان را در داده‌های حوادث ترافیکی در نظر بگیرند. لذا از مدل‌های شبکه‌های عصبی با هدف برآورد تابع بقا استفاده می‌شود. اما این مدل‌ها هم نمی‌توانند مستقیم مدت زمان یک حادثه ترافیکی را پیش‌بینی کنند. بنابراین در این مطالعه مدل شبکه عصبی مبتنی بر تحلیل بقا معرفی می‌شود که بر اساس ساختار یادگیری چند وظیفه‌ای، دو وظیفه برآورد تابع بقا و پیش‌بینی زمان حوادث ترافیکی را همزمان انجام می‌دهد. مدل پیشنهادی می‌تواند وابستگی به مدت زمان را هنگام پیش‌بینی زمان حادثه با برآورد همزمان تابع بقا در نظر بگیرد. نتایج، عملکرد برتر مدل مورد مطالعه را نسبت به مدل‌های آماری و یادگیری ماشین نشان می‌دهد.

واژه‌های کلیدی:تحلیل بقای عمیق، مدل‌های نرخ خطر مبنا، زمان پاکسازی حادثه، تابع بقا، یادگیری چند وظیفه‌ای

کد موضوع‌بندی ریاضی (۲۰۲۰): xxAxx، xxBxx، xxCxx.

---

[۱] فاطمه محمدی، fm357016@gmail.com

۱ مقدمه

حوادث ترافیکی تاثیرات منفی بر ایمنی و کارایی سیستم‌های حمل و نقل دارند. از جمله این تاثیرات می‌توان به ایجاد ترافیک و تأخیر در عبور و مرور، آسیب‌ها و تلفات جانی و بروز حوادث ثانویه اشاره کرد. بنابراین به عنوان یکی از چالش‌های اصلی در مدیریت ترافیک محسوب می‌شوند. سازمان‌های مدیریت ترافیک برای به‌کارگیری استراتژی‌های مناسب و کاهش این اثرات منفی نیاز به پیش‌بینی مدت زمان حادثه دارند. مدت زمان حادثه را می‌توان به چهار مرحله تقسیم کرد: زمان تشخیص، زمان پاسخ، زمان پاکسازی و زمان بازیابی. زمان پاکسازی که زمان ورود بین یک تیم پاسخ و پاکسازی کامل حادثه است طولانی‌ترین مرحله است. یکی از اهداف مدل‌بندی داده‌های حوادث ترافیکی، پیش‌بینی زمان پاکسازی حادثه است. داده‌های زمان پاکسازی حادثه داده‌های مدت زمان هستند. یکی از ویژگی‌های این داده‌ها، وابستگی به مدت زمان است. برای در نظر گرفتن این ویژگی عمدتاً از مدل‌های آماری نرخ خطر مبنا استفاده شده است . نام و منرینگ (۲۰۰۰)؛ جونز و همکاران (۱۹۹۱)؛ چانگ (۲۰۱۰) .نام و منرینگ (۲۰۰۰) مدت زمان حادثه را با استفاده از یک مدل خطر متناسب پارامتری که متغیر وابسته آن نرخ خطر است، تحلیل کردند. آنها گزارش داده‌اند که پارامترهای براورد شده و عملکرد پیش‌بینی مدل زمان شکست شتابیده (AFT) در طول سال‌های مختلف پایدار است. مدل خطرات متناسب کاکس دارای فرض خطرات متناسب هستند که این فرض ممکن است برای همه متغیرهای مستقل برقرار نباشد. در مطالعات قبلی برای پیش‌بینی مدت زمان حادثه از مدل‌های زمان شکست شتابیده (AFT) استفاده شده است. جونز و همکاران (۱۹۹۱) مدل زمان شکست شتابیده (AFT) که متغیر وابسته آن میانگین مدت زمان حادثه بود برای مدل‌بندی داده‌های حوادث ترافیکی به کار بردند. چانگ (۲۰۱۰)مدل زمان شکست شتابیده را برای مجموعه داده‌های حادثه در مقیاس بزرگ در بزرگراه اعمال کرد توسلی حجتی و همکاران (۲۰۱۳) ناهمگنی مشاهده نشده را با اعمال پارامترهای تصادفی در مدل زمان شکست شتابیده در نظر گرفت. مدل زمان شکست شتابیده (AFT) یک مدل پرکاربرد است که اثرات ثابت با زمان عوامل تاثیرگذار را فرض می‌کند. همچنین یکی دیگر از فرضیات این مدل‌ها رابطه خطی متغیرهای مستقل با لگاریتم میانگین زمان پاکسازی است. این دو فرض ممکن است برای داده‌های حوادث ترافیکی و در یک آزادراه برقرار نباشد. بنابراین استفاده از این مدل‌ها دارای محدودیت‌هایی می‌باشد. یکی دیگر از روش‌های رایج برای مدل‌سازی زمان پاکسازی حادثه مدل‌های یادگیری ماشین از جمله مدل جنگل تصادفی، ماشین گرادیان تقویتی و شبکه‌های عصبی مصنوعی است. با این حال مدل‌های یادگیری ماشین میانگین زمان پاکسازی را به عنوان متغیر وابسته در نظر می‌گیرند و ویژگی وابستگی به مدت زمان را نادیده می‌گیرند. مطالعه‌ای اخیر گزارش داده است که عملکرد مدل‌های یادگیری ماشین قابل مقایسه یا بدتر از مدل‌های زمان شکست شتابیده است . وی و همکاران (۲۰۰۷) برای پیش‌بینی مدت زمان حادثه از پیش‌بینی متوالی استفاده کردند که عملکرد مدل‌های شبکه های عصبی مصنوعی را بهبود بخشید. حمد و همکاران (۲۰۲۰) مدل جنگل تصادفی (RF) را برای پیش‌بینی زمان پاکسازی حادثه توسعه دادند. همچنین ما و همکاران (۲۰۱۷) یک مدل GBM را برای پیش‌بینی زمان پاکسازی حادثه و همچنین تحلیل عوامل تاثیرگذار توسعه دادند. نتایج آن‌ها عملکرد پیش‌بینی بهتر GBM را نسبت به RF وANN نشان داد. لی و همکاران (۲۰۲۰) پیش‌بینی مدت زمان حادثه را با استفاده از شبکه عصبی عمیق انجام داد. نتایج آن‌ها نشان داد که DNN بهتر از مدل‌های یادگیری ماشین معمولی با هدف پیش‌بینی میانگین زمان حادثه، عمل می‌کند. مطالعه اخیر تانگ و همکاران (۲۰۲۰) گزارش داد که عملکرد پیش‌بینی مدل‌های یادگیری ماشین قابل مقایسه یا بدتر از مدل‌های زمان شکست شتابیده (AFT) با توزیع های فرضی مختلف بوده است. مطالعات قبلی از مدل‌های یادگیری ماشین و شبکه های عصبی عمیق به عنوان رگرسیون برای زمان پاکسازی حادثه استفاده می‌کردند. اگر هدف این مدل‌ها پیش بینی زمان پاکسازی حادثه باشد نمی‌توانند وابستگی به مدت زمان را که یکی از ویژگی‌های مهم داده‌های زمان تا رویداد

است به طور مستقیم در نظر بگیرند. این مدل‌ها تنها یک وظیفه انجام می‌دهند و آن هم پیش‌بینی میانگین زمان پاکسازی حادثه است. اگر مدل‌های یادگیری ماشین و شبکه‌های عصبی عمیق[1] با هدف برآورد تابع بقا طراحی شوند می‌توانند مدل‌های مناسبی برای تحلیل داده‌های زمان تا رویداد باشند زیرا می‌توانند ویژگی وابستگی به مدت زمان در این داده‌ها را در نظر بگیرند. ایشواران و همکاران (۲۰۰۸) با توسعه جنگل بقای تصادفی که مدل جنگل تصادفی را برای داده‌های زمان تا رویداد گسترش می‌دهد، تابع بقا را برآورد کند. هوثورن و همکاران (۲۰۰۶) مدل ماشین تقویتی گرادیان اصلاح شده[2] را پیشنهاد کردند. کاتزمن و همکاران (۲۰۱۸) Deepsurve را ابداع کردند که با استفاده از شبکه‌های عصبی عمیق روابط غیرخطی و تعاملی را در نظر بگیرد و تابع بقا را برآورد کند. این مدل از مدل‌های خطرات متناسب کاکس و جنگل بقای تصادفی بهتر عمل کرد. لی و همکاران (۲۰۱۸) DeepHitرا پیشنهاد کردند که به طور مستقیم تابع بقا را با استفاده از شبکه‌های عصبی عمیق برآورد می‌کند. با اینکه مدل‌های یادگیری ماشین با هدف براورد تابع بقا می‌توانند وابستگی به مدت زمان را در نظر بگیرند اما به صراحت نمی‌توانند زمان پاکسازی حادثه را پیش‌بینی کنند. در ادامه مدلی ارائه می‌شود که می‌تواند به طور همزمان هم تابع بقا را برآورد کند و هم مدت زمان حادثه را با استفاده از ساختار یادگیری چند وظیفه‌ای پیش‌بینی کند.

۲ تعاریف اولیه

داده‌های زمان پاکسازی حادثه داده‌های مدت زمان هستند که زمانی است تا وقوع رویداد. منظور از رویداد پاکسازی کامل حادثه است. برخی از مفاهیم مورد نیاز در تحلیل داده‌های زمان تا رویداد شامل تابع بقا، تابع نرخ خطر، داده سانسورشده، ویژگی وابستگی به مدت زمان و میانگین طول عمر می‌باشد. در ابتدا لازم است این مفاهیم توضیح داده شود.

تعریف ۱.۲. تابع بقا[3] احتمال این است که زمان بقا بیش از یک زمان مشخص باشد. این تابع، که به نام تابع قابلیت اطمینان نیز مشهور است احتمال زنده ماندن در طول زمان t را محاسبه می‌کند. عموماً این تابع را با استفاده از روش‌های ناپارامتری مانند کاپلان مه یر یا نلسون آلن و یا روش های پارامتریک و از روی تابع توزیع به دست می‌آورند. متغیر تصادفی T که همان زمان بقا است می‌تواند گسسته یا پیوسته باشد اما همواره باید $T \geq ۰$ باشد.

$$S(t) = Pr(T < t) = ۱ - F(t)$$

تعریف ۲.۲. تابع نرخ خطر[4] که با $h(t)$ نشان داده می‌شود، میزان مرگ و میر لحظه‌ای را نشان می‌دهد. این تابع شانس رخداد پیشامد را در لحظه t با شرط آن که واحد آزمایشی تا این لحظه زنده باشد، بیان می‌کند.

- اگر T متغیر تصادفی گسسته باشد:

$$h(t) = Pr(T = t | t \leq T)$$

- اگر T متغیر تصادفی پیوسته باشد:

$$h(t) = \lim_{\Delta t \to ۰} \frac{Pr(t < T < t + \Delta t)}{\Delta t}$$

[1] Deep Neural Network
[2] modified Gradient Boosting Machine
[3] Survival Function
[4] Hazard Function

با استفاده از تابع نرخ خطر می‌توان تابع بقا را برآورد کرد. s متغیر انتگرال‌گیری است و مقدار $[\circ, t]$ را می‌گیرد.

$$S(t) = exp(-\int_{\circ}^{t} h(s)\ \mathrm{ds})$$

با استفاده از تابع نرخ خطر می‌توان تابع نرخ خطر تجمعی را به دست آورد که با نماد $H(t)$ نشان داده می‌شود.

$$H(t) = \int_{\circ}^{t} h(s)\ \mathrm{ds}$$

بنابراین:

$$S(t) = exp(-H(t))$$

تعریف ۳.۲. در کنار داده‌های زمان تا رویداد، وضعیت آن داده هم ثبت می‌شود. منظور از وضعیت داده این است که آیا داده سانسور شده است یا داده کامل است. منظور از داده کامل این است که اطلاعات داده از اول مطالعه تا زمانی که رویداد برایش اتفاق می‌افتد موجود باشد. منظور از داده سانسورشده هم این است که اطلاعات کاملی از زمان دقیق وقوع رویداد برای برخی نمونه‌ها وجود ندارد. سانسورها می توانند انواع مختلفی داشته باشند:

- سانسور راست

  زمانی که رویداد بعد از پایان مطالعه رخ دهد یا به عبارتی دیگر رویداد تا انتهای مطالعه رخ نمی‌دهد.

- سانسور چپ

  زمانی که رویداد قبل از شروع مطالعه رخ داده باشد.

- سانسور فاصله ای

  زمانی که رویداد بین دو بازه زمانی مشخص رخ داده باشد.

تعریف ۴.۲. ویژگی وابستگی به مدت زمان به این معنی است که احتمال پایان پاکسازی حادثه به طول مدت زمانی که پاکسازی به طول انجامیده است بستگی دارد. بدین منظور که در این نوع داده‌ها متغیر وابسته، زمان تا رویداد است. این زمان همواره مثبت است و در مطالعات تحلیل بقا می‌تواند به صورت پیوسته یا گسسته در نظر گرفته شود.

تعریف ۵.۲. مساحت زیر منحنی تابع بقا، میانگین طول عمر گفته می‌شود.

$$E(T) = \int_{\circ}^{\infty} S(t)\mathrm{dt}$$

۳ مدل‌های تحلیل بقا

در آمار برای مدل‌بندی داده‌های زمان تا رویداد از مدل‌های تحلیل بقا استفاده می‌شود این مدل‌ها شامل مدل‌های نرخ خطر مبنا هستند. مدل‌های نرخ خطر مبنا قادر هستند که ویژگی وابستگی به مدت زمان را در نظر بگیرند. این مدل‌ها می‌توانند دو نوع متغیر وابسته

ازجمله میانگین زمان رویداد و تابع نرخ خطر را در نظر بگیرند. مدل‌های نرخ خطر مبنا که متغیر وابسته آن‌ها تابع نرخ خطر است شامل مدل‌های خطرات متناسب کاکس[۵] است. همچنین مدل‌هایی که متغیر وابسته آن‌ها میانگین زمان بقا است شامل مدل‌های زمان شکست شتابیده (AFT) می‌باشد. در حالی‌که هدف مدل خطرات متناسب کاکس برآورد اثرات متغیرهای مستقل بر توابع نرخ خطر یا بقا است، هدف مدل زمان شکست شتابیده برآورد مستقیم اثرات بر روی مدت زمان (یا لگاریتم زمان)است.

- مدل خطرات متناسب کاکس

$$h(t,x) = h_{\circ}(t) \times \exp(\beta_{۱} x_{۱} + \cdots + \beta_p x_p)$$

در مدل کاکس، $h_{\circ}(t)$ تابع نرخ خطر پایه است و نشان می‌دهد زمانی‌که هیچ‌یک از متغیرهای مستقل بر پاسخ تأثیر نداشته باشند، میزان خطر وقوع رویداد چقدر است. همچنین، x_i برای $i = ۱, \cdots, p$ ویژگی‌هایی هستند که بر نرخ خطر تأثیر می‌گذارند و β_i نیز میزان تأثیر هر یک از آن‌ها را مشخص می‌کند. بردار β شامل ضرایب مربوط به متغیرهای مستقل، و $\mathbf{X}$ برداری از این متغیرهای مستقل یا همان ویژگی‌ها است.

این مدل‌ها یک خطر متناسب را فرض می‌کنند که نشان می‌دهد اثر متغیرهای مستقل ثابت از زمان است. منظور از فرض خطرات متناسب این است که نسبت مخاطره برای واحد امi به واحد j ام مستقل از زمان باشد. نرخ خطر پایه مقدار خطر را زمانی‌که همه متغیرهای مستقل صفر هستند، نشان می‌دهد. خطر به عنوان یک تابع ضربی از یک خطر پایه و یک تابعی از متغیرهای مستقل است. مدل کاکس یک مدل نیمه پارامتری است ولی می‌تواند پارامتری هم باشد.

مدل زمان شکست شتابیده این مدل یک رابطه خطی بین لگاریتم زمان بقا و متغیرهای مستقل را فرض می‌کند.

$$\ln t = \beta_{\circ} + \beta_{۱} x_{۱} + \cdots + \beta_p x_p + \epsilon$$

متغیر وابسته این مدل را می‌توان تابع نرخ خطر نیز در نظر گرفت:

$$h(t,x) = exp(-\beta \mathbf{X}) \cdot h_{\circ}(exp(-\beta \mathbf{X}) \times t)$$

می توان توزیع‌های فرضی وایبول، لگ نرمال و لگ لجستیک را در نظر گرفت که در نرخ خظر پایه تاثیرگذار هستند.

مدل‌های شکنندگی

در برخی مطالعات تحلیل بقا، ممکن است زیر گروه‌های جامعه دارای یک متغیر تصادفی مشترک مشاهده نشده باشند که قادر به وارد کردن آنها در مطالعه نیستیم. اگر اثر این متغیرها در نظر گرفته نشود بخشی از تغییرات تابع نرخ خطر که می‌توانست با وجود این عامل توضیح داده شود کاهش می‌یابد و با جمله خطا جمع شده و در نتیجه باعث افزایش تغییرات تابع نرخ خطر نسبت به حالتی که عامل مورد نظر در مدل وجود داشته باشد، می‌گردد که باعث به دست آمدن براوردگرهای اریب برای پارامترهای مدل رگرسیونی می‌گردد. رویکرد پارامتر تصادفی (عامل شکنندگی) روشی است که به طور گسترده برای پرداختن به ناهمگنی مشاهده نشده پذیرفته شده است به طوری‌که به برخی از پارامترها اجازه می‌دهد تا در مشاهدات با توزیع فرضی، تغییر کنند.

$$\beta_n = \beta + \delta_n$$

[5] Cox proportional hazard Model

این عامل شکنندگی را می‌توان هم در مدل‌های خطرات متناسب به کار برد و هم در مدل‌های زمان شکست شتابیده از آن استفاده کرد. اگر این عامل تصادفی در مدل‌های زمان شکست شتابیده به کار رود به نام مدل‌های (Random parameter AFT) شناخته شده و پیش‌بینی و تحلیل دقیق‌تری از عوامل تاثیرگذار مرتبط با ناهمگنی مشاهده نشده ارائه می‌کنند.

۴ مدل‌های یادگیری ماشین

یکی دیگر از روش‌های رایج برای مدل‌سازی داده‌های زمان پاکسازی حادثه، مدل‌های یادگیری ماشین می‌باشد. از جمله می‌توان به مدل‌های شبکه های عصبی مصنوعی (ANN)، جنگل تصادفی(RF) و ماشین گرادیان تقویتی(GBM) اشاره کرد. مدل‌های شبکه های عصبی عمیق(DNN) یک روند اجتناب ناپذیر در زمینه‌های مختلف تحقیقاتی می‌باشد.

۱.۴ جنگل بقای تصادفی

جنگل تصادفی[۶] یکی از الگوریتم‌های یادگیری ماشین مبتنی بر درخت تصمیم است که برای مسائل رگرسیون و طبقه‌بندی استفاده می‌شود. همچنین اگر مدل جنگل تصادفی با هدف برآورد تابع بقا طراحی شود[۷] می‌تواند برای داده‌های زمان تا رویداد مناسب باشد زیرا ویژگی وابستگی به مدت زمان را در این داده‌ها در نظر می‌گیرد. در این نوع الگوریتم‌ها از چندین مدل ضعیف استفاده می‌شود تا عملکرد مدل کلی را بهبود ببخشد. در مدل جنگل تصادفی از چندین درخت تصمیم استفاده می‌شود. همان‌طور که در جنگل تصادفی با هدف رگرسیون مدل‌های ضعیف درخت تصمیم به صورت تصادفی و مستقل آموزش می‌بینند در مدل جنگل بقای تصادفی (RSF) هم درخت‌ها به طور جداگانه با هدف برآورد تابع بقا آموزش می‌بینند. در مدل جنگل بقای تصادفی هر درخت تصمیم به جای پیش‌بینی یک مقدار عددی در گره اخر، تابع بقا را به روش‌های ناپارامتری مثل کاپلان مه یر برآورد می‌کند. پس از آموزش درخت‌ها، برای هر نمونه جدید پیش‌بینی تابع بقا به صورت میانگین از پیش‌بینی تمام درخت‌ها محاسبه می‌شود.

الگوریتم ۱. فرض کنید داده‌ها به صورت زیر باشند:

$$\mathcal{D} = \{(X_i, T_i, \delta_i)\}_{i=1}^{n}$$

که در آن: X_i: بردار ویژگی‌ها، T_i: زمان مشاهده‌شده (کمترین مقدار بین زمان وقوع و سانسور)، و $\delta_i \in \{0, 1\}$: نشانگر وقوع رویداد است. (۱=رخ داده، ۰=سانسور شده)

- ساخت درخت‌ها

۱. یک نمونه بوت‌استرپ از داده‌ها بگیرید.

۲. در هر گره، m ویژگی به‌طور تصادفی انتخاب کنید.

۳. برای هر تقسیم، آماره آزمون لاگ–رتبه را محاسبه کنید.

۴. تقسیم با بیشترین مقدار آماره لاگ–رتبه را انتخاب کنید.

[6] Random Forest

[7] Random Survival Forest

۵. تقسیم گره‌ها تا رسیدن به شرط توقف ادامه می‌یابد.

- برآورد در گره نهایی

در هر گره نهایی، تابع بقا با استفاده از برآوردگر کاپلان–مایر محاسبه می‌شود:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

که: d_j تعداد رخدادها در زمان t_j و r_j تعداد افراد در معرض خطر در زمان t_j ، می‌باشد.

- پیش‌بینی نهایی جنگل

میانگین برآوردهای بقا از تمام درخت‌ها به صورت زیر محاسبه می‌شود:

$$\hat{S}(t \mid X) = \frac{1}{B} \sum_{b=1}^{B} \hat{S}^{(b)}(t \mid X)$$

۲.۴ ماشین گرادیان تقویت‌شده

ماشین گرادیان تقویتی[8] یکی از الگوریتم‌های یادگیری جمعی است . این مدل هم بر مبنا چندین درخت تصمیم است. تفاوت بین مدل‌های Adaboost و ماشین گرادیان تقویتی در نحوه به روزرسانی مدل و نحوه تعامل مدل‌های ضعیف است. Adaboost مدل‌های ضعیف درخت تصمیم را به صورت متوالی تقویت می‌کند و در هر مرحله به مواردی که اشتباه پیش‌بینی شده‌اند وزن‌هایی اختصاص می‌دهد تا در مدل بعدی بیش‌تر به آن‌ها توجه شود. در مدل ماشین گرادیان تقویتی به همه نمونه‌ها وزن‌های یکسانی اختصاص داده می‌شود و با استفاده از روش عددی گرادیان نزولی پیش‌بینی‌ها بهبود می‌بخشد. به عبارت دیگر در هر مرحله در مدل ماشین گرادیان تقویتی هدف بهبود عملکرد کلی مدل در پیش‌بینی است.

۳.۴ شبکه عصبی مصنوعی

شبکه‌های عصبی[9] شامل سه نوع لایه از جمله لایه ورودی، لایه خروجی و لایه‌های پنهان هستند. تعداد نورون‌ها در لایه ورودی برابر با تعداد ویژگی‌های داده‌های ورودی است و تعداد نورون‌ها در لایه خروجی بستگی به هدف پیش‌بینی دارد. تعداد نورون‌های لایه پنهان بین تعداد نورون‌های لایه ورودی و تعداد نورون‌های لایه خروجی قرار می‌گیرد. به‌عبارتی، تعداد نورون‌ها در لایه پنهان ابرپارامتری است که با آزمایش مقادیر مختلف تعیین می‌شود تا بهترین عملکرد آموزشی شبکه عصبی به دست آید.

آموزش شبکه عصبی شامل دو مرحله از پیش‌انتشار و پس‌انتشار است. پیش‌انتشار فرآیندی است که در آن داده‌های ورودی به لایه ورودی ارسال می‌شود، توسط لایه پنهان پردازش می‌شود و در نهایت به لایه خروجی منتقل می‌گردد. جزئیات پیش‌انتشار به‌صورت زیر است:

[8] Gradient Boosting Machine

[9] Neural Network

اگر شبکه عصبی با یک لایه پنهان و یک لایه خروجی داشته باشیم، رابطه بین لایه ورودی و لایه پنهان به‌صورت ریاضی به‌صورت زیر بیان می‌شود:

$$\text{neth}_1 = \sum w_i x_i + b_1$$

که در آن x_i بردار ورودی است، w_j بردارهای وزن بین لایه ورودی و لایه پنهان هستند، و b_1 بایاس بین لایه‌های محاسبه‌شده است که به‌طور انعطاف‌پذیری برای تنظیم عملکرد شبکه عصبی طراحی شده است. سپس، neth_1 از توابع فعالسازی از جمله ReLu عبور می‌کند. تابع فعالسازی ReLu به صورت $\max(0, neth_1)$ است.

$$h_i = g(\text{neth}_1) = \max(0, \text{neth}_1)$$

h_iمقادیر گره‌های لایه پنهان هستند که دوباره در وزن‌های اتصالات بین لایه پنهان و لایه خروجی w_j ضرب شده و با مقدار بایاس هر گره لایه خروجی جمع می‌شود. در نهایت از تابع فعالسازی لایه آخر $g(.)$ عبور کرده و مقدار خروجی هر گره در لایه خروجی که همان Z است، به دست می‌آید. Z به‌صورت زیر محاسبه می‌شود:

$$\text{Net}_z = \sum w_j h_i + b_2$$

و

$$Z = g(\text{Net}_z) = \max(0, \text{Net}_z)$$

که در آن w_j بردارهای وزن بین لایه پنهان و لایه خروجی هستند و b_2 بایاس بین لایه پنهان و لایه خروجی است.

اگر Z با خروجی مورد انتظار مطابقت نداشته باشد، خطای پیش‌بینی با استفاده از تابع ضرر انتخاب‌شده محاسبه خواهد شد. سپس، پس‌انتشار خطای پیش‌بینی را به نورون‌های شبکه باز می‌گرداند و از الگوریتم نزول گرادیان برای اصلاح وزن‌های نورون‌ها به‌منظور کاهش خطاهای پیش‌بینی استفاده می‌کند. هنگامی که خطای پیش‌بینی با انتظارات مطابقت داشت، پیش‌بینی شبکه عصبی انجام می‌شود.

۵ مدل تحلیل بقای عمیق چندوظیفه‌ای

در بحث شبکه‌های عصبی اگر چند وظیفه باهم مرتبط بودند می‌توانیم یک مدل شبکه عصبی طراحی کنیم که همزمان قادر به انجام دو وظیفه باشد. مدل مورد استفاده در این مطالعه یک شبکه عصبی عمیق (منظور تعداد زیاد لایه‌های پنهان) دو وظیفه ای است که می‌تواند ویژگی وابستگی به مدت زمان را با استفاده از اطلاعات مشترکی که از برآورد تابع بقا به دست می‌آورد در پیش‌بینی زمان پاکسازی حادثه در نظر بگیرد. در یادگیری چند وظیفه ای، چندین کار مرتبط به طور همزمان یاد می‌گیرند و از اطلاعات مشترک بین آن‌ها برای بهبود عملکرد کلی استفاده می‌شود. یادگیری چند وظیفه ای به جای آموزش یک مدل جداگانه برای هر کار، یک مدل واحد را برای انجام چندین کار آموزش می‌دهد. با ورود یک بردار متغیرهای مستقل به مدل، چندین خروجی (هدف یا متغیرهای وابسته) به دست می‌آوریم. در

یادگیری چند وظیفه‌ای، برخی از لایه‌ها یا پارامترها در بین وظایف به اشتراک گذاشته می‌شوند و به مدل اجازه می‌دهند ویژگی‌های مشترکی را که برای همه وظایف مفید هستند، یاد بگیرد. این مدل به طور همزمان روی وظایف مختلف آموزش داده می‌شود و لایه‌های مشترک بر اساس زیان ترکیبی از همه وظایف به روز می‌شوند. علاوه بر لایه‌های مشترک، مدل‌های یادگیری چند وظیفه ای معمولا دارای لایه‌های اختصاصی مربوط به وظایف هستند که جنبه‌های منحصر به فرد هر وظیفه را مدیریت می‌کنند. از جمله مزایای یادگیری چند وظیفه‌ای این است که در صورت مرتبط بودن وظایف، عملکرد آنها با اطلاعات مشترک همدیگر بهبود می‌بخشد و مدل یادگیری چند وظیفه‌ای می‌تواند به عنوان یک تنظیم‌کننده عمل کند و از توجه بیش از حد مدل در یک وظیفه جلوگیری کند. همچنین از معایب ساختار یادگیری چند وظیفه‌ای می‌توان گفت با افزایش تعداد وظایف پیچیدگی و هزینه محاسباتی یادگیری چند وظیفه ای می‌تواند به طور قابل توجهی افزایش یابد. نتایج نشان می‌دهند که انجام وظایف چندگانه به طور مکمل عملکرد وظایف مختلف را افزایش می دهد اگر اطلاعات نهفته ای که از مولفه‌های مشترک استخراج می‌شود به دو وظیفه مشابه کمک کند ساختار یادگیری چند وظیفه ای می‌تواند به طور موثر اطلاعات پنهان مشترک را جذب کند. وظیفه پیش‌بینی میانه زمان پاسخگویی بخشی از برآورد تابع بقا است بنابراین دو وظیفه مرتبط هستند. در این مطالعه به دنبال انجام دو وظیفه یعنی برآورد تابع بقا و پیش‌بینی میانگین زمان پاکسازی حادثه به طور همزمان هستیم. وظیفه برآورد تابع بقا مربوط به مدل‌های خطرات متناسب کاکس است. همچنین پیش‌بینی میانگین زمان پاکسازی حادثه وظیفه مدل‌های زمان شکست شتابیده است.

برای لایه اختصاصی مربوط به وظیفه براورد تابع بقا، این مطالعه ۱۰ دقیقه را به عنوان فاصله زمانی برای برآورد تابع بقا با توجه به کاربرد آن در مدیریت حادثه در نظر گرفته است. تابع بقا یک احتمال است، احتمال بقا در هر کدام از بازه‌های ۱۰ دقیقه ای محاسبه می‌شود. برای این منظور تابع فعالسازی لایه آخر سافت مکس[10] در نظر گرفته می‌شود. برای وظیفه برآورد تابع بقا دو تابع زیان در نظر گرفته شده است:

- تابع زیان لگاریتم درستنمایی

در داده‌های زمان تا رویداد داده‌های سانسور شده وجود دارد. این داده‌ها اطلاعات کاملی درباره چگالی به ما نمی‌دهند و بنابراین در تابع درستنمایی باید به همان میزانی که به ما اطلاعات می‌دهند اثر داشته باشند بنابراین باید تابع ماسکی تعریف شود تا داده‌های سانسور شده را از بدون سانسورتشخیص دهد.

$$L_1 = -\sum_{t=1}^{n_{out}} y_t \cdot \log \widehat{y_t}$$

n_{out} تعداد بازه‌هایی است که تابع بقا در آن بازه‌ها برآورد می‌شود. y_t مقدار پاسخ واقعی است که از داده‌های مشاهده شده به دست می‌آید و میزان وقوع رویداد برای یک حادثه در زمان tام است و $\hat{y}_t$ احتمال برآوردشده رخ دادن پاکسازی حادثه در بازه زمانی tام است که خروجی مدلی است که وظیفه برآورد تابع بقا را انجام می‌دهد.

- تابع زیان رتبه بندی[11]

این تابع زیان با استفاده از نرخ خطر تجمعی ترتیب وقوع رویدادها را بررسی می‌کند به طوری‌که حادثه‌هایی که رویداد برای آن‌ها زودتر اتفاق می‌افتد احتمال بقا کمتر و نرخ خطر بیش‌تری نسبت به حادثه‌هایی دارند که همان رویداد را دیرتر تجربه کردند. اگر مدلی بتواند این ترتیب‌ها را درست پیش‌بینی کند مثل این است که توانسته تابع بقا را دقیق برآورد کند. این تابع زیان برای

[10] Softmax

[11] Rank Loss

بررسی این است که آیا مدل ترتیب درست وقوع رویدادها را پیش‌بینی کرده است یا خیر. هر چه این زیان کمتر باشد یعنی براورد تابع بقا دقیق‌تر است. با توجه به عملکرد این تابع زیان باید وقتی که می‌خواهیم این زیان را محاسبه کنیم تابع ماسکی تعریف کنیم تا برای داده‌هایی که در طی زمان اندازه‌گیری شدند هم بتواند ترتیب درست رویدادها را به درستی اعمال کند. این تابع زیان مانند شاخص تطابق وابسته به زمان (C-Index) عمل می کند.

$$L_۲ = \sum_{i \neq j} A_i, j\eta(\hat{H}(t_i|x_{ti}), \hat{H}(t_i|x_{tj}))$$

که در آن:

$$\eta(x, y) = exp(\frac{y - x}{\sigma})$$

و

$$\mathbf{A}_{i,j} = \begin{cases} ۱ & \text{اگر } t_i < t_j \\ ۰ & \text{در غیر این صورت} \end{cases}$$

در لایه آخر شبکه عصبی مربوط به وظیفه پیش‌بینی میانگین زمان پاکسازی حادثه، از تابع فعالسازی خطی استفاده شده است. تابع زیان مربوط به آموزش این وظیفه مجموع میانگین مربعات خطا (MSE) است.

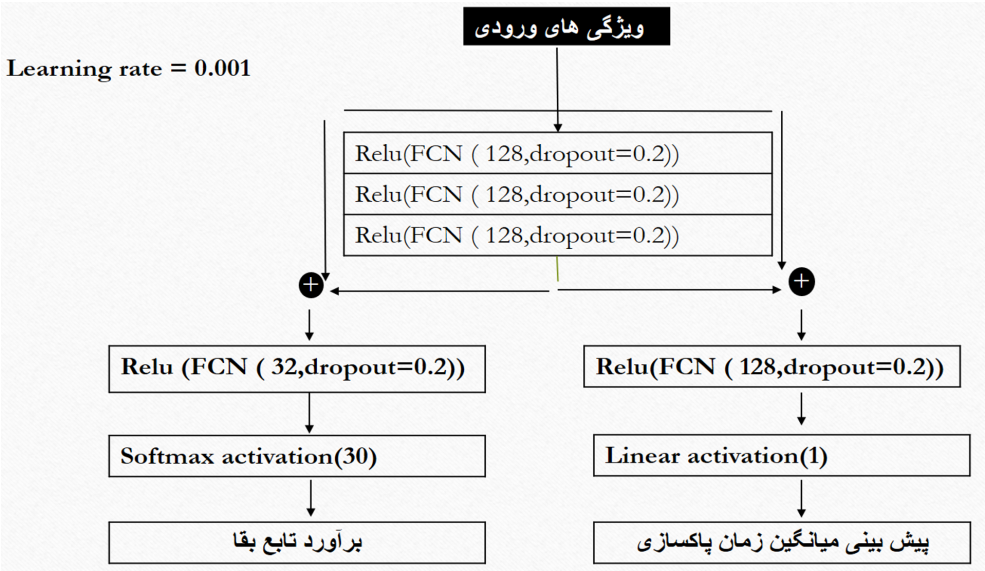
$$MSE = \frac{۱}{K} \sum_{k=۱}^{K} (T_k - \widehat{T_k})^۲$$

در این فرمول Kتعداد حوادث مشاهده شده است. T_k و $\hat{T}_k$ به ترتیب مدت زمان واقعی پاکسازی حادثه و مدت زمان پاکسازی تخمین‌زده شده توسط مدل می باشد.

با تابع زیان ترکیبی زیر لایه‌های مشترک آموزش داده می‌شود:

$$L_{total} = L_۱ + w_۱ \cdot L_۲ + w_۲ \cdot L_۳$$

وزن‌های هر تابع زیان جزو ابر پارامتر هایی است که با استفاده از روش‌های جستجوی شبکه‌ای کالیبره می شوند. از الگوریتم بهینه سازی Adam برای بهینه کردن وزن‌های شبکه عصبی استفاده شده است. نرخ یادگیری در این مدل 0.001 در نظر گرفته شده است. ساختار شبیه‌سازی مدل در شکل ۱ نشان داده شده است. در شکل ۱ بین دو وظیفه، لایه‌های مشترک وجود دارند که ویژگی‌های پنهان مشترک بین دو وطیفه را استخراج می‌کنند. تابع فعالسازی لایه‌های مشترک ReLu می‌باشد و برای هر لایه پنهان در لایه‌های مشترک لایه Dropout در نظر گرفته شده است که روشی برای جلوگیری از بیش‌برازش مدل است. در این روش به طور تصادفی ۲۰ درصد از نورون‌های هر لایه در محاسبات خروجی غیرفعال می‌شود. لایه‌های مشترک با تابع زیان ترکیبی L_{total} آموزش می‌بیند و ویژگی‌های مشترک بین دو وظیفه را استخراج می‌کند. تعداد گره‌ها در هر لایه ۱۲۸ نورون است و لایه‌ها کاملاً به هم متصل هستند. در یادگیری چند وظیفه‌ای باید وظایف باهم مرتبط باشند تا از ویژگی‌های مشترک‌شان برای بهبود عملکرد هر وظیفه بهره برد. بعد از لایه‌های مشترک بین وظایف به وظایف اختصاصی می‌رسیم. شبکه عصبی اول وظیفه برآورد تابع بقا را بر عهده دارد و پارامترهای این شبکه با تابع زیان $L_۱$



شکل ۱: ساختار مدل تحلیل بقای عمیق چند وظیفه‌ای

و L_2 آموزش می‌بیند و تابع فعالسازی لایه آخر این مدل تابع سافت مکس است که برای محاسبه احتمال وقوع پاکسازی حادثه در هر بازه t است. شبکه عصبی دوم با وظیفه پیش‌بینی زمان پاکسازی است که پارامترهای این مدل نیز با تابع زیان L_3 آموزش می‌بیند و تابع فعالسازی لایه آخر این شبکه تابع خطی است.

۶ مثال کاربردی

داده‌های حوادث ترافیکی شامل ۳۹۷۶۰ حادثه با ویژگی‌های زمانی، جغرافیایی و حادثه است. داده‌ها مربوط به تصادف آزادراه‌های کشور کره جنوبی از سال ۲۰۱۴ تا ۲۰۱۹ می‌باشد. به طور تصادفی ۷۰ درصد از داده‌ها برای آموزش و اعتبارسنجی مدل و ۳۰ درصد از داده‌ها برای آزمون مدل استفاده شده است. اغلب برخی ویژگی‌های گمشده یا غیر واقعی در مجموعه داده‌ها وجود دارد که با روش‌های پیش‌پردازش داده‌ها برطرف می‌شوند. فقط از داده‌های کامل حادثه استفاده شده است که شامل تمام ویژگی‌های زمانی، جغرافیایی حادثه و جریان ترافیک است با حذف نقاط پرت در نهایت ۱۶۹۲۹ داده برای توسعه مدل به کار گرفته شده است. در این پژوهش همه داده‌ها کامل هستند و سانسور وجود ندارد. تمام متغیرهای طبقه بندی در طول آموزش مدل به متغیرهای باینری تبدیل می‌شوند و متغیرهای مربوط به ترافیک پیوسته هستند و استاندارد می‌شوند. در این مثال، مدل زمان شکست شتابیده را با در نظر گرفتن توزیع‌های فرضی مختلف برای نرخ خطر پایه، به کار می گیرد .

مدل زمان شکست شتابیده (AFT)(توزیع وایبول)
مدل زمان شکست شتابیده (AFT)(توزیع لگ نرمال)
مدل زمان شکست شتابیده (AFT)(توزیع لگ لجستیک)

در این مطالعه مدل زمان شکست شتابیده با یک پارامتر تصادفی (RPAFT) برای در نظر گرفتن ناهمگنی مشاهده نشده به کار گرفته شد.از۲ مدل (GBM) و (RF) با دو هدف پیش‌بینی میانگین زمان پاکسازی و براورد تابع بقا برای بررسی عملکرد مدل هدف

استفاده شده است. ابرپارامترهای اصلی با استفاده از تکنیک (K-fold) تنظیم شده اند. در این روش هر ترکیب از ابر پارامترها ارزیابی می‌شوند و خطای هر ترکیب به دست می‌آید. ترکیبی که کمترین خطا را در معیارهای میانگین خطای قدر مطلق (MAE) و میانگین درصد خطای قدر مطلق (MAPE) و شاخص تطابق وابسته به زمان (c-td) داشته باشد، انتخاب می شود.

برای جنگل تصادفی (RF) چهار ابرپارامتر از جمله تعداد درخت، حداکثر عمق هر درخت، نرخ نمونه‌گیری برای آموزش هر درخت و تعداد متغیرهای مورد استفاده در هر گره تقسیم وجود دارد. برای ماشین گرادیان تقویتی (GBM) نیز چهار ابر پارامتر از جمله تعداد درخت، حداکثر عمق هر درخت، نرخ یادگیری و تعداد متغیرهای در نظرگرفته‌شده در هر درخت وجود دارد.

مدل هدف، جداگانه با هر کدام از مدل‌های تک وظیفه‌ای شبکه عصبی عمیق (DNN) مقایسه می‌شود تا نشان‌دهنده بهبود عملکرد یادگیری چند وظیفه‌ای باشد. ترکیب ابر پارامترها از جمله تعداد لایه‌ها و نورون‌های هر شبکه، نرخ یادگیری، نرخ dropout و وزن‌ها برای هر زیان به طور گسترده تنظیم شده‌اند تا معماری مدل نهایی را انتخاب کنند. عملکرد مدل برای پیش‌بینی میانگین مدت زمان رویداد با میانگین خطای قدرمطلق (MAE) و میانگین درصد خطای قدرمطلق (MAPE) ارزیابی می‌شود.

$$MAE = \frac{۱}{k}\sum_{k=۱}^{K} |T_k - \hat{T_k}|$$

$$MAPE = \frac{۱}{k}\sum_{k=۱}^{K} |\frac{T_k - \hat{T_k}}{T_k}| \times ۱۰۰$$

عملکرد برای وظیفه برآورد توزیع زمان بقا (تابع بقا) با شاخص C-td ارزیابی می‌شود :

$$\mathbf{C-td} = \frac{\sum_{i\neq j} A_{i,j} \cdot B_{i,j}}{\sum_{i\neq j} A_{i,j}}$$

که در آن:

$$\mathbf{A}_{i,j} = \begin{cases} ۱ & \text{اگر } t_i < t_j \\ ۰ & \text{در غیر این صورت} \end{cases}$$

و

$$\mathbf{B}_{i,j} = \begin{cases} ۱ & \text{اگر } \hat{H}(t_i|(X_{I,t_i}, X_{T,t_i})) > \hat{H}(t_j|(X_{I,t_j}, X_{T,t_j})) \\ ۰ & \text{در غیر این صورت} \end{cases}$$

این شاخص توانایی مدل را در پیش‌بینی صحیح ترتیب وقوع رویدادها، بر اساس این فرض اندازه‌گیری می‌کند که رویدادهایی که مدت بیشتری طول کشیده‌اند، باید مخاطره کمتری نسبت به رویدادهایی داشته باشند که مدت کوتاه‌تری دوام آورده‌اند. شاخص تطابق وابسته به زمان تمام بازه‌های زمانی مخاطره تجمعی را ارزیابی می‌کند، بنابراین می‌تواند مواردی را که اثرات عوامل تاثیرگذار در طول زمان تغییر می‌کنند (وابسته به زمان هستند) را نیز ارزیابی کند. این ویژگی به مدل اجازه می‌دهد تا در شرایط پویا و پیچیده‌تر، مانند شبکه‌های ترافیکی که متغیرهای مستقل ممکن است در زمان‌های مختلف تأثیرات متفاوتی داشته باشند، عملکرد بهتری داشته باشد.

بحث و نتیجه‌گیری

جداول زیر عملکرد پیش‌بینی مدل هدف را در مقایسه با سایر مدل‌ها نشان می‌دهد.

- مدل‌های آماری

مدل‌های آماری	MAE	MAPE	AIC
مدل زمان شکست شتابیده (AFT) (توزیع وایبول)	۱۸.۸۶۵	۵۴.۴۶۵	۲۲۵۵۲.۸۴۴
مدل زمان شکست شتابیده (AFT) (توزیع لگ نرمال)	۱۸.۳۳۵	۵۱.۳۹۱	۱۸۶۱۳.۵۳۶
مدل زمان شکست شتابیده (AFT) (توزیع لگ لجستیک)	۱۸.۴۶۱	۵۲.۱۹۸	۱۸۷۰۰.۴۲۰
مدل شکنندگی (RPAFT) (توزیع لگ لجستیک)	۱۷.۱۱۴	۴۲.۰۶۶	۱۸۲۷۳.۵۸۳

از بین مدل‌های آماری فوق مدل زمان شکست شتابیده با توزیع لگ لجستیک از مدل‌های دیگر AFT ازنظر MAE و MAPE و AICبهتر عمل می‌کند. همچنین مدل شکنندگی با یک پارامتر تصادفی RPAFT بهتر از مدل زمان شکست شتابیده با توزیع فرضی لگ لجستیک عمل می‌کند. این نشان‌دهنده وجود ناهمگنی مشاهده نشده در بین داده‌ها می‌باشد.

- مدل‌های یادگیری ماشین

مدل‌های یادگیری ماشین	MAE	MAPE	ctd
RF-Survival	۱۷.۱۱۱	۴۰.۴۹۱	۰.۶۰۲
RF-Regression	۱۸.۱۷۸	۵۳.۵۵۸	
GBM-Survival	۱۶.۷۶۸	۴۰.۵۷۸	۰.۶۰۷
GBM-Regression	۱۶.۷۹۶	۴۲.۰۰۵	

همه مدل‌های یادگیری ماشین و شبکه‌های عصبی عمیق عملکرد بهتری نسبت به مدل زمان شکست شتابیده با توزیع لگ لجستیک دارند و به این معنی که در نظر گرفتن یک رابطه تعاملی و غیرخطی عملکرد پیش‌بینی را بهبود می‌بخشد. مدل GBM-Regression و DNN-Regression و مدل هدف بهتر از مدل شکنندگی RPAFT عمل می‌کنند. عملکرد هردو مدل جنگل تصادفی (RF) و مدل ماشین گرادیان تقویتی(GBM) با تغییر وظایف‌شان از پیش‌بینی میانگین زمان پاکسازی به برآورد تابع بقا، بهبود یافته است.

- مدل‌های شبکه های عصبی عمیق

مدل‌های شبکه‌های عصبی عمیق	MAE	MAPE	ctd
DNN-Survival	۱۷.۵۵۳	۴۳.۱۹۵	۰.۶۲۴
DNN-Regression	۱۶.۸۰۸	۴۱.۴۵۲	

برخلاف مدل‌های یادگیری ماشین، شبکه عصبی عمیق با هدف برآورد تابع بقا عملکرد پیش‌بینی کننده بهتری نسبت به شبکه عصبی عمیق با هدف پیش‌بینی میانگین زمان پاکسازی حادثه، ندارد.

- مدل هدف

مدل هدف	MAE	MAPE	ctd
Multi task Deep Survival Analysis Model	۱۶.۵۵۱	۳۹.۶۰۹	۰.۶۳۷

برخلاف مدل‌های دیگر که به طور جداگانه وظایف پیش‌بینی مدت زمان حادثه و براورد تابع بقا را انجام می‌دهند مدل هدف این وظایف را به طور همزمان انجام می‌دهد. مدل هدف بهترین عملکرد را هم برای پیش‌بینی میانگین زمان پاکسازی (از نظر MAE و MAPE) و هم برای برآورد تابع بقا(از نظر c-td) نشان می‌دهد. برای پیاده‌سازی مدل روی داده‌ها از نرم افزار پایتون نسخه ۸.۳ اتسفاده شده است.

قدردانی و تشکر

نویسندگان از داوران محترم به خاطر دقت نظر و صرف وقت در مطالعه مقاله و ارائه پیشنهادات برای بهبود آن، تقدیر و تشکر می‌کنند.

مراجع

Nam, D., Mannering, F. (2000). *An exploratory hazard based analysis of highway incident duration*. Transportation Research Part A: Policy and Practice, 34(2), 85-102.

Jones, B., Janssen, L., and Mannering, F. (1991). *Analysis of the frequency and duration of freeway accidents in Seattle.* Accident, Analysis and Prevention, 23(4), 239–255. https://doi.org/10.1016/0001-4575(91)90003-N

Chung, Y. (2010). *Development of an accident duration prediction model on the Korean Freeway Systems.* Accident; Analysis and Prevention, 42(1), 282–289. https://doi.org/10.1016/j.aap.2009.08.005

Tavassoli Hojati, A., Ferreira, L., Washington, S., and Charles, P. (2013). *Hazard based models for freeway traffic incident duration.* Accident; Analysis and Prevention, 52, 171–181. https://doi.org/10.1016/j.aap.2012.12.037

Wei, C., Lee, Y.(2007). *Sequential forecast of incident duration using artifical neural networks models.* Accident; Analysis and Prevention, 39(5), 944–954. https://doi.org/ 10.1016/j.aap.2006.12.017

Hamad, K., Al-Ruzouq, R., Zeiada, W., Abu Dabous, S., and Khalil, M. A. (2020). *Predicting incident duration using random forests.* Transportmetrica A: Transport Science, 16(3), 1269–1293. https://doi.org/10.1080/23249935.2020.1733132

Ma, X., Ding, C., Luan, S., Wang, Y., Wang, Y. (2017). *Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method.* IEEE Transactions on Intelligent Transportation Systems, 18(9), 2303–2310. https://doi.org/10.1109/ TITS.2016.2635719

Li, L., Sheng, X., Du, B., Wang, Y., and Ran, B. (2020). *A deep fusion model based on restricted Boltzmann machines for traffic accident duration prediction. Engineering Applications of Artificial Intelligence, 93(April), 103686.https://doi. org/10.1016/j.engappai.2020.103686*

Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y., and Huang, H. (2020).Statistical and machine-learning methods for clearance time prediction of road incidents: A methodology review. Analytic Methods in Accident Research, 27, 100123. https://doi.org/10.1016/j.amar.2020. 100123

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008).Random survival forests. The Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/ 08-AOAS169

Hothorn, T., Buhlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006).Survival ensembles. Biostatistics (Oxford, England), 7(3), 355–373. https://doi.org/10.1093/ biostatistics/kxj011

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018).DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Medical Research Methodology, 18(1), 24. https://doi.org/10.1186/s12874-018-0482-1

Lee, C., Zame, W. R., Yoon, J., and Van Der Schaar, M. (2018).DeepHit: A deep learning approach to survival analysis with competing risks. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 2314–2321. https://doi. org/10.1609/aaai.v32i1.11842

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

مدل‌بندی تصادفات جاده‌ای استان خراسان جنوبی با استفاده از مدل‌های اتورگرسیو صحیح مقدار

زهره نخعی زاده [۱] ، سارا جمهوری [۲]

[۱,۲] گروه آمار، دانشگاه بیرجند

چکیده: در علم آمار، به دنباله‌ای از مشاهدات که به طور معمول در فواصل زمانی منظم ثبت می‌شوند، ''سری زمانی'' گفته می‌شود. بسیاری از متغیرها، مانند تعداد تصادفات، قربانیان جرایم و سایر موارد مشابه، ماهیتی گسسته دارند. بنابراین، تحلیل و پیش‌بینی این نوع داده‌ها با استفاده از مدل‌های سری زمانی شمارشی، مناسب‌تر از دیگر روش‌ها است. در این مقاله، از مدل‌های سری زمانی شمارشی برای بررسی داده‌های تصادفات روزانه برون‌شهری در شهرستان‌های استان خراسان جنوبی استفاده شده است. هدف از این پژوهش، شناسایی مدل مناسبی است که بهترین برازش را به داده‌ها داشته باشد. پس از مقایسه مدل‌های مختلف، مدلی که بهترین تطابق را با داده‌های موجود نشان می‌دهد، انتخاب شده و بر اساس آن پیش‌بینی‌های لازم انجام می‌شود.

واژه‌های کلیدی: پیش‌بینی، عملگر نازک‌کننده، سری زمانی شمارشی.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62M10.

---

۱ مقدمه

تصادفات جاده‌ای یکی از مهم‌ترین و چالش‌برانگیزترین مسائل در حوزه حمل و نقل مدرن محسوب می‌شوند. این تصادفات بر اثر عوامل مختلفی از جمله افزایش جمعیت، رشد تقاضای سفر و بنابراین استفاده بیشتر از وسیله‌های نقلیه موتوری رخ می‌دهند. در نتیجه، ایمنی جاده‌ها و خیابان‌ها اهمیت ویژه‌ای پیدا می‌کند، چرا که کاهش تصادفات و جلوگیری از وقوع آن‌ها هدف اصلی در ارتقاء ایمنی حمل و نقل است.

در بسیاری از کشورها، از جمله ایران، تصادفات جاده‌ای به عنوان یکی از اصلی‌ترین دلایل مرگ‌ومیر و جراحات تشخیص داده شده است. در نتیجه توجه به این امر یکی از اندیشه‌های اصلی در برنامه‌ریزی‌های حمل و نقل و سیاست‌های ایمنی است. بر اساس آمارهای

---

[۱] سخنران، znakhaeezadeh72@gmail.com

موجود، ایران یکی از کشورهایی است که با بالاترین نرخ مرگ‌ومیر ناشی از تصادفات جاده‌ای مواجه است. تحلیل آمارهای تصادفات می‌تواند به مدیران و کارشناسان این حوزه کمک کند تا روندهای تصادفات را بهتر درک کرده و اقدامات موثری برای کاهش آن‌ها انجام دهند.

برای پیش‌بینی تعداد تصادفات در مناطق خاص و در بازه‌های زمانی مختلف، از مدل‌های پیش‌بینی و سری‌های زمانی استفاده می‌شود. در سال‌های اخیر، تکنیک‌های رگرسیونی که بر فرض استقلال داده‌ها تکیه دارند، کمتر مورد استفاده قرار می‌گیرند، چرا که محدودیت‌هایی در مدل‌سازی سری‌های زمانی پیچیده دارند. در این راستا، محققان در قالب مدل‌های سری زمانی پیوسته مقدار، همچون مدل‌های $ARMA$، $ARIMA$ و $SARIMA$، داده‌های مربوط به تصادفات را مورد مطالعه قرار داده‌اند و نتایج مفیدی در حوزه پیش‌بینی و تحلیل روندهای تصادفات ارائه داده‌اند. مطالعات متعددی در قلمرو مدل‌های سری زمانی انجام شده است که از این میان می‌توان به یوسف‌زاده چابک و همکاران (۲۰۱۶)، آگیمنگ و همکاران (۲۰۲۳)، نصیری و همکاران (۲۰۲۳) و ولوگدین و همکاران (۲۰۲۴) و از میان پژوهش‌های داخلی به توکلی و رحیم‌اف (۱۳۹۰)، عسگری و همکاران (۱۳۹۴) و امیدی و همکاران (۱۳۹۶) اشاره کرد.

به دلیل ماهیت گسسته و غیرمنفی داده‌های تصادفات، مدل‌های سنتی و پیوسته مقدار سری‌های زمانی برای تحلیل این داده‌ها مناسب نیستند. از این‌رو، توسعه مدل‌هایی که به طور خاص برای این نوع داده‌ها طراحی شده‌اند، ضروری است. یکی از ساده‌ترین و در عین حال پرکاربردترین مدل‌ها برای داده‌های شمارشی، مدل $INAR(۱)$ معرفی‌شده توسط مک‌کنزی (۱۹۸۵) است. این مدل نسخه‌ای صحیح‌مقدار از مدل خودرگرسیو مرتبه اول $AR(۱)$، است و با وجود ساختار ساده، از انعطاف‌پذیری بالایی برخوردار است. مدل $INAR(۱)$ نه تنها امکان تحلیل فرآیندهای شمارشی ایستا را فراهم می‌کند، بلکه مبنایی برای توسعه مدل‌های پیشرفته‌تر نیز محسوب می‌شود. همچنین، مفاهیم مهمی مانند برآورد پارامترها، تحلیل مانده‌ها و استنباط آماری را می‌توان به سادگی در قالب این مدل پیاده‌سازی کرد. اساس ساخت مدل‌های $INAR(۱)$ مبتنی بر عملگر نازک‌کننده[1] است. با در نظر گرفتن عملگرهای نازک کننده متفاوت، مدل‌های سری زمانی متنوعی به‌دست می‌آیند. در پژوهش حاضر از سری‌های زمانی شمارشی برای مدل‌سازی داده‌های تصادفات جاده‌ای در سطح استان خراسان جنوبی بهره گرفته شده است.

۲ مدل‌های شمارشی

در این بخش، خانواده مدل‌های سری زمانی شمارشی $INAR(۱)$ و خواص آن مورد بررسی قرار گرفته است. ابتدا مفهوم عملگر نازک‌کننده دوجمله‌ای[2] بیان می‌شود.

تعریف ۱.۲. فرض کنید X یک متغیر تصادفی گسسته و $\{Z_i, i = ۱, ۲, \ldots\}$ دنباله‌ای از متغیرهای تصادفی مستقل و هم‌توزیع با توزیع مشترک برنولی و احتمال موفقیت $\alpha \in (۰, ۱)$ باشند. در این‌صورت توزیع شرطی متغیر تصادفی $\alpha \circ X = \sum_{i=۱}^{X} Z_i$ به شرط X دوجمله‌ای $Bin(X, \alpha)$ است. عملگر $\circ$ عملگر نازک‌کننده دوجمله‌ای نامیده می‌شود.

ال اوش و الزید (۱۹۸۷) مدل $INAR(۱)$ را با استفاده از عملگر دوجمله‌ای به صورت زیر معرفی کردند.

[1] Thinning operator
[2] Binomial thinning

تعریف ۲.۲. فرآیند تصادفی شمارشی $\{X_t, t \in \mathcal{T}\}$ یک فرآیند $INAR(1)$ نامیده می‌شود هرگاه

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z}, \quad \alpha \in (0, 1), \tag{1.2}$$

که در آن جملات نوآوری[3] $\{\varepsilon_t; t \in \mathcal{T}\}$ دنباله‌ای از متغیرهای تصادفی نامنفی، صحیح مقدار، مستقل و هم‌توزیع با میانگین μ_ε و واریانس σ_ε^2 است. همچنین فرض می‌شود که متغیرهای تصادفی موجود در عملگر نازک‌کننده α و ε_t از $\{X_t; t \in \mathcal{T}\}$ مستقل هستند.

ال اوش و الزید (۱۹۸۷) نشان دادند که فرآیند$INAR(1)$، یک زنجیر مارکوف همگن با احتمال انتقال یک مرحله‌ای به صورت زیر است

$$\begin{aligned} p_{k|l} &:= P(X_t = k | X_{t-1} = l) \\ &= \sum_{j=0}^{min\{k,l\}} \binom{l}{j} \alpha^j (1-\alpha)^{l-j} P(\varepsilon_t = k - j). \end{aligned} \tag{2.2}$$

توجه ۳.۲. با انتخاب توزیع‌های مختلفی برای جملات نوآوری، مدل‌های $INAR(1)$ مختلفی حاصل می‌شود. یکی از مدل‌های پرکاربرد، حالتی است که در آن توزیع حاشیه‌ای X_t پواسون با میانگین $\mu = \frac{\lambda}{(1-\alpha)}$ است و $\{\varepsilon_t; t \in \mathcal{T}\}$ دنباله‌ای از متغیرهای تصادفی پواسون با میانگین و واریانس λ می‌باشد.

توجه ۴.۲. با وجود اینکه توزیع پواسون یکی از مهم‌ترین توزیع‌های آماری مرتبط با فرآیندهای شمارشی است، اما استفاده از آن همیشه در مدل‌های سری زمانی شمارشی مناسب نیست. زیرا در این توزیع میانگین و واریانس با یکدیگر برابرند و این خاصیت عملاً در سری‌های زمانی مشاهده نمی‌شود. به همین علت، مدل‌های دیگری مورد مطالعه قرار گرفته است. اگر در (۱.۲) جملات نوآوری از توزیع دوجمله‌ای منفی $NB(n, \pi)$ پیروی کنند، مدل $NBINAR(1)$ حاصل می‌شود.

توجه ۵.۲. در برخی مدل‌های $INAR(1)$ فرض می‌شود که ضریب نازک‌کننده متغیری تصادفی است. در اینصورت مدل به صورت $X_t = \alpha_\phi \circ X_{t-1} + \varepsilon_t$ نوشته می‌شود. جو (۱۹۹۶) با در نظر گرفتن $\phi = \frac{1}{(n+1)}$ و توزیع بتا $Beta(\frac{1-\phi}{\phi}\alpha, \frac{1-\phi}{\phi}(1-\alpha))$ برای α_ϕ وقتی جملات نوآوری دارای توزیع $NB(n(1-\alpha), \pi)$ باشند، مدل $RCNBINAR(1)$ را معرفی کرد. در این مدل، توزیع حاشیه‌ای X_t دوجمله‌ای منفی $NB(n, \pi)$ خواهد بود.

۱.۲ برآورد پارامترهای مدل

روش‌های متعددی مانند روش گشتاورها، برآورد حداقل مربعات شرطی و برآورد ماکزیمم درستنمایی برای تخمین پارامترهای مدل وجود دارد. برای انجام برآورد پارامترها بر پایه روش درستنمایی، می‌توان از احتمال‌های شرطی و خاصیت مارکوفی سری زمانی $INAR(1)$ بهره برد. تابع لگاریتم درستنمایی در مدل $INAR(1)$ با بردار پارامتر $\boldsymbol{\theta}$ عبارت است از

$$\ell(\boldsymbol{\theta}) = \log p_{x_1}(\boldsymbol{\theta}) + \sum_{t=2}^{T} \log p_{x_t|x_{t-1}}(\boldsymbol{\theta}), \tag{3.2}$$

که در آن احتمالات انتقال با توجه به رابطه (۲.۲) محاسبه می‌شوند. پارامتر $\boldsymbol{\theta}$ در مدل‌های مختلف $INAR(1)$ متفاوت است. به عنوان مثال در مدل پواسون $INAR(1)$ که در ملاحظه ۳.۲ معرفی شد $\boldsymbol{\theta} = (\alpha, \lambda)$ و در مدل $NBINAR(1)$ معرفی شده در ملاحظه ۴.۲، $\boldsymbol{\theta} = (\alpha, n, \pi)$ است. در این صورت برآوردگر ماکسیمم درستنمایی پارامتر $\boldsymbol{\theta}$ به صورت $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ell(\boldsymbol{\theta})$ می‌باشد.

[3] Innovations

۲.۲ تشخیص و کفایت مدل

در این بخش، ابزارهای مورد نیاز برای تشخیص مناسب‌ترین مدل برای برازش به داده‌ها معرفی می‌شوند. برای تشخیص ساختار وابستگی سری، نمودار خودهمبستگی جزئی نمونه‌ای $(SPACF)$رسم می‌شود. اگر $\hat{\rho}_{par}(k)$ در تاخیر اول مخالف صفر و در بقیه تاخیرها به سمت صفر میل کند، مدل $INAR(۱)$ برای برازش به داده‌ها مناسب خواهد بود. برای تعیین مدل $INAR(۱)$ مناسب برای داده‌ها، می‌توان از شاخص $\hat{I} = \frac{S^۲}{\bar{X}}$ استفاده کرد که در آن $\bar{X} = \frac{۱}{T}\sum_{t=۱}^{T} X_t$ و $S^۲ = \frac{۱}{T}\sum_{t=۱}^{T}(X_t - \bar{X})^۲$. اگر مدل پواسون برای برازش به داده‌ها مناسب باشد، در اینصورت باید مقدار این آماره تقریبا برابر یک باشد. همچنین این آماره به طور مجانبی از توزیع نرمال با

$$E(\hat{I}) = ۱ - \frac{۱}{T}\frac{۱+\alpha}{۱-\alpha}, \quad V[\hat{I}] \approx \frac{۲}{T}\frac{۱+\alpha^۲}{۱-\alpha^۲},$$

پیروی می‌کند (شوئر و ویب ، ۲۰۱۴). اگر چندین مدل برای برازش به داده‌ها مناسب باشند، از معیارهایی مانند آکائیک و معیار اطلاع بیزی که به صورت زیر هستند

$$AIC = -۲\ell(\boldsymbol{\theta}) + ۲m, \quad BIC = -۲\ell(\boldsymbol{\theta}) + ۲\log T,$$

برای شناسایی مدل مناسب‌تر استفاده می‌شود (m تعداد پارامترهای مجهول مدل و T تعداد مشاهدات است). بعد از شناسایی بهترین مدل، لازم است که کفایت مدل بررسی شود. محاسبه باقی‌مانده‌های استاندارد شده پیرسون اولین رویکرد در کفایت مدل است. این مقادیر طبق رابطه زیر حاصل می‌شود.

$$e_t = \frac{x_t - E(X_t|x_{t-۱})}{\sqrt{V(X_t|x_{t-۱})}}, \quad t = ۲, ۳, \ldots, T.$$

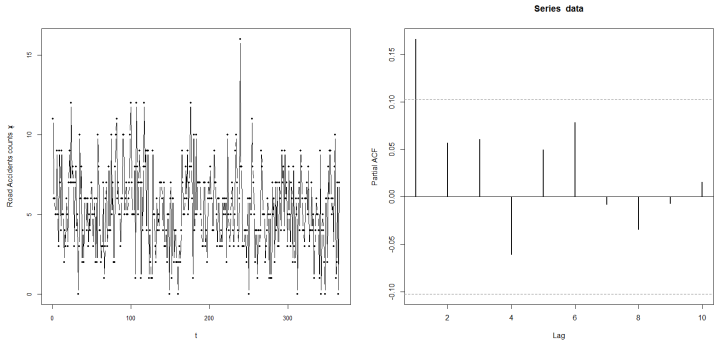
برای یک مدل مناسب انتظار می‌رود که این باقی‌مانده‌ها دارای میانگین و واریانس به ترتیب برابر صفر و یک بوده و با یکدیگر همبستگی نداشته باشد. واریانس بزرگتر (کوچکتر) از یک نشان دهنده وجود بیش‌پراکندگی (کم‌پراکندگی) نسبت به مدل در نظر گرفته شده است (هاروی و فرناندز ، ۱۹۸۹).

۳ پیش‌بینی تصادفات جاده‌ای

در این پژوهش، ابتدا آمار و ارقام مربوط به تصادفات جاده‌ای [۴] در سه سال متوالی ۱۴۰۱، ۱۴۰۲ و ۱۴۰۳ گزارش شده است. این گزارش نشان‌دهنده وضعیت تصادفات، میزان تلفات، محل وقوع و زمان‌بندی آن‌ها است که به منظور شناسایی روندها و ارائه راهکارهای بهبود امنیت جاده‌ای اهمیت فراوان دارد.

در سال ۱۴۰۱، تعداد کل تصادفات ثبت‌شده به ۲۰۵۲ مورد رسید که از این تعداد، ۲۵۸ مورد مرگ و میر ناشی از تصادف گزارش شده است. همچنین، حدود ۳۴ درصد از این تصادفات در جاده‌های روستایی، ۶۰ درصد در جاده‌های اصلی و ۶ درصد در جاده‌های فرعی رخ داده است. در این سال، ۶۴ درصد تصادفات در طول روز و ۳۶ درصد در شب اتفاق افتاده است، که نشان‌دهنده اهمیت زمان‌بندی و دقت در رانندگی در ساعات مختلف شبانه‌روز است. در سال ۱۴۰۲، تعداد تصادفات کمی کاهش یافته و به ۱۹۸۶ مورد رسید. در این سال، تعداد ۲۴۸ فوتی به ثبت رسیده که نسبت به سال قبل کمی کاهش یافته است. همچنین ۵۷ درصد تصادفات در جاده‌های اصلی، ۳۶ درصد در جاده‌های روستایی و ۷ درصد در جاده‌های فرعی رخ داده است. الگوی زمانی همچنان برتری تصادفات

[۴] داده‌ها از پلیس راه استان خراسان جنوبی اخذ شده است.



شکل ۱: تعداد تصادفات روزانه جاده‌ای استان خراسان جنوبی و نمودار ACF داده‌ها

در روز را نشان می‌دهد (۶۳ درصد در روز و ۳۷ درصد در شب). در سال ۱۴۰۳، تعداد تصادفات دوباره کمی کاهش یافته و به ۱۹۵۳ مورد رسیده است. تعداد فوتی‌ها نیز به ۲۰۰ تن کاهش یافته است، که نشان‌دهنده نسبتی پایدار است. در این سال، توزیع جغرافیایی تصادفات نسبت به سال‌های قبل تقریباً ثابت است (۵۷ درصد در جاده‌های اصلی، ۳۵ درصد در جاده‌های روستایی و ۸ درصد در جاده‌های فرعی). همچنین، الگوهای زمانی همچنان بر تصادفات روز تأکید دارند (۶۴ درصد در طول روز و ۳۶ درصد در شب).

بر اساس این آمار، می‌توان نتیجه گرفت که تعداد تصادفات جاده‌ای در طی سه سال مورد بررسی کاهش نسبی یافته است؛ اما با توجه به میزان تلفات، نیاز به تمرکز بیشتر و اتخاذ اقدامات مؤثر در حوزه ایمنی جاده‌ها و آموزش رانندگان احساس می‌شود. همچنین، سهم قابل توجه تصادفات در جاده‌های روستایی و در ساعات روز نشان می‌دهد که برنامه‌ریزی برای ارتقاء زیرساخت‌ها و اقدامات مراقبتی در این فضاها اهمیت دارد. به طور کلی، تمرکز بر کاهش تصادفات در جاده‌های روستایی و کنترل سرعت و رعایت قوانین در زمان‌های مختلف شبانه‌روز می‌تواند نقش مؤثری در بهبود وضعیت ایمنی جاده‌ای کشور ایفا کند. در این بخش مجموعه داده های تصادفات روزانه جاده ای در استان خراسان جنوبی در سال ۱۴۰۳ مورد بررسی قرار می‌گیرد ($T = ۳۶۵$). کدنویسی‌ها به کمک نرم افزار R انجام شده است. شکل (a) ۱، نمودار سری زمانی تعداد تصادفات را نشان می‌دهد. مشاهدات در فاصله صفر تا ۱۶ تصادف در روز قرار دارند. روند یا تغییرات فصلی قابل توجهی در داده‌ها مشاهده نمی‌شود. نوسانات بالا و پایین موجود در این نمودار بیان‌کننده سطح متوسطی از همبستگی است. این موضوع از نمودار ACF نمونه‌ای نیز مشهود است. با توجه به مقدار $\hat{\rho}_{par}(۱) = ۰/۱۶۶۲$ ملاحظه می‌شود که این مقدار به‌طور قابل توجهی از صفر بزرگتر است. بنابراین، مدلی شبیه به $AR(۱)$ برای توصیف مشاهدات مناسب است. به وضوح، مقدار میانگین $\bar{x} \approx ۵/۳۲۰۵$ کمتر از واریانس $s^۲ \approx ۶/۶۵۷۹$ است. لذا، در داده‌ها پراکندگی زیادی وجود دارد. از طرف دیگر، مقدار احتمال برآورد شده در نقطه صفر برابر $p_۰ \approx ۰/۰۲۱۹۱$ است که بزرگتر از مقدار متناظر آن از توزیع پواسون، یعنی $exp(-\bar{x}) \approx ۰/۰۰۴۸$ می‌باشد. بنابراین، برازش یک مدل $INAR(۱)$ به داده‌ها، قابل قبول به نظر می‌رسد. برای انجام آزمون تشخیص بیش‌پراکندگی در مدل $INAR(۱)$، فرضیه صفر اینکه داده‌ها از مدل $INAR(۱)$ با نواوری‌های پواسون [۵] پیروی می‌کنند، آزمون می‌شود. تحت فرضیه صفر، $E(\hat{I}) = ۰/۹۹۶۱$ و $V(\hat{I}) = ۰/۰۷۰۰$ است. با توجه به مقدار $p - value = ۰/۰۰۰۴$ مشخص می‌شود که پراکندگی بیش از حد و احتمالا قابل توجه است. بنابراین مدل $RCNBINAR(۱)$ برای برازش به داده‌ها مناسب‌تر است. این مدل با مدل پواسون $INAR(۱)$ از نظر معیارهای AIC و BIC مقایسه می‌شود. برآورد میانگین و شاخص پراکندگی عوامل نوآوری‌ها به ترتیب $\bar{\mu}_\varepsilon = \bar{x}(۱ - \hat{\alpha}) \approx ۴/۴۳۵۸$ و $\hat{I} = \hat{I}(۱ + \hat{\alpha}) - \hat{\alpha} \approx ۱/۲۸۹۱$ است. از بررسی مقادیر AIC و BIC موجود در جدول

[۵] $PoissonINAR(۱)$

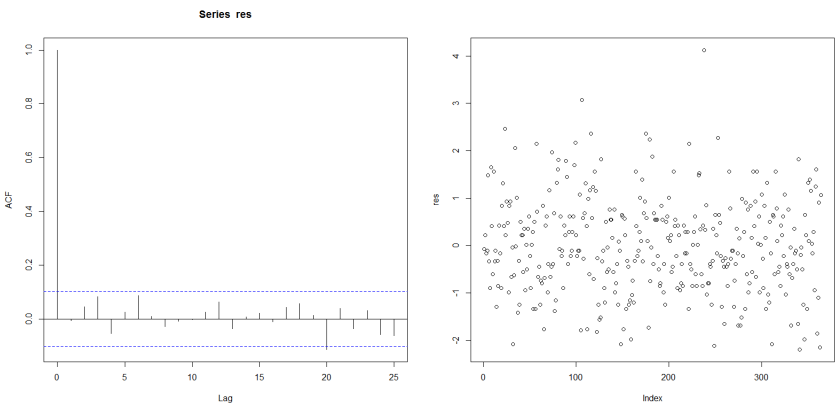
جدول ۱: آمار توصیفی داده‌های تصادفات در شهرهای استان خراسان جنوبی در سال ۱۴۰۳

متغیر	میانگین	میانه	انحراف معیار	مینیمم	ماکزیمم	چولگی	کشیدگی
تصادفات سال ۱۴۰۳	۵٫۳۲۱	۵	۲٫۵۸۰	۰	۱۶	۰٫۳۴۹	۳٫۳۱۹

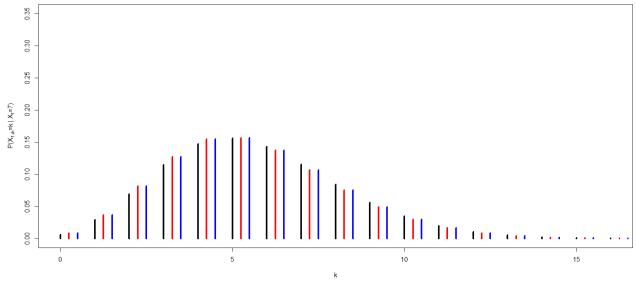
جدول ۲: برآوردهای درستنمایی ماکسیمم پارامترها و AIC و BIC در مدل‌های مختلف

$Model$	Parameter			AIC	BIC
	۱	۲	۳		
$i.i.d.Poisson$	۵٫۳۲۰			۱۷۳۵٫۱۰۷	۱۷۳۹٫۰۰۷
μ	۰٫۱۲۰				
$Poisson.INAR(۱)$	۴٫۶۰۴	۰٫۱۳۵		۱۷۲۶٫۹۱۶	۱۷۳۴٫۷۱۶
(λ, α)	(۰٫۲۴۵)	(۰٫۰۴۱)			
$i.i.d.NB$	۱۹٫۷۲۶	۰٫۷۸۵		۱۷۲۶٫۸۵۱	۱۷۳۴٫۶۵۱
(n, π)	(۷٫۲۴۵)	(۰٫۰۶۱)			
$NBINAR(۱)$	۱۵٫۳۴۰	۰٫۷۷۳	۰٫۱۵۵	۱۷۱۹٫۳۲۹	۱۷۳۱٫۰۲۸
(n, π, α)	(۶٫۳۷۸)	(۰٫۰۶۹)	(۰٫۰۴۹)		
$RCNBINAR(۱)$	۲۱٫۴۵۹	۰٫۰۸۰۱	۰٫۱۵۸	۱۷۱۹٫۲۳۰	۱۷۳۱٫۰۳۰
(n, π, α)	(۹٫۳۱۶)	(۰٫۸۰۱)	(۰٫۱۵۸)		

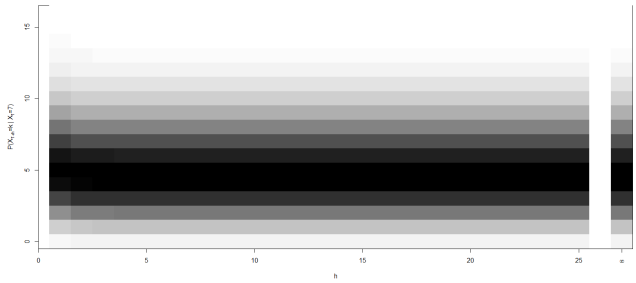
۲ مشخص می‌شود که مدل‌های $INAR(۱)$ بهتر از مدل $i.i.d$ عمل می‌کنند. مقدار برآورد شده α به طور قابل توجهی از صفر بزرگتر است. از مقایسه مدل‌های $INAR(۱)$ برحسب معیارهای AIC و BIC، ملاحظه می‌شود که مدل $RCNBINAR(۱)$ برای برازش به داده‌ها مناسب‌تر است. برای بررسی کفایت این مدل، نمودارهای پراکنش و ACF نمونه‌ای باقیمانده‌های پیرسون در شکل ۲ رسم شده‌اند. با توجه به این شکل، هیچگونه همبستگی معنی‌داری بین باقیمانده‌ها مشاهده نمی‌شود. برای این مدل، توزیع‌های پیش‌بینی h-گام روبه جلو به شرط مقدار آخرین مشاهده ($x_{۳۶۵} = ۷$)، با استفاده از رابطه (۲.۲) به دست می‌آیند. نمودار این توزیع‌ها در شکل ۳ و ۴ رسم شده است. با توجه به شکل ۳ پیش‌بینی مبتنی بر میانه توزیع به ازای سه مقدار $h = ۱$ (مشکی)، $h = ۵$ (قرمز) و $h = ۱۰$ (آبی) در نقطه $k = ۵$ اتفاق می‌افتد. شکل ۴ نیز مقدار احتمال بیشتر را روی مقادیر ۴، ۵ و ۶ نشان می‌دهد. لذا انتظار می‌رود که در روزهای آتی، به طور متوسط ۵ تصادف در روز اتفاق بیافتد.



شکل ۲: نمودارهای پراکنش و ACF نمونه‌ای باقیمانده‌های پیرسون



شکل ۳: هیستوگرام توزیع پیش‌بینی h گام به جلو برای $h = ۱, ۵, ۱۰$



شکل ۴: توزیع پیش‌بینی h گام به جلو برای $h = ۱, \ldots, ۲۵, \infty$

بحث و نتیجه‌گیری

در این مقاله، با استفاده از مدل‌های سری زمانی شمارشی، تعداد تصادفات رانندگی جاده‌ای در سطح استان خراسان جنوبی، مورد مطالعه قرار گرفت. پس از تحلیل، مشخص گردید که مدل $RCNBINAR(۱)$ برای برازش به داده‌ها، مناسب‌تر از سایر مدل‌های رقیب است.

مراجع

امیدی، نبی.، خلیلی، کرم.، امیدی، محمدرضا. و جعفری اسکندری، میثم. (۱۳۹۶). ارائه مدل اقتصادسنجی پیش بینی کشته شدگان و مصدومان تصادفات جاده ای، مطالعات مدیریت ترافیک، شماره ۴۴، ۲۱-۴۴.

توکلی شیرازی، نیما و رحیم اف، کامران. (۱۳۹۰). مدلسازی سری زمانی مجروحین و متوفیات حوادث تصادفات جاده ای ایران با رویکردهای MA، AR، $ARMA$، دومین کنفرانس ملی تصادفات جاده ای، سوانح ریلی و هوایی، زنجان.

عسگری، حشمت اله.، امیدی، محمد رضا.، امیدی، نبی. و مفتاحی، هادی. (۱۳۹۴). پیش بینی بروز تخلفات از نوع سرعت غیرمجاز در آزادراه‌ها، مطالعات مدیریت ترافیک. شماره ۳۸، ۲۵-۳۸.

Agyemang, E. F., Mensah, J. A., Ocran, E., Opoku, E. Nortey, E. N. N. (2023). Time series based road traffic

accidents forecasting via SARIMA and Facebook Prophet model with potential changepoints. *Heliyon*, 9, e22544.

Al-Osh, M. A., Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8(3), 261-275.

Freeland, R. K. (1998). Statistical analysis of discrete time series with application to the analysis of workers' compensation claims data (Doctoral dissertation, University of British Columbia).

Harvey, A.C., Fernandes, C. (1989). Time series models for count or qualitative observations. Journal of Business &*Economic Statistics*, 7(4), 407–417.

Joe, H. (1996). Time series models with univariate margins in the convolution-closed infinitely divisible class. *Journal of Applied Probability*, 33(3), 664–677.

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin*, 21(4), 645–650.

Nasiri, H., Mohammadpour, S. I., Dahaghin, M. (2023). Forecasting time trend of road traffic crashes in Iran using the macro-scale trafiic flow characteristics. *Heliyon*, 9(3), e14481.

Schweer, S., Weiß, C.H. (2014). Compound Poisson INAR(1) processes: stochastic properties and testing for overdispersion. Computational Statistics and Data Analysis, 77, 267–284.

Vologdin, S., Kasatkina, E., Kasatkin, A. (2024). Analyzing and forecasting road traffic accidents and their consequences: a case study of the Udmurt republic. *E3S Web of Conferences*, 471, 06006.

Yousefzadeh-Chabok, S., Ranjbar-Taklimie, F., Malekpouri, R., Razzaghi, A. (2016). A Time Series Model for Assessing the Trend and Forecasting the Road Traffic Accident Mortality. *Arch Trauma Res*, 5(3), e36570.

Modeling Road Traffic Accidents in South Khorasan Province Using Integer-Valued Autoregressive Models

Nakhaeezadeh, Z.[1], **Jomhoori, S.**[2]

[1,2]Department of Statistics, University of Birjand, Iran.

Abstract: In statistics, a sequence of observations that are typically recorded at regular time intervals is referred to as a ”time series”. Many variables, such as the number of accidents, crime victims, and other similar events, exhibit a discrete nature. Therefore, the analysis and forecasting of these types of data are more appropriately conducted using count time series models. In this article, count time series models are applied to analyze daily intercity accident data from the counties of South Khorasan province. These data span a three-year period, from 2021 to 2024. The objective of this research is to identify the most suitable model that provides the best fit to the data. After comparing various models, the one that demonstrates the best match with the existing data is selected, and based on it, the necessary forecasts are made.

Keywords: Counting timeseries , Forecasting, Thinning operator.

Mathematics Subject Classification (2020): 62M10.

پانزدهمین سمینار احتمال و فرآیندهای تصادفی
۸ و ۹ شهریور ۱۴۰۴
دانشگاه کردستان
15th Seminar On Probability and Stochastic Processes
August 30-31, 2025
University of Kurdistan
انجمن آمار ایران

فرایند چوب‌شکنی رگرسیون-بتا با کوواریانس نامانا وابسته به متغیر کمکی

اسماعیل یارعلی [۱] ، فیروزه ریواز [۲] ، مجید جعفری خالدی [۳]

[۱] دانش‌آموخته دکتری آمار، دانشگاه شهید بهشتی، تهران

[۲] گروه آمار، دانشگاه شهید بهشتی، تهران

[۳] گروه آمار، دانشگاه تربیت مدرس، تهران

چکیده: در این مقاله رویکردی جدیدی برای ساخت فرایند چوب‌شکنی[۱] مورد مطالعه قرار می‌گیرد به‌طوری‌که توزیع تصادفی وابسته به متغیر کمکی است. این شیوه مدل‌بندی علاوه برآن‌که نامانایی متاثر از متغیر کمکی را در ساختار وابستگی داده‌ها را پوشش می‌هد، امکان افرازبندی ناحیه مورد مطالعه را نیز فراهم می کند. با در نظر گرفتن این فرایند به عنوان توزیع اثرهای تصادفی در یک چارچوب مدل خطی آمیخته، داده ها مدل شده و استنباط بیزی آن ارائه می شود. سپس با استفاده از یک مجموعه داده‌ی واقعی مربوط به میزان بارش عملکرد مدل معرفی شده در مقایسه با مدل‌های رقیب مورد ارزیابی قرار می‌گیرد.

واژه‌های کلیدی: فرایند چوب‌شکنی، رگرسیون-بتا، نامانایی، متغیر کمکی، ساختار همبستگی.

کد موضوع‌بندی ریاضی (۲۰۲۰): 62H11، 62M30، 62F15.

---

۱ مقدمه

یکی از مؤلفه‌های اصلی تحلیل داده‌های فضایی، تعیین وابستگی فضایی است. این کار عموماً از طریق تابع کوواریانس که متعلق به رده‌ای پارامتری از مدل‌های مانا است، انجام می‌شود. این درحالی است که در بسیاری از کاربردها، فرض مانایی واقع‌بینانه نیست. در واقع، در بسیاری از کاربردها عوامل موضعی مانند توپوگرافیکی منطقه مورد مطالعه بر همبستگی فضایی تاثیرگذار هستند و باعث ایجاد نامانایی در ساختار وابستگی فضایی می‌شوند. لذا ساخت مدل‌های کوواریانس نامانای متأثر از متغیرهای کمکی موضعی بسیار حائز اهمیت است. تاکنون رهیافت‌های مختلفی برای الحاق متغیر کمکی در ساختار مرتبه دوم فرایندهای فضایی ارائه شده است.

---

[۱] سخنران، e_yarali@sbu.ac.ir

[1] Stick-breaking proces

از جمله می‌توان به روش هموارسازی هسته (ریچ و همکاران ، ۲۰۱۱)، روش پیچش فرایند (رایزر و کالدر ، ۲۰۱۵)، روش دگردیسی (اشمیت و همکاران ، ۲۰۱۱)، روش بسط توابع پایه (یارعلی و ریواز ، ۲۰۲۰) و استفاده از معادلات مشتقات جزئی تصادفی (اینگریبستن و همکاران ، ۲۰۱۴) اشاره کرد. گرچه الحاق متغیر تصادفی در ساختار وابستگی باعث بهبود در مدل‌های فضایی هم به لحاظ کاهش بعد تعداد پارامتر و به دنبال آن محاسبات ساده و هم تفسیر راحت‌تر نامانایی شده است، اما در تمامی این روش‌ها برای الحاق متغیر کمکی در ساختار وابستگی فضایی فرض شده است میدان تصادفی مورد مطالعه گاوسی است، در حالی‌که چنین فرضی در عمل به سختی برقرار بوده و داده‌ها شواهدی از تخطی مفروضات توزیع گاوسی را از خود نشان می‌دهند. برای رفع این محدودیت‌ها، در سال‌های اخیر مدل‌های فضایی ناگاوسی متعددی از جمله مدل‌های چوله گاوسی، رهیافت پیچش فرایند گاوسی–لگ گاوسی و فرایندهای گاوسی تبدیل یافته g و h توکی معرفی شده است. علی‌رغم ویژگی‌های مناسب این مدل‌ها، ساختار همبستگی القاء شده توسط آنها مبتنی بر فرض محدودکننده مانایی و همسانگردی میدان تصادفی مورد مطالعه است. درحالی‌که در برخی از کاربردها ارتباط ساختار فضایی با متغیر / متغیرهای کمکی موضعی باعث ایجاد نامانیی در ساختار وابستگی داده ها می شود.

یک رویکرد مناسب استفاده از رهیافت بیزی ناپارامتری است که در آن توزیع داده‌ها نامعلوم و تصادفی در نظر گرفته می شود و برای آن یک پیشینی مناسب انتخاب می شود.در این رویکرد، معمولاً از پیشینی فرایند چوب شکنی استفاده می شود از جمله می‌توان به فرایندهای چوب‌شکنی هسته (ریچ و فوئنتیس ، ۲۰۰۷؛ برزگر و ریواز ، ۲۰۲۰؛ دحدوح و خالدی ، ۲۰۲۰)، فرایندهای چوب‌شکنی پروبیت (رودریگوئز و دانسون ، ۲۰۱۱)، فرایندهای چوب‌شکنی لوجیت (رن و همکارن ، ۲۰۱۱)، فرایندهای چوب‌شکنی تعمیم‌یافته (بارسلا و همکاران ، ۲۰۱۸) و فرایندهای چوب‌شکنی پنهان (رودریگوئز و همکاران ، ۲۰۱۰) اشاره کرد. علی‌رغم ایجاد انعطاف بالا در مدل‌های معرفی شده، ساختار همبستگی القا شده از این مدل‌ها بسیار پیچیده و وابستگی فضایی وابسته به متغیر کمکی نیست. برای رفع این مشکل یارعلی و همکاران (۲۰۲۲) با استفاده از رهیافت بیزی ناپارامتری، رده‌ی جدیدی از فرایند چوب‌شکنی را معرفی کردند به‌طوری‌که نامانایی از طریق متغیر کمکی در همبستگی فضایی القا می‌شود. این شیوه مدل‌بندی علاوه برآن‌که رفتارهای توزیعی شامل عدم تقارن و چند مدی بودن را مدل‌بندی می‌کند، نامانایی متاثر از متغیر کمکی را در ساختار کوواریانس پوشش می‌دهد. همچنین افرازبندی القا شده توسط این پیشینی پیشنهادی وابسته به متغیر کمکی بوده که امکان افرازبندی ناحیه ی مورد مطالعه را بر اساس اطلاعات متغیر کمکی فراهم می کند براین اساس در بخش ۲ فرایند چوب‌شکنی رگرسیون–بتا که توسط یارعلی و همکاران (۲۰۲۲) معرفی شد، بیان می‌شود. در ادامه، با بکارگیری یک مدل بیزی مبتنی بر فرایند چوب‌شکنی رگرسیون–بتا، استنباط‌های پسینی و پیشگویی فضایی ارائه می‌شوند. همچنین با استفاده از یک مجموعه داده‌ی واقعی عملکرد مدل معرفی شده با مدل‌های رقیب از نظر پیشگویی فضایی و افرازبندی ناحیه مورد مطالعه مورد ارزیابی قرار می‌گیرد.

۲ فرایند چوب‌شکنی رگرسیون–بتا

در این بخش با استفاده از رهیافت بیزی ناپارامتری، فرایند فضایی چوب‌شکنی رگرسیون–بتا که توسط یارعلی و همکاران (۲۰۲۲) بیان شد، معرفی می‌شود. بر این اساس فرض کنید $Y(\cdot) = \{Y(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{D} \subset \mathcal{R}^d, d \neq ۱\}$ یک میدان تصادفی باشد که روی ناحیه $\mathcal{D}$ تعریف شده است. به‌علاوه فرض کنید $Y(\boldsymbol{s})$ مقدار مشاهد شده در موقعیت $\boldsymbol{s}$ است که از مدل

$$Y(\boldsymbol{s}) = \boldsymbol{w}^{'}(\boldsymbol{s})\boldsymbol{\beta} + \eta_{\boldsymbol{x}}(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \tag{۱.۲}$$

پیروی می‌کند که در آن $\boldsymbol{w}(\boldsymbol{s})$ بردار متغیرهای کمکی در مکان $\boldsymbol{s}$ و $\boldsymbol{\beta} \in \mathcal{R}^p$ بردار ضرایب رگرسیونی نامعلوم است. همچنین $\epsilon(\boldsymbol{s})$ نشان‌دهنده‌ی خطای اندازه‌گیری است که فرض می‌شود دارای توزیع نرمال با میانگین صفر و واریانس τ^2 است. اکنون برای الحاق متغیر کمکی در ساختار همبستگی فضایی، فرض می‌شود فرایند فضایی $\eta_{\boldsymbol{x}}(\boldsymbol{s})$ به صورت

$$\eta_{\boldsymbol{x}}(\boldsymbol{s}) = G^{-1}(V_{\boldsymbol{x}}(\boldsymbol{s})), \tag{۲.۲}$$

است که در آن $G(\cdot)$ اندازه احتمال تصادفی گسسته است که به صورت $G(\cdot) = \sum_{l=1}^{L} p_l \delta_{\theta_l}(\cdot)$ تعریف می‌شود که در آن $\delta_{\theta_l}(\cdot)$ تابع دیراک[2]، $G^{-1}(\cdot)$ وارون تعمیم‌یافته تابع $G(\cdot)$، و اتم‌های θ_1، $\cdots$، θ_L تحققهایی مستقل از توزیع نرمال $N(0, \sigma^2)$ هستند. همچنین $V_{\boldsymbol{x}}(\cdot) = \{V_{\boldsymbol{x}}(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}^d\}$ یک فرایند فضایی رگرسیون–بتا است که به صورت زیر تعریف می‌شود.

تعریف ۱.۲. فرض کنید $\{Z(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}^d\}$ یک فرایند گاوسی مانا با میانگین صفر، واریانس یک و تابع همبستگی $\rho(\boldsymbol{s}_i, \boldsymbol{s}_j; \lambda)$ باشد. اگر $V_{\boldsymbol{x}}(\boldsymbol{s})$ به شرط $Z(\boldsymbol{s})$، یک فرایند مستقل و دارای توزیع $Beta(\mu_x(\boldsymbol{s})\psi, \psi(1-\mu_x(\boldsymbol{s})))$ با تابع چگالی

$$f(v_x) = \frac{\Gamma[\psi]}{\Gamma[\mu_x(\boldsymbol{s})\psi]\Gamma[\psi(1-\mu_x(\boldsymbol{s}))]} v_x^{\mu_x(\boldsymbol{s})\psi-1}(1-v_x)^{\psi(1-\mu_x(\boldsymbol{s}))-1}, \quad 0 < \mu_x(\boldsymbol{s}) < 1, \qquad \psi > 0, \tag{۳.۲}$$

باشد که در آن $\mu_{\boldsymbol{x}}(\boldsymbol{s}) = \mathbb{E}[V_{\boldsymbol{x}}(\boldsymbol{s})|Z(\boldsymbol{s})]$ به صورت

$$\mu_x(\boldsymbol{s}) = \frac{\exp\left(\boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\gamma} + Z(\boldsymbol{s})\right)}{1 + \exp\left(\boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\gamma} + Z(\boldsymbol{s})\right)}, \tag{۴.۲}$$

تعریف می‌شود به طوری‌که $\boldsymbol{x}(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \cdots, x_q(\boldsymbol{s}))'$ برداری از متغیرهای کمکی در مکان $\boldsymbol{s}$، $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_q)'$ برداری از ضرایب نامعلوم که اثر متغیرهای کمکی را در همبستگی فضایی کنترل می‌کند، آن‌گاه $V_{\boldsymbol{x}}(\cdot) = \{V_{\boldsymbol{x}}(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{R}^d\}$ را یک فرایند فضایی رگرسیون–بتا گویند.

حال، برای هر مجموعه متناهی از موقعیت‌های $\boldsymbol{s}_1$، $\cdots$، $\boldsymbol{s}_n$ روی ناحیه $\mathcal{D}$، احتمال شرطی

$$P_{\boldsymbol{x}}\left(\eta_{\boldsymbol{x}}(\boldsymbol{s}_1) = \theta_{l_1}, \cdots, \eta_{\boldsymbol{x}}(\boldsymbol{s}_n) = \theta_{l_n} | \{v_l\}_{l=1}^{L}, \{\theta_l\}_{l=1}^{L}\right)$$

به صورت

$$\begin{aligned} &P_{\boldsymbol{x}}\left(\eta_{\boldsymbol{x}}(\boldsymbol{s}_1) = \theta_{l_1}, \cdots, \eta_{\boldsymbol{x}}(\boldsymbol{s}_n) = \theta_{l_n} | \{v_l\}_{l=1}^{L}, \{\theta_l\}_{l=1}^{L}\right) \\ &\quad = P\left(\sum_{j=1}^{l_1-1} p_j \le V_{\boldsymbol{x}}(\boldsymbol{s}_1) < \sum_{j=1}^{l_1} p_j; \cdots; \sum_{j=1}^{l_n-1} p_j \le V_{\boldsymbol{x}}(\boldsymbol{s}_n) < \sum_{j=1}^{l_n} p_j\right), \end{aligned} \tag{۵.۲}$$

تعریف می‌شود به طوری‌که l_1،$\cdots$،l_n یکی از مقادیر $\{1, \cdots, L\}$ را می‌گیرد. در این صورت توزیع پیشینی القاشده توسط مدل سلسله‌مراتبی بالا تحت اندازه احتمال $P_{\boldsymbol{x}}$ را یک فرایند چوب‌شکنی رگرسیون–بتا نامیده و با نماد $P_{\boldsymbol{x}} \sim BRSBP(\alpha, \psi, \gamma, \lambda, \sigma^2)$ نمایش می‌دهیم.

یک رهیافت دیگر برای ساخت فرایند چوب‌شکنی رگرسیون–بتا استفاده از فرایندهای برچسب‌گذاری است. برای این منظور، فرض کنید $\xi_{\boldsymbol{x}}(\cdot) = \{\xi_{\boldsymbol{x}}(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{D}\}$ یک فرایند برچسب‌گذاری وابسته به متغیرهای کمکی است که مقدار l را می‌گیرد اگر

$$\sum_{j=1}^{l-1} p_j \le V_{\boldsymbol{x}}(\boldsymbol{s}) < \sum_{j=1}^{l} p_j. \tag{۶.۲}$$

[2] Dirac function

حال برای هر مجموعه متناهی s_1، $\cdots$، s_n وزن‌های توأم وابسته فضایی برای فرایند برچسب‌گذاری به صورت

$$Q_{\boldsymbol{x}(\boldsymbol{s}_1),\cdots,\boldsymbol{x}(\boldsymbol{s}_n)}(l_1,\cdots,l_n) = P\Big(\xi_{\boldsymbol{x}}(\boldsymbol{s}_1)=l_1,\cdots,\xi_{\boldsymbol{x}}(\boldsymbol{s}_n)=l_n|\{v_l\}_{l=1}^{L}\Big)$$
$$= P\Big(\sum_{j=1}^{l_1-1}p_j \le V_{\boldsymbol{x}}(\boldsymbol{s}_1) < \sum_{j=1}^{l_1}p_j;\cdots;\sum_{j=1}^{l_n-1}p_j \le V_{\boldsymbol{x}}(\boldsymbol{s}_n) < \sum_{j=1}^{l_n}p_j\Big), \quad (۷.۲)$$

تعریف می‌شود. در این صورت به شرط فرایند برچسب‌گذاری، اثر تصادفی فضایی (۲.۲) را می‌توان به صورت $\eta_{\boldsymbol{x}}(\boldsymbol{s}) = \theta_{\xi_{\boldsymbol{x}}(\boldsymbol{s})}$ نوشت. بنابراین، با حاشیه‌سازی روی $\xi_{\boldsymbol{x}}$ داریم $\eta_{\boldsymbol{x}}|P_{\boldsymbol{x}} \sim P_{\boldsymbol{x}}$ و $P_{\boldsymbol{x}} \sim BRSBP(\alpha,\psi,\gamma,\lambda,\sigma^2)$. از این‌رو با استفاده از فرایند برچسب‌گذاری (۷.۲) می‌توان یک تعریف معادل و یکسان برای $BRSBP$ ارائه کرد.

۱.۲ ویژگی‌های فرایند چوب‌شکنی رگرسیون-بتا

- فرایند چوب‌شکنی رگرسیون-بتا با توزیع متناهی بعد (۵.۲) در اصول سازگاری کولموگوروف صدق می‌کند.
- فرایند چوب‌شکنی رگرسیون-بتا رده‌ی بزرگتری از فرایندهای چوب‌شکنی است که در حالت خاص فرایند چوب‌شکنی پنهان (رودریگوئز و همکاران ، ۲۰۱۰) را در خود دارد. به بیان واضح‌تر، اگر در مدل $BRSBP$ به جای فرایند فضایی رگرسیون-بتا، فرایند فضایی یکنواخت $U(\boldsymbol{s})$ در نظر گرفته شود آن‌گاه فرایند چوب‌شکنی پنهان ($LaSBP$) حاصل می‌شود. علاوه براین اگر در مدل پیشنهادی، ψ به بینهایت میل کند آن‌گاه مدل $BRSBP$ به یک مدل رگرسیون لجستیک تبدیل می‌شود. همچنین اگر $\mu_{\boldsymbol{x}}(\boldsymbol{s})$ به صورت مدل رگرسیون پروبیت مدل‌بندی شود و ψ به بینهایت میل کند آن‌گاه مدل $BRSBP$ به مدل رگرسیون پروبیت تبدیل می‌شود. بنابراین مدل پیشنهادی رده‌ی بزرگتری است که در حالت خاص مدل‌های رگرسیون لجستیک و پروبیت را در خود دارد. همچنین توزیع بتا نسبت به توزیع یکنواخت انعطاف بیشتری دارد و پارامترهای این توزیع می‌تواند رفتارهای مختلفی از داده‌ها را مدل‌بندی کند.
- افزایشی القا شده توسط مدل $BRSBP$ روی اثر فضایی مبتنی بر متغیرکمکی است که با گسسته‌سازی فرایند فضایی رگرسیون-بتا حاصل می‌شود. همچنین در تعریف مدل $BRSBP$ از یک فرایند گاوسی برای القای وابستگی فضایی استفاده شده است که این کار موجب سادگی در محاسبات و برازش مدل نسبت به فرایندهای چوب‌شکنی دیگر می‌شود.

۳ استنباط بیزی و پیشگویی فضایی

از آن‌جا که هر یک از پارامترها ویژگی خاصی از میدان تصادفی را بیان می‌کنند، فرض می‌شود کلیه‌ی پارامترها مستقل از هم هستند. برای ضرایب رگرسیونی $\boldsymbol{\beta}$ و γ توزیع‌های پیشینی ناآگاهی‌بخش نرمال چندمتغیره در نظر گرفته می‌شود. برای پارامترهای واریانس σ^2 و τ^2 به‌ترتیب پیشینی مزدوج گامای وارون ناآگاهی‌بخش $IG(s_{\sigma^2},r_{\sigma^2})$ و $IG(s_{\tau^2},r_{\tau^2})$ استفاده می‌شود. همچنین پیشینی گاما $G(s_\alpha,r_\alpha)$ و $G(s_\psi,r_\psi)$ به‌ترتیب برای پارامترهای α و ψ در نظر گرفته می‌شود. همچنین برای برازش مدل پیشنهادی، تابع همبستگی نمایی با پارامتر λ به صورت $\rho(\boldsymbol{s},\boldsymbol{s}';\lambda)=\exp\{-\frac{||\boldsymbol{s}-\boldsymbol{s}'||}{\lambda}\}$ فرض می‌شود. برای پارامتر همبستگی λ نیز توزیع گاما $G(a_\lambda,b_\lambda)$ در نظر گرفته می‌شود.

فرض کنید $\boldsymbol{\Theta}=\{\boldsymbol{\beta},\gamma,\sigma^2,\tau^2,\lambda,\alpha,\psi\}$ بردار پارامترهای نامعلوم باشد. با توجه به این‌که توزیع پسینی دارای فرم بسته نیست، با اتخاذ روش نمونه‌گیری گیبز و داده‌افزایی توزیع‌های تمام‌شرطی تعیین و از آن‌ها نمونه‌گیری می‌شود. با توجه به این‌که برخی از

توزیع‌های تمام‌شرطی دارای فرم استاندارد نیستند از الگوریتم متروپولیس–هستینگس برای نمونه‌گیری از آن‌ها استفاده می‌شود. به منظور خلاصه‌سازی جزئیات محاسبات بیان نشده است.

یکی از اهداف اصلی تحلیل داده‌های فضایی پیشگویی فرایند مورد مطالعه در مکان‌های فاقد نمونه است. برای این منظور فرض کنید $y^* = y(\boldsymbol{s}_0)$ و $\boldsymbol{y} = (y(\boldsymbol{s})_1, \cdots, y(\boldsymbol{s}_n))'$ به‌ترتیب نشان‌دهنده‌ی متغیر پاسخ در هر موقعیت دلخواه جدید $\boldsymbol{s}_0$ و برداری از مشاهدات از فرایند Y در مکان‌های $\boldsymbol{s}_1$، $\cdots$، $\boldsymbol{s}_n$ باشند. در این صورت تحت مدل معرفی شده، توزیع پیشگوی فضایی برای مکان $\boldsymbol{s}_0$ به صورت

$$p(y^*|\boldsymbol{y}) = \int p(y^*|\xi_{\boldsymbol{x}}(\boldsymbol{s}_0), \theta_{\xi_{\boldsymbol{x}}(\boldsymbol{s}_0)}, \boldsymbol{\beta}, \tau^2) p(\xi_{\boldsymbol{x}}(\boldsymbol{s}_0), \theta_{\xi_{\boldsymbol{x}}(\boldsymbol{s}_0)}, \boldsymbol{\beta}, \tau^2|\boldsymbol{y}) d\boldsymbol{\beta} d\tau^2 d\xi_{\boldsymbol{x}}(\boldsymbol{s}_0) d\theta_{\xi_{\boldsymbol{x}}(\boldsymbol{s}_0)} \qquad (1.3)$$

است که در آن

$$p(y^*|\xi_{\boldsymbol{x}}(\boldsymbol{s}_0), \theta_{\xi_{\boldsymbol{x}}(\boldsymbol{s}_0)}, \boldsymbol{\beta}, \tau^2) \sim N(\boldsymbol{w}'\boldsymbol{\beta} + \theta_{\xi_{\boldsymbol{x}}(\boldsymbol{s}_0)}, \tau^2). \qquad (2.3)$$

در رابطه (۱.۳) محاسبه‌ی انتگرال به صورت تحلیلی ناممکن است و با استفاده از روش‌های مونت‌کارلو می‌توان آن را تقریب زد. در این راستا، به ازای هر مشاهده پسینی از $\xi_{\boldsymbol{x}}(\boldsymbol{s}_0)$، $\theta_{\xi_{\boldsymbol{x}}(\boldsymbol{s}_0)}$، $\boldsymbol{\beta}$ و τ^2 یک مشاهده از توزیع (۲.۳) تولید می‌شود. سپس با تکرار مراحل فوق به اندازه مورد نیاز، امکان دستیابی به مشاهدات از توزیع پیشگو به صورت $\{y(\boldsymbol{s}_0)^{(m)}, j = 1, \cdots, M\}$ مهیا می‌شود. در نتیجه پیشگویی فضایی بیزی در مکان $\boldsymbol{s}_0$ به صورت زیر بدست می‌آید:

$$\hat{y}(\boldsymbol{s}_0) \equiv \hat{\mathbb{E}}(y(\boldsymbol{s}_0)|\boldsymbol{y}) = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{y}^{(m)}(\boldsymbol{s}_0).$$

۴ مثال کاربردی

در این بخش از مقاله به عملکرد و ارزیابی مدل معرفی شده از دیدگاه پیشگویی فضایی و افرازبندی ناحیه مورد مطالعه در مقایسه با مدل‌های معرفی شده توسط ریچ و همکاران (۲۰۱۱)، رایزر و کالدر (۲۰۱۵)، گرمسی و لی (۲۰۰۸)، پیچ و کوئینتانا (۲۰۱۶) و رودریگوئز و همکاران (۲۰۱۰) پرداخته می‌شود. برای این منظور، این مدل‌ها به اختصار به‌ترتیب به صورت $M1$، $M2$، TGP، PPM، $LaSBP$ و $BRSBP$ نمایش داده می‌شوند. داده‌های مورد بررسی متوسط میزان بارش برحسب میلی‌متر مربوط به ۴۱۳ ایستگاه هواشناسی در سال ۲۰۱۶ برای کشور ایران است. این داده‌ها به وسیله‌ی سازمان هواشناسی و وزارت نیروانداز‌ه‌گیری شده است. در ادامه برای بررسی نامانایی داده‌ها در ساختار دوم فرایند از آزمون نامانایی که توسط سویتربانديوپدی و سوبارائو (۲۰۱۷) بیان گردید، استفاده شده است که در سطح ۵ درصد مانایی داده‌ها رد می‌شود. همچنین با توجه به این‌که میزان بارش به متغیرهای هواشناسی از جمله دما وابسته است، دما به عنوان متغیر کمکی برای برازش مدل‌ها انتخاب می‌کنیم. از آن‌جا که هدف این مقاله بررسی تاثیر متغیر کمکی در همبستگی فضایی است، ابتدا روند از داده‌ها حذف و با داده‌های بدون روند برازش مدل‌ها را انجام می‌دهیم.

در ادامه از معیار میانگین توان دوم خطا[3] ($MSPE$) برای ارزیابی قدرت پیشگویی مدل‌های بیان شده، استفاده می‌شود. برای این منظور ۱۰ مجموعه داده با حذف تصادفی ۱۰ درصد از داده‌ها تولید شده و سپس با استفاده از رهیافت بیزی مدل‌های $M1$، $M2$، TGP، PPM، $LaSBP$ و $BRSBP$ به داده‌ها برازش داده شده است. برای برازش مدل، سه زنجیر با نقاط شروع اولیه

[3] Mean Squared Prediction Error

جدول ۱: مقایسه‌ی عملکرد پیشگویی مدل‌های: M۱، M۲، $LaSBP$، TGP، PPM و $BRSBP$.

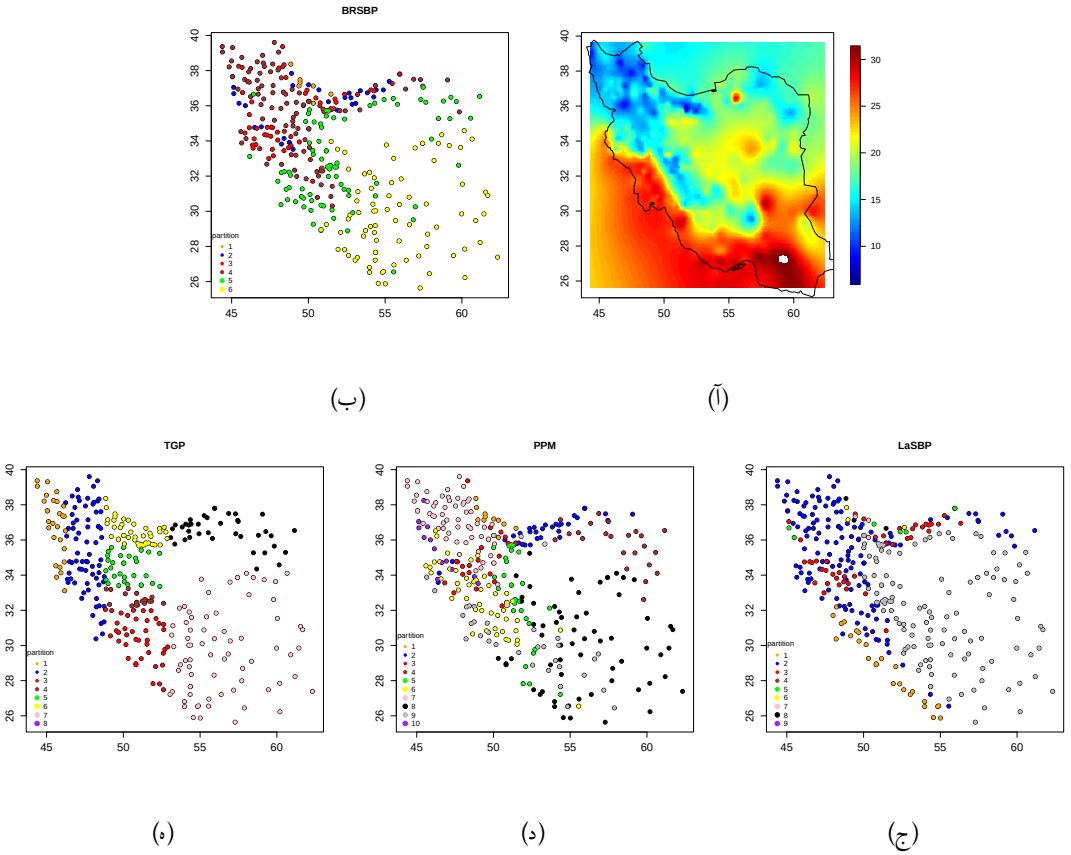
مدل	۱M	۲M	LaSBP	BRSBP	PPM	TGP
MSPE	۱٫۹۲	۱٫۳۶	۱٫۰۸	۰٫۶۴	۰٫۸۸	۰٫۷۶

متفاوت و ۲۰۰۰۰ تکرار الگوریتم MCMC اجرا شد که تعداد ۵۰۰۰ تکرار اولیه به عنوان دوره‌ی داغیدن در نظر گرفته شد. به منظور خلاصه‌سازی، جزئیات انتخاب پیشینی و نحوه‌ی استنباط بیزی این مدل‌ها بیان نشده است. در جدول ۱ نتایج هر یک از مدل‌ها براساس معیار $MSPE$ آمده است. همان‌طور که مشاهده می‌شود مدل $BRSBP$ عملکرد بهتری نسبت به دیگر مدل‌ها دارد. به‌خصوص، مدل‌های M۱ و M۲ عملکرد ضعیف‌تری نسبت به مدل پیشنهادی دارند. در واقع مدل $BRSBP$ مقدار $MSPE$ کمتری نسبت به مدل‌های M۱، M۲، PPM و TGP داراست. مقدار $MSPE$ برای مدل پیشنهادی ۰٫۶۴ و برای مدل $LaSBP$ برابر با ۱٫۰۸ است که تقریباً ۴۱٪ کوچکتر است.

در شکل ۱ نمودار افرازبندی ایستگاه‌های هواشناسی حاصل از برازش مدل $BRSBP$ آمده است. همانطور که مشاهده می‌شود ایستگاه‌هایی که رفتار دمایی یکسانی دارند در خوشه‌های یکسان قرار گرفته‌اند (شکل ب۱). به عنوان مثال، مناطق شرقی و جنوب شرقی ایران که دارای متوسط دمای بالا و میزان بارش کم هستند، در خوشه‌های یکسان قرار گرفته‌اند. همچنین، در شکل‌های ج۱، د۱ و ه۱ به‌ترتیب افرازبندی ایستگاه‌های هواشناسی حاصل از برازش مدل‌های $LaSBP$، PPM و TGP آمده است. همانطور که مشاهده می‌شود، افرازبندی ایجاد شده توسط مدل‌های $LaSBP$ و PPM تقریباً شبیه مدل $BRSBP$ است اما تعداد خوشه‌های ایجاد شده براساس مدل $BRSBP$ کمتر است. اما در مدل TGP ایستگاه‌های هواشناسی نزدیک به هم، در خوشه‌های یکسان قرار گرفته‌اند. با این حال، با توجه به تنوع آب و هوایی که در ایستگاه‌های نزدیک به هم وجود دارد، افرازبندی ایجاد شده توسط مدل TGP منطقی و معقول به‌نظر نمی‌رسد.

بحث و نتیجه‌گیری

در این مقاله با استفاده از فرایند رگرسیون–بتا، رویکرد جدیدی برای ساخت فرایندهای فضایی مورد مطالعه قرار گرفت به‌طوری‌که توزیع تصادفی وابسته به متغیر کمکی است. همانند فرایندهای چوب‌شکنی تعمیم‌یافته تابع کوواریانس مدل معرفی شده نامانا و وابسته به متغیر کمکی است. از ویژگی‌های مهم مدل معرفی شده می‌توان به خوشه‌بندی بر اساس متغیر کمکی اشاره کرد که باعث بهبود در افرازبندی ناحیه مورد مطالعه می‌شود. همچنین با استفاده از یک مجموعه داده واقعی عملکرد مدل از نظر پیشگویی فضایی و خوشه‌بندی مورد ارزیابی قرار گرفت.



شکل ۱: (آ) نمودار رویه دما و (ب، ج، د ، ه) نمودارهای افرازبندی ایستگاه‌های هواشناسی از برازش مدل‌های $BRSBP$، $LaSBP$، PPM و TGP.

مراجع

Bandyopadhyay, S., and Rao, S. S. (2017), *A test for stationarity for irregularly spaced spatial data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79, 95–123.

Barcella, W., M. De Iorio, S. Favaro, and G. L. Rosner (2018), *Dependent generalized dirichlet process priors for the analysis of acute lymphoblastic leukemia*, Biostatistics, 19(3), 342–358.

Barzegar, Z., and Rivaz, F. (2020), *A scalable Bayesian nonparametric model for large spatio-temporal data*, Computational Statistics, 35, 153–173.

Dahdouh, O., and Khaledi, M. J., (2020), *Generalized spatial stick-breaking processes*, Communications in Statistics Simulation and Computation, 1-20.

Dunson, D. B. and J.-H. Park (2008),*Kernel stick-breaking processes*, Biometrika 95(2), 307–323.

Gramacy, R. B. and H. K. H. Lee (2008), *Bayesian treed gaussian process models with an application to computer modeling*. Journal of the American Statistical Association, 103(483), 1119–1130.

Ingebrigtsen, R., F. Lindgren, and I. Steinsland (2014), *Spatial models with explanatory variables in the dependence structure*, Spatial Statistics 8, 20–38.

Page, G. L. and F. A. Quintana (2016), *Spatial product partition models*, Bayesian Analysis 11(1), 265–298. 149.

Reich, B. J. and M. Fuentes (2007), *A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields*, The Annals of Applied Statistics 1(1), 249–264.

Reich, B.J., Eidsvik, J., Guindani, M., Nail, A.J., Schmidt, A.M., (2011), *A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration*, The annals of applied statistics. 5 (4), 2425—2447.

Ren, L., L. Du, L. Carin, and D. B. Dunson (2011), *Logistic stick-breaking process*, Journal of Machine Learning Research 12(1).

Risser, M.D., Calder, C.A., (2015), *Regression-based covariance functions for nonstationary spatial modeling*, Environmetrics 26 (4), 284—297.

Rodriguez, A. and D. B. Dunson (2011), *Nonparametric Bayesian models through probit stick-breaking processes*, Bayesian Analysis 6(1).

Rodríguez, A., Dunson, D.B. and Gelfand, A.E., (2010), *Latent stick-breaking processes*, Journal of the American Statistical Association, 105(490), pp.647-659.

Schmidt, A. M., P. Guttorp, and A. O'Hagan (2011), *Considering covariates in the covariance structure of spatial processes*, Environmetrics 22(4), 487–500.

Yarali, E., Rivaz, F., (2020), *Incorporating covariate information in the covariance structure of misaligned spatial data*, Environmetrics 31 (6), e2623.

Yarali, E., Rivaz, F., and Khaledi, M. J. (2022), *A Bayesian nonparametric spatial model with covariate-dependent joint weights*, Spatial Statistics, 100662.

Beta regression stick-breaking process with covariate dependent nonstationary covariance

Esmail Yarali, Firoozeh Rivaz[1], and Majid Jafari Khaledi[2]

[1]Department of Statistics, Shahid Beheshti University, Tehran.

[2]Department of Statistics, Tarbiat Modares University, Tehran.

Abstract: This paper introduces a new class of stick-breaking processes that incorporate covariate information into the random joint distributions of spatial processes. The proposed modeling approach allows for a nonstationary covariance function that is driven by covariates, and it also induces a covariate-dependent random partitioning scheme. Using this process as a random effect distribution within a linear mixed model framework, the data can be effectively modeled, and a Bayesian inference approach is developed. The performance of the proposed model is evaluated and compared with that of competing spatial models using a real-world example.

Keywords: Stick-breaking process, Beta-Regression, Covariate information, Non-stationarity, Covariance structure.

Mathematics Subject Classification (2020): 62H11, 62M30, 62F15.