دانشکده علوم ریاضی
گروه آمار

پایان‌نامه کارشناسی ارشد

استنباط بیزی تقریبی در فرآیندهای پواسون ناهمگن با کاربرد در تحلیل پیشامدهای بازگشتی

فاطمه کارگر

استاد راهنما

دکتر حسین باغیشنی

شهریور ۱۳۹۵

این پایان نامه تقدیم به آنانی که امروزم مدیون تابش پرتو پر نور

عشق آن ها بر زندگیم است

تقدیم به بهترین موهبت های خداوند در زندگیم

پدر، مادر و همسرم

که در خشش نگاه پر امیدشان روشنی بخش راهم بود، بهترین ها را

برایشان آرزو می کنم.

سپاس‌گزاری...

سپاس خداوندی را که اول همه آثار هستی اوست و قبل از او اولی نبوده و آخر است بی‌آنکه پس از او آخری باشد، خداوندی که دیده بینندگانش از دیدنش قاصر و اندیشه و فهم وصف‌کنندگان از وصفش عاجز است.

برخود لازم می‌دارم از استاد راهنمای بزرگوارم جناب آقای دکتر حسین باغیشنی، که راهنمایی‌های سازنده ایشان فراهم آوردن این تحقیق را میسر نموده است، تشکر نمایم. همچنین مراتب تشکر و قدردانی خود را از آقای دکتر محمد آرشی و دکتر داود شاهسونی که داوری تحقیق حاضر را بر عهده داشته‌اند و از راهنمایی‌های خود اینجانب را بهره‌مند ساخته‌اند، ابراز می‌دارم.

فاطمه کارگر
شهریور ۱۳۹۵

تعهدنامه

این‌جانب **فاطمه کارگر** دانشجوی کارشناسی ارشد رشته **آمار** **دانشکده علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **استنباط بیزی تقریبی در فرآیندهای پواسون ناهمگن با کاربرد در تحلیل پیشامدهای بازگشتی** تحت راهنمایی **دکتر 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 نشده است.
- 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 خواهد 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 کارگر
شهریور ۱۳۹۵

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته‌شده) 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

تحلیل داده‌های بازگشتی در دیدگاه بیزی، به دلیل پیچیدگی مدل‌های معقول برای تحلیل آن‌ها، وابسته به استفاده از روش‌های مبتنی بر نمونه‌گیری، از جمله الگوریتم‌های MCMC، است. اما این الگوریتم‌ها همیشه با مشکلاتی مانند سرعت کند، ناتوانی در تشخیص همگرایی زنجیر و رفتارهای ضعیف آمیختگی، همراه هستند. یک رهیافت جانشین برای این الگوریتم‌ها، روش تقریب لاپلاس آشیانه‌ای جمع‌بسته (INLA) است که توسط رو و همکارانش (۲۰۰۹) ارایه شد. این روش از مشکلات الگوریتم‌های MCMC مبرا بوده و سرعت محاسبات در آن بسیار بالاست. از طرفی، اختلاف دقت نتایج حاصل از آن و الگوریتم‌های MCMC بسیار ناچیز و حتی در مواردی بالاتر است. در این پایان‌نامه، مدل‌بندی بیزی تقریبی پیشامدهای بازگشتی بر اساس فرآیند پواسون ناهمگن و به کمک روش INLA، انجام شده است. عملکرد مدل و روش استنباطی نیز توسط مثال‌های شبیه‌سازی و واقعی، ارزیابی شده‌اند.

کلمات کلیدی: پیشامدهای بازگشتی، تقریب لاپلاس آشیانی جمع‌بسته، فرآیند پواسون ناهمگن، مدل مخاطره متناسب کاکس، میدان تصادفی مارکوفی گاوسی.

پیش گفتار

مطالعه پیشامدهایی که بیش از یک بار رخ می‌دهند، موضوع مورد علاقه در بسیاری از زمینه‌های علمی از جمله پزشکی، قابلیت اطمینان قطعات تولیدی و صنعتی، و جرم‌شناسی است. به این نوع پیشامدها، پیشامدهای چندگانه اطلاق می‌شود. پیشامدهای چندگانه در دو گروه قابل دسته‌بندی هستند: ۱) پیشامدهایی که از یک نوع بوده و برای یک مورد (شیء یا فرد) در طول زمان تکرار می‌شوند و ۲) پیشامدهایی با انواع متفاوت که برای یک فرد یا شیء رخ می‌دهند. آن‌چه که در این پایان‌نامه مورد نظر است، تحلیل پیشامدهای چندگانه نوع اول است که به داده‌های بازگشتی معروف هستند. چند مثال از این نوع پیشامدها، عبارتند از حملات آسم در بیماران تنفسی، بازگشت تومور بعد از جراحی در مطالعات سرطان، حملات قلبی پی در پی، تکرار جرم توسط افراد، و شناسایی مکرر یک خطا در یک سیستم نرم‌افزاری. مطالعات مختلف انجام‌شده برای تحلیل داده‌های حاصل از پیشامدهای بازگشتی، دو دیدگاه کلی را شامل می‌شوند: مدل‌بندی بر اساس تابع شدت (نرخ خطر یا تابع مخاطره) یا مدل‌بندی بر پایه تعداد بازگشت‌ها در یک بازه مشخص.

در سال‌های اخیر، تحلیل داده‌های بازگشتی با استفاده از مدل‌های نیمه‌پارامتری، به دلیل انعطاف بالای آن‌ها در لحاظ کردن همزمان اثرات ایستا و پویای متغیرهای تبیینی، خیلی معمول شده است. تحلیل بیزی این مدل‌ها، به دلیل پیچیدگی مدل‌های معقول برای تحلیل آن‌ها، وابسته به استفاده از روش‌های مبتنی بر نمونه‌گیری، از جمله الگوریتم‌های MCMC، است. اما این الگوریتم‌ها همیشه با مشکلاتی مانند سرعت کند، ناتوانی در تشخیص همگرایی زنجیر و رفتارهای ضعیف آمیختگی، همراه هستند. یک رهیافت جانشین برای این الگوریتم‌ها، روش تقریب لاپلاس آشیانه‌ای جمع‌بسته (INLA) است که در سال ۲۰۰۹ توسط رو و همکارانش ارایه شد. این روش از مشکلات الگوریتم‌های MCMC مبرا بوده و سرعت محاسبات در آن بسیار بالاست. از طرفی، اختلاف دقت نتایج حاصل از آن و الگوریتم‌های MCMC بسیار ناچیز و حتی در مواردی بالاتر است.

در این پایان‌نامه، بر اساس روش آکرکار و همکاران (۲۰۱۲) فرض می‌کنیم پیشامدهای بازگشتی بر اساس یک فرآیند پواسون ناهمگن، که یک مدل نیمه‌پارامتری برای تابع شدت فراهم می‌سازد، رخ می‌دهند. نشان می‌دهیم این مدل عضوی از رده مدل‌های گاوسی پنهان است که روش INLA برای آن‌ها قابل استفاده است. سپس از روش INLA در تحلیل این پیشامدها بهره می‌بریم و عملکرد مدل و روش استنباطی را به کمک مثال‌های شبیه‌سازی و واقعی، ارزیابی می‌کنیم. با توجه به این مقدمه، ساختار پایان‌نامه به صورت زیر تدوین شده است:

- در فصل اول، تعاریف، مفاهیم و مدل‌های مورد نیاز برای ورود به بحث اصلی را بیان می‌کنیم.
- در فصل دوم، روش INLA را برای انجام استنباط بیزی تقریبی در رده مدل‌های گاوسی پنهان معرفی می‌کنیم. همچنین ملاک‌های بیزی ارزیابی مدل‌ها را معرفی می‌کنیم.
- در فصل سوم، برای تحلیل داده‌های بازگشتی، مدل فرآیند پواسون ناهمگن را معرفی می‌کنیم.

نشان می‌دهیم این مدل عضوی از رده مدل‌های گاوسی پنهان است. بنابراین می‌توان از روش INLA برای برازش آن و استنباط بیزی مدل استفاده کرد.

- در فصل چهارم، با استفاده از دو مثال شبیه‌سازی و دو مثال واقعی که داده‌های مربوط به تومور پستان و ایست قلبی را شامل می‌شوند، عملکرد و توانایی رهیافت پیشنهادی پایان‌نامه را ارزیابی می‌کنیم. در پایان نیز به نتیجه‌گیری و ارایه چند پیشنهاد برای آینده تحقیق می‌پردازیم.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مفاهیم، تعاریف و مدل‌های پایه‌ای

در این فصل هدف اصلی پایان‌نامه، مدل‌ها و روش‌های مورد نظر ما برای نیل به این هدف معرفی می‌شوند.

۱.۱ مقدمه

تحلیل بقا یکی از عمده‌ترین تحلیل‌های آماری در حوزه تحقیقات بالینی است. عبارت داده‌های بقا[۱] برای توصیف داده‌هایی اطلاق می‌شود که زمان تا وقوع پیشامد خاصی را اندازه‌گیری می‌کنند. زمان تا وقوع پیشامد، متغیری با مقادیر مثبت و دارای توزیع پیوسته می‌باشد. با توجه به ساختار خاص داده‌های بقا، که شامل زمان‌های سانسور[۲] و بریدگی[۳] می‌باشند، روش‌های تحلیل بقا نیز از دیگر روش‌های آماری مجزا هستند.

داده‌های بقا دارای دو ساختار عمده هستند: داده‌های یک متغیره[۴] و چندمتغیره[۵]. در داده‌های یک متغیره زمان اندازه‌گیری‌شده مربوط به زمان تا وقوع یک پیشامد معین است و زمان‌های مشاهده‌شده بین افراد مورد مطالعه از هم مستقل هستند. عمده‌ترین حادثه مورد مطالعه در داده‌های یک متغیره

[۱] Survival data

[۲] Censoring

[۳] Truncation

[۴] Univariate data

[۵] Multivariate data

مرگ است. در مواردی که پذیره استقلال برقرار نباشد، بحث داده‌های چندمتغیره مطرح می‌شود. این نوع داده‌ها نیز به دو بخش تقسیم می‌شوند: داده‌های موازی[۶] و طولی[۷].

در مطالعاتی که افراد دارای صفات مشترک، همزمان با هم پیگیری می‌شوند، داده‌های حاصل داده‌های موازی یا داده‌های بقای خوشه‌ای نامیده می‌شوند. در این نوع داده‌ها هر فرد یک حادثه را تجربه می‌کند و زمان‌های وقوع به دلیل قرار گرفتن افراد در یک گروه یا خوشه، به هم وابسته هستند. این وابستگی ممکن است به دلیل شرایط محیطی یا ژنتیکی باشد که در بین خوشه‌های مختلف، متفاوت است. به‌عنوان مثال مطالعه زمان بقای اعضای جفتی بدن، مطالعه دوقلوها یا افراد یک خانواده از این نوع هستند.

در مطالعات بالینی گاهی افراد فقط یک حادثه از چندین حادثه ممکن را تجربه می‌نمایند و مشاهده یکی از این حوادث منجر به عدم مشاهده دیگر حوادث ممکن خواهد شد. به‌عنوان مثال برای فردی که در اثر یک تصادف بمیرد، امکان مشاهده مرگ در اثر حمله قلبی وجود ندارد. این نوع داده‌های بقا، داده‌های خطر رقابتی نامیده می‌شوند.

اما در داده‌های طولی بقا، داده‌های زمان چندگانه مطرح می‌شوند و در این مورد یک فرآیند تصادفی با چندین پیشامد مختلف برای هر شخص مشاهده می‌شود. معمولا در مطالعات بیماری‌های مزمن این نوع داده‌ها مشاهده می‌شوند. همچنین مطالعاتی که در آن‌ها هر فرد امکان تجربه متعدد یک پیشامد را در زمان‌های مختلف داشته باشد، داده‌ها از نوع داده‌های طولی بقا هستند و این نوع داده‌ها را داده‌های پیشامدهای بازگشتی[۸] نیز می‌نامند.

محققین بسیاری تحلیل داده‌های بقا را مورد مطالعه قرار داده‌اند. از آن جمله ویلیامز و پیترسون (۱۹۸۱) دو مدل شرطی نوع اول و دوم را در زمینه تحلیل پیشامدهای بازگشتی معرفی کردند و از مدل طبقه‌ای مخاطره متناسب[۹] (یا خطرات متناسب) جهت تبیین وابستگی میان مشاهدات مربوط به هر فرد استفاده کردند. اندرسون و گیل (۱۹۸۲) مدل فرآیندهای شمارشی[۱۰] را معرفی نمودند. در این روش از مدل رایج مخاطره متناسب کاکس[۱۱] استفاده شد و جهت تعدیل اثر وابستگی حاصل از پیشامدهای بازگشتی، یک تعمیم از براورد واریانس استوار، معرفی‌شده توسط وایت (۱۹۸۲)، مورد استفاده قرار گرفت. یک مدل حاشیه‌ای[۱۲] جهت تحلیل بقا در حالت وجود پیشامدهای بازگشتی توسط وی و لین (۱۹۸۹) ارائه گردید. در این روش نیز از مدل طبقه‌ای مخاطره متناسب کاکس استفاده شده است. نلسون (۲۰۰۳) مثال‌های متنوعی از وقوع پیشامدهای بازگشتی در زمینه‌های مختلف همچون پزشکی ارائه کرد. لاولس و نادئو (۱۹۹۵) و لین و همکاران (۲۰۰۰) مدل‌های رگرسیونی خطرات متناسب را در حالت وجود پیشامدهای تکراری معرفی و به‌کار بردند. مدل‌های آماری جهت تحلیل داده‌های پزشکی نیز در حالت پیشامدهای تکراری توسط کوک و لاولس (۲۰۰۲) ارائه شدند.

---

[۶] Parallel data

[۷] Longitudinal data

[۸] Recurrent events

[۹] Proportional hazard

[۱۰] Counting process

[۱۱] Cox proportional hazard

[۱۲] Marginal model

مدل مخاطره متناسب کاکس از رایج‌ترین روش‌های آماری جهت تحلیل داده‌های بقا است که یکی از پذیره‌های اساسی آن استقلال میان مشاهدات (زمان‌های بقا و سانسور) می‌باشد. اما در عمل، شرایطی به وجود می‌آیند که برقراری پذیره استقلال مشاهدات مورد تردید واقع می‌شود. به‌عنوان مثال در داده‌های بازگشتی که ویژگی عمده آن‌ها همبستگی بین مشاهدات است، اگر بدون توجه به این همبستگی از مدل‌های معمول مانند رگرسیون مخاطره متناسب برای تحلیل داده‌ها استفاده شود، برآوردها اریب و واریانس برآورد پارامترها کمتر از مقدار واقعی برآورد می‌شوند. بنابراین منطقی به نظر می‌رسد که با در نظر گرفتن مدلی که شامل اثرات وابسته به و مستقل از زمان است، مطالعه دقیق‌تری از وقوع آن پیشامد داشته باشیم. در همین راستا، مدل نیمه‌پارامتری رگرسیونی را با ترکیبی از ضرایب وابسته به و مستقل از زمان برای مدل‌بندی داده‌های بازگشتی بقا معرفی خواهیم کرد.

در این فصل ابتدا برخی از مفاهیم مورد نیاز را ارائه می‌کنیم و سپس به معرفی رده بزرگی از مدل‌های مخاطره متناسب کاکس برای مدل‌بندی داده‌های بازگشتی می‌پردازیم.

۲.۱ تحلیل بقا

تحلیل بقا مجموعه‌ای از روش‌های آماری برای تحلیل داده‌هایی است که متغیر پاسخ آن‌ها زمان سپری‌شده تا رخداد یک پیشامد است. منظور از زمان در این تعریف می‌تواند زمان شروع یک مطالعه تا زمان رخداد پیشامد مورد نظر یا سن فرد در زمان رخداد پیشامد، بر حسب سال، ماه، هفته و مشابه آن‌ها، باشد. پیشامد مورد نظر می‌تواند مرگ، بروز بیماری، عود بیماری، فروکش کردن نشانه‌های بیماری، بهبودی یا هر تجربه تعریف‌شده‌ای باشد که افراد با آن مواجه می‌شوند. در تحلیل بقا معمولاً به متغیر زمان، عنوان زمان بقا داده می‌شود؛ زیرا بقای هر یک از افراد تحت مطالعه بررسی می‌شود. همچنین به پیشامد مورد نظر عنوان شکست داده می‌شود به این دلیل که نوع پیشامد مورد نظر معمولاً مرگ، بروز بیماری یا برخی دیگر از تجربه‌های منفی برای افراد است. البته ممکن است پیشامد مورد نظر یک تجربه مثبت نیز باشد، مثل بازگشت به کار بعد از انجام یک عمل جراحی. در این تحلیل هر مشاهده دارای یک نقطه شروع، یک نقطه پایانی و یک واحد زمانی تعریف‌شده برای فاصله بین این دو نقطه است. به‌طور کلی اهداف تحلیل بقا عبارتند از: برآورد و تفسیر توابع بقا[۱۳] یا مخاطره، مقایسه توابع بقا یا مخاطره و ارزیابی رابطه بین متغیرهای توضیحی و زمان بقا (زمان شکست).

اولین دیدگاه‌های نظری روش‌های تحلیل بقا که به سرعت گسترش و کاربردهای فراوانی یافتند، مبتنی بر روش‌های کلاسیک مرسوم به جداول عمر هستند که توسط کاپلان و مایر و همکاران آن‌ها در سال ۱۹۵۸ ارائه شدند. آن‌ها در بررسی‌های خویش روشی ناپارامتری را برای برآورد احتمال‌های بقا، در حضور داده‌های سانسور از راست[۱۴]، ارائه کردند. این روش‌ها به سرعت مورد توجه سایر آماردانان قرار گرفتند. از جمله این افراد، می‌توان به اسوردراپ در سال ۱۹۶۵ اشاره کرد. وی نه تنها این روش را به‌کار گرفت بلکه برای اولین بار ضعف‌های این روش را نیز بیان کرد و استفاده از روش‌های مبتنی بر

[۱۳] Survival function

[۱۴] Right censored

فرآیندهای تصادفی را پیشنهاد داد. وی همچنین گفت که افراد در طول زندگی خود تا زمان مرگ، از سطوح مختلفی از سلامتی عبور می‌کنند که زمان توقف هر فرد در هر سطح از سلامتی تصادفی بوده و لذا طول عمر افراد را دنباله‌ای از متغیرهای تصادفی دانست. به دنبال اسوردراپ، هوئم (۱۹۶۹) در مقاله خود ایده‌های به‌کارگیری فرآیندهای تصادفی را در تحلیل بقا مطرح نمود.

نلسون در سال ۱۹۶۹، با ارائه مباحث نظری در مورد تابع مخاطره نه تنها به توسعه مدل پیشنهادی کاپلان و مایر پرداخت، بلکه باعث شد که برای مدتی بحث استفاده از فرآیندهای تصادفی در مطالعات بقا در حاشیه قرار گیرد. وی نقش سانسور بودن داده‌ها را در مدل نشان داد و اثر چنین داده‌هایی را در مدل لحاظ نمود. همچنین با مدل‌بندی ارتباط بین توابع بقا و مخاطره، امکان محاسبه احتمال‌های بقا را در زمان‌های مختلف امکان‌پذیر ساخت. آلتشولر در سال ۱۹۷۰ بحث مخاطرات رقابتی[۱۵] را برای اولین بار مطرح ساخت و در این راستا از روش نلسون استفاده کرد. آلن در سال ۱۹۷۲، برآوردگری را برای تابع مخاطره ارائه نمود که بعدها به برآوردگر نلسون - آلن[۱۶] شهرت یافت. بعدها نشان داده شد که این برآوردگر برای زمان‌های بزرگتر از بزرگترین زمان مشاهده شده چندان مناسب نیست، چرا که اگر در بزرگترین زمان مشاهده‌شده، تمام موارد با مرگ روبرو شوند، مقدار برآورد بقا صفر خواهد شد و اگر سانسورشده در نظر گرفته شوند تابع بقا قابل محاسبه نخواهد بود.

در این سال‌ها اقدامات گسترده‌ای در زمینه استفاده از روش‌های ناپارامتری برآورد توابع بقا صورت پذیرفت. از جمله چگونگی استفاده از آزمون ویلکاکسون تعمیم‌یافته برای مقایسه داده‌های بقا است که توسط گهان (۱۹۶۵) ارائه شد. همچنین افرون (۱۹۶۷) ضرورت استفاده از مدل‌های ناپارامتری در تحلیل بقا را تشریح کرد و روش‌هایی برای رفع مشکلات موجود در مطالعات بقا، به‌ویژه برای رفع مشکلات موجود در استفاده از برآوردگر نلسون - آلن، پیشنهاد کرد. برسلو (۱۹۷۰) روشی را برای مقایسه طول عمر K نمونه مختلف که الگوی سانسور داده‌ها در آن‌ها متفاوت است، بر اساس تعمیم آزمون کروسکال- والیس، ارائه کرد. از جمله سایر تحقیقات می‌توان به گروئن بوم و ولنر (۲۰۰۱)، بانجر و ولنر (۲۰۰۵)، ریچاردسن و هاگهس (۲۰۰۰)، هانگ (۱۹۹۶) و لین و همکاران (۱۹۹۸) اشاره کرد.

شاخص‌های تحلیل بقا

تعریف ۱.۲.۱. داده‌های بقا:
داده‌های $t_1, t_2, \ldots, t_n$ را داده‌های بقا گویند، اگر از متغیر تصادفی بقا T مشاهده شده باشند.

معمولاً این مقادیر نامنفی، زمان شکست یک مولفه فیزیکی (مکانیکی یا الکتریکی) یا زمان مرگ یک واحد زنده (سلول، بیمار، حیوان، انسان) هستند. ممکن است این متغیر، زمان یادگیری یک مهارت باشد یا حتی امکان دارد هیچ ارتباطی با زمان نداشته باشد. برای مثال، متغیر می‌تواند مبلغ پرداختی یک شرکت بیمه در وضعیت خاصی باشد. در برخی موارد، بیمار بهبود یافته و مبلغ کل پرداختی او معلوم است. در موارد دیگر بیماری هنوز ادامه دارد و تنها مبلغ پرداختی تا آن زمان معلوم است.

[۱۵] Competing risks

[۱۶] Nelson-Aalen estimator

تعریف ۲.۲.۱. تابع بقا:
کمیت اصلی که برای بررسی زمان تا پیشامد به‌کار می‌رود، تابع بقا است. احتمال این‌که یک فرد بیش از زمان t عمر کند را تابع بقا می‌نامند و به صورت زیر تعریف می‌شود:

$$S(t) = P(T > t).$$

اگر T یک متغیر تصادفی پیوسته باشد، آن‌گاه $S(t)$ نیز یک تابع پیوسته اکیداً نزولی است. در واقع تابع بقا، متمم تابع توزیع تجمعی است. یعنی

$$S(t) = ۱ - F(t)$$

که در آن

$$F(t) = P(T \leq t).$$

از آن‌جا که تابع بقا از انتگرال‌گیری تابع چگالی احتمال $f(t)$ نیز به‌دست می‌آید، یعنی

$$S(t) = P(T > t) = \int_t^{\infty} f(u)\mathrm{d}u$$

بنابراین

$$f(t) = -\frac{d}{dt}S(t).$$

دقت کنید که $f(t)dt$ را می‌توان به‌عنوان احتمال تقریبی رخداد پیشامد در زمان t در نظر گرفت. یعنی

$$P(T = t) \approx f(t)\mathrm{d}t.$$

منحنی تابع بقا، $S(t)$، می‌تواند شکل‌های مختلفی داشته باشد. نکته مهم این است که همه این شکل‌ها، ویژگی‌های پایه‌ای یکسانی دارند. در واقع همه توابع بقا یکنوا و ناصعودی هستند، در نقطه صفر (زمان شروع) برابر با یک و با پیشرفت زمان به بی‌نهایت، به صفر همگرا می‌شوند.

برای متغیرهای تصادفی گسسته، موضوع به کلی فرق کرده و از فنون متفاوتی برای رفتار با متغیرهای تصادفی گسسته استفاده می‌شود. متغیرهای تصادفی گسسته معمولاً در مواردی مانند وقتی که داده‌های حاصل از یک نمونه‌گیری را گرد می‌کنیم یا وقتی که زمان‌های از کار افتادن را گرد می‌کنیم یا هنگامی که زمان‌های بقا از اعداد صحیح مربوط به اجزا به‌دست می‌آیند، نمایان می‌شوند. اگر T یک متغیر تصادفی گسسته با تابع جرم احتمال

$$p(t_j) = P(T = t_j) \quad j = ۱, ۲, \ldots$$

باشد، تابع بقا برای این متغیر تصادفی گسسته به‌صورت زیر تعریف می‌شود:

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j).$$

تعریف ۳.۲.۱. تابع مخاطره:
یک کمیت اصلی و زیربنایی در تحلیل بقا، تابع مخاطره[۱۷] می‌باشد. این تابع در زمان t به صورت زیر تعریف می‌شود:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

در مباحث قابلیت اعتماد، این تابع به نرخ شکست شرطی نیز معروف است. همچنین در جمعیت‌شناسی به شدت میرایی، در فرآیندهای تصادفی به تابع شدت، در علم همه‌گیرشناسی[۱۸] به نرخ شکست، و در اقتصاد به معکوس نسبت میل، معروف می‌باشد. اگر T یک متغیر تصادفی پیوسته باشد، در این صورت

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln[S(t)].$$

یک کمیت وابسته به مخاطره (نرخ خطر آنی)، تابع مخاطره تجمعی است که به‌صورت زیر تعریف می‌شود:

$$H(t) = \int_0^t h(u) \mathrm{d}u = -\ln[S(t)].$$

بنابراین برای طول عمرهای پیوسته می‌توان نوشت

$$S(t) = \exp[-H(t)] = \exp[-\int_0^t h(u) \mathrm{d}u].$$

از تعریف تابع نرخ خطر، می‌توان مشاهده کرد که $h(t)\Delta t$ احتمال تقریبی این است که یک فرد با سن t پیشامد مورد بررسی را در لحظه‌ای بلافاصله بعد از t تجربه کند. این تابع به‌ویژه در مشخص کردن توزیع شکست مناسب با استفاده از اطلاعات کیفی در مورد نوع شکست یا مرگ مورد استفاده قرار می‌گیرد.

بر خلاف تابع بقا، ممکن است برای تابع مخاطره شکل‌های مختلفی حاصل شود. تابع مخاطره می‌تواند صعودی، نزولی یا حتی ثابت و غیریکنوا باشد. همچنین می‌تواند مشخصات دیگری را که نحوه مرگ یا شکست را بیان می‌کند، به خود بگیرد. مدل‌های با نرخ خطر صعودی، ممکن است بیان‌کننده این باشد که با یک سیر بقای طبیعی سروکار داریم. مانند زمانی که پیشامد مورد نظر، بررسی مرگ در اثر پیری در بین جوامع انسانی باشد.

توابع مخاطره نزولی، در عمل کمتر پیش می‌آیند و ممکن است زمانی رخ دهند که احتمال شکست یا مرگ بسیار زیاد باشد. به‌عنوان نمونه، در بعضی لوازم الکتریکی بسیار حساس، وسیله با کوچک‌ترین نوسان در برق می‌سوزد یا در بعضی از عمل‌های جراحی پیوند اعضا در بیماران، احتمال عفونت بعد از عمل جراحی زیاد است. برای یک تابع مخاطره ثابت می‌توان متغیر تصادفی طول عمر انسان تا مرگ را مثال زد. به‌عنوان مورد آخر، تابع مخاطره‌ای که در ابتدا صعود کند و یک‌باره شروع به نزول کند، یعنی جهت تقعر آن همیشه رو به پایین باشد، بیشتر در مواردی مانند مدل‌بندی داده‌های بقا مربوط به عمل‌های جراحی که بعد از مدتی از عمل، فرد بیمار دچار ریسک زیادی می‌شود که می‌توان در اثر عفونت یا یک عمل جراحی دیگر باشد، به‌وجود می‌آید.

---

[۱۷] Hazard function

[۱۸] Epidemiology

برای متغیرهای تصادفی گسسته تابع مخاطره به‌صورت

$$h(t_j) = ۱ - \frac{S(t_j)}{S(t_{j-۱})} \quad j = ۱, ۲, \ldots$$

تعریف می‌شود. در این صورت، تابع بقا را می‌توان به صورت ضرب احتمال‌های بقای شرطی نیز نوشت. یعنی

$$S(t) = \Pi_{t_j \leq t} \frac{S(t_j)}{S(t_{j-۱})}.$$

۱.۲.۱ ویژگی‌های پایه‌ای داده‌های بقا

داده‌های زمان تا رخداد یک پیشامد، نسبت به این‌که به چه شکلی نتیجه می‌شوند، مسایلی را برای تحلیل ایجاد می‌کنند. یکی از ویژگی‌ها و مشخصه‌های خاص که اغلب در این داده‌ها رخ می‌دهد سانسور است. به‌طور کلی سانسور زمانی اتفاق می‌افتد که می‌دانیم برخی از طول عمرها، فقط در بازه زمانی مشخص اتفاق افتاده‌اند و سایر طول عمرها، به صورت کامل مشاهده شده‌اند. سانسور دارای انواع مختلفی می‌باشد که از آن جمله می‌توان به سانسور چپ، سانسور راست، و سانسور فاصله‌ای اشاره کرد. برای بررسی جامع و کامل داده‌های سانسورشده، باید به نوع سانسور و طرح‌های مواجه با آن که در متون تحلیل بقا مطرح شده‌اند، توجه کنیم.

یکی دیگر از ویژگی‌هایی که در مطالعات بقا بسیار به چشم می‌خورد، پدیده برش است. نمونه‌های بریده آن‌هایی هستند که به ازای آن‌ها مقادیر مشخصی از جامعه کاملاً حذف شده‌اند. شاید صحیح‌تر این است که بیان کنیم برش در جامعه اتفاق می‌افتد نه در نمونه. در حقیقت نمونه بریده، نمونه‌ای از جامعه بریده است. نمونه‌های بریده در آزمایش‌هایی ایجاد می‌شوند که انتخاب نمونه فقط در قسمتی از محدوده متغیر امکان‌پذیر است. مثال‌هایی از این نمونه‌ها مکرراً در صنایع تولیدی اتفاق می‌افتند که از تولیدات غربال‌شده انتخاب می‌شوند.

نمونه‌های بریده بر مبنای این‌که نقطه برش معلوم است یا نه، دسته‌بندی می‌شوند. اگر این نقاط معلوم باشند، فقط باید از نمونه به‌دست آمده، پارامترهای جامعه را برآورد کرد ولی اگر این نقاط نامعلوم باشند، این نقاط نیز پارامترهایی اضافی هستند که باید از روی نمونه برآورد شوند. در این بخش ابتدا انواع سانسور و برش را بیان می‌کنیم و سپس به معرفی تابع درستنمایی طرح‌های سانسور و برش می‌پردازیم.

برخی از انواع سانسور

- داده کامل: داده مربوط به واحد آزمایشی است که وارد مطالعه می‌شود و قبل از یا در پایان مطالعه، پیشامد مطلوب برای آن رخ می‌دهد.

- سانسور راست: گوییم یک مشاهده از راست در C_r سانسور می‌شود، اگر مقدار واقعی آن معلوم نبوده و بدانیم بزرگتر یا مساوی C_r است. یعنی $T \geq C_r$. به‌عنوان مثال در یک مطالعه مربوط به پیوند قلب، بیمار برای ۵۲ هفته بعد از عمل مورد مطالعه قرار می‌گیرد و پیشامد مورد نظر

مرگ بیمار است. حال اگر بنا به دلیلی بیمار از مطالعه خارج شود و در دسترس نباشد، تاریخ آخرین معاینه می‌تواند به‌عنوان زمان سانسور در نظر گرفته شود.

- سانسور چپ: اگر تنها بدانیم که زمان رخداد پیشامد کمتر یا مساوی C_l است، یعنی $T \leq C_l$، ولی زمان رخداد را نشناسیم، سانسور چپ خواهیم داشت. به‌عنوان مثال، در یک مرکز آموزشی کودکان، زمان رخداد پیشامد سنی در نظر گرفته می‌شود که کودک یاد می‌گیرد وظایف مشخصی را به تنهایی انجام دهد. حال اگر کودکی هنگامی که مطالعه در مرکز شروع می‌شود بتواند وظایف مورد نظر را انجام دهد، سانسور چپ رخ داده است.

- سانسور بازه‌ای: اگر تنها بدانیم که مشاهده در بازه‌ای چون (C_l, C_r) قرار دارد، یعنی $C_l < T < C_r$، آن‌گاه نتیجه سانسور فاصله‌ای است. به‌عنوان مثال، در بیماری‌های عفونی که علائم بیماری بلافاصله بعد از ابتلا مشاهده نشده و تنها بین دو ملاقات و انجام آزمایش ابتلا به بیماری تشخیص داده می‌شود، دو زمان ملاقات به‌عنوان زمان‌های سانسور بازه‌ای در نظر گرفته می‌شوند.

در سانسور راست اگر C_i زمان سانسور مربوط به T_i باشد، مشاهدات (Y_i, δ_i) به صورت زیر تعریف می‌شوند:

$$Y_i = min(T_i, C_i)$$

و

$$\delta_i = I(T_i \leq C_i) = \begin{cases} ۱ & T_i \leq C_i \\ ۰ & T_i > C_i \end{cases}$$

بنابراین اگر $\delta_i = ۱$، آن‌گاه Y_i زمان رخداد پیشامد مورد نظر برای فرد i ام و غیرسانسور است. همچنین در سانسور چپ اگر C_i زمان سانسور مربوط به T_i باشد، آن‌گاه (Y_i, ξ_i) مشاهدات مطالعه محسوب می‌شوند، به‌طوری که

$$Y_i = max(T_i, C_i)$$

و

$$\xi_i = I(T_i \geq C_i) = \begin{cases} ۱ & T_i \geq C_i \\ ۰ & T_i < C_i \end{cases}$$

سانسور چپ و سانسور راست هر دو حالت‌های خاصی از سانسور بازه‌ای هستند.

برخی از انواع برش

- نمونه‌های چپ بریده: در این نمونه‌ها فرض می‌کنیم n مشاهده وجود دارند. یک نمونه چپ بریده است، اگر برای هر مشاهده دلخواه t در این نمونه، داشته باشیم $t \geq T_l$ که در آن T_l یک نقطه برش معلوم است؛ یعنی $t \in [T_l, \infty)$. حال اگر چندین نقطه برش $T_۱, \ldots, T_k$ وجود داشته باشند به‌طوری که $T_۱ < \cdots < T_k$، آن‌گاه یک نمونه چپ بریده فزاینده خواهیم داشت.

- نمونه‌های راست بریده: در این نمونه‌ها فرض می‌کنیم n مشاهده وجود دارند. یک نمونه راست بریده است، اگر برای هر مشاهده دلخواه t داشته باشیم $t \leq T_r$ که در آن T_r یک نقطه برش معلوم است؛ یعنی $t \in (۰, T_r]$. حال اگر چندین نقطه برش $T_۱, \ldots, T_m$ وجود داشته باشند به‌طوری که $T_۱ < \cdots < T_m$، آن‌گاه یک نمونه راست بریده فزاینده خواهیم داشت.

تابع درستنمایی بقا

برای داده‌های سانسورشده، نحوه تشکیل و محاسبه تابع درستنمایی بقا با حالت کلاسیک برای داده‌های مستقل بدون سانسور، کاملا متفاوت است. برای داده‌های بقا، سهم تابع درستنمایی داده‌های بدون سانسور، توسط تابع چگالی بقا محاسبه می‌شود و برای داده‌های سانسورشده سهم تابع درستنمایی توسط ارزیابی تابع بقا در یک نقطه از زمان در نظر گرفته می‌شود. برای تشریح نحوه ساخت تابع درستنمایی، فرض کنید n مشاهده داشته باشیم و تابع درستنمایی به شکل $L = \prod_{i=۱}^{n} L_i$ باشد. در این صورت تابع درستنمایی

- برای مشاهدات غیر سانسور به صورت

$$L_i = S(t_i)h(t_i) = h(t_i)exp\{-\int_{۰}^{t_i} h(x)\mathrm{d}x\}$$

محاسبه می‌شود.

- برای مشاهدات با سانسور راست به صورت

$$L_i = S(C_i) = exp\{-\int_{۰}^{C_i} h(x)\mathrm{d}x\}$$

محاسبه می‌شود.

- برای مشاهدات با سانسور چپ به صورت

$$L_i = ۱ - S(C_i) = ۱ - exp\{-\int_{۰}^{C_i} h(x)\mathrm{d}x\}$$

محاسبه می‌شود.

- برای مشاهدات با سانسور بازه‌ای نیز به صورت

$$L_i = S(C_l) - S(C_r) = exp\{-\int_{۰}^{C_l} h(x)\mathrm{d}x\}\{۱ - exp(-\int_{C_l}^{C_r} h(x)\mathrm{d}x)\}$$

محاسبه می‌شود.

نکته اصلی در تحلیل داده‌های بریده‌شده این است که تحلیل‌گر باید از توزیع شرطی، در به‌دست آوردن تابع درستنمایی مربوط به این داده‌ها استفاده کند. یا به‌طور کلی‌تر از روش‌های شرطی کردن برای برآوردیابی بهره گیرد. تابع درستنمایی برای داده‌های بریده در فاصله برش (T_l, T_r)

- برای مشاهداتی که در فاصله برش قرار گرفته و غیر سانسور باشند، به صورت

$$L_i = f(t_i)/[S(T_l) - S(T_r)]$$

است.

- برای مشاهداتی که در فاصله برش قرار گرفته و سانسور باشند، به صورت

$$L_i = S(C_i)/[S(T_l) - S(T_r)]$$

می‌باشد.

- برای برش چپ و غیر سانسور، به صورت

$$L_i = f(t_i)/S(T_l)$$

است.

- برای برش چپ و سانسور، به صورت

$$L_i = S(C_i)/S(T_l)$$

است.

- و برای برش راست، به شکل

$$L_i = f(t_i)/[۱ - S(T_r)]$$

محاسبه می‌شود.

مثال ۴.۲.۱. فرض کنید n مشاهده از توزیع نمایی با میانگین $\lambda^{-۱}$ داشته باشیم. در توزیع نمایی، تابع مخاطره ثابت و برابر λ است. تابع چگالی و تابع بقا برای $t > ۰$ به ترتیب به صورت

$$f(t) = \lambda exp(-\lambda t)$$

و

$$S(t) = exp(-\lambda t)$$

می‌باشند. بنابراین تابع درستنمایی برای وضعیت‌های مختلف داده‌ها به صورت زیر محاسبه می‌شود:

- داده کامل:

$$L_i = f(t_i) = \lambda exp(-\lambda t_i)$$

- سانسور راست:

$$L_i = S(C_r) = exp(-\lambda C_r)$$

- سانسور چپ:

$$L_i = ۱ - S(C_l) = ۱ - exp(-\lambda C_l)$$

- سانسور فاصله‌ای:

$$L_i = S(C_l) - S(C_r) = exp(-\lambda C_l) - exp(-\lambda C_r)$$

- برش چپ و بدون سانسور:

$$L_i = f(t_i)/S(T_l) = \lambda exp\{-\lambda(t_i - T_l)\}$$

- برش چپ و سانسور:

$$L_i = S(t_i)/S(T_l) = exp\{-\lambda(t_i - T_l)\}$$

- برش راست:

$$L_i = f(t_i)/(۱ - S(T_r)) = \lambda exp(-\lambda t_i)/[۱ - exp(-\lambda T_r)]$$

- برش فاصله‌ای:

$$L_i = f(t_i)/[S(T_l) - S(T_r)] = \lambda exp(-\lambda t_i)/[exp(-\lambda T_l) - exp(-\lambda T_r)]$$

۳.۱ مدل‌های بقا

متخصصان آمار از مدل‌ها برای تقریب رابطه بین یک مجموعه از متغیرهای توضیحی و متغیر پاسخ استفاده می‌کنند و اهداف آن‌ها از ساخت مدل، توصیف و پیش‌گویی متغیر پاسخ با استفاده از متغیرهای توضیحی است. متغیر پاسخ در مدل‌های بقا که به مدل‌های زمان شکست هم معروف‌اند، زمان رخداد یک پیشامد است که می‌تواند سانسور شود و این یک ویژگی مهم و در عین حال چالش‌برانگیز داده‌های بقا است.

انتخاب یک مدل برای دسته‌ای از داده‌ها، انتخاب یک توزیع مناسب است که مجموعه‌ای از داده‌های مفروض را مدل‌بندی می‌کند. برای این منظور بایستی اطلاع خوبی از ویژگی‌های توزیع‌های متفاوت داشته باشیم. در بیشتر اوقات، داده‌ها راهنمای خوبی برای انتخاب مدل می‌باشند. هنگامی که داده‌های شکست (طول عمر) را در اختیار داریم می‌توانیم از توزیع‌هایی با تکیه گاه مثبت مانند توزیع‌های وایبل، نمایی، گاما، و لگ‌نرمال برای مدل‌بندی آن‌ها استفاده نماییم. قابل توجه است که اگر داده‌ها مستقل نباشند، برای مدل‌بندی آن‌ها باید از مدل‌هایی شامل توابع توزیع چندمتغیره استفاده کنیم.

انواع توزیع‌های وایبل کاربردهای فراوانی در مدل‌بندی داده‌های بقا دارند. مدل‌های وایبل برای توصیف زمان‌های شکست مشاهده‌شده اجزاء و پدیده‌ها به‌کار می‌روند. این توزیع را می‌توان به‌عنوان

یکی از شناخته‌شده‌ترین توزیع‌های طول عمر نام برد؛ زیرا تابع مخاطره توزیع وایبل در طول زمان ثابت نیست، دارای یک پارامتر اضافه به نام پارامتر شکل است که موجب می‌شود انعطاف‌پذیری آن زیادتر شود، و پیچیدگی‌های سایر مدل‌های پارامتری در برآورد پارامترها را نیز ندارد. تاکنون مقالات و تحقیقات بسیاری در مورد این توزیع‌ها، خواص و کاربردهای آن‌ها ارائه شده‌اند. مورتی و همکاران (۲۰۰۴) انواع مدل‌های وایبل، تبدیلات و کاربردهای آن‌ها را مورد بررسی قرار دادند. هالینان (۱۹۹۳) تاریخچه این توزیع و خواص آن را ارائه داد و همچنین جانسون و همکاران (۱۹۹۴) در کتاب خود یک فصل کامل را به مطالعه این توزیع و خواص آن اختصاص داده‌اند.

گاهی در بررسی تحلیل بقای بیماران ممکن است بیماران با عوامل خطر یکسان، مدت زمان بقای متفاوتی داشته باشند. دلیل این امر می‌تواند تاثیر یک یا چند متغیر توضیحی باشد که در مدل گنجانده نشده‌اند. در این حالت، مدل‌های شکنندگی[۱۹] نیمه‌پارامتری یا پارامتری می‌توانند مورد استفاده قرار گیرند. اگر مولفه شکنندگی در این مدل‌ها معنی‌دار شود، به این مفهوم است که افراد با متغیرهای توضیحی یکسان و ویژگی‌های مشابه ممکن است دارای مدت زمان بقای یکسانی نباشند و ممکن است مدل برای برازش بهتر نیازمند یک یا چند متغیر توضیحی جدید و ناشناخته باشد. به منظور بررسی، مولفه‌ای تصادفی در مدل گنجانده می‌شود که با بررسی معنی‌داری این مولفه می‌توان در مورد میزان تاثیر متفاوت عوامل خطر بر بیماران اظهار نظر نمود. ایده استفاده از اثر تصادفی به گرین‌وود و یول (۱۹۲۰) باز می‌گردد. برد (۱۹۵۹) نخستین بار از اثر تصادفی برای مدل‌بندی داده‌های بقا در حالت یک متغیره استفاده کرد. هدف معرفی اثر تصادفی، بهبود مدل‌بندی مرگ و میر در جامعه بود. برد (۱۹۵۹) از یک توزیع دوپارامتری گاما برای مدل‌بندی طول عمر استفاده کرد. واژه شکنندگی نخستین بار توسط وائوپل و همکاران (۱۹۷۹) برای به‌حساب آوردن ناهمگنی ایجادشده به دلیل نادیده گرفتن اثر متغیرهای توضیحی مهم در مدل برای داده‌های یک متغیره طول عمر (بدون متغیرهای توضیحی) معرفی شد. اما مدل شکنندگی بدون استفاده از واژه شکنندگی توسط کلیتون (۱۹۷۸) برای تبیین همبستگی در داده‌های دومتغیره بقا معرفی شد. آلن (۱۹۹۲) ناهمگنی موجود در داده‌های بقا را با استفاده از توزیع پواسن مرکب مدل‌بندی کرد. اوآکس (۱۹۹۲) نیز در مورد مدل شکنندگی برای پیشامدهای بازگشتی مطالعه کرد.

یکی از روش‌های تحلیل بقا، استفاده از مدل نیمه‌پارامتری مخاطره متناسب کاکس است که می‌توان گفت رایج‌ترین و محبوب‌ترین روش تحلیل بقا محسوب می‌شود؛ دلیل آن هم ملزم نبودن تعیین توزیع احتمال زمان‌های بقا است.

۱.۳.۱ مدل مخاطره متناسب کاکس

تابع مخاطره در حالت کلی هم به زمان و هم به مجموعه‌ای از متغیرهای توضیحی وابسته است که این متغیرها می‌توانند از نوع شاخص‌های جمعیتی (سن، جنس و نژاد)، سابقه بیماری و موارد مشابه باشند. از این‌رو، کاکس (۱۹۷۲) مدل مخاطره متناسب را معرفی کرد که در آن تابع مخاطره در زمان t برای

[۱۹] Frailty models

آزمودنی با بردار متغیرهای توضیحی x به صورت

$$h(t|x) = h_\circ(t)\exp\{G(x,\beta)\}$$

تعریف می‌شود. در این مدل $h_\circ(t)$ تابع مخاطره پایه[۲۰] نامیده می‌شود که سطح عمومی خطر را برای کلیه آزمودنی‌ها مشخص می‌کند و بستگی تابع خطر را با زمان t نشان می‌دهد. همچنین β بردار پارامترهای رگرسیونی است. عبارت دوم هم به صورت نمایی در نظر گرفته شده تا همواره مثبت باشد. در مدل کلاسیک کاکس فرض می‌شود تابع $G(x,\beta)$ خطی است. یعنی

$$h(t|x) = h_\circ(t)\exp(x'\beta).$$

به $\eta = x'\beta$ پیش‌گوی خطی[۲۱] و به $\exp(x'\beta)$ تابع مخاطره نسبی گفته می‌شود.

در مدل رگرسیون کاکس برقراری پذیره‌های استقلال زمان‌های رخداد پیشامدها و همچنین متناسب بودن مخاطرات، ضروری است. منظور از مخاطره‌های متناسب آن است که میزان خطر برای یک فرد نسبت ثابت و مشخصی را با میزان خطر فردی دیگر داشته باشد. به عبارت دیگر

$$\frac{h_{x_i}(t)}{h_{x_j}(t)} = \frac{h_\circ(t)\exp(x_i'\beta)}{h_\circ(t)\exp(x_j'\beta)} = \frac{\exp(x_i'\beta)}{\exp(x_j'\beta)}.$$

حال اگر داده‌ها دارای پیچیدگی باشند، ممکن است همه یا بخشی از این پذیره‌ها برقرار نباشند و استفاده از این مدل را با مشکلات و محدودیت‌هایی همراه سازند. یکی از ویژگیهای مدل کاکس این است که بدون در نظر گرفتن هیچ‌گونه پذیره توزیعی در مورد تابع مخاطره پایه، می‌توان آن را به داده‌های بقا برازش داد. البته در مدل پارامتری کاکس می‌توان برای تابع مخاطره پایه پذیره توزیعی نیز در نظر گرفت. بنابراین در استفاده از مدل کاکس لازم است تابع مخاطره پایه $h_\circ(t)$ با روش ناپارامتری یا پارامتری مدل‌بندی شود. بر همین اساس مدل مخاطره‌های متناسب کاکس به دو صورت پارامتری و نیمه‌پارامتری استفاده می‌شود. قبل از معرفی مدل‌های پارامتری و نیمه‌پارامتری، لازم است ابتدا مدل‌های خطی تعمیم‌یافته[۲۲] را معرفی کنیم.

مدل‌های خطی تعمیم‌یافته

هر مدل خطی تعمیم‌یافته دارای شکل کلی به صورت زیر است:

$$g(\mu) = g[E(Y)] = x'\beta$$

که در آن x بردار متغیرهای توضیحی، β بردار پارامترهای رگرسیونی و $g(\cdot)$ تابع پیوند[۲۳] است. مدل‌های خطی تعمیم‌یافته این امکان را فراهم می‌آورند که بتوان مدل رگرسیون معمولی را به مواردی که متغیر

[۲۰] Baseline hazard function

[۲۱] Linear predictor

[۲۲] Generalized linear models

[۲۳] Link function

پاسخ غیرگاوسی است، بسط و تعمیم داد. در مدل‌های خطی کلاسیک فرض بر این است که جمله خطا، متغیرهای تصادفی مستقل با توزیع نرمال هستند. در حالی‌که در مدل‌های خطی تعمیم‌یافته با استفاده از یک تابع پیوند بین میانگین پاسخ‌ها و متغیرهای توضیحی، ارتباط خطی برقرار می‌باشد و جمله خطا می‌تواند توزیعی غیرنرمال ولی از خانواده توزیع‌های نمایی داشته باشد. واضح است که مدل‌های خطی تعمیم‌یافته رده واحدی از مدل‌های آماری فراهم می‌آورد که مدل‌های خطی کلاسیک را در بر دارد.

هر الگوی خطی تعمیم‌یافته دارای سه جزء اصلی می‌باشد:

۱. مولفه تصادفی: توزیع شرطی متغیر پاسخ را مشخص می‌کند. توزیع پاسخ عضوی از خانواده توزیع‌های نمایی مانند نرمال، دوجمله‌ای و پواسون است. انتخاب مولفه تصادفی بنا به ساختار متغیر پاسخ، متفاوت است.

۲. پیش‌گوی خطی: نقش متغیرهای توضیحی را در قالب یک ترکیب خطی از آن‌ها، مشخص می‌کند. در واقع پیش‌گوی خطی به صورت

$$\eta = x^{'}\beta$$

تعریف می‌شود.

۳. تابع پیوند: تابع پیوند، چگونگی وابستگی $E(Y) = \mu$ را به $x^{'}\beta$ شرح می‌دهد و معمولاً به صورت $g(\mu) = \eta$ یا به‌طور معادل $g^{-1}(\eta) = \mu$ نوشته می‌شود و انتخاب این تابع پیوند تا حدی اختیاری است. تابع پیوند انواع مختلفی دارد که در جدول زیر برخی از توابع پیوند به همراه معکوس آن‌ها فهرست شده‌اند.

جدول ۱.۱: برخی از توابع پیوند به همراه معکوس آن‌ها

توزیع خطا	پیوند	تابع پیوند	تابع میانگین
نرمال	همانی	$X\beta = \mu$	$\mu = X\beta$
نمایی و گاما	وارون	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
نرمال معکوس	مربع وارون	$X\beta = \mu^{-2}$	$\mu = (X\beta)^{-\frac{1}{2}}$
پواسون	لگاریتمی	$X\beta = ln(\mu)$	$\mu = exp(X\beta)$
دوجمله‌ای و چندجمله‌ای	لجیت	$X\beta = ln(\frac{\mu}{1-\mu})$	$\mu = \frac{exp(X\beta)}{1+exp(X\beta)}$

برای جزییات بیشتر در مورد این مدل‌ها می‌توانید به مک‌کلاک و نلدر (۱۹۸۹)، فارمیر و توتز (۲۰۰۱)، و مک‌کلاک و سیرل (۲۰۰۱) اشاره کرد.

مدل‌های پارامتری

در مدل مخاطرات متناسب کاکس، تابع مخاطره پایه را می‌توان تابع مخاطره یکی از توزیع‌های معروف بقا مانند توزیع نمایی، گمپرتز، لگ‌نرمال، و گاما در نظر گرفت. در ادامه برخی از مدل‌های پرکاربرد پارامتری را به اختصار معرفی می‌کنیم.

مثال ۱.۳.۱. مدل نمایی:

فرض کنید $t = (t_1, t_2, \ldots, t_n)'$، n زمان مستقل بقا باشند که هر یک دارای توزیع نمایی با پارامتر λ است. همچنین فرض کنید $\delta = (\delta_1, \delta_2, \ldots, \delta_n)'$ تابع نشانگر برای سانسور باشد. یعنی

$$\delta_i = \begin{cases} 0 & \text{سانسور راست} \\ 1 & \text{در غیر این صورت} \end{cases}$$

توابع چگالی و بقا t_i به ترتیب به شکل‌های

$$f(t_i|\lambda) = \lambda exp(-\lambda t_i)$$

و

$$S(t_i|\lambda) = exp(-\lambda t_i)$$

هستند. برای ساختن یک مدل رگرسیونی، متغیرهای توضیحی از طریق λ و به صورت $\lambda_i = \exp(z_i'\beta)$ معرفی می‌شوند که در آن z_i یک بردار $1 \times p$ بعدی از متغیرهای توضیحی و β بردار $p \times 1$ بعدی از پارامترهای رگرسیونی و در نتیجه $D = (t, Z, \delta)$ داده مشاهده‌شده برای مدل رگرسیونی است، به‌طوری که Z یک ماتریس $n \times p$ ساخته‌شده از متغیرهای توضیحی، معروف به ماتریس طرح، است که i امین سطر آن z_i است. بنابراین تابع درستنمایی به صورت زیر تشکیل می‌شود:

$$\begin{aligned} L(\beta|D) &= \prod_{i=1}^{n} f(t_i|\lambda_i)^{\delta_i} S(t_i|\lambda_i)^{(1-\delta_i)} \\ &= \prod_{i=1}^{n} \exp(z_i'\beta)^{\delta_i} \exp(-\exp(z_i'\beta)t_i)^{\delta_i} \exp(-\exp(z_i'\beta)t_i)^{1-\delta_i} \\ &= \exp\{\sum_{i=1}^{n} \delta_i z_i'\beta\} \exp\{-\sum_{i=1}^{n} t_i \exp(z_i'\beta)\}. \end{aligned}$$

مثال ۲.۳.۱. مدل وایبل:

فرض کنید $t = (t_1, t_2, \ldots, t_n)'$، n زمان مستقل بقا باشند که هر یک دارای توزیع وایبل با پارامتر شکل α ($\alpha > 0$) و پارامتر مقیاس نامنفی λ هستند. بردار δ را نیز مشابه مثال قبلی در نظر بگیرید. تابع چگالی t_i به شکل

$$f(t_i|\alpha, \lambda) = \alpha t_i^{\alpha-1} \exp(\lambda - \exp(\lambda) t_i^{\alpha})$$

و تابع بقا به صورت

$$S(t_i|\alpha, \lambda) = \exp(-\exp(\lambda) t_i^{\alpha})$$

تعریف می‌شوند. بنابراین، تابع درستنمایی برای (α, λ) را می‌توان به صورت

$$\begin{aligned} L(\alpha, \lambda|D) &= \prod_{i=1}^{n} f(t_i|\alpha, \lambda)^{\delta_i} S(t_i|\alpha, \lambda)^{(1-\delta_i)} \\ &= \alpha^{\sum \delta_i} \exp\{\lambda \sum_{i=1}^{n} \delta_i + \sum_{i=1}^{n} (\delta_i(\alpha - 1)\log(t_i) - \exp(\lambda) t_i^{\alpha})\} \end{aligned}$$

نوشت. در این مدل نیز متغیرهای توضیحی از طریق λ و به صورت $\lambda_i = z_i'\beta$ وارد می‌شوند. که در آن مولفه‌های موجود مشابه مدل رگرسیون نمایی معرفی می‌شوند. در این حالت، تابع مخاطره متناظر به شکل

$$h(t_i|\alpha, \lambda_i) = \alpha t_i^{\alpha-1} \exp(\lambda_i)$$

تعریف می‌شود.

مدل‌های نیمه‌پارامتری

کاکس در سال ۱۹۷۲ روشی را ابداع کرد که با به‌کارگیری آن لازم نیست شکل خاصی برای تابع مخاطره پایه تعیین کنیم و بدون برآورد $h_0(t)$ با این روش، می‌توان پارامتر β را در مدل مخاطرات متناسب کاکس برآورد کرد. وی این روش را روش درستنمایی جزئی[۲۴] نام نهاد و ادعا کرد برآوردگر درستنمایی جزئی پارامتر β دارای خواص درستنمایی کامل است.

اما درستنمایی جزئی چیست؟ فرض کنید داده‌های مربوط به n آزمودنی در دسترس هستند و d زمان پیشامد مجزا و $n-d$ زمان سانسورشده داریم. زمان‌های بقا مرتب‌شده را به صورت $y_{(1)} < \cdots < y_{(d)}$ نشان می‌دهیم. مجموعه آزمودنی‌هایی که در زمان $y_{(j)}$، $j = 1, \ldots, d$، در معرض خطر هستند را با $R_{(j)}$ نشان داده و آن را مجموعه مخاطره می‌نامیم. این مجموعه شامل آزمودنی‌هایی است که در زمان $y_{(j)}$ زنده‌اند یا پیشامد مورد نظر را در همین زمان تجربه کرده‌اند. با این نمادگذاری، تابع درستنمایی جزئی کاکس به صورت زیر تعریف می‌شود:

$$PL(\beta|D) = \prod_{j=1}^{d} \frac{\exp(x'_{(j)}\beta)}{\sum_{l \in R_{(j)}} \exp(x'_l\beta)}$$

که در آن $x_{(j)}$ بردار متغیرهای توضیحی برای آزمودنی است که پیشامد مورد نظر را در زمان $y_{(j)}$ تجربه می‌کند. عبارتی که در مخرج کسر قرار دارد، مجموع برای همه آزمودنی‌هایی است که در زمان $y_{(j)}$ در معرض خطر هستند. عمل ضرب نیز روی آزمودنی‌هایی اجرا می‌شود که زمان (مرتب‌شده) پیشامدشان مشاهده شده‌اند و آزمودنی‌هایی که زمان بقا آن‌ها سانسورشده هستند، در این درستنمایی شرکت داده نمی‌شوند. این درستنمایی تنها به مرتب‌سازی زمان پیشامدها وابسته است، زیرا مجموعه مخاطره در هر زمان توسط پیشامدهای مشاهده‌شده تعیین می‌شود.

اندرسون و گیل (۱۹۸۲) ثابت کردند چنان‌چه مسئله را از طریق چارچوب فرآیندهای شمارشی حل نمایند، برآوردگر حاصل از روش درستنمایی جزئی کاکس، به صورت مجانبی، دارای ویژگی برآوردگرهای ماکسیمم درستنمایی کامل است. برآوردگر ماکسیمم درستنمایی جزئی معمولا با استفاده از روش‌های عددی محاسبه می‌شود.

اگر چه بدون تابع مخاطره پایه، $h_0(t)$، می‌توانیم برآوردگرهای ماکسیمم درستنمایی (جزیی) ضرایب رگرسیونی را، که دارای ویژگی‌های بهینه هستند، به دست آوریم، اما برای توصیف و برآورد تابع مخاطره (بقا) افراد باید تابع مخاطره پایه را نیز برآورد کنیم. در مدل مخاطرات متناسب کاکس، $h_0(t)$ به صورت

[۲۴] Partial likelihood

ناپارامتری در نظر گرفته می‌شود. به همین دلیل است که به این مدل یک مدل نیمه‌پارامتری گفته می‌شود؛ زیرا ترکیبی از یک مدل ناپارامتری برای تابع مخاطره پایه و یک بخش پارامتری برای متغیرهای توضیحی است. برای مدل‌بندی ناپارامتری تابع مخاطره پایه، معمولا از روش‌هایی همچون اسپلاین‌ها[۲۵] (گرین، ۱۹۹۶؛ گرین و سیلورمن، ۱۹۹۴)، هسته‌ها[۲۶] و روش تکه‌ای ثابت[۲۷] استفاده می‌کنند.

توابع ناپارامتری، توابع نامتناهی بعد یا با بعد نامتناهی محسوب می‌شوند؛ مانند تابع چگالی احتمال، تابع مخاطره یا یک معادله رگرسیونی ناپارامتری. از طرفی، استنباط آماری مبتنی بر اصل ماکسیمم درستنمایی برای توابع با بعد نامتناهی، دچار چالش است. زیرا تلاش برای ماکسیمم کردن تابع درستنمایی در چنین مواردی، معمولاً به یک مقدار نامتناهی برای درستنمایی یا عدم شناسایی‌پذیری[۲۸] خود تابع نامعلوم منجر می‌شود. در استنباط توابع پارامتری، این مشکل وجود ندارد زیرا در آن‌ها این پذیره برقرار است که تابع مورد نظر یک صورت پارامتری مشخص دارد و در نتیجه استنباط به مساله شناخت تعداد متناهی از پارامترها منتهی می‌شود. در این موارد، یک راه‌حل، در دیگاه بسامدی[۲۹]، استفاده از روش درستنمایی جریمه‌ای[۳۰] است. درستنمایی جریمه‌ای (یا تاوانیده) یک روش ناپارامتری برآورد منحنی است که ما را از پذیره‌های پارامتری در مورد تابع مورد نظر بی‌نیاز می‌کند.

۲.۳.۱ تحلیل داده‌های بازگشتی

برای بررسی وقوع یک پیشامد، نیازمند اندازه‌گیری عوامل یا متغیرهایی هستیم که منجر به وقوع آن پیشامد می‌شوند. برای مثال اگر پیشامد مورد نظر تعداد تصادفات در جاده‌ها باشد، ممکن است با اندازه‌گیری متغیری مانند تعداد خودروها مدلی ایجاد کنیم که رابطه بین این دو را توجیه نماید. گاهی اوقات اندازه‌گیری متغیر مورد نظر بر حسب نوع تحقیق، چندین بار در طول یک بازه زمانی انجام می‌شود که در این حالت داده ایجادشده، یک داده طولی نامیده می‌شود. برخی از مطالعاتی که در علومی مانند مهندسی یا پزشکی انجام می‌شوند، معمولاً در یک بازه زمانی صورت می‌گیرند و متغیرهای مورد نظر در طول این بازه اندازه‌گیری می‌شوند. بنابراین در این علوم، غالباً با داده‌های طولی سروکار داریم.

در داده‌های طولی بقا، داده‌های زمان چندگانه مطرح می‌شوند و در این مورد یک فرآیند تصادفی با چندین پیشامد مختلف برای هر شخص مشاهده می‌شود. معمولاً در مطالعات بیماری‌های مزمن این نوع داده‌ها وجود دارند. همچنین مطالعاتی که در آن‌ها هر فرد امکان تجربه متعدد یک پیشامد را در زمان‌های مختلف داشته باشد، داده‌هایی از نوع داده‌های طولی بقا دارند. این نوع داده‌ها را پیشامدهای بازگشتی نیز می‌نامند. در مطالعات بقا، پیشامدهای بازگشتی از اهمیت بسزایی برخوردار هستند و در بسیاری از زمینه‌های علمی از جمله پزشکی، قابلیت اطمینان قطعات تولیدی و صنعتی، و جرم شناسی رخ می‌دهند.

---

[۲۵] Splines

[۲۶] Kernel methods

[۲۷] Constant piecewise

[۲۸] Identifiablity

[۲۹] Frequentist approach

[۳۰] Penalized likelihood

برای مثال، حملات آسم در بیماران تنفسی، بازگشت تومور بعد از جراحی در مطالعات سرطان، حملات قلبی پی در پی، شناسایی مکرر یک خطا در یک سیستم نرم افزاری، عفونت‌های مکرر ناشی از دستگاه دیالیز، تعداد حملات صرع، اپیزودهای مکرر بیماری‌های روانی یا بستری شدن‌های متوالی در بیمارستان، و عفونت‌های متوالی بیماران ایدزی، پیشامدهای بازگشتی را به وجود می‌آورند (لی و لاگاکوس، ۱۹۹۷). در این نوع داده‌ها ممکن است وقایع مورد نظر مشابه باشند، به‌عنوان مثال، بستری شدن یک فرد دیالیزی در طی مدت یک سال یا دفعات حملات تشنج یک فرد مبتلا به صرع. اما گاهی نیز ممکن است آزمودنی‌های مورد مطالعه، یک حادثه را با شدت‌های مختلف در هر بار وقوع تجربه کنند. به‌عنوان مثال، بازگشت سرطان پستان به صورت متاستاز سرطان تخمدان یا حملات آسمی یک بیمار که در طی یک مدت خاص ممکن است با شدت بیشتری نسبت به دفعات قبل رخ دهد. به این حالت اخیر که ترتیب وقوع حوادث بیانگر پیشینه و پیشرفت بیماری است، داده‌های چندحالتی[۳۱] گفته می‌شود. برای این نوع داده‌ها، احتمال انتقال از یک حالت به حالت‌های دیگر در بیماران مورد توجه است. بر اساس نوع داده‌های بقا از قبیل ترتیب و نوع وقوع حوادث افراد مورد مطالعه، و همچنین همبستگی بین وقوع حوادث یا افراد، روش‌های آماری مختلفی در تحلیل بقا بنا نهاده شده‌اند.

در این پایان‌نامه تمرکز بر روی داده‌های پیشامدهای بازگشتی نوع همسان است. در بسیاری از مطالعات طولی بالینی هر فرد امکان تجربه چندین حادثه پی در پی از یک نوع را دارد. تحلیل داده‌های حوادث بازگشتی در مطالعات بالینی برای ارزیابی اثر درمان و مطالعه روند پیشرفت بیماری‌ها بسیار حائز اهمیت است. بر حسب نظر محققین بالینی، اندازه‌های متفاوت از پیشامدهای بازگشتی مانند تعداد عودها و زمان بین وقوع حوادث مکرر، شاخص‌های مهمی از پیشرفت بیماری هستند. پیشامدهای بازگشتی به‌طور طبیعی دارای ترتیب زمانی هستند و در یک دنباله خاص در طول زمان رخ می‌دهند. بر اساس تعریف‌های مختلف، مقیاس زمانی، مدل‌ها و روش‌های تحلیل این داده‌ها متفاوت خواهند بود. دو مقیاس زمانی مطرح در این زمینه عبارتند از: زمان تا وقوع حوادث و زمان بین پیشامدهای بازگشتی.

اگر تعریف مقیاس زمانی بر حسب زمان تا وقوع باشد، آن‌گاه تحلیل‌های مبتنی بر تابع شدت[۳۲] یا تابع نرخ[۳۳] (که نام‌های دیگری برای تابع مخاطره هستند)، کاربردی‌ترین روش‌های تحلیل خواهند بود. برای مقیاس زمان بین حوادث بازگشتی، تابع مخاطره فواصل زمانی یا توزیع توام فواصل زمانی مدل‌بندی می‌شوند. در مطالعه حاضر مقیاس فواصل زمانی بین حوادث بازگشتی در نظر گرفته شده است. فواصل زمانی بین عودهای متوالی بیانگر حالت ثبات بهبودی در بیماران است. بنابراین، طولانی‌تر شدن این مدت زمان بیانگر کاهش نرخ بازگشت بیماری است و حالت عکس آن، کوتاهتر شدن این مدت زمان بیانگر افزایش نرخ عود بیماری است.

افزون بر مقیاس زمان، مساله مهم دیگر در داده‌های پیشامدهای بازگشتی وجود همبستگی بین زمان‌های پیشامدهای بازگشتی برای هر فرد است. همبستگی بین زمان‌های شکست در پیشامدهای بازگشتی به دو دلیل عمده رخ می‌دهد:

---

[۳۱] Multistate

[۳۲] Intensity function

[۳۳] Rate function

۱. ناهمگنی[۳۴] بین افراد: در بعضی مطالعات برخی افراد دارای نرخ وقوع بالاتر یا کمتری نسبت به افراد دیگر هستند. این امر به دلیل عوامل ناشناخته، غیرقابل اندازه‌گیری یا اندازه‌گیری‌نشده رخ می‌دهد. به‌عنوان مثال، شیوه متفاوت زندگی افراد، عوامل ژنتیکی و عوامل محیطی از جمله مواردی هستند که استعداد وقوع بیماری را در یک فرد افزایش یا کاهش می‌دهند. این امر به ناهمگنی بین افراد و ایجاد همبستگی درون فردی در وقوع حوادث بازگشتی منجر می‌شود.

۲. همبستگی بین پیشامدها: گاهی وقوع یک حادثه منجر به افزایش یا کاهش احتمال وقوع حوادث بعدی می‌شود. این همبستگی بین حوادث ممکن است به دلیل ضعف بیولوژیکی یا مقاومت بیولوژیکی ناشی از حادثه قبل، در افراد مختلف، متفاوت باشد. هر یک از این پدیده‌ها باعث می‌شوند که هر حادثه به‌عنوان تابعی از حوادث قبل ظاهر شود. این حالت نیز منجر به همبستگی درون فردی در این نوع داده‌ها می‌شود.

بنابراین، برای تحلیل پیشامدهای بازگشتی از روش‌هایی استفاده می‌شود که همبستگی بین داده‌ها را در نظر گرفته و تعدیل کنند. در تحلیل این داده‌ها اگر بدون توجه به همبستگی درون فردی، از مدل‌های معمول مانند رگرسیون خطرات متناسب استفاده شود، برآوردها اریب و واریانس برآوردها کمتر از مقدار واقعی برآورد می‌شوند. بنابراین نحوه لحاظ کردن وابستگی بین حوادث و تعدیل همبستگی در مدل، بخش مهمی در تحلیل پیشامدهای بازگشتی است. در این زمینه مدل‌های مختلفی مطرح شده‌اند که عبارتند از: مدل‌های حاشیه‌ای[۳۵]، مدل‌های واریانس تصحیح‌شده، مدل‌های شرطی و مدل‌های شکنندگی که همه آن‌ها تعمیمی از مدل‌های کاکس هستند. هر یک از مدل‌های نامبرده دارای مزایا و معایبی هستند و انتخاب هر یک از آن‌ها بر اساس هدف مطالعه و تعریف مقیاس زمان اندازه‌گیری‌شده، صورت می‌گیرد.

۴.۱ استنباط آماری بیزی

به‌طور کلی، دو دیدگاه برای انجام استنباط آماری در ادبیات آماری وجود دارند: دیدگاه استنباط بسامدی[۳۶] و دیدگاه استنباط بیزی[۳۷]. دیدگاه بسامدی بر این اصل استوار است که احتمال یک فراوانی حدی است، یعنی احتمال یک پیشامد برابر است با فراوانی نسبی وقوع آن پیشامد. این بدین معنی است که بسامدی‌ها احتمال را بر حسب تکرارپذیری پیشامدهای قابل تکرار تعبیر می‌کنند که در آن عدم قطعیت[۳۸] به خاطر تصادفی بودن پیشامدها است. اگر عدم قطعیت وقوع پیشامد به دلیل عدم اطلاع در مورد آن پیشامد باشد، بسامدی‌ها نباید به آن احتمالی نسبت دهند.

در دیدگاه استنباط بیزی، پارامترهای نامعلوم تحقق‌هایی از یک توزیع احتمالی در نظر گرفته می‌شوند.

---

[۳۴] Heterogeneity

[۳۵] Marginal models

[۳۶] Frequentistic inference

[۳۷] Bayesian inference

[۳۸] Uncertainty

یعنی پارامترها نیز متغیر تصادفی محسوب می‌شوند و عدم قطعیت یا درجه اعتقاد[۳۹] نسبت به پارامترها، با استفاده از توزیع‌های احتمالی بیان می‌شود. برای تشریح بیشتر، اگر توزیع احتمالی پارامتر به یک زیرفضا از فضای پارامتر وزن بیشتری بدهد، این بدان معنی است که از دیدگاه بیزی، اعتقاد بیشتری نسبت به تعلق پارامتر به آن زیرفضا وجود دارد. این اعتقاد را اعتقاد پیشین و توزیع احتمال آن را توزیع پیشین می‌نامند. روش استنباط بیزی، بر اساس قاعده بیز، توزیع پیشین را به کمک اطلاعات موجود در داده‌ها (از طریق تابع درستنمایی) به‌روز کرده و توزیع احتمال جدیدی را که به توزیع پسین معروف است، نتیجه می‌دهد.

فرض کنید $X = (X_۱, \ldots, X_n)$ یک نمونه تصادفی از مدل احتمالی پارامتری $f(x|\theta)$ باشد، که در آن $\theta \in \Theta \subset \Re^d$. به Θ فضای پارامتر می‌گوییم. اگر تابع درستنمایی این نمونه را با $L(\theta|x)$ و توزیع پیشین را با $\pi(\theta)$ نمایش دهیم، آن‌گاه توزیع پسین، $\pi(\theta|x)$، به صورت زیر محاسبه می‌شود:

$$\pi(\theta|x) = cL(\theta|x)\pi(\theta) \propto L(\theta|x)\pi(\theta)$$

که در آن c ثابت نرمال‌ساز نامیده می‌شود و برابر معکوس

$$\int_\Theta L(\theta|x)\pi(\theta)d\theta$$

است. در این دیدگاه همه استنباط‌ها مبتنی بر توزیع پسین به‌دست می‌آیند.

صرف نظر از توانایی‌ها و ایرادهایی که بر هر کدام از دو رویکرد استنباط آماری وارد است، ما در این پایان‌نامه از رویکرد بیزی برای استنباط استفاده می‌کنیم.

فاصله اطمینان بیزی یا معتبر[۴۰]

فرض کنید $X = (X_۱, X_۲, \ldots, X_n)'$ یک نمونه تصادفی از خانواده توزیع‌های $\{f(x|\theta) : \theta \in \Theta\}$ باشد، که در آن $\Theta \in \Re$. در دیدگاه بسامدی استنباط آماری، فاصله تصادفی $(L(X), U(X))$ را که

$$P_\theta(L(X) < \theta < U(X) = ۱ - \alpha$$

یک خانواده از فواصل اطمینان سطح $۱ - \alpha$ برای θ می‌دانند. برای مثال، اگر نمونه تصادفی از توزیع $N(\theta, ۱)$ باشد، آن‌گاه خانواده فواصل اطمینان

$$\theta \in (\bar{X} - z_{۱-\alpha/۲}\frac{۱}{\sqrt{n}}, \bar{X} + z_{۱-\alpha/۲}\frac{۱}{\sqrt{n}})$$

به‌دست می‌آید. معنای این فاصله تصادفی این است که این فاصله پارامتر θ را با احتمال $۱ - \alpha$ پوشش می‌دهد. فاصله تصادفی بالا دارای یافته‌های بی‌شماری است که هر کدام به صورت یک فاصله عددی است. اگر احتمال را به طریق فراوانی نسبی تعبیر کنیم و $۱ - \alpha = ٪۹۰$ باشد، می‌توان گفت که نود درصد از این‌گونه فواصل θ را پوشش می‌دهند.

---

[۳۹] Belief

[۴۰] Bayesian confidence or credible interval

در آمار بیزی θ را به‌عنوان یک متغیر تصادفی با توزیع پیشین $\pi(\theta)$ در نظر می‌گیرند. در این حالت θ می‌تواند در یک فاصله عددی با احتمال غیر از صفر و یک نیز قرار گیرد که این فاصله را فاصله اطمینان بیزی می‌نامند و تعبیر این فاصله عددی به‌عنوان فاصله‌ای که با احتمال $\alpha - ۱$ پارامتر θ را دربر دارد کاملا معتبر است.

تعریف ۱.۴.۱. فرض کنید $X = (X_۱, X_۲, \ldots, X_n)'$ یک نمونه تصادفی از توزیع $\{f(x|\theta) : \theta \in \Theta\}$ باشد که در آن θ دارای توزیع پیشین $\pi(\theta)$ و چگالی پسین $\pi(\theta|X = x)$ است. یک فاصله اطمینان بیزی یا فاصله معتبر در سطح $\alpha - ۱$ برای پارامتر $\gamma(\theta)$ فاصله $(L(x), U(x))$ است، به‌طوری که

$$P(L(x) < \gamma(\theta) < U(x)|X = x) = ۱ - \alpha.$$

قضیه ۲.۴.۱. فرض کنید فاصله $(L(x), U(x))$ یک فاصله اطمینان بیزی برای θ در سطح $\alpha - ۱$ باشد. این فاصله دارای کمترین طول خواهد بود، هرگاه عددی مانند $k(\alpha)$ وجود داشته باشد به‌طوری که

$$L(x) < \theta < U(x) \Longleftrightarrow \pi(\theta|x) > k(\alpha).$$

تعریف ۳.۴.۱. فاصله معتبر چگالی پسین رفیع[۴۱] (HPD) $(۱ - \alpha)$٪۱۰۰ برای θ زیر مجموعه C از فضای پارامتر Θ به فرم

$$C = \{\theta \in \Theta|\pi(\theta|x) > k(\alpha)\}$$

است، که در آن $k(\alpha)$ مقداری است که $P(C|x) = ۱ - \alpha$.

۱.۴.۱ تقریب توزیع پسین

مدل‌های معمول بقا برای تحلیل داده‌های پیشامدهای بازگشتی، از جمله مدل‌های پیچیده آماری محسوب می‌شوند که توزیع پسین آن‌ها شکل معلوم و شناخته‌شده‌ای ندارد. بنابراین محاسبه کمیت‌های مبتنی بر این توزیع پسین برای استنباط در مورد پدیده‌های مورد علاقه، ممکن نیست. راهکاری که همیشه در این موارد ارایه می‌شود، تقریب توزیع پسین و استفاده از نسخه تقریبی آن برای محاسبه کمیت‌های مورد نظر مثل میانگین، واریانس یا چندک‌های توزیع است.

روش‌های تقریب توزیع پسین به دو دسته کلی تقسیم می‌شوند: روش‌های عددی و روش‌های مبتنی بر نمونه‌گیری[۴۲]. از جمله روش‌های تقریب عددی می‌توان به تقریب نیوتون- رافسون، روش‌های تربیع‌بندی گاوسی[۴۳] (آبرامویتز و استگن، ۱۹۷۲)، و تقریب لاپلاس[۴۴] (تیرنی و کدین، ۱۹۸۶) اشاره کرد. از معروف‌ترین روش‌های مبتنی بر نمونه‌گیری که به روش‌های شبیه‌سازی مونت کارلویی نیز معروف هستند، می‌توان روش‌های نمونه‌گیری رد و پذیرش[۴۵] (کسلا و برگر، ۲۰۰۲)، نمونه‌گیری نقاط مهم[۴۶]

[۴۱] Highest posterior density
[۴۲] Sampling-based methods
[۴۳] Gaussian quadratures
[۴۴] Laplace approximation
[۴۵] Reject-accept sampling
[۴۶] Importance sampling

(هستربرگ، ۱۹۸۷) و رده بزرگ الگوریتم‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی[۴۷] (MCMC) (هستینگز، ۱۹۷۰؛ رابرت و کسلا، ۲۰۰۴) را نام برد.

در مواردی که بعد توزیع پسین کمتر از ۱۰ باشد، روش‌های عددی خطای قابل قبولی دارند. اما در صورتی که بعد توزیع پسین بالا باشد، تجمع خطاهای گرد کردن در روش‌های عددی بزرگ شده و استفاده از آن‌ها توصیه نمی‌شود. در این مواقع، روش‌های مبتنی بر نمونه‌گیری استفاده می‌شوند. در روش‌های نمونه‌گیری MCMC، خطای تقریب تابعی از بعد توزیع پسین نیست و به ازای هر بعدی از مرتبه $O(n^{-1/2})$ ثابت می‌ماند، که در آن n حجم نمونه مونت کارلویی است. سرعت محاسبات در روش‌های مبتنی بر نمونه‌گیری می‌تواند خیلی کندتر از روش‌های عددی باشد. در روش‌های مبتنی بر نمونه‌گیری، قانون قوی اعداد بزرگ برای نمونه‌های تولیدشده مستقل (کسلا و برگر، ۲۰۰۲) و قضیه ارگودیک برای نمونه‌های تولیدشده وابسته (زنجیر مارکوفی) (بیرخوف، ۱۹۴۲) ضامن همگرایی برآوردگرهای حاصل از نمونه تولیدشده به کمیت‌های مورد نظر از توزیع پسین هستند.

از آن‌جا که تقریب لاپلاس، در این پایان‌نامه، یک بخش مهم در ساخت روش تقریبی مورد نظر ما محسوب می‌شود، در زیربخش بعدی، به‌عنوان یکی از مجموعه روش‌های عددی، به معرفی آن می‌پردازیم.

تقریب لاپلاس

تقریب لاپلاس یک روش عددی برای تقریب انتگرال‌های پیچیده است که در استنباط بیزی از آن زیاد استفاده می‌شود. فرض کنید علاقه‌مند به محاسبه انتگرال

$$\int f(x)\mathrm{d}x = \int \exp(\log f(x))\mathrm{d}x$$

باشیم. بسط سری تیلور تابع $\log f(x)$ تا مرتبه دوم در همسایگی $x = x_۰$ به‌صورت زیر است:

$$\log f(x) \approx \log f(x_۰) + (x - x_۰)\frac{\partial \log f(x)}{\partial x}|_{x=x_۰} + \frac{(x-x_۰)^۲}{۲}\frac{\partial^۲ \log f(x)}{\partial x^۲}|_{x=x_۰}.$$

اگر $x_۰$ برابر با مد و $x^* = \arg\max_x \log f(x)$ باشد، آن‌گاه $\frac{\partial \log f(x)}{\partial x}|_{x=x_۰} = ۰$ و بنابراین

$$\log f(x) \approx \log f(x^*) + \frac{(x-x^*)^۲}{۲}\frac{\partial^۲ \log f(x)}{\partial x^۲}|_{x=x_*}.$$

پس انتگرال مورد نظر به‌صورت زیر تقریب می‌شود:

$$\begin{aligned}\int f(x)\mathrm{d}x &\approx \int \exp\Big(\log f(x^*) + \frac{(x-x^*)^۲}{۲}\frac{\partial^۲ \log f(x)}{\partial x^۲}|_{x=x_*}\Big)\mathrm{d}x \\ &= \exp(\log f(x^*)) \int \exp\Big(\frac{(x-x^*)^۲}{۲}\frac{\partial^۲ \log f(x)}{\partial x^۲}|_{x=x_*}\Big)\mathrm{d}x\end{aligned}$$

[۴۷] Markov chain Monte Carlo sampling

که می‌توان آن را با تابع چگالی توزیع نرمال مرتبط ساخت. با قرار دادن $\sigma^{۲*} = -۱/\frac{\partial^۲ \log f(x)}{\partial x^۲}|_{x=x_*}$ داریم

$$\int f(x)\mathrm{d}x \approx \exp(\log f(x^*)) \int \exp(-\frac{(x-x^*)^۲}{۲\sigma^{۲*}})\mathrm{d}x$$

که انتگرال تابع هسته توزیع نرمال با میانگین x^* و واریانس $\sigma^{۲*}$ است. به‌طور دقیق‌تر، این انتگرال در فاصله (α, β) به صورت

$$\int_\alpha^\beta f(x)\mathrm{d}x \approx f(x^*)\sqrt{۲\pi\sigma^{۲*}}(\Phi(\beta)-\Phi(\alpha))$$

قابل تقریب است، که در این‌جا $\Phi(\cdot)$ تابع توزیع تجمعی $N(x^*, \sigma^{۲*})$ است.

به‌طور مشابه می‌توان نسخه چندمتغیره تقریب لاپلاس را معرفی کرد. فقط مشتق به بردار مشتق‌های جزیی و مشتق دوم به ماتریس مشتق‌های مرتبه دوم که به ماتریس هسیان[۴۸] معروف است، تبدیل می‌شوند. تقریب لاپلاس نیز بر حسب یک توزیع نرمال چندمتغیره نوشته می‌شود.

از این تقریب، به‌ویژه، برای محاسبه میانگین و سایر گشتاورهای توزیع‌های پسین توابعی هموار و مثبت از پارامترها در استنباط بیزی استفاده می‌شود. برای توضیح، فرض کنید $g(\cdot)$ تابعی مشتق‌پذیر و مثبت در فضای پارامتر است. میانگین پسین $g(\theta)$ را می‌توان به صورت

$$E[g(\theta)|x_۱, \ldots, x_n] = \frac{\int g(\theta)e^{\ell(\theta)}\pi(\theta)\mathrm{d}\theta}{\int e^{\ell(\theta)}\pi(\theta)\mathrm{d}\theta} \qquad (۱.۱)$$

نوشت، که در آن $\ell(\theta)$ لگاریتم تابع درستنمایی است. برای محاسبه این میانگین، در روش تقریب لاپلاس مخرج کسر عبارت (۱.۱) با استفاده از تقریب لاپلاس که در بالا توضیح داده شد، محاسبه می‌شود. اما برای صورت کسر، در تقریب لاپلاس، به‌جای بسط تابع زیر انتگرال حول مد توزیع پسین، حول مد خودش بسط داده می‌شود. فرض کنید

$$L(\theta) = \frac{۱}{n}\log(e^{\ell(\theta)}\pi(\theta)) = \frac{۱}{n}(\ell(\theta)+\log\pi(\theta))$$

$$L^*(\theta) = \frac{۱}{n}\log(g(\theta)e^{\ell(\theta)}\pi(\theta)) = \frac{۱}{n}(\log g(\theta)+\ell(\theta)+\log\pi(\theta)).$$

پس میانگین پسین (۱.۱) برابر

$$E[g(\theta)|x_۱, \ldots, x_n] = \frac{\int e^{nL^*(\theta)}\mathrm{d}\theta}{\int e^{nL(\theta)}\mathrm{d}\theta} \qquad (۲.۱)$$

خواهد شد. فرض کنید $\hat\theta$ مد توزیع پسین، یعنی مد $L(\theta)$ و $\sigma^۲ = -۱/L''(\hat\theta)$. روش لاپلاس با استفاده از بسط تیلور حول $\hat\theta$ و $L'(\hat\theta) = ۰$، تقریبی به صورت

$$\begin{aligned}\int e^{nL(\theta)}\mathrm{d}\theta &\approx \int \exp[n(L(\hat\theta)+(\theta-\hat\theta)L'(\hat\theta)+\frac{۱}{۲}(\theta-\hat\theta)^۲L''(\hat\theta))]\mathrm{d}\theta \\ &= \int \exp[nL(\hat\theta)-n(\theta-\hat\theta)^۲/(۲\sigma^۲)]\mathrm{d}\theta \\ &= \sqrt{۲\pi}\sigma n^{-۱/۲}e^{nL(\hat\theta)}\end{aligned}$$

[۴۸] Hessian

برای مخرج کسر (۲.۱) ارائه می‌دهد. به‌طور مشابه، اگر $\hat{\theta}^*$ مد L^* و $\sigma^{*۲} = -۱/L^{*\prime\prime}(\hat{\theta}^*)$ باشد، آن‌گاه تقریب لاپلاس صورت کسر (۲.۱) برابر $\sqrt{۲\pi}\sigma^* n^{-۱/۲} exp\{nL^*(\hat{\theta}^*)\}$ می‌شود. نسبت این دو عبارت، تقریبی به شکل

$$\hat{E}[g(\theta)|x_۱, \ldots, x_n] = (\sigma^*/\sigma)\exp\{n(L^*(\hat{\theta}^*) - L(\hat{\theta}))\}$$

برای میانگین توزیع پسین فراهم می‌آورد.

۲.۴.۱ الگوریتم‌های زنجیر مارکوف مونت کارلویی

در مواردی که توزیع پسین صورت بسته ندارد، باید از روش‌های تقریبی برای محاسبه آن استفاده کرد. یکی از روش‌های تقریب توزیع پسین استفاده از روش‌های شبیه‌سازی مونت کارلویی است. وظیفه الگوریتم‌های شبیه‌سازی مونت کارلویی، تولید نمونه‌هایی از یک توزیع احتمالی مفروض است. زمانی که توزیع‌ها پیچیده یا بعد توزیع پسین (بعد پارامترها) بالا باشد، اجرای این وظیفه به صورت مستقیم کاری سخت و پیچیده است. برای رفع این دشواری، الگوریتم‌های MCMC (متروپولیس و همکاران، ۱۹۵۳؛ هستینگز، ۱۹۷۰) معرفی شدند که به‌طور غیرمستقیم اقدام به تولید نمونه می‌کنند.

مکانیسم الگوریتم‌های MCMC، مبتنی بر تولید نمونه وابسته از یک زنجیر مارکوف است که توزیع ایستای آن، همان توزیع پسین مورد نظر باشد. برای معرفی جزییات الگوریتم‌های MCMC، ابتدا مساله انتگرال‌گیری مونت کارلویی را تشریح می‌کنیم.

انتگرال‌گیری مونت کارلویی

فرض کنید بخواهیم انتگرال

$$H = \int h(x)f(x)\mathrm{d}x \qquad (۳.۱)$$

را محاسبه کنیم، که در آن $h(\cdot)$ تابعی از متغیر تصادفی X با تابع چگالی $f(\cdot)$ است. چنان‌چه نتوان این انتگرال را با استفاده از محاسبات معمولی حل کرد، می‌توان از روش‌های عددی یا شبیه‌سازی برای تقریب H استفاده کرد. در این‌جا به معرفی روش‌های انتگرال‌گیری عددی نمی‌پردازیم. برای دیدن این روش‌ها و جزییات آن‌ها می‌توانید به پرس و همکاران (۲۰۰۷) و لانگ (۲۰۱۰) مراجعه کنید.

با داشتن نمونه $x^{(۱)}, \ldots, x^{(n)}$ از توزیع $f(x)$، تقریب مونت کارلوی H به صورت

$$\hat{H} = \frac{۱}{n}\sum_{i=۱}^{n} h(x^{(i)})$$

محاسبه می‌شود. با استفاده از قانون قوی اعداد بزرگ، زمانی که $n \to \infty$

$$\hat{H} \xrightarrow{a.s} H.$$

دقت این تقریب زمانی که حجم نمونه n بزرگ شود، افزایش می‌یابد و می‌توان آن را به‌صورت زیر ارزیابی کرد:

$$se(\hat{H}) = \sqrt{\frac{\sum_{i=1}^{n}(h(x^{(i)}) - \hat{H})^2}{n(n-1)}}.$$

توجه داشته باشید که این تقریب، معکوس نسبت ریشه دوم اندازه نمونه است. برای مثال اگر بخواهیم خطای استاندارد را نصف کنیم به چهار برابر n اولیه نیاز داریم. همچنین، بنا بر قضیه حد مرکزی (کسلا و برگر، ۲۰۰۲) اگر واریانس X متناهی باشد، چنان‌چه $n \to \infty$، $\hat{H}$ در توزیع به یک متغیر تصادفی نرمال همگرا می‌شود. یعنی

$$\hat{H} \xrightarrow{d} N(H, se(\hat{H})^2).$$

در چارچوب استنباط بیزی، $f(x)$ در (۳.۱) نقش تابع چگالی توزیع پسین مدل، $\pi(\theta|y)$، را دارد و برای انجام استنباط بیزی، باید انتگرال‌گیری توابعی مانند $h(\theta)$ را برحسب توزیع پسین به‌دست آوریم. یعنی باید کمیت‌هایی مانند

$$E(h(\theta)|y) = \int_{\theta\in\Psi} h(\theta)\pi(\theta|y)\mathrm{d}\theta$$

را محاسبه کنیم. بنابراین با داشتن نمونه $\{\theta^{(1)}, \ldots, \theta^{(n)}\}$ از توزیع پسین، می‌توان شاخص‌های پسین مورد نیاز را به طور تقریبی محاسبه کرد. برای مثال احتمال پسینی مجموعه A که به صورت

$$p(\theta \in A|y) = \int_{\theta\in A} \pi(\theta|y)\mathrm{d}\theta$$

تعریف می‌شود، به وسیله نسبت تعداد $(\#)$ نمونه‌های $\theta^{(i)}$ که متعلق به A هستند به صورت

$$\frac{\#(\theta^{(i)} \in A, i = 1, \ldots, n)}{n}$$

تقریب زده می‌شود. شاخص دلخواه دیگری از توزیع پسین، چندک مرتبه q (θ_q) است که به شکل

$$\int_{-\infty}^{\theta_q} \pi(\theta|y)\mathrm{d}\theta = q$$

تعریف می‌شود و با چندک تجربی مرتبه q نمونه مونت کارلویی تقریب زده می‌شود. برای محاسبه این تقریب‌ها تنها کافی است نمونه‌های مستقل از توزیع پسین در اختیار باشند که اگر $\pi(\theta|y)$ توزیع شناخته‌شده‌ای داشته باشد یا تولید مستقیم نمونه از آن راحت باشد، دردسرساز نیست.

تولید غیرمستقیم نمونه

به‌جز مثال‌های ساده و موارد خاص، تولید مستقیم نمونه از توزیع‌های پسین معمولاً کار ساده‌ای نیست. همان‌طور که در بالا اشاره شد، یک راه‌حل معرفی‌شده تولید نمونه به صورت غیرمستقیم است. این راه‌حل به کمک نظریه زنجیرهای مارکوف ارائه شده است. در واقع در رهیافت غیرمستقیم تولید نمونه به کمک زنجیرهای مارکوف، به‌جای شبیه‌سازی نمونه‌های مستقل از توزیع پسین، نمونه‌ای وابسته از زنجیر مارکوفی که توزیع ایستای آن $\pi(\theta|y)$ است، شبیه‌سازی می‌شود. سپس این نمونه می‌تواند

برای محاسبه کمیت‌های پسینی مورد نظر مانند میانگین، چندک‌ها و احتمال‌ها، مورد استفاده قرار گیرد. این روش ترکیبی از نمونه‌گیری مونت کارلویی و زنجیرهای مارکوف است که به روش نمونه گیری MCMC مشهور است.

با این‌که کشف MCMC به مطالعات متروپلیس و همکاران (۱۹۵۳) و هستینگز (۱۹۷۰) برمی‌گردد، اما استفاده شایع از آن به‌عنوان ابزاری برای اجرای استنباط‌های بیزی با کار گلفاند و اسمیت (۱۹۹۰) شروع شد. برای مشاهده تاریخچه پیدایش و گسترش الگوریتم‌های MCMC به رابرت و کسلا (۲۰۱۱) مراجعه کنید. امروزه به دلیل پیشرفت‌های اساسی ایجادشده در منابع محاسباتی ارزان توسط رایانه‌ها، الگوریتم‌های MCMC به هسته اصلی جعبه ابزار استنباط بیزی تبدیل شده‌اند.

برای تشریح الگوریتم‌های MCMC، فرض کنید مجموعه‌ای از متغیرهای تصادفی وابسته

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

داریم که هر یک روی فضای حالت χ تعریف شده است. ایده کلی الگوریتم‌های MCMC ساخت زنجیر مارکوفی است که توزیع ایستای آن توزیع هدف π باشد که مایل به تولید نمونه از آن هستیم؛ به این معنی که اگر $X^{(t)} \sim \pi$ آن‌گاه $X^{(t+1)} \sim \pi$. بنابراین، اگر حجم نمونه تولیدشده از زنجیر مارکوف به اندازه کافی بزرگ باشد، نمونه تولیدشده، تحقق‌هایی از توزیع هدف محسوب می‌شوند. اطمینان از وجود چنین توزیع ایستای منحصر به فردی، مستلزم آن است که زنجیر مارکوف چند ویژگی را داشته باشد:

- تحویل‌ناپذیر[۴۹] باشد. به این معنی که صرف نظر از مقدار اولیه $X^{(0)}$، با احتمال مثبت زنجیر به هر ناحیه از χ دسترسی داشته باشد.
- برگشت‌پذیر[۵۰] باشد. به این معنی که زنجیر به هر حالتی از فضای χ به هر تعدادی که مایل باشد، بتواند منتقل شود. به عبارت دیگر، تعداد دفعات مورد انتظار بازگشت زنجیر برای یک مجموعه $A \subseteq \chi$ نامتناهی باشد.
- نامتناوب[۵۱] باشد. به این معنی که زنجیر در جستجوی فضای حالت، گرفتار هیچ چرخشی نشود. به عبارت دیگر، زنجیر در یک مجموعه موضعی گیر نکند.

تحت این شرایط، توزیع ایستای π یک توزیع حدی هم خواهد بود. به این معنی که $X^{(t)}$ در توزیع به π همگرا می‌شود. یعنی برای $t \to \infty$

$$X^{(t)} \xrightarrow{d} \pi$$

و نقطه شروع تاثیری بر این همگرایی ندارد. بنابراین احتمال $\pi(A) = \int_A \pi(x)\mathrm{d}x$ را می‌توان به وسیله $P(X^{(t)} \in A)$ برای $A \subset \chi$ تقریب زد. نتیجه این ویژگی همگرایی، قضیه ارگودیک[۵۲] است که

[۴۹] Irreducible

[۵۰] Recurrent

[۵۱] Aperiodic

[۵۲] Ergodic theorem

بیان می‌کند میانگین تجربی تابع دلخواه $h(x)$ حاصل از نمونه‌های وابسته تولیدی از زنجیر مارکوف، با احتمال متمایل به ۱، به مقدار واقعی مورد انتظار $E(h(X)) = \int_{\chi} h(x)\pi(x)\mathrm{d}x$ همگرا می‌شود. یعنی

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

این قضیه مشابه قانون قوی اعداد بزرگ برای نمونه‌های مستقل و هم توزیع است. برای ساختن چنین زنجیرهای مارکوفی از دو الگوریتم معروف روش‌های MCMC، یعنی الگوریتم‌های متروپلیس-هستینگز[۵۳] (MH) و نمونه‌گیر گیبز[۵۴] استفاده می‌شود که در ادامه به معرفی آن‌ها می‌پردازیم.

الگوریتم متروپولیس- هستینگز

الگوریتم MH ابتدا توسط متروپولیس و همکاران (۱۹۵۳) پیشنهاد شد و سپس توسط هستینگز (۱۹۷۰) به طور کامل تدوین یافت. تعریفی از این الگوریتم در چیب و گرینبرگ (۱۹۹۵) آمده است. برای توضیح آن، پارامتر θ را یک بعدی در نظر می‌گیریم. برای تعریف الگوریتم، مقدار اولیه (نقطه شروع زنجیر) را برابر $\theta^{(0)}$ قرار می‌دهیم. سپس t امین تکرار الگوریتم شامل مراحل زیر است:

۱. نمونه‌گیری یک مقدار منتخب θ^* از توزیع پیشنهادی با تابع چگالی $q(\theta^*|\theta^{(t-1)})$

۲. محاسبه نرخ پذیرش

$$r = \frac{\pi(\theta^*|y)q(\theta^{(t-1)}|\theta^*)}{\pi(\theta^{(t-1)}|y)q(\theta^*|\theta^{(t-1)})}$$

۳. پذیرش θ^* به‌عنوان مقدار $\theta^{(t)}$ با احتمال $min(1, r)$ و رد آن با احتمال $1 - min(1, r)$ و نسبت دادن مقدار قبلی $\theta^{(t-1)}$ برای $\theta^{(t)}$.

برای اجرای این الگوریتم، تابع چگالی پیشنهادی $q(\cdot|\cdot)$ طوری انتخاب می‌شود که تکیه‌گاه آن با تکیه‌گاه توزیع پسین یکی باشد. دم‌های آن از دم‌های توزیع پسین پهن‌تر و تولید نمونه از آن ساده باشد. معمولاً انتخاب یک تابع چگالی متقارن ارجحیت دارد که در این صورت $q(\theta^*|\theta^{(t-1)}) = q(\theta^{(t-1)}|\theta^*)$. برای مثال، یک توزیع یکنواخت یا توزیع نرمالی که حول مقدار قبلی مرکزی شده باشد، می‌تواند نامزد انتخاب توزیع پیشنهادی باشد. یعنی

$$q(\theta^*|\theta^{(t-1)}) \sim U(\theta^{(t-1)} - \gamma, \theta^{(t-1)} + \gamma)$$

یا

$$q(\theta^*|\theta^{(t-1)}) \sim N(\theta^{(t-1)}, \gamma^2)$$

که در آن پارامتر γ پارامتر تنظیم[۵۵] یا مقیاس است و با توجه به عملکرد الگوریتم انتخاب می‌شود. به الگوریتم MH با این انتخاب توزیع پیشنهادی، الگوریتم MH قدم زدن تصادفی می‌گویند. برای

[۵۳] Metropolis-Hastings

[۵۴] Gibbs sampler

[۵۵] Tuning parameter

مرحله تصمیم‌گیری برای پذیرش یا رد نمونه پیشنهادی، مرحله ۲، در عمل یک عدد تصادفی u از توزیع یکنواخت استاندارد تولید می‌شود. آن‌گاه اگر $u < r$ تصمیم پذیرش و در غیر این‌صورت رد اتخاذ می‌شود. با اجرای این الگوریتم یک زنجیر مارکوف تولید می‌شود که توزیع ایستای آن $\pi(\theta|y)$ خواهد بود (رابرت و کسلا، ۲۰۰۴).

نمونه‌گیر گیبز

روش نمونه‌گیری گیبز ابتدا توسط گمان و گمان (۱۹۸۴) در پردازش مدل‌های تصویر معرفی شد. سپس توسط گلفاند و اسمیت (۱۹۹۰) به‌عنوان روشی بر مبنای نمونه‌گیری برای محاسبه چگالی‌های حاشیه‌ای در چارچوب استنباط بیزی به‌کار گرفته شد. توضیح واضح و ساده این الگوریتم به وسیله کسلا و جورج (۱۹۹۲) بیان شده است. توجه کنید زمانی که توزیع شرطی کامل[۵۶] به‌عنوان توزیع پیشنهادی انتخاب شود، نمونه‌گیر گیبز یک حالت خاص الگوریتم MH است.

برای تشریح این روش، فرض کنید علاقه‌مندیم از توزیع پسین $\pi(\theta|y)$ که $\theta = \{\theta_1, \ldots, \theta_p\}$ یک پارامتر p بعدی است، نمونه تولید کنیم. برای اجرای نمونه‌گیر گیبز باید برای هر پارامتر θ_i توزیع شرطی کامل یعنی توزیع پسین شرطی θ_i به شرط همه پارامترهای دیگر را که با $\pi(\theta_i|\theta_{i\neq j}, y)$ برای $i, j = 1, \ldots, p$ مشخص می‌شود، به دست آوریم. برای شبیه‌سازی از توزیع پسین توام $\pi(\theta|y)$، نمونه‌گیر گیبز مقادیر را به‌طور تکراری از تمام توزیع‌های شرطی کامل تولید می‌کند. برای شروع مقادیر اولیه $\theta^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_p^{(0)}\}$ را انتخاب می‌کنیم. آن‌گاه در t امین تکرار، به شرط داشتن نمونه $\theta^{(t-1)}$ که در گام قبلی به‌دست آمده است، به صورت زیر عمل می‌کنیم:

۱. تولید نمونه $\theta_1^{(t)}$ از توزیع شرطی کامل $\pi(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_p^{(t-1)}, y)$

۲. تولید نمونه $\theta_2^{(t)}$ از توزیع شرطی کامل $\pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_p^{(t-1)}, y)$

.

.

.

p. تولید نمونه $\theta_p^{(t)}$ از توزیع شرطی کامل $\pi(\theta_p|\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, y)$ و قرار دادن

$$\theta^{(t+1)} = \{\theta_1^{(t+1)}, \ldots, \theta_p^{(t+1)}\}.$$

تحت شرایط نظم، ثابت می‌شود t امین تحقق $\theta^{(t)} = \{\theta_1^{(t)}, \ldots, \theta_p^{(t)}\}$ در توزیع به توزیع هدف $\pi = \pi(\theta|y)$ همگرا می‌شود (رابرت و کسلا، ۲۰۰۴).

[۵۶] Full conditional

رهیافت‌های جانشین MCMC

نظریه MCMC با تعداد تکرار بی‌شمار، همگرایی توزیع نمونه‌های شبیه‌سازی‌شده به توزیع هدف (یعنی توزیع پسین) را تضمین می‌کند، اما اجرای نامتناهی یک زنجیر مارکوف عملی نیست. بنابراین باید الگوریتم به تعدادی اجرا شود که همگرایی به توزیع ایستا نتیجه شود. تشخیص همگرایی زنجیر تولیدشده مساله اساسی در الگوریتم‌های MCMC است. ابزارهای تشخیص همگرایی شامل نمودارهای خود همبستگی و اثر، نرخ پذیرش الگوریتم، و برخی آزمون‌ها مانند جی‌وک (جی‌وک، ۱۹۹۲)، گلمن-روبین (گلمن و روبین، ۱۹۹۲) و رفتری-لوئیس (رفتری و لوئیس، ۱۹۹۲) ارائه شده‌اند. نمودار اثر ابزاری است که می‌تواند برای تشخیص آمیختگی[۵۷] زنجیر نیز به‌کار رود. این ویژگی مربوط به سرعت پیمایش توزیع پسین توسط زنجیر مارکوف است. به‌عنوان مثال، اگر آمیختگی زنجیر ضعیف باشد، به معنی سرعت کند پیمایش زنجیر و در نتیجه همبستگی شدید نمونه‌های تولیدشده است. در این صورت، زمان زیادی صرف دستیابی به همگرایی می‌شود. علاوه بر این، همبستگی روی دقت تقریب‌های کمیت‌های پسینی موثر است. گیر (۱۹۹۲) نشان داد که همبستگی بالاتر نمونه‌های تولیدشده، دلیلی برای دقت پایین‌تر تقریب توزیع پسین است. خودهمبستگی را می‌توان به وسیله باریک کردن[۵۸] مجموعه نقاط شبیه‌سازی‌شده یا بهبود الگوریتم بهینه (مثلاً از طریق بازپارامتری کردن نمونه‌گیری گیبز یا به وسیله تنظیم توزیع پیشنهادی MH) کاهش داد. زمانی که استنباط به روش MCMC انجام می‌شود، انرژی و زمان زیادی صرف تنظیم و مشاهده همگرایی می‌شود. به منظور اطلاع از تنظیمات بهتر (پارامتری کردن، توزیع‌های پیشین، مقادیر اولیه و توزیع‌های پیشنهادی MH) که خروجی دقیق‌تر و مطمئن‌تر برای MCMC به‌وجود می‌آورند، به بروکس و همکاران (۲۰۱۱) و مراجع مربوط به اجرای MCMC مراجعه کنید.

مشکل دیگری که نمی‌توان آن را نادیده گرفت، هزینه محاسباتی روش‌های MCMC است. زمانی که با مدل‌های پیچیده، مثل مدل‌های سلسله‌مراتبی[۵۹]، یا مجموعه داده‌های حجیم سروکار داریم، الگوریتم‌های MCMC ممکن است به شدت کند و حتی غیرعملی باشند.

با توجه به این مشکلات ذکرشده برای الگوریتم‌های مبتنی بر نمونه‌گیری MCMC، در سال‌های اخیر محققان متعددی به دنبال یافتن روش‌های جانشین بوده‌اند. یکی از رهیافت‌های جانشین مناسب برای روش‌های MCMC که مورد نظر ما در این پایان‌نامه است و هزینه محاسباتی استنباط بیزی را به شدت کاهش می‌دهد، روش INLA است که توسط رو و همکاران (۲۰۰۹) معرفی شد و در فصل بعدی به معرفی آن می‌پردازیم.

---

[۵۷] Mixing

[۵۸] Thinning

[۵۹] Hierarchical models

فصل ۲

تقریب لاپلاس آشیانی جمع‌بسته

الگوریتم‌های MCMC تا به امروز ابزار استاندارد برای استنباط بیزی مدل‌های با ساختار سلسله مراتبی بوده‌اند. اما وقتی بعد مدل بزرگ باشد، ویژگی‌های همگرایی و آمیختگی این الگوریتم‌ها با مشکل مواجه می‌شوند. از طرفی وجود همبستگی قوی بین مولفه‌های توزیع پسین، این مشکل را دوچندان می‌کند. خوشبختانه استفاده از راهبرد به‌روزرسانی بلوکی[۱] (هلد و رو، ۲۰۰۲؛ رو و هلد، ۲۰۰۵) تا حدی این مشکلات را مرتفع ساخته است، اما مشکل زمان‌بر بودن محاسبات به قوت خود باقی است. رو و همکاران (۲۰۰۹) روش تقریب لاپلاس آشیانی جمع‌بسته، INLA، را معرفی کردند که در تحلیل تقریبی بیزی رده مدل‌های گاوسی پنهان[۲] (LGM)، جایگزین مناسبی برای الگوریتم‌های MCMC است و تقریبی قابل قبول در مدت زمان بسیار کوتاه را جایگزین شبیه‌سازی‌های سنگین MCMC می‌کند. این روش آن‌قدر دقیق عمل می‌کند که نتایج حاصل از آن با نتایج به‌دست آمده از الگوریتم‌های MCMC تفاوت بسیار ناچیزی دارند. در این فصل استنباط بیزی تقریبی مبتنی بر روش INLA در مدل‌های گاوسی پنهان با متغیر پاسخ غیرگاوسی (غیرنرمال) مورد نظر است که در آن برای پارامترهای رگرسیونی و اثرات تصادفی از توزیع پیشین گاوسی استفاده می‌شود. در این روش تقریب لاپلاس (تقریب گاوسی توزیع پسین) به‌عنوان یکی از بخش‌های اصلی روش INLA محسوب می‌شود. پس از بیان گام‌های این روش تقریبی، به جزئیات محاسبه چگالی‌های پسین کناری مورد نظر می‌پردازیم. در بخش آخر نیز به منظور مقایسه مدل‌های مختلف، معیارهای ارزیابی بیزی مدل را معرفی می‌کنیم.

---

[۱] Block updating

[۲] Latent Gaussian models

۱.۲ فرآیند تصادفی گاوسی

در این بخش ابتدا چند تعریف که مورد نیاز است، بیان می‌شوند و سپس به معرفی فرآیند تصادفی گاوسی خواهیم پرداخت.

تعریف ۱.۱.۲. فرآیند تصادفی

خانواده متغیرهای تصادفی $\{X_t, t \in T\}$ را که روی یک فضای احتمال تعریف شده‌اند، یک فرآیند تصادفی با فضای پارامتر T و فضای وضعیت S می‌گویند، هرگاه S مجموعه تمام مقادیری باشد که به ازای هر $t \in T$ متغیر تصادفی X_t می‌تواند اختیار کند.

تعریف ۲.۱.۲. فرآیند قدم زدن تصادفی[۳]

- فرآیند تصادفی w_t را یک فرآیند قدم زدن تصادفی مرتبه اول (rw1) می‌نامند، هرگاه

$$w_t = w_{t-۱} + \epsilon_t$$

که در آن ϵ_t مولفه اغتشاش محض است، یعنی ϵ_t ها ناهمبسته هستند و از توزیعی با میانگین صفر و واریانس ثابت $\sigma_\epsilon^۲$ پیروی می‌کنند.

- فرآیند تصادفی w_t را یک فرآیند قدم زدن تصادفی مرتبه دوم (rw2) گویند، هرگاه

$$w_t = ۲w_{t-۱} - w_{t-۲} + \epsilon_t$$

که در آن ϵ_t مشابه فرآیند قدم زدن تصادفی مرتبه اول تعریف می‌شود.

می‌توان نشان داد که این فرآیند، ایستا نیست و رفتاری به شدت ناپایدار دارد.

تعریف ۳.۱.۲. فرآیند تصادفی گاوسی

فرآیند تصادفی $\{X_t, t \in T\}$ را یک فرآیند گاوسی (نرمال) می‌نامند، هرگاه هر ترکیب خطی متناهی از متغیرهای تصادفی X_t، دارای توزیع نرمال باشد.

در این صورت برای هر $t \in T$، X_t دارای توزیع نرمال است و به‌ویژه $EX_t^۲ < \infty$. فرآیندهای گاوسی دارای ویژگی‌های نظری دلخواه هستند. این فرآیندها همچنین در موارد کاربردی، به‌خصوص در مهندسی و علوم فیزیکی مورد استفاده قرار می‌گیرند. در برخی از متون مانند متون آمار فضایی، به فرآیندهای تصادفی گاوسی میدان تصادفی گاوسی[۴] (GRF) نیز می‌گویند.

---

[۳] Random walk process

[۴] Gaussian random field

۱.۱.۲ مدل‌های رگرسیون جمعی ساختاری

مدل‌های رگرسیون جمعی ساختاری[۵] (STAR) قالبی انعطاف‌پذیر برای مدل‌بندی اثرات خطی و غیرخطی متغیرهای تبیینی شامل مدل‌های خطی تعمیم‌یافته و مدل‌های جمعی تعمیم‌یافته هستند. توزیع متغیر پاسخ y_i، $i = ۱, \ldots, n$ در این مدل‌ها متعلق به خانواده نمایی با لگاریتم تابع چگالی یا جرم احتمال به شکل زیر است:

$$\ln \pi(y_i|u_i, \beta) = y_i \left[\alpha + \sum_{j=۱}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki} \right] - a \left(\alpha + \sum_{j=۱}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki} \right) + c(y_i)$$

که در آن $\{\beta_k\}$ اثرات ثابت خطی متغیرهای تبیینی $z_i = (z_{۱i}, \ldots, z_{n_\beta i})$، $\{f^{(j)}(\cdot)\}$ توابعی از متغیرهای تبیینی $u_i = (u_{۱i}, \ldots, u_{n_f i})$ و $a(\cdot)$ تابعی معلوم، هستند. در این مدل

$$\eta_i = \alpha + \sum_{j=۱}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=۱}^{n_\beta} \beta_k z_{ki} \tag{۱.۲}$$

پارامتر متعارف (کانونی) خانواده توزیع‌های نمایی محسوب می‌شود و به پیش‌گوی جمعی ساختاری معروف است. اگر

$$\mu_i = E[y_i|u_i, \beta] = a'(\eta_i)$$

آنگاه تابع پیوند $g(\cdot)$ میانگین شرطی را به پیش‌گوی جمعی ساختاری به صورت زیر پیوند می‌دهد:

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

توابع $\{f^{(j)}(\cdot)\}$ می‌توانند شامل اثرات غیرخطی مانند روندهای زمانی، اثرات فصلی، اثرات تصادفی و وابسته به زمان یا اثرات هموار در مدل باشند. با تعیین $f^{(j)}(\cdot)$های مختلف می‌توان زیررده‌های متفاوت و پرکاربرد متعددی از این رده از مدل‌ها را معرفی کرد. به‌عنوان مثال برخی از این زیررده‌ها عبارتند از:

الف) مدل‌های رگرسیونی، شامل مدل‌های خطی تعمیم‌یافته، مدل‌های اسپلاین تاوانیده[۶] (لانگ و برزگر، ۲۰۰۴) مدل‌های قدم زدن تصادفی (رو و هلد، ۲۰۰۵) و فرآیندهای گاوسی (چو و قهرمانی، ۲۰۰۵).

ب) مدل‌های پویا[۷] (وست و هریسون، ۱۹۹۷)، که شامل مدل‌هایی هستند که همبستگی زمانی مشاهدات در آن‌ها به کمک یک متغیر تبیینی وارد مدل می‌شود.

[۵] Structured additive regression model

[۶] Penalized spline models

[۷] Dynamic models

ج) مدل‌های فضایی[۸] یا فضایی-زمانی[۹] که شامل مدل‌هایی هستند که ساختار وابستگی فضایی یا فضایی-زمانی داده‌ها از طریق $f^{(j)}$ وارد مدل می‌شود (بنرجی و همکاران، ۲۰۰۸).

مدل‌های گاوسی پنهان

در مدل (۱.۲) فرض کنید میدان تصادفی پنهان[۱۰] $x = \{\alpha, f^{(1)}, \cdots, f^{(n_f)}, \beta_1, \cdots, \beta_{n_\beta}, \eta_i\}$ دارای توزیع نرمال چندمتغیره با بردار میانگین صفر و ماتریس دقت[۱۱] $Q_{\theta_1} = \Sigma_{\theta_1}^{-1}$ باشد، که در آن Σ_{θ_1} ماتریس کواریانس است که به پارامتر θ_1 بستگی دارد. به این مدل که زیررده‌ای از مدل‌های STAR را تشکیل می‌دهد، رده مدل‌های گاوسی پنهان[۱۲] (LGM) می‌گویند.

فرض کنید I مجموعه اندیس‌گذاری باشد که بردار متغیر پاسخ در نقاط این مجموعه مشاهده شده‌اند، یعنی $y = \{y_i; i \in I\}$، و مولفه‌های y به شرط x و θ_2 مستقل از هم و دارای توزیع توام $\pi(y|x, \theta_2)$ باشند. در این صورت، تابع چگالی پسین میدان تصادفی پنهان x و θ به صورت

$$\begin{aligned}
\pi(x, \theta|y) &\propto \pi(\theta)\pi(x \mid \theta)\pi(y|x, \theta) \\
&\propto \pi(\theta)\pi(x|\theta) \prod_{i \in I} (y_i|x_i, \theta) \\
&\propto \pi(\theta)|Q(\theta)|^{1/2} \exp\Big(-\frac{1}{2} x^T Q(\theta) x \Big) \prod_{i \in I} \exp(\log(\pi(y_i|x_i, \theta))) \\
&\propto \pi(\theta)|Q(\theta)|^{1/2} \exp\Big(-\frac{1}{2} x^T Q(\theta) x + \sum_{i \in I} \log(\pi(y_i|x_i, \theta)) \Big)
\end{aligned}$$

خواهد بود، که در آن $\theta = (\theta_1^T, \theta_2^T)$ بردار m بعدی ابرپارامترهای مدل و $\pi(\theta)$ توزیع پیشین θ هستند. به منظور کارا بودن روش INLA، مجموعه I باید تنک[۱۳] باشد. به عبارت دیگر، میدان تصادفی پنهان x، که اغلب دارای بعدی بزرگ (مثلا 10^2 تا 10^5) است، ویژگی استقلال شرطی را می‌پذیرد و بنابراین میدان تصادفی پنهان یک میدان تصادفی مارکوفی گاوسی[۱۴] (GMRF) با ماتریس دقت تنک Q_θ است.

برای مجموعه‌های چگال I مانند داده‌های زمین‌آماری فضایی[۱۵]، به کمک رهیافت معادلات دیفرانسیل جزیی تصادفی[۱۶] می‌توان مجموعه چگال را به کمک مثلث‌سازی به یک مجموعه با ویژگی مارکوفی تبدیل کرد به‌طوری که اجرای روش INLA برای داده‌ها در مجموعه جدید تقریبی، کارا باشد. ویژگی مارکوفی رده LGM باعث افزایش سرعت تجزیه ماتریس‌ها می‌شود که برای اجرای روش‌های نمونه‌گیری یا تقریب زدن توزیع پسین مدل الزامی است. افزایش سرعت نیز مرهون تنک بودن ماتریس دقت است، زیرا

---

[۸] Spatial models
[۹] Spatio-temporal models
[۱۰] Latent
[۱۱] Precision matrix
[۱۲] Latent Gaussian models
[۱۳] Sparse
[۱۴] Gaussian markov random field
[۱۵] Spatial geostatistical data
[۱۶] Stochastic partial differential equations

روش‌های عددی تجزیه ماتریس‌های تنک بسیار سریع‌تر از روش‌های محاسباتی ماتریس‌های چگال هستند.

شرط دیگری که روش INLA باید داشته باشد تا اجرای آن کارا (سریع) باشد، این است که تعداد ابرپارامترها عددی کوچک مانند $m \leq ۶$ باشد. این محدودیت به دلیل استفاده از روش‌های عددی انتگرال‌گیری در روش INLA است که برای کارایی روش‌های عددی انتگرال‌گیری، نباید بعد انتگرال‌ها بیشتر از ۶ باشد. قبل از تشریح روش تقریبی مورد نظر ما، برای درک بهتر آن، رده GMRF را با جزییات بیشتری معرفی می‌کنیم.

۲.۱.۲ فرآیند تصادفی مارکوفی گاوسی

بهترین راه برای تفهیم ویژگی مارکوفی یک میدان تصادفی گاوسی، استفاده از تعریف گراف است. گراف g مجموعه‌ای از راس‌ها است که توسط خانواده‌ای از یال‌ها به هم وصل شده‌اند و به صورت زوج مرتب (ν, ε) نشان داده می‌شود، که در آن ν مجموعه‌ای متناهی و غیرتهی از رئوس و ε یال‌های آن است. اگر راس‌ها به‌صورت $\nu = \{۱, \ldots, n\}$ باشند، گراف را نشاندار می‌نامند.

فرض کنید بردار تصادفی $x = (x_۱, \ldots, x_n)^T$ دارای توزیع نرمال چندمتغیره با میانگین μ و ماتریس کواریانس Σ باشد. همچنین فرض کنید $g = (\nu, \varepsilon)$ گرافی نشاندار باشد که در آن ε شامل همه زوج‌های $\{i, j\}$ است به‌طوری که راس‌های i و j هیچ یال مشترکی نداشته باشند، اگر و تنها اگر $x_i \perp x_j | x_{-ij}$. در این صورت x یک میدان تصادفی مارکوفی گاوسی نسبت به گراف g نامیده می‌شود. قبل از بیان یک تعریف رسمی برای GMRF، ارتباط بین گراف g و پارامترهای توزیع نرمال را بیان می‌کنیم. از آن‌جا که μ تاثیری بر استقلال شرطی دوبه‌دوی درایه‌های x ندارد، می‌توان نتیجه گرفت اطلاعات مربوط به استقلال شرطی فقط در ماتریس کواریانس Σ یا عکس آن یعنی ماتریس دقت، Q، پنهان شده‌اند. پس ماتریس دقت نقش مهمی در تحلیل GMRF بازی می‌کند.

قضیه ۴.۱.۲. (رو و هلد، ۲۰۰۵). اگر x دارای توزیع نرمال چندمتغیره با میانگین μ و ماتریس دقت معین مثبت $Q > ۰$ باشد، آن‌گاه درایه ij ام ماتریس Q، یعنی Q_{ij} برابر صفر است، اگر و تنها اگر $x_i \perp x_j | x_{-ij}$.

با توجه به قضیه ۴.۱.۲، درایه‌های ناصفر ماتریس Q گراف g را مشخص می‌کنند و بر اساس آن‌ها می‌توان استقلال شرطی x_i و x_j را بررسی نمود. این نتیجه دلیل استفاده از ماتریس دقت به‌جای ماتریس کواریانس در رده LGM را نشان می‌دهد.

تعریف ۵.۱.۲. بردار تصادفی $x = (x_۱, \ldots, x_n)^T \in R^n$ یک میدان تصادفی مارکوفی گاوسی تحت گراف نشاندار $G = (\nu, \varepsilon)$ با میانگین μ و ماتریس دقت $Q > ۰$ است، اگر و تنها اگر تابع چگالی آن به‌صورت

$$\pi(x) = (۲\pi)^{-\frac{n}{۲}} |Q|^{\frac{۱}{۲}} exp\{-\frac{۱}{۲}(x-\mu)^T Q (x-\mu)\}$$

باشد و برای هر $i, j \in \nu$ که $i \neq j$

$$Q_{ij} \neq ۰ \iff \{i, j\} \in \varepsilon.$$

قضیه ۶.۱.۲. اگر x یک GMRF تحت گراف $g = (\nu, \varepsilon)$ با میانگین μ و ماتریس دقت $Q > ۰$ باشد، آن‌گاه میانگین، دقت و همبستگی شرطی به ترتیب عبارتند از

$$E(x_i|x_{-i}) = \mu_i - \frac{۱}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}(x_j - \mu_j)$$

$$Prec(x_i|x_{-i}) = Q_{ii}$$

$$Corr(x_i, x_j|x_{-ij}) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}} \quad i \neq j$$

که در آن منظور از $i \sim j$ راس‌های یک یال مشترک است.

درایه‌های قطری ماتریس Q دقت‌های شرطی x_i به شرط x_{-i} و مولفه‌های خارج از قطر اطلاعاتی در مورد همبستگی شرطی بین x_i و x_j به شرط x_{-ij} ارائه می‌کنند. اما از آن‌جا که درایه‌های ماتریس کواریانس $\Sigma = (\Sigma_{ij})$، واریانس $Var(x_i) = \Sigma_{ii}$ و همبستگی

$$Corr(x_i, x_j) = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$$

را نتیجه می‌دهند، ماتریس کواریانس اطلاعاتی در مورد واریانس کناری x_i و همبستگی کناری بین x_i و x_j در اختیار قرار می‌دهد. تفسیر کناری با Σ مستقیماً حاوی اطلاعات مفیدی است و می‌توان از آن برای بیان توزیع‌های کناری یک یا دومتغیره استفاده کرد، در حالی که در مورد Q این کار دشوار است و باید از توزیع توام روی x_{-i} یا x_{-ij} انتگرال گرفت؛ زیرا با توجه به تعریف $Q^{-۱} = \Sigma$، در حالت کلی Σ_{ii} به همه مولفه‌های Q مرتبط است.

در بیشتر موارد Q ماتریسی تنک است و تنها $O(n)$ درایه از $n^۲$ درایه ماتریس Q ناصفر است که این به علت ویژگی مارکوفی میدان تصادفی است. تنک بودن این ماتریس موجب سرعت بخشیدن به اجرای الگوریتم تجزیه $Q = LL^T$ می‌شود که در قالب قضیه زیر بیان می‌شود.

قضیه ۷.۱.۲. (رو و هلد، ۲۰۰۵). فرض کنید x یک GMRF تحت گراف g، با میانگین μ و ماتریس دقت $Q > ۰$ باشد. همچنین فرض کنید برای $۱ \leq i < j \leq n$،

$$F(i, j) = \{i + ۱, \ldots, j - ۱, j + ۱, \ldots, n\}$$

مجموعه اندیس‌های بزرگ‌تر از i به‌جز j باشد. در این‌صورت $x_i \perp x_j|x_{F(i,j)}$، اگر و تنها اگر $L_{ji} = ۰$ که در آن L_{ji} درایه ji ام ماتریس L در تجزیه چولسکی ماتریس Q است.

اگر بتوان تشخیص داد $L_{ji} = ۰$ است، دیگر نیازی به محاسبه آن در الگوریتم تجزیه چولسکی نخواهد بود و این امر موجب کاهش محاسبات می‌شود.

۲.۲ تقریب گاوسی توزیع‌ها

یک شیوه معمول برای تقریب توزیع‌هایی که صورت نامعلومی دارند، استفاده از تقریب گاوسی[۱۷] است. انتخاب یک مقدار اولیه مناسب در این شیوه بسیار مهم است، چرا که در سرعت همگرایی تاثیرگذار است. یک انتخاب مناسب، مد توزیع هدف است. برای این منظور یوون و ویلسون (۲۰۰۹) شیوه مناسبی را برای یافتن مد توزیع هدف (از جمله توزیع‌های پسین حاشیه‌ای پارامترها و میدان تصادفی x) در حداقل تکرار ممکن پیشنهاد دادند، به‌طوری که این امر در تسریع تقریب گاوسی تاثیر بسزایی دارد.

در این بخش ابتدا تقریب گاوسی برای توزیع‌های یک‌متغیره را همراه با یک مثال ساده تشریح می‌کنیم. سپس تعمیم آن برای حالت چندمتغیره را مطرح خواهیم کرد.

- الف) حالت تک‌متغیره:

فرض کنید یک مشاهده از توزیع پواسون با میانگین λ داریم و توزیع پیشین نرمال را برای $\eta = \log\lambda$ با مشخصات $\eta \sim N(\mu, k^{-1})$ انتخاب کنیم. بنابراین توزیع پسین $\pi(\eta|y)$ به صورت

$$\begin{aligned}\pi(\eta|y) &\propto \pi(\eta)\pi(y|\eta)\\ &\propto \exp\{-\frac{k}{2}(\eta-\mu)^2 + y\eta - \exp(\eta)\}\\ &= \exp(f(\eta))\end{aligned}$$

است، که در آن $f(\eta) = -\frac{k}{2}(\eta-\mu)^2 + y\eta - \exp(\eta)$. روشی معمول برای تقریب $f(\eta)$ استفاده از بسط تیلور مرتبه دوم آن حول مد، η_0، به صورت

$$\begin{aligned}f(\eta) &\approx f(\eta_0) + f'(\eta_0)(\eta-\eta_0) + \frac{1}{2}f''(\eta_0)(\eta-\eta_0)^2\\ &= a + b\eta - \frac{1}{2}c\eta^2\end{aligned} \qquad (۲.۲)$$

است، که در آن $a = f(\eta_0) - \eta_0 f'(\eta_0) + \frac{\eta_0^2}{2}f''(\eta_0)$، $b = f'(\eta_0) - f''(\eta_0)\eta_0$ و $c = -f''(\eta_0)$. بنابراین می‌توان $\pi(\eta|y)$ را به‌صورت

$$\tilde{\pi}(\eta|y) \propto \exp(-\frac{1}{2}c\eta^2 + b\eta)$$

تقریب زد، که صورت کانونی توزیع نرمال $N_c(b,c)$ با میانگین $\mu_1(\eta_0) = \frac{b}{c}$ و دقت $k_1(\eta_0) = c$ است.

- ب) حالت چندمتغیره:

[۱۷] Gaussian approximation

فرض کنید n مشاهده مستقل شرطی $y_1, \dots, y_n$ از یک توزیع غیرنرمال در اختیار داریم و x یک GMRF با ماتریس دقت Q و میانگین μ است. در این صورت تابع چگالی شرطی کامل $\pi(x|y,\theta)$ به صورت

$$\pi(x|y,\theta) \propto \exp\Big(-\frac{1}{2}(x-\mu)^T Q(x-\mu) + \sum_{i=1}^{n} \log \pi(y_i|x_i,\theta)\Big)$$

قابل بازنویسی است. مشابه با رابطه (۲.۲)، بسط تیلور مرتبه دوم عبارت

$$g(x_i) = \log \pi(y_i|x_i,\theta)$$

حول مد μ_0 را می‌توان به صورت

$$g(x_i) = a_i + b_i x_i - \frac{1}{2} c_i x_i^2$$

نوشت، که در آن $a_i = g(\mu_0) - \mu_0 g'(\mu_0) + \frac{\mu_0^2}{2} g''(\mu_0)$، $b_i = g'(\mu_0) - g''(\mu_0)\mu_0$ و $c_i = -g''(\mu_0)$. در این صورت تقریب گاوسی چگالی $\pi(x|y,\theta)$ به صورت

$$\begin{aligned}\tilde{\pi}_G(x|y,\theta) &\propto \exp\Big(-\frac{1}{2}x^T Q x + \mu^T Q x + \sum_i (a_i + b_i x_i - \frac{1}{2} c_i x_i^2)\Big) \\ &\propto \exp\Big(-\frac{1}{2}x^T Q x + \mu^T Q x + \sum_i (b_i x_i - \frac{1}{2} c_i x_i^2)\Big)\end{aligned}$$

حاصل می‌شود که وقتی $c_i = -\dfrac{\partial^2 \log \pi(y|x,\theta)}{\partial x_i^2}$ منفی باشد، مقدار آن صفر است. این تقریب را می‌توان به صورت ماتریسی

$$\begin{aligned}\tilde{\pi}_G(x|y,\theta) &\propto \exp\Big(-\frac{1}{2}x^T Q x + \mu^T Q x + b^T x - \frac{1}{2} x^T diag(c) x)\Big) \\ &\propto \exp\Big(-\frac{1}{2}x^T (Q + diag(c)) x + (Q\mu + b)^T x\Big)\end{aligned}$$

نوشت که صورت کانونی توزیع نرمال

$$N_c(Q\mu + b, Q + diag(c))$$

با بردار میانگین $\mu_1 = (Q\mu + b)(Q + diag(c))^{-1}$ و ماتریس دقت $Q + diag(c)$ است که در آن $diag(c)$ یک ماتریس قطری با عناصر روی قطر c_i ها می‌باشد. مشابه با حالت تک‌متغیره b و c به مقدار μ_0 وابسته‌اند و می‌توان با تکرار این فرآیند و بسط آن حول μ_1 که به مد توزیع $\pi(x|y,\theta)$ نزدیک‌تر است، تقریب را بهبود بخشید.

باید توجه داشت که هر چند تقریب گاوسی قابل قبول و اغلب سریع است، اما اگر تقریب توزیع‌های غیرخطی مراتب بالا مورد نظر باشد همگرایی این تقریب با کندی انجام می‌شود. همچنین گاهی نسبت به خطای ناشی از انتخاب نادرست مد و در نظر نگرفتن چولگی توزیع حساسیت زیادی دارد.

۳.۲ تقریب توزیع‌های پسین با روش INLA

در این بخش مراحل انجام روش تقریب INLA را برای محاسبه تقریب توزیع‌های پسین کناری میدان تصادفی پنهان که به صورت

$$\pi(x_i|y) = \int \pi(x_i|\theta, y)\pi(\theta|y)\mathrm{d}\theta, \qquad i = ۱, \ldots, n \tag{۳.۲}$$

است، ارائه می‌دهیم. این تقریب در سه گام اجرا می‌شود:

گام اول: محاسبه تقریب توزیع پسین کناری $\pi(\theta|y)$

گام دوم: محاسبه تقریب $\pi(x_i|\theta, y)$ برای مقادیر منتخب θ

گام سوم: محاسبه تقریب پسین بیزی میدان پنهان با استفاده از مجموع متناهی

$$\tilde{\pi}(x_i|y) = \sum_{k=۱}^{n} \tilde{\pi}(x_i|\theta_k, y)\tilde{\pi}(\theta_k|y)\Delta_k \tag{۴.۲}$$

که در آن Δ_k وزن‌هایی هستند که به مقادیر θ_k اختصاص داده می‌شوند.

ویژگی مخصوص INLA این است که توزیع (۳.۲) را با استفاده از تقریب‌های آشیانی محاسبه می‌کند. بعداً توضیح می‌دهیم که چگونه شیوه آشیانی، تقریب لاپلاس دقیقی را ایجاد می‌کند و موجب تسریع محاسبات می‌شود.

۱.۳.۲ تقریب توزیع پسین توأم ابرپارامترها با راهبرد توری

توزیع پسین توأم بردار ابرپارامترهای مدل به صورت زیر تعریف می‌شود:

$$\begin{aligned}
\pi(\theta|y) &= \frac{\pi(\theta, y)}{\pi(y)} \\
&= \frac{\pi(\theta, y)}{\pi(y)} \frac{\pi(x|\theta, y)}{\pi(x|\theta, y)} \\
&= \frac{\pi(x, \theta, y)}{\pi(y)\pi(x|\theta, y)} \\
&= \frac{\pi(y|x, \theta)\pi(x|\theta)\pi(\theta)}{\pi(y)\pi(x|\theta, y)} \\
&\propto \frac{\pi(y|x, \theta)\pi(x|\theta)\pi(\theta)}{\pi(x|\theta, y)}
\end{aligned} \tag{۵.۲}$$

روش INLA، توزیع پسین کناری (۵.۲) را به صورت

$$\tilde{\pi}(\theta|y) \propto \frac{\pi(y|x, \theta)\pi(x|\theta)\pi(\theta)}{\tilde{\pi}_G(x|\theta, y)}|_{x=x^*(\theta)} \tag{۶.۲}$$

تقریب می‌زند، که در آن مخرج کسر با استفاده از تقریب گاوسی محاسبه شده است و x^* مد توزیع $\pi(x|\theta,y)$ می‌باشد. خطای این تقریب بعد از نرمال‌سازی از مرتبه $O(n^{-3/2})$ است (مارتینو، ۲۰۰۷). عبارت (۶.۲) هم‌ارز با تقریب لاپلاس توزیع پسین کناری بردار θ است (تیرنی و کدین، ۱۹۸۶).

تقریب $\tilde{\pi}(\theta|y)$ درسه مرحله از فرآیند اجرای روش INLA مورد استفاده قرار می‌گیرد. استفاده اصلی آن در محاسبه تقریب توزیع پسین کناری میدان پنهان است. در گام دوم، از (۶.۲) برای محاسبه تقریب کناری درستنمایی $\pi(y)$ استفاده می‌شود. گام سوم به منظور محاسبه کناری پسین برای برخی از عناصر بردار θ یعنی $\pi(\theta_j|y)$، $j = 1,\dots,m$، به کار می‌رود. نحوه به کارگیری (۶.۲) برای محاسبه سه مرحله یادشده، شامل انتگرال‌گیری عددی روی محدوده چندبعدی از θ است. در نتیجه پیش از هر کاری ابتدا باید نقاط انتگرال‌گیری مناسب انتخاب شوند. رو و همکاران (۲۰۰۹) دو راهبرد توری[۱۸] و طرح مرکب مرکزی[۱۹] (CCD) را برای انتخاب نقاط انتگرال‌گیری پیشنهاد کردند. راهبرد توری شامل انتخاب نقاطی از یک توری است که در آن‌ها چگالی $\tilde{\pi}(\theta|y)$ مقدار بیشتری دارد. زمانی که بعد θ بزرگ نباشد (کوچکتر از ۶ باشد)، این شیوه بسیار مناسب و دقیق است. در مواردی که بعد بردار θ بین ۶ تا ۱۲ باشد، راهبرد CCD مفیدتر است. در هر دو راهبرد فرض می‌شود $\tilde{\pi}(\theta|y)$ تک‌مدی است، چرا که در فرآیند انتخاب نقاط انتگرال‌گیری به مد $\tilde{\pi}(\theta|y)$ نیاز داریم که با θ^* نشان داده می‌شود و می‌توان آن را با استفاده از الگوریتم‌های بهینه‌سازی مانند نیوتون-رافسون (تی جالینگ، ۱۹۹۵) یا شبه‌نیوتون[۲۰] (مورای، ۱۹۷۲) محاسبه کرد.

در ادامه این بخش، روش توری شرح داده می‌شود. انتخاب نقاط انتگرال‌گیری با روش توری در سه گام به شرح زیر انجام می‌شود:

گام اول: یافتن مد $\tilde{\pi}(\theta|y)$ با بهینه‌سازی $log\{\tilde{\pi}(\theta|y)\}$ نسبت به θ.

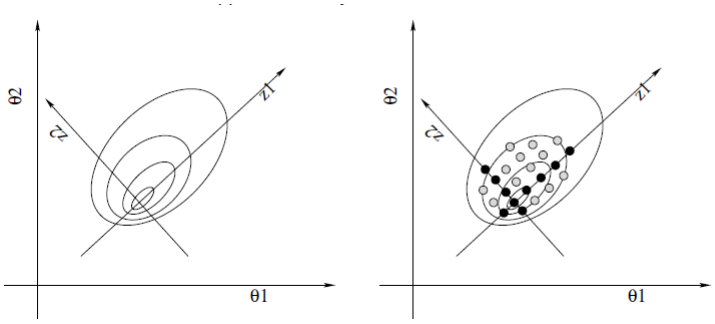
گام دوم: محاسبه ماتریس هسیان در نقطه θ^* با استفاده از تفاضل‌های متناهی[۲۱] (اسکال نیک و گامبینی، ۲۰۰۱). با فرض $\Sigma = H^{-1}$، برای ساده‌سازی انتگرال‌گیری عددی و تصحیح اثر مقیاس و دوران از تبدیل استاندارد z به‌جای θ به صورت

$$\theta(z) = \theta^* + V\Lambda^{1/2}z \qquad (۷.۲)$$

استفاده می‌شود (اسمیت و همکاران، ۱۹۸۷)، که در آن ماتریس‌های V و Λ از تجزیه ماتریس Σ به‌صورت $\Sigma = V\Lambda V^T$ به دست می‌آیند. به‌عنوان مثال، برای حالت دوبعدی $z = (z_1, z_2)$، با قرار دادن $z = (0,0)$ در رابطه (۷.۲)، $\theta(0) = \theta^*$ حاصل می‌شود. اگر تابع چگالی $\tilde{\pi}(\theta|y)$ گاوسی با ماتریس کوواریانس Σ باشد، آن‌گاه $z \sim N(0, I)$.

گام سوم: کاوش $\log\{\tilde{\pi}(\theta|y)\}$ با استفاده از متغیر z به صورت زیر:
نقطه شروع از مختصات جدید در مد تابع چگالی $\log\{\tilde{\pi}(\theta|y)\}$ است. در جهت مثبت محور z_1

[۱۸] Grid
[۱۹] Central composite design
[۲۰] Quasi newton
[۲۱] Finite differences



شکل ۱.۲: انتخاب نقاط انتگرال‌گیری با وسیله راهبرد توری

با گامی به طول δ_z پیش می‌رویم تا جایی که نامساوی

$$log\{\tilde{\pi}(\theta(\circ)|y)\} - log\{\tilde{\pi}(\theta(z)|y)\} < \delta_z \qquad (۸.۲)$$

برقرار شود. سپس به‌طور مشابه همین عمل در جهت محور $z_۲$ انجام می‌شود. شکل ۱.۲ نمودار تراز $\log\{\tilde{\pi}(\theta|y)\}$ را برای $m = ۲$ به همراه محورهای مختصات جدید در مد توزیع پسین را نشان می‌دهد. پس از طی این مراحل، نقاط مشکی رنگ (توپر) در شکل ۱.۲ حاصل می‌شوند. سپس با یافتن مختصاتی که حاصل از ترکیب مختصات نقاط مشکی است موقعیت‌های نقاطی که باید فضای بین نقاط مشکی را پر کنند (نقاط خاکستری رنگ توخالی) به دست می‌آیند به‌طوری که شرط (۸.۲) برای این نقاط نیز برقرار است. به این ترتیب، نقاط θ_k در یک تور منظم مشخص می‌شوند. همه این نقاط برای محاسبه تقریب (۴.۲) وزن‌های یکسانی خواهند داشت. رو و همکاران (۲۰۰۹) مقادیر ۱ و ۲/۵ را برای δ_z پیشنهاد کردند. البته اگر دقت بیشتری در محاسبه چگالی‌های پسین مورد نظر نیاز باشد، می‌توان از $\delta_z = \frac{۱}{۲}$ یا $\delta_z = \frac{۱}{۴}$ استفاده کرد. الگوریتم ۱.۲.۳ نحوه انتخاب نقاط انتگرال‌گیری با این روش را نشان می‌دهد.

الگوریتم ۱.۲.۳ (کاوش $\tilde{\pi}(\theta|y)$ با راهبرد توری):

۱. مد تابع چگالی $log\{\tilde{\pi}(\theta|y)\}$ را بیابید. سپس ماتریس هسیان در مد را محاسبه کرده و قرار دهید $\Sigma = H^{-۱}$.

۲. تجزیه $\Sigma = V\Lambda V^T$ را انجام دهید.

۳. قرار دهید $i = ۱$.

۴. قرار دهید $z = \circ$ و $dir = ۱$.

۵. مادامی که

$$log\{\tilde{\pi}(\theta(\circ)|y)\} - log\{\tilde{\pi}(\theta(z)|y)\} < \delta_z$$

مقادیر $z = z + dir \times \underline{۱}_i$، $\theta(z) = \theta^* + V\Lambda^{\frac{۱}{۲}} z$ و $\tilde{\pi}(\theta(z)|y)$ محاسبه شوند.

۶. قرار دهید $dir = -1$ و به گام ۵ برگردید.

۷. قرار دهید $i = i + 1$ و چنان‌چه $i < m$، به گام ۴ برگردید.

۸. مختصات نقاط خاکستری رنگ را محاسبه کنید.

۲.۳.۲ تقریب توزیع پسین کناری ابرپارامترها

محاسبه توزیع کناری پسین θ_j، $j = 1, \ldots, m$، مستلزم انتگرال‌گیری از تابع چگالی $\tilde{\pi}(\theta|y)$ نسبت به بردار θ_{-j} به صورت

$$\tilde{\pi}(\theta_j|y) = \int \tilde{\pi}(\theta|y)\mathrm{d}\theta_{-j}$$

است. اما اگر بعد بردار θ بزرگ باشد، محاسبه این انتگرال زمان‌بر است. یک شیوه جایگزین، انتگرال‌گیری از $\tilde{\pi}(\theta|y)$ با استفاده از نقاط انتگرال‌گیری است که حاصل از راهبرد توری یا CCD هستند. این روش نه تنها ساده‌تر است بلکه سرعت محاسبات را به مراتب افزایش می‌دهد.

۳.۳.۲ تقریب توزیع پسین کناری عناصر میدان پنهان

پس از انتخاب نقاط انتگرال‌گیری $\{\theta_k\}$ باید توزیع‌های کناری پسین x_i ها محاسبه شوند تا بتوان با استفاده از دو تقریب $\tilde{\pi}(\theta_j|y)$ و $\tilde{\pi}(x_i|\theta_j, y)$ چگالی (۳.۲) را تقریب زد. بنابراین، در این بخش تقریب توزیع پسین $\pi(x_i|\theta_j, y)$ با استفاده از سه روش تقریب گاوسی، لاپلاس و لاپلاس ساده‌شده[۲۲] محاسبه و معایب و محاسن هر یک بیان می‌شوند.

الف) تقریب گاوسی

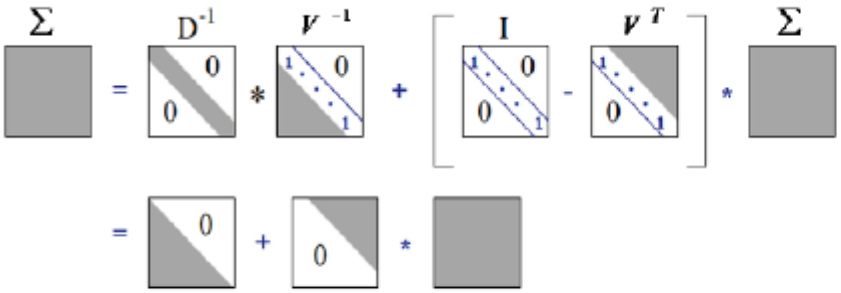
ساده‌ترین راه محاسبه تقریب $\pi(x_i|\theta_j, y)$ استفاده از $\tilde{\pi}_G(x|\theta_j, y)$ و استخراج توزیع کناری مورد نظر از آن است. در مرحله انتخاب نقاط انتگرال‌گیری θ_k ها و محاسبه $\tilde{\pi}(\theta_j|y)$، تقریب گاوسی $\tilde{\pi}_G(x|\theta_j, y)$ محاسبه شد. بنابراین بردار میانگین معلوم است و فقط نیاز به محاسبه واریانس کناری است. رو و مارتینو (۲۰۰۷) واریانس کناری را با استفاده از یک رابطه بازگشتی محاسبه کردند و نشان دادند تقریب

$$\tilde{\pi}_G(x_i|\theta_j, y) \sim N(\mu_i(\theta_j), \sigma_i^2(\theta_j)) \qquad (۹.۲)$$

برای $\pi(x_i|\theta_j, y)$ اغلب نتایج قابل قبولی می‌دهد، که در آن $\sigma^2(\theta_j)$ بردار n بعدی واریانس کناری است که در ادامه نحوه محاسبه آن به تفصیل بیان می‌شود.

محاسبه واریانس کناری در GMRF: همان‌طور که قبلا بیان شد، معمولا یک GMRF با ماتریس دقت Q مشخص می‌شود. برای محاسبه واریانس کناری، می‌توان Q را وارون کرد. چون معمولا بعد میدان تصادفی بزرگ است، وارون کردن ماتریسی با ابعاد بزرگ بسیار سخت و زمان‌بر است. برای این

[۲۲] Simplified Laplace



شکل ۲.۲: نمایش ماتریسی معادله تاکاهاشی

منظور، یک رابطه بازگشتی بر اساس معادلات تاکاهاشی (تاکاهاشی و همکاران، ۱۹۷۳) معرفی شد که واریانس کناری را برای یک GMRF با ابعاد بزرگ، بسیار سریع محاسبه می‌کند. اگر $Q = LL^T$ و $z \sim N(0, I)$ باشد، آن‌گاه $L^T x = z$، که در آن $x \sim N(0, \Sigma)$. معادله $L^T x = z$ را می‌توان به صورت

$$L_{ii} x_i = z_i - \sum_{k=i+1}^{n} L_{ki} x_k \qquad i = 1, \ldots, n$$

نوشت. با ضرب طرفین این معادله در x_j به ازای $j \geq i$ و امید گرفتن از آن، درایه‌های ماتریس Σ از رابطه

$$\Sigma_{ij} = \frac{\delta_{ij}}{L_{ii}} - \frac{1}{L_{ii}} \sum_{k=i+1}^{n} L_{ki} \Sigma_{kj}, \; j \geq i, \; i = 1, \ldots, n \qquad (۱۰.۲)$$

محاسبه می‌شوند، که در آن $\delta_{ij} = 1$ اگر $i = j$ و $\delta_{ij} = 0$ اگر $i \neq j$ باشد. معادله (۱۰.۲) همان معادله تاکاهاشی است که به محاسبه سریع عناصر معینی از وارون یک ماتریس با استفاده از تجزیه چولسکی آن کمک می‌کند، به این ترتیب که اگر $Q = LL^T = VDV^T$، که در آن $L = VD^{1/2}$، D یک ماتریس قطری و V یک ماتریس پایین مثلثی با درایه‌های قطری یک است. بنابراین

$$\Sigma = D^{-1} V^{-1} + (I - V^T)\Sigma. \qquad (۱۱.۲)$$

شکل ۲.۲ نواحی ناصفر را برای رابطه (۱۱.۲) نشان می‌دهد. همان‌طور که واضح است حاصل ضرب $D^{-1}V^{-1}$ یک ماتریس پایین مثلثی است بهطوری که $(D^{-1}V^{-1})_{ii} = (D^{-1})_{ii}$. این امر موجب بی‌نیازی به محاسبه V^{-1} می‌شود. بنابراین درایه‌های ماتریس Σ به صورت

$$\Sigma_{ij} = D_{ij}^{-1} - \sum_{k>j}^{n} V_{kj} \Sigma_{ik}, \qquad i \geq j$$

محاسبه می‌شوند، که پایه‌ای برای محاسبه سریع واریانس‌های کناری x_1 تا x_n فراهم می‌کند.

از معایب تقریب گاوسی این است که امکان دارد این تقریب تحت تاثیر خطای ناشی از تشخیص نادرست مد یا در نظر نگرفتن چولگی توزیع قرار گیرد. به همین دلیل ممکن است یک توزیع چوله با یک توزیع نرمال تقریب زده شود.

ب) تقریب لاپلاس

یک راه مناسب برای بهبود تقریب گاوسی، استفاده از تقریب لاپلاس به صورت

$$\tilde{\pi}_{LA}(x_i|\theta,y) \propto \frac{\pi(x,\theta,y)}{\tilde{\pi}_{GG}(x_{-i}|x_i,\theta,y)}|_{x_{-i}=x^*_{-i}(x_i,\theta)} \qquad (۱۲.۲)$$

است، که در آن $\tilde{\pi}_{GG}$ تقریب گاوسی و $x^*_{-i}(x_i,\theta)$ مد تابع چگالی شرطی $\pi(x_{-i}|x_i,\theta,y)$ است. توجه کنید که $\tilde{\pi}_{GG}$ با تقریب تابع چگالی شرطی متناظر با $\tilde{\pi}_G(x|\theta,y)$ متفاوت است. در $\tilde{\pi}_G$ ماتریس دقت نسبت به x_i ثابت و میانگین آن تابعی خطی از x_i است، در حالی‌که در $\tilde{\pi}_{GG}$ ابتدا مد $x^*_{-i}(x_i,\theta)$ و سپس لگاریتم درستنمایی حول آن محاسبه می‌شود که ماتریس دقت آن با تغییر x_i تغییر می‌کند. از آن‌جا که تابع چگالی $\tilde{\pi}_{GG}$ بر اساس شرطی کردن روی x_i و سپس به‌کار بردن تقریب لاپلاس بر روی بقیه متغیرها محاسبه می‌شود، از این رو از تقریب (۹.۲) که تنها براساس تقریب گاوسی است دقیق‌تر می‌باشد.

محاسبه عبارت (۱۲.۲) مستلزم محاسبه $\tilde{\pi}_{GG}$ برای هر x_i و θ است که بسیار زمان‌بر است. رو و همکاران (۲۰۰۹) دو تعدیل برای آسان کردن محاسبات ارائه دادند. در تعدیل اول به دلیل این‌که میانگین شرطی $E[x_{-i}|x_i,\theta,y]$ و مد شرطی $x^*_{-i}(x_i,\theta)$ به شرط گاوسی بودن توزیع $(x_{-i}|x_i,\theta,y)$ بر هم منطبق خواهند بود و از طرفی به دلیل در نظر گرفتن پیشین گاوسی برای تابع چگالی $\pi(x|\theta)$، گاوسی بودن توزیع $(x_{-i}|x_i,\theta,y)$ خیلی دور از انتظار نیست. بنابراین می‌توان بر اساس

$$x^*_{-i}(x_i,\theta) \approx E_{\pi_G}(x_{-i}|x_i) \qquad (۱۳.۲)$$

از میانگین به‌جای مد استفاده و از گام بهینه‌سازی $\tilde{\pi}_{GG}(x_{-i}|x_i,\theta,y)$ خوداری کرد. هاسیو و همکاران (۲۰۰۴) نشان دادند که استفاده از $E_{\pi_G}(x_{-i}|x_i)$ به جای $x^*_{-i}(x_i,\theta)$ نتایجی تقریباً مشابه را منجر می‌شود.

در تعدیل دوم از این ایده استفاده می‌شود که تنها x_j هایی که در همسایگی x_i هستند بر توزیع کناری x_i اثر می‌گذارند. اگر همبستگی بین x_i و x_j به فاصله بین دو راس i و j تعبیر شود، فقط آن x_j هایی که در شعاع مشخصی از راس i هستند در محاسبه کناری x_i لحاظ خواهند شد، که این شعاع همسایگی با $R(\theta)$ نمایش داده می‌شود. رو و همکاران (۲۰۰۹) نشان دادند که با استفاده از (۱۳.۲) رابطه

$$\frac{E_{\pi_G}(x_j|x_i)-\mu_j(\theta)}{\sigma_j(\theta)} = a_{ij}(\theta)\frac{x_i-\mu_i(\theta)}{\sigma_i(\theta)}$$

برای بعضی a_{ij} که $i \neq j$ برقرار است. از این رو یک روش برای تعیین شعاع همسایگی، ساخت مجموعه $R(\theta)$ به صورت $R(\theta) = \{j : |a_{ij}| > ۰/۰۰۱\}$ است.

ج) تقریب لاپلاس ساده‌شده

در تقریب لاپلاس حتی اگر از میانگین به‌جای مد استفاده شود، تقریب هنوز از لحاظ محاسبات سنگین و زمان‌بر است؛ زیرا لگاریتم مخرج کسر (۱۲.۲) برابر است با

$$\log\{\tilde{\pi}_{GG}(x_{-i}|x_i,\theta,y)\}|_{x_{-i}=E_{\tilde{\pi}_G}(x_{-i}|x_i)} \propto c_0 + \frac{1}{2}\log|Q^* + diag(c)| \qquad (۱۴.۲)$$

که در آن $c_i = -\frac{\partial^2 \pi(y_i|x_i,\theta)}{\partial x_j^2}|_{x_j=E_{\tilde{\pi}_G}(x_{-i}|x_i)}$ و $Q^* = Q_{[-i,i]}$ یک زیرماتریس اصلی Q است. در محاسبه دترمینان موجود در (۱۴.۲) باید برای هر x_i یک ماتریس $(N-1)\times(N-1)$ منظور شود، که خیلی زمان‌بر است. رو و همکاران (۲۰۰۹) تقریب لاپلاس ساده‌شده را معرفی کردند که این مشکل را برطرف می‌کند. در این روش با تعریف $x_i^{(s)} = \frac{x_i - \mu_i(\theta)}{\sigma_i(\theta)}$، صورت کسر (۱۲.۲) حول $x_i = \mu_i(\theta)$ به صورت

$$\begin{aligned}&\log\{\pi(x,\theta,y)\}|_{x_{-i}=E_{\tilde{\pi}_G}(x_{-i}|x_i)} = \\ &-\frac{1}{2}(x_i^{(s)})^2 + \frac{1}{6}(x_i^{(s)})^3 \sum_{j\in I} d_j^{(3)}\{\mu_i(\theta),\theta\}\{\sigma_i(\theta)a_{ij}(\theta)\}^3 + \dots\end{aligned} \qquad (۱۵.۲)$$

بسط داده می‌شود. جملات مرتبه اول و دوم، تقریب گاوسی و جمله مرتبه سوم تصحیحی برای چولگی فراهم می‌کند. در بسط لگاریتم مخرج کسر (۱۲.۲) با توجه به این‌که برای هر ماتریس M، رابطه

$$\frac{\partial}{\partial x_i}\log|M| = Trace(M^{-1}\frac{\partial}{\partial x_i}M)$$

برقرار است (ماردیا و همکاران، ۱۹۸۰)، رابطه

$$\begin{aligned}\frac{d\log|Q^* + diag(c)|}{dx_i} &= Trace[Q^* + diag(c)]^{-1}\frac{d|Q^* + diag(c)|}{dx_i} \\ &= Trace[Q^* + diag(c)]^{-1} diag[d^3(x_i,\theta)] \\ &= \sum_j Var(x_j|x_i)d_j^3(x_i,\theta) \\ &= \sum_j \sigma_j(\theta)\{1 - Corr_{\tilde{\pi}_G}(x_i,x_j|\theta)^2\}d_j^3(x_i,\theta)\end{aligned} \qquad (۱۶.۲)$$

به دست می‌آید، زیرا $Var(x_j|x_i) = \sigma_j(\theta)\{1 - Corr_{\tilde{\pi}_G}(x_i,x_j|\theta)^2\}$. با استفاده از رابطه (۱۶.۲) بسط تیلور مخرج کسر (۱۲.۲) حول $x_i = \mu_i(\theta)$ به صورت

$$\begin{aligned}&\log\{\tilde{\pi}_{GG}(x_{-i}|x_i,\theta,y)\}|_{x_{-i}=E_{\tilde{\pi}_G}(x_{-i}|x_i)} \\ &\propto c_0 - \frac{1}{2}(x_i^{(s)})\sum_{j\in I\setminus i}\sigma_j(\theta)\{1 - Corr_{\tilde{\pi}_G}(x_i,x_j|\theta)^2\} \\ &d_j^3(\mu_i(\theta),\theta)\sigma_j(\theta)a_{ij}(\theta) + \dots\end{aligned} \qquad (۱۷.۲)$$

حاصل می‌شود، که در آن c_0 مقداری ثابت است. اگر با توجه به (۱۵.۲) و (۱۷.۲) روابط

$$\gamma_i^{(1)}(\theta) = \frac{1}{2}\sum_{j\in I\setminus i}\sigma_j^2(\theta)\{1 - Corr_{\tilde{\pi}_G}(x_i,x_j|\theta)^2\}d_j^3(\mu_i(\theta),\theta)\sigma_j(\theta)a_{ij}(\theta)$$

$$\gamma_i^{(۳)}(\theta) = \sum_{j \in I \setminus i} d_j^{(۳)}\{\mu_i(\theta), \theta\}\{\sigma_i(\theta) a_{ij}(\theta)\}^۳$$

تعریف شوند، آن گاه تقریب لاپلاس ساده‌شده به صورت زیر به دست می‌آید:

$$\tilde{\pi}_{SLA}(x_i^s|\theta, y) = c_۰ - \frac{۱}{۲}(x_i^{(s)})^۲ + \gamma_i^{(۱)}(\theta)(x_i^{(s)}) + \frac{۱}{۶}(x_i^{(s)})^۳\gamma_i^{(۳)}(\theta) + \dots$$

۴.۲ معیارهای ارزیابی

برای مقایسه مدل‌ها، باید معیارهای مقایسه مناسب تعریف شوند. در رهیافت بیزی، برای مقایسه مدل‌های سلسله‌مراتبی از معیارهایی نظیر عامل بیزی[۲۳] و معیار انحراف اطلاع[۲۴] (DIC) استفاده می‌شود. همچنین برای مقایسه تقریب‌های توزیع‌های پسین پارامترها با توزیع‌های اصلی، از معیار فاصله کولبک-لیبلر[۲۵] استفاده می‌شود. به‌منظور انتخاب بهترین مدل برای پیش‌گویی از بین مدل‌های برازش داده‌شده مناسب، استفاده از معیارهای اندازه پیشگویی[۲۶] می‌تواند مفید باشد. بسته نرم‌افزاری R-INLA معیارهای ذکرشده را به‌طور خودکار محاسبه و گزارش می‌کند. در ادامه این معیارها را به‌طور خلاصه معرفی می‌کنیم.

۱.۴.۲ اندازه‌های پیشگویی

اندازه‌های پیشگویی کمیت‌هایی هستند که برای مقایسه مدل‌ها در یافتن مدل برتر برای هدف پیش‌گویی به‌کار می‌روند (گلمن و همکاران، ۲۰۰۴). افزون بر این، چنین اندازه‌هایی به‌عنوان ابزاری برای یافتن مشاهدات پرت نیز مفید هستند (پتیت و یانگ، ۱۹۹۰). یکی از این اندازه‌ها، مولفه‌های پیش‌گوی شرطی[۲۷] (CPO) است که برای y_i ها به شکل زیر تعریف می‌شود:

$$CPO_i = \pi(y_i|y_{-i}), \quad i = ۱, \dots, n$$

که در آن y_{-i} همان بردار y به جز درایه i ام است. مقادیر صفر و خیلی کوچک CPO_i بیانگر پرت بودن y_i متناظر است. علاوه بر این از منهای میانگین لگاریتم CPO_i ها برای مقایسه مدل‌ها نیز استفاده می‌شود که آن را معیار نمره لگاریتمی (LogScore) اعتبار سنجی متقابل (گنتینگ و رفتری، ۲۰۰۷) می‌گویند.

یکی دیگر از اندازه‌های پیش‌گویی، تبدیل انتگرالی احتمال[۲۸] (PIT) است که به صورت

$$PIT_i = p(y_i^{new} \leq y_i|y_{-i})$$

[۲۳] Bayes factor
[۲۴] Deviance information criterion
[۲۵] Kullback-Leibler distance
[۲۶] Predictive measures
[۲۷] Conditional predictive ordinates
[۲۸] Probability integral transform

تعریف می‌شود. با استفاده از این معیار نیز می‌توان به پرت بودن یک مشاهده پی برد. همچنین اگر هیستوگرام PIT ها تقریباً یکنواخت باشد، به معنی مناسب بودن مدل مورد نظر برای پیشگویی است.

۲.۴.۲ معیار انحراف اطلاع

معیار DIC اندازه‌ای از پیچیدگی و برازش مدل است که برای مقایسه مدل‌های پیچیده استفاده می‌شود (اشپیگل‌هالتر و همکاران، ۲۰۰۲). با تعریف DIC بر اساس لگاریتم تابع درستنمایی به شکل

$$D(X,\theta) = -۲\sum_{i\in I}\log\pi(y_i|x_i,\theta) + c$$

که در آن c مقداری ثابت است (دمپستر، ۱۹۷۴)، DIC به صورت

$$DIC = \bar{D} + p_D$$

تعریف می‌شود، که در آن

$$\bar{D} = E_{\theta|y}(D)$$

میانگین پسین انحراف‌ها و p_D تفاضل میانگین انحراف‌ها و انحراف میانگین‌ها است که به صورت

$$p_D = E_{\theta|y}(D) - D(E_{\theta|y}(\theta)) = \bar{D} - D(\bar{\theta})$$

تعریف می‌شود و بیانگر میزان پیچیدگی مدل است. مدل‌های با DIC کوچکتر داده‌ها را بهتر پشتیبانی می‌کنند.

فصل ۳

تحلیل بیزی پیشامدهای بازگشتی

در داده‌های پیشامدهای بازگشتی که در آن یک حادثه برای فردی چندین بار تکرار می‌شود، همبستگی به‌وجود می‌آید. در این داده‌ها بررسی اثرات عوامل خطر بر فاصله زمانی بین حوادث بیشتر مورد توجه است. برای تحقق این هدف، استفاده از مدل مخاطره‌های متناسب کاکس به دلیل پیش‌فرض استقلال داده‌ها مناسب نیست. بنابراین با در نظر گرفتن تعمیمی از این مدل که وابستگی پیشامدها را در مدل لحاظ می‌کند، استنباط‌های کاراتری به‌دست می‌آید. در این فصل نشان می‌دهیم که داده‌های بازگشتی را با مدل فرآیند پواسون ناهمگن به‌عنوان زیررده‌ای از مدل‌های گاوسی پنهان، که استنباط در آن‌ها با روش INLA قابل انجام است، می‌توان مدل‌بندی و تحلیل کرد.

۱.۳ فرآیندهای شمارشی

فرآیندهای تصادفی متعددی در طبیعت وجود دارند که ممکن است در پزشکی، زیست‌شناسی، فیزیک، جغرافیا، اقتصاد یا روان‌شناسی رخ دهند. اگر یک محقق بخواهد طبیعت احتمالی یک پدیده را بشناسد، بدون شک باید از فرآیندهای تصادفی استفاده کند. فرآیندهای شمارشی از جمله فرآیندهای تصادفی بسیار مفید هستند که در زمینه‌های مختلف کاربردی برای مدل‌بندی پدیده‌ها استفاده می‌شوند. این نوع فرآیند، وظیفه شمارش تعداد رخدادهای مورد نظر محققین را بر عهده دارند.

فرآیند تصادفی $\{N(t), t \geq 0\}$ را یک فرآیند شمارشی گویند، اگر $N(t)$ نشان‌دهنده تعداد حوادثی باشد که در فاصله زمانی $(0, t]$ رخ می‌دهند. تعریف دقیق‌تر آن به همراه چند ویژگی مفید این نوع فرآیند در ادامه بیان می‌شوند.

تعریف ۱.۱.۳. فرآیند تصادفی $\{N(t), t \geq ۰\}$ را یک فرآیند شمارشی می‌نامند، هرگاه در شرایط زیر صدق کند:

۱. مقادیر $N(t)$ صحیح باشند.

۲. $N(t) \geq ۰$

۳. اگر $t_۱ < t_۲$، آن‌گاه $N(t_۱) \leq N(t_۲)$

۴. برای $t_۱ < t_۲$، $N(t_۲) - N(t_۱)$ تعداد رخدادها در فاصله $(t_۱, t_۲]$ باشد.

تعریف ۲.۱.۳. فرآیند شمارشی دارای نموهای مستقل است، اگر تعداد پیشامدها در فاصله‌های زمانی جدا از هم، مستقل باشند. یعنی برای هر t و s تعداد پیشامدها تا زمان t، یعنی $N(t)$، مستقل از تعداد پیشامدها بین زمان‌های t و $t+s$، یعنی $N(t+s) - N(t)$ است.

تعریف ۳.۱.۳. یک فرآیند شمارشی دارای نموهای مانا است، هرگاه به ازای هر $h > ۰$ و $۰ \leq t_۱ \leq t_۲ \leq \ldots \leq t_n$ بردارهای

$$(N(t_۱), N(t_۲) - N(t_۱), \ldots, N(t_n) - N(t_{n-۱}))$$

و

$$(N(t_۱ + h), N(t_۲ + h) - N(t_۱ + h), \ldots, N(t_n + h) - N(t_{n-۱} + h))$$

هم‌توزیع باشند.

یکی از زیررده‌های فرآیندهای شمارشی، فرآیندهای پواسون هستند که کاربردهای متنوع و فراوانی در علوم پیدا کرده‌اند. در بخش بعدی به تشریح این نوع فرآیند شمارشی می‌پردازیم.

۱.۱.۳ فرآیندهای پواسون

یکی از مباحث مهم آمار، نظریه توزیع‌هاست. آشنایی هر چه بیشتر با توزیع‌ها و خواص آن‌ها ما را در استفاده بهتر و دقیق‌تر از آن‌ها یاری خواهد داد. یک موضوع اساسی در زمینه تحلیل داده‌ها، انتخاب و تعیین مدل مناسب جهت برآوردیابی و پیش‌بینی مشاهدات آینده است. برای تعیین مدل مناسب آماری و تحلیل‌های مربوط، لازم است در ابتدا توزیع‌هایی را که به نظر محقق در تحلیل داده‌ها مناسب هستند مورد بررسی قرار داده و مناسب‌ترین آن‌ها را انتخاب کنیم. هر پدیده تصادفی را که در آن نوعی شمارش مورد توجه باشد، در صورت برابری میانگین و واریانس، می‌توان با استفاده از توزیع پواسون مدل‌بندی کرد. این شمارش ممکن است تعداد تصادفات رانندگی منجر به مرگ در طی یک سال، تعداد ذرات رادیواکتیو منتشرشده در واحد زمان، تعداد مرگ و میر ناشی از یک بیماری در بازه زمانی معین یا به‌طور کلی تعداد پیشامدهای رخ‌داده در یک محدوده زمانی یا ناحیه مکانی مشخص باشد. ضابطه تابع احتمال توزیع پواسون به صورت زیر بیان می‌شود:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = ۰, ۱, \ldots$$

فرآیند پواسون یک فرآیند شمارشی است که در حالت عملی، تعداد وقوع یک پیشامد خاص را در طول زمان می‌شمارد و این تعداد از توزیع پواسون پیروی می‌کند. مثال‌هایی از این نوع فرآیند عبارتند از تعداد مشتریان واردشده به یک بانک در یک روز، تعداد وقوع زلزله در یک ناحیه معین، و تعداد دفعات قطعی برق طی یک ماه. این فرآیند یک ابزار مدل‌سازی بدیهی برای بسیاری از مسائل کاربردی است. با توجه به اسناد چاپی موجود، در سال ۱۷۶۷، میچل برای اولین بار در ستاره‌شناسی از ویژگی‌های فرآیند پواسون استفاده کرد. اما استفاده از نام فرآیند پواسون، توسط فرای به سال ۱۹۲۸ باز می‌گردد. این فرآیند نه تنها بسیاری از پدیده‌های اطراف زندگی بشر را مدل‌بندی می‌کند، بلکه امکان انجام تحلیل‌های ریاضی را هم فراهم می‌آورد. تعریف ریاضی این فرآیند به صورت زیر است:

تعریف ۴.۱.۳. فرآیند شمارشی $\{N(t), t \geq ۰\}$ یک فرآیند پواسون با نرخ λ است، هرگاه

۱. $N(۰) = ۰$

۲. فرآیند دارای نموهای مستقل و مانا باشد.

۳. تعداد پیشامدها در هر فاصله دلخواه به طول t دارای توزیع پواسون با میانگین λt باشد. یعنی به ازای هر $s, t \geq ۰$، برای هر $n = ۰, ۱, ۲, \ldots$، بتوان نوشت

$$P(N(t+s) - N(s) = n) = \frac{exp(-\lambda t)(\lambda t)^n}{n!}$$

تعریف معادل دیگری از فرآیند پواسون را می‌توان به صورت زیر تعریف کرد.

تعریف ۵.۱.۳. فرآیند شمارشی $\{N(t), t \geq ۰\}$ یک فرآیند پواسون با نرخ λ نامیده می‌شود، اگر

۱. $N(۰) = ۰$

۲. فرآیند دارای نموهای مستقل و مانا باشد.

۳. اگر $h \to ۰$ آن‌گاه $P(N(h) = ۱) = \lambda h + o(h)$.

۴. اگر $h \to ۰$ آن‌گاه $P(N(h) \geq ۲) = o(h)$.

۲.۱.۳ فرآیندهای پواسون ناهمگن

در نظریه مخاطره، معمولاً از فرآیندهای پواسون همگن برای مدل‌بندی فراوانی ادعاها استفاده می‌شوند که گاهی به دلیل ثابت بودن نرخ شدت ادعا، نتایج حاصل دور از ذهن به‌دست می‌آیند. یک مدل جامع‌تر وابسته به زمان، فرآیند پواسون ناهمگن است که نرخ شدت آن تابعی از زمان است. در مدلی که در این وضعیت به‌کار می‌رود باید یک تابع وابسته به زمان مناسب یا فرآیندی تصادفی مانند $\{\lambda(t)\}$ را جایگزین پارامتر ثابت λ کنیم.

فرآیندهای پواسون ناهمگن با تابع‌های شدت وابسته به زمان برای توصیف نوسانات مخاطره پدیده‌هایی مناسب‌اند که تغییرات شدت رخدادهای آن‌ها به صورت فصلی است. بسیاری از پدیده‌های طبیعی در یک محیط متناوب و تحت شرایط فصلی ظاهر می‌شوند.

تشابه بین تابع‌های نرخ شدت و نرخ شکست به ما کمک می‌کنند که کاربردهای متفاوت فرآیندهای پواسون ناهمگن را بیابیم. بعضی از ویژگی‌های فرآیندهای پواسون ناهمگن با نرخ شکست (تک) دوره‌ای، توسط چوکوا و همکاران (۱۹۹۳) استخراج شده‌اند. این ویژگی‌ها توسط گاریدو و همکاران (۱۹۹۶) در نظریه مخاطره به کار گرفته شده‌اند. برگ و هابرمن (۱۹۹۴) از یک فرآیند زاد مارکوف ناهمگن به منظور پیش‌بینی روند ادعاهای بیمه عمر استفاده کرده‌اند که فرآیند پواسون ناهمگن حالت خاصی از آن است. بعضی از مسائل مربوط به ورشکستگی در فضاهای تناوبی نیز توسط آسمیوسن و رولسکی (۱۹۹۴) مورد بررسی قرار گرفته‌اند.

تعریف ۶.۱.۳. فرآیند شمارشی $\{N(t), t \geq ۰\}$ را یک فرآیند پواسون ناهمگن با تابع شدت $\lambda(t)$، برای $t \geq ۰$ گویند، هرگاه

الف) $N(۰) = ۰$

ب) برای هر $t > ۰$، $N(t)$ دارای توزیع پواسون با میانگین $m(t) = \int_{۰}^{t} \lambda(s)ds$ می‌باشد.

ج) برای هر $۰ \leq t_۱ < t_۲ < \ldots < t_m$

$$N(t_۱), N(t_۲) - N(t_۱), \ldots, N(t_m) - N(t_{m-۱})$$

مستقل از هم هستند.

در این تعریف، $m(t)$ تابع میانگین فرآیند پواسون ناهمگن است. برخی از ویژگی‌های فرآیند پواسون ناهمگن را می‌توان به صورت زیر فهرست کرد:

۱. فرآیند پواسون ناهمگن با $\lambda(t) = \lambda$ برای هر $t \geq ۰$ یک فرآیند پواسون همگن است.

۲. فرآیند پواسون ناهمگن یک فرآیند مارکوفی است.

۳. برای هر $۰ \leq s < t$، $N(t) - N(s)$ دارای توزیع پواسون با میانگین

$$m(t) - m(s) = \int_{s}^{t} \lambda(x)\mathrm{d}x$$

است. در واقع می‌توان گفت، برای هر $۰ \leq t_۱ < t_۲ < \cdots < t_m$ و هر $k_۱, \ldots, k_m \geq ۰$

$$\begin{aligned} &P\left(N(t_۱) = k_۱, N(t_۲) = k_۲, \ldots, N(t_m) = k_m\right) = \\ &P\left(N(t_۱) = k_۱, N(t_۲) - N(t_۱) = k_۲ - k_۱, \ldots\right. \\ &\left.N(t_m) - N(t_{m-۱}) = k_m - k_{m-۱}\right) \\ &= \frac{e^{-m(t_۱)}(m(t_۱))^{k_۱}}{k_۱!} \frac{e^{-(m(t_۲)-m(t_۱))}(m(t_۲) - m(t_۱))^{k_۲-k_۱}}{(k_۲ - k_۱)!} \cdots \\ &\frac{e^{-(m(t_m)-m(t_{m-۱}))}(m(t_m) - m(t_{m-۱}))^{k_m-k_{m-۱}}}{(k_m - k_{m-۱})!}. \end{aligned}$$

احتمال این‌که دقیقاً n پیشامد در فاصله $(۰,t)$ رخ بدهد، برابر است با

$$P(N(t)=n)=\frac{[m(t)]^n}{n!}e^{-m(t)},\ n=۰,۱,۲,\dots$$

که در آن $m(t)=E[N(t)]=\int_۰^t \lambda(s)\mathrm{d}s$ و $\lambda(t)$ تابع شدت است.

۴. تابع بقا $R(t)$ در فاصله زمانی $(۰,t)$ به صورت

$$R(t)=P(N(t)=۰)=e^{-m(t)}$$

تعریف می‌شود و در حالت کلی تابع بقا در فاصله $(t,t+x)$ به صورت

$$R(x|t)=P(N(t+x)-N(t)=۰)=e^{-[m(t+x)-m(t)]}$$

تعریف می‌شود.

۵. تابع چگالی در این فرآیند به صورت

$$f(x)=\lambda(t+x)e^{-[m(t+x)-m(t)]}$$

است که در آن $\lambda(x)=\dfrac{\partial}{\partial x}[m(x)]$.

۶. واریانس مدل فرآیند پواسون ناهمگن نیز به صورت زیر است:

$$var(N(t))=\int_۰^t \lambda(s)\mathrm{d}s.$$

مثال ۷.۱.۳. یک فرآیند پواسون ناهمگن با تابع شدت زیر را در نظر بگیرید:

$$\lambda(t)=\begin{cases}۵ & t\in(۱,۲],(۳,۴],\dots\\ ۳ & t\in(۰,۱],(۲,۳],\dots\end{cases}$$

برای محاسبه احتمال این‌که تعداد پیشامدها در فاصله زمانی $[۱/۲۵,۳)$، یعنی $N(۳)-N(۱/۲۵)$، بیشتر از دو باشد ابتدا باید بدانیم $N(۳)-N(۱/۲۵)$ دارای توزیع پواسون با میانگین

$$m(۳)-m(۱/۲۵)=\int_{۱/۲۵}^{۳}\lambda(t)\mathrm{d}t=\int_{۱/۲۵}^{۲}۵\mathrm{d}t+\int_{۲}^{۳}۳\mathrm{d}t=۶/۷۵$$

است. بنابراین

$$P(N(۳)-N(۱/۲۵)>۲)=۱-e^{-۶/۷۵}(۱+۶/۷۵+(۶/۷۵)^۲/۲)=۰/۹۶۴.$$

۲.۳ نمایش فرآیند پواسون ناهمگن برای مدل مخاطره متناسب کاکس

مدل مخاطره‌های متناسب کاکس روشی است که علاوه بر برآورد احتمال‌های بقا در زمان‌های مختلف، به بررسی تاثیر متغیرهای کمکی بر روی این احتمال‌ها نیز می‌پردازد. این مدل به صورت زیر بیان می‌شود

$$h_i(t|z_i) = h_\circ(t)\exp(\beta^T z_i)$$

که در آن t زمان برای هر یک از آزمودنی‌های تحت مطالعه، z_i بردار متغیرهای کمکی قابل مشاهده، $h_\circ(t)$ تابع مخاطره پایه مربوط به زمان t، و β بردار ضرایب رگرسیونی نامعلوم است که تاثیر متغیرهای کمکی را نشان می‌دهد.

همان‌طور که در فصل اول گفته شد، مدل مخاطره‌های متناسب کاکس به دو صورت پارامتری و نیمه‌پارامتری در نظر گرفته می‌شود که به دلیل انعطاف بیشتر، در این پایان‌نامه، رویکرد نیمه‌پارامتری مد نظر بوده و از مدل تکه‌ای ثابت برای مدل‌بندی تابع مخاطره پایه استفاده می‌کنیم. برای این کار، محور زمان را به J فاصله مجزا به صورت $\circ = s_\circ < s_1 < \ldots s_J = T$ تقسیم کرده و k امین فاصله را به صورت $I_k = (s_{k-1}, s_k]$ نمایش می‌دهیم. تابع مخاطره پایه را در هر فاصله، تابعی ثابت و به شکل زیر در نظر می‌گیریم:

$$h_\circ(t) = \lambda_k, \quad t \in I_k = (s_{k-1}, s_k].$$

لازم به ذکر است که بریسلو (۱۹۷۲) و بریسلو (۱۹۷۴) استفاده از توزیع نمایی تکه‌ای را برای مدل‌بندی تابع مخاطره پایه پیشنهاد کردند. هولفورد (۱۹۸۰) و لیرد و الیور (۱۹۸۱) در کارهای خود، به‌طور مستقل، به معادل بودن مدل مخاطره تکه‌ای با یک مدل رگرسیون پواسون اشاره کردند.

به‌طور کلی، مخاطره برای i امین آزمودنی در j امین فاصله به صورت زیر تعریف می‌شود:

$$\begin{aligned} h_i^{(j)}(t) &= h_\circ(t)exp(\beta^T z_i) \\ &= exp\{\beta^T z_i + \log h_\circ(t)\} \\ &= exp\{\beta^T z_i + \log \lambda_j.\} \end{aligned}$$

فرض کنید $\eta_i^{(j)} = \beta^T z_i + \log \lambda_j$ و $\delta_i^{(j)}$ تابع نشانگر شکست i امین آزمودنی در j امین فاصله باشد (۱ برای شکست و $\circ$ برای سانسور). همچنین فرض کنید t_i زمان بقای i امین آزمودنی برای $i = 1, 2, \ldots, n$ باشد. در این صورت، بقای پایه برای $t \in I_k$ به صورت

$$\begin{aligned} S_{\circ i}(t) &= exp\{-\int_\circ^t h_\circ(u)\mathrm{d}u\} \\ &= exp\{-\sum_{j=1}^{k-1}(s_{j+1} - s_j)\lambda_j - (t - s_k)\lambda_k\} \end{aligned}$$

تعریف می‌شود. بنابراین، لگاریتم تابع درستنمایی برای i امین مشاهده با زمان t، $t \in I_k$، به صورت

$$\log[h_i(t)^{\delta_i} S_i(t)] = \delta_i^{(k)} \eta_i^{(k)} - \left\{ \sum_{j=1}^{k-1} (s_{j+1} - s_j) exp(\eta_i^{(j)}) + (t - s_k) exp(\eta_i^{(j)}) \right\}$$

$$= \delta_i^{(j)} \eta_i^{(k)} - (t - s_k) exp(\eta_k) - \sum_{j=1}^{k-1} (s_{j+1} - s_j) exp(\eta_i^{(j)})$$

قابل بازنویسی است، که در آن عبارت

$$\delta_i^{(j)} \eta_i^{(k)} - (t - s_k) exp(\eta_i^{(k)})$$

همان لگاریتم تابع درستنمایی پواسون با میانگین $(t - s_k) exp(\eta_k)$ برای مشاهده ۰ یا ۱، بر اساس تابع نشانگر $\delta_i^{(j)}$، است. همچنین

$$- \sum_{j=1}^{k-1} (s_{j+1} - s_j) exp(\eta_i^{(j)})$$

تابع درستنمایی $k-1$ متغیر پواسون با میانگین $(s_{j+1} - s_j) exp(\eta_i^{(j)})$ است، زمانی که برای همه آن‌ها ۰ مشاهده شده است. برای فهم بهتر به مثال زیر توجه کنید.

مثال ۱.۲.۳. داده‌های مورد استفاده در این مثال متشکل از زمان بقای مربوط به ۹۰ بیمار مبتلا به سرطان معده است که به دو گروه درمانی تقسیم شده‌اند. یک گروه شامل افرادی است که درمان آن‌ها ترکیبی از شیمی‌درمانی و پرتودرمانی است و گروه دیگر، شامل افرادی است که فقط بر مبنای شیمی‌درمانی تحت درمان قرار گرفته‌اند (استابلین و همکاران، ۱۹۸۱). اثر درمان، تنها متغیر کمکی استفاده شده می‌باشد. برای ایجاد مدل مخاطره با ثابت تکه‌ای، محور زمان به دو فاصله با طول مساوی تقسیم می‌شود؛ بنابراین تابع بقای پایه برای i امین آزمودنی، برای $t \in (0, 900]$ برابر

$$S_{0i}(t) = exp\{-t\lambda_1\}$$

و برای $t \in (900, 1800)$ برابر

$$S_{0i}(t) = exp\{-900\lambda_1 - (t - 900)\lambda_2\}$$

تعریف می‌شود. همچنین تابع مخاطره به صورت

$$h_i(t) = exp\{\beta_0 + trt_i\beta_1 + \log \lambda_i^{(j)}\}$$

است که β_0 عرض از مبدا، β_1 ضریب اثر درمان و $\lambda_i^{(j)}$ مخاطره پایه i امین آزمودنی در j امین فاصله، برای $i = 1, \ldots, 90$ فرد و $j = 1, 2$ زیرفاصله، است. بنابراین، تابع درستنمایی برای $0 \leq t_i \leq 900$ برابر

$$l_i = \delta_i^{(1)} \eta_i^{(1)} - t_i exp(\eta_i^{(1)})$$

و برای $t_i \geq ۹۰۰$ برابر

$$l_i = \delta_i^{(۲)}\eta_i^{(۲)} - ۹۰۰exp(\eta_i^{(۱)}) - (t_i - ۹۰۰)exp(\eta_i^{(۲)})$$

است، که در آن

$$\eta_i^{(j)} = \beta_۰ + trt_i\beta_۱ + \log\lambda_j, \quad j = ۱, ۲$$

.

در مدل مخاطره‌های متناسب کاکس، در حالت عدم وجود پیشامدهای بازگشتی، زمان‌های رخداد و سانسور حاصل از افراد مختلف، مستقل فرض می‌شوند. حال با توجه به این مطلب که در پیشامدهای بازگشتی، مشاهدات مربوط به هر فرد وابسته‌اند، پس برای تحلیل پیشامدهای بازگشتی با استفاده از مدل مخاطره‌های متناسب کاکس باید به نحوی این وابستگی در مدل وارد شود که در ادامه به بیان نحوه ورود وابستگی خواهیم پرداخت.

۱.۲.۳ ورود وابستگی پیشامدهای بازگشتی

فرض کنید $N_i(T)$ تعداد پیشامدهای رخ داده برای i امین آزمودنی در زمان T باشد و $N_{ik} = N_i(s_k) - N_i(s_{k-۱})$ تعداد پیشامدهای رخداده برای i امین آزمودنی در فاصله I_k باشند. در نتیجه، تعداد کل پیشامدهای رخداده برای i امین آزمودنی در طول زمان مطالعه، به صورت $N_i = \sum_k N_{ik}$ خواهد بود.

اگر رخدادهای فرد i ام در طول زمان $(۰, T]$ مشاهده شود، آن‌گاه فرآیند تعداد پیشامدهای فرد i ام در طول زمان مشاهده‌شده، $\{N_i(t), t \geqslant ۰\}$، یک فرآیند پواسون با تابع شدت $= h(t|z)$ $h_۰(t)exp(\beta^T z)$ است.

تحت فرآیند پواسون انتظار می‌رود که میانگین و واریانس داده‌ها با یکدیگر برابر باشند. اما در عمل معمولاً این شرایط برقرار نیست و داده‌ها دارای واریانس بزرگتری هستند. این مقدار واریانس اضافی به عامل یا عواملی نسبت داده می‌شود که سبب وابستگی داده‌ها با یکدیگر شده و پذیره مستقل بودن آن‌ها را نقض می‌کنند. در این صورت، یک متغیر غیرقابل مشاهده w_i به‌عنوان نماینده همه عوامل ناشناخته، اندازه‌گیری‌نشده یا غیرقابل اندازه‌گیری که مخصوص فرد بوده و باعث همبستگی داده‌های یک فرد می‌شوند، بایستی وارد مدل می‌گردد.

فرض می‌شود این متغیر تصادفی (یا شکنندگی[۱]) دارای اثر ضربی روی تابع شدت فرآیند است. در این حالت فرآیند $N_i(t)$ به شرط معلوم بودن z_i و w_i، یک فرآیند پواسون ناهمگن با تابع شدت زیر است:

$$\begin{aligned} h(t \mid z_i, w_i) &= h_۰(t)w_i exp(\beta^T z_i) \\ &= exp(\log(h_۰(t)) + \log(w_i) + \beta^T z_i) \\ &= exp(b_k + \alpha_i + \beta^T z_i) \end{aligned}$$

[۱] Frailty

که برای $t \in I_k = (s_{k-1}, s_k]$ برابر

$$h(t \mid z_i, w_i) = exp(b_k + \alpha_i + \beta^T z_i)$$

است، که در آن $\alpha_i = \log(w_i)$ و $b_k = \log(\lambda_k)$ می‌باشد. بنابراین $\eta_{ik} = b_k + \alpha_i + \beta^T z_i$.
تعداد پیشامدها، N_{ik}، به شرط معلوم بودن z_i و w_i دارای توزیع پواسون

$$N_{ik} \sim P(exp(\eta_{ik})(s_k - s_{k-1}))$$

است. با فرض استقلال N_{ik} و $N_{ik'}$، برای $k \neq k'$، لگاریتم درستنمایی فرد i ام به صورت

$$l_i \propto \sum_{k=1}^{K} \{E_{ik} \log(exp(\eta_{ik})(s_k - s_{k-1})) - exp(\eta_{ik})(s_k - s_{k-1})\}$$

بیان می‌شود.

۳.۳ نمایش مدل گاوسی پنهان

اگر در مدل فرآیند پواسون ناهمگن، توزیع‌های پیشین خاصی را برای پارامترهای رگرسیونی، اثر شکنندگی و تابع مخاطره پایه تکه‌ای انتخاب کنیم، مدل بیزی نتیجه‌شده عضوی از خانواده LGM خواهد شد که می‌توان از روش INLA برای برازش آن استفاده کرد.

به‌طور جزیی‌تر، برای ضرایب رگرسیونی β توزیع پیشین گاوسی را انتخاب می‌کنیم. برای لگاریتم اثر شکنندگی، یعنی α_i، نیز توزیع پیشین گاوسی را در نظر می‌گیریم. تابع مخاطره پایه در هر فاصله زمانی $(s_{k-1}, s_k]$ به صورت

$$h_0(t) = \lambda_k = exp(b_k) \quad t \in (s_{k-1}, s_k]$$

تعریف شد. برای b_k ها از فرآیند پیشین همبسته قدم زدن تصادفی مرتبه اول، rw1، استفاده می‌کنیم. در واقع، فرض می‌کنیم $(b_1, \ldots, b_J)$ دارای نموهای مرتبه اول مستقل و نرمال هستند. یعنی

$$b_{j+1} - b_j \sim N(0, \tau_b^{-1}) \qquad j = 1, \ldots J-1.$$

همچنین برای ابرپارامترهای $\theta = (\tau_\alpha, \tau_b)$ از توزیع‌های پیشین گاما با پارامترهای معلوم استفاده می‌کنیم.

با این انتخاب‌ها، میدان تصادفی پنهان

$$x = \{\alpha_1, \ldots, \alpha_n, b_1, \ldots, b_J, \beta, \eta_{11}, \ldots, \eta_{nJ}\}$$

دارای توزیع نرمال چندمتغیره با بردار میانگین صفر و ماتریس دقت $Q(\theta)$ است. بنابراین مدل مخاطره متناسب کاکس با حضور اثر شکنندگی برای لحاظ کردن وابستگی بین پیشامدهای بازگشتی که به صورت یک فرآیند پواسون ناهمگن نمایش داده شد، عضوی از رده مدل‌های گاوسی پنهان محسوب می‌شود. به عبارت دیگر، تابع درستنمایی تعداد وقایعی که برای فرد i در فاصله k رخ می‌دهند، به میدان گاوسی پنهان x از طریق پیشگوی η_{ik} وابسته است.

فصل ۴

ارزیابی عملکرد روش INLA در تحلیل داده‌های بازگشتی

به منظور ارزیابی استنباط بیزی تقریبی پیشنهادی با روش INLA در تحلیل داده‌های پیشامدهای بازگشتی، در این فصل دو مثال شبیه‌سازی را ارایه می‌دهیم. همچنین دو مجموعه داده واقعی در مورد داده‌های تومور پستان موش‌های آزمایشگاهی و داده‌های ایست قلبی را با روش پیشنهادی مورد تحلیل قرار می‌دهیم. نتایج حاصل از این مثال‌ها، کارایی آماری و محاسباتی روش استنباط پیشنهادی را اثبات می‌کنند.

تمامی محاسبات توسط بسته R-INLA در نرم‌افزار R انجام شده‌اند. مطالب نظری مرتبط و راهنمای استفاده از توابع این بسته به همراه مثال‌های مختلف، در وب‌سایت www.r-inla.org بارگذاری شده‌اند. همه محاسبات نیز بر روی یک رایانه دو هسته‌ای 2.53GHz اجرا شده‌اند. کدهای R لازم برای بازتولید مثال واقعی و شبیه‌سازی اول در پیوست آ داده شده است.

۱.۴ مطالعه شبیه‌سازی

در این بخش، دو مثال شبیه‌سازی را ارایه می‌کنیم. در مثال اول، داده‌ها از یک فرآیند پواسون ناهمگن با یک تابع مخاطره (شدت) تکه‌ای، شبیه‌سازی شدند و در مثال دوم، تابع مخاطره پایه واقعی یک مدل پارامتری وایبل در نظر گرفته شد تا عملکرد روش برازش پیشنهادی (یک روش نیمه‌پارامتری) در حالتی که تابع مخاطره پایه واقعی، پارامتری است نیز مورد بررسی قرار گیرد.

۱.۱.۴ مثال اول: فرآیند پواسون ناهمگن

در این مثال، پیشامدهای بازگشتی را برای ۲۰۰ نفر از یک فرآیند پواسون ناهمگن با رهیافت نمونه‌گیری تصادفی پیشنهادی در کتاب راس (۲۰۰۲)، در بازه زمانی $(۰, ۱۰]$، تولید کردیم. تابع مخاطره واقعی (که در فرآیندهای پواسون به تابع شدت معروف است) به صورت

$$h(t) = ۳\cos^۲(t) \tag{۱.۴}$$

در نظر گرفته شده است. نحوه شبیه‌سازی پیشامدهای غیربازگشتی با این تابع شدت در مثالی در کتاب ریزو (۲۰۰۸) تشریح شده است. با دقت در تابع (۱.۴) و مقایسه با مدل مخاطره متناسب کاکس، به‌راحتی می‌توان دریافت که تابع مخاطره پایه $h_۰(t) = \cos^۲(t)$ است و متغیر تبیینی وجود ندارد. به عبارت دیگر متغیر تبیینی برای همه افراد، ثابت است. نمونه‌ای تصادفی از پیشامدهای شبیه‌سازی‌شده برای ۱۰ نفر در شکل ۱.۴ نمایش داده شده است.

هدف از این مثال شبیه‌سازی، ارزیابی کارایی برآورد تابع مخاطره پایه با استفاده از رهیافت پیشنهادی در این پایان‌نامه است. برای این منظور، ابتدا بازه زمانی $(۰, ۱۰]$ را به K زیرفاصله تقسیم کردیم و چهار حالت مختلف $K = ۲۰, ۴۰, ۶۰, ۸۰$ را در نظر گرفتیم. سپس تعداد پیشامدهای E_{ik} را برای $i = ۱, \ldots, ۲۰۰$ و $k = ۱, \ldots, K$، تحت مدل فرآیند پواسون ناهمگن پیشنهادی در فصل سوم مدل‌بندی کردیم. در مدل پیشنهادی فرآیند پیشین قدم زدن تصادفی مرتبه اول را برای لگاریتم تابع مخاطره پایه در نظر گرفتیم. همچنین تمام پیشن‌های انتخاب‌شده، توزیع‌های پیشین پیش‌فرض در بسته R-INLA هستند.

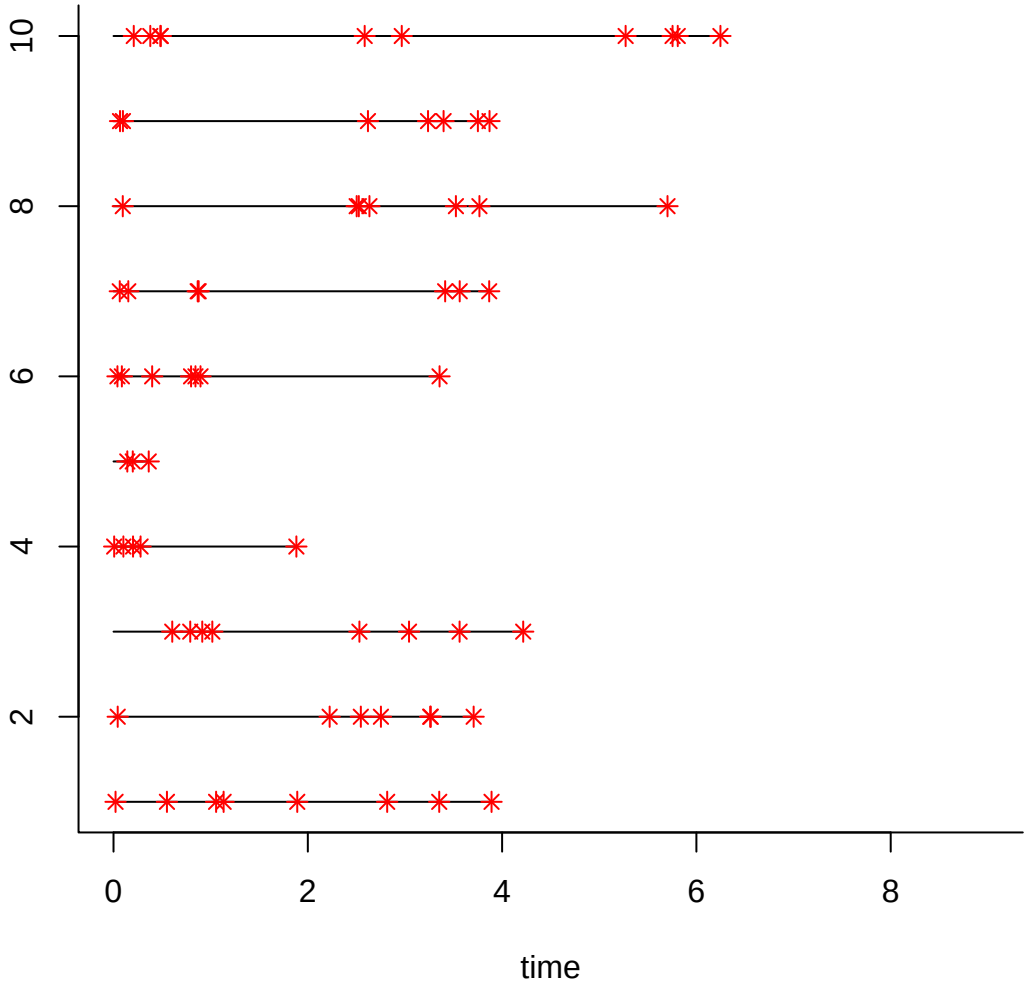
منحنی واقعی لگاریتم تابع مخاطره و برآورد بیزی (میانگین توزیع پسین) آن برای مقادیر مختلف K در شکل ۲.۴ نشان داده شده است. همان‌طور که مشهود است، تابع مخاطره پایه واقعی با تابع پایه تکه‌ای برآوردشده حاصل از روش INLA به‌طور قابل قبولی بازیابی شده است. افزون بر این، با افزایش تعداد زیرفاصله‌ها، یعنی K، تقریب حاصل بهتر هم شده است. البته دقت در زیرفاصله‌های ابتدایی خیلی بالاتر از زیرفاصله‌های انتهایی بازه تحت مطالعه است. دلیل آن هم وجود مشاهدات (پیشامدهای بازگشتی) بیشتر در این فواصل است. بنابراین در زیرفاصله‌هایی که تعداد پیشامدها به اندازه کافی باشد، دقت تقریب حاصل از روش پیشنهادی برای تابع مخاطره پایه کاملا خوب است. همچنین تفاوت چندانی در تقریب‌های حاصل از $K = ۶۰$ و $K = ۸۰$ نیست.

۲.۱.۴ مثال دوم: مدل مخاطره متناسب پارامتری

در این مثال، برای ارزیابی عملکرد برآورد ضرایب رگرسیونی و تابع مخاطره پایه، پیشامدهای بازگشتی را از یک مدل مخاطره متناسب با یک تابع مخاطره پایه وایبل، شبیه‌سازی کردیم.

فاصله زمانی تحت مطالعه را $(۰, ۲]$ و تعداد افراد را $n = ۳۰$ در نظر گرفتیم. پیشامدهای بازگشتی در فاصله تحت مطالعه را برای فرد i ام از مدل

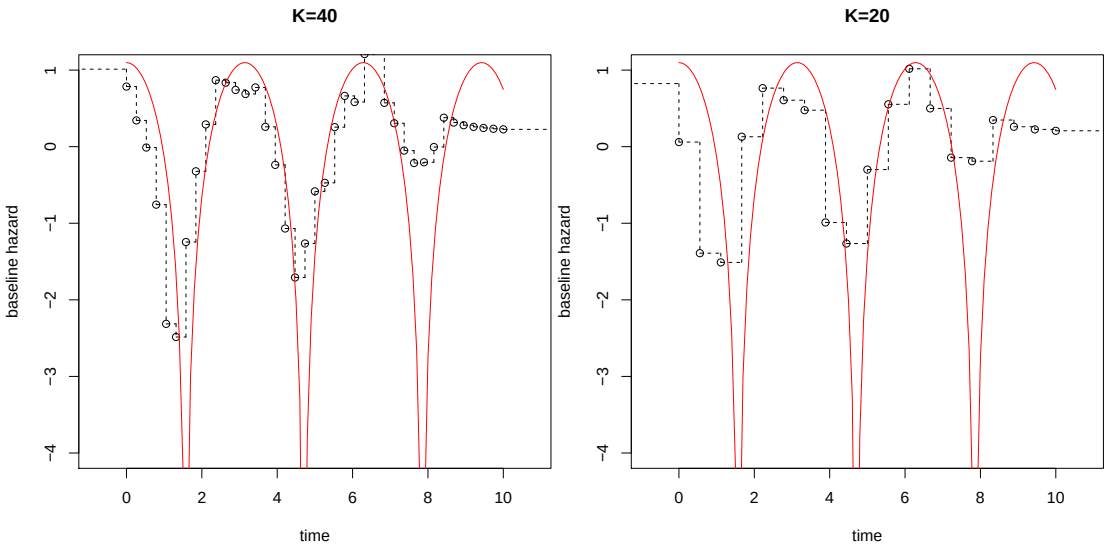
$$h_i(t) = h_۰(t)\exp(\beta_۱ x_{۱i} + \beta_۲ x_{۲i} + \log z_i) \tag{۲.۴}$$



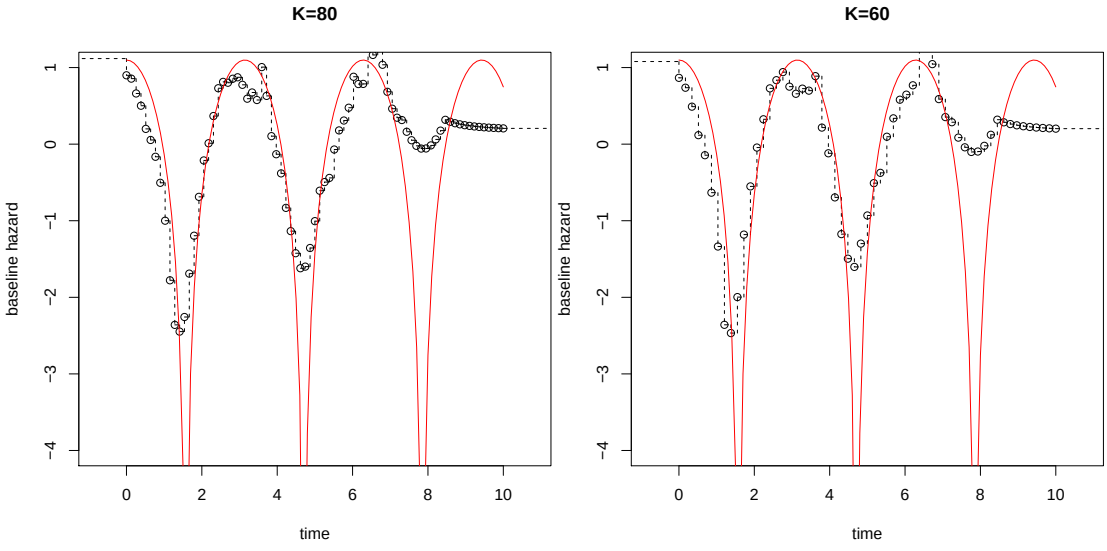
شکل ۱.۴: پیشامدهای بازگشتی شبیه‌سازی‌شده برای یک نمونه ۱۰ تایی. علامت‌های ستاره متناظر با زمان رخداد پیشامدها هستند.

تولید کردیم، که در آن

- مقادیر واقعی پارامترها برابر $(\beta_1, \beta_2) = (-0/7, 1/5)$ انتخاب شدند.
- هر دو متغیر تبیینی x_1 و x_2 از توزیع نرمال استاندارد تولید شدند.
- متغیر شکنندگی z_i از توزیع نرمال با میانگین صفر و واریانس $\sigma^2 = 0/5$ تولید شد. این متغیر برای در نظر گرفتن اثر بازگشتی بودن پیشامدهای فرد i ام به مدل افزوده می‌شود.



(الف) (ب)



(ج) (د)

شکل ۲.۴: برآورد تابع مخاطره پایه در مثال شبیه‌سازی اول: لگاریتم تابع مخاطره پایه واقعی (خط ممتد) به همراه برآورد بیزی حاصل از میانگین توزیع پسین حاشیه‌ای تابع مخاطره پایه (خط چین)، برای تعداد زیرفاصله‌های الف) $K = ۲۰$، ب) $K = ۴۰$، ج) $K = ۶۰$، و د) $K = ۸۰$.

- تابع مخاطره پایه، مدل پارامتری وایبل با پارامتر مقیاس $s = ۱$ و پارامتر شکل $\nu = ۳$ و ضابطه

$$h_۰(t) = s\nu t^{\nu-۱}$$

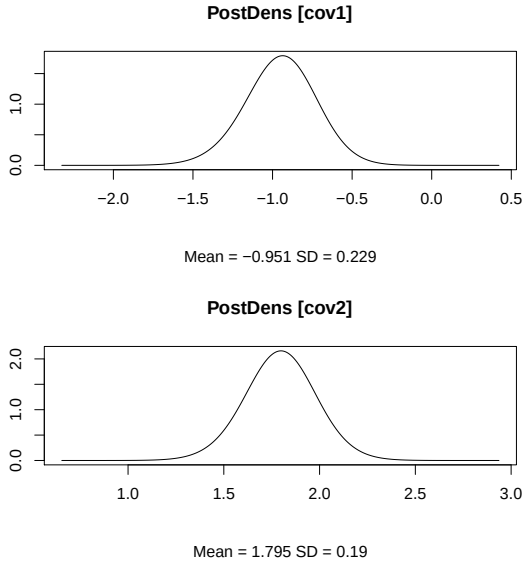
در نظر گرفته شد.

مجموع داده‌های شبیه‌سازی‌شده برای ۳۰ نفر نیز برابر ۲۳۱۲ به دست آمد.

با داده‌های شبیه‌سازی حاصل، مدل پیشنهادی را با تقسیم فاصله $[۰, ۲)$ به $K = ۵۰$ قسمت، و در نظر گرفتن توزیع پیشین قدم زدن تصادفی با پارامتر دقت پیش‌فرض تابع inla در بسته R-INLA برای تابع مخاطره پایه، برازش دادیم. نتایج برازش در جدول ۱.۴ گزارش شده‌اند. علی‌رغم آن‌که مدل واقعی دارای تابع مخاطره پایه وایبل است و تعداد افراد ۳۰ در نظر گرفته شده است، برآوردهای بیزی به مقادیر واقعی نزدیک هستند و فواصل اعتبار بیزی HPD[۱] نیز سازگاری برآوردها را به شکلی نمایش می‌دهند. البته برآورد پارامتر وابستگی متغیر شکنندگی، اریبی قابل ملاحظه‌ای دارد که با افزایش حجم نمونه n این اریبی قابل کاهش است.

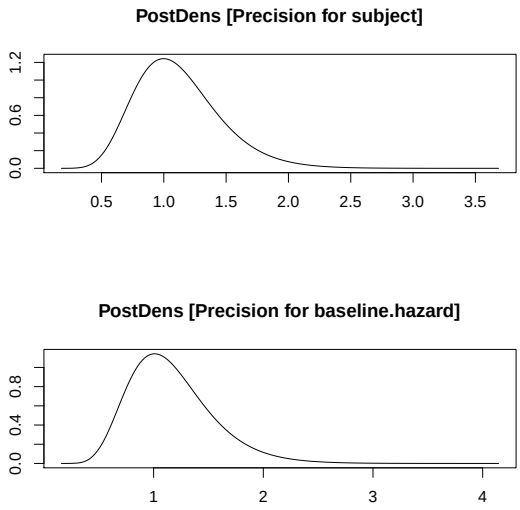
جدول ۱.۴: نتایج برآورد مدل (۲.۴) شامل میانگین توزیع پسین و فاصله اعتبار HPD ۹۵٪

پارامتر	مقدار واقعی	برآورد	فاصله اعتبار HPD
$\beta_۱$	$-۰/۷$	$-۰/۹۵$	$(-۰/۵۱, -۱/۴۰)$
$\beta_۲$	۱/۵	۱/۷۹	$(۱/۴۲, ۲/۱۷)$
$\sigma^۲$	۰/۵	۰/۹۸	$(۰/۴۶, ۱/۶۱)$



شکل ۳.۴: نمودار چگالی پسین حاشیه‌ای پارامترهای رگرسیونی

[۱] Highest posterior density



شکل ۴.۴: نمودار چگالی پسین حاشیه‌ای پارامترهای دقت متغیر شکنندگی و پیشین قدم زدن تصادفی برای تابع مخاطره پایه

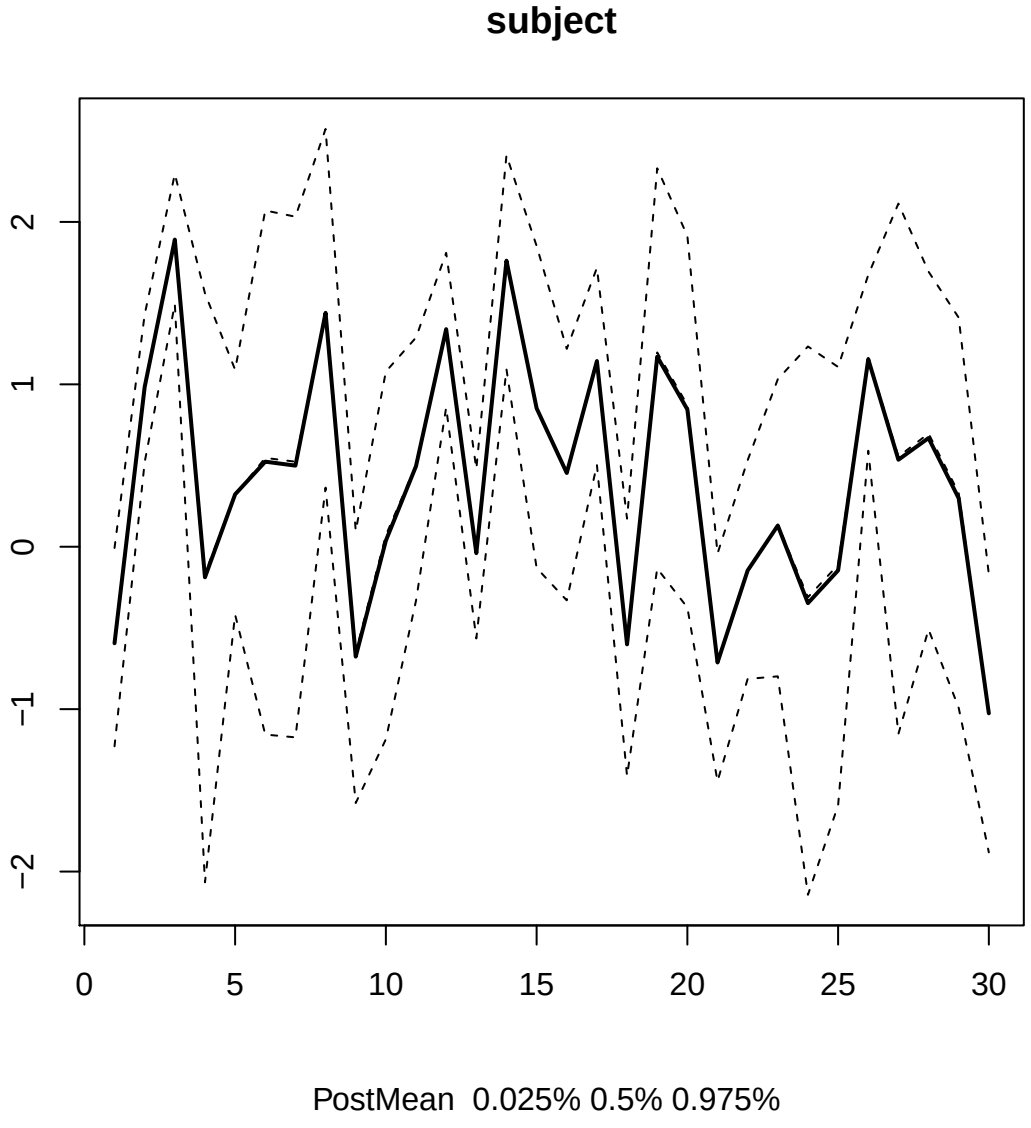
شکل‌های ۳.۴ و ۴.۴ توابع چگالی پسین حاشیه‌ای تقریبی پارامترهای رگرسیونی مدل و پارامترهای وابستگی را نشان می‌دهند که حاوی تمام اطلاعات پسین حاشیه‌ای هستند. عدم قطعیت موجود در برآوردهای بیزی از روی این نمودارهای چگالی، بهتر قابل مشاهده هستند. شکل ۵.۴ نیز برآورد لگاریتم مولفه شکنندگی را نشان می‌دهد که حاکی از ناهمپراشی قوی در پیشامدهای مشاهده‌شده بین افراد است. به عبارت دیگر، حضور مولفه شکنندگی کاملا معنی‌دار است و بین پیشامدهای بازگشتی افراد، وابستگی قوی وجود دارد.

شکل ۶.۴ نیز تابع مخاطره پایه واقعی (وایبل) را به همراه برآورد بیزی متناظر حاصل از برازش INLA که بر اساس میانگین توزیع پسین حاشیه‌ای به‌دست آمده است، نشان می‌دهد. کران‌های اعتبار بیزی ۹۵٪ نیز در شکل نمایش داده شده‌اند. همان‌طور که ملاحظه می‌کنید، تقریب حاصل از مدل فرآیند پواسون ناهمگن حاصل از روش INLA به منحنی واقعی نزدیک است و در کران‌های اعتبار بیزی قرار دارد. بنابراین عملکرد مناسب روش پیشنهادی در این مثال شبیه‌سازی در هر دو مقوله برآورد پارامترهای مدل و تابع مخاطره پایه، مورد تایید قرار می‌گیرد.

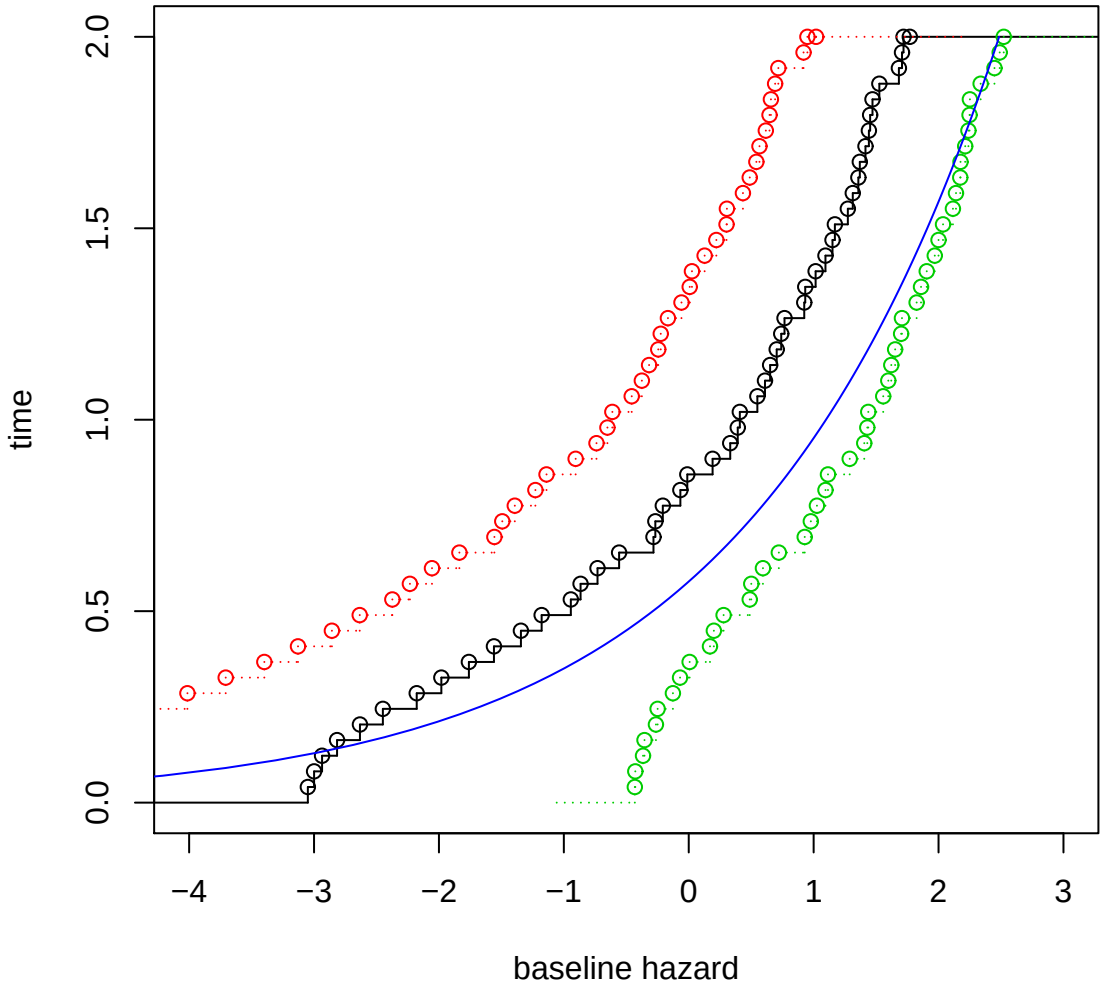
زمان لازم برای برازش مدل با روش INLA در این مثال، نزدیک به ۷ ثانیه است!!!

۲.۴ داده‌های تومور پستان

در این بخش، داده‌های تومور پستان تعدادی از موش‌های آزمایشگاهی را که توسط گیل (۱۹۸۰) گزارش و تحلیل شده‌اند، با روش بیزی تقریبی پیشنهادی مورد تحلیل قرار داده‌ایم. پیشامد مورد نظر در مورد این داده‌ها، زمان پیشرفت سرطان پستان در موش‌هاست. تعداد ۴۸ موش آزمایشگاهی در این مطالعه



شکل ۵.۴: برآورد بیزی (میانگین توزیع پسین) لگاریتم متغیر شکنندگی با فاصله اعتبار ۹۵٪



شکل ۶.۴: منحنی تابع مخاطره پایه وایبل واقعی (منحنی ممتد)، به همراه برآورد بیزی (میانگین پسین) لگاریتم تابع مخاطره پایه (منحنی تکه‌ای ممتد) و کران‌های اعتبار بیزی ۹۵٪ (محنی تکه‌ای نقطه‌ای).

مورد بررسی قرار گرفته‌اند. در این مطالعه ابتدا یک ماده سرطان‌زا به همه موش‌ها تزریق شده و سپس به‌طور تصادفی به دو گروه تقسیم شدند که یک گروه تحت درمان با دارو قرار گرفتند و گروه دوم، گروه کنترل را تشکیل دادند. از روز ۶۲ تا روز ۱۸۲، پیدایش تومورها هفته‌ای دو بار بررسی شد. همه موش‌ها نیز بعد از روز ۱۸۲ از راست سانسور شدند (به این معنی که پایان مطالعه در روز ۱۸۲ ثبت شد).

شکل ۷.۴ زمان‌های رخداد تومور را برای همه ۴۸ موش مورد مطالعه، نشان می‌دهد. زمان رخدادها بر حسب روز می‌باشد. تنها متغیر تبیینی مورد استفاده در این مطالعه، ”گروه“ است که $group = ۱$ گروه تحت درمان و $group = ۲$ گروه کنترل را بیان می‌کند.

برای اجرای روش مدل فرآیند پواسون ناهمگن پیشنهادی، ابتدا محور زمان را به ۵ زیرفاصله با طول برابر تقسیم کردیم. در واقع تعداد پیشامدهای E_{ik} برای $i = ۱, \ldots, ۴۸$ و $k = ۱, \ldots, ۵$، با مدل پیشنهادی در فصل سوم مدل‌بندی شدند. بنابراین، تابع مخاطره (شدت) متناسب شرطی برای موش i ام در زیرفاصله k ام، عبارتست از

$$h(t|z_i, w_i) = \exp\{\beta_۰ + \beta_۱ group_i + b_k + \alpha_i\},\ t \in I_k$$

به‌طوری که مشابه مباحث فصل سوم، $s_k - s_{k-۱}$ برای زیرفاصله‌های زمانی یکسان ثابت است.

برای تکمیل مدل، توزیع‌های پیشین زیر را انتخاب کردیم:

- با قرار دادن $\lambda = (\lambda_۱, \ldots, \lambda_K)$، برای $\log(\lambda) = b$ یک فرآیند قدم زدن تصادفی مرتبه اول با پارامتر دقت τ_b تنظیم شد.
- برای هر دو پارامتر $\beta_۰$ و $\beta_۱$ توزیع پیشین نرمال با میانگین صفر و واریانس ۱۰۰۰ انتخاب شد. این توزیع پیشین یک نوع توزیع پیشین مبهم[۲] محسوب می‌شود که با توجه به واریانس خیلی بزرگ آن تقریبا ناآگاهی‌بخش[۳] محسوب می‌شود و در عین حال یک توزیع پیشین سره[۴] است.
- برای $\alpha_i = \log(w_i)$ یک توزیع نرمال با میانگین صفر و پارامتر دقت τ_α انتخاب شد.
- برای ابرپارامترهای τ_b و τ_α نیز توزیع گاما $\Gamma(۱, ۰/۰۰۱)$ در نظر گرفته شد، به‌طوری که منظور از $\Gamma(q, r)$ توزیع گاما با میانگین q/r و واریانس $q/r^۲$ است.

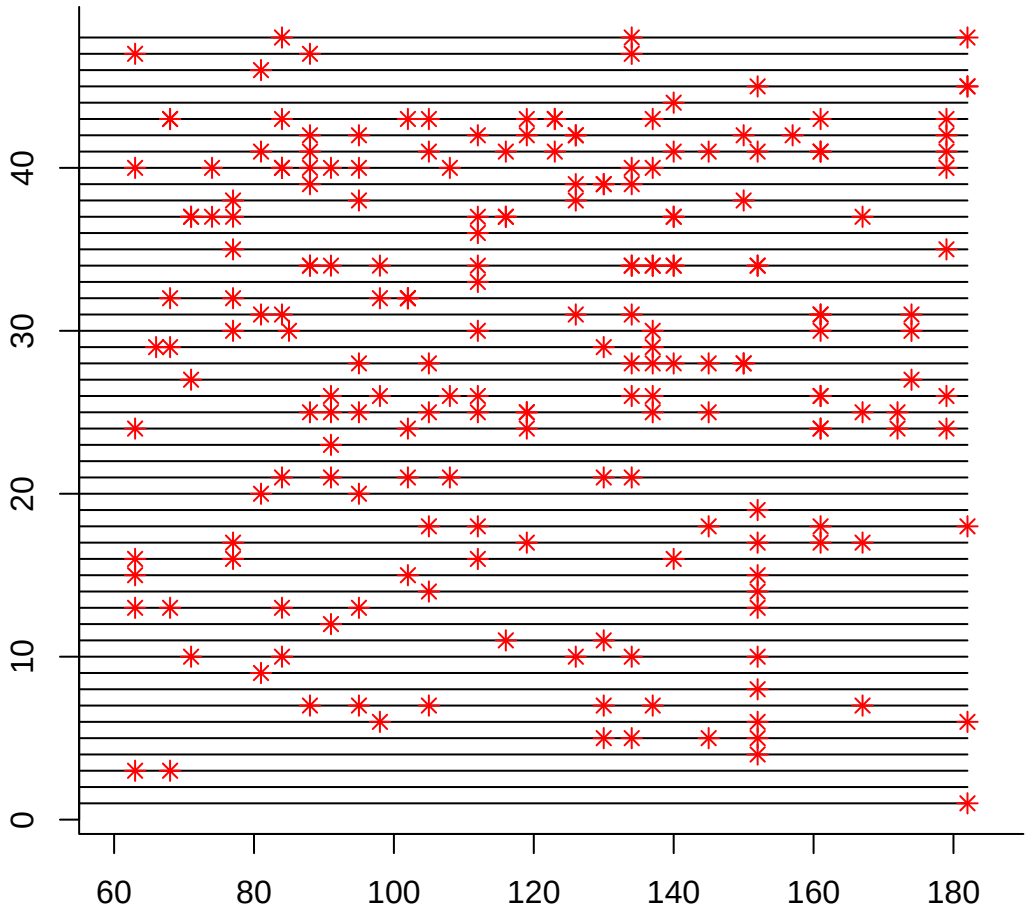
پس از تکمیل مدل‌بندی، آن را به داده‌ها برازش دادیم. مدت زمان برازش مدل با روش INLA نزدیک به ۱۷ ثانیه طول کشید که با توجه به حجم داده‌ها، یعنی ۲۵۳، در مقابل با روش‌های MCMC بسیار سریع‌تر است.

برآوردهای لگاریتم تابع مخاطره پایه و لگاریتم شکنندگی در شکل ۸.۴ نمایش داده شده‌اند. همان‌طور که از روی شکل مشهود است، لگاریتم تابع مخاطره پایه تقریبا در طول بازه زمانی تحت مطالعه ثابت است. در مقابل اثر برآوردشده مولفه شکنندگی، شهودی قوی از ناهمپراشی بین موش‌ها را نشان

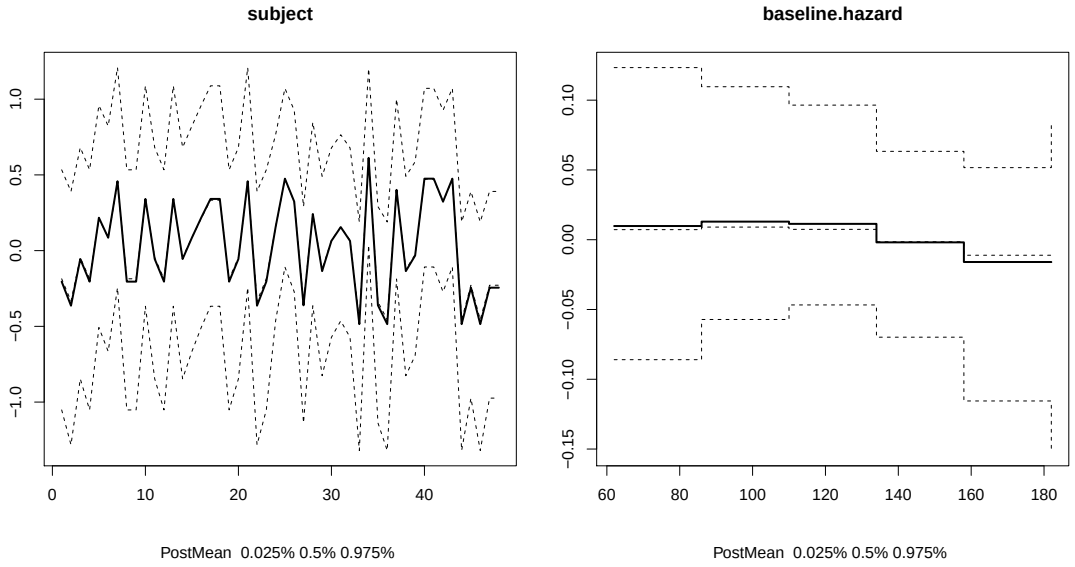
[۲] Vague prior

[۳] Noninformative

[۴] Proper



شکل ۷.۴: زمان‌های رخداد تومور برای ۴۸ موش آزمایشگاهی: علامت ستاره حاکی از زمان رخداد تومور است



(ب) (الف)

شکل ۸.۴: برآورد میانگین پسین لگاریتم تابع مخاطره پایه (الف) و لگاریتم مولفه شکنندگی (ب)

می‌دهد. این ناهمپراشی به آن معنی است که بعضی از موش‌ها نسبت به سایر موش‌ها در گروه خود، بیشتر از خطر ابتلا به تومور برخوردار هستند و رنج می‌برند.

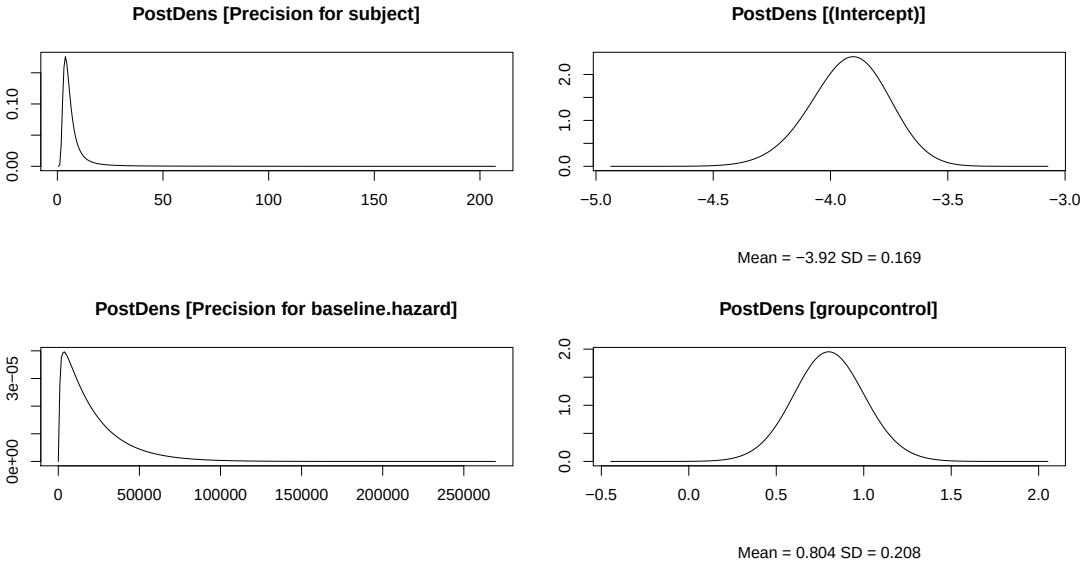
میانگین و انحراف معیار پسین برآوردشده برای اثر درمان با دارو (اثر متغیر گروه) به ترتیب برابر ۰٫۸ و ۰٫۲ هستند. فاصله اعتبار HPD ۹۵٪ برای این اثر (۰٫۳۹۷, ۱٫۲۱) به دست آمد. این نتیجه، معنی‌داری اثر درمان را بیان می‌کند. به این معنی که درمان با دارو در کاهش مخاطره رخداد تومور موثر است. این نتایج با نتایج منتشرشده در سینها (۱۹۹۳) مطابقت دارد.

شکل ۹.۴ نیز توابع چگالی پسین حاشیه‌ای برآوردشده پارامترهای رگرسیونی β_0 و β_1 و پارامترهای دقت مولفه شکنندگی τ_α و توزیع پیشین قدم زدن تصادفی برای لگاریتم تابع مخاطره پایه τ_b را که با روش INLA برآورد شده‌اند، نمایش می‌دهد.

۱.۲.۴ حضور اثر شکنندگی

برای بررسی اهمیت حضور مولفه شکنندگی در مدل، برای لحاظ کردن وابستگی بین پیشامدهای بازگشتی هر فرد، مدل را یک بار بدون اثر شکنندگی به داده‌ها برازش دادیم. میانگین و انحراف معیار برآوردشده اثر درمان با دارو به ترتیب برابر ۰٫۸۱ و ۰٫۱۵ به دست آمدند. همچنین فاصله اعتبار HPD ۹۵٪ برابر (۰٫۵۲, ۱٫۱۱) نتیجه شد. نتایج حاصل باز هم معنی‌داری اثر درمان را نشان می‌دهند. اما به‌وضوح انحراف معیار برآوردشده یک کم‌برآوردی قابل ملاحظه را به همراه دارد. بنابراین، در نظر نگرفتن اثر شکنندگی در مدل باعث القای اریبی قابل ملاحظه در برآورد دقت اثر درمان می‌شود.

برای مقایسه این دو مدل بر اساس معیارهای انتخاب مدل، مقدار DIC را برای هر دو مدل محاسبه کردیم. مقدار این معیار برای مدل با اثر شکنندگی برابر ۵۹۰٫۳۰ و برای مدل بدون اثر شکنندگی برابر



(ب) (الف)

شکل ۹.۴: توابع چگالی پسین حاشیه‌ای برآوردشده با روش INLA برای پارامترهای رگرسیونی (الف) و پارامترهای دقت مولفه شکنندگی و پیشین لگاریتم تابع مخاطره پایه (ب)

۶۱۴٫۴۸ حاصل شد. نتیجه مقایسه این معیار برای دو مدل نیز تاییدی بر برتری مدل با حضور اثر شکنندگی است.

۳.۴ داده‌های ایست قلبی

هدف اصلی در این مثال، ارزیابی اثر شرایط آب و هوایی بر روی ایست‌های قلبی ثبت‌شده برای مردم ساکن شهر تروندهیم[۵] کشور نروژ است. این شهر در میانه کشور نروژ در شمال اروپا واقع شده است و دارای آب و هوایی سرد و بارانی و همچنین شب‌های طولانی در زمستان و جوی معتدل و روزهای طولانی در تابستان است.

داده‌های مورد نظر شامل رخداد ایست‌های قلبی (به‌عنوان پیشامد مورد علاقه)، طی دوره زمانی ژانویه ۱۹۹۰ تا نوامبر ۱۹۹۸، است که توسط خدمات درمانی اورژانسی مورد درمان قرار گرفته‌اند (کوالوی و اسکوگول، ۲۰۰۷). مجموعه داده‌ها، شامل ۳۲۵۶ روز و ۸۰۹ ایست قلبی ثبت‌شده طی این دوره زمانی می‌باشد. جزییات تعداد پیشامدها به ازای هر روز در جدول ۲.۴ گزارش شده‌اند. در این داده‌ها، تعداد پیشامدهای ثبت‌شده کوچک هستند و تقریبا در ۷۸٪ روزها هیچ پیشامدی ثبت نشده است.

کوالوی و اسکوگول (۲۰۰۷) متغیرهای تبیینی مختلفی را در نظر گرفتند. اما ما در این مثال، تنها متغیرهای موثری را که آن‌ها گزارش کردند، در نظر می‌گیریم. به‌طور مشخص، متغیرهای تبیینی مورد

[۵] Trondheim

جدول ۲.۴: تعداد پیشامدها به ازای هر روز

ایست قلبی	۰	۱	۲	۳
تعداد روزها	۲۵۳۶	۶۳۶	۷۹	۵

نظر ما عبارتند از: دمای هوا (temp)، رطوبت نسبی هوا (humidity)، سرعت وزش باد (windspeed)، میزان بارش (precipitation)، بارش برف (snow)، که یک متغیر نشانگر با مقادیر ۱ به معنی بارش برف و ۰ به معنی سایر بارش‌ها است، و شماره روز (day). متغیر شماره روز، شماره روز در یک سال را بیان می‌کند. متغیرهای تبیینی دمای هوا، رطوبت نسبی، و سرعت باد، چندین بار طی یک روز اندازه‌گیری شده‌اند و ما از میانگین روزانه این اندازه‌ها برای مقادیر این متغیرها استفاده کرده‌ایم.

مطابق با پیشنهاد کوالوی و اسکوگول (۲۰۰۷)، فاصله زمانی تحت مطالعه را ۲۴ ساعت در نظر گرفتیم. زیرا منطقی است که رخداد ایست قلبی تحت تاثیر این دوره زمانی ۲۴ ساعته، $[۰, ۲۴]$، باشد. در این مثال، هر آزمودنی (مشابه با هر موش در مثال قبلی) یک روز است. بنابراین ۳۲۵۶ آزمودنی داریم و هدف سنجش اثر متغیرهای تبیینی (شرایط آب و هوایی) بر تعداد ایست‌های قلبی در یک روز است.

برای این داده‌ها، فرض کنید E_{ik} بیانگر تعداد ایست‌های قلبی در روز i ام در زیرفاصله k ام باشد، به‌طوری که فاصله زمانی ۲۴ ساعت یک روز را به $K = ۱۲$ زیرفاصله مساوی تقسیم کردیم. بنابراین، پیشگوی خطی مدل برای روز i ام، $i = ۱, \ldots, ۳۲۵۶$، در زیرفاصله k ام، $k = ۱, \ldots, ۱۲$، به صورت زیر است:

$$\begin{aligned} \eta_{ik} &= \beta_۰ + \beta_۱ windspeed_i + \beta_۲ snow_i + f^{(temp)}(temp_i) \\ &+ f^{(precipitation)}(precipitation_i) + f^{(humidity)}(humidity_i) \\ &+ f^{(day)}(day_i) + b_k. \end{aligned} \tag{۳.۴}$$

در مدل (۳.۴)، مشابه مثال‌های قبلی، لگاریتم تابع مخاطره (شدت) پایه در زیرفاصله k ام به صورت $b_k = \log(\lambda_k)$ تعریف می‌شود. در این مدل، اثر متغیرهای سرعت وزش باد و بارش برف را خطی در نظر گرفتیم و فرض کردیم سایر متغیرهای تبیینی و لگاریتم تابع مخاطره پایه، اثری هموار و غیرخطی دارند. به‌طور خاص، یک توزیع پیشین قدم زدن تصادفی مرتبه اول RW1 برای لگاریتم تابع مخاطره پایه و توزیع‌های پیشین قدم زدن تصادفی مرتبه دوم RW2 برای متغیرهای دمای هوا، میزان بارش، رطوبت نسبی و شماره روز، در نظر گرفتیم (فصل سوم کتاب رو و هلد (۲۰۰۵) را برای جزییات این فرآیندها ببینید). علاوه بر این، چون داده‌ها به ترتیب زمانی، طی تقریبا ۹ سال متوالی، گردآوری و در تحلیل وارد شده‌اند، متغیر شماره روز به‌عنوان یک متغیر دوره‌ای[۶] تعریف شد. توزیع پیشین برای همه ابرپارامترهای مدل نیز گاما با پارامترهای معلوم انتخاب شد.

برآوردهای بیزی و فاصله اعتبار بیزی ۹۵٪ اثرات خطی دو متغیر سرعت وزش باد و بارش برف، در جدول ۳.۴ گزارش شده‌اند. با توجه به فواصل اعتبار بیزی برآوردشده برای این دو متغیر، می‌توان عدم معنی‌داری هر دو را نتیجه گرفت.

[۶] Cyclic

جدول ۳.۴: برآوردهای بیزی (میانگین پسین) و فواصل اعتبار بیزی اثرات خطی متغیرهای سرعت وزش باد و بارشِ برف

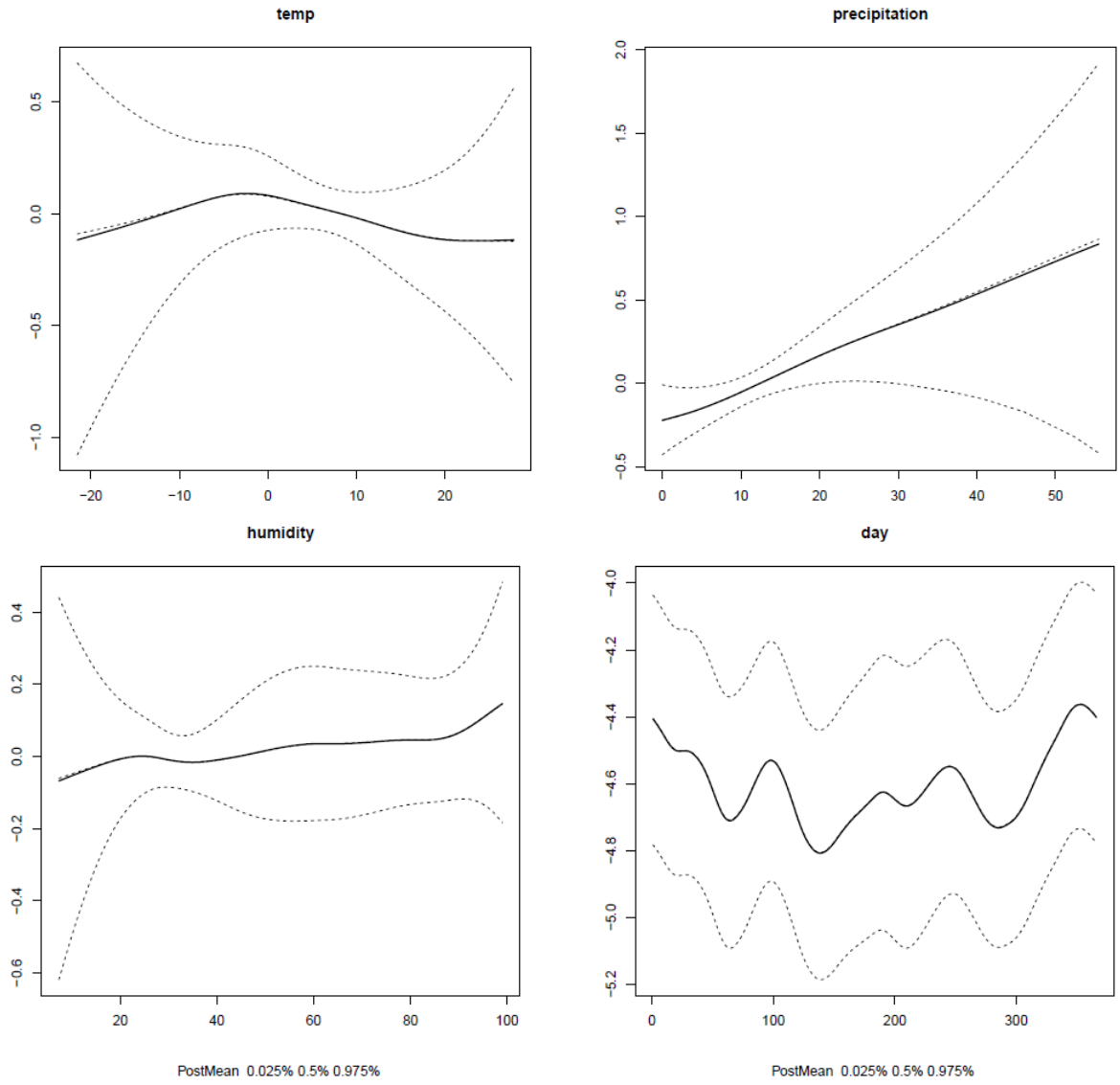
متغیر تبیینی	میانگین پسین	انحراف معیار	فاصله اعتبار ۹۵٪
windspeed	−۰/۰۰۵	۰/۰۱۵	(−۰/۰۳, ۰/۰۲)
snow	۰/۱۶۲	۰/۱۰۵	(−۰/۰۴, ۰/۳۷)

برآوردهای اثرات متغیرهای دمای هوا، میزان بارش، رطوبت نسبی و شماره روز نیز در شکل ۱۰.۴ گزارش شده‌اند. از روی برآوردهای گزارش‌شده در این شکل، می‌توان نتایج زیر را خلاصه کرد:

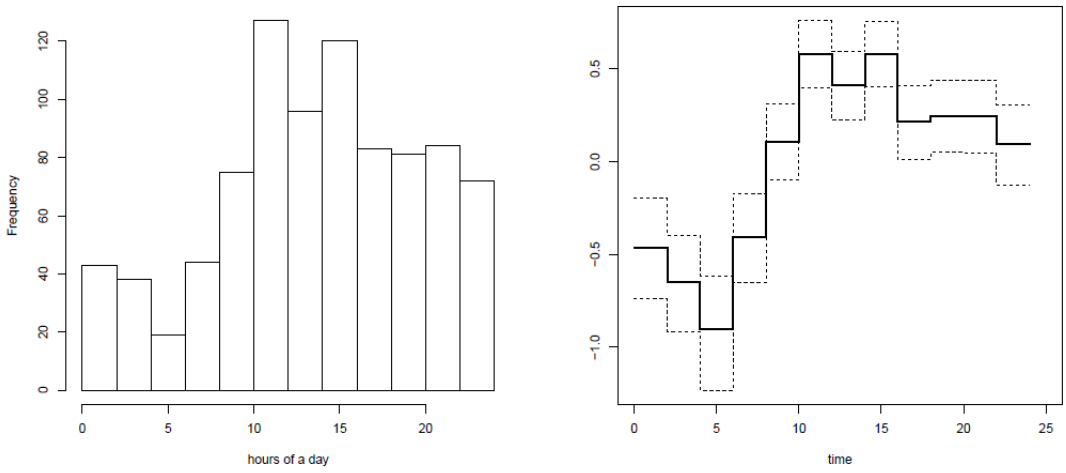
- زمانی‌که دمای هوا کمی پایین‌تر از صفر است، شدت (تعداد) ایست‌های قلبی ماکسیمم می‌شود و با افزایش دما این شدت کاهش می‌یابد.
- اثر میزان بارش خطی است و هر چه میزان بارش بیشتر می‌شود، شدت ایست قلبی هم افزایش می‌یابد.
- به نظر می‌رسد اثر رطوبت نسبی هوا نسبت به زمان ثابت است.
- متغیر شماره روز برای در نظر گرفتن تغییرات دوره‌ای (فصلی) وارد مدل شده است. اثر این متغیر در سرتاسر سال در حال تغییر است، اما در فصل زمستان ماکسیمم و در فصل بهار می‌نیمم می‌شود.
- از روی کران‌های اعتبار بیزی در این شکل، واضح است که عدم قطعیت برآوردها در کران‌های فاصله زمانی ۲۴ ساعت که تعداد مشاهدات پیشامدها در آن زمان‌ها کم هستند، بزرگ می‌شود.

نمودار سمت چپ شکل ۱۱.۴ هیستوگرام زمان‌های واقعی ایست‌های قلبی را طی ۲۴ ساعت نشان می‌دهد. این تعداد ایست‌های قلبی، کمترین تعداد (حدود ۵ تا) را در صبح نشان می‌دهد و با افزایش سطح فعالیت روزانه افراد تا حدود ساعت ۴ بعد از ظهر، افزایش می‌یابد. سپس کاهش می‌یابد و پایدار می‌شود. برآورد تابع لگاریتم مخاطره پایه همراه با کران‌های اعتبار ۹۵٪ نیز در نمودار سمت راست شکل ۱۱.۴ نمایش داده شده است.

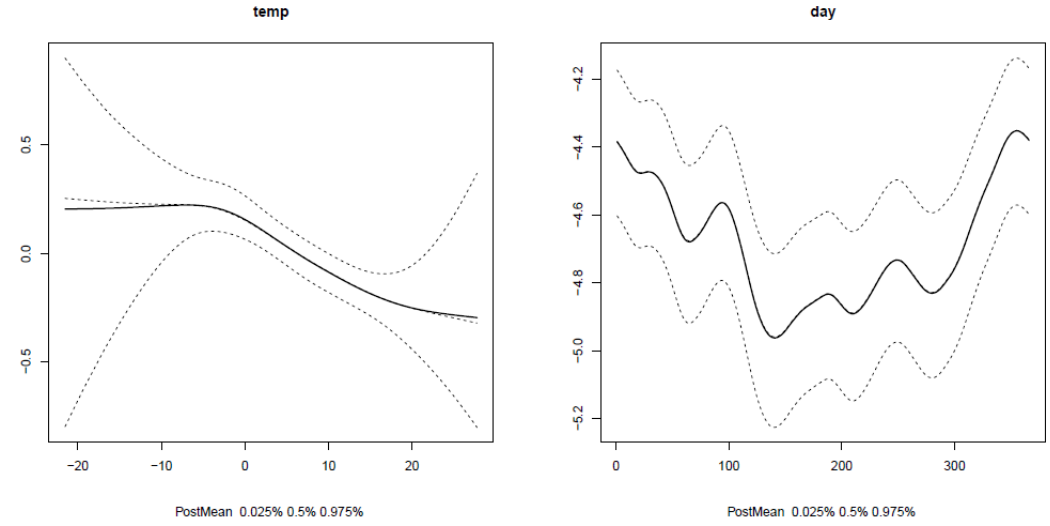
از شکل ۱۰.۴ واضح است که رخداد پیشامد ایست قلبی تحت تاثیر هر دو متغیر دمای هوا و شماره روز است. تعداد رخدادهای ایست قلبی در زمستان یا زمانی که دمای هوا منفی است، بیشتر است. از آن‌جایی که وابستگی بین دمای هوا و شماره روز قوی است، بررسی اثر این دو متغیر به‌طور جداگانه می‌تواند مفید باشد. بنابراین برای مدل‌بندی تعداد ایست‌های قلبی، دو مدل متفاوت در نظر گرفتیم: یکی با حضور فقط شماره روز و دیگری با حضور فقط دمای هوا. برای هر دو متغیر نیز اثری غیرخطی و هموار (با یک پیشین RW2) در نظر گرفتیم. برآوردهای بیزی اثرات آن‌ها به همراه کران‌های اعتبار ۹۵٪ آن‌ها در شکل ۱۲.۴ گزارش شده‌اند. معنی‌داری اثر هر دو متغیر از روی برآوردهای این شکل، مشهود است. زمانی که دمای هوا زیر صفر درجه سانتیگراد است، شدت ایست‌های قلبی ثابت ولی در بالاترین حد خود است. اما با افزایش دما به شدت کاهش می‌یابد. از طرف دیگر، اثر شماره روز در



شکل ۱۰.۴: برآوردهای میانگین پسین اثر (ناپارامتری) متغیرهای دمای هوا temp، میزان بارش precipitation، رطوبت نسبی humidity، و شماره روز day به همراه کران‌های اعتبار ۹۵٪



شکل ۱۱.۴: هیستوگرام زمان‌های واقعی ایست‌های قلبی (شکل سمت چپ) و برآورد بیزی (میانگین توزیع پسین) و کران اعتبار ۹۵٪ لگاریتم تابع مخاطره پایه (شکل سمت راست)



شکل ۱۲.۴: برآوردهای بیزی (میانگین پسین) و کران‌های اعتبار ۹۵٪ اثر متغیرهای دما و شماره روز وقتی به‌طور جدا در مدل وارد شوند

سراسر سال متغیر است. اما می‌توان گفت که شدت رخداد ایست‌های قلبی طی زمستان زیاد و در بهار کمترین است. دلیل آن هم می‌تواند شرایط آب و هوایی بد فصل زمستان در این شهر باشد.

۱.۳.۴ معنی‌داری متغیرها: انتخاب مدل

برای سنجش معنی‌داری همه متغیرها، چند مدل دیگر را نیز بررسی کردیم. مشابه قبل برای متغیرهای سرعت وزش باد و بارش برف، اثر خطی و برای سایر متغیرها اثری هموار با فرآیند RW2 در نظر گرفتیم. افزون بر این، برای متغیر شماره روز یک اثر دوره‌ای فرض کردیم.

برای مقایسه مدل‌ها نیز از معیار انتخاب مدل بیزی DIC استفاده کردیم. این مقایسه برای مدل‌هایی که متغیرهای تبیینی اشاره‌شده در آن‌ها حضور داشته‌اند، در جدول ۴.۴ گزارش شده‌اند.

جدول ۴.۴: معیار DIC مدل‌های مختلف با متغیرهای تبیینی مختلف برای داده‌های ایست قلبی

مدل	متغیرهای تبیینی حاضر	DIC
۱	-	۷۷۴۰٫۶
۲	temp	۷۷۲۶٫۶۲
۳	precipitation	۷۷۳۷
۴	day	۷۷۲۹٫۳
۵	temp+snow	۷۷۲۵٫۲
۶	snow+day	۷۷۲۸
۷	temp+precipitation	۷۷۲۱٫۳
۸	day+precipitation	۷۷۲۶٫۲۷
۹	temp+humidity	۷۷۳۰٫۱۵
۱۰	day+humidity	۷۷۴۸٫۹
۱۱	temp+precipitation+snow	۷۷۲۰٫۱۵
۱۲	day+precipitation+snow	۷۷۲۴٫۲۵
۱۳	temp+precipitation+humidity	۷۷۲۵٫۴۸
۱۴	day+precipitation+humidity	۷۷۵۹٫۲۵
۱۵	temp+precipitation+snow+humidity+windspeed+day	۷۷۲۶٫۶۲

بر اساس نتایج جدول ۴.۴، بهترین مدل بر حسب DIC (یعنی مدل با DIC برابر ۷۷۲۰٫۱۵) مدل شماره ۱۱ یعنی مدلی است که متغیرهای تبیینی دمای هوا، میزان بارش و بارش برف را در خود دارد. همچنین از مقایسه مدل‌های مختلف، می‌توان فهمید که مدل‌هایی که متغیر تبیینی شماره روز را دارا هستند، مقدار DIC بزرگتری نسبت به مدل‌هایی که متغیر دمای هوا را دارا بوده و سایر متغیرهایشان یکسان هستند، دارند. بنابراین، شاید بتوان مدعی شد که دمای هوا یک متغیر با معنی‌داری بیشتری نسبت به شماره روز است.

از روی نتایج مقایسه مدل‌ها، می‌توان گفت که برای مطالعه تاثیر متغیرهای آب و هوایی بر شدت ایست‌های قلبی، وارد کردن متغیرهای دمای هوا، میزان بارش، و بارش برف به مدل کافی است.

همچنین نتایج تحلیل این داده‌ها مشخص می‌کند که تعداد ایست‌های قلبی در شرایط بد جوی، افزایش می‌یابد.

۴.۴ نتیجه‌گیری و آینده تحقیق

در این پایان‌نامه، یک مدل بیزی نیمه‌پارامتری بر اساس مدل مخاطره متناسب برای مدل‌بندی پیشامدهای بازگشتی معرفی شد. به دلیل پیچیدگی محاسباتی مبتنی بر الگوریتم‌های MCMC و مشکلات سرعت و آمیختگی این روش‌ها، از رهیافت استنباط بیزی تقریبی مبتنی بر روش INLA استفاده کردیم. برای آن‌که بتوانیم از روش INLA استفاده کنیم، مدل را، به شرط متغیرهای تبیینی و اثر مولفه شکنندگی، بر اساس یک فرآیند پواسون ناهمگن که حالت خاصی از فرآیندهای گاوسی پنهان است، بازنویسی کردیم. این نمایش مدل مخاطره متناسب نیمه‌پارامتری بر اساس مدل فرآیند پواسون ناهمگن، یا به عبارتی بر اساس مدل فرآیند گاوسی پنهان، ما را قادر می‌سازد تا از تمام ویژگی‌های خوب محاسباتی روش INLA بهره ببریم.

روش محاسباتی INLA از نظر محاسباتی کارا است و نسبت به روش‌های نمونه‌گیری MCMC به شدت سریع‌تر است. این روش مشکلات همگرایی کند و آمیختگی ضعیف زنجیرهای MCMC را نیز دارا نیست، زیرا یک روش تقریبی قطعی (غیرتصادفی) محسوب می‌شود.

در رهیافت پیشنهادی، تابع مخاطره پایه بر اساس یک تابع تکه‌ای ثابت مدل‌بندی می‌شود و از نسخه‌های پارامتری استفاده نشده است. اما در مطالعه شبیه‌سازی نشان دادیم که حتی زمانی که تابع مخاطره پایه واقعی یک تابع پارامتری مثل وایبل است، روش برازش پیشنهادی خوب و قابل قبول عمل می‌کند و خطای تقریب قابل صرفنظر کردن است. سایر مثال‌های شبیه‌سازی و واقعی، عملکرد خوب و کارای روش پیشنهادی را نشان می‌دهند. بنابراین، استفاده از آن را در مدل‌بندی داده‌های پیشامدهای بازگشتی را توصیه می‌کنیم.

با توجه به مطالب مطرح‌شده در این پایان‌نامه، برای آینده تحقیق می‌توانیم موارد زیر را برشمریم:

۱. در این پایان‌نامه به مقوله استنباط بیزی پیشامدهای بازگشتی پرداختیم. زمینه‌های استنباطی که به آن‌ها پرداخته شد، برآورد نقطه‌ای و فاصله‌ای و معیارهای انتخاب مدل (آزمون فرضیه) را شامل می‌شد. پیش‌گویی زمان رخداد پیشامد بعدی در حوزه تحلیل پیشامدهای بازگشتی، بحث پرطرفداری است که از اهمیت ویژه‌ای برخوردار است. اجرای این پیش‌گویی با روش پیشنهادی می‌تواند جالب توجه آماردان‌ها و محققین مختلفی باشد.

۲. در مدل نیمه‌پارامتری مذکور، پذیره تابع مخاطره پایه، یک تابع تکه‌ای ثابت (در زیرفاصله‌ها) است. استفاده از سایر رهیافت‌های معمول برای مدل‌بندی ناپارامتری این تابع، مانند اسپلاین‌ها، و برازش مدل حاصل با روش INLA نیز موضوع مهم دیگری برای تحقیقات آینده است.

۳. امکان مدل‌بندی و تحلیل پیشامدهای چندگانه (پیشامدهای چندمتغیره برای هر فرد) با رهیافتی مشابه، موضوع بسیار جالبی است که بررسی آن از موضوعات مورد نظر آینده است.

مراجع

[1] Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. The Annals of Applied Probability, 951-972.

[2] Aalen, O. O. (1972). Estimering av Risikorater for Prevensjonsmidlet 'Spiralen'. Norwegian.) Graduate thesis in statistics. Institute of Mathematics, University of Oslo.

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

[4] Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. Mathematical Biosciences, 6, 1-11.

[5] Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. The Annals of Statistics, 10, 1100-1120.

[6] Asmussen, S., Rolski, T. (1994). Risk theory in a periodic environment: the Cramer-Lundberg approximation and Lundberg's inequality. Mathematics of Operations Research, 19(2), 410-433.

[7] Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. J. R. Statist. Soc. B, 70, 825–848.

[8] Banerjee, M., Wellner, J. A. (2005). Confidence intervals for current status data. Scandinavian Journal of Statistics, 405-424.

[9] Beard, R.E. (1959) Note on some mathematical mortality models. In: G.E.W. Wolstenholme and M.O'Conner (eds.). The Lifespan of Animals. Ciba Foundation Colloquium on Ageing. Little, Brown, Boston, 302 - 311

[10] Berg, M. P., Haberman, S. (1994). Trend analysis and prediction procedures for time non-homogeneous claim processes. Insurance: Mathematics and Economics, 14(1), 19-32.

[11] Birkhoff, G.D. (1942), What is the ergodic theorem?, American Mathematical Monthly, 49(4), 222-226.

[12] Breslow, N. (1974). Covariance analysis of censored survival data. Biometrics, 30:89–99.

[13] Breslow, N. (1972). Discussion on regression models and life-tables (by d. r. cox). Journal of the Royal Statistical Society, Series B, 34:216–217.

[14] Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika, 57(3), 579-594.

[15] Brooks, S., Gelman, A., Jones, G., and Meng, X. (2011), Handbook of Markov Chain Monte Carlo, CRC Press, Taylor Francis Group.

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

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

[18] Chib, S. and Greenberg, E. (1995), Understanding the Metropolis–Hastings algorithm, The American Statistician, 49, 327–335.

[19] Chu, W. and Z. Ghahramani (2005). Gaussian processes for ordinal regression. Journal of Machine Learning Research 6, 1019–1041.

[20] Chukova, S., Dimitrov, B., Garrido, J. (1993). Renewal and nonhomogeneous Poisson processes generated by distributions with periodic failure rate. Statistics probability letters, 17(1), 19-25.

[21] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika, 65(1), 141-151.

[22] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society, Series B, 34:187–220.

[23] Cook, R. J., Lawless, J. F. (2002). Analysis of repeated events. Statistical Methods in Medical Research, 11(2), 141-166.

[24] Dempster, M. A. H. (1974). An application of quantile arithmetic to the distribution problem of stochastic linear programming. Bulletin of the Institute of Mathematics and Its Applications, 10, 186-194.

[25] Efron, B. (1967). The two sample problem with censored data. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 4, No. University of California Press, Berkeley, CA, pp. 831-853).

[26] Fahrmeir, L. Tutz, G. (2001). Multivariate statistical modelling based on generalized linear models, 2nd edn. Springer-Verlag, Berlin.

[27] Gail, M. H., Santner, T. J., and Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times of tumor. Biometrics, 36:255–266.

[28] Garrido, J., Dimitrov, B. and Chukova, S. (1996) modelling for compound nonstationary processes with periodic claim intensity rate", Technical Report No. 2/96, Concordia University, Montreal.

[29] Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika, 52, 203-223.

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

[31] Gelman, A., Meng, X. L. (Eds.). (2004). Applied Bayesian modeling and causal inference from incomplete-data perspectives. John Wiley Sons.

[32] Gelman, A. and Rubin, D. (1992), Inference from Iterative Simulation using Multiple Sequences, Statistical Science, 7, 457–511.

[33] 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.

[34] Geweke, J. (1992), Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments, In Bayesian Statistics 4, Clarendon Press, Oxford, UK.

[35] Geyer, C.J. (1992), Practical Markov chain Monte Carlo, Statistical Science, 7(4), 473–483.

[36] Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation.Journal of the American Statistical Association, 102, 359-378.

[37] Green, W. D. K. (1996). The thin-plate spline and images with curving features. Image fusion and shape variability, 79-87.

[38] Green, P., Silverman, B. W. (1994). The penalized likelihood approach.

[39] Greenwood, M., Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. Journal of the Royal Statistical Society, 83(2), 255-279.

[40] Groeneboom, P. and Wellner, J. A. (2001). Computing Chernoff's distribution. Journal of Computational Graphical Statistics, 10, 388-400.

[41] Hallinan, A. J. (1993). A review of the Weibull distribution. Journal of Quality Technology, 25, 85-85.

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

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

[44] Hoem, J. M. (1969). Fertility rates and reproduction rates in a probabilistic setting. Biométrie-Praximétrie, 10(1), 38-66.

[45] Holford, T. (1980). The analysis of rates and of survivorship using log-linear models. Biometrics, 36:299–305.

[46] Hsiao, C.K., Huang, S.Y., and Chang, C.W. (2004), Bayesian marginal inference via cadidate's formula, Statistics and Computing, 14(1):59–66.

[47] Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. The Annals of Statistics, 24(2), 540-568.

[48] Johnson, R. A., Haskell, J. H. (1983). Sampling properties of estimators of a Weibull distribution of use in the lumber industry. Canadian Journal of Statistics, 11(2), 155-169.

[49] Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457-481.

[50] Knorr-Held, L. and Rue, H. (2002), On block updating in Markov random field models for disease mapping, shape Scandinavian Journal of Statistics, 29(4), 597- 614.

[51] Kvaløy, J. T. and Skogvoll, E. (2007). Modelling seasonal and weather dependency of cardiac arrests using the covariate method. Statistics in Medicine, 26:3315 – 3329.

[52] Laird, N. and Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. Journal of the American Statistical Association, 76:231–240.

[53] Lange, K. (2010), Numerical Analysis for Statisticians, Springer.

[54] Lang, S. and A. Brezger (2004). Bayesian P-splines. Journal of Computational and Graphical Statistics 13 (1).

[55] Lawless, J. F. and Nadeau, J. C. (1995). Some simple robust methods for the analysis of recurrent events. Technometrics, 37, 158-168.

[56] Li, Q. H., Lagakos, S. W. (1997). Use of the Wei–Lin–Weissfeld method for the analysis of a recurring and a terminating event. Statistics in Medicine, 16(8), 925-940.

[57] Lin, D. Y., Oakes, D., Ying, Z. (1998). Additive hazards regression with current status data. Biometrika, 85(2), 289-298.

[58] Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. Journal of Royal Statistical Society Ser B, 62, 711-730.

[59] Mardia, K. V., Kent, J. T., Bibby, J. M. (1980). Multivariate analysis.

[60] Martino, S. (2007). Approximate Bayesian inference for latent Gaussian models. Ph. D. thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology.

[61] Martino, S. and Rue, H. (2010), Implementing approximate Bayesian inference using integrated nested Laplace approximation: a manual for the inla program, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway. Compiled on April 8, 2010.

[62] McCulluagh, P. and Nelder, J. A. (1989). Generalized linear models. Chapman and Hall, London.

[63] McCulloch, C. E., Searle, S. R. (2001). Generalized, Linear, and Mixed Models (Wiley Series in Probability and Statistics).

[64] 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.

[65] Murthy, D. P., Bulmer, M., Eccleston, J. A. (2004). Weibull model selection for reliability modelling. Reliability Engineering System Safety, 86(3), 257-267.

[66] Murray, W. A. (1972). Numerical methods for unconstrained optimization.

[67] Nelson, W. (1969). Hazard plotting for incomplete failure data. Journal of Quality Technology, 1(1), 27-52.

[68] Nelson, W. B. (2003). Recurrent events data analysis for product repairs, disease recurrences, and other applications. ASA-SIAM Series on Statistics and Applied Probability, 10.

[69] Oakes, D. (1992). Frailty models for multiple event times. in Survival Analysis: State of the Art, eds. J.P. Klein and P. K. Goel, pages 371–379.

[70] Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981).On the regression analysis of multivariate failure time data', Biometrika, 68, 373-379 .

[71] Pettit, L. I., Young, K. D. S. (1990). Measuring the effect of observations on Bayes factors. Biometrika, 77(3), 455-466.

[72] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007), Numerical Recipes. The Art of Scientific Computing, Cambridge University Press.

[73] Raftery, A. E., Lewis, S. M. (1992). [Practical Markov Chain Monte Carlo]: comment: one long run with diagnostics: implementation strategies for Markov Chain Monte Carlo. Statistical Science, 7(4), 493-497.

[74] Richardson, B. A., Hughes, J. P. (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. Biostatistics, 1(3), 341-354.

[75] Rizzo, M. (2008). Statistical computing with R. Chapman Hall.

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

[77] Robert, C. and Casella, G. (2011), A short history of Markov chainMonte Carlo: Subjective recollections from incomplete data, Statistical Science, 26, 102–115.

[78] Ross, S. M. (2002). Simulation. Academic Press, California.

[79] Rue, H. and Held, L. (2005). Gaussian Markov Random Fields: Theory and Applications, volume 104 of Monographs on Statistics and Applied Probability. Chapman Hall, London.

[80] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. Journal of the Royal Statistical Society, Series B, 71(2):319–392.

[81] Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. Journal of Statistical Planning and Inference, 137(10):3177–3192. Special Issue: Bayesian Inference for Stochastic Processes.

[82] Scolnik, H. D., Gambini, J. M. (2001). A new method to compute second derivatives. Journal of Computer Science Technology, 1.

[83] Sinha, D. (1993). Semiparametric bayesian analysis of multiple event time data. Journal of the American Statistical Association, 88:979–983.

[84] Smith, A. F. M., Skene, A. M., Shaw, J. E. H. and Naylor, J. C. (1987) Progress with numerical and graphical methods for practical Bayesian statistics. Statistician, 36, 75–82.

[85] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and der Linde, A. (2002), Bayesian measures of model complexity and fit (with discussion), Journal of the Royal Statistical Society (series B), 64, 583–639.

[86] Stablein, D. M., Carter,W. H., and Novak, J.W. (1981). Analysis of survival data with non-proportional hazard functions. Controlled Clinical Trials, 2:149–159.

[87] Sverdrup, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. Scandinavian Actuarial Journal, 1965(3-4), 184-211.

[88] Takahashi, K., Fagan, J., and Chen, M.S. (1973). Formation of a sparse bus impedance matrix and its application to short circuit study. In 8th PICA Conference proceedings, pages 63–69.

IEEE Power Engineering Society. Papers presented at the 1973 Power Industry Computer Application Conference in Minneapolis, Minnesota.

[89] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, 81(393):82–86.

[90] Tjalling J. Ypma, Historical development of the Newton-Raphson method, SIAM Review 37 (4), 531–551, 1995.

[91] Vaupel, J. M., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography, 16:439–454.

[92] Wei, L. J., Lin, D. Y., Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association, 84(408), 1065-1073.

[93] West, M. and Harrison J., (1997), Bayesian Forecasting and Dynamic Models (2nd ed.), New York: Springer-Verlag.

[94] White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica, 50, 1-25.

[95] Yoon, J. W., Wilson, S. P. The Efficient Gaussian Approximation for a class of Latent Gaussian model.

پیوست آ

کدهای R

آ.۱ اجرای روش

بسته R- INLA در R ، برای استنباط تقریبی بیزی به وسیله روش INLA به‌کار می‌رود (مارتینو و رو، ۲۰۱۰). این بسته، برنامه مستقل INLA را در بسته GMRFLib جایگزین می‌کند (مارتینو و رو، ۲۰۱۰) و می‌توان این بسته را با نوشتن دستور زیر در R دانلود و نصب کرد

```
> source("http://www.math.ntnu.no/inla/givemeINLA.R")
```

این بسته را می‌توان به صورت زیر فراخوانی کرد

```
> library(INLA)
```

دستور

```
> inla.upgrade(testing=TRUE)
```

برای به‌روزرسانی این بسته به آخرین نسخه به‌کار می‌رود. برای برنامه‌نویسی با R-INLA ، فرض کنید مثالی با دو متغیر تبیینی $x = (x_1, x_2)$ و تابع $f(.)$ برای متغیر تبیینی سوم z_1 داشته باشیم، پیشگوی خطی از طریق دستور زیر ایجاد می‌شود:

```
> formula <- y ~1 + x1 + x2 + f(z1, model="...")
```

که در آن y ، x۱، x۲ و z۱ ستون‌های ساختار داده‌ای هستند که داده‌های مورد نظر را در خود دارند. عبارت ۱ اجباری نیست و به‌عنوان عرض از مبدا می‌باشد. در تابع $f()$ ، "..."=model نوع تابع $f(.)$ را مشخص می‌کند. پیش‌فرض آن "iid" = model است. لیست مدل‌های دیگر به‌صورت زیر است

```
> names(inla.models()$latent)
[1]   "linear"     "iid"          "mec"         "meb"              "rgeneric"
[6]   "rw1"        "rw2"          "crw2"         "seasonal"          "besag"
[11] "besag2"    "bym"         "bym2"        "besagproper"    "besagproper2"
[16] "ar1"        "ar"           "ou"          "generic"          "generic0"
[21] "generic1"  "generic2"   "generic3"    "spde"              "spde2"
[26] "spde3"      "iid1d"       "iid2d"        "iid3d"             "iid4d"
[31] "iid5d"      "2diid"       "z"            "rw2d"             "rw2diid"
[36] "slm"         "matern2d" "copy"         "clinear"           "sigm"
[41] "revsigm"
```

در نهایت الگوریتم INLA به‌وسیله تابع inla به‌صورت زیر اجرا می‌شود:

```
> inla(formula, family = "...", data)
```

که در آن family توزیع داده‌ها (درستنمایی) را مشخص می‌کند. برخی از توزیع‌های موجود

```
>  names(inla.models()$likelihood) #output partially omitted
[1] "poisson"                                    "gpoisson"
[3] "binomial"                                   "testbinomial1"
[5] "gamma"                                     "beta"
[7] "betabinomial"                              "cbinomial"
[9] "nbinomial"                                  "simplex"
[11] "gaussian"                                  "normal"
[13] "circularnormal"                           "wrappedcauchy"
[15] "iidgamma"                                 "iidlogitbeta"
[17] "sas"                                           "loggammafrailty"
[19] "logistic"                                      "skewnormal"
[21] "sn"                                             "gev"
[23] "laplace"                                       "lognormal"
[25] "exponential"                                 "coxph"
[27] "weibull"                                       "loglogistic"
[29] "zeroinflatednbinomial0"
```

هستند.

در مورد داده‌های بازگشتی، به‌عنوان مثال فرض کنید پیشگوی خطی را به‌صورت زیر داشته باشیم

$$\eta_{ij} = \beta_{۰} + \beta_{sex} sex_i + \beta_{age} age_i + \beta_{dis۲} dis۲_i + \beta_{dis۳} dis۳_i + \beta_{dis۴} dis۴_i + \log(u_i)$$

در این صورت پیشگوی خطی از طریق دستور زیر ایجاد می‌شود:

```
>formula = inla.surv(time, event, subject)~ age + sex + dis2 + dis3 + dis4 +
f (subject, model="iid", param=c(1, 1))
```

و در نهایت الگوریتم INLA به‌وسیله تابع inla به‌صورت زیر اجرا می‌شود:

```
>mod=inla(formula, family="coxph", data=data,
control.data=list(param=c(1,1)), control.fixed=list(prec=1))
```

که در آن تابع "inla.surv()" برای توصیف داده سانسور در مدل‌های تحلیل بقا استفاده می‌شود. "param" پارامترهای پیشین‌های گاما را مشخص می‌کند و "prec" دقت پیشین بردار β را تعیین می‌کند.

۲.آ کدهای مربوط به مثال تومور پستان

```
>rm(list=ls())

>library(INLA)

## data entry
>time=c(182,182,63,68,182,152,182,130,134,145,152,182,98,152,182,88,95,
105,130,137,167,182,152,182,81,182,71,84,126,134,152,182,116,130,
182,91,182,63,68,84,95,152,182,105,152,182,63,102,152,182,63,77,
112,140,182,77,119,152,161,167,182,105,112,145,161,182,152,182,81,
95,182,84,91,102,108,130,134,182,182,91,182,63,102,119,161,161,
172,179,182,88,91,95,105,112,119,119,137,145,167,172,182,91,98,
108,112,134,137,161,161,179,182,71,174,182,95,105,134,137,140,145,
150,150,182,66,68,130,137,182,77,85,112,137,161,174,182,81,84,126,
134,161,161,174,182,68,77,98,102,102,102,182,112,182,88,88,91,98,112,
134,134,137,137,140,140,152,152,182,77,179,182,112,182,71,71,74,77,112,
116,116,140,140,167,182,77,95,126,150,182,88,126,130,130,134,182,63,74,
84,84,88,91,95,108,134,137,179,182,81,88,105,116,123,140,145,152,161,
161,179,182,88,95,112,119,126,126,150,157,179,182,68,68,84,102,105,119,
123,123,137,161,179,182,140,182,152,182,182,81,182,63,88,134,182,84,134,
182)
>event=c(1,0,1,1,0,1,0,1,1,1,1,0,1,1,1,1,1,1,1,1,1,0,1,0,1,0,1,1,1,1,1,0,1,1,0,1,
0,1,1,1,1,1,0,1,1,0,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,0,1,1,
1,1,1,1,0,0,1,0,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,
1,0,1,1,0,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,0,1,1,
1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,0,1,1,1,1,0,1,
1,1)
>subject=c(1,2,3,3,3,4,4,5,5,5,5,5,6,6,6,7,7,7,7,7,7,7,8,8,9,9,10,10,10,10,10,
```

```
10,11,11,11,12,12,13,13,13,13,13,13,14,14,14,15,15,15,15,16,16,16,16,
16,17,17,17,17,17,17,18,18,18,18,18,19,19,20,20,20,21,21,21,21,21,21,
21,22,23,23,24,24,24,24,24,24,24,24,25,25,25,25,25,25,25,25,25,25,25,
25,26,26,26,26,26,26,26,26,26,26,27,27,27,28,28,28,28,28,28,28,28,28,
29,29,29,29,29,30,30,30,30,30,30,30,31,31,31,31,31,31,31,31,32,32,32,
32,32,32,32,33,33,34,34,34,34,34,34,34,34,34,34,34,34,34,34,34,35,35,35,
36,36,37,37,37,37,37,37,37,37,37,37,37,38,38,38,38,38,39,39,39,39,39,
39,40,40,40,40,40,40,40,40,40,40,40,40,41,41,41,41,41,41,41,41,41,41,
41,41,42,42,42,42,42,42,42,42,42,42,43,43,43,43,43,43,43,43,43,43,43,
43,44,44,45,45,45,46,46,47,47,47,47,48,48,48)
>group=c(rep(1,81),rep(2,172))
>rats.data <- as.data.frame(cbind(time,event,subject,group))
>rats.data$group <- factor(rats.data$group,labels=c('treatment','control'))
>cutpoints = seq( 62,182, len=6)
>formula = inla.surv(time, event, subject=subject) ~ group +
f(subject, model="iid",param=c(1,0.001))
>model = inla(formula,family="coxph",control.hazard=list(cutpoints=cutpoints),
control.inla = list(int.strategy="grid", diff.logdens=15, dz=0.2),
data=rats.data)
>h = inla.hyperpar(model, dz=0.2, diff.logdens=15)

>sigma.subj <- inla.tmarginal(function(x) sqrt(1/x), h$marginals.hy[[1]])
#sigma.basline <- inla.tmarginal(function(x) 1/exp(x), h$marginals.hy[[2]])

>inla.emarginal(function(x) x, sigma.subj)
>inla.qmarginal(c(0.025, 0.5, 0.975), sigma.subj)
>inla.hpdmarginal(0.95, sigma.subj)
```

آ.۳ کدهای مربوط به مثال اول شبیه‌سازی

```
>rm(list=ls())

>set.seed(2321)

>lambda <- 3
>upper <- 4

>T.rec <- list()
>m <- 200

>for (i in 1:m){
N <- rpois(1, lambda * upper)
```

```
Tn <- rexp(N, lambda)
Sn <- cumsum(Tn)
Un <- runif(N)
keep <- (Un <= cos(Sn)^2) #indicator, as logical vector
T.rec[[i]] <- Sn[keep]
}

>n.obs <- sapply(T.rec, length)
>seq.max <- seq_len(max(n.obs))
>T.rec.mat <- sapply(T.rec, "[", i = seq.max)
>len <- length(T.rec.mat[,1])
>T.rec.vec <- cbind(rep(1:m, each = len), as.vector(T.rec.mat),
rep(1,length(T.rec.mat[,1])), rep(1,length(T.rec.mat[,1])))
>T.rec.f <- T.rec.vec[!is.na(T.rec.vec[,2]),]

#rbinom(length(T.rec.mat[,1]),size=1,prob=0.8)

>dat <- as.data.frame(cbind(T.rec.f[,2], T.rec.f[,3], T.rec.f[,1],
T.rec.f[,4]))
>names(dat) <- c("time", "event", "subject", "covariate")
>dat

>library(INLA)

>n.cut = 20 # 40, 60, 80
>cutpoints = seq( 0, 10, len=n.cut)
>formula = inla.surv(time, event, subject=subject) ~ 1 +
f(subject, model="iid")
>model = inla(formula, family="coxph",
control.hazard=list(cutpoints=cutpoints),
#control.inla = list(int.strategy="grid", diff.logdens=15, dz=0.2),
        data=dat)
>h = inla.hyperpar(model, dz=0.2, diff.logdens=15)
#summary(h)


>st <- st.low <- st.up <- rep(NA, n.cut)
>for (i in 1:n.cut){
st[i] <- median(h$marginals.random$baseline.hazard[[i]][,1])
st.low[i] <- quantile(h$marginals.random$baseline.hazard[[i]][,1],
probs=0.025)
st.up[i] <- quantile(h$marginals.random$baseline.hazard[[i]][,1],
probs=0.975)
}
```

```
>temp=seq(0,10,length.out=(n.cut-1))
>graph<-stepfun(temp,st,f=0)
>graph.low <- stepfun(temp,st.low,f=0)
>graph.up <- stepfun(temp,st.up,f=0)

>plot(graph, ylim=c(-4,1), main="", xlab="time", ylab="baseline hazard",
lty=2)
#lines(graph.low, col=3, lty=3)
#lines(graph.up, col=3, lty=3)
>curve(log(3*cos(x)^2), from=0, to=10, ylab=expression(hat(h)(t)),
col="red", xlab="t", add=TRUE)

#library(NHPoisson)
#lambda <- function(t) 3*cos(t)^2
#tt <- seq(0,4,length.out=10)
#yy <- lambda(tt)
#aux<-simNHP.fun(lambda=yy)
#aux$posNH
```

Aabstract

Due to the complexity of appropriate statistical models for Bayesian analysis of recurrent event data, we need to use sampling-based methods such as MCMC algorithms. However, MCMC methods applied to these models come with a wide range of problems in terms of convergence, mixing properties and computational time. An alternative method is integrated nested Laplace approximation (INLA) method, introduced by Rue et al. (2009). The INLA method is very fast and does not suffer the same problems as MCMC. Moreover, the approximations described by INLA to be extremely accurate so that, in order for any bias to be detected, the MCMC algorithm would have to run for much longer time than it is usually done in practice. In this thesis, we assume that recurrent events occur according to a non-homogeneous Poisson process and then perform approximate Bayesian inference using INLA. We illustrate our approach using both simulated and real life data.

Keywords: Recurrent events, integrated nested Laplace approximation, non-homogeneous Poisson process, Cox proportional hazard model, Gaussian Markov random field.

Shahrood University of Technology

Faculty of Mathematical Sciences

Approximate Bayesian inference in non-homogeneous Poisson processes with application to analysis of recurrent events

Fatemeh Kargar

Supervisor

Dr. Hossein Baghishani

September 2016