
بسم الله الرحمن الرحیم

دانشکده علوم ریاضی

رشته آمار، گرایش آمار ریاضی

پایان‌نامه کارشناسی ارشد

استنباط بیزی تقریبی مدل‌های الگوی نقطه‌ای فضایی پیچیده با روش تقریب لاپلاس

نگارنده: فاطمه شیاسی

استاد راهنما

دکتر حسین باغیشنی

استاد مشاور

دکتر نگار اقبال

بهمن ۱۳۹۶

شماره:
تاریخ:

بسمه تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای فاطمه شیاسی با شماره دانشجویی ۹۴۱۰۶۸۴ رشته آمار گرایش آماریاضی تحت عنوان استنباط بیزی تقریبی مدل‌های الگوی نقطه‌ای فضایی پیچیده با روش تقریب لاپلاس که در تاریخ ۱۳۹۶/۱۱/۹ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه: ..خیلی خوب....) ☒ مردود ☐

نوع تحقیق: نظری ☒ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر حسین باغیشنی	استادیار	
۲- استادراهنمای دوم			
۳- استاد مشاور	دکتر نگار اقبال	استادیار	غایب
۴- نماینده تحصیلات تکمیلی	دکتر احمد نزاکتی	دانشیار	
۵- استاد ممتحن اول	دکتر محمدرضا ربیعی	استادیار	
۶ـ استاد ممتحن دوم	دکتر داود شاهسونی	دانشیار	

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشم تاریخ و امضاء و مهر دانشکده

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می‌تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

معاونت تحقیقات و فناوری
دانشکده ریاضی

تقدیم به پدر و مادرم :

خدای را بسی شاکرم که از روی کرم، پدر و مادری فداکار نسیبم ساخته تا در سایه درخت پربار وجودشان بیاسایم و از ریشه آنها شاخ و برگ گیرم و از سایه وجودشان در راه کسب علم و دانش تلاش نمایم. والدینی که بودنشان تاج افتخاری است بر سرم و نامشان دلیلی است بر بودنم، چرا که این دو وجود، پس از پروردگار، مایه هستی ام بوده اند دستم را گرفتند و راه رفتن را در این وادی زندگی پر از فراز و نشیب آموختند، آموزگارانی که برایم زندگی، بودن و انسان بودن را معنا کردند....

«هر خشت زسر منزل امید به جایی‌ست از بس که زمین دل ما زلزله دارد»

تقدیر و تشکر:

سپاس و ستایش مر خدای را که آثار قدرت او بر چهره روز روشن، تابان است و انوار حکمت او در دل شب تار، درفشان. آفریدگاری که خویشتن را به ما شناساند و درهای علم را بر ما گشود و عمری و فرصتی عطا فرمود تا بدان، بنده ضعیف خویش را در طریق علم و معرفت بیازماید. حال که به یاری پروردگار، تحقیق حاضر به پایان رسیده بر خود لازم می‌دانم از زحمات تمامی بزرگوارانی که در طی دوران تحصیل و اجرای این پایان‌نامه همراه و پشتیبانم بودند، سپاس‌گزاری نمایم.

از استاد با کمالات و شایسته و دلسوز؛ جناب آقای دکتر حسین باغیشنی که در کمال سعه صدر، با حسن خلق و فروتنی، از هیچ کمکی در این عرصه بر من دریغ ننمودند و بدون حضور ایشان اتمام این رساله برایم ناممکن بود، نهایت تشکر و امتنان را دارم و برای آن بزرگوار آرزوی سلامتی و کامیابی از درگاه خداوند تبارک و تعالی می‌نمایم.

از سرکار خانم دکتر نگار اقبال که زحمت مشاوره این رساله را متقبل شدند؛ کمال تشکر و قدردانی را دارم.

از داوران جناب آقای دکتر محمدرضا ربیعی و جناب آقای دکتر داود شاهسونی که زحمت مطالعه و داوری پایان‌نامه این‌جانب را تقبل کردند صمیمانه کمال تشکر را دارم. از کلیه اساتید دانشگاه صنعتی شاهرود که به مدد تحصیل در این مقطع افتخار شاگردی‌شان نصیب بنده شد کمال تشکر و قدردانی را دارم.

همچنین از دوست عزیزم خانم فاطمه شیربره که تا پایان راه خالصانه در کنارم بودند کمال تشکر را دارم.

فاطمه شیاسی

بهمن ۱۳۹۶

تعهدنامه

این‌جانب **فاطمه شیاسی** دانشجوی کارشناسی ارشد رشته **آمار علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **استنباط بیزی تقریبی مدل‌های الگوی نقطه‌ای فضایی پیچیده با روش تقریب لاپلاس** ، تحت راهنمایی **حسین باغیشنی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط این‌جانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده‌اند.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘دانشگاه صنعتی شاهرود’’ یا ‘‘Shahrood University of Technology’’ به چاپ خواهند رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده‌اند.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده) شده است، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

فاطمه شیاسی
بهمن ۱۳۹۶

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

داده‌های الگوی نقطه‌ای فضایی، کاربردهای متنوعی در دنیای واقعی دارند. به‌عنوان نمونه، می‌توان شناسایی نقاط زلزله‌خیز و بررسی گونه‌های مختلف گیاهان را نام برد. در تحلیل این نوع از داده‌ها ویژگی اصلی مورد علاقه، تابع شدت است. برای مدل‌بندی این تابع رهیافت‌های متفاوتی مانند فرآیندهای نقطه‌ای دوجمله‌ای، پواسن و کاکس توسط آماردان‌های مختلف معرفی شده‌اند. با توجه به ماهیت این داده‌ها و انعطاف‌پذیری فرآیندهای کاکس، در این پایان‌نامه، زیررده‌ای از این فرآیندها معروف به فرآیندهای لگ‌گاوسی را در نظر می‌گیریم. فرآیندهای لگ‌گاوسی، به دلیل سادگی تعبیه کردن متغیرهای کمکی درون خود، بسیار منعطف عمل می‌کنند. استنباط کلاسیک در فرآیندهای لگ‌گاوسی نیازمند محاسبات پیچیده برای به‌دست آوردن تقریبی از تابع درست‌نمایی است. به همین دلیل، رهیافت معمول محققان برای برازش تابع شدت با فرآیندهای لگ‌گاوسی، بیزی است. با توجه به صریح نبودن شکل توزیع پسین در این فرآیندها ابزار متداول برای برازش مدل، الگوریتم‌های مونت کارلوی زنجیر مارکوفی (MCMC) هستند. اجرای این الگوریتم‌ها نیاز به مهارت‌های برنامه‌نویسی دارد و از لحاظ همگرایی و زمان محاسبات، آلوده به مشکلات جدی است. یک رهیافت جانشین، استفاده از یک روش تقریبی بیزی به‌نام تقریب لاپلاس آشیانه‌ای جمع‌بسته (INLA) است. این تقریب در مقایسه با الگوریتم‌های MCMC بسیار سریع است و نتایج آن از نظر دقت برابری می‌کند.

در این پایان‌نامه، از روش تقریبی INLA برای برازش مدل‌های کاکس لگ‌گاوسی استفاده می‌کنیم. برای نمایش کاربست مباحث نظری مطرح‌شده، داده‌های الگوی نقطه‌ای زلزله ناحیه شمال غرب ایران را با مدل کاکس لگ‌گاوسی تحلیل می‌کنیم.

کلمات کلیدی: الگوهای نقطه‌ای، فرآیندهای کاکس لگ‌گاوسی، تابع شدت، تابع K، تقریب لاپلاس آشیانه‌ای جمع‌بسته.

فهرست مقالات مستخرج از پایان‌نامه

۱. شیاسی، فاطمه، باغیشنی، حسین، اقبال، نگار، (۱۳۹۶)، تحلیل بیزی تقریبی الگوی نقطه‌ای فضایی زمین‌لرزه‌های شمال غرب ایران با فرآیندهای کاکس لگ‌گاوسی، مجموعه مقالات دومین سمینار آمار فضایی و کاربردهای آن، دانشگاه صنعتی شاهرود، شاهرود، ۱۵۹-۱۶۶.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

الگوهای نقطه‌ای فضایی: مفاهیم و تعاریف پایه‌ای

در این فصل ابتدا داده‌های فضایی و مفاهیم پایه‌ای الگوهای نقطه‌ای فضایی را معرفی می‌کنیم. سپس استنباط خود را با رویکرد بیزی انجام می‌دهیم. بخش‌های از این فصل برگرفته از کتاب دکتر محمدزاده (۱۳۹۸)، پایان‌نامه کارشناسی ارشد جلیلیان (۱۳۹۵) و کتاب بدلی (۲۰۱۵) است.

۱.۱ آمار فضایی

برخی از داده‌ها به گونه‌ای هستند که در یک ناحیه فضایی گردآوری شده‌اند و دارای مشخصه‌های مکانی هستند و به نوعی وابستگی آن‌ها ناشی از موقعیت و مکان قرار گرفتن آن‌ها در فضای مورد بررسی است. به‌عنوان مثال، غلظت یک نوع ماده معدنی در خاک یک ناحیه، مکان تصادفات جاده‌ای در یک شهر بزرگ و مکان درختان در یک جنگل، داده‌هایی هستند که دارای مشخصه‌های مکانی هستند. برای تحلیل این نوع داده‌ها از روش‌های کلاسیک آماری که بر پذیره استقلال و هم‌توزیعی مشاهدات استوارند، نمی‌توان استفاده کرد. در واقع این نوع داده‌ها دارای همبستگی‌های فضایی هستند. آمار فضایی شاخه‌ای از آمار است که به تحلیل داده‌های با مشخصه‌های مکانی که وابستگی آن‌ها تابعی از فاصله بین موقعیت‌های مشاهدات

باشد می‌پردازد. در این حالت مشاهدات نزدیک به‌هم وابسته‌تر و مشاهدات دورتر از هم، وابستگی کم‌تری دارند. این نوع داده‌ها، داده‌های فضایی نامیده می‌شوند. یکی از اهداف آمار فضایی بررسی و مطالعه ساختار وابستگی فضایی بین مشاهدات و مدل‌بندی آن است.

اولین نشانه‌های آماری داده‌های فضایی به هالی جغرافیدان و ستاره‌شناس انگلیسی و سال ۱۶۸۶ برمی‌گردد. سپس پیدایش و گسترش این شاخه از آمار توسط زمین‌شناسان به منظور بررسی داده‌های ذخایر معدنی صورت گرفت. ماترون (۱۹۶۲) از جمله نخستین زمین‌شناسانی بود که به بررسی داده‌های فضایی با رویکرد آمار فضایی پرداخت. ریپلی با نگارش کتاب «آمار فضایی» در سال ۱۹۸۱ دریچه تازه‌ای بر این شاخه از آمار گشود. او با مطرح کردن پیش‌گویی فرآیندهای تصادفی و انقلاب رایانه، زمینه را برای روش‌های مطرح‌شده در آمار فضایی مهیا کرد. با این حال آمار فضایی با انتشار کتاب «آمار برای داده‌های فضایی» به قلم کرسی در سال ۱۹۹۳ بیشتر شناخته شد و پژوهشگران مختلفی از آمار و سایر رشته‌های مرتبط با آن به تحقیق در این زمینه پرداختند.

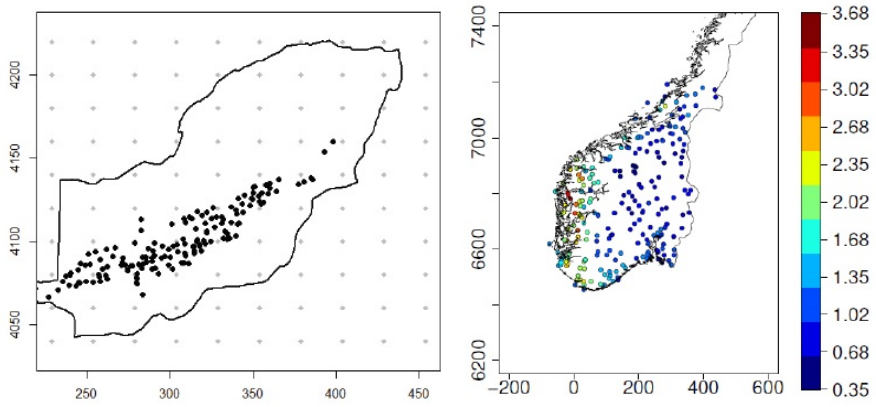
۲.۱ انواع داده‌های فضایی

داده‌های فضایی بر اساس این‌که موقعیت مشاهدات و متغیر مورد اندازه‌گیری به چه صورت باشند به سه گروه داده‌های زمین‌آماری[۱]، مشبکه‌ای[۲] و الگوهای نقطه‌ای[۳] تقسیم می‌شوند.

۱. **داده‌های زمین‌آماری:** اگر ناحیه جغرافیایی مورد نظر چگال[۴] باشد، یعنی بین هر دو موقعیت امکان مشاهده موقعیت‌های دیگر هم وجود داشته باشد، داده‌های حاصل از این ناحیه، داده‌های زمین‌آماری نامیده می‌شوند. به‌عنوان مثالی از این نوع داده‌ها، می‌توان به تعداد نوعی جانور دریایی که در طول یک ساحل، در مکان‌های مختلف نمونه‌گیری شده‌اند اشاره کرد. به‌عنوان مثال‌های دیگر (شکل ۱.۱) پراکنش جغرافیایی ایستگاه‌های اندازه‌گیری حجم بارش در جنوب نروژ و ایستگاه‌های سنجش شاخص‌های آب زیرزمینی در استان گلستان را نشان می‌دهد. به‌طور معمول هدف از تحلیل داده‌های زمین‌آماری، پیش‌گویی مقدار متغیر مورد بررسی در موقعیت‌های جدید است.

۲. **داده‌های مشبکه‌ای:** این نوع داده‌ها زمانی که نواحی مورد بررسی غیرچگال یا به عبارتی مکان‌ها ناحیه‌ای باشند، مشاهده می‌شوند. این مکان‌ها ممکن است منظم یا نامنظم باشند. به‌طور مثال، کشور ایران دارای ۳۱ استان است و اگر داده‌های تعداد مبتلایان به یک نوع سرطان در چند سال گذشته در این استان‌ها موجود باشند، این نوع داده‌ها از نوع مشبکه‌ای خواهند بود. تعداد حوادث رانندگی در ۲۲ منطقه تهران و تعداد

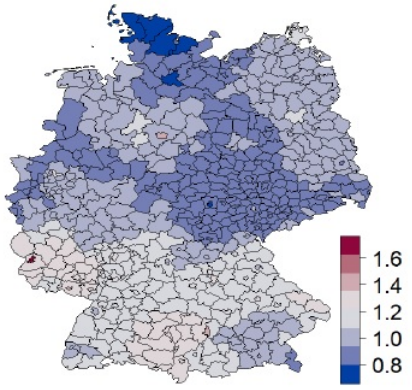
[۱] Geostatistical data
[۲] Lattice data
[۳] Point pattern
[۴] Dense



(ب) (الف)

شکل ۱.۱: الف) نقشه ۲۳۳ ایستگاه اندازه‌گیری بارش در جنوب نروژ، ب) پراکنش جغرافیایی ایستگاه‌های سنجش شاخص‌های آب زیرزمینی در استان گلستان

مرگ و میر بیماران مبتلا به سرطان حنجره در ۵۴۴ ناحیه در آلمان (شکل ۲.۱) از دیگر مثال‌های داده‌های مشبکه‌ای است. در تحلیل داده‌های فضایی مشبکه‌ای، اغلب مدل‌بندی احتمالاتی مشاهدات، مد نظر است.



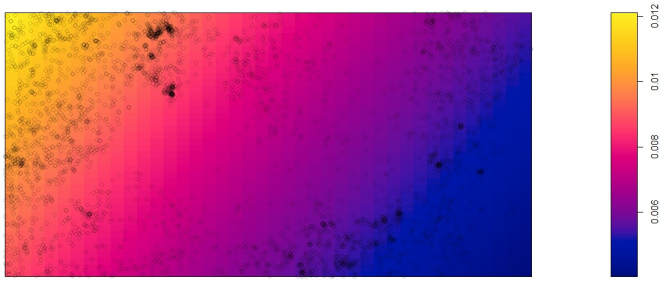
شکل ۲.۱: تعداد مرگ و میر بیماران مبتلا به سرطان حنجره در ۵۴۴ ناحیه در آلمان

۳. **الگوهای نقطه‌ای:** در داده‌های الگوی نقطه‌ای، موقعیت مشاهده‌شده همان مقدار متغیر تصادفی مورد علاقه است. به‌طور دقیق‌تر یک ناحیه فضایی را در نظر بگیرید که امکان وقوع یک پیشامد خاص در هر نقطه از این فضا وجود داشته باشد. با انتخاب یک زیرمجموعه کراندار از این فضا و ثبت مکان وقوع پیشامدهای مد نظر در آن، با این نوع داده‌ها مواجه می‌شویم. به زیرمجموعه کراندار پنجره مشاهده و به مجموعه مکان‌های پیشامدهای مشاهده‌شده یک الگوی نقطه‌ای فضایی گفته می‌شود. به‌عنوان مثال، مکان‌های وقوع زلزله که در یک پنجره مشاهده از پیش تعیین‌شده ثبت شده‌اند، یک الگوی نقطه‌ای فضایی را شکل می‌دهند یا مکان درختان در یک جنگل بارانی استوایی.

۱.۲.۱ برخی از کاربردهای الگوهای نقطه‌ای

علاقه به روش‌های تحلیل داده‌های الگوهای نقطه‌ای در بسیاری از زمینه‌های علمی، به‌ویژه در بوم‌شناسی، همه‌گیرشناسی[۵]، علوم زمین، نجوم، اقتصادسنجی و تحقیقات جرم‌شناختی، به سرعت در حال گسترش است. برای درک بهتر وسعت کاربردهای این نوع تحلیل‌ها در این بخش چند مثال را ارایه می‌دهیم.

مثال ۱.۲.۱. شکل ۳.۱ یک پنجره مشاهده با ابعاد ۱۰۰۰ متر در ۵۰۰ متر، الگوی نقطه‌ای مکان ۳۶۰۴ درخت از گونه‌های بیلسچرندیا پندوال[۶] را نشان می‌دهد. این مجموعه داده



شکل ۳.۱: الگوی نقطه‌ای مکان یک گونه درخت

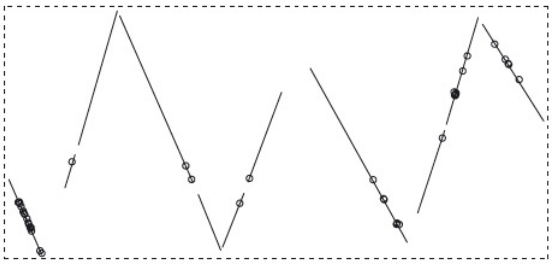
بخشی از مجموعه داده بسیار بزرگ‌تر حاوی مکان صدها هزار درخت متعلق به صدها گونه است. جزئیات بیشتر در مورد مجموعه داده‌ها و تحلیل آن‌ها در هوبل و فوستر (۱۹۸۳)، کندیت و همکاران (۱۹۹۶) و کندیت (۱۹۹۸) قابل مشاهده هستند.

این‌که آیا داده‌ها از الگوی خاصی پیروی می‌کنند، سوال اساسی است که همواره در الگوهای نقطه‌ای مطرح می‌شود. همان‌طور که در شکل مشاهده می‌شود، درختان این گونه در برخی از بخش‌های پنجره مشاهده، فراوانی بیشتر و در برخی از بخش‌های آن فراوانی کمتری دارند. این ناهمگنی در توزیع فضایی مکان درختان می‌تواند به علت تغییرات مواد غذایی مورد نیاز و رقابت با سایر گونه‌ها باشد. هم‌چنین علت پراکنش درختان به‌صورت خوشه‌ای در شکل می‌تواند به دلیل فرآیند تکثیر درختان باشد. پاسخ به سوال فوق و بررسی آماری و مدل‌بندی الگوی نقطه‌ای این مثال، مورد توجه پژوهشگران جنگل‌داری و بوم‌شناسی است.

مثال ۲.۲.۱. شکل ۴.۱ موقعیت ۵۵ نهنگ کوچک سیاه را که در قسمتی از اقیانوس اطلس شمالی نزدیک به شهر اسپیتزبرگن[۷] در کشور نروژ مشاهده شده‌اند، نشان می‌دهد. موقعیت نهنگ‌ها توسط یک کشتی که به دنبال آن‌ها حرکت می‌کردند، ثبت شده است. این مشاهدات یک الگوی نقطه‌ای را نشان می‌دهند. الگوی نقطه‌ای مذکور را می‌توان به‌عنوان مشاهداتی

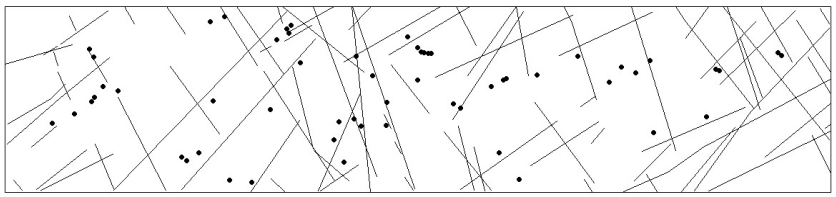
[۵] Epidemiology
[۶] Beilschrnedia pendula
[۷] Spitzbergen

ناقص از تمامی مکان‌های نهنگ‌ها در نظر گرفت؛ زیرا موقعیت این نهنگ‌ها را تنها در نزدیکی کشتی می‌توان ثبت کرد. هم‌چنین ممکن است به دلیل شرایط نامناسب جوی یا حرکت نهنگ‌ها در زیر آب و عدم مشاهده آن‌ها تمام موقعیت‌شان ثبت نشده باشد (جزئیات بیشتر در مورد مجموعه داده‌ها و تحلیل آن‌ها در اسکاوی و کوین (۲۰۰۴) و واگه پی‌ترسن و همکاران (۲۰۰۶) قابل مشاهده است). احتمال مشاهده یک نهنگ تابعی نزولی از فاصله بین کشتی و نهنگ است و برای مسافت بیشتر از ۲ کیلومتر، به علت مسافت زیاد نهنگ‌ها از کشتی و عدم مشاهده آن‌ها، این احتمال برابر با صفر است. بنابراین پنجره مشاهدات ناحیه‌ای چهارگوش به ضلع ۴ کیلومتر است که کشتی در مرکز آن قرار می‌گیرد.



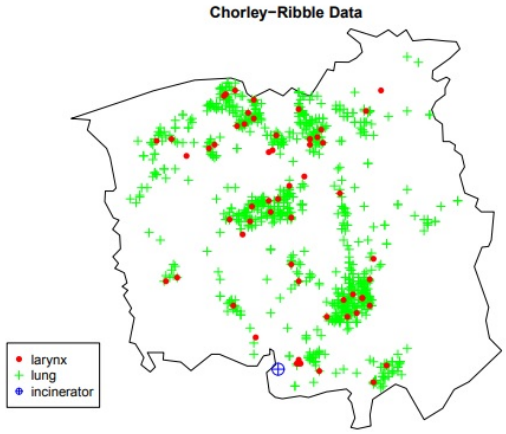
شکل ۴.۱: موقعیت حرکت نهنگ‌ها

مثال ۳.۲.۱. شکل ۵.۱ در مورد یک بررسی کانی‌شناسی است. در این شکل خطوط مستقیم، گسل زمین‌شناسی و نقاط ذخایر مس هستند. گسل‌ها از طریق تصاویر ماهواره‌ای به‌راحتی قابل مشاهده هستند اما ذخایر مس به‌سختی پیدا می‌شوند. سوال اساسی در این مثال این است که آیا گسل برای پیش‌بینی معدن مس مناسب است؟ به‌عنوان مثال، احتمال دارد میزان مس موجود در نزدیکی گسل‌ها کمتر یافت شود.



شکل ۵.۱: کانسارهای مس (نقاط) و گسل زمین‌شناسی (خطوط مستقیم)

مثال ۴.۲.۱. شکل ۶.۱ نقشه‌ای از مکان‌های اقامتگاه ۵۸ بیمار مبتلا به سرطان حنجره و ۹۷۸ بیمار مبتلا به سرطان ریه می‌باشد که توسط مقامات بهداشتی منطقه مورد نظر تهیه شده است. در این منطقه زباله‌سوز صنعتی متروکه‌ای نیز وجود دارد. سوال اساس مطرح‌شده در این مثال حاکی از این است که آیا با وجود گروه کنترل (سرطان ریه) ارتباطی بین سرطان حنجره و مکان زباله‌سوز صنعتی وجود دارد و آیا کسانی که سرطان حنجره دارند در معرض خطر سرطان ریه نیز هستند؟



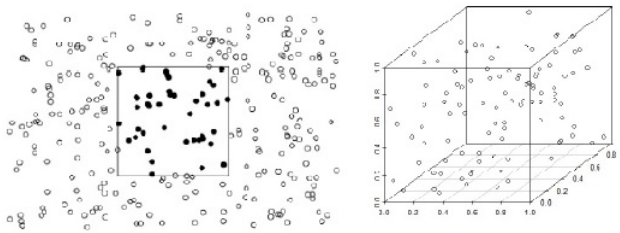
شکل ۶.۱: مکان‌های اقامتگاه ۵۸ بیمار مبتلا به سرطان حنجره در نزدیکی یک زباله‌سوز صنعتی متروکه

در ادامه، ابتدا برخی از تعاریف مورد نیاز برای پرداختن به تحلیل الگوهای نقطه‌ای را ارائه می‌دهیم.

۳.۱ مفاهیم الگوهای نقطه‌ای

تعریف ۱.۳.۱. فضایی را که الگوهای نقطه‌ای در آن تولید می‌شوند، **فضای وضعیت**[۸] نامیده و با S نشان می‌دهند، به‌طوری که $S \subset \mathbb{R}^d$.

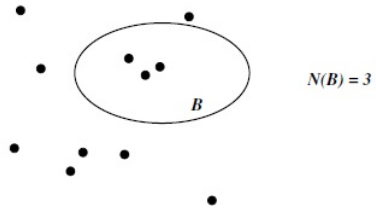
تعریف ۲.۳.۱. زیر مجموعه‌ای کراندار از S که مشاهده الگوی نقطه‌ای در آن مورد نظر است را **پنجره مشاهدات**[۹] نامیده و با W نشان می‌دهند. شکل ۷.۱ مثالی از پنجره مشاهدات را در فضای $\mathbb{R}^2$ و $\mathbb{R}^3$ نشان می‌دهد.



شکل ۷.۱: نمایشی از مفهوم پنجره مشاهدات

[۸] State Space
[۹] Observation Window

تعریف ۳.۳.۱. مجموعه $B \subset S$ را یک **الگوی نقطه‌ای** در S گویند، هرگاه B یک زیرمجموعه متناهی موضعی S باشد. الگوهای نقطه‌ای را معمولا با حروف بزرگ لاتین مثل X و B نشان می‌دهند، درحالی که از حروف یونانی مانند ζ و η برای نمایش اعضای S استفاده می‌کنند. شکل ۸.۱ نمایشی از الگوی نقطه‌ای B و اعضای آن است.



شکل ۸.۱: نمایشی از الگوی نقطه‌ای B و اعضای آن

۱.۳.۱ انواع الگوهای نقطه‌ای

دو تعمیم از الگوهای نقطه‌ای را می‌توان در دو دسته الگوهای نقطه‌ای نشان‌دار[۱۰] و الگوهای نقطه‌ای چندنوعی[۱۱] قرار داد.

الگوهای نقطه‌ای نشان‌دار

در برخی از الگوهای نقطه‌ای ممکن است به هر مکان (نقطه) اطلاعات دیگری با عنوان نشان، مانند شکل و اندازه، ضمیمه شود و آن را نشان‌دار کند. یک الگو نقطه‌ای نشان‌دار را می‌توان به صورت زوج (ζ, x) نمایش داد، که در آن x اطلاعات ضمیمه را نشان می‌دهد.

تعریف ۴.۳.۱. (بدلی، ۲۰۰۴): یک فرآیند نقطه‌ای نشان‌دار روی فضای S با نشان‌ها در فضای M یک فرآیند نقطه‌ای Y روی فضای $S \times M$ است، به‌طوری که برای هر زیرمجموعه فشرده $k \subset S$ تقریبا همه جا $N_y(k \times M) < \infty$ برقرار باشد. فضای نشان M، می‌تواند صورت‌های بسیار گسترده داشته باشد. برای مثال، ممکن است به‌صورت مجموعه‌ای متناهی، بازه‌ای از اعداد حقیقی یا فضای بسیار پیچیده مانند مجموعه کلیه چندضلعی‌های محدب باشد.

الگوهای نقطه‌ای چندنوعی

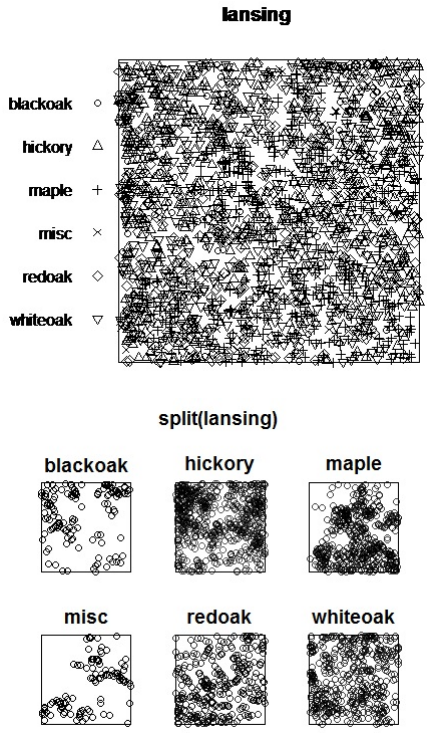
در شکل ۹.۱ زیر یک نمونه از مکان‌های داده‌های یک جنگل نمایش داده‌شده است که در آن شش نوع گونه درختی وجود دارد و مکان‌های هر گونه را نیز به‌طور جداگانه نمایش داده‌ایم. در این موارد با الگوهای نقطه‌ای چندنوعی روبرو هستیم. یک روش جایگزین برای تحلیل این داده‌ها، تقسیم کردن آن‌ها به شش الگوی نقطه‌ای است که هر الگو یک گونه درخت را بررسی می‌کند.

۴.۱ فرآیندهای نقطه‌ای

با توجه به تعریف فضای وضعیت، S فضایی است که الگوهای نقطه‌ای در آن رخ می‌دهند و W بخشی از S است که به‌عنوان پنجره مشاهدات مورد بررسی قرار می‌گیرد. به‌طور شهودی مجموعه‌ای از نقاط که به‌طور تصادفی در یک ناحیه مانند $X = \{u_۱, u_۲, ...\}$ توزیع شده‌اند را یک فرآیند نقطه‌ای می‌نامند. یک فرآیند نقطه‌ای فضایی، یک الگوی تصادفی از نقاط در فضای $d > ۲$ بعدی است (معمولا $d = ۲$ یا $d = ۳$). فرآیندهای نقطه‌ای فضایی به‌عنوان مدل‌های آماری در تحلیل الگوهای مشاهده‌شده نقاط مفید هستند که تعریف دقیق‌تر آن به‌صورت زیر است.

[۱۰] Marked point processes

[۱۱] Multivariate point processes



شکل ۹.۱: نمایشی از الگوهای نقطه‌ای چندنوعی

تعریف ۱.۴.۱. (مولر و واگه‌پترسن، ۲۰۰۴). مجموعه تصادفی X را یک فرآیند نقطه‌ای بر S گویند، هرگاه برای هر مجموعه کراندار و متناهی $B \subset S$، $X_B = X \cap B$ متناهی باشد.

در واقع هر الگوی نقطه‌ای، یک تحقق از X مانند $X = \{u_1, u_2, \ldots\}$ است.

تعریف ۲.۴.۱. $N(\cdot)$ را تابع شمارش X گویند، هرگاه برای هر $B \subset S$، $N(B) = n(X_B)$ تعداد نقاط فرآیند نقطه‌ایX در B باشد. در واقع $N(B)$ شمارشی (صحیح مقدار و نامنفی) است.

احتمال این‌که فرآیند نقطه‌ای X در مجموعه B هیچ نقطه‌ای نداشته باشد یا به عبارت دیگر پوچ باشد را با $\nu(B)$ نشان می‌دهند و آن را تابع احتمال پوچی[۱۲] فرآیند نقطه‌ای X می‌نامند. در واقع تابع احتمال پوچی به صورت زیر قابل تعریف است:

$$\nu(B) = P\{N(B) = 0\}\ ,\ B \subset S.$$

باید توجه داشت که فرآیند نقطه‌ای X و تابع شمارش $N(\cdot)$ به‌طور یکتا یکدیگر را مشخص‌سازی[۱۳] می‌کنند. همچنین توزیع فرآیند نقطه‌ای X به‌صورت یکتا توسط تابع احتمال پوچی مشخص می‌شود.

[۱۲] Void probability function
[۱۳] Characterization

به کمک گشتاورهای تابع شمارش $N(\cdot)$ می‌توان گشتاورهای فرآیند نقطه‌ای X را تعریف کرد. به‌عنوان مثال، شدت یک ویژگی توصیفی اولیه یک فرآیند نقطه‌ای است. تابع شدت[۱۴] میزان غلظت بین داده‌ها را در واحد حجم یا مساحت نشان می‌دهد. بررسی شدت یک الگوی نقطه‌ای یکی از اولین و مهم‌ترین مراحل در تحلیل داده‌ها است که می‌توان آن را با استفاده از گشتاورهای تابع شمارش $N(\cdot)$ تعریف کرد.

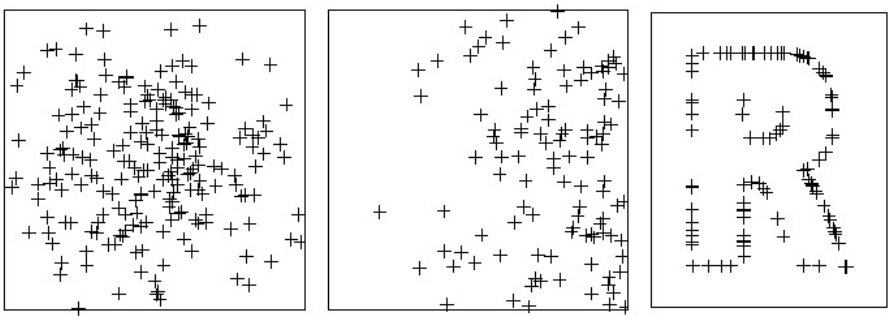
۱.۴.۱ تابع شدت

تابع $\lambda : S \longrightarrow [۰, \infty)$ را تابع شدت فرآیند نقطه‌ای X گویند، هرگاه برای هر $B \subset S$،

$$E[N(B)] = \int_B \lambda(u)du.$$

تابع شدت مشخصه مهمی در فرآیندهای نقطه‌ای است. هر چقدر $\lambda(u)$ بزرگ‌تر باشد، احتمال وجود نقاط در همسایگی‌های کوچک u برای فرآیند نقطه‌ای X افزایش می‌یابد. اگر λ تابعی ثابت باشد، فرآیند نقطه‌ای X را همگن[۱۵] (مانای مرتبه اول) با شدت λ و در غیر این‌صورت آن را ناهمگن[۱۶] گویند. در یک فرآیند ناهمگن بر S نقاط به‌صورت ناهمگون پراکنده می‌شوند، به‌طوری‌که ممکن است حتی در یک زیرناحیه از S هیچ نقطه‌ای وجود نداشته باشد.

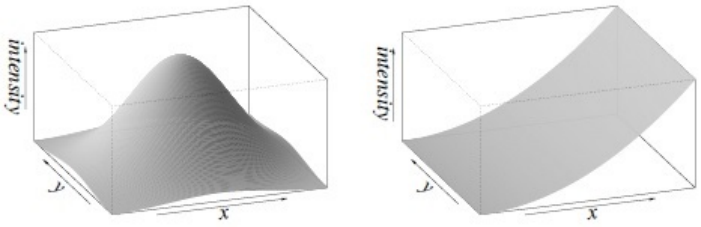
نمونه‌هایی از الگوهای ناهمگن در شکل ۱۰.۱ نشان داده شده‌اند.



شکل ۱۰.۱: الگوهای نقطه‌ای شبیه‌سازی‌شده با شدت‌های مختلف. چپ: شدت پخش، بیشتر در مرکز طرح است. میانه: شدت پخش، بیشتر در لبه سمت راست طرح است و راست: شدت منحصر به فرد بر لبه‌های حرف R است.

شکل ۱۱.۱ تابع شدت متناظر با الگوهای نقطه‌ای شبیه‌سازی‌شده سمت چپ و میانه در شکل ۱۰.۱ را نشان می‌دهد. مهمترین مسئله علمی در تحلیل الگوهای نقطه‌ای این است که آیا تابع شدت همگن است یا خیر؟ برای پاسخ به این سوال روش‌های مختلفی وجود دارند.

[۱۴] Intensity function
[۱۵] Homogeneous
[۱۶] Heterogeneous

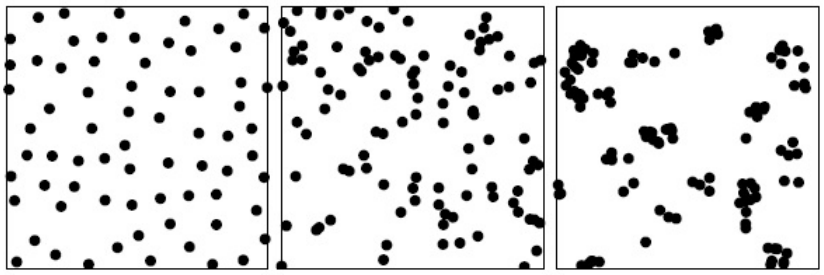


شکل ۱۱.۱: تابع شدت متناظر با الگوهای نقطه‌ای شبیه‌سازی‌شده سمت چپ و میانه در شکل ۱۰.۱

تابع شدت ناهمگن را می‌توان به‌صورت ناپارامتری با استفاده از روش‌های شمارش پنجره‌های مربعی[۱۷] (بدلی، ۲۰۱۵) و برآورد هسته[۱۸](بیتل، ۱۹۹۰ و دیگل، ۱۹۸۵) برآورد کرد.

۲.۴.۱ تابع K

بیشترین انگیزه برای تحلیل داده‌های الگوی نقطه‌ای برای پاسخ به این سوال است که آیا نقاط مستقل از یکدیگر قرار گرفته‌اند یا نوعی وابستگی متقابل دارند. شکل ۱۲.۱ سه الگوی نقطه‌ای «منظم[۱۹]» (که در آن نقطه‌ها تمایل به اجتناب از یکدیگر دارند)، «مستقل[۲۰]» (تصادفی فضایی کامل[۲۱]) و «خوشه‌ای[۲۲]» (که در آن نقاط تمایل به همسایگی نزدیک دارند) را نشان می‌دهد.



شکل ۱۲.۱: الگوهای منظم (چپ)، مستقل (میانه) و خوشه‌ای (راست)

به‌طور کلی روش‌های شناخت الگوی توزیع نقاط به دو رده چگالی محور[۲۳] و فاصله‌ای

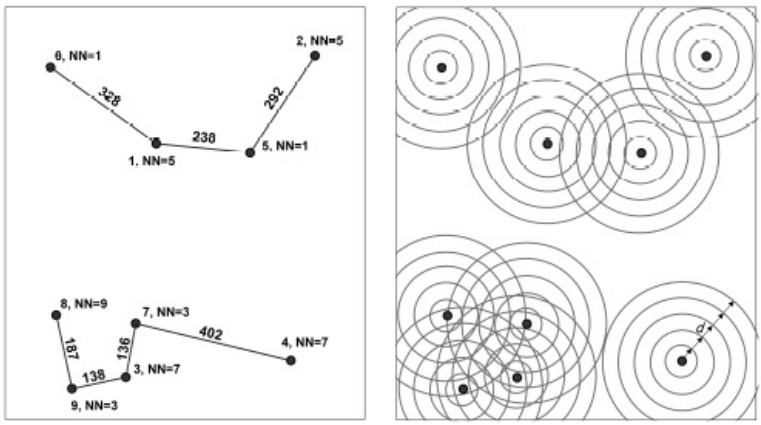
[۱۷] Quadrat counts
[۱۸] Kernel estimation
[۱۹] Regularity
[۲۰] Independence
[۲۱] Complete spatial randomness
[۲۲] Clustering
[۲۳] Axis density

محور[۲۴]axis Distance دسته‌بندی می‌شوند. روش‌های شمارش مربعی و برآورد هسته چگالی، از روش‌های چگالی محور محسوب می‌شوند و نمونه‌هایی از روش‌های فاصله‌ای محور عبارتند از روش میانگین نزدیک‌ترین همسایه، تابع K[۲۵]، تابع G[۲۶] و تابع F[۲۷]. ما در این پایان‌نامه به تفسیر روش‌های میانگین نزدیک‌ترین همسایه و تابع K می‌پردازیم.

در روش میانگین نزدیکترین همسایه برای نقاط موجود در منطقه مورد مطالعه، فاصله هر نقطه با نزدیکترین همسایه آن محاسبه می‌گردد و مجموع این فواصل بر تعداد نقاط تقسیم می‌شود. چنانچه فاصله هر نقطه p_i با نزدیکترین همسایه‌اش را $d_{min}p_i$ بنامیم، آن‌گاه تابع فوق به شکل زیر خواهد بود:

$$\frac{\sum_{i=1}^{n} d_{min}p_i}{n}$$

که در آن n تعداد نقاط موجود در منطقه مورد مطالعه است (شکل ۱۳.۱ سمت چپ).



شکل ۱۳.۱: سمت چپ) محاسبه فاصله نزدیکترین همسایه و سمت راست) مبنای محاسبه تابع K

از معایب این تابع، دور ریختن مقادیر زیادی از اطلاعات مرتبط با الگو است، زیرا خروجی این تابع تنها یک عدد خواهد بود و قضاوتی کلی را در خصوص پراکندگی نقاط در منطقه مورد مطالعه ارائه می‌دهد. یک روش بسیار محبوب برای تحلیل همبستگی فضایی در الگوهای نقطه‌ای، تابع K است که اولین بار توسط ریپلی (۱۹۸۱) معرفی گردید.

فرض کنید سوال اصلی تحقیق مربوط به فاصله بین نقاط در الگوی نقطه است. پس محاسبه فاصله $d_{ij} = ||x_i - x_j||$ بین تمام جفت‌های مشخص نقاط متمایز x_i و x_j در الگوی نقطه‌ای X طبیعی خواهد بود.

۲۴

۲۵K-function

۲۶G-function

۲۷F-function

این تابع بر پایه شمارش نقاط موجود در تمامی فواصل موجود میان نقاط قرار دارد. بنابراین مانند روش نزدیک‌ترین همسایه از می‌نیمم فاصله استفاده نمی‌شود. همان‌طور که در شکل ۱۳.۱ دیده می‌شود، دامنه‌های فاصله از قبل تعیین شده و به‌صورت حریم‌هایی در اطراف هر یک از نقاط ترسیم شده است. شمارش تعداد نقاطی که در هر یک از حریم‌ها قرار می‌گیرد، مبنای محاسبه تابع K خواهد بود. ریپلی تابع K را بر اساس رابطه زیر تعریف کرد:

$$K(d) = \frac{A}{N^{۲}} \sum_{i} \sum_{j} I(d_{ij}) \qquad (۱.۱)$$

که در آن $K(d)$ مقدار تابع K در فاصله d است و اگر فاصله نقطه i تا j کمتر از مقدار d باشد، $I(d_{ij}) = ۱$ خواهد بود و در غیر این صورت $I(d_{ij}) = ۰$. همچنین A مساحت منطقه و N تعداد کل نقاط موجود در منطقه مورد مطالعه است.

تابع K برای یک فرآیند نقطه‌ای X به‌عنوان تعداد مورد انتظار rـ همسایگی یک نقطه معمولی X تعریف می‌شود. برای این کار باید فرض کنیم X ثابت است. به این معنا که به ازای بردار v توزیع X با توزیع فرآیند تغییر یافته $X + v$ یکسان است یعنی X دارای شدت همگن λ است. بنابراین برای هر $r > ۰$ و هر مکان u تعریف می‌کنیم:

$$K(r) = \frac{۱}{\lambda} E[u \text{ دارای یک نقطه در موقعیت} | \text{تعداد r همسایگی } X] \qquad (۲.۱)$$

از آن‌جا که فرآیند ثابت است، این تعریف به موقعیت u بستگی ندارد. در تعریف ۲.۱ نماد « | » نشان می‌دهد که این انتظارات شرطی هستند. به‌طور مستقیم، فرض می‌کنیم یک نقطه تصادفی از X در مکان u وجود دارد؛ با توجه به این، تعداد قابل توجهی از نقاط دیگر X را در فاصله r قرار می‌دهیم و در نهایت بر شدت λ تقسیم می‌کنیم تا $K(r)$ را به‌دست آوریم. برای هر موقعیت فضایی تعریف می‌کنیم

$$t(u, r, X) = \sum_{j=۱}^{n(X)} I\{۰ < ||u - x_j|| \leq r\}. \qquad (۳.۱)$$

رابطه (۳.۱) تعداد نقاط الگوی نقطه‌ای X را که در فاصله r از مکان u قرار دارند و شامل خود u نمی‌شوند، محاسبه می‌کند.

تعریف ۳.۴.۱. اگر X یک فرآیند نقطه‌ای ثابت با شدت $\lambda > ۰$ باشد، برای هر $r \geq ۰$

$$K(r) = \frac{۱}{\lambda} E[t(u, r, X) | u \in X]$$

تابع K فرآیند نقطه‌ای X نامیده می‌شود. این تابع به موقعیت u بستگی ندارد.

برای چند مدل فرآیند نقطه‌ای فرمول‌های صریح برای تابع K استخراج شده‌اند. به‌عنوان نمونه، برای یک فرآیند نقطه‌ای پواسون تصادفی کامل، از آن‌جا که نقاط مستقل هستند، به‌طور

شهودی حضور یک نقطه تصادفی در موقعیت u هیچ تاثیری بر حضور نقاط در مکان‌های دیگر ندارد. بنابراین

$$E[t(u,r,X)|u \in X] = E[t(u,r,X)].$$

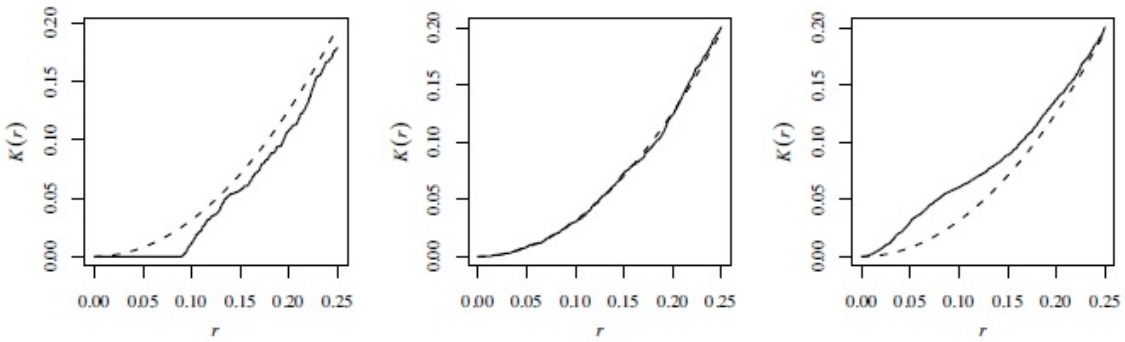
اما $t(u,r,X)$ تعداد نقاط X است که در فاصله دایره $b(u,r)$ به شعاع r و مرکز u قرار دارند و $|b(u,r)|$ مساحت دایره به شعاع r و مرکز u است. تعداد مورد انتظار چنین نقاطی

$$\lambda \times |b(u,r)| = \lambda \pi r^2$$

است که با تقسیم کردن بر λ برای هر فرآیند پوآسون همگن به‌صورت $K_{pois}(r) = \pi r^2$ نتیجه می‌شود. دقت کنید این محاسبه برای $\mathbb{R}^2$ معتبر است.

۳.۴.۱ تابع K تجربی

تابع تجربی $\hat{K}(r)$ خلاصه‌ای از فاصله‌های جفتی در مجموعه داده‌های الگوی نقطه‌ای است که به ما امکان مقایسه مقادیر مختلف داده‌ها را می‌دهد. اما سوال کلیدی در مورد هر آمار خلاصه برای یک الگوی نقطه‌ای به این معنا است که فرآیند نقطه‌ای، کدام الگوی نقطه‌ای را تولید می‌کند. در تحلیل ابتدایی یک مجموعه داده الگوی نقطه‌ای، فرض بر همگنی تابع شدت است. برای بررسی این همگنی، می‌توان تابع K تجربی محاسبه شده از داده‌ها، یعنی $\hat{K}(r)$ را با تابع K نظری فرآیند پواسون همگن یعنی $K_{pois}(r) = \pi r^2$ که به‌عنوان معیار «عدم همبستگی» عمل می‌کند، مقایسه کرد.

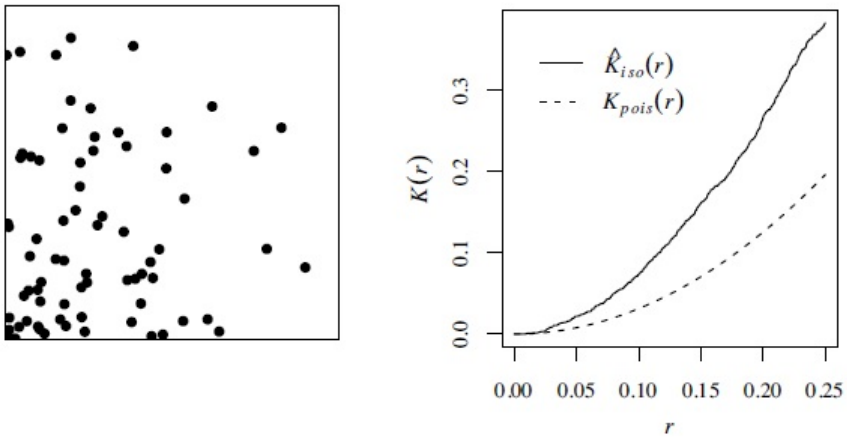


شکل ۱۴.۱: تابع تجربی K (منحنی خط پیوسته) برای هر یک از سه الگو در شکل ۱۲.۱ و تابع K نظری برای یک فرآیند پوآسون (منحنی خط‌چین).

شکل ۱۴.۱ به وضوح وجود و نوع همبستگی بین نقاط در یک الگوی نقطه‌ای را نشان می‌دهد. در تصویر سمت چپ، منحنی برای تابع K تجربی (منحنی خط پیوسته) کمتر از منحنی نظری برای الگوی کاملا تصادفی (منحنی خط‌چین) است، $\hat{K}(r) < K_{pois}(r)$، که نشان می‌دهد نقاط دارای همسایه‌های کمتری نسبت به الگوی مورد انتظار است. بنابراین

می‌توان گفت این الگو همسو با الگوی منظم است. به‌طور مشابه، در تصویر سمت راست منحنی تجربی بالاتر از منحنی نظری است، $\hat{K}(r) > K_{pois}(r)$، که نشان می‌دهد نقاط نسبت به الگوی مورد انتظار دارای همسایه‌های بیشتری است. بنابراین الگو سازگار با یک الگوی نقطه‌ای خوشه‌ای است. تابع K یک ابزار قدرتمند برای بررسی الگوهای نقطه‌ای است، اما باید به خاطر داشت که «همبستگی علیت[۲۸] نیست». برای توضیح بیشتر، اگر تحلیل نشان دهد $\hat{K}(r) > K_{pois}(r)$، محقق نمی‌تواند دقیق بگوید که این نشان دهنده یک الگوی خوشه‌ای است، اما آن را «سازگار» با خوشه‌بندی می‌داند یا نشان می‌دهد که بین نقاط «ارتباط مثبت» وجود دارد.

تابع K بر اساس این پذیره که فرآیند نقطه‌ای ثابت است، تعریف و محاسبه می‌شود. اگر فرآیند ثابت نباشد، اختلاف توابع $\hat{K}(r)$ تجربی و K_{pois} نظری لزوما شهودی بر تعامل نقطه‌ای نیستند، زیرا ممکن است مرتبط با تغییر شدت باشند.



شکل ۱۵.۱: شکل چپ) الگوی نقطه‌ای پواسون ناهمگن و شکل راست) تابع K تجربی

شکل ۱۵.۱ تحقق یک فرآیند پواسون ناهمگن و تابع K تجربی متناظر آن را نشان می‌دهد. طرح تابع K بیان‌گر آن است که به‌طور متوسط یک نقطه در این الگو دارای همسایگی بیشتری نسبت به الگوی تصادفی کامل مورد انتظار است. با این حال، این اتفاق رخ می‌دهد زیرا تراکم بیشتری از نقاط در یک گوشه پنجره وجود دارد. در واقع، نقاط به‌طور مثبت مرتبط هستند، اما نه به این دلیل که آن‌ها به‌طور معمول خوشه‌ای هستند.

۴.۴.۱ اثر مرزی

اثر مرزی[۲۹] در تحلیل الگوهای نقطه‌ای فضایی، زمانی مطرح می‌شود که ناحیه W که الگوی نقطه‌ای روی آن مشاهده شده است، قسمتی از یک ناحیه بزرگ باشد که فرآیند اصلی روی آن

[۲۸] causation
[۲۹] Edge effect

توزیع شده است. برای اهداف استنباطی باید به دو حالت زیر توجه کنیم:

الف) داده‌های X تحققی از یک فرآیند نقطه‌ای متناهی هستند که فقط درون W تعریف شده‌اند (حالت کراندار).

ب) داده‌های X یک تحقق مشاهده شده جزئی $y \cap W$ از یک فرآیند نقطه‌ای Y هستند که در کل ناحیه $\mathbb{R}^d$ توسعه یافته است و تنها در پنجره W مشاهده شده است (حالت غیر کراندار).

تصحیح مرزی

تصحیح مرزی[۳۰] یک استراتژی برای حذف اثر مرزی است که با یک طرف کشیده (اریب) می‌شود. یکی از راه‌کارهای ساده، روش لبه[۳۱] است. هنگام برآورد $K(r)$ برای یک فاصله مشخص r، توجه خود را به مواردی که دایره به شعاع r به‌طور کامل در داخل پنجره قرار دارد، محدود می‌کنیم تا اثر مرزی رخ ندهد. بنابراین در شمارش و محاسبه تابع K، توجه خود را به نقاطی محدود می‌کنیم که برای آن‌ها $b(x_i, r)$ به طور کامل در داخل W قرار بگیرد. برای چنین نقاطی $t(x_i, r, X \cap W) = t(x_i, r, X)$. بنابراین تعداد واقعی r همسایگی قابل مشاهده است. شکل ۱۶.۱ این نوع تصحیح را نشان می‌دهد.

شکل ۱۶.۱: روش مرزی اصلاح لبه برای محاسبه تابع K.

در شکل ۱۶.۱ هنگام برآورد $K(r)$، فاصله‌های d_{ij} فقط برای نقاط x_iای محاسبه می‌شوند که حداقل r واحد دور از مرز هستند. این نقاط آن‌هایی هستند که در قسمت خاکستری قرار دارند.

[۳۰] Border correction
[۳۱] Border

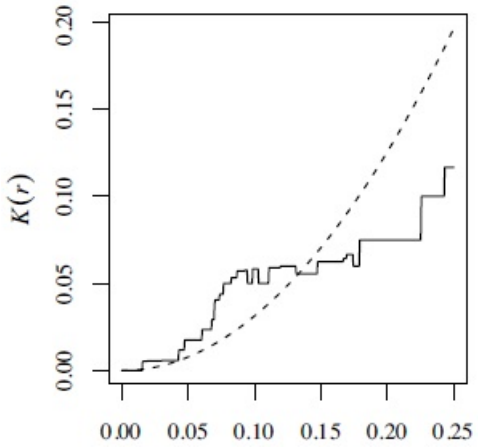
اگر $d(u,\partial W)$ نشان دهنده کوتاه‌ترین فاصله از یک مکان u به مرز پنجره W باشد، توجه خود را به نقاط x_i محدود می‌کنیم که $d(u,\partial W) \geq r$. این نقاط داده‌هایی هستند که در مجموعه نقاط فرسایش‌شده[۳۲] قرار می‌گیرند. یعنی

$$W_{\ominus r} = W_{\ominus} b(\circ, r) = \{u \in W \ : \ d(u,\partial W) \geq r\}.$$

مجموعه $W_{\ominus r}$ متشکل از مکان‌هایی در W هستند که حداقل r واحد از مرز ∂W دور هستند. مجموعه فرسایش شده در شکل ۱۶.۱ قابل مشاهده است. بنابراین، تابع $K(r)$ با اصلاح مرز به‌صورت زیر برآورد می‌شود:

$$\hat{K}_{bord}(r) = \frac{\sum_{x_i \in X \cap W_{\ominus r}} t(x_i, r, X)}{\bar{\lambda} n(X \cap W_{\ominus r})} = \frac{\sum_{i=1}^{n} I\{b_i \geq r\} t(x_i, r, X)}{\bar{\lambda} \sum_{i=1}^{n} I\{b_i \geq r\}} \qquad (۴.۱)$$

که در آن $b_i = d(x_i, \partial W)$ فاصله داده x_i با مرز پنجره است و $\bar{\lambda}$ برآورد شدت است که معمولا $\bar{\lambda} = n(x \cap W)/|W|$. برآوردگر (۴.۱) طوری تعریف شده است که مخرج صفر نشود، یعنی تا زمانی که $r < max_i b_i$. برآورد اصلاح مرزی تابع K نسبتا ساده است. با این حال، در مجموعه داده‌های کوچک، در مقایسه با روش های دیگر می تواند نادرست باشد، زیرا تعداد زیادی از داده‌ها را از بین می‌برد.



شکل ۱۷.۱: برآورد اصلاح مرزی تابع K (منحنی خط کامل) برای ۲۰ نقطه تصادفی یکنواخت در مربع واحد.

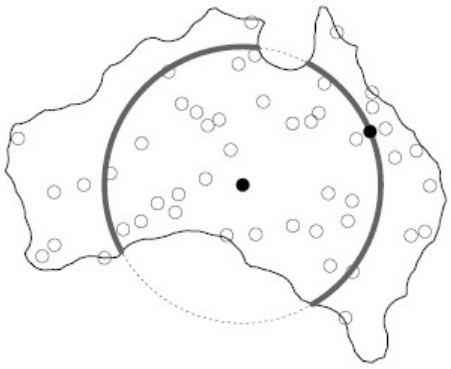
به‌عنوان مثال اگر W مربع واحد باشد و $r = ۰/۱$، پنجره فرسایش شده $W_{\ominus r}$ یک مربع به طول ۰/۸ و مساحت ۰/۶۴ است که تقریبا ۴۰ درصد از داده‌ها برای تخمین $K(۰/۱)$ از بین می‌روند. در مجموعه داده‌های کوچک یا در شرایط غیر معمول، نمودار تابع $\hat{K}_{bord}(r)$ می‌تواند یک برآورد ناپایدار داشته باشد. همان‌طور که در شکل ۱۷.۱ نشان داده شده است، برای مقادیر بزرگ r، مخرج رابطه (۴.۱) تابع K باید یک تابع افزایشی از r باشد، اما برای

[۳۲] eroded set

برآورد اصلاح مرزی نیازی نیست. وقتی تعداد داده (نقاط) زیاد است، روش لبه معمولا به روش‌های با محاسبات بیشتر ترجیح داده می‌شود، زیرا فاصله‌های r کوچک هستند و از دست دادن کارایی آماری در روش لبه ناچیز است. روش لبه یک درمان مفید برای اثرات مرزی در بسیاری از مسائل فضایی است که به «اصل دانش محلی[۳۳] » مورفولوژی ریاضی مرتبط است (بدلی و همکاران، ۲۰۱۵).

تصحیح همسان‌گرد

یک رویکرد جایگزین برای اثر مرزی، تجسم کل فرآیند نقطه‌ای X است که در فضای دو بعدی وجود دارد. یک جفت نقطه x و x' از X را در نظر بگیرید و فرض کنید x در پنجره مشاهده قرار می‌گیرد. با توجه به موقعیت اولیه نقطه x و فاصله $d = ||x - x'||$، نقطه دوم x' باید در دایره $b(x, d)$ به شعاع d و مرکز x قرار گیرد. به‌طور کلی همسان‌گرد، تنها بخشی از این دایره در داخل پنجره قرار دارد شکل ۱۸.۱. اگر فرآیند نقطه‌ای نیز همسان‌گرد باشد (تحت چرخش، خواص یکسان داشته باشد)، پس احتمالا نقطه دوم x' به‌طور مساوی در هر نقطه از این دایره قرار دارد.



شکل ۱۸.۱: نمایش تصحیح همسان‌گرد

احتمال این‌که x در داخل W قرار گیرد،

$$p(x, d) = \frac{\ell(W \cap \partial b(x, d))}{۲\pi d}$$

تعریف می‌شود، که در آن ℓ نشان‌دهنده طول است. یک استراتژی استاندارد برای اصلاح

[۳۳] local knowledge principle

نمونه‌گیری اریب، برآوردگر هورویتز تامپسون[۳۴] است، که به صورت زیر تعریف می‌شود:

$$\hat{K}_{iso}(r) = \frac{۱}{\bar{\lambda} n} \sum_{i=۱}^{n} \sum_{j=۱, j \neq i}^{n} I\{d_{ij} \leq r\} \frac{۱}{p(x_i, d_{ij})}. \qquad (۵.۱)$$

این استراتژی را می‌توان برای برآوردگرهای تابع K استفاده کرد. اگر هر جفت نقطه x_i و x_j با احتمال $p(x_i, d_{ij})$ وزن‌دار شود، به برآوردگر تصحیح همسان‌گرد[۳۵] متنهی می‌شود که توسط ریپلی (۱۹۸۱) معرفی شد. برآوردگر تصحیح همسان‌گرد بر خلاف برآوردگر تصحیح مرزی یک تابع غیر کاهشی از r است.

سایر روش‌های تصحیح مرز نیز وجود دارند که می‌توان به تصحیح انتقال[۳۶] (بدلی و همکاران، ۲۰۱۵) اشاره کرد.

۵.۴.۱ تابع همبستگی زوجی

تعریف ۴.۴.۱. تابع $g : S \times S \longrightarrow [۰, \infty)$ را تابع همبستگی زوجی[۳۷] فرآیند نقطه‌ای X گویند، هرگاه به ازای هر $A, B \subset S$ که $A \cap B = \varnothing$،

$$Cov(N(A), N(B)) = \int_A \int_B \lambda(u)\lambda(v)[g(u, v) - ۱]dudv.$$

تابع همبستگی زوجی کاربردهای فراوانی در علوم اخترشناسی و اخترفیزیک دارد و در تحلیل اکتشافی الگوهای نقطه‌ای ابزار بسیار مهمی تلقی می‌شود. برای تفسیر g نیاز به درک شهودی آن می‌باشد. فرض کنید u و v دو نقطه در S و A_u و A_v همسایگی‌های کوچک u و v به مساحت d_u و d_v باشند. در این صورت

$$g(u, v) = ۱ + \frac{Cov(N(A_u), N(A_v))}{\lambda(u)\lambda(v)d_u d_v} \approx \frac{E[N(A_u), N(A_v)]}{\lambda(u)\lambda(v)d_u d_v} \approx \frac{P\{N(A_u) > ۰, N(A_v) > ۰\}}{P\{N(A_u) > ۰\}P\{N(A_v) > ۰\}}.$$

با توجه به این تقریب می‌توان گفت $g(u, v)$ تقریبا با احتمال این‌که فرآیند نقطه‌ای X به صورت هم‌زمان نقاطی در نزدیکی u و v داشته باشد، تقسیم بر حاصل‌ضرب احتمال‌های این‌که فرآیند نقطه‌ای X یک نقطه در نزدیکی u و یک نقطه در نزدیکی v داشته باشد، برابر است. بنابراین، می‌توان موارد زیر را به $g(u, v)$ نسبت داد:

- اگر $g(u, v) \approx ۱$، برای همسایگی‌های کوچکی از u و v

$$P\{N(A_u) > ۰, N(A_v) > ۰\} \approx P\{N(A_u) > ۰\}P\{N(A_v) > ۰\}.$$

[۳۴] Horvitz Thompson
[۳۵] Isotropic correction
[۳۶] Translation correction
[۳۷] Pair correlation function

یعنی نقاطی از فرآیند X که در نزدیکی u هستند با نقاطی از X که در نزدیکی v هستند، اثر متقابل ندارند. در واقع پیشامدهای $\{N(A_u) > ۰\}$ و $\{N(A_v) > ۰\}$ مستقل هستند.

- اگر $g(u,v) > ۱$، برای همسایگی‌های کوچکی از u و v

$$P\{N(A_u) > ۰, N(A_v) > ۰\} > P\{N(A_u) > ۰\}P\{N(A_v) > ۰\}.$$

یعنی علاوه بر عدم استقلال دو پیشامد $\{N(A_u) > ۰\}$ و $\{N(A_v) > ۰\}$، احتمال توام آن‌ها بزرگ‌تر از حاصل‌ضرب احتمال‌های آن‌ها است. این بدان معنی است که نقاطی از فرآیند X که در نزدیکی u هستند با نقاطی از X که در نزدیکی v هستند، اثر متقابل جذبی[۳۸] دارند. به‌عبارت دیگر نقاط مذکور متمایل‌اند در مجاورت یکدیگر باشند و تشکیل یک ساختار خوشه‌ای بدهند.

- اگر $g(u,v) < ۱$، برای همسایگی‌های کوچکی از u و v

$$P\{N(A_u) > ۰, N(A_v) > ۰\} < P\{N(A_u) > ۰\}P\{N(A_v) > ۰\}.$$

یعنی علاوه بر عدم استقلال دو پیشامد $\{N(A_u) > ۰\}$ و $\{N(A_v) > ۰\}$، احتمال توام آن‌ها کوچک‌تر از حاصل‌ضرب احتمال‌های آن‌ها است. این بدان معنی است که نقاطی از فرایند X که در نزدیکی u هستند با نقاطی از X که در نزدیکی v هستند، اثر متقابل دفعی[۳۹] دارند. به عبارت دیگر نقاط مذکور تمایلی به قرار گرفتن در مجاورت هم از خود نشان نمی‌دهند.

تابع همبستگی زوجی ابزار مناسبی برای سنجش همبستگی بین نقاط فرآیند است، زیرا با توجه به ملاحظات بالا $g(u,v)$ نوع اثر متقابل هر دو نقطه u و v از فرآیند را مشخص می‌کند.

۶.۴.۱ مانایی و همسان‌گردی فرآیندهای نقطه‌ای

در بسیاری از روش‌های آماری تحلیل الگوهای نقطه‌ای، مانایی[۴۰] و همسان‌گردی شرط لازم و اساسی است. به‌طور کلی، هرگاه توزیع فرآیند نقطه‌ای نسبت به انتقال پایا باشد آن را فرآیند نقطه‌ای مانا گویند و هرگاه نسبت به دوران پایا باشد، فرآیند نقطه‌ای همسان‌گرد است.

تعریف ۵.۴.۱. (مولر و واگه‌پترسن، ۲۰۰۴). فرآیند نقطه‌ای X را در $S \subset \mathbb{R}^۲$ مانا می‌نامند، هرگاه برای هر بردار ثابت $u \in S$، $X + u \overset{d}{=} X$ که در آن $X + u = \{v + u; v \in X\}$.

[۳۸] Absorption Interaction
[۳۹] Interactive interaction
[۴۰] Stationary

تعریف ۶.۴.۱. (مولر و واگه‌پترسن، ۲۰۰۴) فرآیند نقطه‌ای X را همسان‌گرد گویند، هرگاه برای هر دوران حول مبدا مانند O، $OX \overset{d}{=} X$ که در آن $OX = \{Ou; u \in X\}$. منظور از دوران حول مبدا ماتریس دوران زیر است که در آن $\theta \in [0, 2\pi)$ زاویه دوران را مشخص می‌کند:

$$O = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}.$$

از تعریف مانایی می‌توان نتیجه گرفت که همه مشخصه‌های یک فرآیند نقطه‌ای مانا نسبت به انتقال یکسان باقی می‌ماند. یعنی تابع شدت یک فرآیند نقطه‌ای مانا با توجه به این‌که برای هر $u, v \in \mathbb{R}^2$، $\lambda(u) = \lambda(u + v)$، یک تابع ثابت است. پس یک فرآیند مانا همواره همگن است. اما عکس این مطلب در حالت کلی برقرار نیست، یعنی همگن بودن فرآیند لزوما مانایی را نتیجه نمی‌دهد. علاوه بر مشخصه تابع شدت، تابع همبستگی زوجی نیز نسبت به انتقال پایاست. یعنی برای هر $u, v \in \mathbb{R}^2$، $g(u + \varepsilon, v + \varepsilon) = g(u, v)$. پس تابع همبستگی زوجی یک فرآیند نقطه‌ای مانا تنها تابعی از اختلاف بین نقاط است و تنها به $u - v$ بستگی دارد.

با توجه به تعریف همسان‌گردی، اگر فرآیند نقطه‌ای علاوه بر مانایی همسان‌گرد نیز باشد، همه مشخصه‌های آن نسبت به دوران پایاست. به‌عنوان مثال، تابع همبستگی زوجی یک فرآیند مانا و همسان‌گرد تنها تابعی از فاصله بین دو نقطه است. بنابراین برای یک فرآیند نقطه‌ای مانا و همسان‌گرد، تابع شدت یک مقدار ثابت $\lambda > 0$ و تابع همبستگی زوجی تابعی یک متغیره $g(r)$ است و قطعا کار با آن‌ها ساده‌تر از کار با تابع دو متغیره یا تابع چهار متغیره $g(u, v)$ است.

۵.۱ استنباط بیزی

در رویکرد کلاسیک پارامترهای تابع همبستگی زوجی و تابع شدت ثابت اما مجهول در نظر گرفته می‌شوند و برآورد آن‌ها باید بر اساس مشاهدات X_S انجام شود. علاوه بر این، استنباط کلاسیک برای برآورد پارامترها مبتنی بر رهیافت درستنمایی است. با توجه به بسته نبودن شکل درستنمایی بعضی از فرآیندها، استفاده از این روش کار مشکل و پیچیده‌ای است (ایلیان و همکاران، ۲۰۱۲). در مقابل در رویکرد بیزی فرض می‌شود پارامترهای تابع همبستگی زوجی و تابع شدت تحقق‌هایی از توزیع‌های احتمالی با نام توزیع‌های پیشین هستند. با این نگاه می‌توان میدان تصادفی ζ را نیز به‌عنوان یک پارامتر با توزیع پیشین گاوسی در نظر گرفت. در این رویکرد تمام استنباط‌ها شامل برآورد پارامترها و پیش‌گویی میدان تصادفی ζ مبتنی بر توزیع پسین مدل ساخته می‌شوند. بنابراین با توجه به این‌که میدان تصادفی ζ یک میدان پنهان به حساب می‌آید و پیش‌بینی آن بر اساس مشاهدات X با کمیت $E\Big(\zeta(s)\Big|X\Big)$، $s \subseteq S$ انجام می‌شود، استفاده از رویکرد بیزی برای استنباط آماری فرآیندهای LGC مناسب‌تر است.

از آن‌جایی که محاسبه توزیع پسین بر حسب تابع درستنمایی برخی از فرآیندها کار

پیچیده‌ای است و با انتخاب پیشین مناسب ممکن است پیچیده‌تر شود، نیاز به روش‌های تقریبی توزیع پسین اجتناب‌ناپذیر است. روش‌های عددی و روش‌های مبتنی بر نمونه‌گیری، دو رده کلی برای تقریب توزیع پسین هستند. از معروف‌ترین روش‌های عددی می‌توان به تقریب نیوتون – رافسون، روش‌های تربیع‌بندی گاوسی[۴۱] (آبراموویتز و استگن، ۱۹۷۲) و تقریب لاپلاس[۴۲]اشاره کرد و در بین روش‌های مبتنی بر نمونه‌گیری (روش‌های شبیه‌سازی مونت کارلویی) روش‌های نمونه‌گیری رد و پذیرش[۴۳] (کسلا و برگر، ۲۰۰۲)، نمونه‌گیری نقاط مهم[۴۴] (هستربرگ، ۱۹۸۷؛ رابرت و کسلا، ۲۰۰۴) و الگوریتم‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی[۴۵] (MCMC) (هستینگز، ۱۹۷۰؛ رابرت و کسلا، ۲۰۰۴) را نام برد.

روش‌های عددی زمانی که بعد توزیع پسین کمتر از ۱۰ باشد، خطای قابل قبولی دارند اما در ابعاد بالا استفاده از این روش‌ها به علت بزرگ بودن خطای گرد کردن پیشنهاد نمی‌شود. در این موارد رهیافت جایگزین استفاده از روش‌های مبتنی بر نمونه‌گیری است. با وجود سرعت کند محاسباتی روش‌های نمونه‌گیری MCMC نسبت به روش‌های عددی، از آن‌جایی که خطای تقریب تابعی از بعد توزیع پسین نیست و به ازای هر بعدی ثابت و از مرتبه $O(n)^{-\frac{1}{4}}$ است استفاده از آن‌ها در ابعاد بالا برای تقریب توزیع پسین توصیه می‌شود. در روش‌های مبتنی بر نمونه‌گیری دو قضیه قانون قوی اعداد بزرگ و قضیه ارگودیک به ترتیب برای نمونه‌های تولید شده مستقل و وابسته (زنجیر مارکوفی) برای تضمین همگرایی برآوردگرها به کمیت مورد نظر از توزیع پسین استفاده می‌شوند. در این پایان‌نامه در بین روش‌های عددی، تقریب لاپلاس را که بیان آن برای تشریح تقریب لاپلاس آشیانه‌ای جمع‌بسته[۴۶] (INLA) ضروری است و در روش‌های نمونه‌گیری، الگوریتم‌های مونت کارلوی زنجیر مارکوفی را معرفی می‌کنیم.

۱.۵.۱ تقریب لاپلاس

تقریب لاپلاس برای هر تابع همواری که درون دامنه خود دارای نقطه ماکسیمم است، یک تقریب به کمک تابع توزیع نرمال پیشنهاد می‌کند. تقریب لاپلاس یک روش عددی برای تقریب انتگرال‌های پیچیده است (تیرنی و کدین، ۱۹۸۶). مثلا فرض کنید انتگرال مورد نظر به‌صورت زیر باشد:

$$\int L(x)dx = \int e^{\ln(L(x))}dx$$

اگر $x_0 = x_{max}$ نشان‌دهنده نقطه ماکسیمم تابع $L(x)$ باشد، با توجه به بسط سری تیلور

[۴۱] Gauss quadrature methods
[۴۲] Laplace approximation
[۴۳] Reject – accept sampling
[۴۴] Importance sampling
[۴۵] Markov chain Monte Carlo
[۴۶] Integrated Nested Laplace Approximation

تا مرتبه دوم می‌توان نوشت:

$$\ln L(X) \approx \ln L(x_\circ) + \underbrace{(x-x_\circ)\frac{\partial \ln L(x_\circ)}{\partial x}\Big|_{x_\circ=x_{max}}}_{\circ} + \frac{(x-x_\circ)^2}{2}\frac{\partial^2 lnL(x_\circ)}{\partial x^2}\Big|_{x_\circ=x_{max}}$$

$$\ln L(X) \approx \ln L(x_\circ) + \frac{(x-x_\circ)^2}{2}\frac{\partial^2 \ln L(x_\circ)}{\partial x^2}\Big|_{x_\circ=x_{max}}$$

بنابراین با جایگذاری این مقدار، انتگرال به صورت زیر تقریب زده می‌شود:

$$\begin{aligned}\int L(x)dx &\approx \int exp\Big[lnL(x_\circ) + \frac{(x-x_\circ)^2}{2}\frac{\partial^2 \ln L(x_\circ)}{\partial x^2}\Big|_{x_\circ=x_{max}}\Big]dx \\ &= \underbrace{e^{\ln L(x_{max})}}_{\text{مقدار ثابت}}\int \exp\Big[\frac{(x-x_\circ)^2}{2}\underbrace{\frac{\partial^2 \ln L(x_\circ)}{\partial x^2}\Big|_{x_\circ=x_{max}}}_{\text{مقدار ثابت}}\Big]dx\end{aligned}$$

حال اگر قرار دهیم $\sigma^2 = -(\frac{\partial^2 \ln L(x=x_{max})}{\partial x^2})^{-1}$ انتگرال تابع هسته توزیع نرمال با میانگین $x^* = x_{max}$ و واریانس σ^2 به صورت زیر به‌دست می‌آید:

$$\int L(x)dx = e^{\ln L(x^*)}\int \exp\Big[-\frac{(x-x^*)^2}{2\sigma^2}\Big]dx.$$

تقریب لاپلاس برای حالت چندمتغیره نیز قابل محاسبه است. در این حالت در بسط تیلور و انتگرال‌ها، کمیت‌ها به‌صورت بردار و ماتریس (دترمینان ماتریس) خواهند بود. جزییات بیشتر در کاس و استیفی (۱۹۸۹) قابل مشاهده است.

۲.۵.۱ الگوریتم‌های مونت کارلوی زنجیر مارکوفی

یکی از رده‌های تقریب توزیع پسین، روش‌های شبیه‌سازی مونت کارلویی است. این روش‌ها نمونه‌هایی از یک توزیع احتمالی مفروض تولید می‌کنند، اما زمانی که توزیع‌ها پیچیده باشند یا بعد توزیع پسین (بعد پارامترها) بالا باشد، تولید نمونه به‌صورت مستقیم مشکل خواهد بود. برای رفع این مشکل از روش‌های تولید نمونه به‌صورت غیرمستقیم مانند الگوریتم‌های MCMC استفاده می‌کنند.

روش‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی توسط متروپولیس و همکاران (۱۹۵۳) و هستینگز (۱۹۷۰) معرفی شدند. این روش‌ها یک چارچوب کلی از مجموعه روش‌هایی است که برای انتگرال‌گیری به روش مونت کارلوی مورد استفاده قرار می‌گیرند. برای درک بهتر و معرفی روش MCMC نیاز به معرفی زنجیره مارکوف و برخی از ویژگی‌های آن‌ها داریم. مطالب این بخش برگرفته از رابرتز و کسلا (۲۰۰۴) است.

زنجیر مارکوف

دنباله متغیرهای تصادفی $\{X^{(t)} : t \geq ۰\}$، که همه $X^{(t)}$ها در مجموعه شمارای S مقدار می‌گیرند، را «زنجیر مارکوف همگن» با فضای وضعیت S گویند، هرگاه در شرط زیر صدق کند:

$$\begin{aligned} P(X^{(t+۱)} = j | X^{(t)} = i, X^{(t-۱)} = i_{t-۱}, \cdots, X^{(۱)} = i_۱, X^{(۰)} = i_۰) \\ = P(X^{(t+۱)} = j | X^{(t)} = i) = p_{ij}. \end{aligned}$$

در این تعریف p_{ij} را احتمال انتقال یک‌مرحله‌ای از وضعیت i به وضعیت j می‌گویند، و ماتریس انتقال P را به‌صورت زیر تعریف می‌کنند:

$$P = [p_{ij}]_{i,j \in S}.$$

بدیهی است که

$$\forall i, j \in S : \; p_{ij} > ۰; \; \sum_{j \in S} p_{ij} = ۱$$

یعنی هر سطر ماتریس P یک احتمال شرطی بر فضای وضعیت S تعریف می‌کند. به‌طور مشابه اگر تعریف شود:

$$P^{(m)} = [p_{ij}^{(m)}]_{i,j \in S} \;\; p_{ij}^{(m)} = P(X^{(t+m)} = j | X^{(t)} = i)$$

آن‌گاه $p_{ij}^{(m)}$ را احتمال انتقال m مرحله‌ای از وضعیت i به وضعیت j و $P^{(m)}$ را ماتریس احتمال انتقال m مرحله‌ای زنجیره‌ی $\{X^{(t)}\}$ می‌گویند. از آن‌جایی که

$$\begin{aligned} p_{ij}^{(۲)} = P(X^{(۲)} = j | X^{(۰)} = i) &= \sum_{k \in S} P(X^{(۲)} = j, X^{(۱)} = k | X^{(۰)} = i) \\ &= \sum_{k \in S} P(X^{(۲)} = j | X^{(۱)} = k) P(X^{(۱)} = k | X^{(۰)} = i) \\ &= \sum_{k \in S} p_{ik} p_{kj} \end{aligned}$$

بنابراین $P^{(۲)} = P \times P$. به همین ترتیب می‌توان نشان داد

$$\begin{aligned} P^{(m)} &= \underbrace{P \times \cdots \times P}_{m} \\ P^{(m)} \times P^{(n)} &= P^{(m+n)}. \end{aligned} \tag{۶.۱}$$

روابط (۶.۱) به برابری‌های «چپمن- کلموگروف[۴۷]» مشهورند و بیان می‌کنند که با داشتن ماتریس احتمال انتقال یک‌مرحله‌ای، ماتریس انتقال چندمرحله‌ای را می‌توان به‌دست آورد.

[۴۷] Chapman-Kolmogorov

ساختار احتمالی زنجیر

فرض کنید $X^{(0)}$ وضعیت اولیه زنجیر باشد با داشتن توزیع $X^{(0)}$، توزیع تمام $X^{(t)}$ها وجود خواهند داشت. برای اثبات این واقعیت فرض کنید $X^{(0)}$ دارای توزیع $\Pi^{(0)}$ باشد. این توزیع را (برای سادگی نمادگذاری) به صورت زیر نمایش می‌دهیم:

$$\Pi_i^{(0)} = P(X^{(0)} = i)$$
$$\Pi^{(0)} = [\pi_i^{(0)}]_{i \in S}.$$

حال اگر توزیع $X^{(1)}$ با $\Pi^{(1)}$ نشان داده شود، در این صورت

$$\Pi^{(1)} = [\pi_j^{(1)}]_{j \in S}$$
$$\Pi_j^{(1)} = P(X^{(1)} = j) = \sum_{i \in S} P(X^{(1)} = j | X^{(0)} = i) P(X^{(0)} = i) = \sum_{i \in S} \pi_i p_{ij}.$$

با نمادگذاری ماتریسی می‌توان نوشت:

$$(\Pi^{(1)})^T = (\Pi^{(0)})^T \times P.$$

به‌طور مشابه، اگر توزیع $X^{(t)}$ با $\Pi^{(t)}$ نشان داده شود، روابط زیر برقرار هستند:

$$\Pi^{(t)} = [\pi_j^{(t)}]_{j \in S}$$
$$(\Pi^{(t)})^T = (\Pi^{(t)})^T \times P$$

که نشان می‌دهد توزیع $X^{(t)}$ تنها به ماتریس P و توزیع $\Pi^{(0)}$ وابسته است.

توزیع حدی زنجیر

همان‌طور که مشاهده شد، برای زنجیر مارکوف با ماتریس احتمال انتقال یک مرحله‌ای P، با انتخاب یک توزیع احتمال برای وضعیت اولیه $\{X^{(0)}\}$، دنباله توزیع‌های احتمال$\{\Pi^{(t)}\}_{t>0}$ برای $X^{(t)}$ها به‌دست می‌آید. برای دنباله توزیع‌های $\Pi^{(t)}$ چنانچه Πای وجود داشته باشد

$$\lim_{t \longrightarrow \infty} \Pi^{(t)} = \Pi$$

و Π یک توزیع احتمال بر S باشد، آن‌گاه Π را توزیع حدی زنجیر می‌نامند.

تعریف ۱.۵.۱. زنجیر $\{X^{(t)}\}$ را «ارگودیک» گویند، هرگاه برای هر توزیع ابتدایی $\Pi^{(0)}$، دنباله توزیع‌های احتمال به‌دست آمده $\{\Pi^{(t)}\}_{t>0}$ به توزیع احتمال Π همگرا باشد، یعنی

$$\forall\ \Pi^{(0)} : \lim_{t \longrightarrow \infty} \Pi^{(t)} = \Pi.$$

توزیع مانای زنجیر

اگر برای وضعیت اولیه زنجیر مارکوف، $X^{(\circ)}$، توزیع $\Lambda = [\lambda_i]_{i\in S}$ را بتوان در نظر گرفت به‌طوری‌که

$$(\Lambda)^T = (\Lambda)^T \times P$$

آن‌گاه برای هر $t \geq \circ$ توزیع $X^{(t)}$ نیز $\Lambda = [\lambda_i]_{i\in S}$ خواهد بود. در این حالت Λ را توزیع مانای زنجیره $\{X^{(t)}\}$ (یا متناظر توزیع مانای ماتریس P) می‌گویند.

زنجیرهای تحویل‌ناپذیر و نامتناوب

برای هر $i, j \in S$، گویند وضعیت j از وضعیت i قابل دسترسی است، اگر

$$\exists \in N : \ p_{ij}^{(k)} = P(X^{(t+k)} = j | X^{(t)} = i) > \circ.$$

به عبارت دیگر از وضعیت i پس از مدت زمان متناهی بتوان به وضعیت j رفت. دو وضعیت i و j را مرتبط گویند، هرگاه j از i قابل دسترسی و i از j قابل دسترسی باشد. زنجیر مارکوف $\{X^{(t)}\}$ را تحویل‌ناپذیر [۴۸] گویند، هرگاه تمام وضعیت‌های آن مرتبط باشند (به این معنی که صرف نظر از مقدار اولیه $X^{(\circ)}$، زنجیر با احتمال مثبت به هر ناحیه از S دسترسی داشته باشد). برای هر $i \in S$ دوره تناوب وضعیت i را با $d(i)$ نشان داده و آن را به‌صورت زیر تعریف می‌کنند:

$$d(i) = gcd\{k \in N : p_{ii}^{(k)} > \circ\}$$

که در آن gcd نشان‌دهنده بزرگترین مقسوم علیه مشترک یک مجموعه از اعداد می‌باشد. اگر $d(i) = ۱$ وضعیت i را نامتناوب[۴۹] می‌گویند. به عبارت دیگر وضعیت i نامتناوب است، هرگاه از i در هر زمانی بتوان دوباره به i برگشت. به‌طور کلی زنجیر را نامتناوب گویند، هرگاه هر وضعیت آن نامتناوب باشد.

زنجیر زمان برگشت‌پذیر

زنجیر مارکوف $\{X^{(t)}\}$ را زمان برگشت‌پذیر[۵۰] گویند، هرگاه یک توزیع $\Pi = [\pi_i]_{i\in S}$ بر S یافت شود که

$$\forall i, j \in S \ : \ \pi_i p_{ij} = \pi_j p_{ij}. \qquad (۷.۱)$$

اگر زنجیر مارکوف $\{X^{(t)}\}$ به معنی فوق زمان برگشت‌پذیر باشد، آن‌گاه Π توزیع مانای زنجیر بوده و یکتاست. زیرا، با جمع بر روی i دو طرف تساوی (۷.۱) می‌توان نتیجه گرفت:

$$\forall j \in S : \sum_{i\in S} \pi_i p_{ij} = \pi_j.$$

[۴۸] Irreducible
[۴۹] Aperiodic
[۵۰] Recurrent

در واقع $(\Pi)^T = (\Pi)^T \times P$ است.

۳.۵.۱ انتگرال‌گیری مونت کارلویی

برای تقریب انتگرال

$$H = \int_S h(x)f(x)dx \qquad (۸.۱)$$

که در آن $h(\cdot)$ تابعی از متغیر تصادفی x با تابع چگالی $f(\cdot)$ است، به روش مونت کارلویی با تولید مشاهدات مستقل و هم‌توزیع از چگالی $f(x)$ و استفاده از $\hat{H}_{Mc} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$ کمیت H را تقریب می‌زنند. با استفاده از قانون قوی اعداد بزرگ، زمانی که $n \longrightarrow \infty$، $\hat{H}_{Mc} \xrightarrow{a.s} H$. در چارچوب استنباط بیزی، $f(x)$ در (۸.۱) نقش تابع چگالی توزیع پسین مدل، $\pi(\boldsymbol{\theta}|y)$ را دارد و برای انجام استنباط بیزی، باید انتگرال‌گیری توابعی مانند $h(\boldsymbol{\theta})$ را برحسب توزیع پسین به‌دست آورد. در واقع نیاز به محاسبه کمیت‌هایی مانند کمیت زیر داریم:

$$E(h(\boldsymbol{\theta})|y) = \int_{\boldsymbol{\theta}\in\psi} h(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|y)d\boldsymbol{\theta}.$$

بنابراین می‌توان با داشتن نمونه $\{\theta^{(1)}, \cdots, \theta^{(n)}\}$ از توزیع پسین، شاخص‌های مورد نیاز را به‌طور تقریبی محاسبه کرد.

اما تولید نمونه از تابع چگالی $f(x)$ یا توزیع پسین، مشکل اساسی روش‌های مونت کارلویی است. برای این کار روش‌های مختلفی از جمله روش رد و پذیرش وجود دارد اما به علت محدودیت‌های این روش‌ها به دنبال روش قدرتمندتری برای استخراج نمونه هستیم. تولید نمونه به صورت غیرمستقیم و با استفاده از نظریه زنجیرهای مارکوف می‌تواند رهیافت مناسبی باشد. در واقع در این روش به‌جای شبیه‌سازی نمونه‌های مستقل از توزیع پسین، نمونه‌ای وابسته از زنجیر مارکوفی که توزیع مانای آن $\pi(\boldsymbol{\theta}|y)$ است، شبیه‌سازی می‌شود. پس این نمونه به پشتوانه قضیه ارگودیک (والتر، ۱۹۸۲)، می‌تواند برای محاسبه کمیت‌های پسینی مورد نظر مانند میانگین، چندک‌ها و احتمال‌ها مورد استفاده قرار گیرد.

۶.۱ الگوریتم‌های MCMC

ایده کلی الگوریتم‌های MCMC برای متغیرهای تصادفی وابسته

$$\{X^{(0)}, X^{(1)}, \cdots, X^{(t)}, \cdots\}$$

ساخت زنجیر مارکوفی است که توزیع مانای آن توزیع هدف π باشد که علاقه‌مند به تولید نمونه از آن هستیم. برای اطمینان از وجود چنین توزیع مانای منحصر به فردی، نیاز به یک زنجیر

مارکوف تحویل‌ناپذیر، برگشت‌پذیر و نامتناوبی داریم. تحت چنین شرایطی، توزیع مانای π یک توزیع حدی خواهد بود (در واقع وقتی $t \longrightarrow \infty$، $X^{(t)} \xrightarrow{d} \pi$ و نقطه شروع تاثیری در همگرایی ندارد). با این وجود احتمال $\pi(A) = \int_A \pi(x)dx$ را می‌توان به‌وسیله $P(X^{(t)} \in A)$ برای $A \in S$ تقریب زد. از این ویژگی همگرایی می‌توان به قضیه ارگودیک که مشابه قانون قوی اعداد بزرگ برای نمونه‌های مستقل و هم‌توزیع است، اشاره کرد. در این قضیه میانگین تجربی تابع دلخواه $h(x)$ که از نمونه‌های وابسته (که از زنجیر مارکوف تولید شده‌اند) حاصل شده است، با احتمال نزدیک به یک به مقدار واقعی مورد انتظار همگرا می‌شود. در واقع

$$\frac{1}{n}\sum_{i=1}^{n} h(X^{(t)}) \xrightarrow{a.s} E(h(x)) = \int_S h(x)\pi(x)dx.$$

اما مسئله اصلی، ساختن زنجیر مارکوف تحویل‌ناپذیر، نامتناوب و زمان برگشت‌پذیر با تابع چگالی مانای حدی یکتای $\pi(x)$ است. در ادامه به‌طور مختصر برای تولید چنین زنجیری، دو الگوریتم متروپلیس ـ هستینگز[۵۱] (چیب و گرینبرگ، ۱۹۹۵) و نمونه گیری گیبز[۵۲] (گمان گمان، ۱۹۸۴) را معرفی می‌کنیم.

الگوریتم متروپلیس ـ هستینگز

اولین بار متروپولیس و همکاران (۱۹۵۳) الگوریتم متروپلیس ـ هستینگز را معرفی کردند سپس هستینگز (۱۹۷۰) این روش را کامل کرد. با انتخاب یک توزیع پیشنهادی مناسب (که تکیه‌گاه آن با تکیه‌گاه توزیع پسین یکی باشد، دم‌های آن از دم‌های توزیع پسین پهن‌تر و تولید نمونه از آن ساده باشد) مانند $q(X^*|X^{(t)})$ و با داشتن مشاهده $X^{(t)}$، پس از سه گام، مشاهده $X^{(t+1)}$ معلوم می‌شود. بنابراین برای به‌دست آوردن سایر مشاهدات نیاز به یک مقدار اولیه داریم که آن را برابر $X^{(0)}$ می‌گیریم. با داشتن $X^{(t)} = x^{(t)}$، الگوریتم مقدار $X^{(t+1)}$ را به صورت زیر تولید می‌کند:

۱) مقدار پیشنهادی X^* را از چگالی $q(.|X^{(t)})$ تولید می‌کنیم.

۲) نسبت متروپلیس ـ هستینگز را که به‌صورت $R(u,v) = \frac{f(v)q(u|v)}{f(u)q(v|u)}$ تعریف شده است را برای $x^{(t)}, X^*$ محاسبه می‌کنیم.

۳) $X^{(t+1)}$ را به صورت زیر انتخاب می‌کنیم:

- با احتمال $min\{1, R(x^{(t)}, X^*)\}$; $X^{(t+1)} = X^*$
- با احتمال $1 - min\{1, R(x^{(t)}, X^*)\}$; $X^{(t+1)} = x^{(t)}$

[۵۱] Metropolis-Hastings

[۵۲] Gibbs sampler

این الگوریتم را تا زمانی که تعداد مشاهدات مورد نیاز به‌دست آید، ادامه می‌دهیم. بنابراین این الگوریتم با انتخاب توزیع پیشنهادی مناسب و مقدار اولیه $x^{(0)}$ به ما یک زنجیر مارکوف با توزیع مانای مورد نظر می‌دهد (رابرت و کسلا، ۲۰۰۴).

نمونه‌گیری گیبز

گیبز برای تقریب یک انتگرال در فیزیک، الگوریتمی را معرفی کرد که به الگوریتم گیبز مشهور شد. گلفاند و اسمیت (۱۹۹۰) از این روش برای محاسبه چگالی‌های حاشیه‌ای در چارچوب استنباط بیزی استفاده کردند و در نهایت کسلا و جورج (۱۹۹۲) این الگوریتم را به‌صورت ساده و واضح بیان کردند. در الگوریتم متروپلیس‌ـ هستینگز زمانی که توزیع شرطی کامل به‌عنوان توزیع پیشنهادی انتخاب شود، نمونه‌گیری گیبز یک حالت خاصی از آن می‌شود.

برای تقریب (۸.۱) به روش مونت کارلویی وقتی که با روش‌های معمول نتوان نمونه‌ای از $f(x)$ استخراج کرد، الگوریتم متروپلیس‌ـ هستینگز روشی عمومی برای تولید یک زنجیر ارگودیک با چگالی مانای حدی یکتای $f(x)$ می‌باشد که تحقق‌های این زنجیر را می‌توان به عنوان نمونه‌ای تقریبی از $f(x)$ فرض کرد. اما اجرای این الگوریتم نیازمند چگالی پیشنهادی مناسب برای تولید یک مشاهده به شرط داشتن مشاهده قبلی است و به‌طور کلی نمی‌توان قاعده‌ای برای انتخاب توزیع پیشنهادی ارایه داد. بنابراین بهتر است بر اساس مسائلی که به روش MCMC می‌خواهیم حل کنیم، الگوریتم‌های مخصوصی را طرح کرده و زنجیرهای تولیدشده با این الگوریتم‌ها را دسته‌بندی کنیم. به‌طور مثال در الگوریتم متروپلیس‌ـ هستینگز، یک دسته‌بندی بر اساس انواع زنجیرهایی که می‌توان با این الگوریتم تولید کرد، زنجیرهای مستقل است.

حال با طرح مسئله‌ای، الگوریتم دیگری را برای حل این مسئله مطرح می‌کنیم:

فرض کنید $X = (X_1, \cdots, X_p) \sim f(x)$ و

$$X_{-i} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_p) \; ; \; i = 1, \cdots, p.$$

فرض کنید برای $i = 1, \cdots, p$ توزیع شرطی $X_i|X_{-i} = x_{-i}$ دارای تابع چگالی $f(x_i|x_{-i})$ باشد. می‌خواهیم (۸.۱) را به روش مونت کارلویی تقریب بزنیم، اما به روش‌های معمول نمی‌توانیم از $f(x)$ نمونه استخراج کنیم. چنانچه برای $i = 1, \cdots, p$، $f(x_i|x_{-i})$ طوری باشد که با داشتن x_{-i} بتوان از آن نمونه تولید کرد، آن‌گاه با انتخاب وضعیت اولیه $X^{(0)}$، الگوریتم زیر با داشتن $X^{(t)}$، $X^{(t+1)}$ را تولید می‌کند:

۱) یک ترتیب از مولفه‌های $X^{(0)}$ را انتخاب می‌کنیم.

۲) برای $i = 1, \cdots, p$، X_i^* را از $f(x_i|X_{-i}^{(t)})$ تولید کرده، تا $X^* = (X_1^*, \cdots, X_p^*)$ به‌دست آید.

۳) قرار می‌دهیم: $X^{(t+1)} = X^*$

زنجیر مارکوفی که با این الگوریتم تولید می‌شود، تحویل‌ناپذیر، نامتناوب و زمان برگشت‌پذیر و در نتیجه ارگودیک است. بنابراین این الگوریتم نیز یکی از روش‌های MCMC است. از این روش، زمانی استفاده می‌شود که توزیع‌های شرطی کامل از نوع شناخته‌شده‌ای باشند. در غیر این صورت با استفاده از الگوریتم‌های دیگری مانند الگوریتم متروپلیس ـ هستینگز، نمونه‌گیری رد و پذیرش و نمونه‌گیری سازوار[۵۳] و نمونه‌گیری برشی[۵۴] (نیل، ۲۰۰۳؛ ارشدی‌پور، ۱۳۹۳) نمونه‌گیری از توزیع‌های شرطی کامل امکان‌پذیر است. نرم‌افزارهای کاربردی برای اجرای روش‌های MCMC عبارتند از: WinBUGS (لون و همکاران، ۲۰۰۰)، JAGS (دپاولی، ۲۰۱۶) و BayesX (بلیتز و همکاران، ۲۰۱۲؛ برزگر، ۲۰۰۵) است که اجرای آن‌ها در نرم‌افزار R با استفاده از بسته‌های R2WinBUGS (استورز، ۲۰۰۵)، R2JAGS[۵۵] و R2BayesX[۵۶] انجام می‌شود.

الگوریتم‌های MCMC علی‌رغم محاسن زیادی که دارند برای برخی از داده‌های حجیم با مشکل همگرایی کند، آمیختگی ضعیف زنجیر، سرعت کند و در نتیجه هزینه بالا مواجه هستند. یکی از رهیافت‌های جانشین مناسب برای روش‌های MCMC که در این پایان‌نامه مورد نظر است، روش INLA می‌باشد که توسط رو و همکاران (۲۰۰۹) معرفی شد. این روش با سرعت بالای محاسباتی، دقتی برابر با روش‌های MCMC دارد و هزینه‌های محاسباتی استنباط بیزی را به شدت کاهش می‌دهد. در فصل ۲ به معرفی این روش می‌پردازیم.

[۵۳] Adjustable sampling

[۵۴] Slie Sampling

[۵۵] https://cran.r-project.org/web/packages/R2jags/R2jags.pdf

[۵۶] https://cran.r-project.org/web/packages/R2BayesX/R2BayesX.pdf

فصل ۲

روش بیزی تقریبی برای فرآیندهای کاکس لگ‌گاوسی

۱.۲ فرآیندهای نقطه‌ای معروف

در این زیر بخش چند فرآیند نقطه‌ای معروف و پر کاربرد را معرفی می‌کنیم.

فرآیند دوجمله‌ای

فرض کنید $n \in N, \phi \neq B \in \mathcal{B}$ و P یک اندازه احتمال تعریف شده بر فضای اندازه پذیر $(B, \mathcal{B}(B))$ باشد که در آن $\mathcal{B}$ سیگماجبر تعریف شده برای تمام مجموعه‌های B است. فرآیند نقطه‌ای X، متشکل از یک نمونه n تایی مستقل و هم‌توزیع، از توزیع P را یک فرآیند دوجمله‌ای[۱] n نقطه‌ای با توزیع P برمجموعه B نامیده و با نماد $X \sim bin(B, n, P)$ نشان می‌دهند. اگر P نسبت به اندازه لبگ κ مطلقا پیوسته باشد و $f = \frac{dp}{d\kappa}$ در این صورت X را یک فرآیند دوجمله‌ای n نقطه‌ای با چگالی f بر مجموعه B نامیده و با نماد $X \sim bin(B, n, f)$ نمایش می‌دهند.

[۱] Binomial point process

فرآیند نقطه‌ای پواسون

برای بسیاری از پدیده‌های طبیعی توزیع پواسون یکی از اساسی‌ترین و کاربردی‌ترین توزیع‌ها است. به همین دلیل فرآیند پواسون[۲] نیز نقشی اساسی در نظریه فرآیندهای نقطه‌ای دارد و به دلیل ویژگی‌های منحصر به فردش مورد توجه پژوهشگران در استنباط فضایی قرار می‌گیرد. به‌طور شهودی، تعداد نقاط در یک فرآیند نقطه‌ای پواسون در هر مجموعه کراندار $B \subset S$ که با $N(B)$ نشان داده می‌شود، یک متغیر تصادفی پواسون است و نقاط این فرآیند اثر متقابلی با یکدیگر ندارند.

تعریف ۱.۱.۲. (مولر و واگه‌پترسن، ۲۰۰۴). فرآیند نقطه‌ای پواسون با تابع شدت $\lambda(\cdot)$ یک فرآیند نقطه‌ای پواسن X روی S می‌باشد، هرگاه

۱) برای هر مجموعه کراندار $B \subset S$، $\int_B \lambda(u)du < \infty$ و متغیر تصادفی $N(B)$ دارای توزیع پواسون باشد.

۲) برای هر $n \in N$، $B \subset S$ و $E[N(B)] < \infty$ مشروط بر این‌که $\{N(B) = n\}$، نقاط واقع در X_B مستقل، هم توزیع و دارای تابع چگالی توام زیر هستند:

$$f(u_1, u_2, \cdots, u_n | N(B) = n) = \prod_{i=1}^{n} f(u_i | N(B) = n) = \prod_{i=1}^{n} \frac{\lambda(u_i)}{\int_B \lambda(u_i)du_i}$$

که به اختصار آن را با $X \sim Poisson(S, \lambda)$ نشان می‌دهیم.

تابع احتمال پوچی برای فرآیند نقطه‌ای پواسون X با تابع شدت $\lambda(\cdot)$ به‌صورت زیر است:

$$\nu(B) = P\{N(B) = 0\} = \exp\{-\int_B \lambda(u)du\}, \quad B \subset S.$$

تابع همبستگی زوجی یک فرآیند نقطه‌ای پواسون با توجه به ویژگی ۲ همواره ثابت و برابر یک است ($g(u, v) = 1$). ثابت بودن تابع شدت (همگنی) یک فرآیند پواسون شرط لازم و کافی برای مانایی و همسان‌گردی این فرآیند است. فرآیند پواسون استاندارد یا فرآیند پواسون با نرخ واحد، همان فرآیند پواسون همگن با شدت یک است.

در ادامه فرآیند نقطه‌ای پواسون همگن و فرآیند نقطه‌ای پواسون ناهمگن را به صورت مختصر شرح خواهیم داد.

- **فرآیند نقطه‌ای پواسون ناهمگن** مهمترین مدل برای بسیاری از اهداف عملی، فرآیند نقطه‌ای پواسون ناهمگن است. در واقع این فرآیند یک فرآیند تصادفی فضایی کامل است که در آن میانگین تراکم نقاط فضایی متفاوت است. فرض کنید تابع $\lambda(u)$ برای موقعیت فضایی u، میانگین تراکم الگوی نقطه‌ای باشد. بخشی از پنجره مشاهدات مثل

[۲] Poisson process

B را در نظر بگیرید که به پیکسل‌های کوچک تقسیم شده باشند. برای یک پیکسل با موقعیت فضایی u، انتظار داریم تعداد نقاط موجود در آن $\lambda(u)\Delta u$ باشد که در آن Δu مساحت پیکسل است. تعداد کل نقاط مورد انتظار در B، مجموع مقادیر $\lambda(u)\Delta u$ که برای همه پیکسل‌ها با مراکز u در داخل B است. اگر مساحت پیکسل‌ها به سمت صفر میل کند، تعداد مورد نظر برابر $\int_B \lambda(u)du$ می‌شود. بنابراین تابع شدت $\lambda(u)$ فراوانی کلی نقاط و توزیع فضایی آن‌ها را تعیین می‌کند. این مدل بسیار عمومی است، زیرا هیچ محدودیتی در تابع $\lambda(u)$ وجود ندارد. بنابراین، در مطالعات «توزیع فضایی نقاط» برای مجموعه داده‌های مختلف، اغلب مدل فرآیند پواسون ناهمگن قابل استفاده است.

پذیره اصلی در الگوهای نقطه‌ای پواسون این است که نقاط مستقل از یکدیگر هستند. با توجه به این‌که دقیقا n نقطه در منطقه B وجود دارد، این نقاط مستقل هستند و هر نقطه دارای یک توزیع احتمالی روی B با تابع چگالی احتمال $f(u) = \lambda(u)/\mu$ است.

یکی از ویژگی‌های فرآیندهای الگوی نقطه‌ای پواسونی ویژگی نازک شدن[۳] است. این ویژگی بیان می‌کند، اگر یک فرآیند پواسون ناهمگن با تابع شدت $\lambda(u)$ تنک شود، به‌طوری که احتمال حفظ نقاط در موقعیت u، $p(u)$ باشد و سرنوشت هر نقطه مستقل از سرنوشت نقاط دیگر باشد، فرآیند حاصل از نقاط حفظ‌شده پواسون با تابع شدت $p(u)\lambda(u)$ است.

یکی دیگر از ویژگی‌های فرآیندهای الگوی نقطه‌ای پواسونی آن است که اگر X و Y فرآیندهای نقطه‌ای پواسون ناهمگن مستقل به ترتیب با توابع شدت $\lambda_X(u)$ و $\lambda_Y(u)$ باشند، آن‌گاه $X \cup Y$ نیز یک فرآیند پواسون با تابع شدت $\lambda_X(u) + \lambda_Y(u)$ است.

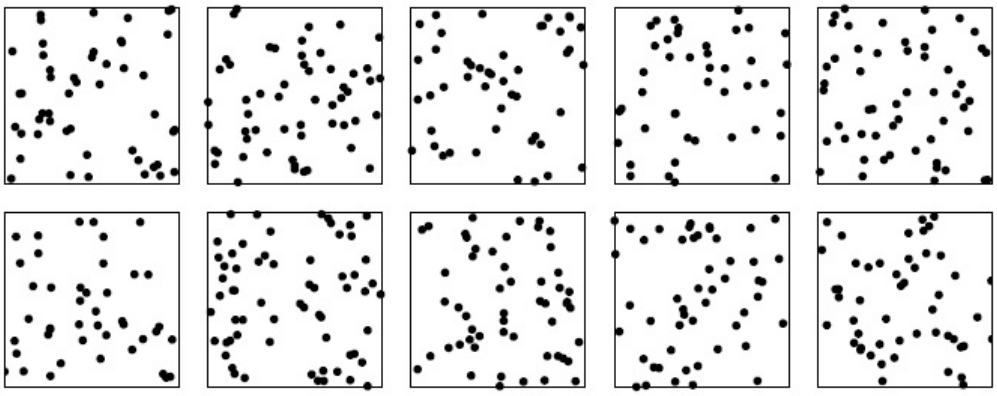
- **فرآیند نقطه‌ای پواسون همگن**

فرآیندهای نقطه‌ای فضایی کاملاً تصادفی[۴] (CSR) هم ارز با فرآیندهای پواسون همگن در $\mathbb{R}^d$ می‌باشد. در این فرآیندها $N(B)$ تعداد نتایج نقاط است که به‌طور مستقل و یکنواخت بر روی ناحیه B توزیع شده‌اند. CSRها کاربردهای متفاوتی دارند. این فرآیندها یک مدل واقعی از برخی پدیده های فیزیکی مانند رادیواکتیو، حوادث نادر و رویدادهای شدید هستند. و به عنوان معیار یا مدل مرجع استاندارد برای الگوهای کاملا تصادفی، که در آن الگوهای دیگر را می توان مقایسه کرد استفاده می‌شوند. در بسیاری از آزمون‌های آماری، این مدل‌ها به عنوان فرضیه صفر عمل می کنند. شکل ۱.۲ نمایشی از الگوهای کاملا تصادفی است. علاوه بر ساخت بسیاری از مدل‌های دیگر از CSR مفاهیم ریاضی زیادی نسبت به آن تعریف شده‌است.

---

[۳] Thinning

[۴] Complete spatial randomness



شکل ۱.۲: الگوهای کاملا تصادفی: ۱۰ شبیه‌سازی از فرآیند نقطه‌ای پواسون با شدت ۵۰ در واحد مربع.

در این‌جا همگنی به این معنی است که تعداد نقاط مورد انتظار در منطقه B به‌طور متوسط باید متناسب با $|B|$ باشد:

$$En(X \cap B) = \lambda |B|$$

λ میانگین تعداد تصادفی نقاط در هر واحد است و به عنوان شدت فرآیند نقطه‌ای شناخته می‌شود که دربخش ۱.۴.۱ به طور کامل به آن پرداخته شد.

فرآیند نقطه‌ای کاکس

در بسیاری از پدیده‌های طبیعی که با فرآیندهای نقطه‌ای مدل‌سازی می‌شوند، نقاط فرآیند با یکدیگر اثر متقابل از نوع جذبی، دفعی یا ترکیبی از این دو را دارند و تحقق آن‌ها به شکل الگوهای نقطه‌ای خوشه‌ای یا منظم هستند. در فرآیند پواسون برخلاف سادگی و انعطاف‌پذیری بالا نقاط اثر متقابلی ندارند. بنابراین این مسئله یکی از معایب این فرآیند محسوب می‌شود. رهیافت جانشین برای این مسئله، استفاده از فرآیند نقطه‌ای پواسون تعمیم‌یافته[۵] است به‌طوری که بتوان اثر متقابل بین نقاط را لحاظ کرد. یکی از این تعمیم‌ها استفاده از فرآیند کاکس[۶] است. کاکس (۱۹۵۵) فرآیند پواسونی را معرفی کرد که در آن اندازه شدت، یک اندازه تصادفی بود. او این فرآیند را پواسون دومرحله‌ای[۷] نامید که به افتخار او این فرآیند را کاکس می‌نامند.

گردایه‌ای از متغیرهای تصادف $Z = \{Z(u) : u \in S\}$ یک میدان تصادفی بر S است، هرگاه برای هر $u \in S$، $Z(u)$ یک متغیر تصادفی باشد. حال فرآیند نقطه‌ای کاکس به‌صورت زیر قابل تعریف است.

[۵] Generalized poisson process

[۶] Cox point process

[۷] Two-step poisson process

تعریف ۲.۱.۲. هرگاه به شرط میدان هادی[۸] Z، X یک فرآیند پواسون با تابع شدت تصادفی Z باشد، فرآیند نقطه‌ای X را یک فرآیند کاکس می‌نامند.

با این تعریف، فرآیند کاکس همان فرآیند پواسون ناهمگنی است که تابع شدت آن توسط میدان تصادفی Z تعیین می‌شود و آن را فرآیند پواسون تصادفی دو مرحله‌ای نیز می‌نامند. در مرحله اول، میدان تصادفی Z تابع شدت فرآیند را تولید می‌کند و در مرحله دوم با داشتن تابع شدت، فرآیند پواسون نقاط فرآیند را تولید می‌کند. این فرآیند در مواردی مانند وجود اطلاعات پیشین و وجود یک فرآیند پنهان، در زمین آمار و تنک‌سازی تصادفی استفاده می‌شود.

تابع احتمال پوچی X به‌صورت زیر تعریف می‌شود:

$$\nu(B) = P\{N(B) = \circ\} = E\big[\exp\{-\int_B Z(u)du\}\big]$$

که در آن امید ریاضی نسبت به میدان تصادفی Z، محاسبه می‌شود و شکل بسته‌ای ندارد. تابع شدت این فرآیند به‌صورت زیر محاسبه می‌شود:

$$\int \lambda(u)du = E\big[N(B)\big] = E\big[E\big[N(B)|Z\big]\big] = E\big[\int_B Z(u)du\big] = \int_B E[Z(u)]du.$$

بنابراین

$$\lambda(u) = E[Z(u)].$$

با شرطی کردن بر میدان تصادفی Z، تابع همبستگی زوجی X به‌صورت زیر است:

$$g(u,v) = \frac{E[Z(u)Z(v)]}{E[Z(u)]E[Z(v)]} = ۱ + \frac{Cov(Z(u),Z(v))}{\lambda(u)\lambda(v)} = ۱ + \frac{c(u,v)}{\lambda(u)\lambda(v)}.$$

بنابراین می‌توان گفت با استفاده از تابع میانگین و تابع کواریانس میدان تصادفی هادی، تابع شدت و تابع همبستگی زوجی یک فرآیند کاکس تعیین می‌شوند. در این فرآیند مانایی و همسان‌گردی تنها به مانایی و همسان‌گردی میدان تصادفی هادی آن بستگی دارد.

فرآیند کاکس لگ‌گاوسی

فرآیندهای کاکس لگ‌گاوسی در رویکرد آماری نخستین بار توسط مولر و همکاران (۱۹۹۸) معرفی شدند. این فرآیندها رده وسیعی از فرآیندهای نقطه‌ای هستند که برای مدل‌سازی الگوهای نقطه‌ای بسیار انعطاف‌پذیر هستند. برای تعریف دقیق فرآیندهای کاکس لگ‌گاوسی که به اختصار آن را با LGCP[۹] نشان می‌دهیم، نیاز به مقدماتی است که ابتدا به آن‌ها می‌پردازیم.

---

[۸] Induced field

[۹] Log-Gaussian Cox Process

۲.۲ میدان تصادفی گاوسی

یک تعریف کلی از میدان تصادفی گاوسی را در پیوست (آ.۱) آورده‌ایم. در این‌جا تعریفی جزیی‌تر از این میدان را ارایه می‌دهیم.

تعریف ۱.۲.۲. (ادلر، ۱۹۸۱). هرگاه برای هر $n \in \mathbb{N}$ و هر انتخاب

$$u_1, u_2, \cdots, u_n \in S$$

بردار تصادفی $(\zeta(u_1), \zeta(u_2), \cdots, \zeta(u_n))$ دارای توزیع نرمال n متغیری با بردار میانگین

$$(m(u_1), m(u_2), \cdots, m(u_n))$$

و ماتریس واریانس

$$\Sigma = \begin{bmatrix} c(u_1, u_1) & \cdots & c(u_1, u_n) \\ c(u_2, u_1) & \cdots & c(u_2, u_n) \\ \vdots & \ddots & \vdots \\ c(u_n, u_1) & \cdots & c(u_n, u_n) \end{bmatrix}$$

باشد، $\zeta = \{\zeta(u) : u \in S\}$ را یک میدان تصادفی گاوسی می‌گویند.

به‌طور یکتا توزیع میدان تصادفی گاوسی ζ توسط تابع میانگین و تابع کواریانس نامنفی مشخص می‌شود. شبیه‌سازی یک میدان تصادفی گاوسی با بسته $RandomField$ در نرم افزار R امکان‌پذیر است.

یک میدان گاوسی تحت شرایط زیر مانای مرتبه دوم است:

- تابع میانگین $m(u)$ ثابت باشد.
- تابع کواریانس $c(u, v)$ تنها به اختلاف $u - v$ بستگی داشته باشد.

در صورت مانا بودن و همسان‌گردی یک میدان تصادفی گاوسی، می‌توان نوشت

$$c(u, v) = Cov(\zeta(u), \zeta(v)) = c(||u - v||) = c(r).$$

باید توجه داشت که مانایی یک فرآیند تصادفی گاوسی از مانایی مرتبه دوم نتیجه می‌شود. معمولا انتخاب رایج برای تابع کواریانس میدان گاوسی ζ از خانواده‌های نمایی، گاوسی، مترن[۱۰]، دایره‌ای، کوشی و پایدار است. هرگاه تمام تحقق‌های ζ انتگرال‌پذیر موضعی باشد، یعنی $P\{\int_B \zeta(u)du\} = ۱$، آن‌گاه میدان تصادفی ζ را تقریبا همه جا انتگرال‌پذیر موضعی می‌دانند. انتگرال‌پذیری یک میدان گاوسی در گرو پیوسته بودن تابع کواریانس آن است (ادلر، ۱۹۸۱). با توجه به خانواده‌های مورد استفاده برای تابع کواریانس، میدان گاوسی موجود در فرآیند کاکس لگ‌گاوسی که در ادامه بیشتر معرفی می‌شود، تقریبا همه جا انتگرال‌پذیر موضعی است.

[۱۰] Matern

۳.۲ فرآیند کاکس لگ‌گاوسی

مولر و همکاران (۱۹۹۸) فرآیند کاکسی را معرفی کردند که در آن لگاریتم میدان هادی آن یک میدان تصادفی گاوسی است. از آن‌جایی که یک میدان گاوسی مقادیر منفی را اختیار می‌کند، نمی‌توان به‌طور مستقیم از آن به‌عنوان میدان هادی یک فرآیند کاکس استفاده کرد. برای رفع این مشکل، کافی است از تبدیل نمایی بر روی میدان گاوسی استفاده کرد تا یک میدان تصادفی نامنفی به‌دست آید.

تعریف ۱.۳.۲. (مولر و همکاران، ۱۹۹۸). میدان هادی $\Lambda(s) = exp(\zeta(s))$ را یک فرآیند LGC بر S گویند، هرگاه

$$\boldsymbol{\zeta} = \{\zeta(s); s \in S \subset \mathbb{R}^d\}$$

یک میدان تصادفی گاوسی با تابع میانگین $m(s)$ و تابع کواریانس $c(s,v)$ باشد. این فرآیند به‌طور قریب به یقین، انتگرال‌پذیر موضعی است. شایان ذکر است که مانایی و همسان‌گردی یک فرآیند LGC از مانایی و همسان‌گردی میدان گاوسی متناظر خود نتیجه می‌شود.

تابع احتمال پوچی در یک فرآیند LGC به صورت

$$\nu(B) = E\Big[\exp\Big(-\int_B \exp(\boldsymbol{\zeta}(s))ds\Big)\Big],\ B \subseteq S$$

تعریف می‌شود و این امید ریاضی نسبت به میدان تصادفی گاوسی محاسبه می‌شود که در آن متغیر تصادفی $\zeta(s)$ برای هر $s \in S$ دارای توزیع نرمال است اما توزیع انتگرال تصادفی

$$\int_B \exp(\zeta(s))ds$$

شکل بسته‌ای ندارد و نمی‌توان آن را تعیین کرد.

برای هر $s, v \in S$، تابع همبستگی زوجی و تابع شدت به ترتیب به صورت زیر تعریف می‌شوند:

$$\begin{aligned} g(s,v) &= \exp(c(s,v)) \\ \lambda(s) &= \exp(m(s) + \frac{c(s,v)}{۲})\cdot \end{aligned}$$

با توجه به دو تابع همبستگی زوجی و شدت، می‌توان نتیجه گرفت که توزیع یک فرآیند LGC توسط این دو تابع تعیین می‌شود. معمولا برای تعیین تابع همبستگی زوجی $g(\cdot,\cdot)$، یا معادل آن تابع کواریانس $c(\cdot,\cdot)$، و تابع شدت $\lambda(\cdot)$ از مدل‌های پارامتری استفاده می‌شود. بنابراین برای تعیین یک فرآیند کاکس لگ‌گاوسی کافی است بردار پارامترهای تابع همبستگی زوجی، تابع شدت و همچنین میدان تصادفی ζ برآورد شوند.

برای پیش‌بینی میدان تصادفی گاوسی یک تور بر روی ناحیه W (پنجره مشاهدات) در نظر گرفته و میدان گاوسی ζ_W را بر آن تقریب می‌زنند. فرض کنید X یک فرآیند کاکس لگ‌گاوسی با میدان گاوسی ζ باشد. در این صورت تابع درست‌نمایی مشاهدات X_W به شکل زیر است:

$$L(\boldsymbol{\theta}|X_W) = f(X_W) = E\Big[\exp\Big(\int_W (1 - \exp(\zeta(s)))ds\Big)\prod_1^n \exp(\zeta(s_i))\Big]$$
$$= E\Big[\exp\Big(|W| - \int_W \exp(\zeta(s))ds + \sum_{s\in W}^{max}(\zeta(s))\Big)\Big]$$

که در آن $\boldsymbol{\theta}$ بردار پارامترهای مدل است. با توجه به شکل تابع درست‌نمایی، امید ریاضی باید به روش مناسبی تقریب زده شود. در بخش‌های بعدی دو روش را برای این کار معرفی و تشریح می‌کنیم.

فرآیند کاکس لگ‌گاوسی یک مورد خاص از رده مدل‌های گاوسی پنهان LGMs است و مدل‌های گاوسی پنهان[۱۱] زیر رده گسترده و انعطاف‌پذیری از مدل‌های رگرسیون جمعی ساختاری[۱۲](STAR) هستند.

۱.۳.۲ مدل‌های آمیخته خطی

در مدل‌های خطی[۱۳]LMs تابع رگرسیونی یا به عبارتی میانگین شرطی $Y|X$، که در آن Y متغیر پاسخ و X ماتریس طرح است، یک ترکیب خطی از X به صورت $E(Y|X) = X\boldsymbol{\beta}$ می‌باشد. در این مدل β ضرایب رگرسیونی یا اثرات ثابت‌اند. یکی از پذیره‌های مدل‌های خطی، استقلال بین مشاهدات است. زمانی که برای پاسخ‌ها این پذیره برقرار نباشد با مدل‌های آمیخته خطی[۱۴] (LMMs) که تعمیمی از مدل‌های LM است، روبرو می‌شویم. در این مدل‌ها وابستگی بین پاسخ‌ها با استفاده از متغیرهای پنهان یا اثرات تصادفی وارد مدل می‌شوند و میانگین پاسخ این مدل‌ها به صورت زیر گزارش می‌شود:

$$E[Y|X, Z, \alpha] = X\boldsymbol{\beta} + Z\alpha$$

که در آن Z ماتریس طرح متناظر اثرات تصادفی α است (مک‌کالاک و سیرل، ۲۰۰۱).

۲.۳.۲ مدل‌های آمیخته خطی تعمیم‌یافته

مدل‌های خطی زمانی که متغیر پاسخ پیوسته نیست، مثل پاسخ‌های دودویی یا شمارشی، مناسب نیستند و نمی‌توان از توزیع نرمال به‌عنوان مدل خطا استفاده کرد، زیرا مدل‌های

[۱۱] Latent Gaussian Models
[۱۲] Structured additive regression model
[۱۳] Linear Model
[۱۴] Linear Mixed Model

خطی، محدودیت موجود در میانگین پاسخ این داده‌ها را در نظر نمی‌گیرند. رهیافت موجود برای این داده‌ها مدل‌های خطی تعمیم‌یافته[۱۵] (GLMs) می‌باشد که توسط نلدر و ودربرن (۱۹۷۲) معرفی شد. در این مدل‌ها که برای تحلیل پاسخ‌های عضو خانواده توزیع‌های نمایی معرفی شدند، متغیرهای پاسخ می‌توانند غیرنرمال باشند و همچنین میانگین پاسخ لزوما به صورت ترکیب خطی از متغیرهای تبینی نیست و تابعی از آن است که آن را تابع پیوند می‌نامند.

۳.۳.۲ مدل‌های رگرسیون جمعی ساختاری

مدل‌های STAR برای مدل‌بندی اثرات غیرخطی متغیرهای تبیینی شامل مدل‌های خطی تعمیم‌یافته و مدل‌های جمعی تعمیم‌یافته[۱۶] قالبی بسیار انعطاف‌پذیر است. توزیع متغیر پاسخ $y_i,\ i = ۱,...,n$ متعلق به خانواده توزیع‌های نمایی با تابع چگالی یا جرم احتمال به صورت زیر است:

$$\begin{aligned} f(y_i|\boldsymbol{u_i},\boldsymbol{\beta}) = \exp\{y_i(\beta_۰ + \sum_{j=۱}^{n_f} f^{(j)}(u_{ij}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}) \\ -b(\alpha + \sum_{j=۱}^{n_f} f^{(j)}(u_{ij}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}) + c(y_i)\} \end{aligned} \tag{۱.۲}$$

که در آن $\boldsymbol{z}_i = (z_{۱i}, \cdots, z_{n_\beta i})$ و $\boldsymbol{u}_i = (u_{۱i}, \cdots, u_{n_f i})$ متغیرهای تبیینی، $\{\beta_k\}$ اثرات خطی مولفه‌های z_i، $\{f^{(j)}(u_{ij})\}$ توابعی از مولفه‌های u_i و $b(\cdot)$ و $c(\cdot)$ توابعی معلوم هستند. در این مدل

$$\beta_۰ + \sum_{j=۱}^{n_f} f^{(j)}(u_{ij}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}$$

پارامترهای کانونی در خانواده توزیع‌های نمایی محسوب می‌شوند. اگر

$$\mu_i = E[Y_i|u_i,\boldsymbol{\beta}] = g^{-۱}(\eta_i)$$

آن‌گاه

$$g(\mu_i) = \eta_i = \beta_۰ + \sum_{j=۱}^{n_f} f^{(j)}(u_{ij}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki}$$

که در آن $g(\cdot)$ تابع پیوند[۱۷] است. از آن‌جا که توابع $f^{(j)}(\cdot)$ می‌توانند شامل اثرات غیرخطی مانند روندهای زمانی، اثرات فصلی، همبستگی‌های فضایی، شیب‌ها و عرض از مبداهای تصادفی در مدل باشند، این مدل‌ها انعطاف‌پذیری بالایی دارند.

[۱۵] Generalized Linear Model
[۱۶] Generalized additive model
[۱۷] Link function

میدان تصادفی مارکوفی گاوسی

میدان تصادفی مارکوفی گاوسی[۱۸] (GMRF) بر اساس گراف تعریف می‌شود. یک گراف شامل مجموعه‌ای از راس‌هاست که توسط یک یا چند یال به هم وصل شده‌اند. گراف را به صورت زوج مرتب (ν, ξ) نشان می‌دهند که در آن ν مجموعه‌ای متناهی و غیرتهی از رئوس و ξ یال‌های آن است. اگر راس‌ها به صورت $\nu = \{1, \cdots, n\}$ گزارش شوند، گراف را نشان‌دار می‌نامند.

تعریف ۲.۳.۲. فرض کنید $g = (\nu, \xi)$ گرافی نشان‌دار باشد که در آن ξ شامل تمام زوج‌های $\{i, j\}$است به‌طوری که راس‌های i و j هیچ یال مشترکی نداشته باشند و بردار تصادفی

$$\zeta = (\zeta_1, \cdots .\zeta_n)^T$$

دارای توزیع نرمال چندمتغیره با میانگین μ و ماتریس کواریانس Q^{-1} باشد. اگر ζ_i از ζ_j به شرط بقیه مولفه‌ها، ζ_{-ij}، مستقل باشد، در این صورت ζ یک میدان تصادفی مارکوفی گاوسی نسبت به گراف g نامیده می‌شود (رو و هلد، ۲۰۰۵).

در یک GMRF توزیع شرطی کامل هر مولفه ζ به‌صورت توزیع شرطی آن مولفه به شرط سایر مولفه‌ها تعیین می‌شود در واقع $\pi(\zeta_i|\zeta_{-i})$ که در آن ζ_{-i} بردار ζ به جز مولفه ζ_i است، از خاصیت مارکفی توزیع شرطی کامل، ζ_i، $i = 1, \cdots, n$ تنها به تعداد کمی از مولفه‌های ζ بستگی دارد. اگر این مجموعه از مولفه‌ها را با δ_i مشخص کنیم که تشکیل‌دهنده مجموعه‌ای از همسایگی‌های مولفه i ام است، آن‌گاه $\pi(\zeta_i|\zeta_{-i}) = \pi(\zeta_i|\zeta_{\delta_i})$ به‌طوری که فرض می‌شود جملات ζ_i و ζ_{-i,δ_i} مستقل هستند. برای هر $i = 1, \cdots, n$ این رابطه استقلال شرطی می‌تواند به صورت $\zeta_i \perp \zeta_{-i,\delta_i}|\zeta_{\delta_i}$ نوشته شود.

رو و هلد (۲۰۰۵) نشان دادند که درایه ijام ماتریس دقت معین مثبت[۱۹] Q صفر است اگر و تنها اگر $\zeta_i \perp \zeta_j|\zeta_{-i,j}$ و درایه‌های غیر صفر ماتریس توسط ساختار همسایگی میدان مشخص می‌شود یعنی اگر $j \in \{i, \delta_i\}$ آن‌گاه $Q_{ij} \neq 0$، بنابراین ماتریس دقت Q به صورت یک ماتریس تنک[۲۰] می‌باشد که محاسبات را بسیار ساده می‌کند. بنابراین میدان تصادفی مارکوفی گاوسی با استفاده از یک ماتریس دقت تنک تعریف می‌شود و باعث می‌شود که بتوانیم از روش‌های موثر عددی در محاسبات استفاده کنیم.

۴.۳.۲ مدل گاوسی پنهان

فرض کنید میدان تصادفی پنهان $\zeta = \{\alpha, \beta, \beta_1, \cdots, \beta_{n_\beta}, f^{(1)}, \cdots, f^{(n_f)}, \eta_i\}$ در مدل (۱.۲) دارای توزیع نرمال چندمتغیره با بردار میانگین صفر و ماتریس دقت $Q_\theta = \Sigma_\theta^{-1}$ باشد، که در آن Σ_θ ماتریس کواریانس است که به پارامتر θ بستگی دارد. به این مدل که زیررده‌ای

[۱۸] Gaussian Markov Random Field

[۱۹] Positive definite precision matrix

[۲۰] Sparse matrix

از مدل‌های STAR است مدل گاوسی پنهان می‌گویند. مدل‌های LGMs رده وسیعی است که شامل مدل‌های خطی تعمیم‌یافته، مدل‌های خطی آمیخته، مدل‌های جمعی تعمیم‌یافته، مدل‌های جمعی تعمیم‌یافته آمیخته، مدل‌های فضایی خطی، مدل‌های فضایی خطی تعمیم‌یافته، مدل‌های فضایی ـ زمانی[۲۱] و رگرسیون اسپلاین می‌باشند و در زمینه‌های مختلفی کاربرد دارند.

اکثر مدل‌های گاوسی پنهان دارای دو ویژگی اساسی هستند. اول این‌که میدان تصادفی پنهان ζ، که اغلب ابعاد بزرگی دارد، در ویژگی‌های مارکوفی دو به دو، مارکوفی موضعی و مارکوفی فراموضعی (رو و هلد، ۲۰۰۵) صدق می‌کنند. بنابراین میدان تصادفی پنهان یک میدان تصادفی گاوسی مارکوفی با یک ماتریس دقت تنک Q_θ است. به همین دلیل می‌توان از روش‌های عددی مربوط به ماتریس‌های تنک که بسیار سریع‌تر از روش‌های محاسباتی ماتریس‌های چگال[۲۲] هستند، استفاده کرد. ویژگی دوم این است که تعداد ابرپارامترها مقداری کوچک، مانند $m \leq ۶$، است که شرایط لازم برای استنباط‌های سریع را فراهم می‌سازند (ادزویک و همکاران، ۲۰۰۹).

مدل‌های کاکس لگ‌گاوسی در رده LGM قرار می‌گیرند. در این پایان‌نامه از دیدگاه بیزی برای استنباط در این مدل‌ها استفاده می‌کنیم.

۴.۲ تقریب لاپلاس آشیانه‌ای جمع‌بسته

در بخش ۴.۳.۲ مشاهده کردیم که کلیه مولفه‌های پنهان مورد نظر برای استنباط در مجموعه پارامترها به‌صورت $\zeta = \{\beta_۰, \boldsymbol{\beta}, \boldsymbol{f}\}$ تعریف می‌شوند و بردار m ابرپارامتر به‌صورت

$$\boldsymbol{\theta} = \{\theta_۱, \ldots, \theta_m\}$$

مشخص می‌شود. توزیع n مشاهده با فرض استقلال شرطی، به وسیله درستنمایی زیر به‌دست می‌آید:

$$P(y|\zeta, \boldsymbol{\theta}) = \prod_{i=۱}^{n} P(y_i|\zeta_i, \boldsymbol{\theta}) \qquad (۲.۲)$$

که در آن هر داده y_i تنها به یک عضو ζ_i در میدان پنهان ζ مرتبط باشد، بحث می‌شود. مارتینز و همکاران (۲۰۱۳) موردی را که داده‌ها به چندین توزیع یعنی درستنمایی‌های چندگانه تعلق دارند، مد نظر قرار دادند. فرض می‌کنیم پیشین ζ دارای توزیع نرمال چندمتغیره با بردار میانگین ۰ و ماتریس دقت $Q(\theta) = \sum_\theta^{-۱}$ یعنی $\zeta \sim N(۰, Q(\theta))$ با تابع چگالی زیر باشد

$$P(\zeta|\boldsymbol{\theta}) = (۲\pi)^{-n/۲}|Q(\theta)|^{۱/۲} \exp(-\frac{۱}{۲}\zeta' Q(\theta)\zeta). \qquad (۳.۲)$$

[۲۱] Spatio-temporal model
[۲۲] Dense matrix

فرض می‌شود مولفه‌های میدانی گوسی پنهان ζ به‌طور شرطی مستقل هستند، با این نتیجه که $Q(\theta)$ ماتریس دقت تنک است. این ویژگی با عنوان میدان تصادفی مارکوف گوسی (GMRF)، شناخته می‌شود (رو و هلد، ۲۰۰۵). توجه کنید زمانی‌که استنباط GMRF به‌کار می‌رود، تنک بودن ماتریس دقت باعث فواید محاسباتی می‌شود. در واقع، عملیات جبر خطی برای ماتریس‌های تنک به‌کار می‌روند که به محاسبات سرعت قابل توجهی می‌دهند (رو و هلد، ۲۰۰۵).

توزیع پسین توام ζ و $\boldsymbol{\theta}$ به وسیله ضرب درستنمایی و چگالی $GMRF$ و توزیع پیشین ابرپارمتر $p(\boldsymbol{\theta})$ به‌دست می‌آید:

$$\begin{aligned} p(\boldsymbol{\zeta},\boldsymbol{\theta}|y) &\propto p(\boldsymbol{\theta}) \times p(\boldsymbol{\zeta}|\boldsymbol{\theta}) \times p(y|\boldsymbol{\zeta},\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta}) \times p(\boldsymbol{\zeta}|\boldsymbol{\theta}) \times \prod_{i=1}^{n} p(y_i|\zeta_i,\theta_i) \\ &\propto p(\boldsymbol{\theta}) \times |Q(\theta)|^{1/2} \exp\big(-\frac{1}{2}\boldsymbol{\zeta}' Q(\theta)\boldsymbol{\zeta}\big) \times \prod_{i=1}^{n} exp(log(p(y_i|\zeta_i,\theta_i))) \\ &\propto p(\boldsymbol{\theta}) \times |Q(\theta)|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{\zeta}' Q(\theta)\boldsymbol{\zeta} + \sum_{i=1}^{n} log(p(y_i|\zeta_i,\theta_i))\right) \end{aligned} \tag{۴.۲}$$

اهداف استنباط بیزی به‌دست آوردن توزیع‌های پسین حاشیه‌ای برای هر عضو بردار پارامتر

$$p(\zeta_i|y) = \int p(\zeta_i,\boldsymbol{\theta}|y)d\theta = \int p(\zeta_i|\boldsymbol{\theta},y)p(\boldsymbol{\theta}|y)d\boldsymbol{\theta} \tag{۵.۲}$$

و بردار ابرپارامتر

$$p(\theta_k|y) = \int p(\boldsymbol{\theta}|y)d\boldsymbol{\theta}_{-k} \tag{۶.۲}$$

هستند. که در آن θ_{-j} بردار حاصل از حذف درایه jام $\boldsymbol{\theta}$ و n_d بعد میدان ζ است. بنابراین باید مراحل زیر را اجرا کنیم:

(۱) $P(\theta_i|y)$ را محاسبه کرده و سپس با استفاده از این چگالی، می‌توان پسین‌های حاشیه‌ای $p(\theta_k|y)$ را به‌دست می‌آوریم.

(۲) $p(\zeta_i|\boldsymbol{\theta},y)$ را محاسبه می‌کنیم که با استفاده از این چگالی، می‌توان پسین حاشیه‌ای $p(\zeta_i|y)$ را به‌دست آورد.

(۳) انتگرال توزیع پسین میدان پنهان را با استفاده از جمع متناهی محاسبه می‌کنیم

$$\tilde{\pi}(\zeta_i|y) = \sum_k \tilde{\pi}(\zeta_i|\theta_k,y)\tilde{\pi}(\theta_k|y)\Delta_k. \tag{۷.۲}$$

روش INLA فرض‌هایی از مدل را برای تقریب عددی پسین‌های مورد نظر برمبنای روش تقریب لاپلاس که در بخش ۱.۵.۱ معرفی شد، به‌کار می‌برد (تیرنی و کدین، ۱۹۸۶).

۱.۴.۲ تقریب چگالی پسین توام بردار ابرپارامتر

چگالی پسین توام بردار ابرپارامتر مدل به صورت

$$\begin{aligned}\pi(\boldsymbol{\theta}|y) &= \frac{\pi(\boldsymbol{\theta},y)}{\pi(y)}\\ &= \frac{\pi(\boldsymbol{\theta},y)}{\pi(y)}\frac{\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}{\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}\\ &= \frac{\pi(\boldsymbol{\zeta},\boldsymbol{\theta},y)}{\pi(y)\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}\\ &= \frac{\pi(y|\boldsymbol{\zeta},\boldsymbol{\theta})\pi(\boldsymbol{\zeta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(y)\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}\\ &\propto \frac{\pi(y|\boldsymbol{\zeta},\boldsymbol{\theta})\pi(\boldsymbol{\zeta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}\end{aligned} \tag{۸.۲}$$

است. تقریب لاپلاس آشیانه‌ای جمع‌بسته معادله را برای هر مقدار θ به صورت:

$$\tilde{\pi}(\boldsymbol{\theta}|y) \propto \frac{\pi(y|\boldsymbol{\zeta},\boldsymbol{\theta})\pi(\boldsymbol{\zeta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}_G(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}^*(\boldsymbol{\theta})} \tag{۹.۲}$$

تقریب می‌زند، که در آن مخرج کسر با استفاده از تقریب گوسی محاسبه شده است. در تقریب گوسی با استفاده از بسط تیلور، توزیع شرطی کامل

$$\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y) \propto \exp\Big(-\frac{۱}{۲}\boldsymbol{\zeta}^T Q\boldsymbol{\zeta} + \sum_{i\in I}\log\pi(y_i|\zeta_i,\boldsymbol{\theta})\Big)$$

حول مُد توزیع شرطی کامل $\boldsymbol{\zeta}$ یعنی $\boldsymbol{\zeta}^*(\boldsymbol{\theta}) = (\zeta_۱^*(\theta),\cdots,\zeta_n^*(\theta))$ به صورت:

$$\begin{aligned}\tilde{\pi}(\boldsymbol{\zeta}|y,\boldsymbol{\theta}) &\propto \exp\Big(-\frac{۱}{۲}\boldsymbol{\zeta}^T Q\boldsymbol{\zeta} + \sum_{i\in I} g_i(\zeta_i)\Big)\\ &\propto \exp\Big(-\frac{۱}{۲}\boldsymbol{\zeta}^T Q\boldsymbol{\zeta} + \sum_{i\in I}(a_i + b_i\zeta_i - \frac{۱}{۲}c_i\zeta_i\Big)\\ &\propto \exp\Big(-\frac{۱}{۲}\boldsymbol{\zeta}^T(Q+diag(c))\boldsymbol{\zeta} + b^T\boldsymbol{\zeta}\Big)\end{aligned} \tag{۱۰.۲}$$

به‌دست می آید که در آن $c_i = -g''(\zeta_i^*(\boldsymbol{\theta}))$ و $b_i = g'(\zeta_i^*(\boldsymbol{\theta})) + \zeta_i^*(\boldsymbol{\theta})c_i$ است و a_i از تناسب $\pi(\boldsymbol{\theta}|y) = \frac{\pi(\boldsymbol{\zeta},\boldsymbol{\theta},y)}{\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}$ حذف می‌شود (رو و هلد، ۲۰۰۵)، سپس تقریب لاپلاس چگالی پسین با بسط مخرج کسر حول $\boldsymbol{\zeta} = \boldsymbol{\zeta}^*(\boldsymbol{\theta})$ به صورت

$$\pi_{\tilde{LA}}(\boldsymbol{\theta}|y) \propto \frac{\pi(\boldsymbol{\zeta},\boldsymbol{\theta},y)}{\pi(\boldsymbol{\zeta}|\boldsymbol{\theta},y)}|\boldsymbol{\zeta} = \boldsymbol{\zeta}^*(\boldsymbol{\theta}) \tag{۱۱.۲}$$

محاسبه می‌شود.

تقریب $\tilde{\pi}_{LA}(\boldsymbol{\theta}|y)$ در سه مرحله از فرآیند استنباطی روش INLA مورد استفاده قرار می‌گیرد. استفاده اصلی آن در محاسبه تقریب بیزی پسین میدان پنهان است. در مرحله دوم از (۹.۲)

برای محاسبه تقریب حاشیه‌ای درستنمایی $\pi(y)$ استفاده می‌شود. مرحله سوم به منظور محاسبه چگالی حاشیه‌ای پسین برای برخی از بردار ابرپارامترها، $\boldsymbol{\theta}$ ، یعنی $\pi(\theta_j|y)$ به‌کار می‌رود. نحوه به کارگیری معادله (۹.۲) برای محاسبه سه مرحله بیان شده، شامل انتگرال گیری عددی روی محدوده چندبعدی از $\boldsymbol{\theta}$ است. لذا پیش از هر عملی ابتدا باید نقاط انتگرال‌گیری مناسب انتخاب شوند. رو و همکاران (۲۰۰۹) دو راهبرد توری و طرح مرکب مرکزی[۲۳] یا CCD را برای انتخاب نقاط انتگرال‌گیری معرفی کردند. راهبرد توری شامل انتخاب نقاطی از یک توری است که چگالی $\tilde{\pi}_{LA}(\boldsymbol{\theta}|y)$ حجم بیشتری در آن نقاط دارد.

۲.۴.۲ تقریب چگالی حاشیه‌ای پسین درایه‌های ابرپارامتر

محاسبه چگالی حاشیه‌ای پسین θ_k مستلزم انتگرال‌گیری از چگالی $\tilde{\pi}(\boldsymbol{\theta}|y)$ نسبت به بردار ζ_{-j} است (معادله (۶.۲))، اما اگر بعد $\boldsymbol{\theta}$ بزرگ باشد، محاسبه این انتگرال مشکل و زمان‌بر است. رهیافت‌های جایگزین انتگرال‌گیری عبارتند از: راهبرد توری، الگوریتم‌های محاسبه $\tilde{\pi}(\theta_k|y)$، درون‌یابی گوسی‌نامتقارن و الگوریتم‌انتگرال‌گیری عددی $\boldsymbol{\theta}$ که برای محاسبه چگالی حاشیه‌ای پسین ابرپارامترهامورد استفاده قرار می‌گیرند.

۳.۴.۲ تقریب چگالی حاشیه‌ای پسین عناصر میدان‌پنهان

پس از انتخاب نقاط $\{\theta_k\}$ باید چگالی حاشیه‌ای پسین ζ_i ها را محاسبه کنیم تا بتوان با استفاده از دو تقریب $\tilde{\pi}(\theta_k|y)$ و $\tilde{\pi}(\zeta_i|\theta_k,y)$ چگالی (۵.۲) را تقریب زد. تقریب پسین $\tilde{\pi}(\zeta_i|\theta_k,y)$ با استفاده از روش‌های تقریب گوسی، لاپلاس و لاپلاس ساده شده قابل محاسبه است.

به‌طور کلی روش INLA با استفاده از تبدیل‌های لاپلاس و انتگرال‌گیری عددی محاسباتی سریع، تقریبی دقیق را جایگزین الگوریتم‌های MCMC می‌کند. توزیع پسین کناری در (۵.۲) و (۶.۲) با استفاده از تقریب لاپلاس به صورت

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{\pi(\zeta,\boldsymbol{\theta},\boldsymbol{y})}{\tilde{\pi}_G(\zeta|\boldsymbol{\theta},\boldsymbol{y})}\right|_{\zeta=\zeta^*(\boldsymbol{\theta})}$$

محاسبه می‌شود، که در آن $\tilde{\pi}_G(\zeta|\boldsymbol{\theta},\boldsymbol{y})$ تقریب گاوسی $\pi(\zeta_i|\zeta,\boldsymbol{\theta},\boldsymbol{y})$ و $\zeta^*(\boldsymbol{\theta})$ مد توزیع $\pi(\zeta_i|\zeta,\boldsymbol{\theta},\boldsymbol{y})$ است.

۵.۲ معادلات دیفرانسیل جزئی تصادفی

ایده رهیافت معادلات دیفرانسیل جزئی تصادفی[۲۴] (SPDE) توسط لیندگرین و همکاران (۲۰۱۱) معرفی شد. آن‌ها در این رهیافت فرآیند فضایی پیوسته را با استفاده از یک فرآیند تصادفی

[۲۳] Central Composite Design

[۲۴] stochastic partial differential equation

گسسته نمایش داده و با استفاده از ماتریس دقت تنکی که در ادامه معرفی می‌شود، ساختار مارکوفی را وارد مسئله می‌کنند. در واقع، با این رهیافت تقریبی از یک میدان تصادفی چگالی توسط یک میدان تصادفی تنک (مارکوفی) ارایه می‌شود که باعث افزایش سرعت محاسبات خواهد شد.

برای تشریح رهیافت SPDE فرض کنید $Z(s) = \zeta(s), s \in S \subseteq \mathbb{R}^d$ یک میدان تصادفی گاوسی مانای مرتبه دوم و همسان‌گرد با تابع کواریانس مترن با پارامترهای همواری ν و مقیاس κ باشد. علاوه بر این، فرض کنید تحققی از فرآیند $X(s_i)$ در d موقعیت $s_1, \cdots, s_d$ مشاهده شده باشد. روش SPDE یک GMRF با همسایگی فضایی و ماتریس دقت تنک Q می‌یابد که به بهترین وجه نشان‌دهنده میدان مترن مورد نظر باشد. در این روش میدان تصادفی مترن به‌صورت یک ترکیب خطی از توابع تکه‌ای خطی که بر یک مثلث‌بندی[۲۵] از دامنه S تعریف شده است، بیان می‌شود. به عبارت دیگر دامنه فضای S به مجموعه‌ای از مثلث‌ها تقسیم‌بندی می‌شود که در بیشتر از یک یال مشترک یا زاویه مشترک متقاطع نیستند. رئوس مثلث‌بندی اولیه در موقعیت‌های $s_1, \cdots, s_d$ قرار می‌گیرند و سپس رئوس دیگر به منظور رسیدن به یک مثلث‌بندی مناسب برای پیش‌گویی فضایی و سایر اهداف اضافه می‌شوند.

میدان‌های تصادفی مترن تنها جواب‌های پایدار برای یک معادله دیفرانسیل جزئی تصادفی هستند، به‌طوری که یک میدان تصادفی گاوسی $\zeta(s)$ با تابع کواریانس مترن و پارامترهای هموارسازی ν و مقیاس κ به صورت زیر قابل تعریف است:

$$C(h) = (\frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}})(\kappa h)^\nu K_\nu(\kappa h) \qquad (۱۲.۲)$$

که در آن $h = ||s_i - s_j||$، $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu+\frac{d}{2})(4\pi)^{\frac{d}{2}}\kappa^{2\nu}}$ و K_ν تابع بسل نوع دوم هستند (ویتل، ۱۹۵۴ و ۱۹۶۳).

یک جواب پایدار برای معادله دیفرانسیل جزئی تصادفی به‌صورت زیر است:

$$(\kappa^2 - \Delta^2)^{\frac{\alpha}{2}}\tau(s)\zeta(s) = W(s),\ s \in R^2, \alpha = \nu + \frac{d}{2},\ \kappa > 0, \nu > 0 \qquad (۱۳.۲)$$

که در آن W میدان تصادفی فضایی نوفه سفید[۲۶] و Δ عمل‌گر دیفرانسیلی مرتبه دوم یا همان عمل‌گر لاپلاس[۲۷] $\Delta = \sum_{i=1}^d \frac{\partial^2}{\zeta_i^2}$ است و τ واریانس را کنترل می‌کند. جمله $(\kappa^2 - \Delta^2)^{\frac{\alpha}{2}}$ یک عمل‌گر دیفرانسیل[۲۸] است، به‌طوری که با به کار بردن تعریف فوریه یک لاپلاس کسری[۲۹] به صورت

$$\{F(\kappa^2 - \Delta^2)^{\frac{\alpha}{2}}\phi\}(K) = (\kappa^2 + ||K||^2)^{\frac{\alpha}{2}}(F\phi)(K)$$

[۲۵] triangulation
[۲۶] White noise
[۲۷] Laplacian operator
[۲۸] Pesudo-differential operator
[۲۹] Fractional Laplacian

در نظر گرفته می‌شود، که در آن ϕ تابعی در $\mathbb{R}^d$ است. لیندگرین و همکاران (۲۰۱۱) ابتدا یک صورت ضعیف تصادفی[۳۰] معادله دیفرانسیل جزئی (۱۳.۲) را برای یک میدان تصادفی مارکوفی گاوسی که بیان‌گر یک میدان تصادفی مترن روی شبکه مثلث‌بندی باشد، در نظر گرفتند. شبکه بندی مثلثی انعطاف‌پذیری هندسی بیشتری نسبت به روش‌های مبتنی بر شبکه‌های معمول و سنتی دارد.

روش عناصر محدود[۳۱] روشی عددی برای حل تقریبی معادلات دیفرانسیل جزئی است و مثلث‌بندی یک راه حل معمول برای حل معادلات دیفرانسیل جزئی با استفاده از روش عناصر محدود است (برنر و اسکات، ۲۰۰۷)، به‌طوری که دقت جواب‌ها به نحوه مثلث‌بندی بستگی دارد. بنابراین لیندگرین و همکاران (۲۰۱۱) برای نمایش عناصر محدود حل معادله دیفرانسیل جزئی، یک ترکیب خطی از توابع پایه $\phi_t(s)$ به شکل

$$\zeta(s) = \sum_{g=1}^{G} \phi_g(s)\hat{\zeta}_g \qquad (۱۴.۲)$$

ایجاد کردند، که در آن $\{\phi_g\}$ مجموعه‌ای از توابع پایه[۳۲] و $\{\hat{\zeta}\}$ وزن‌های دارای توزیع گاوسی با میانگین صفر هستند و G تعداد کل رئوس مثلث‌بندی است. با توجه به هدف رسیدن به ساختار مارکوفی، توابع پایه به‌صورتی انتخاب می‌شوند که در راس g، ۱ و سایر رئوس صفر باشند.

در بسته نرم‌افزاری INLA برای $d = 2$، معادله دیفرانسیل جزئی (۱۳.۲) به صورت عمومی

$$(\kappa^2(s) - \Delta^2)^{\frac{\alpha}{2}}\tau(s)\zeta(s) = W(s),\ s \in R^2, \alpha = \nu + 1, \nu > 0 \qquad (۱۵.۲)$$

در نظر گرفته می‌شود که میدان‌های تصادفی نامانا را شامل می‌شود. یعنی میدان‌هایی که واریانس $W(s)$ و پارامتر κ در آن‌ها ثابت نیستند. اما برای سادگی واریانس $W(s)$ را ثابت می‌گیرند و برای $\zeta(s)$ پارامتر مقیاس $\tau(s)$ فرض می‌شود، به‌طوری که اگر یک یا هر دو پارامتر مقیاس τ و κ ثابت نباشند، میدان نامانا خواهد بود و طبق رابطه (۱۲.۲)، $\sigma^2 = \frac{1}{4\pi\tau^2\kappa^2}$ می‌شود. لیندگرین و همکاران (۲۰۱۱) نشان دادند که وزن‌های گاوسی درنظرگرفته شده در (۱۴.۲) یک ساختار مارکوفی دارند و میدان گاوسی مترن را به یک میدان تصادفی مارکوفی گاوسی تبدیل می‌کنند. همچنین ثابت کردند ماتریس دقت میدان تصادفی مارکوفی گاوسی تولیدشده Q یعنی، (برای $\alpha = 2$) برای بردار وزن‌های گاوسی $\varsigma = \{\hat{\zeta}_1, \cdots, \hat{\zeta}_G\}$ با استفاده از شرایط مرزی نیومن[۳۳] (چنگ و همکاران، ۲۰۰۵) به‌صورت زیر است:

$$Q = \tau^2(\kappa^2 C + 2\kappa^2 G + GC^{-1}G)$$

[۳۰] Stochastic weak formulation
[۳۱] Finite element method
[۳۲] Basis functions
[۳۳] Newman boundary conditions

که در آن درایه ماتریس قطری C به‌صورت

$$C_{ii} = \int \phi_i(s)ds$$

و درایه ماتریس تنک G، به‌صورت

$$G_{ij} = \int \nabla\phi_i(s)\nabla\phi_g(s)ds$$

است به‌طوری‌که منظور از ∇ عملگر گرادیان است. همچنین κ پارامتر مقیاس و τ واریانس را کنترل می‌کند. از آن‌جایی که ماتریس دقت Q که درایه‌های آن به τ و κ وابسته‌اند، تنک است (دارای درایه‌های زیاد صفر است)، ξ یک فرآیند تصادفی گسسته با توزیع نرمال $N(\circ, G^{-1})$ است و در واقع این ماتریس راه‌حل تقریبی روش SPDE را بیان می‌کند.

فصل ۳

فرآیندهای فضایی و فضایی-زمانی کاکس لگ‌گاوسی

در این فصل، برخی از مباحث تکمیلی در مورد فرآیندهای کاکس لگ‌گاوسی ارایه می‌شود. ابتدا برازش یک مدل LGCP از دیدگاه کلاسیک را بیان می‌کنیم. سپس به جزییات مدل‌بندی متغیرهای نشان‌دار در یک فرآیند LGC نشان‌دار می‌پردازیم. در پایان نیز مدل‌های فضایی-زمانی LGCPs و نحوه مدل‌بندی ساختار وابستگی فضایی-زمانی آن‌ها را تشریح می‌کنیم.

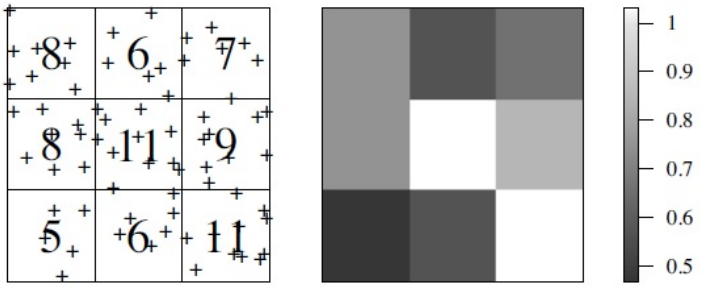
۱.۳ دیدگاه کلاسیک مدل‌بندی یک فرآیند LGC

روش معمول برای مدل‌بندی فرآیندهای LGC به صورت کلاسیک، شمارش پنجره‌های مربعی است. که در آن، پنجره مشاهدات W به زیرناحیه‌های $B_1, \cdots, B_m$ تقسیم می‌شوند، که مربع نامیده می‌شوند. برای سادگی، فرض کنید مناطق دارای مساحت مساوی هستند. تعداد نقاط را در هر مربع می‌شماریم. برای $j = 1, \cdots, m$ قرار می‌دهیم:

$$n_j = n(X \bigcap B_j).$$

اگر مقادیر مورد انتظار $E[n(X \bigcap Bj)]$، همه جا برابر باشند، تابع شدت همگن است.
اندازه پنجره ها می‌تواند بزرگ یا کوچک انتخاب شود. برای تعیین اندازه آن‌ها دو مسئله

اریبی و تنوع مطرح است. انتخاب مربع‌های بزرگتر خطای نسبی شمارش تعداد نقاط در هر مربع را کاهش می‌دهد اما تنوع فضایی تابع شدت در هر مربع را از بین می‌برد. شمارش پنجره‌های مربعی در نرم افزار R در بسته spatstat (بدلی و همکاران، ۲۰۰۵) توسط تابع quadratcount قابل اجرا است (بدلی، ۲۰۱۵).



شکل ۱.۳: شمارش پنجره‌های مربعی برای یک نوع داده. چپ: شمارش پنجره‌های مربعی و راست: برآورد تابع شدت (تعداد در هر متر مربع).

شمارش پنجره‌های مربعی یک تکنیک ابتدایی برای تجزیه الگوهای نقطه فضایی است و شدت متوسط در هر مربع، یک برآورد ساده از تابع شدت است. در واقع تابع intensity() یک تابع عمومی برای محاسبه شدت یک مجموعه داده فضایی یا مدل فرآیند نقطه‌ای فضایی است. میانگین تراکم نقاط در واحد سطح یا شدت تجربی یک الگو نقطه‌ای با تابع intensity.ppp() و شدت نظری آن با intensity.ppm() محاسبه می‌شود. اگر الگو نقطه‌ای X یک مدل فرآیند نقطه‌ای پواسون باشد، شدت فرآیند دقیقا محاسبه می شود اما اگر X یکی دیگر از مدل‌های فرآیند نقطه‌ای گیبز [۱] (آ.۲) باشد، شدت به‌طور تقریبی با استفاده از تقریب زین اسبی ـ پواسون[۲] محاسبه می‌شود (بدلی و همکاران، ۲۰۱۲a؛ اندرسن و همکاران؛ ۲۰۱۴؛ بدلی و همکاران، ۲۰۱۶). این تقریب برای مدل‌های تعامل جفتی در دسترس است.جزئیات بیشتر در اندرسن و همکاران (۲۰۱۴) قابل مشاهده است.

۲.۳ مدل‌بندی متغیرهای نشان‌دار

در داده‌های الگو نقطه‌ای هر گاه علاوه بر نقاط، ویژگی خاصی به هر نقطه ضمیمه شود آن ویژگی را «نشان» و الگو نقطه‌ای را «نشان‌دار» گویند. متغیرهای نشان‌دار را می‌توان وارد مدل‌سازی کرد و برازش دقیق‌تری را نتیجه گرفت. در ادامه به نحوه مدل‌بندی این متغیرها می‌پردازیم. برای برازش مدل فرآیند LGC، پنجره مشاهدات را با گسسته‌سازی به یک مشبکه

[۱] Gibbs point process

[۲] Poisson-saddlepoint approximation

با

$$N = n_{row} \times n_{col}$$

سلول $\{s_{ij}\}$ تقسیم می‌کنیم، به‌طوری که مساحت هر سلول برابر $|s_{ij}|$، برای $i = ۱, \ldots, n_{row}$ و $j = ۱, \ldots, n_{col}$، می‌باشد. بنابراین نقاط درون الگو را می‌توان به شکل $\{\xi_{ijk_{ij}}\}$ نمایش داد، به‌طوری که $k_{ij} = ۱, \ldots, y_{ij}$ و y_{ij} تعداد نقاط مشاهده‌شده در سلول s_{ij} است. تعداد نقاط مشاهده‌شده در سلول s_{ij} به شرط مفروض بودن تابع شدت $\Lambda(s_{ij}) = \exp(\zeta(s_{ij})) = \exp(\zeta_{ij})$ از توزیع پواسون پیروی می‌کند؛ به عبارت دیگر

$$y_{ij}|\zeta_{ij} \sim Po(|s_{ij}| \exp(\zeta_{ij}))$$

و ζ_{ij} به‌صورت زیر مدل‌بندی می‌شود:

$$\zeta_{ij} = \beta_{۰} + f_z(s_{ij}) + f_l(s_{ij}) + \cdots + f_p(s_{ij}) + u_{ij} \qquad (۱.۳)$$

که در آن $f_l(s_{ij}) + \cdots + f_p(s_{ij})$ اثر ساختاریافته فضایی و منعکس‌کننده بزرگ مقیاس فضایی[۳] است همچنین $z(s_{ij})$ نشان‌دهنده یک متغیر کمکی است. متغیرهای کمکی ویژگی‌های خلاصه برای هر مکان در یک پنجره مشاهده هستند و منعکس‌کننده رفتار محلی فردی مانند تعامل محلی یا رقابت هستنداز آن‌جایی که ممکن است رابطه بین متغیر کمکی و لگاریتم تابع شدت خطی نباشد، ارتباط دقیق آن‌ها را با تابع هموار $f(z(s_{ij}))$ نشان می‌دهیم. در این معادله u_{ij} جمله خطا مدل است. حال برای مدل‌بندی متغیرهای نشان‌دار الگوی نقطه‌ای فضایی $X = (\xi_{۱}, \cdots, \xi_n)$ را با دو نشان $\boldsymbol{m}_{۱} = (m_{۱۱}, \cdots, m_{۱n})$ و $\boldsymbol{m}_{۲} = (m_{۲۱}, \cdots, m_{۲n})$ در نظر بگیرید. تعداد نشان‌ها می‌تواند بیشتر هم باشد. فرض می‌شود که $\boldsymbol{m}_{۱}$ و $\boldsymbol{m}_{۲}$ به ترتیب از توزیع خانواده نمای $F_{۱\theta_{۱}}$ و $F_{۲\theta_{۲}}$ با بردارهای پارامتر $\boldsymbol{\theta}_{۱} = (\theta_{۱۱}, \cdots, \theta_{۱q})$ و $\boldsymbol{\theta}_{۲} = (\theta_{۲۱}, \cdots, \theta_{۲q})$ می‌باشند که هر دو وابسته به شدت الگو نقطه‌ای هستند. نشان $\boldsymbol{m}_{۲}$ علاه بر شدت الگو نقطه‌ای به نشان $\boldsymbol{m}_{۱}$ نیز وابسته است. بدون از دست دادن عمومیت، پارامترهای $\boldsymbol{\theta}_{۱۱}$ و $\boldsymbol{\theta}_{۲۱}$ به ترتیب پارامترهای موقعیت‌های $F_{۱}$ و $F_{۲}$ هستند .

با توجه به گسسته‌سازی پنجره مشاهدات ζ_{ij} به صورت زیر مدل می‌شود:

$$\zeta_{ij} = \beta_{۰} + \beta_{۱} \cdot f_s(s_{ij}) + u_{ij} \qquad (۲.۳)$$

برای نشان‌ها مدلی می‌سازیم که در آن نشان $\boldsymbol{m}_{۱}$ با فرض وابستگی به اثر ساختار یافته فضایی $f_s(s_{ij})$ به الگو بستگی داشته باشد. بنابراین با در نظر گرفتن

$$m_{۱}(\xi_{ijk_{ij}})|\kappa_{ijk_{ij}} \sim F_{۱\theta_{۱}(\kappa_{ijk_{ij}}, \theta_{۱۲}, \cdots, \theta_{۱q})}$$

داریم:

$$\kappa_{ijk_{ij}} = \beta_{۰۲} + \beta_{۲} \cdot f_s(s_{ij}) + \varsigma_{ijk_{ij}}, \qquad (۳.۳)$$

[۳] Large scale spatial

که در آن $\varsigma_{ijk_{ij}}$ جمله خطا است. نشان $m_۲$ از طریق $f_s(s_{ij})$، هم به الگوی نقطه‌ای و هم به نشان $m_۱$ بستگی دارد. بنابراین برای

$$m_۲(\xi_{ijk_{ij}})|\nu_{ijk_{ij}} \sim F_{۲\theta_۲(\nu_{ijk_{ij}},\theta_{۲۲},\cdots,\theta_{۲q})}$$

داریم:

$$\nu\kappa_{ijk_{ij}} = \beta_{۰۳} + \beta_{۳} \cdot f_s(s_{ij}) + \beta_{۴} \cdot m_{۱(\xi_{ijk_{ij}})}\omega_{ijk_{ij}}. \qquad (۴.۳)$$

که در آن $\omega_{ijk_{ij}}$ جمله خطا است.

۱.۲.۳ فرآیندهای قدم زدن تصادفی

یک روش مدل‌بندی ناپارامتری متغیرهای نشان‌دار استفاده از فرآیندهای قدم زدن تصادفی[۴] است که مشابه اسپلاین‌ها عمل می‌کنند. در این بخش ابتدا قدم زدن تصادفی مرتبه اول RW1 و سپس قدم زدن تصادفی مرتبه دوم RW2 را معرفی می‌کنیم و نشان می‌دهیم، این فرآیندها ساختار مارکوفی دارند که استفاده از آن‌ها در رهیافت INLA مفید است.

قدم زدن تصادفی مرتبه اول

قدم زدن تصادفی مرتبه اول با فرض مستقل بودن نمو ساخته می‌شود:

$$\Delta x_i \overset{i.i.d}{\sim} N(۰, \kappa^{-۱}),\ i = ۱, \cdots, n-۱ \qquad (۵.۳)$$

در واقع به این معنی است که:

$$x_j - x_i \sim N(۰, (j-۱)\kappa^{-۱}),\ i < j$$

همچنین اگر تقاطع بین $\{i, \cdots, j\}$ و $\{k, \cdots, l\}$ تهی باشد برای $i < j$ و $k < l$ داریم:

$$Cov(x_j - x_i, x_l - x_k) = ۰$$

تابع چگالی x از نمو $(n-۱)$ معادله (۵.۳) حاصل می‌شود:

$$\begin{aligned}\pi(x|\kappa) &\propto \kappa^{\frac{n-۱}{۲}} \exp\Big(-\frac{\kappa}{۲}\sum_{i=۱}^{n-۱}(\Delta x_i)^۲\Big)\\ &= \kappa^{\frac{n-۱}{۲}} \exp\Big(-\frac{\kappa}{۲}\sum_{i=۱}^{n-۱}(x_{i+۱} - x_i)^۲\Big)\\ &= \kappa^{\frac{n-۱}{۲}} \exp\Big(-\frac{۱}{۲}\boldsymbol{x}^T Q \boldsymbol{x}\Big)\end{aligned}$$

[۴] Random walk

که در آن $Q = \kappa R$ و R را یک ساختار ماتریسی می‌نامیم:

$$R = \begin{pmatrix} ۱ & -۱ & & & & & \\ -۱ & ۲ & -۱ & & & & \\ & -۱ & ۲ & -۱ & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -۱ & ۲ & -۱ & \\ & & & & -۱ & ۲ & -۱ \\ & & & & & -۱ & ۱ \end{pmatrix}$$

ماتریس R از معادله زیر به راحتی محاسبه می‌شود:

$$\sum_{i=۱}^{n-۱} (\Delta x_i)^۲ = (D\boldsymbol{x})^T (D\boldsymbol{x}) = \boldsymbol{x}^T D^T D \boldsymbol{x} = \boldsymbol{x}^T R \boldsymbol{x}$$

که D ماتریس $n \times (n-۱)$ بعدی است.

$$D = \begin{pmatrix} -۱ & ۱ & & & \\ & -۱ & ۱ & & \\ & & \ddots & \ddots & \\ & & & -۱ & ۱ \end{pmatrix}$$

قدم زدن تصادفی مرتبه دوم

هر گاه از مرتبه دوم نمو

$$\Delta^۲ x_i \sim N(۰, \kappa^{-۱}),\ i = ۱, \cdots, n-۲ \qquad (۶.۳)$$

استفاده کنیم، مدل به‌عنوان قدم زدن تصادفی مرتبه دوم شناخته می‌شود که تابع چگالی x به‌صورت زیر قابل نمایش است:

$$\begin{aligned} \pi(x) &\propto \kappa^{\frac{n-۲}{۲}} \exp\big(-\frac{\kappa}{۲} \sum_{i=۱}^{n-۲} (x_i - ۲x_{i+۱} + x_{i+۲})^۲\big) \\ &= \kappa^{\frac{n-۲}{۲}} \exp\big(-\frac{۱}{۲} \boldsymbol{x}^T Q \boldsymbol{x}\big) \end{aligned}$$

و ماتریس دقت آن به صورت زیر است:

$$Q = \kappa \begin{pmatrix} ۱ & -۲ & ۱ & & & & & & \\ -۲ & ۵ & -۴ & ۱ & & & & & \\ ۱ & -۴ & ۶ & -۴ & ۱ & & & & \\ & ۱ & -۴ & ۶ & -۴ & ۱ & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & ۱ & -۴ & ۶ & -۴ & ۱ & \\ & & & & ۱ & -۴ & ۶ & -۴ & ۱ \\ & & & & & ۱ & -۴ & ۵ & -۲ \\ & & & & & & ۱ & -۲ & ۱ \end{pmatrix}$$

۲.۲.۳ رهیافت SPDE

با توجه به روش SPDE که در بخش (۵.۲) به‌طور کامل معرفی شد متغیر نشان‌دار را می‌توان به‌صورت یک فرآیند با رهیافت SPDE وارد یک مدل LGCP کرد. در روش SPDE داده‌ها به جای متوسل شدن به سلول‌ها با توجه به مکان دقیق خود، مدل‌سازی می‌شوند. انعطاف‌پذیری در شبکه‌بندی یکی از مزایای این روش محسوب می‌شود و آن را نسبت به روش پنجره‌های مربعی که در بخش (۱.۳) معرفی شد ممتاز می‌کند، زیرا برای کنترل مناطق غیر مستطیلی نیازی به تلاش‌های محاسباتی ندارد و نقاط با شبکه‌بندی بهینه مدل‌سازی می‌شوند. همچنین با توجه به این‌که روش SPDE موقعیت نقاط را در نظر می‌گیرد دقت و کیفیت پیشگویی بالاتری دارد.

۳.۳ الگوهای نقطه‌ای فضایی – زمانی

در برخی مطالعات علمی در هواشناسی، محیط زیست، زمین‌شناسی و اقیانوس‌شناسی با پدیده‌های فضایی داده‌هایی روبرو هستیم که در طول زمان مشاهده می‌شوند و به‌طور زمانی نیز به یکدیگر وابسته‌اند. به مشاهداتی که علاوه بر موقعیت فضایی به موقعیت زمانی نیز وابسته باشند، داده‌های فضایی – زمانی می‌گویند. تحلیل این نوع داده‌ها نیازمند تعیین دو نوع ساختار وابستگی یعنی وابستگی فضایی و وابستگی زمانی است. با توجه به ماهیت این نوع داده‌ها نیازمند دانستن مفاهیم و روش‌های تحلیل مختص به آن‌ها هستیم. ساختار وابستگی مربوط به این نوع داده‌ها، ساختار وابستگی فضایی – زمانی است که معمولا به وسیله تابع کواریانس مدل‌بندی می‌شود.

در سال‌های اخیر، بررسی‌های زیادی برای تعیین این ساختار وابستگی، مدل‌سازی و تحلیل چنین داده‌هایی انجام شده‌است. با توسعه آمار بیزی آمار فضایی توسعه و رشد چشم‌گیری داشته است. نخستین بار کیتانیدس (۱۹۸۶)، با اشاره به مسائل و دشواری‌های

موجود در تحلیل به روش‌های کلاسیک داده‌های فضایی، پیشنهاد استفاده از روش‌های بیزی برای تحلیل این نوع از داده‌ها را داد. اما به علت محاسبات پیچیده و دشوار رهیافت بیزی برای تحلیل داده‌های فضایی مورد استقبال قرار نگرفت. اما با توسعه و گسترش آمار بیزی و ابداع روش‌های محاسبات بیزی پیشرفته مانند روش‌های MCMC، تحلیل بیزی داده‌های فضایی نیز رشد چشم‌گیری یافت.

علت اصلی رشد قابل توجه داده‌های فضایی و فضایی – زمانی، وجود ابزار محاسباتی است که ما را قادر می‌سازد داده‌های واقعی را از طریق GPS ، ماهواره و ... جمع‌آوری کنیم. امروزه محققان در علوم مختلف با داده‌های از جغرافیا (شامل اطلاعاتی درباره فضا و زمان) سروکار دارند.

روش‌های بیزی برای استنباط داده‌های فضایی و فضایی – زمانی در سال ۲۰۰۰ با توسعه روش‌های شبیه‌سازی مونت‌کارلویی زنجیر مارکوفی شروع شد (کسلا و جورج، ۱۹۹۲؛ گیلکس و همکاران، ۱۹۹۶). قبل از این سال، روش بیزی فقط برای مدل‌های نظری به‌کار می‌رفت و کاربردهای کمی در موارد مطالعات واقعی، به دلیل عدم وجود ابزارهای شبیه‌سازی یا عددی برای محاسبه توزیع‌های پسین، داشت. اما پیدایش MCMC برای محققان این امکان را فرآهم آورد که مدل‌های پیچیده را با داده‌های حجیم، بدون نیاز به ساختارهای ساده شده توسعه دهند. روش‌های MCMC به طور گسترده برای استنباط بیزی به کار می‌روند، اما محاسبات این روش‌ها سنگین هستند. زمانی که با مجموعه داده‌های پیشرفته و حجیم که می‌توانند از داده‌های فضایی و زمانی با بعد بالا باشند روبرو هستیم این مشکل بیشتر نمایان می‌شود و اجرای روش MCMC زمان زیادی می‌برد. برای غلبه بر این مشکل از تقریب INLA که در فصل دوم بیان شد استفاده می‌کنیم.

برای تشریح این نوع داده‌ها نیاز به تعریف میدان تصادفی فضایی ـ زمانی، توابع کواریانس و چگونگی ساخت آن داریم.

تعریف ۱.۳.۳. (میدان تصادفی فضایی – زمانی). هر میدان تصادفی فضایی را می‌توان به صورت

$$Y(s,t) = u(s,t) + \varepsilon(s,t)$$

تجزیه کرد که در آن $u(s,t)$ برای موقعیت فضایی $s \in D \subseteq \mathbb{R}^d$ و نقطه زمانی

$$t \in T \subseteq \mathbb{R}$$

یک میدان حالت است و می‌تواند حالت‌های مختلفی داشته باشد و $\varepsilon(s,t) \sim N(0, \sigma_\varepsilon^2)$ خطای اندازه‌گیری فضایی – زمانی را نشان می‌دهد. میدان حالت به صورت‌های مختلف تجزیه می‌شود. تجزیه مورد نظر در این پایان نامه به صورت $u(s,t) = \mu(s,t) + \sigma(s,t)$ که در آن $\mu(s,t)$ تغییرات بزرگ مقیاس یا روند نام‌گذاری می‌شود. به‌طور معمول برای مدل‌بندی روند داده‌های فضایی – زمانی روش‌های متعددی را می‌توان به کار گرفت. ما در این پایان‌نامه

روند را به صورت ترکیبی خطی از متغیرهای تبینی فضایی و زمانی و بدون متغیرهای تبیینی در نظر می‌گیریم. بنابراین $\mu(s,t)$ می تواند تابعی از متغیرهای کمکی باشد که به صورت $\mu(s_i,t) = z(s_i,t)\beta$ نشان داده می شود. می‌توان گفت $z(s_i,t) = (z_1(s_i,t),\cdots,z_p(s_i,t))'$ برداری با بعد p از متغیرهای کمکی در نقاط موقعیت s_i و در زمان t است و $\beta = (\beta_1,\cdots,\beta_p)'$ بردار ضرایب رگرسیونی است. حال اگر تحقق میدان تصادفی فضایی زمانی را اتو رگرسیو[۵] (AR^1) یا $SPDE$ و ماتریس کواریانس را مترن بگیریم، میدان پنهانی شامل ابرپارامترها، σ_ε^2 پارامتر واریانس تصادفی و ضرایب AR^1 یا پارامترهای $SPDE$ خواهد شد. که با فرض استقلال پیشین ابرپارامترها توزیع پسین را محاسبه می‌کنیم و مجددا به علت بسته نبودن شکل توزیع پسین ناچار به استفاده از روش‌های تقریبی هستیم.

تعریف ۲.۳.۳. (میدان تصادفی فضایی – زمانی گاوسی). اگر به ازای $k \geq 1$، توزیع‌های متناهی‌بعد یک میدان تصادفی فضایی – زمانی، نرمال k متغیره باشند، میدان تصادفی حاصل گاوسی نامیده می‌شود.

تعریف ۳.۳.۳. ضرب کرونکر[۶] دو ماتریس A و B را با نماد $A\otimes B$ نشان داده و به‌صورت ضرب هر درایه ماتریس A در همه ماتریس B تعریف می‌شود. به عبارتی اگر

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

و B یک ماتریس دلخواه باشد، آن‌گاه

$$A\otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

۱.۳.۳ فرآیندهای فضایی_زمانی LGC

برای تشریح فرآیندهای فضایی _ زمانی LGC ابتدا به معرفی فرآیندهای فضایی_زمانی پواسون، فرآیندهای فضایی_زمانی کاکس و ساختارهای وابستگی فضایی_زمانی جداپذیر می‌پردازیم و در پایان فرآیندهای فضایی_زمانی LGC را معرفی خواهیم‌کرد.

[۵] Auto regressive

[۶] Kroneker product

فرآیندهای فضایی–زمانی پواسون

فرض کنید AT نشان‌دهنده یک ناحیه فضایی–زمانی باشد؛ عموما در کاربرد، AT برای برخی مناطق فضایی S و فاصله زمانی $(\circ, T)$ به فرم $S \times (\circ, T)$ است. همچنین فرض کنید $N(AT)$ نشان‌دهنده تعداد رخدادها در AT است. یک فرآیند پواسون ناهمگن فضایی–زمانی با شدت $\lambda(x,t)$ به شرح موارد بدیهی زیر، تعریف می‌شود:

- $N(AT)$ دارای توزیع پواسون با میانگین $\int_{\circ}^{T}\int_{A} \lambda(x,t)dxdt$ است.
- در رابطه $N(AT) = n$، رخدادهای n در $A \times (\circ, T)$ یک نمونه تصادفی مستقل از یک توزیع روی AT با توزیع تجربی نسبی برای $\lambda(x,t)$ است.

به‌طور مشابه، تابع لگاریتم درستنمایی همسو، برای داده‌های (x_i, t_i) به صورت زیر است:

$$L(\lambda) = \sum_{i=1}^{n} \log \lambda(x_i, t_i) \int_{\circ}^{T}\int_{A} \lambda(x,t)dxdt \qquad (۷.۳)$$

فرآیندهای فضایی–زمانی کاکس

همان‌طور که در بخش (۱.۲) اشاره شد فرآیند کاکس یک فرآیند پواسون ناهمگن است که نوعی فرآیند تصادفی نامنفی است. در محیط فضایی–زمانی، شدت فرآیند را به صورت $\Lambda(x,t)$ می‌نویسیم. در ادامه امید $\Lambda(x,t)$ با فرض مانایی ساختار کواریانس معرفی می‌شود. یک تبدیل پارامتری مناسب می‌تواند به شرح زیر باشد:

$$\Lambda(x,t) = \lambda(x,t)R(x,t) \qquad (۸.۳)$$

که در آن $R(x,t)$ فرآیند مانا با امید ۱ و تابع کوواریانس $\gamma(u,v) = \sigma^2 c(u,v)$ است.

ساختارهای وابستگی فضایی–زمانی جداپذیر

ساختارهای وابستگی در فضایی–زمانی به دو دسته کلی ساختارهای وابستگی جداپذیر و جداناپذیر فضایی–زمانی رده‌بندی می‌شوند. زمانی که ما با داده‌های فضایی روبرو هستیم با توجه به تعریف تابع کواریانس داریم:

$$C(u) = C(-u)$$

ولی این ویژگی در حالت فضایی – زمانی وجود ندارد.در واقع در حالت کلی

$$\forall\ u, v;\ C(u,v) = C(-u,v)\ \&\ C(u,v) = C(u,-v)$$

برقرار نیست. در صورت برقراری این ویژگی تابع کواریانس را متقارن گویند. اگر

$$\forall u, v \ \frac{C(u,v)}{C(u,\circ)} = \frac{C(\circ,v)}{C(\circ,\circ)}$$

که در آن $C(\circ,\circ)$ یک مقدار ثابت است، تابع کوواریانس را جدایی‌پذیر گویند. اگر شرط جداپذیری برقرار باشد، به راحتی می‌توان کوواریانس فضایی– زمانی را به‌صورت یک کوواریانس فضایی و یک کوواریانس زمانی نوشت. برای مدل کردن تابع کوواریانس فضایی – زمانی از مدل‌های مختلفی مانند مدل‌های ضربی، خطی و جمعی – ضربی می‌توان استفاده کرد. جزئیات بیشتر در رودریگز و همکاران (۲۰۱۰) قابل مشاهده است. معمولاً برای سادگی تابع کوواریانس جداپذیر را به‌صورت ضربی

$$C(u,v) = C(u,\circ)C(\circ,v)$$

محاسبه می‌کنند. بنابراین در (۸.۳)، $\Lambda(x,t)$ جداپذیر مرتبه اول است اگر:

$$\lambda(x,t) = \lambda(x)\mu(t)$$

و جداپذیر مرتبه دوم است اگر:

$$\gamma(u,v) = \sigma^{۲} c(u,\circ)c_{۲}(\circ,v).$$

از این نوع ساختار برای تحلیل فضایی–زمانی داده‌های زلزله در فصل (۴)استفاده می‌کنیم.
ساختارهای جداناپذیر در کرسی (۱۹۹۳)، ما (۲۰۰۸،۲۰۰۳) و رودریگز و دیگل (۲۰۱۰) قابل مطالعه است.

فرآیندهای فضایی–زمانی LGC

در یک فرآیند کاکس لگ‌گاوسی فضایی–زمانی، $\log \Lambda(x,t)$ یک فرآیند گاوسی است و بر اساس میانگین و ساختار کواریانس تعریف می‌شود. برای این فرآیندها با استفاده از معادله (۸.۳) می‌توان نوشت:

$$R(x,t) = \exp\{S(x,t)\}$$

که در آن $S(x,t)$ یک فرآیند گاوسی با میانگین $-\circ/۵\nu^{۲}$، واریانس $\nu^{۲}$ و بنابراین همان‌طور که انتظار داشتیم $E[R(x,t)] = ۱$ و تابع همبستگی $g(u)$ است. بنابراین می‌توان نتیجه گرفت که واریانس و تابع کواریانس $R(x,t)$ به ترتیب $\sigma^{۲} = \exp(\nu^{۲}) - ۱$ و $c(x,t) = \exp\{\nu^{۲} g(u,v)\} - ۱$ است.

هر خانواده معتبر از توابع همبستگی فضایی–زمانی می‌تواند برای تعریف یک کلاس معتبر از فرآیندهای کاکس لگ‌گاوسی مورد استفاده باشد. مطالعه چنین خانواده‌هایی از توابع همبستگی، منجر به تولید یک زیر موضوع در هر بحث مرتبط با خود شده است، که در

کتاب گنیتینگ و گوتورپ[۷] (۲۰۱۰) مورد بررسی قرارگرفته‌است. همان‌طور که قبلا مطرح شد، یک راه برای تایید صحت واستفاده از یک خانواده تفکیک پذیر $C(u,v) = C(u,\circ)C(\circ,v)$ است. برای مثال، هریک از $C(u,\circ)$ و $C(\circ,v)$ می‌تواند برای قرار گرفتن درون کلاس مترن انتخاب شوند.این امر یک راه مناسب است و گاهی یک راه منطقی تجربی برای برازش داده‌ها محسوب می‌شود، اما به‌طور کلی یک راه خاص و ویژه از دیدگاه مکانیسمی نیست.

[۷] Gneiting and Guttorp

فصل ۴

تحلیل بیزی الگوی نقطه‌ای زمین‌لرزه‌های شمال غرب ایران

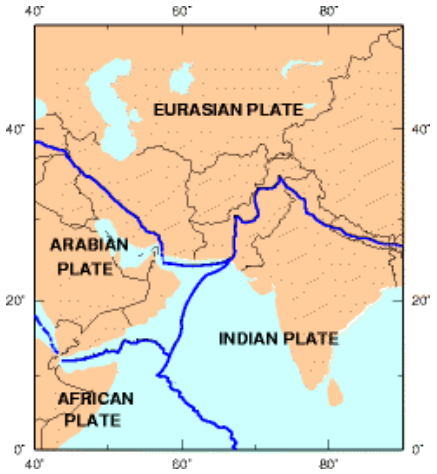
۱.۴ مقدمه

ایران کشوری زلزله‌خیز است و جایگاه ویژه‌ای در میان مناطق زلزله‌خیز جهان دارد. درواقع ایران در مسیر کمربند زلزله‌خیز آلپاید قرار گرفته است (استوکلین، ۱۹۷۷). این کمربند از میانه اقیانوس اطلس شروع شده و پس از عبور از دریای مدیترانه، شمال ترکیه، ایران، هند، چین و جزایر فیلیپین به کمربند دیگری که اقیانوس آرام را دور می‌زند متصل می‌شود. بر اساس نظریه زمین‌ساخت صفحه‌ای[۱] در ایران، همه ساله پوسته سخت عربستان سعودی به میزان متوسط ۵ سانتیمتر در سال به طرف فلات ایران حرکت می‌کند (شکل ۱.۴).

این حرکت حداقل از اواخر دوره میوسین و بیش از ۱۰ میلیون سال قبل تا کنون به صورت آخرین مرحله فعالیت‌های کوه‌زایی در ایران ادامه داشته و موجب فشردگی فلات ایران مابین دو صفحه عربستان[۲] در جنوب و توران[۳] در شمال شده است (درویش‌زاده، ۱۳۷۰).

پوسته کره زمین در ایران بر اثر عوارض اعمال فشارهای ذکر شده به قطعات کوچک‌تر

[۱] Plate Tectonic
[۲] Arabian Plate
[۳] Eurasian Plate



شکل ۱.۴: موقعیت و حرکت صفحه عربستان به سمت فلات ایران

تقسیم شده که حرکات فرعی متفاوتی را نسبت به یکدیگر نشان می‌دهد. وقوع زمین‌لرزه‌های بزرگ و مخرب در گذشته در سراسر ایران به‌ویژه سلسله کوه‌های البرز در شمال و زاگرس در جنوب غرب ایران و همچنین منطقه آذربایجان بر اثر اعمال فشاری کمربند زلزله‌خیزی آلپاید است (وگنر، ۱۹۱۵). در اواخر کوه‌زایی آلپ که از جمله پدیده‌های زمین‌شناسی حاصل از فشردگی است، تحولات زمین‌ساختی جوان[۴] در منطقه آذربایجان اثرات بارزی از خود به‌جای گذاشته که مهم‌ترین آن‌ها ایجاد فرونشست‌های[۵] حد فاصل دریاچه ارومیه تا ایران مرکزی و دریاچه مازندران است (ذکاء، ۱۳۶۸).

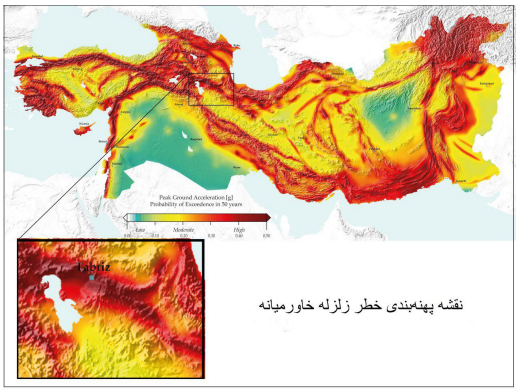
از عوارض مهم حرکت‌های زمین‌ساخت جوان، ایجاد شکستگی‌های فعال در مرز بین فرونشست‌ها و برآمدگی‌ها[۶] می‌باشد. جابه‌جایی‌های نسبی قائم و گاهی افقی در امتداد این گسل‌ها[۷] اغلب توام با آزاد شدن مقدار زیادی انرژی به صورت الاستیک زلزله‌های مخرب است. تحت شرایطی این‌گونه حرکات می‌تواند تدریجی و به آرامی[۸] صورت بگیرد و بدون ایجاد زلزله‌ای بزرگ موجب جابه‌جایی و برش سنگ‌ها و هرگونه تاسیسات ساخت دست بشر که از مسیر آن‌ها عبور کند، گردد.

در منطقه آذربایجان بر اثر حرکت‌های زمین‌ساخت جوان علاوه بر تعداد زیادی گسل‌های فرعی، تعدادی گسل‌های عمده و قابل توجه از جمله گسل‌های شمال تبریز، سلماس، دریک، سراسکند، بزقوش، پیرانشهر، ماسوله، شرق اردبیل، هرآباد، آستارا، سلطانیه و دشت مغان به‌وجود آمده‌اند که نقش مهمی از نظر وضعیت لرزه زمین‌ساختی[۹] و لرزه‌خیزی[۱۰] منطقه دارا

[۴] Neotectonic
[۵] Depression
[۶] Uplift
[۷] Fault
[۸] Creeping
[۹] Seismotectonic
[۱۰] Seismicity

هستند.

بارزترین پدیده زمین‌ساختی منطقه آذربایجان گسل جوان شمال تبریز است. این گسل را دنباله سیستم گسل آناتولی که سرتاسر شمال ترکیه را می‌پیماید تلقی نموده‌اند (وگنر، ۱۹۱۵). با در نظر گرفتن طول و سابقه زلزله‌خیزی و پتانسیل گسل شمال تبریز که به موازات جاده اصلی بستان‌آباد ـ تبریز امتداد یافته، یکی از مهمترین عوامل زمین‌ساختی نامطلوب در منطقه آذربایجان و به‌ویژه شهر تبریز است. منطقه مورد نظر در این پایان‌نامه مربوط به گسل مذکور است. موقعیت مورد مطالعه یکی از مناطق زلزله‌خیز ایران است و بر اساس نقشه تهیه‌شده توسط زارع و همکاران (۲۰۱۶) جزو مناطق پرخطر لرزه‌ای دسته‌بندی شده است (شکل ۲.۴). شهر تبریز بر اثر وقوع زمین‌لرزه‌های مخرب گذشته ناشی از این گسل، بارها مورد خسارت شدید و گسترده و در بعضی موارد به همراه چند ده هزار کشته، به کلی ویران شده است.



شکل ۲.۴: نقشه پهنه‌بندی خطر زلزله در ناحیه خاورمیانه (گزارش‌شده توسط زارع و همکاران، ۲۰۱۶)

زمین‌لرزه یا زلزله، لرزش و جنبش زمین است که به علت آزاد شدن انرژی ناشی از گسیختگی سریع در گسل‌های پوسته زمین در مدتی کوتاه روی می‌دهد. برای آگاهی از میزان تاثیر هر پدیده لازم است تا بتوانیم به نحوی آن را به صورت کمی بیان کنیم. برای کمی کردن اندازه زلزله، از دو رهیافت مختلف استفاده می‌شود؛ یک رهیافت بر اساس اندازه‌گیری دستگاهی (بزرگای زلزله) و دیگری به‌واسطه تاثیرپذیری دست‌سازهای بشر از زلزله (شدت زلزله). شدت زلزله در هر مکان متفاوت است و با دور شدن از کانون زلزله کم می‌شود، در حالی که بزرگای زلزله همواره ثابت است و ربطی به دور شدن از کانون ندارد (چرا که به کل انرژی آزادشده مربوط است).

- **شدت زمین‌لرزه:** شدت یك زلزله در یك مکان خاص بر مبنای اثرهای قابل مشاهده زمین‌لرزه در آن مکان تعیین می‌شود. دقت در تعیین شدت زلزله به دقت مشاهده‌کننده وابسته است. مقیاس‌های مختلفی برای تعیین شدت زمین‌لرزه همانند مقیاس مرکالی اصلاح شده، MSK و $EMS98$ ارائه شده‌اند (وگنر، ۱۹۱۵). برای تعیین شدت زمین‌لرزه برای هر کدام از مقیاس‌ها جداولی تهیه شده‌اند و بر اساس آن‌ها میزان آسیب‌های ناشی

از زلزله بر سازه‌های مختلف ارائه گردیده است و مشاهده‌گر با تطبیق خسارت‌های به‌وجود آمده از زلزله با موارد ذکرشده در جدول‌ها، شدت زلزله را تعیین می‌کند.

- **بزرگای زلزله:** به منظور اندازه‌گیری زمین‌لرزه و به‌دست آوردن معیاری برای مقایسه و سنجش زمین‌لرزه‌ها، از بزرگای زلزله استفاده می‌شود که می‌توان آن را با در نظر گرفتن دامنه نوسانات روی نگاشت محاسبه نمود. مقیاس‌های متفاوتی برای اندازه گیری بزرگای زلزله وجود دارند. اولین مقیاس بزرگا، توسط چارلز ریشتر در سال ۱۹۳۵ برای زلزله‌های جنوب کالیفرنیا تعریف شد که بزرگای محلی نامیده می‌شود. علاوه بر مقیاس ریشتر، مقیاس‌های مختلف دیگری نیز وجود دارند که هر کدام کاربردهای خاص خود را در مهندسی زلزله و زلزله‌شناسی ایفا می‌کنند. هر زلزله فقط و فقط یک بزرگا دارد و بزرگا با فاصله از محل وقوع زلزله تغییر نمی‌یابد.

۲.۴ معرفی داده‌ها

داده‌های مورد نظر این پایان‌نامه، داده‌های لرزه‌ای ثبت‌شده در ناحیه شمال غرب ایران مرتبط با گسل شمال تبریز و گسلش‌های آن است. این داده‌ها از سایت پژوهشکده بین‌المللی زلزله برای دوره زمانی چهار ساله ۲۰۱۵-۲۰۱۲ که توسط دستگاه لرزه‌نگار ثبت شده‌اند، دریافت شده است. تحلیل‌ها بر اساس یک زیرمجموعه به حجم ۵۰۰ از این داده‌ها گزارش شده‌اند. داده‌های لرزه‌ای و نقاط مرزی ناحیه در یک چهارگوش با رئوس

۱. (طول جغرافیایی۴۸’۷۶، عرض جغرافیایی۳۹’۲۲)

۲. (طول جغرافیایی۴۴’۴۷، عرض جغرافیایی۳۹’۲۲)

۳. (طول جغرافیایی۴۸’۷۶، عرض جغرافیایی۳۶’۲۵)

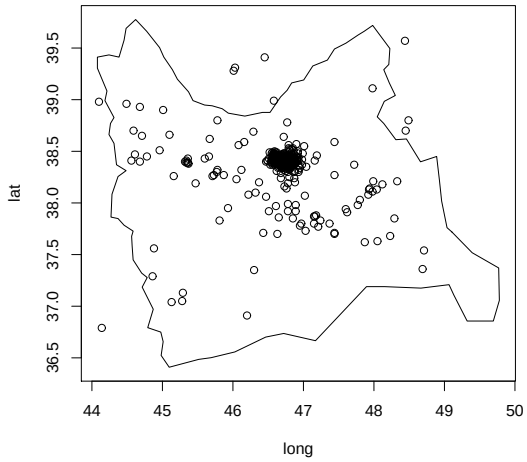
۴. (طول جغرافیایی ۴۸’۷۶، عرض جغرافیایی۳۹’۲۲)

قرار دارند (شکل ۳.۴).

مجموعه داده لرزه‌ای شامل موقعیت‌های جغرافیایی نقاط زلزله با طول و عرض مشخص، تاریخ و زمان دقیق وقوع لرزه‌ها و اندازه زمین‌لرزه (اندازه‌گیری‌شده به واحد ریشتر) است.

۳.۴ مدل‌های مختلف فرآیند LGC برای تحلیل داده‌ها

برای تحلیل داده‌ها با توجه به اطلاعات در دسترس، از مدل‌های مختلفی استفاده کردیم که هر کدام را به تفکیک بیان می‌کنیم.



شکل ۳.۴: نمایش ۵۰۰ موقعیت لرزه‌ای شمال غرب ایران

۱.۳.۴ مدل پواسون برای داده‌های حاصل از پنجره‌های مربعی

در این استنباط هدف برآورد تابع شدت است. قبلا متذکر شدیم که شدت تجربی یک مجموعه داده، تراکم متوسط (میانگین نقاط در واحد سطح یا حجم) است. تابع $intensity$ در نرم‌افزار R و در بسته $spatstat$ به طور کلاسیک شدت را محاسبه می‌کند. مقادیر شدت و شمارش پنجره‌های مربعی در شکل ۴.۴ قابل مشاهده است. نقاط زلزله در پنجره مشاهدات S، مشخص‌شده در شکل ۳.۴، پراکنده شده‌اند. برای برازش مدل فرآیند LGC، پنجره مشاهدات را با گسسته‌سازی به یک مشبکه با $N = n_{row} \times n_{col}$ سلول $\{s_{ij}\}$ تقسیم می‌کنیم، به‌طوری که مساحت هر سلول برابر $|s_{ij}|$، برای $i = ۱, \ldots, n_{row}$ و $j = ۱, \ldots, n_{col}$، می‌باشد. بنابراین نقاط درون الگو را می‌توان به شکل $\{\xi_{ijk_{ij}}\}$ نمایش داد، به‌طوری که $k_{ij} = ۱, \ldots, y_{ij}$ و y_{ij} تعداد نقاط مشاهده‌شده در سلول s_{ij} است. تعداد نقاط مشاهده‌شده در سلول s_{ij} به شرط مفروض بودن تابع شدت $\Lambda(s_{ij}) = \exp(\zeta(s_{ij})) = \exp(\zeta_{ij})$ از توزیع پواسون پیروی می‌کند؛ به عبارت دیگر

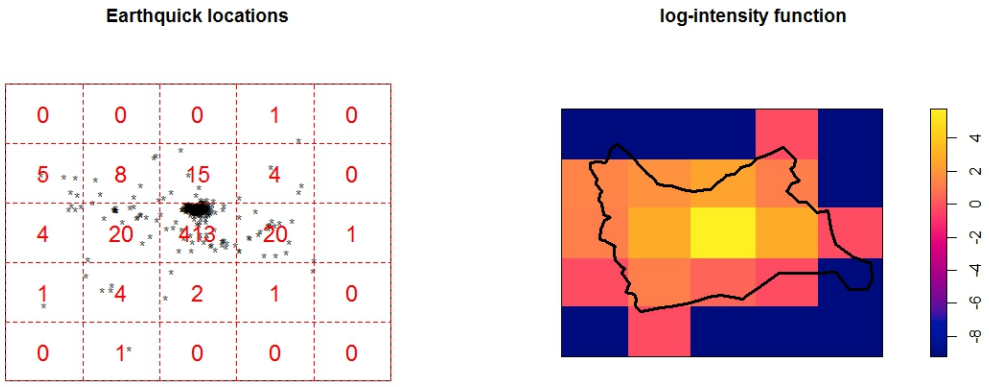
$$y_{ij}|\zeta_{ij} \sim Po(|s_{ij}| \exp(\zeta_{ij})) \qquad (۱.۴)$$

جزئیات بیشتر در(بدلی، ۲۰۱۲) قابل مشاهده است. از برآورد شدت در مکان‌های مختلف و بر اساس شکل به وضوح می‌توان مشاهده کرد که قسمت میانی منطقه مورد نظر شدت بیشتری دارند و چون در وابستگی فضایی همسایه‌ها به یکدیگر شباهت بیشتری دارند.

در این قسمت هدف برازش مدل کاکس لگ‌گاوسی به داده‌های لرزه‌ای می‌باشد. این کار با استفاده از دستور $lgcp.estK()$ در بسته $spatstat$ امکان‌پذیر است. این الگوریتم با استفاده از روش مینیمم مقابله و تابع K از الگوی نقطه‌ای، یک مدل فرآیند کاکس لگ‌گاوسی را به مجموعه داده‌های الگو نقطه‌ای برازش می‌دهد. در این الگوریتم شکل کوواریانس فرآیند LGC

جدول ۱.۴: برآورد شدت به روش کلاسیک

	[۴۳/۵, ۴۴/۸)	[۴۴/۸, ۴۶/۱)	[۴۶/۱, ۴۷/۴)	[۴۷/۴, ۴۸/۷)	[۴۸/۷, ۵۰]
[۳۹/۵, ۴۰/۵]	۰/۰۰۰۰۰۰۰	۰/۰۰۰۰۰۰۰	۰/۰۰۰۰۰۰۰	۰/۷۶۹۲۳۰۸	۰/۰۰۰۰۰۰۰
[۳۸/۵, ۳۹/۵)	۳/۸۴۶۱۵۳۸	۶/۱۵۳۸۴۶۲	۱۱/۵۳۸۴۶۱۵	۳/۰۷۶۹۲۳۱	۰/۰۰۰۰۰۰۰
[۳۷/۵, ۳۸/۵)	۳/۰۷۶۹۲۳۱	۱۵/۳۸۴۶۱۵۴	۳۱۷/۶۹۲۳۰۷۷	۱۵/۳۸۴۶۱۵۴	۰/۷۶۹۲۳۰۸
[۳۶/۵, ۳۷/۵)	۰/۷۶۹۲۳۰۸	۳/۰۷۶۹۲۳۱	۱/۵۳۸۴۶۱۵	۰/۷۶۹۲۳۰۸	۰/۰۰۰۰۰۰۰
[۳۵/۵, ۳۶/۵)	۰/۰۰۰۰۰۰۰	۰/۷۶۹۲۳۰۸	۰/۰۰۰۰۰۰۰	۰/۰۰۰۰۰۰۰	۰/۰۰۰۰۰۰۰



شکل ۴.۴: نمایشی از تابع شدت و طیف رنگی آن به روش کلاسیک

باید مشخص شود، پیش فرض تابع کوواریانس نمایی است، اما دیگر مدل های کوواریانس را نیز می‌توان انتخاب کرد.

الگوریتم برازش یک مدل فرآیند نقطه‌ای کاکس لگ‌گاوسی به داده‌ها نیازمند یافتن پارامترهای مدل فرآیند LGC است که نزدیکترین تطابق بین تابع نظری K و تابع مشاهده شده K مدل فرآیند LGC را نشان می‌دهد. تابع K برای $LGCP$، عبارت است از $K(r) = \int_{\circ}^{r} ۲\pi s\, exp(C(s))ds$ و همان‌طور که در بخش ۳.۲ اشاره شد تابع کواریانس و تابع شدت به ترتیب به صورت زیر تعریف می‌شوند:

$$c(r) = \sigma^{۲} c(-r/\alpha)$$

$$\lambda(s) = \exp(m(s) + \frac{c(s,v)}{۲})$$

که $\sigma^{۲}$ و α به ترتیب پارامترهای واریانس میدان گاوسی و پارامتر مقیاس خودهمبستگی را دارند و c یک تابع کوواریانس شناخته شده است و شکل کوواریانس را تعیین می‌کند. در این الگوریتم ابتدا روش مینیمم مقابله برای پیدا کردن مقادیر بهینه پارامترهای $\sigma^{۲}$ و α استفاده

می‌شود. سپس پارامتر μ از برآورد شدت ρ استنتاج می‌شود. روش مینیمم مقابله مبتنی بر روش گشتاوری در استنباط آماری کلاسیک است، از مقابله برآوردهای تجربی و مشخصه‌های نظری فرآیند برای برآورد پارامترها استفاده می‌کنند.

به عنوان مثال، برای برآورد پارامترهای تابع همبستگی زوجی، $\psi = (\sigma^2, \alpha)$ ، می‌توان از توان دوم اختلاف بین مقدار تجربی (برآورد ناپارامتری) تابع همبستگی زوجی و مقدار نظری آن که تابعی بر حسب ψ است به صورت

$$M(\psi) = \int_0^a (log(\hat{g}) - log(g(r;\psi)))^2 dr$$

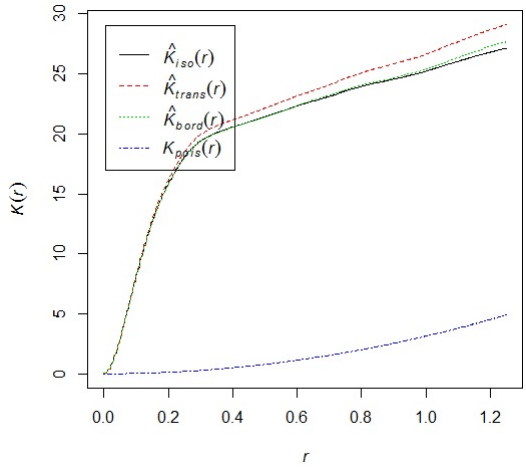
استفاده کرد که در آن $\hat{g}(r)$ برآورد تابع همبستگی زوجی به روش برآورد هسته و g عبارت است از $g(r;\psi) = \exp(c(r;\psi))$ که $c(r,\psi)$ تابع کواریانس میدان گاوسی ζ است. مقدار $a > 0$ توسط کاربر تعیین می‌شود و اغلب برابر با حداکثر فاصله‌ای که نقاط فرآیند ممکن است با یکدیگر اثر متقابل داشته باشند در نظر گرفته می‌شود. حال با مینیمم کردن $M(\psi)$ نسبت به ψ یک برآورد به روش مینیمم مقابله برای ψ حاصل می‌شود. به‌طور مثال برآورد پارامترهای مورد نظر به این روش برای داده‌های زلزله در جدول زیر آمده است.

جدول ۲.۴: برآورد پارامترهای کواریانس و برآورد پارامترهای روش مینیمم مقابله

σ^2	α	var	scale
۶/۷۵۰۸۷۲۴	۰/۳۱۳۴۳۴۹	۶/۷۵۰۸۷۲۴	۰/۳۱۳۴۳۴۹

باید توجه داشت که روش مینیمم مقابله یک روش ساده است و کارایی آن در مقایسه با روش برآورد ماکسیمم درست‌نمایی بسیار کمتر است. به علاوه این روش هیچ راه‌کاری برای پیش‌بینی فضایی میدان گاوسی ζ ارایه نمی‌دهد. در بخش‌های بعد روش‌های برآوردی را به کار خواهیم گرفت که علاوه بر کارایی بیشتر، پیش‌بینی فضایی میدان گاوسی ζ را امکان‌پذیر می‌کند. برای تسریع در انجام محاسبات می‌توان از دستور $lgcp.estpcf()$ به‌جای دستور $lgcp.estK()$ استفاده نمود.

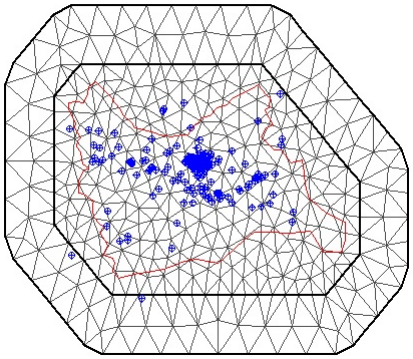
با استفاده از تابع K می‌توان نحوه پراکندگی نقاط را بررسی کرد. با رسم تابع K تجربی محاسبه شده از داده‌ها $\hat{K}(r)$ در مقابل تابع K نظری فرآیند پواسون همگن (K_{pois}) مشاهده می‌شود که $K(r) > K_{pois}(r)$ به این معنا که ارتباط بین نقاط مثبت است و همبستگی وجود دارد علاوه بر این می‌توان پراکندگی داده‌ها را سازگار با فرآیند خوشه‌ای دانست.



شکل ۵.۴: تخمین تابع K

۴.۴ فضایی بدون متغیرهای نشان‌دار با رهیافت LGCP SPDE

اولین قدم برای برازش مدل فرآیند LGC، مثلث‌بندی پنجره مشاهدات است. شکل ۶.۴ مثلث‌سازی پنجره مشاهدات، که منطقه شمال غرب ایران است را نشان می‌دهد. ابتدا ما یک مدل، تنها با وجود اثر ساختار فضایی به داده‌ها برازش می‌دهیم.



شکل ۶.۴: نقشه شمال غرب ایران و مثلث‌سازی آن به روش SPDE

مدل $LGCP$ بدون متغیر نشان‌دار به صورت معادله زیر است:

$$\zeta_{ij} = \beta_0 + f_{spat}(s_{ij}) + u_{ij} \qquad (۲.۴)$$

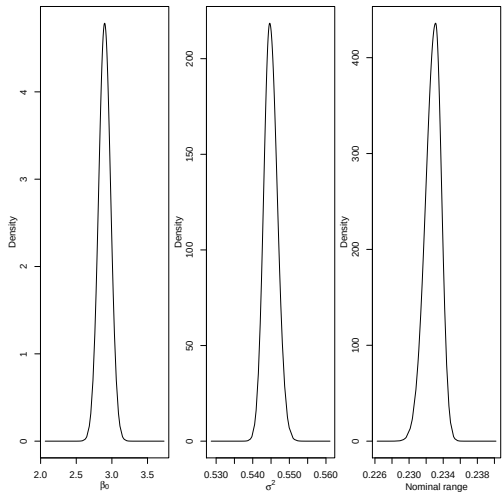
که در آن β_0 پارامتر عرض از مبدا، $f_{spat}(s_{ij})$ اثر فضایی تابع شدت است و با رهیافت معادلات دیفرانسیل تصادفی جزئی وارد مدل می‌شود. همچنین u_{ij} نیز جمله خطای ناهمبسته مدل

است که فرض می‌کنیم دارای توزیع گاوسی با میانگین صفر است. برآوردهای پسین (میانگین، انحراف معیارو چندک‌ها) β_0 و پارامترهای میدان پنهان در جدول زیر قابل مشاهده است. در رهیافت SPDE میدان تصادفی مارکوفی گاوسی یک میدان تصادفی مترن شامل پارامتر دامنه و انحراف معیار است که در این پایان‌نامه به ترتیب آن‌ها را با Range و Stdev نشان می‌دهیم.

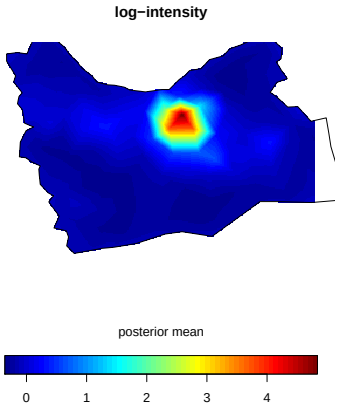
جدول ۳.۴: برآورد β_0 و پارامترهای مدل

			فواصل اعتبار		
	میانگین	انحراف معیار	۰/۰۲۵	۰/۵	۰/۹۷۵
β_0	۲/۸۹۸۳	۰/۰۸۳۴	۲/۷۳۴۶	۲/۸۹۸۲	۳/۰۶۱۸
Range	۰/۲۳۲۸	۰/۰۰۰۹	۰/۲۳۰۸	۰/۲۳۲۸	۰/۲۳۴۴
Stdev	۰/۵۴۵۰	۰/۰۰۱۸	۰/۵۴۱۵	۰/۵۴۴۹	۰/۵۴۸۷

با توجه به برآوردهای حاصل می‌توان نتیجه گرفت داده‌هایی که فاصله آن‌ها بیشتر از ۰/۲۳ کیلومتر است تقریبا مستقل از یکدیگرند.

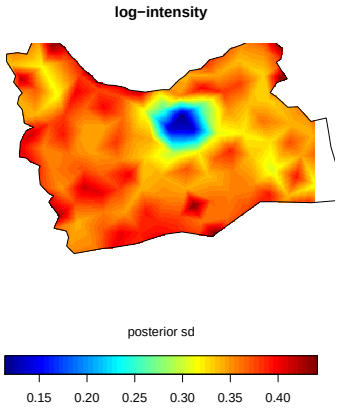


شکل ۷.۴: چگالی‌های حاشیه‌ای پسین برای پارامترهای مدل



شکل ۸.۴: نقشه پهنه‌بندی زمین‌لرزه‌های شمال غرب ایران بر اساس میانگین پسین اثر فضایی

شکل ۸.۴ نقشه پهنه‌بندی زمین‌لرزه‌های شمال غرب ایران را بر اساس میانگین پسین اثر فضایی نشان می‌دهد. با توجه به این شکل بیشترین تعداد زلزله در منطقه میانی (قسمت قرمز رنگ)، روی گسل شمال تبریز است و این منطقه در معرض خطر می‌باشد. هرچه از این منطقه پر خطر دورتر می‌شویم تعداد زلزله‌ها و بنابراین خطر زلزله کمتر می‌شود.



شکل ۹.۴: نقشه پهنه‌بندی زمین‌لرزه‌های شمال غرب ایران بر اساس انحراف‌معیار پسین اثر فضایی

با توجه به شکل ۹.۴ میزان خطا برآورد در مناطق میانی (قسمت آبی رنگ) نسبت به دیگر مناطق کمتر است زیرا تعداد نقاط بیشتری در منطقه میانی وجود دارد.

۵.۴ LGCP فضایی با متغیرهای نشان‌دار

در بخش ۲.۳ به نحوه مدل‌بندی الگو نقطه‌ای نشان‌دار اشاره شد. با توجه به اطلاعات موجود در داده‌های مورد بررسی، مدل‌بندی را به‌صورت نشان‌دار ادامه می‌دهیم. بنابراین می‌توان ζ_{ij} در معادله (۱.۳) را به‌صورت زیر تعریف کرد:

$$\zeta_{ij} = \beta_0 + m + f_{spat}(s_{ij}) + u_{ij} \qquad (۳.۴)$$

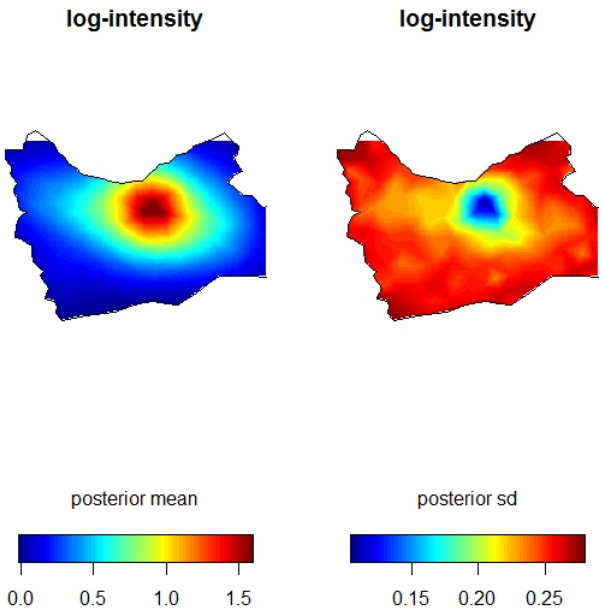
که در آن m یک متغیر نشان‌دار است و برای داده‌های ما اندازه زمین‌لرزه (به واحد ریشتر) است. برآورد پارامترها، ابرپارامتر و متغیر نشان‌دار را می‌توان در جدول ۵.۴ مشاهده کرد.

جدول ۴.۴: برآورد مقادیر انحراف معیار، فواصل اطمینان و میانگین پارامترها و متغیر نشان‌دار

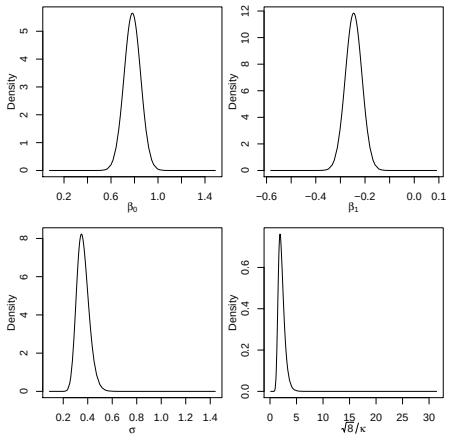
			فواصل اعتبار		
	میانگین	انحراف معیار	۰/۰۲۵	۰/۵	۰/۹۷۵
β_0	۰/۷۸۱۸	۰/۰۷۰۶	۰/۶۴۳۲	۰/۷۸۱۸	۰/۹۲۰۴
m	−۰/۲۴۶۲	۰/۰۳۳۷	−۰/۳۱۲۳	−۰/۲۴۶۲	−۰/۱۸۰۱
Range	۲/۱۹۰	۰/۶۱۹۷	۱/۲۷۳۳	۲/۰۸۶۳	۳/۶۷۸۸
Stdev	۰/۳۶۲	۰/۰۵۰۴	۰/۲۷۴۷	۰/۳۵۷۷	۰/۴۷۲۶

با توجه به مقدار برآورد متغیر کمکی می‌توان تاثیر آن را منفی دانست. این تفسیر به وضوح در شکل ۱۰.۴ که نقشه پهنه‌بندی اثر متغیر نشان‌دار زمین‌لرزه‌های ناحیه شمال غرب ایران است، مشخص می‌باشد. در واقع در محدوده‌ای که بزرگای زلزله مقدار قابل توجه‌ای داشته است تعداد نقاط زلزله کم بوده است.

شکل ۱۱.۴ چگالی‌های حاشیه‌ای پسین پارامترها را نشان می‌دهد.



شکل ۱۰.۴: نقشه پهنه‌بندی زمین‌لرزه‌های ناحیه شمال غرب ایران، بر اساس میانگین پسین متغیر نشان‌دار (سمت چپ)، بر اساس انحراف معیار پسین متغیر نشان‌دار(سمت راست).



شکل ۱۱.۴: چگالی حاشیه‌ای پسین پارامترها

مدل‌بندی متغیرهای نشان‌دار با فرآیندهای قدم زدن تصادفی

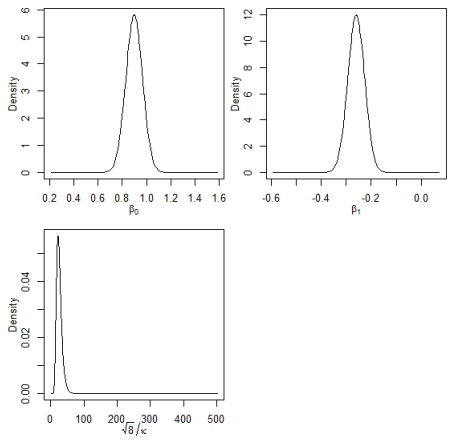
متغیر کمکی را می‌توان به صورت قدم زدن تصادفی مرتبه اول (RW1) وارد مدل کرد (ایلیان و همکاران، ۲۰۱۲). در این حالت نتایج به صورت زیر قابل مشاهده است.

در جدول ۱۲.۴، ϖ پارامتر دقت قدم زدن تصادفی مرتبه اول است.

شکل ۱۲.۴ چگالی‌های حاشیه‌ای پارامترها را زمانی که متغیر نشان‌دار «rw۱» مدل شده است، نشان می‌دهد.

جدول ۵.۴: برآورد پارامتر β_0 ،پارامتر دقت و متغیرنشان‌دار با «rw۱»

	برآورد	معیار انحراف	چندک ۰/۰۲۵	۰/۵	۰/۹۷۵
β_0	۰/۸۹۸۱	۰/۰۶۸۷	۰/۷۶۳۲	۰/۸۹۸۱	۱/۰۳۲۹
m	−۰/۲۶۰۰	۰/۰۳۳۳	−۰/۳۲۵۳	−۰/۲۶۰۰	−۰/۱۹۴۸
ϖ	۲۶/۰۲	۸/۰۷۷	۱۳/۸۴	۲۴/۷۹	

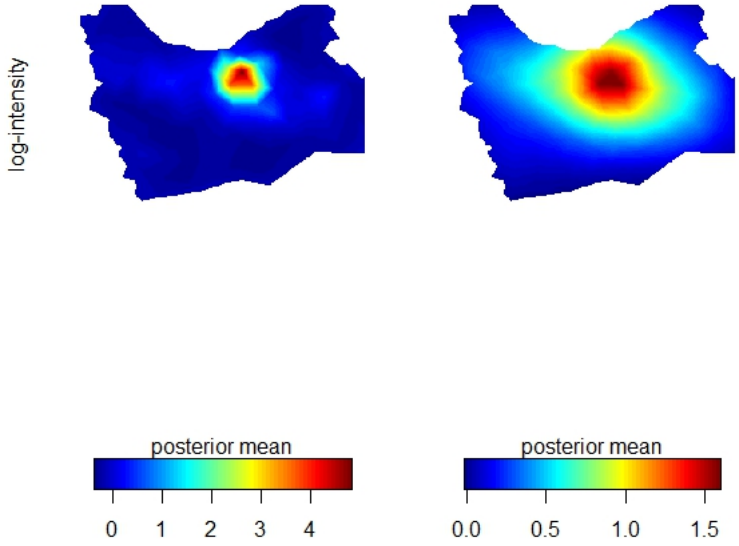


شکل ۱۲.۴: چگالی‌های حاشیه‌ای پارامترها با استفاده از روش INLA

مقایسه فرآیند LGC بدون نشان با فرآیند LGCنشان‌دار

با مقایسه مقادیر دو جدول ۳.۴ و ۵.۴ که به ترتیب مربوط به مدل‌بندی فرآیند LGC بدون نشان و فرآیند LGC نشان‌دار است متوجه می‌شویم که علاوه بر افزایش برآورد پارامتر Range، پارامتر خطا کاهش یافته است. همچنین وابستگی فضایی با وجود متغیرنشان‌دار از ۰/۲۳ کیلومتر به ۳/۶۷ کیلومتر افزایش یافته است.

با توجه به شکل ۱۳.۴ می‌توان به این نتیجه رسید که با وجود متغیر نشان‌دار محدوده وسیع‌تری به‌هم وابسته‌اند و وابستگی فضایی نقاط بر اساس طیف رنگی ملموس‌تر است هر چه از مرکز زلزله دورتر می‌شویم تعداد زلزله‌ها کمتر می‌شوند.



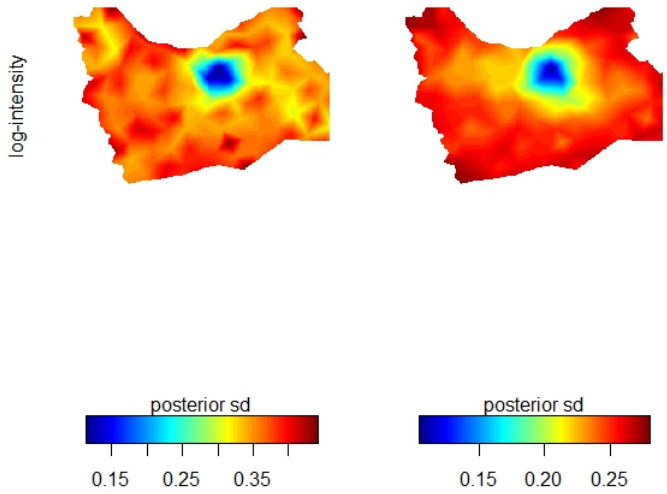
شکل ۱۳.۴: مقایسه دو مدل فرآیند LGC بدون نشان (سمت چپ) و مدل فرآیند LGC نشان‌دار (سمت راست) بر اساس میانگین پسین آن‌ها.

با مقایسه این دو مدل بر اساس انحراف معیار برآوردها، با وجود متغیر نشان‌دار متوجه کاهش چشم‌گیر خطا و برآورد دقیق‌تر پارامترها خواهیم شد که در شکل ۱۴.۴ نیز قابل مشاهده است.

۱.۵.۴ فرآیند LGC فضایی ـ زمانی

با توجه به مطالبی که در فصل سوم بحث شد گاهی در عمل تحلیل داده‌هایی مورد نظر است که علاوه بر وابستگی فضایی، در طول زمان و در فضای مورد مطالعه نیز به یکدیگر وابسته هستند. بنابراین نیاز است که از مدل سلسله مراتبی فضایی – زمانی استفاده شود. در این مدل‌ها مدل‌بندی متغیرهای پاسخ با استفاده از یک میدان تصادفی گاوسی انجام می‌شود و فرض می‌کنیم خطای اندازه‌گیری به صورت یک میدان تصادفی گاوسی نوفه سفید است. وابستگی فضایی – زمانی با یک میدان پنهان تصادفی، فضایی –زمانی که اصولا به صورت میدان حالتی که با زمان تغییر می‌کند و اتو رگرسیو مرتبه اول است، وارد مدل می‌شود. میدان مذکور را معمولا گاوسی با تابع کواریانس مترن در نظر می‌گیرند.

برای انجام یک مدل فضایی – زمانی ابتدا آن را به عنوان یک مدل SPDE برای حوزه فضایی و یک مدل AR(۱) برای اندازه‌گیری زمان در نظر می‌گیریم. مدل فضایی ـ زمانی با استفاده از ضرب کرونکر بین ماتریس دقت مکان و زمان انجام می‌شود. چون زمان در الگوهای نقطه‌ای

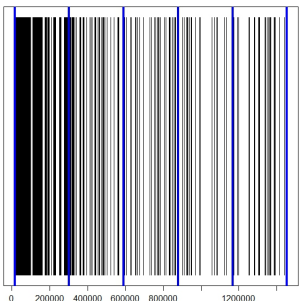


شکل ۱۴.۴: مقایسه دو مدل فرآیند LGC بدون نشان (سمت چپ) و مدل فرآیند LGC نشان‌دار (سمت راست) بر اساس انحراف معیار پسین آن‌ها

پیوسته است ما می‌توانیم از طریق گسسته‌سازی به روش SPDE و با استفاده از توابع مبنا تحلیل مورد نظر را انجام دهیم. برای برازش مدل پیوسته فضایی ـ زمانی با استفاده از روش SPDE، نیاز به تعریف مجموعه‌ای از گره‌ها در طول زمان داریم. تعداد گره انتخابی ما برای داده‌های مورد نظر ۶ عدد می‌باشد که اگر مبدا زمان را ۲۰۱۲/۸/۱ در نظر بگیریم زمان‌های زیر را شامل می‌شود:

جدول ۶.۴: گره‌های انتخابی

t	2012/8/11	2013/2/27	2013/9/15	2014/4/3	2014/10/19	2015/5/7



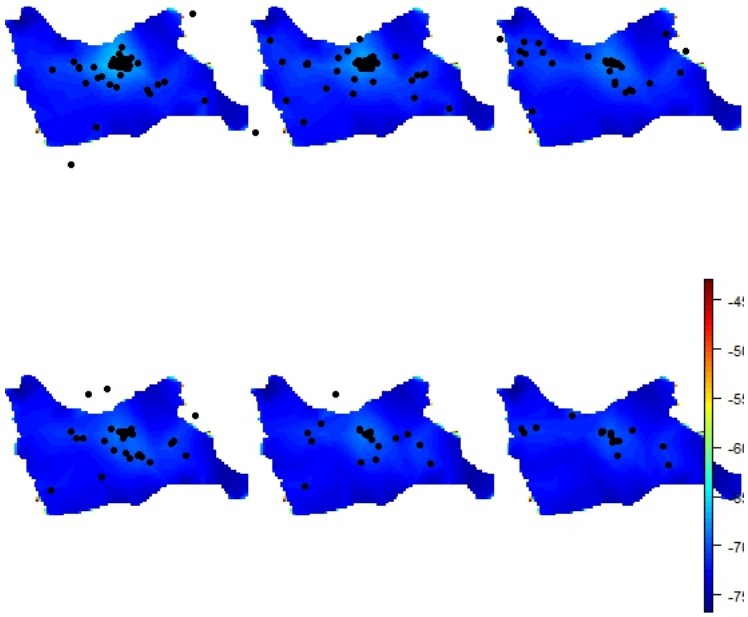
شکل ۱۵.۴: زمان رخ داد هر زلزله (سیاه) و گره‌های مورد استفاده برای استنتاج (آبی).

LGCP فضایی ـ زمانی بدون متغیر نشان‌دار

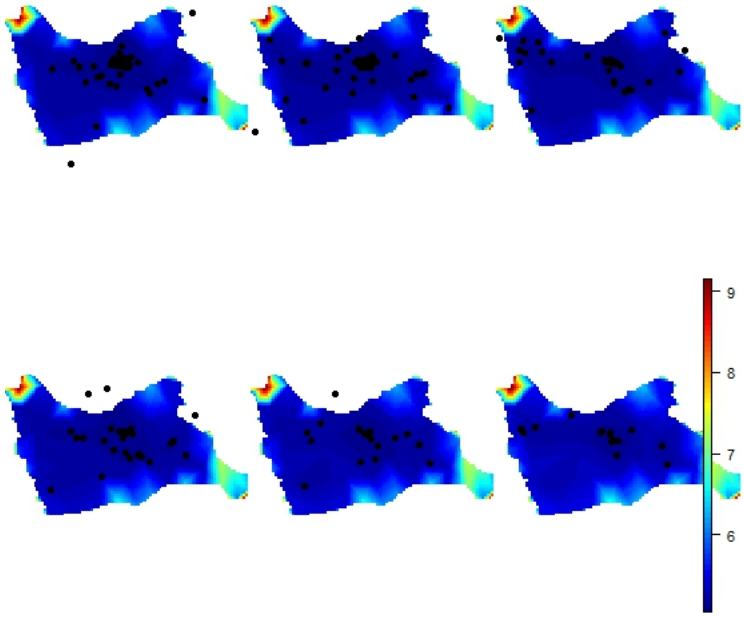
اکنون مدل را تنها با اثر فضایی–زمانی، بدون متغیر نشان‌دار برازش می‌دهیم. در این حالت داریم:

$$\zeta = \beta_0 + f(s,t) + u \qquad (۴.۴)$$

با توجه به شکل۱۶.۴ که پهنه‌بندی نقاط زلزله بر اساس میانگین پسین اثر فضایی–زمانی در زمان‌های اول تا ششم است می‌توان گفت که بیشترین زلزله در زمان‌های اول تا چهارم است. همچنین برآورد بیزی اثر فضایی در این شش زمان نشان می‌دهد در ناحیه مرکزی جاذبه وجود دارد و تعداد زلزله‌ها، بیشتر در مکان‌هایی که تجمع نقاط بیشتر است رخ می‌دهد. پهنه‌بندی انحراف معیار پسین اثر فضایی–زمانی در شکل ۱۷.۴ قابل مشاهده است.



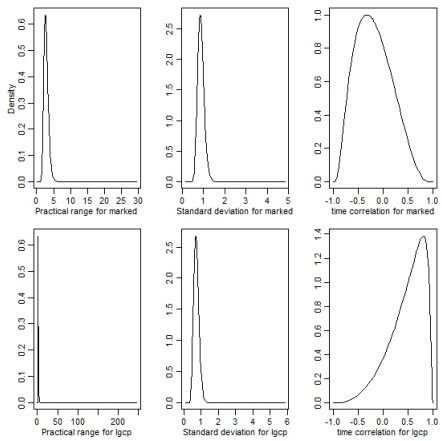
شکل ۱۶.۴: پهنه‌بندی میانگین پسین اثر فضایی–زمانی در زمان‌های اول تا ششم



شکل ۱۷.۴: پهنه‌بندی انحراف معیار پسین اثر فضایی–زمانی در زمان‌های اول تا ششم

LGCP فضایی ـ زمانی با متغیر نشان‌دار

در این مدل علاوه بر اثر فضایی–زمانی، متغیر نشان‌دار که بزرگای زلزله است را وارد مدل می‌کنیم و استنباط را ادامه می‌دهیم. متغیر نشان‌دار به روش SPDE مدل‌بندی می‌شود. شکل ۱۸.۴ چگالی حاشیه‌ای پسین پارامترها را نشان می‌دهد.

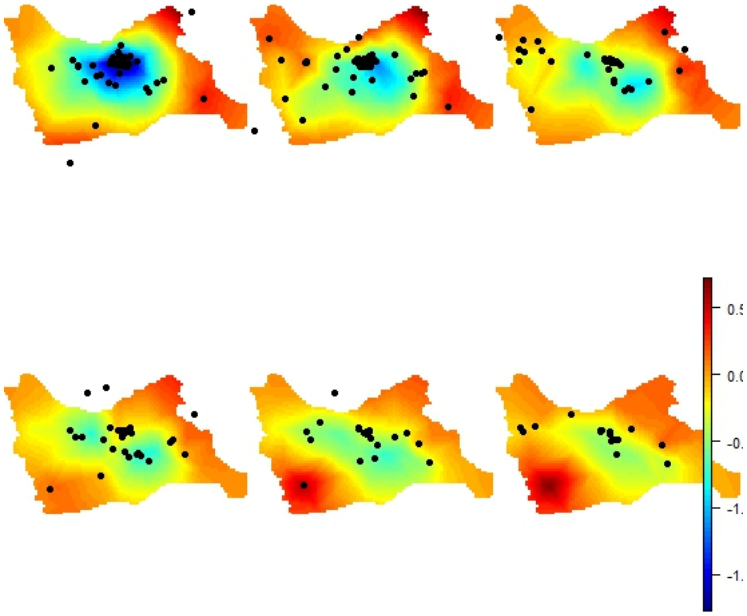


شکل ۱۸.۴: چگالی حاشیه‌ای پسین پارامترها

با اثر دادن متغیر نشان‌دار در مدل، خوشه‌ای بودن در شکل‌های ۱۹.۴ و ۲۰.۴ محسوس

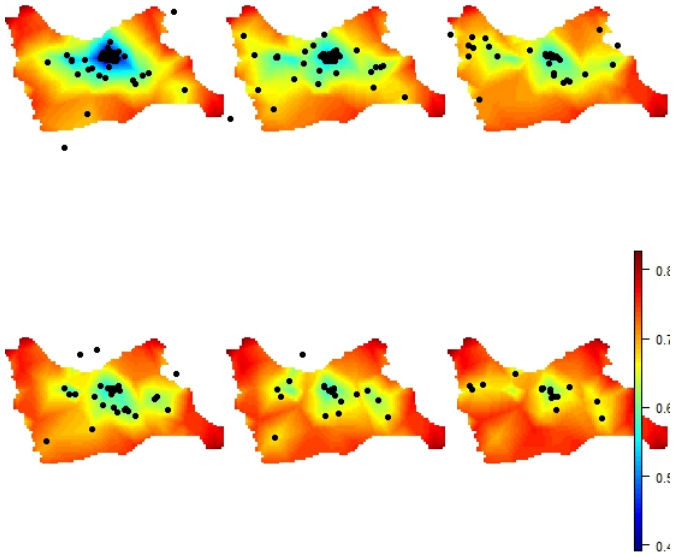
است و با مقایسه مدل نشان‌دار و مدل بدون متغیر نشان‌دار می‌توان نتیجه گرفت که نقاط در مناطق میانی در هر دو مدل تمایل به قرار گرفتن در کنار یکدیگر دارند که خوشه‌ای بودن الگو را نشان می‌دهد.

این تفاسیر را در پهنه‌بندی زمین‌لرزه‌های ناحیه شمال غرب ایران بر اساس میانگین پسین تابع شدت نیز می‌توان مشاهده کرد (شکل (۲۱.۴)).

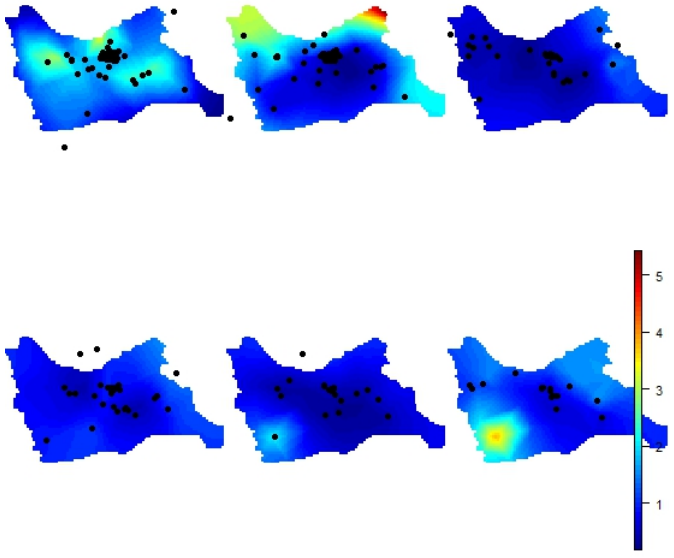


شکل ۱۹.۴: پهنه‌بندی میانگین پسین اثر فضایی_زمانی در زمان‌های اول تا ششم

با توجه به نتایج حاصل شده از این تحقیق و مطابقت پهنه‌بندی‌های حاصل با گسل‌های مربوط به شمال غرب ایران می‌توان این‌گونه استنباط کرد که بیشترین تمرکز در جهت شمال غربی، جنوب شرقی، از خوی تا سراب در رابطه با گسل شمال تبریز و همچنین در راستای شمالی جنوبی در شمال تبریز به طرف مرز ایران و روسیه و تا چند کیلومتری قفقاز قابل ملاحظه می‌باشد در صورتی که در نواحی دیگر این تمرکز کمتر و لرزه‌ها به صورت پراکنده می‌باشد.



شکل ۲۰.۴: پهنه‌بندی انحراف معیار پسین اثر فضایی_زمانی در زمان‌های اول تا ششم



شکل ۲۱.۴: پهنه‌بندی زمین‌لرزه‌های ناحیه شمال غرب ایران بر اساس میانگین پسین تابع شدت

۶.۴ نتیجه‌گیری

در این پایان‌نامه، فرآیندهای LGC به‌عنوان فرآیندهای منعطف برای مدل‌بندی الگوهای نقطه‌ای فضایی معرفی شدند. از تحلیل بیزی این فرآیندها استقبال بیشتری شده است. اما استنباط بیزی معمول در این فرآیندها به کمک الگوریتم‌های MCMC انجام می‌شود. روش تقریبی INLA جانشین جدید کارایی است که در کنار سرعت بالا، از مشکلات همیشگی همراه الگوریتم‌های MCMC، شامل همگرایی کند و آمیختگی ضعیف، دور است. با استفاده از این روش تقریبی، مدل‌بندی نقاط زلزله‌خیز ناحیه شمال غرب ایران با یک فرآیند LGC نشان‌دار را اجرا و نتایج رو تفسیر کردیم.

پیوست آ

آ.۱ فرآیند تصادفی گاوسی

فرآیند تصادفی

فرآیند تصادفی مجموعه‌ای از متغیرهای تصادفی با اندیس مرحله یا زمان است که وضعیت یک پدیده آزمایش تصادفی را در طول یک دوره نمایش می‌دهد. متغیرهای تصادفی و اندیس آن‌ها می‌توانند از نوع گسسته و پیوسته و نیز چند بعدی باشند. فرض کنید X_t متغیر تصادفی متناسب با t باشد که در فضای پارامتر T و فضای وضعیت S تعریف شده باشد. در این صورت، مجموعه متغیرهای تصادفی $\{X_t; t \in T\}$ را یک فرآیند تصادفی گویند که در آن X_t به ازای هر $t \in T$ یک متغیر تصادفی است.

فرآیند تصادفی گاوسی

فرآیند تصادفی $\{X_t, t \in T\}$ را در نظر بگیرید. هرگاه هر ترکیب خطی متناهی از متغیرهای تصادفی X_t دارای توزیع نرمال باشد، X_t را یک فرآیند گاوسی (نرمال) یا به عبارت دیگر میدان تصادفی گاوسی[۱] (GRF) می‌نامند. این فرآیندها کاربردهای مختلفی در علوم مختلف دارند.

[۱] Gaussian Random Field

آ.۲ فرآیند نقطه‌ای گیبز

فرآیندهای نقطه‌ای گیبز، اولین بار در نظریه فیزیک آماری ظهور کرد. در آمار ریاضی، این فرآیندها به‌عنوان مدل‌هایی برای مطالعه الگوهای نقطه‌ای فضایی به‌کار می‌روند (بدلی و همکاران، ۲۰۱۲a).

آ.۳ دستورات نرم‌افزار R مربوط به بخش ۴.۴

- فراخوانی بسته‌هایی لازم

```
library(INLA)
library(raster)
library(sp)
library(ggplot2)
library(spatstat)
library(fields)
library(rgeos)
```

- ساختن مشبکه

```
(nv <- (mesh <- inla.mesh.2d(
loc.domain=burbdy, offset=c(.3, 1),
max.edge=c(.5, 0.7), cutoff=.05))$n)
plot(mesh)
lines(burbdy, col="red")
points(locs, col=4, pch=10); lines(burbdy, col=2)

spde <- inla.spde2.pcmatern(
mesh=mesh, alpha=2, ### mesh and smoothness parameter
prior.range=c(0.05, 0.01), ### P(practic.range<0.05)=0.01
prior.sigma=c(1, 0.01)) ### P(sigma>1)=0.01
sum(diag(spde$param.inla$M0))
```

```
sum(w <- sapply(1:length(dmesh), function(i) {
if (gIntersects(dmesh[i,], domainSP))
return(gArea(gIntersection(dmesh[i,], domainSP)))
else return(0)
}))
```

- ماتریس مشاهدات

```
n <- nrow(locs)
y.pp <- rep(0:1, c(nv, n))
e.pp <- c(w, rep(0, n))

lmat <- inla.spde.make.A(mesh, locs)
imat <- Diagonal(nv, rep(1, nv))
A.pp <- rBind(imat, lmat)

stk.pp <- inla.stack(data=list(y=y.pp, e=e.pp),
A=list(1,A.pp), tag='pp',
effects=list(list(b0=rep(1,nv+n)), list(i=1:nv)))
```

- برازش مدل

```
pp.res <- inla(y ~ 0 + b0 + f(i, model=spde),
family='poisson', data=inla.stack.data(stk.pp),
control.predictor=list(compute=TRUE, A=inla.stack.A(stk.pp)),
E=inla.stack.data(stk.pp)$e, verbose=TRUE,
control.inla=list(strategy='gaussian', int.strategy='eb'))
```

مراجع

[۱] ارشدی‌پور، ا، (۱۳۹۳)، «روش نمونه‌گیری برشی به‌عنوان یکی از روش‌های MCMC»، **مجله اندیشه آماری**، جلد ۱۴، شماره ۲، ص ۱۵-۳.

[۲] جلیلیان، م، (۱۳۹۵)، پایان‌نامه ارشد: «پیش‌بینی فضایی برای فرآیندهای کاکس لگ‌گاوسی»، دانشکده علوم‌پایه، دانشگاه رازی کرمانشاه.

[۳] درویش‌زاده، ع، (۱۳۷۰)، «**زمین‌شناسی ایران**»، نشر دانش امروز، تهران.

[۴] ذکاء، ی، (۱۳۶۸)، «**زمین‌لرزه‌های تبریز**»، جلد اول، چاپ اول، انتشارات کتاب‌سرا، ص ۸.

[۵] شیاسی، ف، باغیشنی، ح، اقبال، ن، (۱۳۹۶)، «تحلیل بیزی تقریبی الگوی نقطه‌ای فضایی زمین‌لرزه‌های شمال غرب ایران با فرآیندهای کاکس لگ‌گاوسی»، مجموعه مقالات دومین سمینار آمار فضایی و کاربردهای آن، دانشگاه صنعتی شاهرود، شاهرود، ص ۱۶۶-۱۵۹.

[۶] محمدزاده، م، (۱۳۹۱)، «**آمار فضایی و کاربردهای آن**»، چاپ اول، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

[7] Abramowitz, M. and Stegun, I. (1972), "**Handbook of Mathematical Functions**", Courier Dover Publications.

[8] Adler, R. J. (1981), **The Geometry of Random Fields (Wiley Series in Probability and Statistics)**.

[9] Anderssen, R.S., Baddeley, A., DeHoog, F.R. and Nair, G.M. (2014), " Solution of an integral equation arising in spatial point process theory", **Journal of Integral Equations and Applications**, 26 (4) 437–453.

[10] Baddeley, A. annd Turner,R. (2005), "spatstat: An R Package for Analyzing Spatial Point Patterns", **Journal of Statistical Software**, 12.

[11] Baddeley, A. and Nair, G. (2012a), ”Fast approximation of the intensity of Gibbs point processes”, **Electronic Journal of Statistics**, 6 1155–1169.

[12] Baddeley, A. and Nair, G. (2016), Poisson-saddlepoint approximation for spatial point processes with infinite order interaction. **Journal of Applied Probability**, 54(04):1008-1026.

[13] Baddeley, A., Rubak, E and Turner R. (2015), **Spatial Point Patterns Methodology and Applications with R**, 203-220.

[14] Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf. N. (2012),” BayesX – Software for Bayesian inference in structured additive regression model”, **Journal of Statistical Software** Version 2.1.Available from http://www.stat.uni-muenchen.de/bayesx.

[15] Bernner, S.C. and Scott, R. (2007), **”The Mathematical Theory of Finite Element Methods”**, Springer,New York.

[16] Bithell, J.F. (1990), ”An application of density estimation to geographical epidemiology”, **Statistics in Medicine**, 9:691–701.

[17] Brezger, A., Kneib, T., and Lang. S. (2005), ”BayesX Manuals, Ludwig-Maximilians-University”, Munich, URL http://www.stat.uni-muenchen.de/ bayesx/.

[18] Casella,G. and George, E. (1992), ”Explaining theGibbs Sampler”, **The American Statistician**, 46, 167–174.

[19] Casella, G. and Berger, R. (2002), **Statistical Inference**, Duxbury, Thomson Learning.

[20] Cheng, A. H.-D and Cheng, D. T. (2005), ”Heritage and early history of the boundary element method”, **Engineering Analysis with Boundary Elements**, 29 (3), 268.

[21] Chib, S. and Greenberg, E. (1995), ”Understanding the Metropolis-Hastings algorithm” , **Journal of the American Statistician**, 49,327-335.

[22] Condit, R., Hubbell, S. P. and Foster, R. B. (1996), ”Changes in tree species abundance in a neotropical forest: impact of climate change”, **Journal of Tropical Ecology**. 12, 231–256.

[23] Condit, R. (1998), **Tropical Forest Census Plots**, Springer-Verlag and R. G. Landes Company, Berlin,Germany and Georgetown, Texas.

[24] Cox, D. R. (1955), ”Some Statistical Methods Connected with Series of Events”, **Journal of the Royal Statistical Society**, 17 (2), 129–164

[25] Cox, D. R. (1964), “On the Estimation of the Intensity Function of a Stationary Point Process”, **Journal of the Royal Statistical Society**. Ser. B, 27, 332–337.

[26] Cressie, N. (1993), **Statistics for Spatial Data**, John Wiley and Sons.

[27] Depaoli, S., Clifton, J. P. and Cobb, P. R. (2016). “Just Another Gibbs Sampler (JAGS) Flexible Software for MCMC Implementation”, **Journal of Educational and Behavioral Statistics**, 1076998616664876.

[28] Diggle, P.J. (1985),” A kernel method for smoothing point process data”, **Journal of the Royal Statistical Society**, Series C (Applied Statistics), 34:138–147.

[29] Diggle, P. J. (2003), **Statistical Analysis of Spatial Point Pattern**, New York, Oxford University Press.

[30] Eidsvik,J., Martino, S. and Rue, H. (2009), ”Approximate Bayesian ference in Spatial Generalized Linear Mixed Models”, **Scandinavian Journal of Statistics**, 36, 1-22.

[31] Gelfand, A. and Smith, A. (1990), ”Sampling-based approaches to calculating marginal densities”, **Journal of the American Statistical Association**, 85, 398-409.

[32] Geman, S. and Geman, D. (1984), ”Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images”, **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 6, 721-741.

[33] Gilks, W.R, Richardson, S. and Spiegelhalter, D.J. (1996), **”Markov Chain Monte Carlo in Practice”**, London : Chapman and Hall.

[34] Gneiting, T. and Guttorp, P. (2010). **Continuous parameter spatio-temporal processes**. In Handbook of Spatial Statistics, eds. A.E. Gelfand, P.J. Diggle, M. Fuentes, P. Guttorp, 427–436. Boca Raton: Chapman and Hall/CRC Press.

[35] Hastings, W. (1970), **Monte Carlo sampling methods using Markov chains and their application**, Biometrika, 57, 97-109.

[36] Hesterberg, T.C. (1987), ”Importance Sampling in Multivariate Problems”,**Proceedings of the Statistical Computing Section of the American Statistical Association**, 412-417.

[37] Hubbell, S. P. and Foster, R. B. (1983), ”Diversity of canopy trees in a neotropical forest and implications for conservation. In Tropical rain forest: ecology and management (eds S. L.

Sutton, T. C. Whitmore and A. C. Chadwick)", **Blackwell Scientific Publications**, Oxford, 25–41.

[38] Illian, J. B. and Rue, H. (2012), "A toolbox for fitting complex spatial point process models using integrated Laplace transformation (INLA)", **Institute of Mathematical Statistics**, 6, 1499–1530.

[39] K. Kitanidis, P. (1986), "Parameter uncertainty in estimation of spatial functions:Bayesian analysis", **Water Resources Research**, 22(4), 499-507.

[40] Lindgren, F., Rue, H. and Lindström, J. (2011), "An explicit link between Gaussian felds and Gaussian Markov random felds: the stochastic partial differential equation approach". **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, 73(4), 423-498.

[41] Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000), **WinBUGS – a Bayesian modelling framework**: Concepts, structure, and extensibility, Statistics and Computing, 10, 325–337.

[42] Ma, C. (2003), "Families of spatio-temporal covariance models", **Journal of Statistical Planning and Inference**, 116, 489–501.

[43] Ma, C. (2008), "Recent developments in the construction of spatio-temporal covariance models, Stochastic Environmental Research and Risk Assessment", 22, S39–S47.

[44] Matheron, G. (1962), **Traite de geostatistique appliquee, Tome I: Memoires du Bureau de Recherches Geologiques et Minieres**, no. 14, Editions Technip, Paris, 333 p.

[45] Martins, T.G., Simpson, D., Lindgren, F. and Rue, H. (2013), **Bayesian computing with INLA: New features,Computational Statistics and Data Analysis**, 67, 68–83.

[46] Metropolis, N. Rosenbluth, A., Teller, A. and Teller, E. (1953), "Equation of state calculations by fast computing machines", **Journal of Chemical Physics**, 21(6), 1087–1092.

[47] Moller, J. and Waagepetersen, R. P. (2004), **Statistical inference and simulation for spatial point processes**. CRC Press.

[48] Moller, J. Syversveen, A. R. and Waagepetersen, R. P. (1998), "Log gaussian cox processes", **Scan dinavian journal of statistics**, 25(3), 451-482.

[49] Neal R. N. (2003), ”Slie sampling, Annals of Statistis”, **Institute of Mathematical Statistics**, 31(3), 705-767.

[50] Nelder, J.A. and Wedderburn, R.W.M. (1972), ”Generalized linear models”. **Journal of the Royal Statistical Society**, A, 135, 370-84.

[51] Ntzoufras, I. (2009), **Bayesian Modeling Using WinBUGS**, John Wiley and Sons.

[52] Ripley, B. D. (1981), **Spatial Statistics. Willey**, New York.

[53] Robert, C. and Casella, G. (2004), **Monte Carlo Statistical Methods**, Springer.

[54] Rodrigues, A. and Diggle, P.J. (2010), ”A class of convolution-based models for spatio-temporal processes with non-separable covariance structure”.**Scandinavian Journal of Statistics**, 37, 553–567.

[55] Rue, H. and Held, L. (2005), **Gaussian Markov Random Fields: Theory and Applications, London**: Chapman and HallCRC Press.

[56] Rue, H. and Martino, S. (2007), ”Approximate Bayesian inference for hierarchical Gaussian Markov random fields models”. **Journal of Statist**. Planng Inf., 137, 3177–3192.

[57] Rue, H., Martino, S., and Chopin, N. (2009), ”Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion)”. **Journal of the Royal Statistical Society** (series B), 71, 319-392.

[58] Skaug, H. J., Øien, N., Schweder, T. and Bøthun, G. (2004), ”Abundance of minke whales (balaneoptera acutorostrata) in the northeast Atlantic: variability in time and space”, **Canadian Journal of Fisheries and Aquatic Science**. 61,870–886.

[59] Stocklin J. (1977), ”Structural Correlation of the Alpine ranges between Iran and Central Asia”, **Soc.Geol.Fr., Mem. Hors**. Ser, p 333.353.

[60] Sturtz, S., Ligges, U. and Gelman, A. (2005), ”R2WinBUGS: A Package for Running WinBUGS from R”, **Journal of Statistical Software**,Volume 12, Issue 3, 13534.

[61] Takahashi, K., Fagan, J. and Chen, M.S. (1973). ”Formation of a sparse bus impedance matrix and its application to short circuit study”. IEEE Power Engineering Society.pages 63–69. Papers presented at the 1973 Power Industry Computer Application Conference in Minneapolis, Minnesota.

[62] Tierney, L. and Kadane, J.B. (1986), ”Accurate approximations for posterior moments and marginal densities”, **Journal of the American Statistical Association**, 81, 82–86.

[63] Waagepetersen, R. and Schweder, T. (2006), Likelihood-based inference for clustered line transect data, **Journal of Agricultural, Biological, and Environmental Statistics**, 11, 264–279.

[64] Walters, P. (1982), ”An Introduction to Ergodic Theory”, **Springer. ISBN**, 0-387-95152-0.

[65] Wegener, A. (1915), ”Die Entstehung der Kontinente and Ozeane”. Vieweg, Braunschweig.

[66] Whittle, P. (1954), ”On stationary processes in the plane”, **Biometrika**, 41(3/4), 434–449.

[67] Whittle, P. (1963), ”Stochastic Processes in Several Dimensions”, **Bulletin of theInternational Statistical Institute**, 40, 974–994.

واژه‌نامه انگلیسی به فارسی

فرایند دوجمله‌ای ..... Binomial point process
مشخص‌سازی ..... Characterization
خوشه ..... clustering
به‌طور کاملا تصادفی فضایی ..... Complete spatial randomness
فرایند نقطه‌ای کاکس ..... Cox point process
چگال ..... Dense
همه‌گیرشناسی ..... Epidemiology
داده‌های زمین‌آماری ..... Geostatistical data
همگن ..... Homogeneous
ناهمگن ..... Heterogeneous
تابع شدت ..... Intensity function
برآورد هسته ..... Kernel estimation
داده‌های مشبکه‌ای ..... Lattice data
فرایند کاکس لگ‌گاوسی ..... log Gaussian Cox process
الگو نقطه‌ای نشان‌دار ..... Marked point processes
الگو نقطه‌ای چندنوعی ..... Multivariate point processes
پنجره مشاهدات ..... Observation Window
تصویر کوچک ..... Pixel
فرایند نقطه‌ای پواسون ..... poisson point process
الگوهای نقطه‌ای ..... Point pattern
شمارش پنجره‌های مربعی ..... Quadrat counts
منظم ..... Regularity
فضای وضعیت ..... State Space
استریوگرافیک قطبی جهانی ..... Universal polar stereographic
تابع احتمال پوچی ..... Void probability function
منطقه ..... Zone

Abstract

Spatial point pattern data has various applications in the real world. For example, identification of earthquake points and Study of different species plants. Intensity function is the main Property in these data analysis. Several statisticians suggeste different Approaches to modeling of this function such as, Binomial point processes, Poisson and Cox. According to nature of these data and cox processes flexibility, in this study we considered Subcategories of these processes as known log gaussian cox processes. log gaussian cox processes act very flexible because of simple improvise of auxiliary variables in its inside. Classic deduction in Leggings processes need to complicated calculations to earn approximate of the likelihood function.So because of this, researcher's usual approaches for fit intensity function by Leggings processes is Bayesian. Since of posterior distribution form in these processes isn't clear, the usual tools to fit modeling is Algorithms of the Markov Chain Mont Carlo (MCMC). Performance of these algorithms need to Programming skills and is serious in Convergence and computation time. A replacement approach is using of Bayesian approximate method as known Laplace approximation (INLA). This approximation in compare with MCMC algorithms is very quick and its result is the same in accuracy. In this thesis, we use the approximation method INLA for fitting the log gaussian cox processes models. To illustrate the application of the theoretical topics, we analyze the data of the earthquake point pattern of the northwest region of Iran with thelog gaussian cox processes model.

Keyword: Point Processes, Log Gaussian Cox Process, Intensity Function, Function K, Integrated Nested Laplace Approximation.

Shahrood University of Technology

Faculty Of Mathematical Sciences

MSc Thesis in: Statistics

Approximate Bayesian inference of complex spatial point pattern models using Laplace approximation

By: Fateme Shiasi

Supervisor

Dr. Hossein Baghishani

Advisor

Dr. Negar Eghbal

January 2018