
بسم الله الرحمن الرحيم

دانشکده علوم ریاضی

رساله دکتری

تحلیل بیزی داده‌های رسته‌ای با ساختارهای فضایی و فضایی-زمانی

نگارنده: لیلا برموده

استاد راهنما

دکتر حسین باغیشنی

اسفند ۱۴۰۰

شماره: ۳-۲/۲- ۲۵۱
تاریخ: ۱۴۰۰/۱۲/۱۵
ویرایش:

باسمه تعالی

مدیریت تحصیلات تکمیلی

پیوست شماره ۲

دانشکده : ریاضی

گروه : آمار

رساله دکتری خانم لیلا برموده

شماره دانشجویی: ۹۵۰۰۰۸۵

تحت عنوان: تحلیل بیزی داده های رسته ای با ساختارهای فضایی و فضایی ـ زمان

در تاریخ ۱۴۰۰/۱۲/۰۴ توسط کمیته تخصصی زیر جهت اخذ مدرک رساله دکتری ارزیابی گردید و با درجه عالی مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی: دکتر حسین باغیشنی		نام و نام خانوادگی:	-----
نام و نام خانوادگی: ------	-----	نام و نام خانوادگی: ------	-----

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی: دکتر احمد نزاکتی رضازاده	از طرف	نام و نام خانوادگی: دکتر مهدی قوتمند	از
نام و نام خانوادگی: دکتر افشین فلاح	از طرف		
نام و نام خانوادگی: دکتر امید کریمی	از طرف		

نام و نام خانوادگی رئیس دانشکده : دکتر مهدی قوتمند

تاریخ، امضاء و مهر دانشکده:

وزارت علوم، تحقیقات و فناوری
دانشکده ریاضی

پیشکشی ناقابل به شهید مدافع حرم
مهندس مصطفی زال نژاد
امیدوارم که این رساله از عهده پیشکش‌شدن
به حضورش برآید.

حمد خالق، نعت مخلوق

حمد مخصوص خداست. حمدی که حدش بی نهایت و حسابش بی شمار و مقدارش نامتناهی و زمانش نامنقطع باشد. خدایا تو را شکر می کنم که از همان آغاز کودکی‌ام مرا با درد آشنا کردی تا به ارزش کیمیایی درد پی ببرم و ناخالصی‌های وجودم را در آتش درد بسوزم و هنگام راه رفتن بر روی زمین خاطرم آرام باشد و موجودیت وجود خود را احساس کنم. در آغاز، سپاس را مخصوص وجود پدر و مادری می دانم که صبورانه مرا در این راه یاری نموده و مرا از سرچشمه محبت خویش بهره‌مند ساختند.
همچنین از زحمات دکتر حسین باغیشنی که راهنمایی این رساله را برعهده داشتند تشکر و قدر دانی می‌نمایم.
در پایان نیز وظیفه خود می‌بینم از برادر عزیزم که پیوسته مشوق من برای پیمودن مسیر بوده است، تشکر نمایم.
از داوران محترم که داوری این رساله را تقبل فرموده اند، نیز کمال تشکر را دارم.

لیلا برموده
اسفند ۱۴۰۰

تعهد نامه

این‌جانب **لیلا برموده** دانشجوی دکتری رشته **آمار** در دانشکده **علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده رساله با عنوان **تحلیل بیزی داده‌های رسته‌ای با ساختارهای فضایی و فضایی ـ زمانی** ، تحت راهنمایی آقای دکتر **حسین باغیشنی** متعهد می‌شوم:

- تحقیقات در این رساله 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 رساله، 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 رساله تاثیرگذار بوده‌اند، در مقالات مستخرج از رساله رعایت می‌شود.
- در تمام مراحل انجام این رساله، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این رساله، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

لیلا برموده
اسفند ۱۴۰۰

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 رساله بدون ذکر منبع مجاز نیست.

چکیده

متغیر رسته‌ای، متغیری است که مشاهدات را گروه‌بندی می‌کند. برای مثال، می‌توان گروه‌بندی افراد با ویژگی‌های متفاوت و مشترک بر اساس گروه خونی را نام برد. امروزه گسترش علوم مختلف، تحلیل کارای داده‌های رسته‌ای را ضروری کرده است، به‌طوری که شاخه‌های مختلفی از علوم و کاربردهای آن‌ها به دنبال راهی برای تحلیل دقیق‌تر این نوع داده‌ها هستند. در آمار کلاسیک روش‌های متعددی برای تحلیل داده‌های رسته‌ای ارائه شده‌اند. در رهیافت بیزی تحلیل و استنباط داده‌های رسته‌ای نیازمند استفاده از روش‌های مبتنی بر نمونه‌گیری مانند الگوریتم‌های MCMC است. اما این روش‌ها برای داده‌های رسته‌ای با ساختارهای وابستگی پیچیده از کارایی لازم برخوردار نیستند. رهیافت جانشین در این رساله، استفاده از روش تقریبی بیزی لاپلاس آشیانی جمع‌بسته (INLA) است که از مشکلات عمده الگوریتم‌های MCMC در استنباط بیزی مدل‌های فضایی برحذر است. برای توسعه مدل بیزی مناسب با مشکل ناشناسایی ذاتی در مدل‌های چندجمله‌ای روبرو هستیم که برای رفع آن، از دو نوع قید برای پارامترهای مدل استفاده و کارایی آن‌ها را با هم مقایسه می‌کنیم. همچنین از تبدیل چندجمله‌ای ـ پوآسن استفاده خواهیم کرد تا امکان برازش مدل در چارچوب مدل‌های گاوسی پنهان فراهم شود. به‌عنوان یک مدل و راه چاره‌ای دیگر، با توجه به نوع پاسخ، از مدل لوجیت تکی‌شده استفاده می‌کنیم. سرانجام کارایی مدل‌های پیشنهادی را به کمک مثال‌های شبیه‌سازی و واقعی در هر کدام از روش‌ها به‌طور جداگانه مورد بررسی و ارزیابی قرار می‌دهیم.

واژه‌های کلیدی: استنباط بیزی، شناسایی‌پذیری، مدل چندجمله‌ای فضایی، مدل چندجمله‌ای کسری، لوجیت تکی‌شده، تبدیل چندجمله‌ای ـ پواسن، ساختار وابستگی فضایی ـ زمانی، روش INLA.

پیش‌گفتار

توسعه و پیشرفت‌های متعدد در حوزه‌های مختلف کاربردی، گردآوری و تحلیل داده‌ها را بیش از پیش مهم و حیاتی کرده است. به خوبی می‌توان دید که صنایعی همچون گردشگری و حمل و نقل و علومی مانند علوم اجتماعی، زمین‌فیزیک و پزشکی به شدت نیازمند تحلیل داده‌ها هستند. خوشبختانه با پیشرفت ابزارهای ثبت داده‌ها و همچنین ایجاد فضاهای کافی برای ذخیره آن‌ها، تا حدی از چالش‌های پیش رو در زمینه گردآوری و ثبت داده‌ها کاسته شده است. امروزه حتی می‌توان از به‌روز رسانی دقیقه‌ای و ثانیه‌ای داده‌های با حجم بالا استفاده کرد. گردآوری داده‌ها و سازماندهی آن‌ها تا زمانی که به تصمیم‌گیری و نتیجه خاصی منجر نشود، کار چندان مفیدی نخواهد بود. در واقع، داده‌ها زمینه‌ای را فراهم می‌آورند تا با پردازش و تحلیل آن‌ها درباره پدیده تحت بررسی و قضاوت، تصمیم‌گیری و برنامه‌ریزی کنیم. در این میان، نمی‌توان از اهمیت داده‌های رسته‌ای در تحلیل داده‌ها غافل شد زیرا علاوه بر ذات طبیعی رده‌ای بودن بسیاری از داده‌ها، همواره با داده‌های روبرو هستیم که در تعیین و گردآوری آن‌ها به آزمون و خطا نیاز داریم که خود موجب رده‌ای شدن داده‌ها می‌شود. برای مثال، فرض کنید بخواهیم میزان دوز مؤثر یک دارو را بیابیم، برای این کار باید در ابتدا رده‌هایی از آن را مورد بررسی قرار دهیم.

از طرف دیگر، پیشرفت در علم داده‌ها موجب شده است تا هم‌زمان داده‌های مربوط به مکان و نواحی مختلف در زمان‌های متوالی و مختلف در دسترس باشند. برای بررسی بهتر و جامع‌تر تفاوت‌های منطقه‌ای داده‌ها و کشف روند تغییرات و حرکت آن‌ها باید اثرات مربوط به نواحی مورد مطالعه و زمان آن‌ها را نیز در نظر بگیریم، زیرا تجربه محققان در طول سالیان دراز نشان داده است که داده‌ها در نواحی مختلف و در بازه‌های زمانی مختلف از تفاوت‌های چشمگیری برخوردارند. اگرچه در نظر گرفتن آن‌ها به میزان زیادی بر پیچیدگی مدل‌ها خواهد افزود.

برای تحلیل و بررسی دقیق داده‌ها نیازمند یافتن و به‌کارگیری مدل‌هایی هستیم که بتواند همه اطلاعات موجود در داده‌ها را به روشنی بیان کند؛ زیرا یک مدل خوب باید علاوه بر توانایی توصیف داده‌ها، قدرت پیشگویی بالایی داشته باشد. افزودن هم‌زمان پیچیدگی مدل با در نظر گرفتن اثرات مربوط به مکان داده‌ها و زمان گردآوری آن‌ها، نباید ما را از مسئله قابل شناسایی بودن مدل غافل کند. به‌عبارت دیگر، مدل‌ها باید توانایی تشخیص پارامترهای مدل را به‌خوبی داشته باشند.

با توجه به نکاتی که ذکر شد، در این رساله هدف اصلی مدل‌بندی بیزی داده‌های رسته‌ای با تمرکز بر داده‌های (اسمی) چندجمله‌ای، به دلیل اهمیت و فراوانی آن‌ها در علوم مختلف، با حضور اثرات مربوط به مکان و زمان است. یکی از مسایل مهم مورد نظر ما در این راستا، بررسی و اعمال شرایط شناسایی‌پذیری مدل در کنار کارایی محاسباتی و آماری برازش مدل است. در استنباط بیزی معمولا از الگوریتم‌های نمونه‌گیری MCMC برای تحلیل مدل‌های

آماری پیچیده استفاده می‌کنند، اما این الگوریتم‌ها به‌ویژه برای داده‌های رسته‌ای چندجمله‌ای از چالش‌های جدی مانند همگرایی ضعیف زنجیر مارکوف رنج می‌برند. از طرف دیگر، افزودن اثرات مربوط به مکان و زمان داده‌ها به وضوح بر این دشواری می‌افزاید. برای دوری از مشکلات عمده الگوریتم‌های MCMC در استنباط بیزی مدل‌های فضایی و فضایی–زمانی، رهیافت پیشنهادی در این رساله استفاده از روش تقریبی بیزی لاپلاس آشیانی جمع‌بسته (INLA) است. اهداف این رساله را می‌توان به‌صورت زیر خلاصه کرد:

- افزایش قدرت شناسایی‌پذیری مدل به‌طوری که دقت برآوردها افزایش یابد.
- افزایش کارایی و سرعت در برازش مدل.
- توسعه مدل‌های رسته‌ای با در نظر گرفتن اثرات ذاتی فضایی و فضایی–زمانی در چارچوب بیزی.
- توسعه مدل رگرسیون جمعی–ساختاری برای داده‌های رسته‌ای در چارچوب بیزی.
- مقایسه مدل پیشنهادی با سایر مدل‌های متداول برای تحلیل داده‌های رسته‌ای.

با توجه به نکاتی که در بالا ذکر شد، محتوای فصل‌های رساله را می‌توان به‌صورت زیر خلاصه کرد:

فصل اول به معرفی خلاصه‌ای از داده‌های رسته‌ای و مدل‌های متداول به‌کار رفته در تحلیل آن‌ها می‌پردازد. ساختارهای وابستگی خاص مانند ساختار فضایی و فضایی–زمانی نیز در این فصل معرفی می‌شوند.

فصل دوم رهیافت INLA را به‌طور خلاصه شرح می‌دهد.

فصل سوم به تشریح کامل مفهوم شناسایی‌پذیری در حالت کلی و به‌طور خاص در رویکرد بیزی می‌پردازد. همچنین مشکلات شناسایی‌پذیری را معرفی و راه‌حل‌های آن را به همراه نتائج نظری ارائه می‌کند.

فصل چهارم مدل چندجمله‌ای با ساختار جمعی–رگرسیونی با در نظر گرفتن قید شناسایی در حالت فضایی را ارزیابی کرده و به کمک مثال‌های شبیه‌سازی و داده واقعی نتایج حاصل را با مدل متداول چندجمله‌ای کسری مقایسه و تحلیل می‌کند.

فصل پنجم مدل لجیت تکی‌شده برای رفع مشکل شناسایی‌پذیری مدل چندجمله‌ای را در حالتی بررسی می‌کند که داده‌ها فراوانی ندارند و اثر فضایی از نوع زمین‌آماری است.

فصل ششم مسئله شناسایی‌پذیری را برای مدل چندجمله‌ای جمعی–ساختاری فضایی–زمانی بررسی می‌کند.

فهرست مقالات مستخرج از رساله

1. Barmoudeh, L., Baghishani, H., Martino, S. (2022). Bayesian spatial analysis of crash severity data with the INLA approach: assessment of different identification constraints. Accident Analysis and Prevention. To appear.

2. Barmoudeh, L., Baghishani, H. (2022). Spatial Bayesian individualized logit models with application to water quality data analysis. In preparation.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

تحلیل داده‌های رسته‌ای

۱.۱ مقدمه

تحلیل داده‌های رسته‌ای[۱] به دلیل کاربردهای بسیاری که در علوم مختلف از جمله علوم زیستی، پزشکی و علوم اجتماعی دارد، سال‌ها است که علاقه‌مندان زیادی را در این علوم و همچنین علم آمار به خود جذب کرده است. در این راستا در دهه‌های اخیر، با گسترش روش‌های استنباط آماری، کارهای متعددی در این زمینه از دانشمندان علوم مختلف به خصوص در پزشکی به چشم می‌خورد. متغیرهای رسته‌ای را می‌توان به متغیرهایی نسبت داد که فقط با تعدادی از مقادیر یا رسته اندازه‌گیری می‌شوند که با تعریف متغیر تصادفی پیوسته که می‌تواند تعداد نامتناهی از مقادیر را اختیار کند، تفاوت دارد، اما گاهی حتی با داده‌های پیوسته به‌صورت داده‌های رسته‌ای برخورد می‌شود (دنیل، ۱۹۹۹). برای مثال، سن اگرچه به‌طور ذاتی یک متغیر پیوسته است، اما به دلایل کاربردی با آن مانند یک متغیر رسته‌ای برخورد می‌کنند. مثلا ممکن است گروه‌های سنی به‌صورت فاصله‌های پنج ساله نمایش داده شوند. به‌عبارتی می‌توان گفت متغیر رسته‌ای، متغیری است که مشاهدات را گروه‌بندی می‌کند. در واقع در بسیاری از علوم با داده‌های خامی روبرو هستیم که برای تحقیق و بررسی بهتر با آن‌ها رسته‌ای برخورد می‌شود. نقطه مشترک در بین داده‌های مربوط به تولد، مرگ، ازدواج، طلاق، آموزش و پرورش، استخدام، اشتغال و مهاجرت، رسته‌ای بودن آن‌ها است که به‌طور گسترده در علوم

[۱] Categorical

اجتماعی مورد بررسی قرار می‌گیرند. در حقیقت، بیشتر مشاهدات در تحقیقات علوم اجتماعی به‌طور رسته‌ای بررسی می‌شوند. توجه کنید حتی سوال ما با این عنوان که آیا در تحقیقات از متغیر رسته‌ای استفاده می‌شود یا نه، خود یک اندازه‌گیری رسته‌ای محسوب می‌شود.

۲.۱ انواع داده‌های رسته‌ای

بسیاری از متغیرهای رسته‌ای فقط شامل دو رسته هستند؛ مانند متغیرهایی که در آن فقط بر رسته‌های پیروزی و شکست، تاکید دارد که متغیرهای دودویی[۲] نامیده می‌شوند. در بسیاری از بخش‌های علوم مختلف، خروجی‌های حاصل از آزمایش تنها دو مقدار را می‌پذیرد. به‌عنوان مثال، می‌توان به برد و باخت یک بازی بسکتبال، اصابت یا عدم اصابت ضربه توپ توسط یک بازیکن، قبولی یا رد شدن یک دانش‌آموز در امتحان و جواب دادن و ندادن بدن یک بیمار به نوعی از داروی جدید در طول دوره درمان اشاره کرد. در همه این مثال‌ها برای راحتی کار، برآمد آزمایش را به‌صورت پیروزی و شکست بیان می‌کنند. در حالت کلی برای پیروزی پاسخ $Y = ۱$ و برای شکست پاسخ $Y = ۰$ را در نظر می‌گیرند. وقتی که متغیرهای رسته‌ای بیشتر از دو رسته داشته باشند، آن‌ها را به دو دسته اسمی[۳] و ترتیبی[۴] تقسیم می‌کنند.

۱.۲.۱ متغیرهای رسته‌ای اسمی

مشاهدات این نوع متغیرها مقادیر کمّی ندارند و برای نامیدن یک متغیر به‌کار می‌روند. برای مثال

- جنسیت (زن و مرد)
- رنگ مو (قهوه‌ای، سیاه، روشن)
- محل زندگی (شمال استوا، جنوب استوا، هیچکدام)
- نظرات دانش‌آموزان درباره یک برنامه کلاسی (بی‌تفاوت، معمولی، عالی) و
- انتخاب یکی از چهار برند در تولید یک محصول خاص مثل ماست

از جمله این نوع داده‌ها هستند.

---

[۲] Binary
[۳] Nominal
[۴] Ordinal

۲.۲.۱ متغیرهای رسته‌ای ترتیبی

گونه دیگری از داده‌های رسته‌ای دارای رسته‌های ترتیبی هستند که بر اساس مقیاس ترتیبی اندازه‌گیری می‌شوند. از جمله مثال‌های این نوع داده‌ها می‌توان موارد زیر را فهرست کرد:

- بیان شرایط یک بیمار بعد از مصرف نوع جدیدی از دارو به‌صورت خوب، متعادل، بد و بحرانی که به ترتیب شدت را نشان می‌دهند.
- بیان میزان رضایت از خدمات یک هتل (خیلی ناراضی، ناراضی، تا حدی راضی، راضی، خیلی راضی)
- تعیین طبقه اجتماعی افراد (پایین، متوسط، بالا)
- سطح سواد افراد (راهنمایی، دبیرستان، دانشگاه) و
- رتبه‌بندی نظر یک بیننده نسبت به یک برنامه تلویزیونی (ضعیف، متوسط، خوب).

لازم به ذکر است در این رساله به علت کاربردهای متعدد و بسیاری که در طبیعت و مطالعات گذشته از داده‌های اسمی به چشم می‌خورد، تمرکز بر تحلیل داده‌های رسته‌ای اسمی است. به همین منظور، تنها در این فصل، به اختصار به تحلیل داده‌های ترتیبی اشاره خواهیم کرد. توجه داشته باشید که در این رساله برای نشان دادن بردار و ماتریس از صورت پر رنگ حروف لاتین استفاده می‌شود.

۳.۱ توزیع‌های داده‌های رسته‌ای

تحلیل و استنباط داده‌ها نیازمند داشتن فرض‌هایی در مورد سازوکار تولید تصادفی داده‌ها است. برای مثال، در تحلیل داده‌های پیوسته، توزیع نرمال، نقش کلیدی را بازی می‌کند. در این بخش، سه توزیع کلیدی برای تحلیل پاسخ‌های رسته‌ای، یعنی توزیع دوجمله‌ای، توزیع چندجمله‌ای و توزیع پوآسن را مرور خواهیم کرد.

۱.۳.۱ توزیع دوجمله‌ای

عموما هدف محققان از تحلیل داده‌های دودویی، برآورد و پیشگویی احتمال پیروزی یا شکست در حضور برخی متغیرهای تبیینی است. در متغیرهای دودویی واحدهای آزمایش به‌صورت تکی بررسی می‌شوند، به‌طوری که برای هر واحد تنها دو رسته پیروزی و شکست امکان‌پذیر است. چنین آزمایشی را آزمایش برنولی[۵] می‌نامند (اگرستی، ۲۰۱۳) که تنها دارای یک پارامتر

[۵] Bernoulli

برای پیروزی است. تابع جرم احتمال برنولی

$$\mathrm{P}(Y=y|\pi)=\pi^{y}(۱-\pi)^{۱-y}\quad y=۰,۱,$$

است، که در آن $\mathrm{P}(Y=۱)=\pi$ احتمال پیروزی را نشان می‌دهد. فرض کنید $(y_۱,y_۲,\ldots,y_n)'$ بردار مشاهدات حاصل از n آزمایش مستقل برنولی باشد. مجموع تعداد موفقیت‌ها یعنی $Y=\sum_{i=۱}^{n}Y_i$ دارای توزیع دوجمله‌ای[۶] با تابع احتمال

$$\mathrm{P}(Y=y)=\binom{n}{y}\pi^{y}(۱-\pi)^{۱-y}\quad y=۰,۱,۲,\ldots,n,$$

است. چون $\mathrm{E}(Y_i)=\mathrm{E}(Y_i^{۲})=\pi$، بنابراین امید ریاضی و واریانس توزیع دوجمله‌ای برای Y به ترتیب برابر با

$$\mathrm{E}(Y)=n\pi\qquad \mathrm{Var}(Y)=n\pi(۱-\pi)$$

است. در حالت کلی و به‌طور خلاصه آزمایش دوجمله‌ای باید دارای ویژگی‌های زیر باشد:

۱. آزمایش شامل n آزمایش برنولی یکسان است، به این معنی که احتمال موفقیت، π، برای هر آزمایش برنولی ثابت است.

۲. هر آزمایش دارای دو برآمد است که یکی از برآمدها را موفقیت و دیگری را شکست می‌نامند.

۳. آزمایش‌ها مستقل هستند، یعنی نتیجه یک آزمایش تأثیری در نتایج آزمایش‌های دیگر ندارد.

۴. متغیر تصادفی Y برابر با تعداد موفقیت‌های دیده‌شده در n آزمایش است.

مثال ۱.۳.۱. بسیاری از وقایع در زندگی واقعی را می‌توان با توزیع احتمال دوجمله‌ای توضیح داد. در زیر به دو کاربرد اشاره می‌کنیم.

- توزیع دوجمله‌ای به‌طور عمده در زمینه داروسازی استفاده می‌شود. برای مثال، زمانی که یک دارو برای معالجه یک بیماری خاص کشف یا اختراع می‌شود، نتایج حاصل از آزمایش با دو برآمد نمایش داده می‌شود؛ یعنی دارو در درمان بیماری مد نظر مفید است یا خیر. بنابراین با دو نتیجه موفقیت و شکست روبرو هستیم که برای مدل‌بندی موثر بودن دارو بر اساس تعداد بهبودیافته‌ها از توزیع دوجمله‌ای استفاده می‌کنیم.
- یک کاربرد توزیع دوجمله‌ای می‌تواند در مدل‌بندی تعداد پیامک‌های ناشناس دریافتی افراد باشد، زیرا در دریافت هر پیامک دو نتیجه ممکن وجود دارد: یک پیامک ناشناس است یا نیست.

[۶] Binomial

۲.۳.۱ توزیع چندجمله‌ای

حال فرض کنید برای n آزمایش مستقل، برآمدها از $J(\geq ۲)$ رسته باشند. همچنین فرض کنید اگر برآمد حاصل از آزمایش i متعلق به رسته j باشد، آن‌گاه $Y_{ij} = ۱$ و در غیر این صورت $Y_{ij} = ۰$. بنابراین $\boldsymbol{y}_i = (y_{i۱}, y_{i۲}, \ldots, y_{iJ})'$ نشان‌دهنده مشاهده یک آزمایش چندجمله‌ای با $\sum_j y_{ij} = ۱$ است. فرض کنید $N_j = \sum_i Y_{ij}$ تعداد برآمدهای مربوط به رسته jام باشد. بنابراین $(N_۱, N_۲, \ldots, N_J)'$ دارای توزیع چندجمله‌ای است. حال فرض کنید در هر بار آزمایش، احتمال این‌که نتیجه متعلق به رسته jام باشد برابر با $\pi_j = \mathrm{P}(Y_{ij} = ۱)$ است. در این صورت، تابع جرم احتمال چندجمله‌ای به‌صورت

$$\mathrm{P}(N_۱ = n_۱, \ldots, N_J = n_J) = \binom{n!}{n_۱! \ldots n_J!} \pi_۱^{n_۱} \pi_۲^{n_۲} \ldots \pi_J^{n_J}$$

است که در آن $\sum_j n_j = n$. با توجه به این‌که $n_J = n - (n_۱ + n_۲ + \cdots + n_{J-۱})$، پس توزیع $J-۱$ بعدی است. توزیع دوجمله‌ای حالت خاصی از توزیع چندجمله‌ای با $J = ۲$ است. میانگین، واریانس و کوواریانس برای این توزیع به‌صورت

$$\mathrm{E}(N_j) = N\pi_j, \quad \mathrm{Var}(N_j) = N\pi_j(۱ - \pi_j), \quad \mathrm{Cov}(N_j, N_k) = -N\pi_j\pi_k,$$

هستند. می‌توان نشان داد که به ازای $j = ۱, \ldots, J$ توزیع کناری N_j، دوجمله‌ای است (اگرستی، ۲۰۱۳). مثال‌های زیر که از اگرستی (۲۰۰۲) برگرفته شده‌اند، برخی کاربردهای توزیع چندجمله‌ای را نشان می‌دهند.

- یک زیست‌شناس در نظر دارد تا بداند گرایش سوسمارهای موجود در یک برکه به چه نوعی از غذا بیشتر است. برای این کار، او در ابتدا انواع غذاها را رده‌بندی می‌کند. بنابراین متغیر پاسخ در این مثال، تمایل سوسمارها به انواع غذاها، دارای توزیع چندجمله‌ای است و عواملی مثل اندازه سوسمار و عوامل محیطی دیگر را می‌توان به‌عنوان متغیرهای تبیینی در نظر گرفت.

- در علوم مربوط به بازاریابی[۷] به‌طور عمده از توزیع چندجمله‌ای استفاده می‌کنند (چن و همکاران، ۲۰۰۱). با این هدف، تعدادی از برندهای خاص (مثل لوازم خانگی) را به‌عنوان متغیر پاسخ انتخاب کرده و عوامل تأثیرگذار مختلف را بر روی آن می‌سنجند.

۳.۳.۱ توزیع پوآسن

گاهی، داده‌های شمارشی[۸] نتیجه تعداد ثابتی از آزمایش‌ها نیست. برای مثال، اگر متغیر Y تعداد افراد کشته‌شده در تصادفات جاده هراز شهرستان آمل در هفته آینده باشد، آن‌گاه حد

[۷] Marketing
[۸] Count data

بالایی برای تعداد n وجود ندارد. یک توزیع پرکاربرد برای این نوع متغیر، توزیع پوآسن است. تابع جرم احتمال پوآسن به‌صورت

$$\mathrm{P}(Y=y)=\frac{\mathrm{e}^{-\lambda}\lambda^{y}}{y!},\quad y=0,1,\ldots,$$

است. توزیع پوآسن دارای میانگین و واریانس یکسان و برابر با λ است. از توزیع پوآسن می‌توان برای تقریب توزیع دوجمله‌ای زمانی که اندازه n بزرگ و π کوچک باشد، نیز استفاده کرد.

۴.۱ مدل‌های رگرسیونی برای داده‌های رسته‌ای

به روشنی مدل‌بندی داده‌ها به ما کمک می‌کند تا ارزیابی کاراتری از تفسیر روند داده‌ها و تأثیرات مربوط به آن‌ها داشته باشیم. امروزه با گسترش روش‌های آماری هم‌زمان با حضور نرم‌افزارهای قوی ما را به سمت موقعیت‌هایی می‌کشاند که می‌توان از سایر توزیع‌ها به‌جز توزیع نرمال برای متغیر پاسخ استفاده کرد. بنابراین نیازمند روش‌های فراگیری در تحلیل داده‌ها هستیم. در طول دهه‌های اخیر، پیشرفت‌های چشمگیری در تعریف و به‌کارگیری مدل‌ها اتفاق افتاده است. یکی از این پیشرفت‌ها در طول تاریخ رشته آمار، تشخیص این بوده است که می‌توان بسیاری از ویژگی‌های خوب توزیع نرمال را در خانواده توزیع‌های نمایی[۹] نیز یافت. در زیر به معرفی این خانواده و برخی مدل‌های عمومی تأثیرگذار در آن برای تحلیل داده‌های رسته‌ای خواهیم پرداخت.

۱.۴.۱ خانواده توزیع‌های نمایی

متغیر تصادفی Y با مشاهدات $\{y_1,\ldots,y_n\}$ را که توزیع احتمال آن به پارامتر θ وابسته است، در نظر بگیرید. این توزیع متعلق به خانواده توزیع‌های نمایی است، اگر تابع (چگالی) احتمال آن به صورت

$$\mathrm{p}(y_i|\theta_i)=a(\theta_i)b(y_i)\exp(s(y_i)Q(\theta_i))$$

باشد، که در آن $a(\cdot)$، $b(\cdot)$، $s(\cdot)$ و $Q(\cdot)$ توابع معلومی هستند. در این توزیع، $Q(\theta_i)$ پارامتر طبیعی نامیده می‌شود. شکل دیگر توزیع به‌صورت

$$\mathrm{P}(Y_i=y_i,\theta_i)=\exp(s(y_i)Q(\theta_i)+c(\theta_i)+d(y_i)) \qquad (۱.۱)$$

است که برای $s(y_i)=y_i$ خانواده را کانونی[۱۰] می‌نامند (اگرستی، ۲۰۰۲). بسیاری از خانواده توزیع‌های معروف مانند پوآسن، نرمال و دوجمله‌ای در این خانواده قرار می‌گیرند. بعضی از آن‌ها را در جدول ۱.۱ می‌توان مشاهده کرد.

[۹] Exponential family of distributions
[۱۰] Cononical

جدول ۱.۱: توزیع‌های پوآسن، دوجمله‌ای و نرمال به‌عنوان مثالی از خانواده توزیع‌های نمایی در حالت تک متغیره

توزیع	پارامتر طبیعی	c	d
پوآسن	$\log(\theta)$	$-\theta$	$-\log(y!)$
نرمال	$\frac{\mu}{\sigma^2}$	$\frac{-\mu}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$	$\frac{-y^2}{2\sigma^2}$
دوجمله‌ای	$\log(\frac{\pi}{1-\pi})$	$n\log(1-\pi)$	$\log\binom{y}{n}$

۲.۴.۱ مدل‌های خطی تعمیم‌یافته

مدل‌های خطی تعمیم‌یافته[11] شامل دامنه گسترده‌ای از مدل‌های آماری هستند که می‌توان آن‌ها را به مدل‌هایی با متغیرهای پاسخ گسسته تعمیم داد و اغلب در مواردی به‌کار می‌روند که توزیع متغیر پاسخ نرمال نیست. در این مدل‌ها با پذیرفتن استقلال مشاهدات، با استفاده از یک تابع پیوند بین میانگین مشاهدات و متغیرهای تبیینی ارتباط برقرار می‌شود. همچنین توزیع پاسخ (توزیع خطا) عضوی از خانواده توزیع‌های نمایی و تابع پیوند یکنوا و مشتق‌پذیر است. فرض کنید $\{y_i,\ \ i = 1,\ldots,n\}$ مشاهدات مستقل از یک خانواده از توزیع‌های نمایی باشند که دارای تابع جرم احتمال (۱.۱) هستند. در یک مدل خطی تعمیم‌یافته، مقادیر پارامتر θ_i، $i = 1,\ldots,n$، را به‌صورت تابعی از متغیرهای تبیینی در نظر می‌گیرند. بین پارامتر θ_i و میانگین (شرطی) پاسخ، μ_i، در GLM رابطه‌ای وجود دارد که توسط تابعی به اسم پیوند تعیین می‌شود. اگر $\{Y_i,\ \ i = 1,\ldots,n\}$ یک نمونه تصادفی از متغیر پاسخ باشد، آن‌گاه

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{k=1}^{l} \beta_k x_{ik}, \quad i = 1,\ldots,n$$

که در آن β_0 پارامتر عرض از مبدا، β_kها ضرایب رگرسیونی و x_kها متغیرهای تبیینی هستند و $g(\cdot)$ تابع پیوند است. تابع پیوندی که میانگین را به پارامتر طبیعی متصل کند، پیوند کانونی نامیده می‌شود. در این حالت $g(\mu_i) = Q(\theta_i)$. ترکیب خطی متغیرهای تبیینی، یعنی η، پیشگوی خطی نامیده می‌شود.

به‌طور خلاصه، مدل‌های خطی تعمیم‌یافته دارای سه مولفه است:

۱. توزیع خطا (یا همان توزیع پاسخ) که عضوی از خانواده توزیع‌های نمایی است و توزیع متغیر پاسخ Y_i را مشخص می‌کند.

۲. مولفه سیستماتیک که متغیرهای تبیینی را به‌صورت ترکیب خطی $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ در نظر می‌گیرد، که در آن $\boldsymbol{x}_i = (1, x_{i1},\ldots,x_{il})'$ بردار متغیرهای تبیینی برای مشاهده iام و $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_l)'$ بردار ضرایب رگرسیونی است. اغلب مولفه سیستماتیک را پیشگوی خطی می‌گویند.

[11] Generalized linear models (GLM)

۳. تابع پیوند یکنوا که بین میانگین شرطی توزیع پاسخ و پیشگوی خطی به‌صورت

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta},$$

پیوند برقرار می‌کند.

مثال ۱.۴.۱. مدل خطی نرمال: معروفترین مدل خطی تعمیم‌یافته، همان مدل خطی نرمال به‌صورت

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2), \quad \mathrm{E}(Y_i|\boldsymbol{x}_i) = \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta},$$

است. در این‌جا تابع پیوند، تابع همانی $g(\mu_i) = \mu_i$ است. این مدل معمولا به شکل

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

نوشته می‌شود، که در آن ϵ_i جمله خطا با توزیع $\mathrm{N}(0, \sigma^2)$ است و فرض می‌شود ϵ_iها مستقل و هم‌توزیع هستند.

مثال ۲.۴.۱. نرخ مرگ و میر: در یک جامعه بزرگ، احتمال پیدا کردن فردی که در یک زمان خاص در حال مرگ باشد، کوچک است. با فرض این‌که مرگ و میر حاصل از بیماری‌های غیرعفونی مستقل هستند، می‌توان تعداد مرگ و میر، Y، را با یک توزیع پوآسن $Y \sim \mathrm{Pois}(\lambda)$ مدل‌بندی کرد، که در آن $\lambda = \mathrm{E}(Y)$ متوسط تعداد مرگ در یک زمان خاص مانند یک سال است. این متوسط می‌تواند به عوامل متعددی مانند اندازه جامعه، سن و جنسیت افراد بستگی داشته باشد. برای وارد کردن اطلاعات متغیرهای تبیینی $\boldsymbol{x}$ در مدل، می‌توان از مدل

$$\log(\lambda) = \log(n) + \boldsymbol{x}'\boldsymbol{\beta},$$

استفاده کرد، که در آن n اندازه جامعه و $\boldsymbol{x}'\boldsymbol{\beta}$ نرخ مرگ به ازای کل جمعیت با اندازه n در هر سال است.

تا این‌جا دریافتیم مدل خطی تعمیم‌یافته، تعمیم مدل خطی است که در آن مشاهدات می‌توانند گسسته یا رسته‌ای باشند (مک کلاح و نلدر، ۱۹۸۹). در واقع GLM برای یکی کردن مدل‌های آماری مختلف به‌کار می‌رود تا نتایج جامعی برای رده بزرگی از متغیرهای پاسخ حاصل شود. در مدل‌های خطی، میانگین تابع خطی از پارامترهای رگرسیونی است و واریانس مربوط به مشاهدات ثابت است، در حالی‌که در مدل خطی تعمیم‌یافته، میانگین به کمک یک تابع پیوند تابعی از متغیرهای تبیینی است. در این صورت، بسیاری از مدل‌های آماری معروف که در آن بین متغیر پاسخ و سایر متغیرهای تبیینی رابطه ناخطی وجود دارد را می‌توان به کمک این تعمیم به‌کار گرفت. برای تحلیل داده‌های رسته‌ای اغلب از توابع پیوند لوجیت[۱۲] و پروبیت استفاده می‌کنند که در ادامه مدل‌های مرتبط را معرفی می‌کنیم.

[۱۲] Logit

تابع پیوند لوجیت

اولین بار بارتلت (۱۹۳۷)، شکل عمومی

$$\log(\frac{\pi}{۱-\pi}) \tag{۲.۱}$$

را در مدل رگرسیون مطرح کرد. در کتاب جدول‌های آماری که در سال ۱۹۳۸ منتشر شد، فیشر این تبدیل احتمال را برای پارامترهای دوجمله‌ای در تحلیل داده‌های دوجمله‌ای پیشنهاد کرد. در سال ۱۹۴۴، فیزیکدان و آماردان جوزف برکسون اصطلاح لوجیت را برای این تبدیل پیشنهاد داد. مدل لوجیت به‌طور گسترده در علوم زیستی و علوم اجتماعی به‌کار می‌رود. این مدل به‌ویژه برای مطالعات جمعیت‌شناسی برای ارزیابی تاثیر متغیرهای تبیینی بر مخاطره نسبی برآمدها مثل باروری و مرگ و میر کاربرد فراوانی دارد. در ریاضیات و آمار، تابع لجیت عدد حقیقی $۰ < \pi < ۱$ به‌صورت

$$\text{logit}(\pi) = \log\left(\frac{\pi}{۱-\pi}\right) = \log\pi - \log(۱-\pi)$$

تعریف می‌شود. در مدل‌های با توزیع پاسخ دودویی و چندجمله‌ای، استفاده از رگرسیون میانگین برابر با مدل‌بندی روی احتمال است. دامنه رگرسیون میانگین اعداد حقیقی است، در حالی که احتمال مقداری بین صفر و یک است. بنابراین، به کمک مدل خطی تعمیم‌یافته از تابع پیوند لوجیت استفاده می‌کنیم تا بتوانیم دامنه تغییرات میانگین را حفظ کنیم.

مدل لوجیت برای داده‌های دودویی

در تعریف این مدل از منابعی همچون اگرستی (۲۰۱۳)، کانگدان و همکاران (۲۰۰۵) و اگرستی (۲۰۱۹) استفاده کردیم. برای تعریف مدل یک نمونه nتایی را در نظر بگیرید. فرض‌های زیر را برای هر یک از اعضای نمونه می‌پذیریم:

۱. بردار $\boldsymbol{x}_i = (۱, x_{i۱}, \ldots, x_{il})'$، بردار متغیرهای تبیینی iامین عضو از نمونه است.

۲. متغیر Y_i یک پاسخ دودویی برای i امین عضو از نمونه است که مقادیر $\{۰,۱\}$ را اختیار می‌کند و از توزیع برنولی پیروی می‌کند. بنابراین

$$Y_i|\boldsymbol{x}_i \sim \text{Ber}(\pi_i), \quad \text{E}(Y_i|\boldsymbol{x}_i) = \text{P}(Y_i = ۱|\boldsymbol{x}_i) = \pi_i.$$

ساده‌ترین نوع تابعی که بتوان به‌کار گرفت، تابع خطی

$$\pi(\boldsymbol{x}_i) = \beta_۰ + x_{i۱}\beta_۱ + \cdots + x_{il}\beta_l = \boldsymbol{x}_i'\boldsymbol{\beta},$$

است. اما متأسفانه این مدل ممکن است نتایجی که در آن مقادیر متغیرهای تبیینی و ضرایب رگرسیونی بستگی دارد را به دنبال داشته باشد که در آن π کمتر از صفر یا بیشتر از ۱ شود.

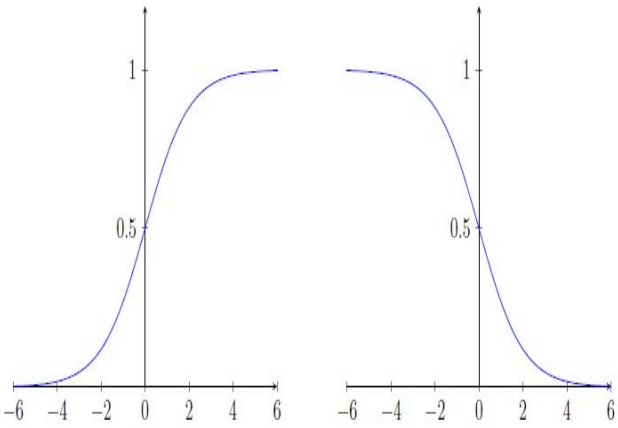
خوشبختانه، توابع پیوند ناخطی متعددی در دسترس هستند که π_i را بین صفر و یک نگه می‌دارند. متداول‌ترین آن‌ها، تابع پیوند لوجیت (۲.۱) است. در این صورت

$$\boldsymbol{\pi}(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{۱ + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}.$$

تفسیر ضرایب رگرسیونی در این مدل بر حسب نسبت بخت‌ها به ازای یک واحد افزایش در متغیر تبیینی، به شرط ثابت بودن سایر متغیرهای تبیینی، است. برای درک این مساله می‌توان نوشت

$$\log\left(\frac{\pi(\boldsymbol{x}_i)}{۱ - \pi(\boldsymbol{x}_i)}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

ضریب مثبت یک متغیر تبیینی باعث افزایش احتمال پیروزی می‌شود. شکل ۱.۱ حالتی را نشان می‌دهد که تنها یک متغیر تبیینی در مدل لوجیت حضور دارد. مشاهده می‌کنید که برای مقادیر مثبت ضریب رگرسیونی، احتمال پیروزی افزایشی و برای مقادیر منفی، کاهشی است. جهت نمودار به علامت x بستگی دارد.



شکل ۱.۱: منحنی رگرسیون احتمال پیروزی به ازای یک متغییر تبیینی، در حالتی که ضریب رگرسیونی مثبت (سمت چپ) یا منفی (سمت راست) است.

مدل لوجیت در آزمایش‌های زیستی به‌ویژه برای تعیین میزان مصرف دوز دارو کاربرد گسترده‌ای پیدا کرده است. مدل‌های با پاسخ دوجمله‌ای عامل اصلی انگیزه برای تحلیل داده‌های مربوط به آزمایش‌هایی است که مقدار مختلفی از یک دارو یا مواد آزمایشی ترکیب‌شده را برای دسته‌ای از موارد مورد آزمایش به‌کار می‌گیرند. برای مثال، فرض کنید در آزمایشی برای تعیین دوز حشره‌کش، تعدادی از حشره‌کش‌ها را برای تعیین سطح دوز مصرفی در سطح معینی می‌آزمایند. در روزهای ابتدایی هیچکدام یا برخی از حشره‌ها از بین نمی‌روند. اما در روزهای بعد ممکن است نابود شوند که حاکی از تاثیر دوز مصرفی بر روی آستانه تحمل حشره دارد.

مدل لوجیت برای داده‌های چندجمله‌ای

در تعریف این مدل از اگرستی (۲۰۱۳)، دو و همکاران (۲۰۰۴)، ؛کانگدان (۲۰۰۵) و سید (۲۰۲۰) استفاده کردیم. داده‌های اسمی با J رسته را در نظر بگیرید. در این حالت، مدل لجیت لگاریتم بخت، یعنی $\log(\frac{\pi}{۱-\pi})$، را برای هر جفت از رده‌ها در نظر می‌گیرد. برای متغیرهای تبیینی $\boldsymbol{x}' = (x_۱, \ldots, x_l)$، فرض کنید

$$\pi_{ij} = \pi_j(\boldsymbol{x}_i) = \mathrm{P}(Y_i = j|\boldsymbol{x}_i), \quad j = ۱, \ldots, J; \quad \sum_{j=۱}^{J} \pi_{ij} = ۱.$$

در حالت کلی برای داده‌های با پاسخ چند رده‌ای، لگاریتم بخت را به‌صورت مقایسه زوجی بین رده‌ها به‌کار می‌گیرند. برای مثال، $\frac{\pi_{ij}}{\pi_{iJ}}$ بخت موفقیت بین دو رده j و J است. در این حالت به رده J رده پایه می‌گویند که انتخاب آن دلخواه است. بنابراین برای هر $j = ۱, \ldots, J$ که $j \neq J$ و $\boldsymbol{\beta}_j = (\beta_{۱j}, \beta_{۲j}, \ldots, \beta_{lj})'$ داریم

$$\begin{aligned} \log(\frac{\pi_{ij}(\boldsymbol{x}_i)}{\pi_{iJ}(\boldsymbol{x}_i)}) &= \log\left(\frac{\pi_{ij}(\boldsymbol{x}_i)}{۱ - \sum_{j=۱}^{J-۱} \pi_{ij}(\boldsymbol{x}_i)}\right) = \beta_{۰j} + \beta_{۱j}x_{i۱} + \cdots + \beta_{lj}x_{il} \qquad (۳.۱) \\ &= \beta_{۰j} + \sum_{k=۱}^{l} x_{ik}\beta_{kj} = \beta_{۰j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j = \eta_{ij} \end{aligned}$$

و همچنین

- برای هر i، y_{ij} مشاهده مربوط به jامین رسته از متغیر تصادفی Y_{ij} است.
- احتمال مشاهده jامین رسته از متغیر پاسخ برای iامین مورد، π_{ij} است و
- β_{kj} ضریب رگرسیونی kامین متغیر تبیینی است.

با بازنویسی معادله (۳.۱) به کمک ویژگی‌های تابع لگاریتم، می‌توان نوشت

$$\pi_{ij} = \pi_{iJ}\exp(\eta_{ij}).$$

با توجه به این‌که $\sum_{j=۱}^{J} \pi_{ij} = ۱$، پس

$$۱ = \pi_{iJ}\exp(\beta_{۰۱} + \boldsymbol{x}_i'\boldsymbol{\beta}_۱) + \pi_{iJ}\exp(\beta_{۰۲} + \boldsymbol{x}_i'\boldsymbol{\beta}_۲) + \cdots + \pi_{iJ}.$$

بنابراین

$$\pi_{iJ} = \frac{۱}{۱ + \sum_{j=۱}^{J-۱} \exp(\beta_{۰j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j)}.$$

از این رو، عبارت عمومی برای π_{ij}، به شکل

$$\pi_{ij} = \frac{\exp(\beta_{۰j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j)}{۱ + \sum_{j=۱}^{J-۱} \exp(\beta_{۰j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j)} = \frac{\eta_{ij}}{۱ + \sum_{j=۱}^{J-۱} \eta_{ij}}$$

به‌دست می‌آید. برای تشریح مدل به مثال زیر برگرفته از اگرستی (۲۰۱۳) توجه کنید.

مثال ۳.۴.۱. پژوهشگران می‌خواهند عواملی را که بر انتخاب غذا توسط سوسمار تأثیر می‌گذارند، بیابند. برای این منظور ۲۱۹ سوسمار از ۴ دریاچه در ایالت فلوریدا را انتخاب کردند. متغیر پاسخ اسمی (یعنی نوع غذای دلخواه) را بر اساس بیشترین حجم غذای اولیه‌ای که از شکم سوسمار می‌یابند، تعیین می‌کنند. با این توصیف، متغیر پاسخ دارای پنج رده ماهی (F)، پرنده‌ها (B)، بی‌مهرگان (I)، خزنده‌ها (R) و سایر (O) است. سوسمارها در دو اندازه مختلف از سه دریاچه مختلف انتخاب شدند. بنابراین اندازه سوسمار و نوع دریاچه را به‌عنوان متغیرهای تبیینی می‌توان در نظر گرفت. متغیرهای تبیینی s، h، o و t به ترتیب نشان‌دهنده اندازه سوسمار، دریاچه اول، دریاچه دوم و دریاچه سوم هستند که همه متغیرهای نشانگر دوسطحی هستند (اندازه سوسمارها به دو رده اندازه بزرگ و کوچک دسته‌بندی شده‌اند). با در نظر گرفتن ماهی به‌عنوان رده پایه، نتایج برآوردهای ماکسیمم درستنمایی پارامترهای مدل لوجیت چندجمله‌ای در جدول ۲.۱ گزارش شده‌اند. مقادیر خطای معیار برآورد اثرات متغیرهای تبیینی داخل پرانتز جلوی هر برآورد ضریب نوشته شده‌اند. برای مثال، معادله پیشگو برای مقایسه

جدول ۲.۱: پارامترهای برآوردشده مدل لوجیت چندجمله‌ای در مثال سوسمارها

مدل لوجیت	عرض از مبدأ	اثر اندازه	اثر دریاچه اول	اثر دریاچه دوم	اثر دریاچه سوم
$\text{logit}(\frac{\pi_I}{\pi_F})$	−۱/۵۵	۱/۴۶ (۰/۴۰)	−۱/۶۶ (۰/۶۱)	۰/۹۴ (۰/۴۷)	۱/۱۲ (۰/۴۹)
$\text{logit}(\frac{\pi_R}{\pi_F})$	−۳/۳۱	−۰/۳۵ (۰/۵۸)	۱/۲۴ (۱/۱۹)	۲/۴۶ (۱/۱۲)	۲/۹۴ (۱/۱۲)
$\text{logit}(\frac{\pi_B}{\pi_F})$	−۲/۰۹	−۰/۶۳ (۰/۶۴)	۰/۷۰ (۰/۷۸)	−۰/۶۵ (۱/۲۰)	۱/۰۹ (۰/۸۴)
$\text{logit}(\frac{\pi_O}{\pi_F})$	−۱/۹۰	۰/۳۳ (۰/۴۵)	۰/۸۳ (۰/۵۶)	۰/۰۱ (۰/۷۸)	۱/۵۲ (۰/۶۲)

خوراک بی‌مهرگان در مقابل با ماهی‌ها به‌صورت زیر است:

$$\text{logit}(\frac{\hat{\pi}_I}{\hat{\pi}_F}) = -۱/۵۵ + ۱/۴۶s - ۱/۶۶h + ۰/۹۴o + ۱/۱۲t.$$

در این مدل، به ازای دریاچه مشخص، احتمال انتخاب غذای بی‌مهرگان برای سوسمارهای با اندازه کوچک $\exp(۱/۴۶) = ۴/۳$ برابر زمانی است که اندازه سوسمار بزرگ باشد. همچنین، احتمال انتخاب خوراک بی‌مهرگان در دریاچه‌های دوم و سوم بیشتر از دریاچه اول است.

مدل لوجیت برای داده‌های ترتیبی

در بیان این مدل از مک کلاک (۱۹۸۰) استفاده کردیم. در تحلیل داده‌های ترتیبی به کمک تابع پیوند لوجیت، از شکل تجمعی آن بهره می‌گیرند. فرض کنید برای متغیر تصادفی رسته‌ای Y، J تعداد رسته‌ها باشد. اگر نمونه‌های $n_۱, \ldots, n_J$ را فراوانی در هر رسته در نظر بگیریم، تعداد کل مشاهدات $n = \sum_j n_j$ است. برای یک مشاهده که به‌طور تصادفی از جامعه انتخاب شده است، فرض کنید π_j احتمال پاسخ در رسته j باشد. احتمال تجمعی رده j به‌صورت

$$F_j = \text{P}(Y \le j) = \pi_۱ + \cdots + \pi_j, \quad j = ۱, ۲, \ldots, J$$

است که ترتیب رسته‌ها را به‌صورت

$$۰ < F_۱ < F_۲ < \cdots < F_J = ۱$$

نشان می‌دهد.

برای J رسته با احتمال‌های $\pi_۱, \ldots, \pi_J$، مدل لوجیت تجمعی[۱۳] به‌صورت

$$\begin{aligned} \text{logit}\{\text{P}(Y \le j)\} &= \log \frac{\text{P}(Y \le j)}{۱ - \text{P}(Y \le j)} \\ &= \log \frac{\pi_۱ + \cdots + \pi_j}{\pi_{j+۱} + \cdots + \pi_J}, \quad j = ۱, \ldots, J-۱ \end{aligned}$$

تعریف می‌شود. این مدل باعث می‌شود تا پاسخ‌ها به دو قسمت $Y \le j$ و $Y > j$ تفکیک شوند که خود، نوعی داده ترتیبی دودویی به شمار می‌آید. هر لوجیت تجمعی از اطلاعات همه J رسته استفاده می‌کند.

مدل لوجیت تجمعی با متغیر تبیینی

فرض کنید Y_i متغیر پاسخ رسته‌ای با J رسته و $\boldsymbol{x}_i$ بردار متغیرهای تبیینی باشد. مدل لوجیت تجمعی در این حالت، هم‌زمان از لوجیت تجمعی $(J-۱)$ رسته استفاده می‌کند که به‌صورت

$$\text{logit}\{\text{P}(Y_i \le j)\} = \beta_{۰j} + \boldsymbol{x}_i'\boldsymbol{\beta}, \quad j = ۱, \ldots, J$$

تعریف می‌شود. این مدل یک احتمال شرطی به شرط متغیرهای تبیینی است، بنابراین باید در واقع $\text{P}(Y_i \le j)$ را با $\text{P}(Y_i \le j|\boldsymbol{x}_i)$ جایگزین کنیم. یک مدل معادل برای لوجیت تجمعی به‌صورت

$$\text{P}(Y_i \le j) = \frac{\exp(\beta_{۰j} + \boldsymbol{x}_i'\boldsymbol{\beta})}{۱ + \exp(\beta_{۰j} + \boldsymbol{x}_i'\boldsymbol{\beta})}$$

است. در این مدل‌ها بر خلاف داده‌های رسته‌ای اسمی، پارامترهای رگرسیونی برای همه رسته‌ها یکسان است.

مثال ۴.۴.۱. در تحلیل اختلالات روانی افراد در یک آسایشگاه روانی، شدت اختلالات روانی افراد یک متغیر پاسخ ترتیبی است. رده‌های پاسخ عبارتند از خوب، با علائم خفیف، با علائم متوسط و با اختلال روانی جدی. متغیرهای تبیینی نیز شامل سابقه‌ای از وقایع زندگی افراد $(x_۱)$ و وضعیت اجتماعی‌ـاقتصادی $(x_۲)$ هستند. مدل لوجیت تجمعی به‌صورت

$$\text{logit}\{\text{P}(Y_i \le j)\} = \beta_{۰j} + \beta_۱ x_۱ + \beta_۲ x_۲, \quad j = ۱,۲,۳,۴,$$

است. برآورد ماکسیمم درستنمایی ضرایب رگرسیونی $\beta_۱$ و $\beta_۲$ به ترتیب برابر با $-۰/۳۱۹$ و $۱/۱۱۱$ است. برآوردها نشان می‌دهد که داشتن وضعیت اجتماعی‌ـاقتصادی بهتر (یعنی

[۱۳] Cumulative

مشاهده بزرگتر برای این متغیر) احتمال تجمعی را افزایش می‌دهد و این به آن معنی است که احتمال داشتن شدت اختلال بالاتر کم خواهد شد. نتیجه برای متغیر تبیینی دوم، با توجه به ضریب منفی آن، برعکس است. داده‌های این مثال در http://www.stat.ualberta.ca قابل دسترسی هستند.

تابع پیوند پروبیت

نوع دیگری از تابع پیوند برای تحلیل داده‌های رسته‌ای، تابع پیوند پروبیت[۱۴] است که مدل نتیجه‌شده یکی از مدل‌های معروف در تحلیل داده‌های رسته‌ای محسوب می‌شود. این تابع پیوند، اولین‌بار توسط فردی به نام بلیس در سال ۱۹۳۴ معرفی شد.

مدل پروبیت برای داده‌های دودویی

در این حالت نیز مدل ناخطی در π به یک مدل خطی از π نسبت به متغیرهای تبیینی توسط یک تابع یکنوا تغییر شکل می‌دهد. احتمال iامین مشاهده یعنی π_i، با استفاده از توزیع نرمال استاندارد به‌صورت

$$\pi_i = \int_{-\infty}^{\eta_i} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)du = \Phi(\eta_i),$$

است که از آن داریم

$$\eta_i = \Phi^{-1}(\pi_i).$$

بنابراین مدل پروبیت را می‌توان به‌صورت

$$\Phi^{-1}(\pi_i) = \eta_i = \beta_0 + \sum_{k=1}^{l} x_{ik}\beta_k,$$

یا

$$\pi_i = \Phi(\beta_0 + \sum_{k=1}^{l} x_{ik}\beta_k),$$

نوشت.

مدل پروبیت تجمعی برای داده‌های رسته‌ای ترتیبی

مدل پروبیت تجمعی به‌صورت

$$\Phi^{-1}(\mathrm{P}(Y_i \le j)) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}, \quad j = 1, \ldots, J-1,$$

است که در علوم مختلف مدل پروبیت ترتیبی نامیده می‌شود. همانند مدل لجیت، اثر $\boldsymbol{\beta}$ برای همه رسته‌ها یکسان است. مدل پروبیت تجمعی، احتمالات تجمعی را به‌صورت

$$\mathrm{P}(Y_i \le j) = \Phi(\beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}), \quad j = 1, \ldots, J-1,$$

[۱۴] Probit

توصیف می‌کند.

۵.۱ ساختارهای وابستگی

تا به اینجا دانستیم که می‌توان به کمک رده مدل‌های خطی تعمیم‌یافته، بسیاری از داده‌های رسته‌ای را با فرض این‌که احتمال وقوع یک رسته تابعی از متغیرهای تبیینی باشد، مدل‌بندی و تحلیل کرد. اگرچه پیشرفت بزرگی به‌نظر می‌رسد، اما گاهی اوقات با داده‌هایی روبرو هستیم که در بین آن‌ها ساختار همبستگی وجود دارد. در این حالت، باید به دنبال مدلی باشیم که همبستگی موجود در داده‌ها را برای به‌دست آوردن کارایی بالاتر در تحلیل لحاظ کند. یکی از مدل‌های مناسب، مدل آمیخته خطی تعمیم‌یافته[۱۵] (فارمیر و توتز، ۲۰۰۱) نام دارد.

۱.۵.۱ مدل‌های آمیخته خطی تعمیم‌یافته

در حالتی که بین مشاهدات همبستگی وجود دارد، در مدل خطی تعمیم‌یافته، پذیره استقلال مشاهدات به استقلال شرطی تعدیل و همبستگی بین آن‌ها با اضافه‌کردن اثرات تصادفی[۱۶] از طریق متغیرهای پنهان[۱۷] به مدل در نظر گرفته می‌شود. در این صورت مدل خطی تعمیم‌یافته به مدل آمیخته خطی تعمیم‌یافته تبدیل می‌شود و پیشگوی خطی در آن، در قالب ماتریسی، به‌صورت

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u},$$

بیان می‌شود، که در آن $\boldsymbol{\eta}' = (\eta_1, \ldots, \eta_n)$، $\boldsymbol{\beta}$ بردار $l+1$ بعدی از اثرات ثابت، $\boldsymbol{u}$ بردار q بعدی از اثرات تصادفی، $\boldsymbol{X}$ ماتریس طرح $n \times (l+1)$ بعدی و $\boldsymbol{Z}$ ماتریس طرح $n \times q$ بعدی است. معمولا در این رده از مدل‌ها، فرض می‌شود بردار $\boldsymbol{u}$ دارای توزیع گاوسی است و وابستگی مشاهدات پاسخ از طریق همین بردار تصادفی وارد مدل می‌شود. در حالت خاصی که $q = 1$ و $\boldsymbol{Z}$ یک بردار n بعدی از ۱ هاست، مدل حاصل به مدل با عرض از مبدا تصادفی[۱۸] معروف است.

۲.۵.۱ مدل آمیخته جمعی‌ـ‌ساختاری تعمیم‌یافته

با پیشرفت‌های مختلفی که در علوم رخ داده است، داده‌هایی در طبیعت مشاهده می‌کنیم که همواره در بین آن‌ها رد پای اثرات مختلف را می‌توان جستجو کرد. برای مثال، در تحلیل داده‌ها ممکن است برخی متغیرهای تبیینی به‌صورت ناخطی روی متغیر پاسخ تاثیر داشته باشند. ممکن است اثرات متقابل در بین برخی از متغیرها وجود داشته باشد که تغییرات مربوط

[۱۵] Generalized linear mixed model (GLMMs)
[۱۶] Random effects
[۱۷] Latent variables
[۱۸] Random intercept model

به آن‌ها نیز هموار نیست. همچنین در بعضی از داده‌ها اثر مکان رخداد پاسخ یا نقطه‌ای که داده متعلق به آن است را می‌توان یافت. افزون بر این، از اثرات مربوط به قرار گرفتن مشاهدات در یک بازه زمانی خاص نمی‌توان غافل شد. به‌طور مثال، اگر داده‌ها مربوط به مقایسه پیشرفت یا مبتلا شدن افراد به یک بیماری خاص در یک بازه زمانی مشخص و در ناحیه‌های مختلف یک کشور باشد، آن‌گاه مدل‌بندی مربوط به تأثیر اقلیم و نواحی محل زندگی در طول زمان ابتلا به این بیماری، گریزناپذیر است. در این صورت باید به دنبال مدلی باشیم تا از همه این اطلاعات بهره ببرد.

هستی و تیبشرانی (۱۹۹۰) رده مدل‌های جمعی تعمیم‌یافته[۱۹] را معرفی کردند که می‌توان در آن اثر ناخطی متغیرهای تبیینی را وارد مدل کرد. با برقراری پذیره‌های مدل خطی تعمیم‌یافته، در این رده از مدل‌ها، پیشگوی خطی به یک پیشگوی جمعی به‌صورت

$$g(\mu) = \eta = \beta_\circ + \sum_{j=1}^{l} f_j(x_j),$$

تبدیل می‌شود که در آن $f_j(\cdot)$ها اثرات ناخطی (ناپارامتری) متغیرهای تبیینی هستند که معمولا شرایط همواری مشخصی برای آن‌ها در نظر گرفته می‌شود. تعمیمی دیگر از مدل GAM مدلی است که شامل هر دو اثرات خطی و ناخطی به شکل

$$g(\mu) = \eta = \beta_\circ + \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{c=1}^{p} f_c(z_c),$$

است، که در آن $\boldsymbol{x}$ دارای اثرات خطی و z_cها دارای اثرات ناخطی هستند. این نوع مدل، در واقع، یک مدل نیمه‌پارامتری است.

تعمیمی دیگر بر مدل نیمه‌پارامتری، تحت عنوان مدل آمیخته جمعی ـ ساختاری تعمیم‌یافته[۲۰]، به پیشگوی جمعی اثرات تصادفی را نیز اضافه می‌کند که برای پاسخ‌های با ساختار وابستگی می‌تواند مناسب و به اندازه کافی منعطف باشد. در این حالت

$$g(\mu) = \eta = \beta_\circ + \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{c=1}^{p} f_c(z_c) + f(u), \qquad (۴.۱)$$

که در آن $f(u)$ اثر تصادفی با معمولا توزیع نرمال است.

معمولا یک حالت کلی‌تر برای پیشگوی (۴.۱) مطرح است که به آن مدل رگرسیون جمعی ـ ساختاری[۲۱] می‌گویند (فاهرمیر و توتز، ۲۰۰۱). در این نوع از مدل‌بندی، پیشگوی خطی با یک پیشگوی کامل‌تر جمعی ـ ساختاری به صورت

$$\eta = \beta_\circ + \sum_{r=1}^{n_f} f(u_r) + \sum_{k=1}^{l} x_k\beta_k,$$

[۱۹] Generalized additive models (GAMs)
[۲۰] Generalized additive mixed models (GAMMs)
[۲۱] Structured additive regression model (STAR)

جایگزین می‌شود. به کمک توابع $f(\cdot)$ می‌توان اثرات ساختاری مختلف را وارد مدل کرد. این اثرات می‌توانند شامل

- اثرات ناخطی توابع پیوسته
- رویه‌های هموار دوبعدی (یا چندبعدی)
- اثر فضایی
- اثر فضایی‌ـ‌زمانی
- متغیرهای با اثرات متغیر زمانی یا فضایی

شوند. در این مدل، متغیرهای u می‌توانند متغیرهای اندیس‌گذار زمان، مکان، زمان‌ـ‌مکان یا مشاهدات یک متغیر تبیینی پیوسته باشند.

با معرفی مدل‌های بالا، دریافتیم که برای لحاظ کردن اثرات فضایی و فضایی‌ـ‌زمانی (که مد نظر این رساله است) در تحلیل داده‌های رسته‌ای، باید از چه رویکردهایی در مدل‌بندی استفاده کنیم. در ادامه، به‌طور خلاصه به تعریف این اثرها خواهیم پرداخت.

۶.۱ ساختارهای وابستگی خاص

در فرآیند تحلیل داده‌ها، ممکن است با داده‌هایی روبرو شویم که از نظر مکانی و زمانی به یکدیگر وابسته هستند. به‌ویژه این‌که در دهه‌های اخیر، به‌خاطر پیشرفت در ابزارهایی مانند ماهواره‌ها و GPS، جمع‌آوری داده‌های واقعی زمانی ممکن شده است. در واقع، امروزه در شاخه‌های علمی مختلفی مانند محیط زیست، اقلیم‌شناسی، علوم اجتماعی و علوم پزشکی، داده‌ها همراه اطلاعات فضا و زمان رخداد آن‌ها در اختیار محققان قرار دارند. به‌عنوان مثال، فرض کنید میزان وقوع یک بیماری خاص مثلا سرطان ریه در تمام کشور یا نواحی بزرگی از آن در سال‌های مختلف، هدف مطالعه باشد. به نظر شما چه نوع مدلی این امکان را به ما می‌دهد تا جای ممکن از تمام اطلاعات موجود در داده‌ها به خوبی استفاده کنیم. برای این کار، ابتدا باید الگوی جغرافیایی بیماری شناسایی و تعیین شود، به‌طوری که ناحیه‌های نزدیک به هم الگوهای مشابهی از شاخص‌های جغرافیایی را نمایش دهند. این وظیفه توسط مدل‌های فضایی[۲۲] انجام می‌شود. به‌عبارتی، احتمال وقوع رخداد بیماری در ناحیه‌های همسایه به هم نزدیک‌تر است. همچنین سوال مهم دیگر این می‌تواند باشد که احتمال وقوع در طول زمان چگونه تغییر می‌کند؟ یا این‌که بخواهیم مسئله را هم از نظر جغرافیایی و هم از نظر زمانی بررسی کنیم که نتیجه به مدل‌های فضایی‌ـ‌زمانی[۲۳] منتهی می‌شود. امروزه مدل‌های فضایی

[۲۲] Spatial models
[۲۳] Spatio-temporal models

و فضایی–زمانی به‌طور وسیعی مورد استفاده قرار می‌گیرند. از همین رو، مقدمه مختصری از برخی مدل‌های مربوط به ساختار وابستگی که امروزه به وفور گسترش یافته‌اند، بیان می‌کنیم. این مدل‌ها در فصل‌های آینده به‌کار می‌روند.

۱.۶.۱ ساختار وابستگی زمانی

در بسیاری از موقعیت‌های کاربردی با داده‌هایی مواجه می‌شویم که از نظر زمانی به یکدیگر وابسته هستند. رویکردهای مختلفی برای مدل‌بندی ساختار وابستگی زمانی معرفی شده‌اند. به‌عنوان نمونه، ساختار وابستگی زمانی را می‌توان با مدل‌های اتورگرسیو[۲۴]، میانگین متحرک[۲۵] یا ترکیب آن‌ها یعنی اتورگرسیو میانگین متحرک[۲۶] وارد فرآیند تحلیل داده‌ها کرد (باکس و همکاران، ۱۹۹۴). به‌عنوان مثال، فرآیند اتورگرسیو به صورت زیر تعریف می‌شود.

تعریف ۱.۶.۱. دنباله $\{X_t,\ t \in \mathcal{T}\}$ یک فرآیند اتورگرسیو مرتبه p است، هرگاه بتوان آن را به‌صورت

$$X_t = \psi_1 X_{t-1} + \psi_2 X_{t-2} + \cdots + \psi_p X_{t-p} + \epsilon_t$$

نوشت، که در آن ψها پارامترهای مدل، ϵ_t یک فرآیند نوفه سفید (نرمال) و $\mathcal{T}$ مجموعه (گسسته) اندیس‌گذار زمانی است. زمانی‌که p برابر با یک باشد، فرآیند اتورگرسیو از مرتبه اول است و ساختار وابستگی ماتریس کوواریانس، Σ، به‌صورت $\sigma^2\rho^{|i-j|}$ است، که در آن σ^2 واریانس فرآیند نوفه سفید است. در این حالت (برای سه نقطه زمانی گسسته)

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}.$$

از سایر مدل‌های زمانی معروف و پرکاربرد، می‌توان به فرآیندهای قدم‌زدن تصادفی مرتبه اول[۲۷] (RW1) و مرتبه دوم (RW2) اشاره کرد (رو و هلد، ۲۰۰۵).

تعریف ۲.۶.۱. دنباله $\{\nu_t,\ t \in \mathcal{T}\}$ یک فرآیند قدم‌زدن تصادفی مرتبه اول است، هرگاه رابطه

$$\nu_t | \nu_{t-1} \sim \mathrm{N}(\nu_{t-1}, \sigma),$$

برقرار باشد که در آن σ انحراف معیار شرطی مدل است. به‌طور مشابه، اگر

$$\nu_t | \nu_{t-1}, \nu_{t-2} \sim \mathrm{N}(2\nu_{t-1} + \nu_{t-2}, \sigma),$$

[۲۴] Autoregressive (AR)
[۲۵] Moving-average (MA)
[۲۶] Autoregressive moving-average (ARMA)
[۲۷] First order random walk

آن‌گاه فرآیند را قدم‌زدن تصادفی مرتبه دوم می‌نامند. دقت کنید که مثلا برای مدل قدم‌زدن تصادفی مرتبه اول $\nu_t|\nu_{t-1},\ldots$ با $\nu_t|\nu_{t-1}$ هم‌توزیع است.

۲.۶.۱ ساختار وابستگی فضایی

در بررسی‌های محیطی، اغلب با داده‌هایی مواجه می‌شویم که از یکدیگر مستقل نیستند و وابستگی آن‌ها ناشی از موقعیت و مکان قرار گرفتن داده‌ها در فضای مورد بررسی است. این‌گونه داده‌ها، داده‌های فضایی نامیده می‌شوند. داده‌های زمین آماری[۲۸] و داده‌های مشبکه‌ای[۲۹] از انواع داده‌های فضایی محسوب می‌شوند.

داده‌های زمین‌آماری

این نوع داده‌ها در موقعیت‌های ثابت و مشخص در ناحیه‌ای پیوسته (چگال) مشاهده می‌شوند. متغیر مورد بررسی ممکن است گسسته یا پیوسته باشد. به‌عنوان مثال، برای حالت پیوسته می‌توان غلظت مواد معدنی در یک معدن و مقدار ریزش باران ثبت‌شده در ایستگاه‌های هواشناسی و برای حالت گسسته می‌توان تعداد جانوران دریایی در یک مجموعه از مکان‌های نمونه‌گیری شده در طول یک ساحل را نام برد (محمدزاده، ۱۳۹۱). برای مدل‌بندی داده‌های زمین آماری از میدان تصادفی استفاده می‌کنند که در زیر معرفی شده است.

تعریف ۳.۶.۱. میدان تصادفی، مجموعه‌ای از متغیرهای تصادفی مانند $\{Y(s); s \in D\}$ است که در آن مجموعه اندیس‌گذار D یک زیرمجموعه از فضای اقلیدسی d بعدی، $d \geq 1$ است. برای این میدان تصادفی، میانگین در موقعیت s و کوواریانس در موقعیت‌های s_1 و s_2 به‌ترتیب به‌صورت

$$\begin{aligned} \mu(s) &= \mathrm{E}(Y(s)), \quad s \in D, \\ C(s_1, s_2) &= \mathrm{Cov}(Y(s_1), Y(s_2)), \quad s_1, s_2 \in D, \end{aligned}$$

تعریف می‌شوند.

اگر میدان تصادفی نرمال[۳۰] باشد، آن‌گاه برای هر $n \geq 1$ و برای هر مجموعه از موقعیت‌های $\{s_1, \ldots, s_n\}$، بردار $(Y(s_1), \ldots, Y(s_n))'$ دارای توزیع نرمال n متغیره با بردار میانگین $\boldsymbol{\mu} = (\mu(s_1), \ldots, \mu(s_n))'$ و ماتریس کوواریانس $\boldsymbol{\Sigma}$ است که مولفه‌های آن توسط تابع کوواریانس $C(\cdot, \cdot)$ تعیین می‌شوند.

[۲۸] Geostatistical data
[۲۹] Lattice data
[۳۰] Gaussian random field (GRF)

ماتریس کوواریانس با ساختار فضایی را می‌توان به کمک توابع کوواریانس فضایی پارامتری مانند مترن[۳۱]، نمایی، گوسی، کروی[۳۲]، دایره‌ای و درجه ۳ به‌دست آورد. به‌عنوان مثال، تابع کوواریانس مترن برای موقعیت‌های s_i و s_j به‌صورت

$$\mathrm{Cov}(Y(s_i), Y(s_j)) = \mathrm{Cov}(Y_i, Y_j) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}}(\rho||s_i - s_j||)^{\lambda}\kappa_{\lambda}(\rho||s_i - s_j||),$$

تعریف می‌شود، که در آن، $||s_i - s_j||$ فاصله اقلیدسی بین موقعیت‌های s_i و s_j، σ^2 واریانس کناری میدان مترن، $\kappa_{\lambda}(\cdot)$ تابع بسل[۳۳] نوع دوم (آبراموویچ و همکاران، ۱۹۷۲) با مرتبه $\lambda > 0$ و $\Gamma(\cdot)$ تابع گاما است. پارامتر λ نیز درجه هموار فرآیند را اندازه‌گیری می‌کند و معمولا ثابت نگه داشته می‌شود. همچنین ρ پارامتر مقیاس است. در عمل پارامترهای تابع کوواریانس بر اساس مشاهداتی از میدان و با روش‌های درستنمایی ماکسیمم یا بیزی برآورد می‌شوند (کرسی، ۱۹۹۳؛ محمدزاده، ۱۹۹۱).

داده‌های مشبکه‌ای

این نوع داده‌ها مربوط به مکان‌های ناحیه‌ای هستند، که ممکن است منظم یا نامنظم باشند. تصاویر ماهواره‌ای از سطح زمین که به تعدادی عنصر تصویر[۳۴] کوچک به‌صورت شبکه منظم در $\mathbb{R}^2$ تقسیم شده‌اند، مثالی برای داده‌های مشبکه‌ای منظم است. تعداد حوادث رانندگی، تعداد اتباع خارجی غیر مجاز ساکن یا نرخ رشد اقتصادی هر استان مثال‌های دیگر از داده‌های مشبکه‌ای هستند (محمدزاده، ۱۳۹۱). برای مدل‌بندی ساختار فضایی داده‌های مشبکه‌ای می‌توان از مدل اتورگرسیو شرطی[۳۵] ؛ بی‌سیگ، ۱۹۷۴) استفاده کرد. مدل CAR برای بردار تصادفی $(Y_1, \ldots, Y_n)'$ که دارای توزیع نرمال باشد، به صورت زیر تعریف می‌شود، که در آن چگالی‌های شرطی تعیین می‌شوند و از روی آن‌ها می‌توان پارامترهای توزیع توام بردار را کاملا به‌صورت

$$Y_i|\boldsymbol{Y}_{-i}, \tau, \boldsymbol{W} \sim \mathrm{N}\left(\frac{\sum_{m\sim i}^{n}\omega_{im}Y_m}{\sum_{m\sim i}^{n}\omega_{im}}, \frac{1}{\tau\sum_{m\sim i}^{n}\omega_{im}}\right) \qquad (۵.۱)$$

مشخص کرد، که در آن $\boldsymbol{Y}_{-i}$ بردار بدون مولفه i ام است، منظور از $i \sim m$ همسایه بودن دو ناحیه i و m است، τ پارامتر دقت مدل است و ماتریس همسایگی $\boldsymbol{W} = \{\omega_{im}, i \sim m, i > m\}$ دارای مولفه‌های صفر و یک است که یک‌ها مشخص می‌کنند هر ناحیه با کدام نواحی دیگر همسایه است. در این حالت امید شرطی، میانگین اثرهای تصادفی در نواحی همسایه و واریانس شرطی، عکس نسبت همسایه‌ها است.

[۳۱] Matern
[۳۲] Spherical
[۳۳] Bessel
[۳۴] Pixel
[۳۵] Conditional autoregressive (CAR)

این مدل با میانگین‌گیری اثرات همسایه‌ها، نوعی هموواری را در میدان تصادفی القا می‌کند که برای کاربردهای آمار فضایی مشبکه‌ای مطلوب هم هست. مدل (۵.۱) با توجه به این‌که یک چگالی سره را نتیجه نمی‌دهد، به مدل اتورگرسیو شرطی ذاتی[۳۶] معروف است (رو و هلد، ۲۰۰۵). نسخه‌های سره برای مدل‌های CAR نیز وجود دارند که استفاده از مدل ICAR که یکی از این نسخه‌ها است، در اغلب کاربردها برتری دارد. برای جزییات بیشتر می‌توانید به کتاب رو و هلد (۲۰۰۵) مراجعه کنید.

۳.۶.۱ وابستگی فضایی‌ـ‌زمانی

تحلیل داده‌های فضایی‌ـ‌زمانی مستلزم تعیین ساختار همبستگی فضایی‌ـ‌زمانی آن‌ها از طریق تابع کوواریانس است. در دو دهه اخیر بررسی‌های زیادی برای تعیین ساختار همبستگی، مدل‌بندی و تحلیل چنین داده‌هایی انجام شده‌اند (ویکل و کرسی، ۲۰۱۱). ساختار همبستگی این داده‌ها به‌وسیله کوواریانس فضایی‌ـ‌زمانی تعیین می‌شود. این تابع که نقش به‌سزایی در پیشگویی موقعیت‌های فضایی یا زمانی فاقد مشاهده دارد، به‌طور معمول نامعلوم است و باید بر اساس مشاهدات برآورد شود. ساخت توابع کوواریانس فضایی‌ـ‌زمانی معتبر را می‌توان به دو صورت تفکیک‌پذیر[۳۷] و تفکیک‌ناپذیر بررسی کرد که معمولا برای سهولت در برازش مدل معتبر از صورت تفکیک‌پذیر آن استفاده می‌کنند.

تعریف ۴.۶.۱. یک میدان تصادفی فضایی‌ـ‌زمانی، مجموعه‌ای از متغیرهای تصادفی مانند $Y(\cdot,\cdot) = \{Y(s,t);\ (s,t) \in D \times \mathcal{T}\}$ است، که در آن s موقعیت فضایی در مجموعه $D \subseteq \mathbb{R}^d$، $d \geq ۱$، و t لحظه زمانی در مجموعه $\mathcal{T} \in \mathbb{R}^+$ است.

تعریف ۵.۶.۱. میدان تصادفی فضایی‌ـ‌زمانی $Y(\cdot,\cdot)$ تفکیک‌پذیر نامیده می‌شود، اگر توابع کوواریانس فضایی و زمانی C_s و C_r وجود داشته باشند، به‌طوری که

$$\mathrm{Cov}(Y(s,t), Y(s',t')) = C_s(s,s')C_r(t,t'), \quad (s,t),(s',t') \in D \times \mathcal{T}.$$

برای یک میدان تصادفی فضایی‌ـ‌زمانی تفکیک‌پذیر، ماتریس کوواریانس $\boldsymbol{\Sigma}$ برابر حاصلضرب کرونه‌کری ماتریس کوواریانس فضایی $n \times n$ بعدی $\boldsymbol{\Sigma_s} = (\sigma_{ij})$ و ماتریس کوواریانس زمانی $m \times m$ بعدی $\boldsymbol{\Sigma_t}$ به صورت

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma_s} \otimes \boldsymbol{\Sigma_t} = \begin{bmatrix} \sigma_{۱۱}\boldsymbol{\Sigma_t} & \cdots & \sigma_{۱۱}\boldsymbol{\Sigma_t} \\ \vdots & \ddots & \vdots \\ \sigma_{n۱}\boldsymbol{\Sigma_t} & \cdots & \sigma_{nn}\boldsymbol{\Sigma_t} \end{bmatrix}$$

[۳۶] Intrinsic CAR (ICAR)
[۳۷] Seperable

است. با انتخاب یک مدل معتبر برای Σ_s که در بخش ساختار وابستگی فضایی معرفی کردیم و به طریق مشابه انتخاب یک مدل معتبر برای Σ_t از بخش ساختار وابستگی زمانی، می‌توان مدل‌های معتبر فضایی‌ـ‌زمانی به‌دست آورد (محمد زاده، ۱۳۹۱).

فصل ۲

تقریب لاپلاس آشیانه‌ای جمع‌بسته

مسأله کلیدی در استنباط بیزی، حل انتگرال‌های با بعد بالا است. یک روش قدیمی، تقریب لاپلاس است که تاریخ آن به لاپلاس (۱۷۷۴) برمی‌گردد. این ایده ساده، انتگرال را با بسط تیلور مرتبه دوم تابع زیر انتگرال تقریب می‌زند و آن را به‌صورت عددی حل می‌کند. با توسعه آشیانه‌ای این روش کلاسیک و ترکیب آن با روش‌های عددی مدرن برای ماتریس‌های تنک[۱]، شیوه تقریب لاپلاس آشیانه‌ای جمع‌بسته[۲] توسط رو و همکاران (۲۰۰۹) معرفی شد تا استنباط بیزی را برای مدل‌های گوسی پنهان[۳] تقریب بزند. امروزه LGMها مدل‌های بسیار مفیدی برای تحلیل بیزی به شمار می‌روند و اساس بسیاری از مدل‌های آماری را تشکیل می‌دهند.

در این رساله، سعی کرده‌ایم تا از قدرت و شگفتی‌های روش INLA بهره گرفته تا به برازش هرچه کاراتر و در عین حال سریع و ساده مدل‌های چندجمله‌ای بپردازیم. در این فصل به معرفی و نحوه به‌کارگیری این روش کارا پرداخته‌ایم.

۱.۲ مقدمه‌ای بر INLA

یک مانع کلیدی همیشگی بر سر راه آماردانان و پژوهشگرانی که آمار بیزی را برای تحلیل و مدل‌بندی داده‌های خود انتخاب می‌کنند، اجرای استنباط بیزی است. از دیدگاه ریاضی،

[۱] Sparse
[۲] Integrated nested Laplace approximation (INLA)
[۳] Latent Gaussian models (LGMs)

استنباط بیزی با اصول اولیه زیر تعریف می‌شود:

۱. به کمک اطلاعات موجود در داده‌های مشاهده‌شده، اطلاعات پیشین خود را نسبت به پارامترهای نامعلوم به روز می‌کنیم.

۲. توزیع پسین پارامترهای مدل را به‌دست می‌آوریم.

۳. آماره‌های مورد نیاز را بر اساس توزیع پسین مانند چندک‌ها، توزیع‌های پسین کناری، میانگین و واریانس محاسبه می‌کنیم.

در عمل، گفتن این اصول آسانتر از انجام آن است. با پیداش و توسعه عملی روش‌های نمونه‌گیری مونت‌کارلوی زنجیر مارکوفی[۴] (رابرت و کسلا، ۱۹۹۹) پیشرفت‌های چشمگیری در به‌کارگیری رویکرد بیزی در مسایل مختلف کاربردی ایجاد شد. این روش‌ها یک دستورالعمل کلی برای تولید نمونه‌هایی از توزیع پسین با ساختن زنجیر مارکوف و در نظر گرفتن پسین هدف به عنوان توزیع مانا عرضه می‌کنند. همچنین در سال‌های اخیر با پیدایش و توسعه ابزارهای نرم‌افزاری همچون WinBUGS (اسپیگل‌هاتر، ۱۹۹۵)، JAGS (پلامر، ۲۰۱۶) و Stan (تیم استان، ۲۰۱۵) بر توفیقات این روش‌ها برای استنباط بیزی افزوده شده است. از زمانی که استنباط بیزی به این روش‌ها مجهز شد، آمار بیزی رشد سریعی را در محبوبیت تجربه کرده است به‌طوری که آمار بیزی هم اکنون در بسیاری از رشته‌ها، شاخه‌های آماری و مجله‌های مربوط به آن‌ها به گستردگی خودنمایی می‌کند. با این وجود و به عقیده بسیاری از کاربران آمار بیزی این انقلاب در آمار بیزی، هنوز نیازمند توجه بیشتری است. چرا که به نظر می‌رسد همچنان با استنباط دست و پاگیری روبرو هستیم که به رایانه‌های پرسرعت و زمانی از کاربر که در رسیدن به مدل مناسب و کارا هدر می‌رود، نیازمند است. حتی اجرای مجدد بسیاری از مدل‌ها را دست و پاگیر و تکرار فرآیند انتخاب مدل را ناممکن کرده است (باکس و تیائو، ۱۹۷۳). بنابراین گاهی استنباط بیزی، به دلایل عملی نبودن به‌کارگیری آن، مورد استقبال قرار نمی‌گیرد و کنار گذاشته می‌شود. این در حالی است که روش تقریبی INLA، استنباط را در یک فاصله زمانی معقول انجام می‌دهد و در بسیاری از موارد از روش معمول MCMC سریعتر و کاراتر است. اگرچه INLA در عمل به مدل‌های LGM محدود است، اما این رده از مدل‌ها گستره بزرگی از مدل‌های کارآمد و پرمصرف را شامل می‌شوند و افزون بر آن، امروزه بسیاری از مدل‌های آماری را می‌توان براساس LGMها تعریف کرد.

قصه از آن‌جا آغاز می‌شود که رو و همکاران (۲۰۰۹) توانستند نشان دهند با استفاده از تقریب لاپلاس آشیانه‌ای تودرتو می‌توان چگالی‌های کناری پسین را با دقت خیلی خوبی محاسبه کرد. این روش روی مدل‌هایی تمرکز می‌کند که بردار متغیر پنهان در آن‌ها، یک میدان تصادفی گوسی مارکوفی[۵] است. ایده اصلی بر این اساس پایه‌گذاری شد که می‌توان انتگرال توزیع پسین را به‌صورت حاصلضرب آشیانه‌ای انتگرال‌های با بعد پایین نوشت و سپس

[۴] Markov chain Monte Carlo (MCMC)
[۵] Gaussian Markov random field (GMRF)

به‌صورت تقریبی اما با دقت بالا انتگرال‌ها را به‌صورت عددی حل کرد. روش INLA، در اغلب موارد، با دقتی عمل می‌کند که حتی اختلاف‌های بین نتایج حاصل از آن و نمونه‌گیری MCMC، جزئی و غیرقابل تشخیص هستند. فایده اصلی این تقریب از نظر محاسباتی این است که برآوردهای دقیقی را برای تحلیل‌های بیزی در چند ثانیه یا دقیقه انجام می‌دهد. برای مثال، در مدل‌های معمولی فضایی که بعد n چند هزار است، تقریبا این روش برای همه پسین‌های کناری در یک دقیقه یا چند دقیقه محاسبه می‌شود. در مقابل، الگوریتم‌های MCMC مربوط به آن‌ها ساعت‌ها یا روزها طول می‌کشد تا چگالی‌های کناری پسین را محاسبه کند. همچنین اریبی تقریب خیلی کمتر از خطای MCMC و در عمل ناچیز است. از جمله مهم‌ترین دلایلی که می‌توان برای گسترش فراوان استفاده از روش INLA بیان کرد، فراهم کردن تحلیل‌های بیزی کامل با اجرای سریع و کدنویسی راحت و سریع است بدون آن‌که نیاز به نوشتن الگوریتم‌های نمونه‌گیری باشد. اما شاید بتوان علت اصلی این گسترش را قابل اجرا بودن برای رده وسیعی از مدل‌ها (مدل‌های گوسی پنهان) دانست چرا که امروزه بسیاری از مدل‌های پویا و مدل‌های فضایی در این رده قرار می‌گیرند.

توسعه نرم‌افزاری این روش، نتیجه چندین سال تلاش و کار سخت بوده که توسط هاوارد رو و تیمش انجام شده است. طی سال‌های ۲۰۰۲ تا ۲۰۰۴، هاوارد به همراه هلد، شروع به درک اهمیت مدل‌هایی را کردند که می‌توان با INLA انجام داد. در سال ۲۰۰۵ آن‌ها کتابی را با عنوان

Gaussian Markov Random Fields: Theory and Applications

در اهمیت این گونه مدل‌ها نوشتند. در سال ۲۰۰۷، اولین اجرای روش INLA در محیط C انجام شد، اما نیاز به فایل‌های دستی ورودی داشت. در سال ۲۰۰۸ دانشجوی دکتری هاوارد به نام سارا مارتینو اولین نسخه آن را برای اجرا در نرم‌افزار R تهیه کرد. برای داشتن مروری بر مثال‌های به‌کارگرفته شده با INLA می‌توانید به سایت www.r-inla.org و مقاله مروری رو و همکاران (۲۰۱۷) مراجعه کنید. برای درک این روش ابتدا به تعریف مفاهیم زیر می‌پردازیم:

۱. تقریب لاپلاس

۲. مدل‌های گوسی پنهان

۳. میدان‌های تصادفی گوسی مارکوفی

۲.۲ تقریب لاپلاس

تقریب لاپلاس یک روش قدیمی در تقریب انتگرال‌ها به شمار می‌رود. فرض کنید هدف تقریب انتگرال

$$I_n = \int_{\chi} \exp(nf(x))dx$$

باشد در حالتی که $n \longrightarrow \infty$. فرض کنید $x_\circ$ نقطه‌ای باشد که $f(x)$ در آن بیشترین مقدار را دارد. بنابراین

$$\begin{aligned} I_n &\approx \int_\chi \exp\left(n(f(x_\circ) + \frac{1}{2}(x-x_\circ)^2 f''(x_\circ))\right)dx \\ &= \exp(nf(x_\circ))\sqrt{\frac{2\pi}{-nf''(x_\circ)}} = \tilde{I}_n. \end{aligned}$$

در این‌جا روش لاپلاس، ایده ساده اما قدرتمند تقریب تابع زیر انتگرال با چگالی توزیع گوسی را انجام می‌دهد. اگر $nf(x)$ را لگاریتم درستنمایی و x را پارامتر نامعلوم در نظر بگیریم و قضیه حد مرکزی وقتی $n \longrightarrow \infty$ برقرار باشد، آن‌گاه تقریب گوسی دقیق است. برای انتگرال‌های با بعد بالا، خطای تقریب $\mathrm{O}(n^{-1})$ است. بنابراین

$$I_n = \tilde{I}_n(1 + \mathrm{O}(n^{-1})).$$

این یک نتیجه خوب است، چرا که خطای نسبی از مرتبه n^{-1} در مقابل $n^{-\frac{1}{2}}$ در به‌کارگرفتن روش‌های معمول استنباط بر پایه شبیه‌سازی است.

تقریب لاپلاس یک ابزار کلیدی در زمان‌های قبل از عمومی شدن الگوریتم‌های MCMC به شمار می‌رود. فرض کنید بخواهیم توزیع کناری $\pi(\gamma_1)$ را از توزیع توأم $\pi(\gamma)$ به‌دست آوریم، که در آن $\gamma = (\gamma_1, \ldots, \gamma_n)'$. می‌توان نوشت:

$$\begin{aligned} \pi(\gamma_1) &= \frac{\pi(\gamma)}{\pi(\gamma_{-1}|\gamma_1)} \\ &\approx \frac{\pi(\gamma)}{\pi_G(\gamma_{-1}; \boldsymbol{\mu}(\gamma_1), \boldsymbol{Q}(\gamma_1))}|_{\gamma_{-1}=\boldsymbol{\mu}(\gamma_1)} \end{aligned} \tag{۱.۲}$$

در واقع $\pi(\gamma_{-1}|\gamma_1)$ با چگالی توزیع گوسی تقریب زده می‌شود که در آن $\pi_G(\cdot;\boldsymbol{\mu},\boldsymbol{Q})$ تقریب گوسی $\pi(\cdot)$ است. تیرنی و کادان (۱۹۸۶) نشان دادند اگر $\pi(\gamma) \propto \exp(nf_n(\gamma))$ یعنی اگر $f_n(\gamma)$ میانگین لگاریتم درستنمایی باشد، خطای نسبی تقریب (۱.۲) برابر است با $\mathrm{O}(n^{-\frac{3}{2}})$.

۱.۲.۲ مدل‌های گوسی پنهان

مدل‌های گوسی پنهان را به کمک چارچوب مدل‌های سلسله‌مراتبی توضیح می‌دهیم. با فرض استقلال مشاهدات $\boldsymbol{y} = (y_1, \ldots, y_n)'$ با تابع (چگالی) احتمال $p(\cdot|\boldsymbol{\theta}_1)$، به شرط میدان تصادفی گوسی تصادفی $\boldsymbol{x}$ و بردار ابرپارامتر $\boldsymbol{\theta}_1$، می‌توان نوشت

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) \sim \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}_1)$$

به‌طوری که

$$\boldsymbol{x}|\boldsymbol{\theta}_2 \sim \mathrm{N}(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)).$$

بردار $\boldsymbol{x}$ شامل همه قسمت‌های تصادفی مدل که نشان‌دهنده ساختار وابستگی موجود در مدل هستند، می‌شود. بنابراین درستنمایی دارای ابرپارامترهای $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ است. بنابراین، توزیع پسین مدل به‌صورت

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} p(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{\frac{1}{2}} \exp[-\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i=1}^{n} \log(p(y_i|x_i, \theta))]\end{aligned}$$

نتیجه می‌شود. توجه داشته باشید که $\boldsymbol{x}$ توزیع توام مربوط به همه پارامترهای موجود در پیشگوی خطی است و $\boldsymbol{\theta}$ بردار ابرپارامترهای مدل که لزوما دارای توزیع گوسی نیست. بنابراین، این مدل یک LGM است اگر و تنها اگر توزیع توام پارامترهای مدل گوسی باشد نه ابرپارامترها. در واقع مدل‌های گوسی پنهان همان مدل‌های STAR هستند که اثرات ساختاری آن‌ها گوسی است. در حالت کلی برای کارایی روش تقریب INLA در مدل‌های گوسی پنهان، باید شرایط زیر برقرار باشند:

۱. تعداد ابرپارامترها کم و معمولا بین ۲ تا ۵ است و نباید از ۲۰ بیشتر شود.

۲. توزیع میدان پنهان $\boldsymbol{x}|\boldsymbol{\theta}$ گوسی و با ویژگی مارکوفی است و n می‌تواند بزرگ بین 10^3 تا 10^5 باشد.

۳. داده‌های $\boldsymbol{y}$، به شرط داشتن $\boldsymbol{x}$ و $\boldsymbol{\theta}$، مستقل شرطی هستند. در نتیجه مشاهدات y_i به عناصر میدان پنهان یعنی x_i ها بستگی دارند.

این شرایط هم به دلایل محاسباتی و هم به خاطر ایجاد اطمینان بالا از داشتن یک تقریب دقیق، ضروری است.

۲.۲.۲ میدان‌های تصادفی گوسی مارکوفی

همان‌طور که پیشتر بیان شد، یکی از شرایط لازم برای کارایی روش INLA مارکوفی بودن میدان تصادفی $\boldsymbol{x}$ است. فرض کنید بردار تصادفی $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ دارای توزیع نرمال چندمتغیره با میانگین $\boldsymbol{\mu}$ و ماتریس دقت معین مثبت $\boldsymbol{Q}$ با تابع چگالی

$$\mathrm{p}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{Q}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{Q}|^{\frac{1}{2}} \exp\{(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})\}$$

است. بردار $\boldsymbol{x}$ دارای ویژگی مارکوفی است، هرگاه x_i و x_j به شرط سایر اعضای باقیمانده $\boldsymbol{x}_{-ij}$، برای تعداد کمی از $\{i, j\}$ها از یکدیگر مستقل باشند. اهمیت این ویژگی در آن است که اگر دو عضو x_i و x_j مستقل شرطی باشند، آن‌گاه درایه متناظر با آن‌ها در ماتریس دقت صفر خواهد بود. یعنی $Q_{ij} = 0$. نتیجه عملی مارکوفی بودن میدان تصادفی آن است که ماتریس دقت $\boldsymbol{Q}$

تنک خواهد شد و روش‌های بسیار کارایی برای تجزیه ماتریس‌های تنک وجود دارند (رو و هلد، ۲۰۰۵) که باعث افزایش سرعت محاسبات می‌شود.

بنابراین در عمل برای داشتن کارایی محاسباتی، میدان تصادفی پنهان نه تنها گوسی بلکه باید مارکوفی (تنک) هم باشد (رو و هلد، ۲۰۰۵؛ هلد و رو ۲۰۱۰). در واقع یک GMRF یک میدان گوسی با ویژگی اضافه‌تر استقلال شرطی است. شاید ساده‌ترین مثال از یک مدل مارکوفی، مدل اتورگرسیو مرتبه اول با معادله

$$x_t = \phi x_{t-1} + \epsilon_t, \quad t = 1, 2, \ldots, m,$$

باشد. در این مدل، همبستگی بین x_t و x_s برابر است با $\phi^{|s-t|}$ و ماتریس کوواریانس $m \times m$ چگال است. با این وجود، x_t و x_s برای $|s-t| > 1$ به شرط سایر xها از هم مستقل‌اند و بنابراین ماتریس دقت حاصل، یک ماتریس تنک سه قطری است. به همین دلیل است که در ادبیات INLA در میدان گوسی از ماتریس دقت برای نمایش توزیع استفاده می‌شود تا ماتریس کوواریانس.

در نظر گرفتن GMRF در محاسبات خیلی کاراتر و بهتر است زیرا در این صورت به جای یک ماتریس چگال که هزینه محاسباتی بالایی دارد، با یک ماتریس تنک سروکار داریم. در مثال اتورگرسیو، ماتریس دقت، یک ماتریس سه قطری است که تجزیه آن به زمان محاسباتی $\mathrm{O}(m)$ نیاز دارد. در حالی که در حالت کلی تجزیه یک ماتریس چگال $m \times m$ به محاسبات با مرتبه $\mathrm{O}(m^3)$ نیاز دارد. در مدل‌های با ساختار فضایی نیز زمان محاسبات به $\mathrm{O}(m^{3/2})$ و نیاز به حافظه به $\mathrm{O}(m\log(m))$ کاهش می‌یابد.

۳.۲ مراحل روش تقریب INLA

فرض کنید بردار $\boldsymbol{y} = (y_1, \ldots, y_n)'$، n مشاهده از توزیع پاسخ با خانواده توزیع‌های نمایی با میانگین μ_i باشد. همچنین فرض کنید پیشگوی جمعی–ساختاری مدل نسبت به میانگین، تابعی از متغیرهای تبیینی باشد. توجه داشته باشید که پیشگوی ساختاری می‌تواند تابعی از اثرات ساختاری مانند توابع ناخطی، اثر تصادفی و اثر فضایی نیز باشد، که این جملات بردار پنهان $\boldsymbol{x}$ را می‌سازند. توزیع متغیر پاسخ به بردار ابرپارامترهای $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ نیز وابسته است. مشاهدات پاسخ به شرط بردار متغیرهای پنهان $\boldsymbol{x}$ از یکدیگر مستقل هستند. بنابراین تابع درستنمایی مدل به‌صورت

$$\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i \in I} \mathrm{p}(y_i|x_i, \boldsymbol{\theta}),$$

است، که در آن $I \subseteq \{1, \ldots, n\}$ نشان‌دهنده اندیس مشاهدات پاسخ است. در استنباط بیزی، هدف به‌دست آوردن توزیع پسین پارامترهای مدل و ابرپارامترها است که روش به‌دست آوردن آن در INLA در سه گام اجرا می‌شود:

گام اول

برای توزیع شرطی کامل

$$\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) \propto \exp\{-\frac{۱}{۲}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in I}\log p(y_i|x_i,\boldsymbol{\theta})\},$$

با استفاده از بسط تیلور $\log p(y_i|x_i,\boldsymbol{\theta})$ حول مد توزیع شرطی کامل $\boldsymbol{x}$، یعنی

$$\boldsymbol{x}^*(\boldsymbol{\theta}) = (x_۱^*(\boldsymbol{\theta}),\ldots,x_{n_d}^*(\boldsymbol{\theta}))',$$

که در آن n_d بعد بردار پنهان $\boldsymbol{x}$ است، تقریب گوسی به صورت

$$\begin{aligned}\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) &\propto \exp(-\frac{۱}{۲}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in I} g_i(x_i)) \\ &\propto \exp(-\frac{۱}{۲}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} + \sum_{i\in I}(a_i + b_i x_i - \frac{۱}{۲}c_i x_i^۲),\end{aligned}$$

به‌دست می‌آید، که در آن $c_i = -g''(x_i^*(\boldsymbol{\theta}))$ و $b_i = g_i'(x_i^*(\boldsymbol{\theta})) + x_i^*(\boldsymbol{\theta})c_i$. جمله a_i را می‌توان از تناسب حذف کرد (رو و همکاران، ۲۰۰۵). سپس تقریب لاپلاس چگالی پسینی

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})},$$

با بسط مخرج کسر حول $\boldsymbol{x} = \boldsymbol{x}^*(\boldsymbol{\theta})$ به‌صورت

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

محاسبه می‌شود.

گام دوم

تقریب لاپلاس چگالی شرطی

$$\pi(x_i|\boldsymbol{\theta},\boldsymbol{y}) = \frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x_{-i}}|x_i,\boldsymbol{\theta},\boldsymbol{y})},$$

با بسط مخرج کسر حول $\boldsymbol{x}_{-i} = \boldsymbol{x}_{-i}^*(x_i,\boldsymbol{\theta})$ به‌صورت

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

محاسبه می‌شود، که در آن $\boldsymbol{x}_{-i}^*(x_i,\boldsymbol{\theta})$ مد چگالی $\pi(\boldsymbol{x}_i|x_i,\boldsymbol{\theta},\boldsymbol{y})$ و $\tilde{\pi}_{GG}(\boldsymbol{x}_i|x_i,\boldsymbol{\theta},\boldsymbol{y})$ تقریب گوسی $\pi(\boldsymbol{x}_i|x_i,\boldsymbol{\theta},\boldsymbol{y})$ است. همچنین $\boldsymbol{x}_{-i}$ بردار حاصل از حذف درایه iام $\boldsymbol{x}$ است.

گام سوم

در نهایت جمع‌های متناهی

$$\begin{aligned}\tilde{\pi}(x_i|\boldsymbol{y}) &= \sum_{k=۱}^{n_{p_۱}} \tilde{\pi}(x_i|\boldsymbol{\theta}_k,\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\triangle_k; \quad i=۱,\ldots,n_d,\\ \tilde{\pi}(\theta_j|\boldsymbol{y}) &= \sum_{q=۱}^{n_{p_۲}} \tilde{\pi}(\boldsymbol{\theta}_q|\boldsymbol{y})\triangle_{jq}; \quad j=۱,\ldots,\ell\end{aligned}$$

تقریب زده می‌شوند، که در آن $n_{p_۱}$ و $n_{p_۲}$ تعداد $\boldsymbol{\theta}$های انتخاب‌شده از تکیه‌گاه ابرپارامترها است، ℓ بعد بردار ابرپارامترها است و وزن‌های $\triangle_k$ و $\triangle_{jq}$ با استفاده از رو و همکاران (۲۰۰۹) به صورت

$$\begin{aligned}\triangle_k &= \{(n_{p_۱})(f_۰^۲-۱)[۱+\exp(\frac{-mf_۰^۲}{۲})]\}^{-۱}\\ \triangle_{jq} &= \{(n_{p_۲})(f_۰^۲-۱)[۱+\exp(\frac{-(m-۱)h_۰^۲}{۲})]\}^{-۱},\end{aligned}$$

محاسبه می‌شوند که در آن $f_۰$ و $h_۰$ هر مقدار ثابت بزرگتر از یک و m حاصل انتگرال $\boldsymbol{\theta}'\boldsymbol{\theta}$ است. در واقع این جمع‌ها به ترتیب تقریب چگالی‌های پسین کناری

$$\begin{aligned}\pi(x_i|\boldsymbol{y}) &= \int \pi(x_i|\boldsymbol{y},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \quad i=۱,\ldots,n_d,\\ \pi(\theta_i|\boldsymbol{y}) &= \int \pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}_{-j}, \quad j=۱,\ldots,\ell\end{aligned}$$

هستند، که در آن $\boldsymbol{\theta}_{-j}$ بردار حاصل از حذف درایه jام $\boldsymbol{\theta}$ است.

به‌طور خلاصه، رویکرد INLA بر اساس ساختار رده LGM طراحی شده است؛ بعد $\boldsymbol{\theta}$ پایین است؛ $\boldsymbol{x}|\boldsymbol{\theta}$ یک GMRF است و مشاهدات پاسخ مستقل شرطی هستند، به‌طوری که y_i تنها به x_i و $\boldsymbol{\theta}$ بستگی دارد. برای فهم بیشتر از فرآیند تقریب، مثال زیر از رو و همکاران (۲۰۱۷) را تشریح می‌کنیم.

مثال ۱.۳.۲. مدل

$$\begin{aligned}y_i|\lambda_i &\sim \text{Pois}(\lambda_i), \quad i=۱,\ldots,n,\\ \log(\lambda_i) &= \eta_i = g(\beta)u_{j(i)},\end{aligned}$$

را در نظر بگیرید، که در آن $\beta \sim \text{N}(۰,۱)$، $\boldsymbol{u} \sim \text{N}(۰,\boldsymbol{Q}^{-۱})$ و $g(\cdot)$ یک تابع یکنوا است. اندیس $j(i)$ طوری ساخته شده است که بعد $\boldsymbol{u}$ ثابت باشد و به n بستگی نداشته باشد و u_jها به همان تعداد دفعه مشاهده شده‌اند. محاسبه پسین کناری β و u_j با مشکل روبرو است، زیرا ضرب دو مدل گوسی و ناگوسی را داریم. راهکار INLA این است که این مسئله را به چند مسئله کوچکتر که تقریبا گوسی هستند، تبدیل کند و در نهایت تقریب لاپلاس را برای آن‌ها به‌کار

گیرد. ایده کلیدی در این جا، نوشتن توزیع شرطی است. برای این مسئله توزیع شرطی β به صورت

$$\pi(\beta|\boldsymbol{y}) \propto \pi(\beta) \int \prod_{i=1}^{n} p(y_i|\lambda_i) \times p(\boldsymbol{u})d\boldsymbol{u}, \tag{۲.۲}$$

است که در آن $\lambda_i = \exp(g(\beta)u_{j(i)})$. برای تقریب خوب انتگرال، باید تابع زیر انتگرال به چگالی گوسی نزدیک باشد. توزیع پسین کناری u_j نیز به صورت

$$\pi(u_j|\boldsymbol{y}) = \int \pi(u_j|\beta, \boldsymbol{y}) \times \pi(\beta|\boldsymbol{y})d\beta. \tag{۳.۲}$$

است. توجه داشته باشید که در این‌جا می‌توان انتگرال را به‌طور مستقیم حساب کرد، زیرا β یک بعدی است. مشابه با (۲.۲) می‌توان نوشت

$$\pi(\boldsymbol{u}|\beta, \boldsymbol{y}) \propto \prod_{i=1}^{n} p(y_i|\lambda_i) \times p(\boldsymbol{u}) \tag{۴.۲}$$

که باید نزدیک به چگالی گوسی باشد. تقریب زدن $\pi(u_j|\beta, \boldsymbol{y})$، همان تقریب انتگرال مربوط به $\pi(\boldsymbol{u}|\beta, \boldsymbol{y})$ اما با یک بعد کمتر است، زیرا u_j ثابت است که باز هم به چگالی گوسی نزدیک است.

یک نکته کلیدی این است که می‌توان مسئله را به سه مسئله ریزتر زیر تبدیل کرد:

۱. تقریب $\pi(\beta|\boldsymbol{y})$ به کمک (۲.۲).

۲. تقریب $\pi(u_j|\beta, \boldsymbol{y})$ برای همه jها و βهای دلخواه.

۳. محاسبه $\pi(u_j|\boldsymbol{y})$ با استفاده از ترکیب دو تقریب در مراحل ۱ و ۲ در ترکیب با (۴.۲).

اگرچه در ظاهر هزینه‌ای که پرداخت می‌کنیم افزایش می‌یابد و مسئله پیچیده‌تر می‌شود، زیرا باید مرحله دوم را برای تمام مقادیر βها حساب کنیم، اما با کمی تأمل درخواهیم یافت که با جایگزین کردن وابستگی‌های پیچیده با شرطی کردن و انتگرال‌های عددی، مسئله به تقریب لاپلاس توزیع‌هایی که به گوسی نزدیک هستند منتهی می‌شود.

حال یک سوأل مهم این است که آیا می‌توان این اصول را برای LGMها به‌کار گرفت به‌طوری که از دقت و کارایی لازم برخوردار باشد؟ پاسخ، مثبت است. می‌توان اصول بالا را برای مدل‌های LGM نیز به‌کار گرفت به‌طوری که β را با $\boldsymbol{\theta}$ و $\boldsymbol{u}$ را با $\boldsymbol{x}$ جایگزین کرد. تقریبی که در نتیجه آن به‌دست می‌آید، هر چند از محاسبات سریعی برخوردار است اما کمی دقت خود را از دست می‌دهد.

۴.۲ مثال‌های بیشتر برای مقایسه

برای مقایسه سرعت تقریب حاصل از روش INLA با سایر نرم‌افزارهای استنباط بیزی، دو مثال را تشریح می‌کنیم.

مثال ۱.۴.۲. در این مثال، مشاهدات پاسخ را از مدل رگرسیون خطی ساده

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = ۱, \ldots, n,$$

شبیه‌سازی کردیم، که در آن ϵ_i دارای توزیع نرمال استاندارد با دقت $\sigma^{-۱}$ است. متغیر تبیینی x را از یک توزیع نرمال با میانگین ۶ و واریانس ۲ تولید کردیم و مقادیر واقعی پارامترهای رگرسیونی با تولید از یک توزیع نرمال استاندارد، $\beta = -۰/۶۷۴$ و $\alpha = ۱/۰۹$ انتخاب شدند. مقدار واقعی $\sigma^{-۱}$ را نیز برابر با یک در نظر گرفتیم. برای مقایسه نتایج حاصل از رهیافت INLA، مدل رگرسیون بیزی را با همان پیشین‌های نرمال در روش INLA برای ضرایب رگرسیونی α و β و پیشین گاما برای پارامتر دقت، با ابرپارامترهای یکسان، به زبان برنامه‌نویسی JAGS در این نرم‌افزار معرفی و شبیه‌سازی از آن را اجرا کردیم. برآوردهای میانگین پسینی پارامترها توسط هر دو روش INLA و JAGS برای $n = ۱۰۰$ در جدول ۱.۲ گزارش شده‌اند. برآوردها با هر دو روش به مقادیر واقعی نزدیک هستند. مدت زمان اجرای مدل با دو روش برای اندازه‌های نمونه برابر با ۱۰۰, ۵۰۰, ۵۰۰۰, ۲۵۰۰۰ در جدول ۲.۲ گزارش شده‌اند. از روی نتایج این جدول، به روشنی سرعت بالا و کارایی محاسباتی روش INLA مشهود است.

جدول ۱.۲: نتایج برازش مدل رگرسیون ساده

	مقدار واقعی	JAGS	INLA
α	$-۰/۶۷۴$	$-۰/۶۶۹$	$-۰/۶۷۱$
β	۱/۰۹	۱/۰۰	۱/۰۶

جدول ۲.۲: مقایسه زمان اجرا برای مدل رگرسیون ساده

n	JAGS	INLA
۱۰۰	۴/۱۹	۰/۱۷۶
۵۰۰	۱۸/۱۴۱	۱/۲۴۳
۵۰۰۰	۳۸۱/۵۷۳	۲/۷۸۷
۲۵۰۰۰	۲۲۰۳/۶۷۹	۱۳/۲۷
۱۰۰۰۰۰	۸۸۷۳/۸۳۶	۵۲/۷۸۷

مثال ۲.۴.۲. معمولا برای نمایش همزمان وابستگی اندازه دور سر جنین به سن مادر و تفاوت بین رشد دور سر در جنین‌ها، از مدل عرض از مبدا و شیب تصادفی

$$y_{ft} = (\beta_۰ + u_{۰f}) + (\beta_۱ + u_{۱f})\text{tga} + e_{ft}, \quad u_{of} \sim N(۰, \epsilon), \quad u_{۱f} \sim N(۰, \epsilon_۱), \quad e_{ft} \sim N(۰, \sigma),$$

استفاده می‌کنند، که در آن y_{ft} پاسخ طولی (دور سر جنین) برای جنین f در زمان t و tga سن مقیاس‌بندی‌شده مادر است.

جدول ۳.۲: نمایی از ساختار داده‌های مربوط به مثال اندازه دور سر جنین

شماره مشاهده	id	y	ga	tga
۱	۱	۲۱۱	۲۳/۰۰	۱۶/۸۳
۲	۱	۲۷۴	۲۸/۴۳	۱۹/۰۰
۳	۱	۳۱۴	۳۳/۴۳	۲۰/۳۹
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
۳۰۹۵	۷۰۷	۲۹۲	۲۹/۵۴	۱۹/۳۰
۳۰۹۶	۷۰۷	۳۴۸	۳۷/۵۷	۲۱/۱۱

نمایی از ساختار داده‌ها در جدول ۴.۲ قابل مشاهده است، که در آن id کد شناسایی مادران، y اندازه دور سر جنین (پاسخ)، ga سن مادر و tga همان سن مادر است که برای راحتی برازش در ۱۹/۵۰ سال مرکزی شده است. برای دسترسی کامل به این مجموعه داده می‌توانید به لینک http://BendixCarstensen.com/Bayes/Cph-2012/data/fetal.csv مراجعه کنید. به منظور مقایسه روش INLA با یک روش مبتنی بر درستنمایی کلاسیک و MCMC، به کمک بسته‌های نرم‌افزاری lme4 و JAGS مدل را به داده‌ها برازش دادیم و نتایج را با نتایج حاصل از روش INLA مقایسه کردیم. برآوردهای متناظر برای اثر ثابت، پارامتر دقت اثرات تصادفی، انحراف استاندارد جمله باقیمانده و همبستگی بین شیب و عرض از مبدا، به ترتیب در جدول‌های ۴.۲، ۵.۲، ۶.۲ و ۷.۲ گزارش شده‌اند.

جدول ۴.۲: برآوردهای عرض از مبدأ و شیب در مدل

	lme4	JAGS	INLA
عرض از مبدا	۲۸۶/۲۷	۲۸۵/۵۸	۲۸۶/۲۹
tga	۲۶/۴۸	۲۶/۳۴	۲۶/۴۸

جدول ۵.۲: برآوردهای پارامترهای دقت اثرات تصادفی

	lme4	JAGS	INLA
عرض از مبدا	۷۴/۴۹	۷۶/۲۳	۶۷/۹۳
tga	۱/۲۱	۱/۱۵	۱/۰۰۲

جدول ۶.۲: برآورد پارامتر دقت مشاهدات

lme4	JAGS	INLA
۸/۵۸	۸/۶۳	۸/۶۴

جدول ۷.۲: برآورد همبستگی بین شیب و عرض از مبدأ

INLA	JAGS	lme4
۰/۷۸	۰/۷۶	۰/۷۴

در تمام نتایج، هر سه روش عملکرد مشابهی دارند و نمی‌توان تفاوت چشمگیری بین آن‌ها تعیین کرد. مدت زمان محاسبات برای JAGS و INLA به ترتیب ۱۱۴/۵۹ و ۱۴/۳ دقیقه است، که به روشنی کارایی روش INLA قابل درک است.

با همه ویژگی‌های خوبی که روش INLA دارد، متاسفانه به دلیل عدم امکان تعبیه هسته توزیع چندجمله‌ای در بسته R-INLA نمی‌توان پاسخ‌های رسته‌ای اسمی را با آن مدل‌بندی و تحلیل کرد. بنابراین، به کمک تبدیلی به نام تبدیل چندجمله‌ای‌ـ‌پوآسن[۶] (بیکر، ۱۹۹۴)، ابتدا هسته چندجمله‌ای را به پوآسن تبدیل کرده و سپس به کمک روش INLA به تحلیل پاسخ‌های رسته‌ای می‌پردازیم. در فصل چهارم این تبدیل را تشریح می‌کنیم.

در این فصل سعی کردیم تا نشان دهیم روش INLA یک تکنیک برای اجرایی کردن برازش مدل‌های پیچیده است و نه حتی یک روش شبیه‌سازی مانند MCMC و نه صرفا یک روش کدنویسی، بلکه روشی است که کارایی محاسباتی به همراه سادگی نسبی کدنویسی در آن موجب شده است تاکنون افراد زیادی برای به‌کارگیری مدل‌های فضایی یا فضایی‌ـ‌زمانی پیچیده یا هر مدل پیچیده دیگری که در چارچوب رده LGM قرار گیرد، با کاربردهای مختلف و فراوان در آن اقدام کنند.

[۶] Multinomial-Poisson transformation (MPT)

فصل ۳

شناسایی‌پذیری

شناسایی‌پذیری[۱] یکی از پیش‌نیازهای اساسی در مدل‌های آماری به شمار می‌رود و استنباط بدون شناسایی‌پذیری بی‌معنی است. بنابراین، لازم است قبل از برازش مدل شناسا بودن آن را مورد بررسی قرار دهیم. در این فصل مفهوم شناسایی‌پذیری در مدل‌ها را بررسی و تشریح می‌کنیم. پس از درک مفاهیم و شناخت راه‌حل‌های پیش رو برای رفع مشکل عدم شناسایی‌پذیری، روش پیشنهادی برای رفع آن در مدل‌های چندجمله‌ای با پیشگوی جمعی‌ـ ساختاری را در مدل مشابه آن به چالش می‌کشیم. در انتها برخی نتایج نظری را بیان می‌کنیم.

۱.۳ مقدمه

استنباط آماری به ما می‌آموزد تا ’’چگونه’’ از داده‌ها بیاموزیم در حالی که تحلیل شناسایی ’’چیزی’’ را که ما می‌توانیم از داده‌ها بیاموزیم، توضیح می‌دهد. اگرچه از نظر منطقی ’’چه چیزی’’ بر ’’چگونه’’ مقدم است، اما متاسفانه مفهوم شناسایی‌پذیری توجه نسبتاً کمی را در جامعه آماری به خود اختصاص داده است و ما باید نسبت به مفاهیمی از آمار که به امکان‌پذیر بودن تشخیص منحصر به فرد پارامترها توسط توزیع یا متغیر تصادفی مشاهده‌شده می‌پردازند، توجه بیشتری داشته باشیم (باسه و بوجینو، ۲۰۲۰). مسئله شناسایی‌پذیری در بسیاری از علوم مانند اقتصادسنجی و زیست‌شناسی که افراد در آن مجبور به مدل‌بندی ساختارهای

[۱] Identifiability

موجود در طبیعت هستند، مطرح است. این گوناگونی و گستردگی موجب شده است تا بسیاری از افرادی که رشته اصلی آن‌ها آمار نبوده، در پی نیاز به مدل‌بندی، به توضیح و مفهوم‌سازی مناسب برای این مسئله روی آورند. تاریخ اولین پژوهش‌ها در این زمینه را به دانشمندان علم اقتصاد به حداقل قبل از سال ۱۹۳۰ نسبت می‌دهند. اولین بار کوپمن (۱۹۴۹) مفهوم شناسایی‌پذیری را مطرح کرد. بعد از آن هرویچ (۱۹۵۰) و کوپمن و ریرسول (۱۹۵۰) به تعریف ریاضی این مفهوم پرداخته و نظریه مربوط به آن را در مدل‌های آماری توسعه دادند. به دنبال آن افرادی از رشته‌های مختلف برای مدل‌بندی سیستم‌های بیولوژی، بوم‌شناسی، مدل‌های پارامتری و ناپارامتری مطرح شدند. این افراد به ترتیب شامل گودمن (۱۹۵۹)، رتنبرگ (۱۹۷۱)، هیسایو (۱۹۸۳)، جاکز و پری (۱۹۹۰)، پولانیو و دپراگانکا بریرا (۱۹۹۴)، کروس و مانسکی (۲۰۰۲)، ماتزکین (۲۰۰۷) و مانسکی (۲۰۰۹) است. این واگرایی و فقدان نظریه یکپارچه، موجب شده است تا این بخش از علم آمار توسعه جدی را به خود نبیند.

در منظر علم آمار، مدل مفید بر سه پایه استوار است (وایلند و همکاران، ۲۰۲۱). یک مدل خوب در درجه اول باید با یک دقت منطقی داده‌ها را توصیف کند. دوم این که باید قدرت پیش‌گویی بالایی داشته باشد. مدل‌هایی که دارای این دو ویژگی باشند، خوب محسوب می‌شوند. در درجه سوم یک مدل باید توانایی برازش شدن به سیستم‌های بیولوژیکی را داشته باشد.

در یک فرآیند مدل‌بندی، معمولا افراد از ساده‌ترین مدل شروع می‌کنند که توانایی توصیف داده‌ها را ندارد و بد تلقی می‌شود. افزودن همزمان پیچیدگی به مدل به‌گونه‌ای که قابل برازش باشد و توانایی توصیف داده را داشته باشد، نیازمند توجه به مسئله قابل شناسایی بودن مدل است. یعنی مدل هم از قدرت پیشگویی بالایی برخوردار بوده و هم توانایی تعیین پارامترها به کمک داده‌ها را داشته باشد که همان شناسایی‌پذیری مدل است. مفهوم شناسایی‌پذیری با قدرت زیاد به انتقال مدل از بد به خوب مربوط می‌شود. به عبارتی دیگر، تحلیل شناسایی برای داشتن یک مدل خوب، ضروری است.

یکی از اهداف ما در این رساله افزایش سطح شناساگر بودن مدل لوجیت چندجمله‌ای است، به‌گونه‌ای که همزمان از ساختارهای پیچیده فضایی و فضایی‌ـ‌زمانی در یک چارچوب بیزی به کمک روش INLA بهره گرفته باشد.

۲.۳ شناسایی‌پذیری مدل

فرض کنید مدل آماری (پارامتری) $(\mathcal{Y}, \wp, \mathrm{P})$ را داشته باشیم که در آن $\mathcal{Y}$ فضای نمونه، $\wp$ سیگما‌ـ‌میدان تعریف‌شده روی آن و P خانواده‌ای از اندازه احتمال روی $(\mathcal{Y}, \wp)$ است که با $\mathrm{P} = \{\mathrm{P}_\vartheta : \vartheta \in \Theta\}$ مشخص شده و Θ فضای پارامتر است.

تعریف ۱.۲.۳. (پائولینو و پریرا، ۱۹۹۴). دو نقطه از Θ از نظر مشاهده معادل هستند، اگر به

ازای هر $A \in \wp$، $\mathrm{P}_{\vartheta_0}(A) = \mathrm{P}_{\vartheta_1}(A)$. در این صورت می‌توان نوشت $\vartheta_0 \sim \vartheta_1$.

تعریف ۲.۲.۳. (اسوارتز و همکاران، ۲۰۰۴). فرض کنید $F = \{F_\theta,\ \theta \in \Theta\}$ خانواده‌ای از توابع توزیع باشد. همچنین فرض کنید A متغیر تصادفی قابل مشاهده[۲] از این خانواده باشد. اگر حداقل یک جفت (θ, θ')، $\theta \neq \theta'$ از Θ وجود داشته باشد که به ازای هر a از تکیه‌گاه متغیر A، $F_\theta(a) = F_{\theta'}(a)$، می‌توان گفت θ با A قابل شناسایی نیست.

در واقع، بنا به تعریف ۲.۲.۳، در این حالت تابع درستنمایی برای همه داده‌ها ثابت است.

تعریف ۳.۲.۳. (اسوارتز و همکاران، ۲۰۰۴). اگرچه مفهوم نزدیک ناشناس بودن[۳] از قدرت کمتری برخوردار است، ولی در اشاره به تعریف قبل می‌توان گفت θ تقریبا ناشناس است اگر برای همه مقادیر m در تکیه‌گاه متغیر M، $F_\theta(m) \approx F_{\theta'}(m)$.

در یک کلام می‌توان گفت ناشناسا بودن مدل به داده‌ها اجازه نمی‌دهد تا تفاوتی بین مقادیر پارامترها تشخیص داده شود.

مثال ۱.۲.۳. فرض کنید X دارای توزیع نرمال با میانگین $\mathrm{E}(X) = \mu_1 - \mu_2$ باشد. بنابراین $\mu_1 - \mu_2$ را می‌توان با مشاهداتی از X برآورد کرد، اما پارامترهای μ_1 و μ_2 به‌طور منحصر به فرد قابل شناسایی و برآورد نیستند. در حقیقت می‌توان به تعداد نامتناهی از جفت‌های

$$\{(\mu_i, \mu_j),\ \ i, j = 1, 2, \ldots (i \neq j),\}$$

فکر کرد به‌گونه‌ای که $\mu_i - \mu_j = \mu_1 - \mu_2$. فقط در صورتی می‌توان μ_1 و μ_2 را به‌طور منحصر به فرد برآورد کرد که قابل شناسایی باشند.

همان‌طور که در مثال بالا قابل مشاهده است، در بسیاری از مسائل آماری فرض می‌شود توزیع آماری مورد مطالعه به‌جز برای مجموعه‌ای از پارامترها کاملا معلوم است. بدیهی است قبل از هر گونه استنباط باید مدل از نظر شناسایی‌پذیری مورد بررسی قرار گیرد. در این صورت مطمئن می‌شویم که متغیر تصادفی مورد نظر از توزیع‌های متفاوت با پارامترهای متفاوت پیروی نمی‌کند. گاهی اوقات ممکن است پارامتر θ خود قابل شناسایی نباشد، اما یک تابع ناثابت از آن مانند $\varphi(\theta)$ قابل شناسایی باشد. یعنی برای هر θ و θ' که متعلق به Θ باشد، به ازای تمام مقادیر a از تکیه‌گاه A، $F_\theta(a) = F_{\theta'}(a)$ ما را به $\varphi(\theta) = \varphi(\theta')$ می‌رساند. در این صورت می‌توان گفت θ به‌طور جزئی شناسایی‌پذیر[۴] است. در مثال قبل F_θ یک خانواده از توزیعهای نرمال، $\theta = (\mu_1, \mu_2)$ و $\varphi(\theta) = \mu_1 - \mu_2$ است.

برای تبدیل θ به یک متغیر شناسایی‌پذیر، می‌توان متغیر تصادفی Z را به مدل اضافه کرد به‌طوری که θ توسط متغیر تصادفی تقویت‌شده (A, Z) قابل شناسایی باشد. در این میان خود

[۲] Observable
[۳] Near nonidentifiability
[۴] Partially identifiable

متغیر تصادفی Z ممکن است نتواند θ را تشخیص دهد. در این صورت مسئله ناشناسا بودن قابل اصلاح[۵] است. در مثال زیر این موضوع را تشریح می‌کنیم.

مثال ۲.۲.۳. فرض کنید $X_۱$ و $X_۲$ دو متغیر تصادفی مستقل از توزیع نمایی با تابع توزیع

$$F_i(x) = ۱ - \exp(-\lambda_i x), \quad \lambda_i > ۰, \quad x > ۰ \; i = ۱, ۲,$$

باشند. به سادگی می‌توان دید $U = \min(X_۱, X_۲)$ دارای توزیع نمایی با پارامتر $\lambda_۱ + \lambda_۲$ است. در این‌جا $\lambda_۱$ و $\lambda_۲$ با U قابل شناسایی نیستند، زیرا تعداد نامتناهی از جفت متغیرهای تصادفی مستقل با توزیع نمایی وجود دارند که مجموع پارامترهای آن‌ها می‌تواند $\lambda_۱ + \lambda_۲$ باشد. در مقابل، $\theta = (\lambda_۱, \lambda_۲)$ به‌طور جزئی قابل شناسایی است، زیرا $\varphi(\theta) = \lambda_۱ + \lambda_۲$ همیشه قابل شناسایی است.

برمن (۱۹۶۳) نشان داد اگر X_iها به‌طور مستقل توزیع شده باشند و بتوان می‌نیمم را شناسایی کرد، آن‌گاه می‌توان توابع توزیع F_i را به‌طور منحصر به فرد بر اساس توابع یکنوا

$$\mathrm{H}_k(x) = \mathrm{P}(U \leqslant x, I = k), \quad k = ۱, ۲, \ldots, l,$$

که در آن I پارامتر اضافه است و شناسایی‌پذیری مدل را اصلاح می‌کند، تعیین کرد. به‌عنوان مثال، می‌توان قالب

$$\mathrm{F}_k(x) = ۱ - \exp\Big\{ - \int_{-\infty}^{x} \Big(۱ - \sum_{j=۱}^{l} H_j(t) \Big)^{-۱} dH_k(t) \Big\}, \quad k = ۱, ۲, \ldots, l,$$

را در نظر گرفت. نتایج برمن در مثال قبل نشان می‌دهد که مسئله ناشناسا بودن وقتی از (U, I) استفاده کنیم، قابل اصلاح است.

اگر پارامترها برآورد نشوند یا یک راه منحصر به فرد برای برآورد آن‌ها وجود نداشته باشد، آن‌گاه گوییم مسئله ناشناسا[۶] است. اگر یک راه منحصر به فرد برای شناسایی پارامترها وجود داشته باشد، آن‌گاه گوییم پارامترها شناسایی[۷] می‌شوند. اگر پارامترها در بیشتر از یک راه مستقل شناسایی شوند، گوییم پارامترها بیش‌شناسا[۸] هستند. در استنباط کلاسیک، مدل‌های زیادی را می‌توان یافت که نیاز به رفع ناشناسایی دارند. در زیر به مثالی از یک مدل کلاسیک در اسوارتز (۲۰۰۴) اشاره می‌کنیم.

مثال ۳.۲.۳. مدل طرح عاملی یک‌طرفه

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = ۱, \ldots, k, \quad j = ۱, \ldots, n_i,$$

[۵] Rectifiable
[۶] Unidentified or under identified
[۷] Just identified
[۸] Over identified

را در نظر بگیرید، که در آن، μ میانگین کلی، α_iها سطوح عاملی و ϵ_{ij} مولفه خطا است که از توزیع $\mathrm{N}(۰, \sigma^۲)$ پیروی می‌کند. بردار پارامتر $(\mu, \alpha_۱, \ldots, \alpha_k, \sigma)$ قابل شناسایی نیست زیرا برای همه مقادیر ثابت و معلوم γ، دو بردار $(\mu, \alpha_۱, \ldots, \alpha_k, \sigma)$ و $(\mu - \gamma, \alpha_۱ - \gamma, \ldots, \alpha_k - \gamma, \sigma)$ توزیع احتمال یکسانی از داده‌ها را نتیجه می‌دهند.

۳.۳ شناسایی‌پذیری و اهمیت آن در استنباط بیزی

در سال‌های اخیر، پیشرفت در ساختارهای نرم‌افزاری و کامپیوتری و امکان برازش مدل‌های مختلف آماری موجب شده است تا مدل‌های پیچیده با استقبال بیشتری از سوی آماردانان و کاربران آن‌ها روبه رو شوند. نتیجه این امر آن است که ممکن است برخی مدل‌ها به قدر کافی درک نشده و ناشناسایی به‌طور ناخواسته وارد مدل و باعث خسارت شود.

رویارویی استنباط بیزی با مسئله شناسایی‌پذیری مدل را می‌توان از دو جنبه فلسفی و کاربردی مورد توجه قرار داد. نگرانی که موجب می‌شود تا استنباط بیزی از نظر کاربردی در چالش ناشناسا بودن یا نزدیک ناشناسا بودن قرار گیرد، گرایش شدید پارامترهای توزیع پسین به یکدیگر است. در واقع، اهمیت شناسایی‌پذیری در استنباط بیزی از آن‌جا آغاز می‌شود که با حضور مسئله شناسایی‌پذیری، تمایل به داشتن همبستگی شدید بین پارامترها در توزیع پسین به‌طور چشمگیری افزایش می‌یابد و در نتیجه زنجیر MCMC در یک زمان منطقی با همگرایی ضعیف همراه می‌شود یا به همگرایی نمی‌رسد (اسوارتز، ۲۰۰۴). نگرانی دوم این است که در صورت وجود مسئله شناسایی‌پذیری حتی با نمونه با اندازه بزرگ، توانایی غلبه بر اطلاعات پیشین به‌دست نمی‌آید (جانسون، ۲۰۰۱). علاوه بر این، مسئله شناسایی‌پذیری را در مدل‌های سلسله‌مراتبی نیز می‌توان یافت (باسو، ۱۹۸۳). در زیر مثالی از اسوارتز (۲۰۰۴) ذکر شده است که به فهم بیشتر ما از این مسئله کمک می‌کند.

مثال ۳.۳.۱. فرض کنید $\boldsymbol{Y} = (Y_۱, \ldots, Y_n)'$ نمونه‌ای تصادفی از توزیع نرمال با میانگین μ و واریانس ۱ باشد، به‌طوری که $\mu|\alpha \sim \mathrm{N}(\alpha, ۱)$ و $\alpha \sim \mathrm{N}(\alpha_۰, ۱)$، که در آن $\alpha_۰$ معلوم است. هدف، تعیین توزیع پسین μ و α است. با انجام عملیاتی جبری، می‌توان نشان داد توزیع توام پسین (μ, α) نرمال دومتغیره با بردار میانگین $(\frac{۲n\bar{y}+\alpha_۰}{۲n+۱}, \frac{n\bar{y}+(۲n+۱)\alpha_۰}{۲n+۱})'$ و ماتریس کوواریانس با عناصر $\sigma_{۱۱} = \frac{۲}{۲n+۱}$، $\sigma_{۱۲} = \sigma_{۲۱} = \frac{۱}{۲n+۱}$ و $\sigma_{۲۲} = \frac{n+۱}{۲n+۱}$ است. در این‌جا $\bar{y}$ میانگین مشاهدات است. این مدل سلسله‌مراتبی شناسایی‌پذیر نیست، زیرا برای هر $\alpha \neq \alpha'$، $F(\mu, \alpha|\boldsymbol{y}) = F(\mu, \alpha'|\boldsymbol{y})$ که در آن $\boldsymbol{y}$ بردار مشاهدات پاسخ و $F(\cdot, \cdot|\boldsymbol{y})$ در این‌جا تابع توزیع پسین توام (μ, α) است.

یکی از دلایل اهمیت ناشناسایی این است که همیشه پارامتربندی منحصر به فرد مدل‌ها برای داشتن قابلیت تفسیرپذیری و برآوردی سازگار، امری ضروری محسوب می‌شود.

۴.۳ حل مسئله شناسایی‌پذیری

فرض کنید یک مدل آماری خاص با p پارامتر نامعلوم داریم که به نظر می‌رسد برای داده‌های ما مناسب باشد. اما این مدل قابل شناسایی نیست، به این معنی که مقادیر چندگانه‌ای برای بردار پارامتر با توزیع یکسان برای داده‌های مشاهده‌شده وجود دارند. افراد در طول زمان از روش‌های مختلفی برای حل این مسئله در علم آمار استفاده کرده‌اند. در زیر به برخی از آن‌ها اشاره می‌کنیم.

۱. یکی از راه حل‌های متداول این مشکل در استنباط بیزی، انتخاب توزیع پیشین است (اسوارتز، ۲۰۰۴). اما این راه حل محکمی نیست، زیرا در این صورت نه تنها داده‌ها بلکه توزیع پیشین نیز توانایی شناسایی پارامترهای مدل را ندارند. از طرفی، افراد در این روش به انتخاب توزیع‌های پیشین خاص محدود می‌شوند.

۲. یک راه حل استاندارد به اعتقاد تحلیل‌گرانی همچون گستافسون (۲۰۰۴)، استفاده از قیدهای شناسایی مناسب است. با قیدبندی کردن پارامترها، مدل با قدرت شناسایی بالا به‌دست می‌آید (گاس و همکاران، ۲۰۲۰). به عبارتی، در این حالت می‌توان فضای پارامتر شدنی Ω را به $\tilde{\Omega} \subset \Omega$ برای همه $\Theta \subset \Omega$ محدود کرد. در این رساله، هدف ما پیدا کردن قید مناسب برای شناسایی مدل لجیت چندجمله‌ای است.

۳. راه دیگر استفاده از قضیه زیر است.

قضیه ۱.۴.۳. (پائولینو و پریرا، ۱۹۹۴). فرض کنید $\psi(\vartheta)$ تابع شناساگر باشد. تابع $\lambda(\vartheta)$ قابل شناسایی است اگر و تنها اگر تابعی از $\psi(\vartheta)$ باشد.

ما در این رساله، راه حل دوم، قید مجموع مساوی با صفر[۹] را برای رهایی از مشکل شناسایی‌پذیری در مدل لوجیت چندجمله‌ای پیشنهاد می‌کنیم و نتایج حاصل را با قید متداول رده پایه[۱۰] یا رده گوشه[۱۱] مقایسه می‌کنیم.

۱.۴.۳ تاریخچه استفاده از قیدها در آمار

استنباط آماری بر اساس قیدها در طول دهه‌های گذشته مسیر رو به جلو را طی کرده است که می‌توان آغاز آن در سال‌های اولیه را به کودو (۱۹۶۳) و میروین (۱۹۹۴) نسبت داد. بارمی و دیکسترا (۱۹۹۸) روش قید مدل باکس کانوکس بر روی مدل لوجیت چندجمله‌ای را پیشنهاد کردند. لین (۱۹۹۷) روش قید روی عناصر واریانس در مدل خطی تعمیم‌یافته را آزمود. هال و

[۹] Sum to zero
[۱۰] Baseline category
[۱۱] Corner point category

پریستارت (۲۰۰۱)، قیدهای ترتیبی را بر روی مدل‌های خطی تعمیم‌یافته و ناخطی تعمیم‌یافته به‌کار گرفتند. همچنین بسیاری از افراد روش‌های متفاوت قیدبندی را در شاخه‌های مختلف علم آمار به‌کار گرفتند که در نتیجه آن دقت پیشگویی در مدل‌ها به‌طور جدی بالا رفت (سید، ۲۰۲۰). پیشرفت توام با نیاز روزانه به مدل‌های این چنینی موجب شده است تا افراد مختلفی وارد حوزه استنباط آماری با قیدها شوند. از جمله آن‌ها می‌توان به گاس و همکاران (۲۰۲۰)، سید (۲۰۲۰) و وینالد و همکاران (۲۰۲۱) اشاره کرد.

علی‌رغم پیشرفت عمده در علم داده‌ها و هوش مصنوعی و در نتیجه نیاز به پیش‌بینی دقیق رده چه از منظر پاسخ و چه از منظر احتمال، کمتر کسی را می‌توان یافت که در قیدبندی مدل لوجیت کار کرده باشد. برای مثال، می‌توان به سید (۲۰۲۰) در رهیافت کلاسیک در قید بر روی پارامترهای احتمال مدل لجیت با هدف ارتقاء مدل برای هوش مصنوعی اشاره کرد. ما در این رساله با بهره گرفتن از قیدها، علاوه بر رفع مشکل شناسایی‌پذیری در مدل لجیت، به این بخش از استنباط آماری نیز توجه کرده‌ایم. اگرچه استفاده از قیدها نیازمند به‌کارگیری مدل‌های پیچیده‌تر از نظر محاسباتی و زمان اجرا در برنامه‌های کدنویسی است، اما تجربه محققان مختلف در این زمینه نشان داده است که به‌کار گرفتن قیدها موجب افزایش کارایی و دقت در نتایج به‌دست آمده شده است (سید، ۲۰۲۰).

در این مرحله، بحث را درباره عملکرد دو نوع پارامتربندی، بر اساس قیدهای مجموع مساوی با صفر و رده گوشه، در مدل لوجیت چندجمله‌ای پی می‌گیریم به‌گونه‌ای که یک مدل با کیفیت مشابه با مدل لوجیت چندجمله‌ای که مشکل شناسایی‌پذیری در آن‌ها نیز یکسان باشد را با روش‌های نامبرده مورد بررسی قرار می‌دهیم.

در قید رده پایه، معمولا افراد پارامترهای مربوط به یکی از رده‌ها (رده پایه) را برابر صفر قرار می‌دهند و در قید مجموع مساوی با صفر، که پیشنهاد ما در این رساله است، حاصل جمع پارامترهای مد نظر برای رده‌ها برابر صفر در نظر گرفته می‌شود. عملکرد این دو نوع قید برای مدل چندجمله‌ای در فصل چهارم به‌طور کامل تشریح شده است.

۵.۳ مدل نرمال با ویژگی ناشناسایی مشابه با مدل لوجیت چندجمله‌ای

در این بخش برای ارزیابی و یافتن قید مناسب، سعی کردیم قیدهای مد نظر را در یک مدل نرمال که مشکل شناسایی در آن مشابه مدل چندجمله‌ای است، به‌کار بگیریم. یک توزیع پیوسته مانند توزیع نرمال که در آن پاسخ‌ها به صورت رده‌ای مرتب شده باشند، مشکل شناسایی‌پذیری در آن مشابه مشکل شناسایی‌پذیری در مدل لوجیت چندجمله‌ای است (اگرستی، ۲۰۰۲). بر این اساس، سعی کردیم تا با یک مدل شبیه‌سازی‌شده، عملکرد انواع قیدها را بررسی کرده و در حد امکان نتایج محکمی را ارائه کنیم.

فرض کنید یک توزیع نرمال با میانگین عاملی در سه رده داشته باشیم. فرض کنید $\boldsymbol{Y}$ یک متغیر تصادفی نرمال با میانگین $\boldsymbol{Z\beta}$ است، که در آن $\boldsymbol{Z}$ یک متغیر عاملی با سه سطح، و $\boldsymbol{\beta} = (\beta_۰, \beta_۱, \beta_۲, \beta_۳)'$ بردار پارامترهای نامعلوم مدل با مولفه $\beta_۰$ به‌عنوان عرض از مبدا است. بنابراین $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{Z\beta}, \boldsymbol{Q}_y^{-۱})$ که در آن، $\boldsymbol{Q}_y = \tau_y I$ ماتریس دقت مدل است. همچنین توزیع نرمال با میانگین صفر و ماتریس دقت $\boldsymbol{Q}_\beta = \tau_\beta I$ را به‌عنوان توزیع پیشین برای $\boldsymbol{\beta}$ در نظر می‌گیریم. هدف ما به‌دست آوردن توزیع پسین $(\beta_۱, \beta_۲, \beta_۳)'$ است. به همین منظور، برای به‌دست آوردن توزیع پسین، ابتدا قید رده پایه یا رده گوشه را در نظر می‌گیریم. برای این کار فرض می‌کنیم $\eta_{ij} = \beta_۰ + \beta_j$ پیشگوی خطی مربوط به میانگین آزمودنی i و رده j باشد، به‌طوری که در آن $i = ۱, \ldots, n$ و $j = ۱, ۲, ۳$. حال فرض کنید $\eta^*_{ij} = \beta_۰ + \beta^*_j$، به‌طوری که برای اعمال قید، قرار می‌دهیم $\beta^*_۳ = ۰$. با انجام محاسباتی، می‌توان نوشت

$$\boldsymbol{Z}_۱\boldsymbol{\beta^*} = \begin{bmatrix} ۱ & ۱ & ۰ \\ ۱ & ۰ & ۱ \\ ۱ & ۰ & ۰ \end{bmatrix} \begin{bmatrix} \beta_۰ \\ \beta^*_۱ \\ \beta^*_۲ \end{bmatrix}.$$

توزیع $\boldsymbol{\beta^*} \sim \mathrm{N}(۰, \boldsymbol{Q}^{-۱}_{\beta^*})$ را به‌عنوان توزیع پیشین $\boldsymbol{\beta^*}$ می‌پذیریم، که در آن $\boldsymbol{Q}_{\beta^*} = \tau_\beta I_{۳\times ۳}$. بنابراین می‌توان نوشت

$$\begin{aligned} \pi(\boldsymbol{\beta^*}|y) &\propto |\boldsymbol{Q}_y|^{۱/۲} e^{\frac{-۱}{۲}(y-\boldsymbol{Z}_۱\boldsymbol{\beta^*})'\boldsymbol{Q}_y(y-\boldsymbol{Z}_۱\boldsymbol{\beta^*})} |\boldsymbol{Q}_{\beta^*}|^{۱/۲} e^{\frac{-۱}{۲}\boldsymbol{\beta^*}'\boldsymbol{Q}_{\beta^*}\boldsymbol{\beta^*}} \\ &\propto e^{\frac{-۱}{۲}[\boldsymbol{\beta^*}-(\boldsymbol{Z}_۱'\boldsymbol{Q}_y\boldsymbol{Z}_۱+\boldsymbol{Q}_{\beta^*})^{-۱}\boldsymbol{Z}_۱'\boldsymbol{Q}_y y]'(\boldsymbol{Z}_۱'\boldsymbol{Q}_y\boldsymbol{Z}_۱+\boldsymbol{Q}_{\beta^*})^{-۱}[\boldsymbol{\beta^*}-(\boldsymbol{Z}_۱'\boldsymbol{Q}_y\boldsymbol{Z}_۱+\boldsymbol{Q}_{\beta^*})^{-۱}\boldsymbol{Z}_۱'\boldsymbol{Q}_y y]}. \end{aligned}$$

در نتیجه $\boldsymbol{\beta^*} = (\beta_۰, \beta^*_۱, \beta^*_۲)'$ دارای توزیع پسین نرمال با میانگین $(\boldsymbol{Z}_۱'\boldsymbol{Q}_y\boldsymbol{Z}_۱+\boldsymbol{Q}_{\beta^*})^{-۱}\boldsymbol{Z}_۱'\boldsymbol{Q}_y y$ و ماتریس دقت $(\boldsymbol{Z}_۱'\boldsymbol{Q}_y\boldsymbol{Z}_۱+\boldsymbol{Q}_{\beta^*})$ است. با انجام محاسبات ساده، میانگین برای رده‌ها به‌صورت

$$\beta_۱ = \frac{-۱}{۳}\beta^\star_۱ + \frac{-۱}{۳}\beta^\star_۲, \quad \beta_۲ = \frac{۲}{۳}\beta^\star_۱ + \frac{-۱}{۳}\beta^\star_۲, \quad \beta_۳ = \frac{-۱}{۳}\beta^\star_۱ + \frac{۲}{۳}\beta^\star_۲.$$

به‌دست می‌آید. در گام دوم برای به‌کار گرفتن قید مجموع مساوی با صفر باید $\sum_{m=۱}^{۳}\beta_m = ۰$ را به مدل وارد کنیم. در واقع ما به توزیع پسین $\boldsymbol{\beta}$ تحت قید خطی $\boldsymbol{A\beta} = ۰$ نیاز داریم، که در آن $\boldsymbol{A} = (۰, ۱, ۱, ۱)$ و در نتیجه $\boldsymbol{A\beta} = \beta_۱ + \beta_۲ + \beta_۳ = ۰$. توزیع شرطی $\boldsymbol{\beta}$ به شرط $\boldsymbol{A\beta} = ۰$ نرمال با میانگین صفر و ماتریس دقت $\tau_\beta^{-۱}(I - \boldsymbol{B})$ است که در آن $\boldsymbol{B} = \boldsymbol{Q}_\beta^{-۱}\boldsymbol{A'}(\boldsymbol{A}\boldsymbol{Q}_\beta^{-۱}\boldsymbol{A'})^{-۱}\boldsymbol{A}$. در ابتدا تعریف می‌کنیم

$$\boldsymbol{Z} = \begin{bmatrix} ۱ & ۱ & ۰ & ۰ \\ ۱ & ۰ & ۱ & ۰ \\ ۱ & ۰ & ۰ & ۱ \end{bmatrix}.$$

پیشگوی خطی برای $i = ۱, \ldots, n$ و $j = ۱, ۲, ۳$ برابر است با $\eta_{ij} = \beta_۰ + \beta_j$ بنابراین توزیع پسین $\boldsymbol{\beta}$ به صورت

$$\pi(\boldsymbol{\beta}|y) \propto |\boldsymbol{Q}_y|^{۱/۲} e^{\frac{-۱}{۲}(\boldsymbol{y}-\boldsymbol{Z\beta})'\boldsymbol{Q}_y(\boldsymbol{y}-\boldsymbol{Z\beta})} |\tau_\beta^{-۱}(I - \boldsymbol{B})|^{۱/۲} e^{\frac{-۱}{۲}\boldsymbol{\beta}'(\tau_\beta^{-۱}(I-\boldsymbol{B}))\boldsymbol{\beta}}$$

$$\propto \; e^{\frac{-1}{2}[\boldsymbol{\beta}-(\boldsymbol{Z'Q}_y\boldsymbol{Z}+\tau_\beta^{-1}(I-\boldsymbol{B})^{-1})^{-1}\boldsymbol{Z'Q}_y\boldsymbol{y}]'(\boldsymbol{Z'Q}_y\boldsymbol{Z}+\tau_\beta^{-1}(I-\boldsymbol{B}))[\boldsymbol{\beta}-(\boldsymbol{Z'Q}_y\boldsymbol{Z}+\tau_\beta^{-1}(I-\boldsymbol{B}))^{-1}\boldsymbol{Z'Q}_y\boldsymbol{y}]}.$$

نتیجه می‌شود. در نتیجه یک توزیع نرمال با میانگین $(\boldsymbol{Z'Q}_y\boldsymbol{Z}+\tau_\beta^{-1}(I-\boldsymbol{B}))^{-1}\boldsymbol{Z'Q}_y\boldsymbol{y}$ و ماتریس دقت $\boldsymbol{Z'Q}_y\boldsymbol{Z}+\tau_\beta^{-1}(I-\boldsymbol{B})$, توزیع پسین برای $\boldsymbol{\beta}$ است، که در آن $(I-\boldsymbol{B})$ یک ماتریس تکین است (رو و هلد، ۲۰۰۵). بنابراین باید از روش معکوس ماتریس تعمیم‌یافته[۱۲] و تجزیه مقدار تکین[۱۳] استفاده کنیم.

در شکل‌های ۱.۳ و ۲.۳ ویژگی‌های توزیع‌های پسین حاصل از دو نوع پارامتربندی به‌عنوان تابعی از τ_β مقایسه و نمایش داده شده‌اند. شکل ۱.۳ نسبت میانگین‌های پسین را با دقت دو رقم اعشار برای $\tau_\beta \in \{۰/۰۰۱, ۰/۰۰۵, ۰/۲۵, ۰/۵, ۰/۷۵, ۱\}$ نشان می‌دهد. تفاوتی بین میانگین‌های پسین مشاهده نمی‌شود. شکل ۲.۳ دقت برآوردشده را برای $\tau_y = ۱$، $n = ۳۰$ و مقادیر یکسان τ_β مانند قبل نشان می‌دهد. تفاوت خیلی کمی بین دقت‌های پسینی تنها وقتی که مقادیر τ_β کوچکی دارد، به چشم می‌خورد. در پایان، واریانس حاصل از دو مدل با $\tau_y = ۱$ و $\tau_\beta = ۰/۰۰۱$ به‌طوری که اندازه n رو به افزایش است، در شکل ۳.۳ به نمایش گذاشته شده است. همان‌طور که در شکل مشهود است، وقتی اندازه نمونه افزایش یابد، واریانس‌های پسین حاصل از مدل‌ها به یکدیگر همگرا می‌شوند.

با در نظر گرفتن نتایج شبیه‌سازی مدل نرمال عاملی، دو قضیه زیر را بیان و ثابت می‌کنیم. همان‌طور که نتایج شبیه‌سازی نشان دادند، میانگین‌های حاصل از توزیع‌های پسین برای دو قید متفاوت تا دو رقم اعشار با هم برابر هستند که در زیر این برابری را به صورت رسمی نیز اثبات می‌کنیم.

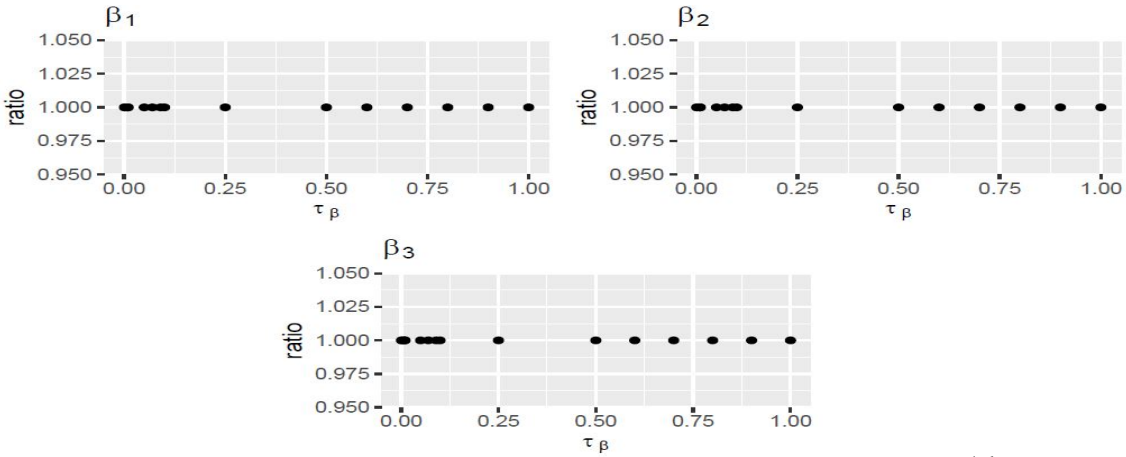
قضیه ۱.۵.۳. مدل نرمال با میانگین عاملی $\boldsymbol{Z\beta}$ و دقت $\boldsymbol{Q}_y = \tau_y I$ را در نظر بگیرید. فرض کنید برای شناسایی مدل از دو قید رده گوشه و مجموع مساوی با صفر استفاده کنیم که ماتریس‌های دقت توزیع‌های پیشین نرمال متناظر آن‌ها به ترتیب $\boldsymbol{Q}_{\beta^*}$ و $\boldsymbol{Q}_\beta$ هستند. دو میانگین حاصل از پسین برای دو قید رده گوشه و قید مجموع مساوی با صفر، تقریبا اختلاف ناچیز دارند.

برهان. مدل را بدون عرض از مبدا (برای سبکی فرمول‌ها) در نظر بگیرید. قبل از شروع برهان در ابتدا تعریف می‌کنیم

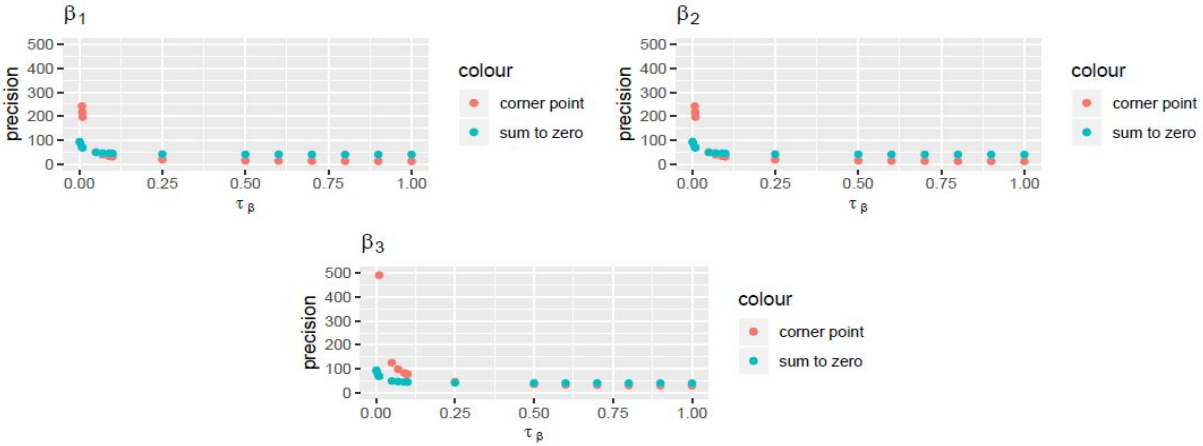
$$\boldsymbol{Z} = \begin{bmatrix} ۱ & ۰ & ۰ \\ ۰ & ۱ & ۰ \\ ۰ & ۰ & ۱ \end{bmatrix},$$

$$\boldsymbol{Z}_۱ = \begin{bmatrix} ۱ & ۰ & ۰ \\ ۰ & ۱ & ۰ \\ ۰ & ۰ & ۰ \end{bmatrix}.$$

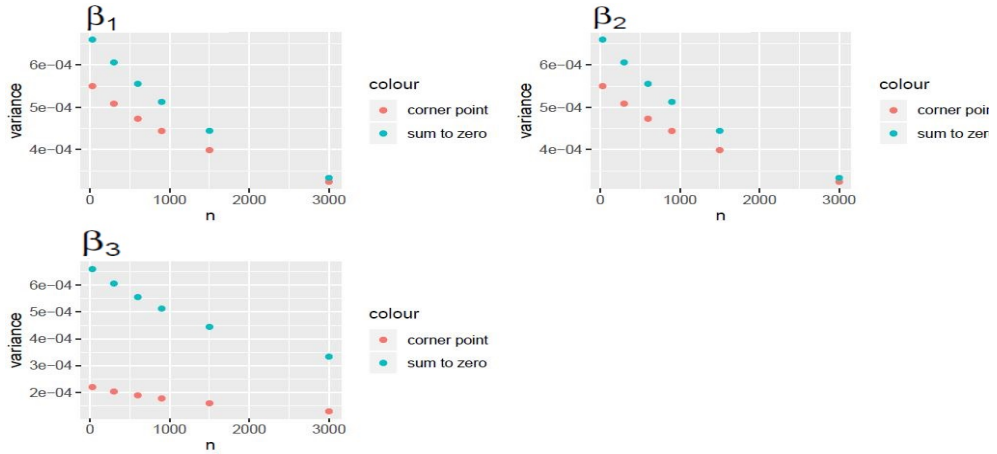
[۱۲] Moore-Penrose
[۱۳] Singular value dicomposition (svd)



شکل ۱.۳: نسبت میانگین پسین با در نظر گرفتن قیدهای شناسایی‌پذیری در مدل نرمال عاملی.



شکل ۲.۳: مقایسه دقت با در نظر گرفتن قیدهای شناسایی‌پذیری در مدل نرمال عاملی برای $n = ۳۰$ و $\tau_y = ۱$. دقت‌های برآوردشده در هر رده وقتی τ_β از ۰/۰۰۱ تا ۱ تغییر می‌کند، تقریبا به یکدیگر نزدیک هستند.



شکل ۳.۳: مقایسه واریانس با در نظر گرفتن قیدهای شناسایی‌پذیری در مدل نرمال عاملی وقتی تعداد مشاهدات رو به افزایش است.

در حالت کلی قرار می‌دهیم $\boldsymbol{Q}_\beta = \boldsymbol{Q}$. می‌دانیم میانگین‌های مدل‌های قید مجموع مساوی با صفر و قید رده گوشه به ترتیب برابر است با:

$$(\boldsymbol{Z'Q}_y\boldsymbol{Z} + \boldsymbol{Q}^{-1}(I - \boldsymbol{B}))^{-1}\boldsymbol{Z'Q}_y\boldsymbol{y},$$

و

$$(\boldsymbol{Z}_1'\boldsymbol{Q}_y\boldsymbol{Z}_1 + \boldsymbol{Q}_{\beta^*})^{-1}\boldsymbol{Z}_1'\boldsymbol{Q}_y\boldsymbol{y}.$$

با مقایسه دو میانگین در می‌یابیم که تفاوت اصلی بین دو میانگین در ماتریس دقت است. برای نشان دادن میزان این تفاوت نسبت دترمینان ماتریس‌های دقت را به صورت

$$\frac{|\boldsymbol{Z'}Q_y\boldsymbol{Z} + \boldsymbol{Q}^{-1}(I - \boldsymbol{B})|}{|\boldsymbol{Z}_1'\boldsymbol{Q}_y\boldsymbol{Z}_1 + \boldsymbol{Q}_{\beta^*}|} = \frac{|\boldsymbol{Z'Q}_y\boldsymbol{Z}(I + \boldsymbol{Q}^{-1}\boldsymbol{Z'Q}_y^{-1}\boldsymbol{Z}(I - \boldsymbol{B}))|}{|\boldsymbol{Z}_1'\boldsymbol{Q}_y\boldsymbol{Z}_1 + \boldsymbol{Q}_{\beta^*}|}$$

می‌نویسیم. داریم

$$\frac{|\boldsymbol{Z'Q}_y\boldsymbol{Z}(\boldsymbol{Q} + \boldsymbol{Z'Q_y}^{-1}\boldsymbol{Z} - \boldsymbol{Z'Q_y}^{-1}\boldsymbol{ZB})|}{|\boldsymbol{Q}_{\beta^*} + \boldsymbol{Z}_1'\boldsymbol{Q}_y\boldsymbol{Z}_1|}$$

که برابر با

$$\frac{|\boldsymbol{Z'Q}_y\boldsymbol{Z}(\boldsymbol{Q} + \boldsymbol{Z'Q_y}^{-1}\boldsymbol{Z})(I - (\boldsymbol{Q} + \boldsymbol{Z'Q_y}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z'Q_y}^{-1}\boldsymbol{ZB})|}{|\boldsymbol{Q}_{\beta^*} + \boldsymbol{Z}_1'\boldsymbol{Q}_y\boldsymbol{Z}_1|},$$

است. با توجه به صورت کسر باید نشان دهیم $\boldsymbol{B}/\tau_y$ نسبت به $\boldsymbol{Q} + I/\tau_y$ از مقدار کوچکی برخوردار است. به عبارتی پرانتز دوم در صورت کسر تقریبا مقدار صفر می‌گیرد. حال تجزیه ویژه[۱۴] $\boldsymbol{Q} = \boldsymbol{V}\Lambda\boldsymbol{V'} + \epsilon\boldsymbol{NN'}$ را در نظر بگیرید که در آن $\boldsymbol{N} \in \mathbb{R}^{n\times w}$ و $\boldsymbol{V} \in \mathbb{R}^{n\times(n-w)}$ به‌گونه‌ای که w تعداد قیدها و n بعد ماتریس دقت است. برای $\boldsymbol{R} \in \mathbb{R}^{w\times w}$، می‌توان نوشت $\boldsymbol{A} = \boldsymbol{RN'}$ و در نتیجه بنابر رو و هلد (۲۰۰۵) داریم: $\boldsymbol{V'A'} = \boldsymbol{V'NA'} = 0$ و همچنین $\boldsymbol{V'NN'} = 0$.

$$\boldsymbol{Q}^{-1}\boldsymbol{A'} = \boldsymbol{V\Lambda}^{-1}\boldsymbol{V'A'} + \epsilon^{-1}\boldsymbol{NN'NR'} = \epsilon^{-1}\boldsymbol{A'}.$$

بسطی از $\boldsymbol{B}$ را به صورت

$$\begin{aligned}\boldsymbol{B} &= Q^{-1}\boldsymbol{A'}(\boldsymbol{AQ}^{-1}\boldsymbol{A'})\boldsymbol{A}\\ &= \epsilon^{-1}\boldsymbol{NR'}(\epsilon^{-1}\boldsymbol{RN'NR'})'\boldsymbol{RN'}\\ &= \boldsymbol{NN'}, \qquad (۱.۳)\end{aligned}$$

می‌توان نوشت. بنابراین

$$\begin{aligned}(\boldsymbol{Q} + \boldsymbol{Z'Q_y}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z'Q_y}^{-1}\boldsymbol{ZB} &= 1/\tau_y(\boldsymbol{Q} + 1/\tau_y I)^{-1}\boldsymbol{B}\\ &= 1/\tau_y\boldsymbol{V}(\Lambda + 1/\tau_y I)^{-1}\boldsymbol{V'NN'} + 1/\tau_y(\epsilon + 1/\tau_y)^{-1}\boldsymbol{NN'}\\ &= (1 - \frac{\epsilon}{1/\tau_y + \epsilon})\boldsymbol{NN'}. \qquad (۲.۳)\end{aligned}$$

[۱۴] eigendecomposition

پس

$$(I-(\boldsymbol{Q}+\boldsymbol{Z'Q_y^{-1}Z})^{-1}\boldsymbol{Z'Q_y^{-1}ZB}) = (I-1/\tau_y(\boldsymbol{Q}-1/\tau_y I)^{-1}\boldsymbol{B}) \quad (۳.۳)$$
$$= (\frac{\epsilon}{1/\tau_y+\epsilon})^w.$$

توجه داشته باشید که (۳.۳) به ابرپارامترهای نامعلوم بستگی دارد. اما در مدل، $\epsilon \ll \tau_y$ و بنابراین (۳.۳) تقریبا برابر صفر است. اختلاف در میانگین با اختلاف ناچیز بین دو ماتریس دقت $\boldsymbol{Q}$ و $\boldsymbol{Q}_{\beta^*}$ کنترل می‌شود. □

در ادامه قصد داریم نشان دهیم در این حالت دقت مدل با قید مجموع مساوی با صفر از قید رده گوشه بیشتر است.

قضیه ۲.۵.۳. مدل بیان‌شده در قضیه ۱.۵.۳ را در نظر بگیرید. در این حالت دقت مدل حاصل از قید مجموع مساوی با صفر از مدل حاصل از قید رده گوشه بیشتر است.

برهان. می‌دانیم قید رده گوشه دقت برآورد مدل را نسبت به مدل بدون قید کاهش می‌دهد. بنابراین کافی است نشان دهیم که دقت مدل با قید مجموع مساوی با صفر از مدل بدون قید بیشتر است ابتدا تابع درستنمایی مربوط به مدل بدون قید را به صورت

$$\pi(\boldsymbol{\beta}|y) \propto |\boldsymbol{Q}_y|^{1/2} e^{\frac{-1}{2}(\boldsymbol{y}-\boldsymbol{Z\beta})'\boldsymbol{Q}_y(\boldsymbol{y}-\boldsymbol{Z\beta})} |\boldsymbol{Q}|^{1/2} e^{\frac{-1}{2}\boldsymbol{\beta Q\beta}}$$
$$\propto e^{\frac{-1}{2}[\boldsymbol{\beta}-(\boldsymbol{Z'Q_yZ+Q})^{-1}\boldsymbol{Z'Q_y y}]'(\boldsymbol{Z'Q_yZ+Q})^{-1}[\boldsymbol{\beta}-(\boldsymbol{Z'Q_yZ+Q})^{-1}\boldsymbol{Z'Q_y y}]}.$$

داریم. از برهان قضیه قبل به خاطر داریم $\boldsymbol{B}=\boldsymbol{NN'}$، $\boldsymbol{Q}=\boldsymbol{V'}+\epsilon\boldsymbol{NN'}$ و $I=\boldsymbol{VV'}+\boldsymbol{NN'}$. معکوس مور–پنروس $(I-\boldsymbol{B})\boldsymbol{Q}^{-1}$ را با $[(I-\boldsymbol{B})\boldsymbol{Q}^{-1}]^+$ نشان می‌دهیم. بنابراین داریم

$$([(I-\boldsymbol{B})\boldsymbol{Q}^{-1}]^+ + \boldsymbol{Z'Q_yZ})^{-1} = \boldsymbol{V}(\boldsymbol{\Lambda}+\tau_y I)^{-1}\boldsymbol{V'} + \tau_y^{-1}\boldsymbol{NN'}$$

و در نتیجه اختلاف دو دقت برابر است با

$$([(I-\boldsymbol{B})\boldsymbol{Q}^{-1}]^+ + \boldsymbol{Z'Q_yZ})^{-1} - (\boldsymbol{Q}+\boldsymbol{Z'Q_yZ})^{-1} = \frac{\epsilon}{\tau_y(\tau_y+\epsilon)}\boldsymbol{NN'} > 0$$

که نشان می‌دهد دقت مدل بدون قید از دقت مدل با قید مجموع مساوی با صفر کمتر است. از طرفی می‌دانیم دقت مدل با قید رده گوشه از مدل بدون قید کمتر است. بنابراین دقت مدل با قید مجموع مساوی با صفر از قید رده گوشه بیشتر است. □

قضیه دوم نشان می‌دهد قید مجموع مساوی با صفر یکی از اهداف ما از پیشنهاد این قید که همان افزایش دقت برآورد است را به خوبی ادا کرده است. لازم به ذکر است که این قضایا در حالت کلی ثابت شده‌اند و به هر مدل LGM دیگری نیز قابل تعمیم هستند. بنابراین می‌توان این نتایج را به هر مدلی که بتوان آن را به صورت یک LGM در آورد، نسبت داد.

فصل ۴

مدل چندجمله‌ای جمعی‌ـ‌ساختاری با در نظر گرفتن قید شناسایی

در این فصل، مدل چندجمله‌ای بیزی با ساختار رگرسیونی جمعی را برای داده‌های رسته‌ای فضایی معرفی و عملکرد آن را ارزیابی می‌کنیم. با هدف برازش سریع یک مدل با کیفیت بالا، از تقریب INLA استفاده می‌کنیم. برای این کار، در ابتدا با کمک تبدیل چندجمله‌ای‌ـ‌پوآسن (بیکر، ۱۹۹۴) به مدل چندجمله‌ای قابلیت برازش و استنباط با روش INLA می‌بخشیم. برای رفع مشکل شناسایی ذاتی در پارامترهای مدل از دو نوع قید پیشنهادی در فصل قبل استفاده کرده و کارایی آن را به کمک مثال‌های شبیه‌سازی و واقعی ارزیابی می‌کنیم. مثال واقعی شامل داده‌های مربوط به شدت تصادفات رانندگی در استان مازندران است. تاکنون مدل‌های لوجیت چندجمله‌ای به‌طور وسیعی برای تحلیل داده‌های رسته‌ای مربوط به تصادفات جاده‌ای به‌کار گرفته شده‌اند. برای داشتن یک مقایسه جامع با مدل‌های به‌کار گرفته‌شده در این زمینه، یک نسخه بیزی از مدل چندجمله‌ای کسری[۱] (تیل، ۱۹۶۰، مولاهی، ۲۰۱۰، لی، ۲۰۱۸) را توسعه داده و نتایج حاصل از برازش مدل پیشنهادی خود را با این مدل در مثال واقعی مقایسه و نتیجه‌گیری می‌کنیم.

[۱] Fractional split multinomial model

۱.۴ تعریف مدل

در این بخش مدل چندجمله‌ای بیزی را در قالب یک رگرسیون جمعی‌ـ‌ساختاری با مولفه فضایی، با در نظر گرفتن دو نوع قید شناسایی، معرفی می‌کنیم.

۱.۱.۴ رگرسیون جمعی‌ـ‌ساختاری برای داده‌های رسته‌ای اسمی

فرض کنید در هر آزمایش $i = ۱, \ldots, n$، یک انتخاب از بین J رده داشته باشیم. بردار مشاهدات برای هر i به صورت $\boldsymbol{y}_i = (y_{i۱}, \ldots, y_{iJ})'$ است، به‌گونه‌ای که $\sum_{j=۱}^{J} y_{ij} = n_i$. همچنین فرض کنید $\boldsymbol{\pi}_i = (\pi_{i۱}, \ldots, \pi_{iJ})'$ که در آن π_{ij} احتمال انتخاب jامین رده در iامین آزمایش است. درستنمایی مدل چندجمله‌ای برای هر آزمایش i به صورت

$$\mathrm{p}(\boldsymbol{y}_i | n_i, \boldsymbol{\pi}_i) = \binom{n_i}{\boldsymbol{y}_i} \pi_{i۱}^{y_{i۱}} \ldots \pi_{iJ}^{y_{iJ}}, \tag{۱.۴}$$

است، که در آن برای هر $i = ۱, \ldots, n$، $\sum_j \pi_{ij} = ۱$. می‌دانیم به کمک مدل لجیت چندجمله‌ای، می‌توان متغیرهای تبیینی را به صورت زیر وارد مدل کرد

$$\pi_{ij} = \frac{\exp(\eta_{ij})}{\sum_{s=۱}^{J} \exp(\eta_{is})} \tag{۲.۴}$$

که در آن η_{ij} پیشگوی خطی و به صورت

$$\eta_{ij} = \alpha_j + x_{۱ij}\beta_{۱j} + \cdots + x_{lij}\beta_{lj}, \quad i = ۱, \ldots, n, \quad j = ۱, \ldots, J, \tag{۳.۴}$$

است. برای jامین رده، α_j همان عرض از مبدا است و $(x_{۱j}, \ldots, x_{lj})'$ برداری از متغیرهای تبیینی برای همه آزمایش‌ها و $\boldsymbol{\beta}_j = (\beta_{۱j}, \ldots, \beta_{lj})'$ بردار ضرایب رگرسیونی است.

احتمال انتخاب یک گزینه از بین چند گزینه تحت تاثیر سایر گزینه‌ها نیست (لوکه، ۱۹۵۹). در اصطلاح ریاضی، این به آن معنی است که احتمال انتخاب گزینه c از تعداد J گزینه برابر با

$$\pi_c = \frac{w_c}{\sum_j w_j},$$

است، که در آن w_c وزن گزینه c است و به ماکسیمم نرم[۲] معروف است (لوکه، ۱۹۵۹). این نمایش از احتمال رابطه خیلی نزدیکی با مفهوم مدل لجیت در آمار دارد (لوکه، ۱۹۷۷). ما از این حقیقت استفاده می‌کنیم تا قید $\sum_j \pi_{ij} = ۱$ را وارد مدل کنیم. در واقع $\sum_{s=۱}^{J} \exp(\eta_{is})$ در رابطه (۲.۴) که به قانون انتخاب لوکه معروف است (لوکه، ۱۹۵۹)، تضمین می‌کند تا مجموع احتمالات در همه رده‌ها یک باشد. به‌عبارتی دیگر با استفاده از مدل لجیت توانستیم قید شناسایی‌پذیری برای احتمال را وارد مدل کنیم. استفاده از تابع نمایی در تابع لجیت برای

[۲] Soft max

سادگی ریاضی و راحتی کار است. یک ویژگی خوب این تابع این است که مشتق آن با خود تابع یکسان است.

با در نظر گرفتن $\boldsymbol{u} = (u_1, \ldots, u_{n_f})'$، برای نمایش اطلاعات مربوط به اثرات متغیرهای ساختاری می‌توان پیشگوی خطی را به صورت

$$\eta_{ij} = \alpha_j + \sum_{r=1}^{n_f} f_j(u_{ri}) + \sum_{k=1}^{l} x_{kij}\beta_{kj}, \tag{۴.۴}$$

توسعه داد، که در آن $f(\cdot)$ نشان‌دهنده تابعی ناخطی از متغیرهای ساختاری (تبیینی پیوسته) است که می‌تواند به صورت اثرات ناپارامتری، زمانی، فضایی و فضایی‌ـ‌زمانی ظاهر شود. اگر به پارامترهای $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_J')'$، $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)'$ و $\{f(\cdot)\}$ توزیع پیشین گوسی نسبت دهیم، آن‌گاه با یک مدل LGM روبرو هستیم.

۲.۱.۴ مدل‌بندی ساختار وابستگی فضایی

بسته به پیوسته (چگال) یا گسسته (ناچگال) بودن ناحیه تحت مطالعه، مدل‌های متفاوتی برای مدل‌بندی ساختار وابستگی فضایی معرفی شده‌اند. مدل‌های معروف پیشنهادی برای ناحیه اندیس‌گذاری‌شده گسسته، برای استفاده در INLA توصیه و همچنین در بسته نرم‌افزاری `R-INLA` تعبیه شده‌اند؛ کاربردهای متعددی از آن‌ها را می‌توان در اسکرود و همکاران (۲۰۱۲)، اگرته و همکاران (۲۰۱۴) و بلانگیاردو و کاملتی (۲۰۱۵) مشاهده کرد.

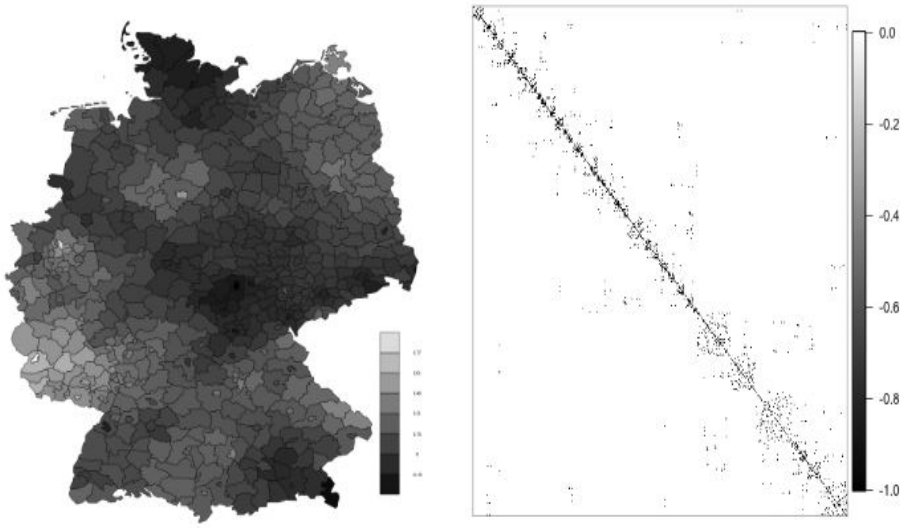
یکی از معروفترین مدل‌های فضایی ناحیه گسسته، مدل بی‌سیگ (بی‌سیگ، ۱۹۹۱) یا همان ICAR است که به افتخار آماردانی با همین نام، ژولین بی‌سیگ، نامگذاری شده است. ما در این رساله از عبارت CAR برای هر مدل گوسی چندمتغیره که از توزیع‌های شرطی ساخته شده باشد، استفاده می‌کنیم. بنابراین، در این رساله مدل بی‌سیگ، یک مثال خاص از مدل CAR محسوب می‌شود.

مدل بی‌سیگ

صورت کلی مدل ICAR را در فصل اول معرفی کردیم. این مدل، اثر فضایی در هر ناحیه، u_i، را روی یک مجموعه از نواحی که با $\{1, \ldots, n\}$ اندیس‌گذاری و روی نواحی همسایه شرطی شده است، مدل‌بندی می‌کند. معمولا دو ناحیه را همسایه در نظر می‌گیرند، اگر مرز مشترکی داشته باشند (شکل ۱.۴ را ببینید). توزیع شرطی برای u_i به صورت

$$u_i|\boldsymbol{u}_{-i}, \tau_u \sim \mathrm{N}\left(\frac{1}{d_i}\sum_{j\sim i} u_j, \frac{1}{d_i\tau_u}\right), \tag{۵.۴}$$

است، که در آن $j \sim i$ همسایه بودن دو ناحیه i و j را نشان می‌دهد و d_i تعداد همسایه‌های ناحیه i است. همچنین τ_u پارامتر دقت مدل است. در این صورت، توزیع توأم $\boldsymbol{u} = (u_1, \ldots, u_n)'$ به



شکل ۱.۴: مدل بی‌سیگ. شکل سمت چپ مربوط به داده‌های یک مثال از نواحی کشور آلمان است. شکل سمت راست ماتریس دقت تنک $\boldsymbol{Q}$ متناظر با این نواحی را نشان می‌دهد.

صورت

$$\boldsymbol{u}|\tau_u \sim \mathrm{N}\left(۰, \frac{۱}{\tau_u}\boldsymbol{Q}^{-۱}\right),$$

است، که در آن $\boldsymbol{Q}$ یک ماتریس دقت ساختاری است و مولفه‌های آن عبارتند از

$$Q_{i,j} = \begin{cases} d_i, & i = j \\ -۱, & i \sim j \\ ۰, & \text{سایر موارد.} \end{cases} \qquad (۶.۴)$$

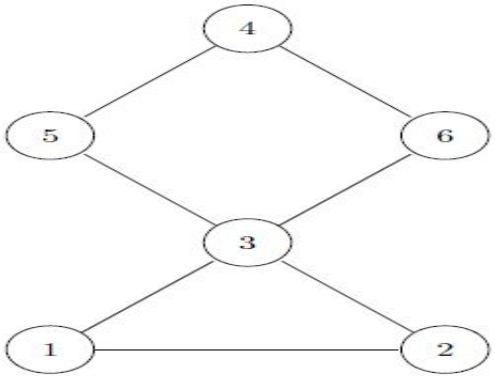
این ماتریس دقت ساختاری، به‌طور مستقیم ساختار همسایگی نواحی را لحاظ می‌کند و با داشتن $\mathrm{O}(n)$ عضو ناصفر، تنک است (پنل راست شکل ۱.۴ را ببینید). توجه داشته باشید مدل بی‌سیگی که در آن وزن‌های مثبت برای هر جفت از ناحیه‌های همسایه منظور شده باشد، به‌طور مشابه قابل تعریف است (رو و هلد، ۲۰۰۵). راه‌های مختلفی برای تعیین مولفه‌های ماتریس دقت در (۶.۴) وجود دارند که به‌عنوان نمونه می‌توان به دو روش زیر اشاره کرد.

۱. از طریق ماتریس همسایگی متقارن با بعد $n \times n$

۲. با استخراج مستقیم ساختار از فایل شیپ[۳] به کمک بسته‌های نرم افزاری مانند بسته‌های `maptools` (بیوند و لوین کح، ۲۰۱۷) و `spdep` (بیوند و پیراس، ۲۰۱۵؛ بیوند و همکاران، ۲۰۱۳) در محیط `R`.

[۳] Shape file

تعریف مدل بی‌سیگ با اختصاص یک گراف شروع می‌شود. گراف از مجموعه‌ای از نقاط و یال، که نواحی و همسایگی بین آن‌ها را نشان می‌دهند، تشکیل می‌شود. یک گراف متصل‌شده است اگر مسیر وجود داشته باشد (یعنی مجموعه‌ای از یال‌های به‌هم پیوسته داشته باشد) که هر نقطه را حداقل به یک نقطه دیگر متصل می‌کند. در یک گراف متصل‌شده، رابطه همسایگی واضح است (هر نقطه حداقل یک همسایه دارد). در شکل ۲.۴ نمونه‌ای از یک گراف (فرنی‌ـ استرتینو، ۲۰۱۸) را می‌توانید ببینید. مدل بی‌سیگ از چنین گرافی استفاده می‌کند.



شکل ۲.۴: نمونه‌ای از یک گراف متصل‌شده.

مشکل شناسایی‌پذیری در مدل بی‌سیگ

مدل معرفی‌شده در (۵.۴) و (۶.۴)، ذاتی (ناسره) است به این معنی که میانگین کلی در آن نامشخص رها شده و نیاز به شناسایی دارد. در حالت کلی، دو راه برای شناسایی آن وجود دارد:

۱. **استفاده از قید خطی**: فرنی‌ـ اسرینو در سال ۲۰۱۸ نشان داد که با استفاده از یک قید خطی می‌توان مدل را قابل شناسایی کرد. با توجه به قابلیت INLA در به‌کار گرفتن قید خطی، استفاده از قید مجموع مساوی با صفر را پیشنهاد کرده و به‌کار می‌گیریم، به‌طوری که مجموع اثرات فضایی برابر صفر شود، یعنی $\sum_{i=1}^{n} u_i = 0$.

۲. **اضافه کردن یک مقدار خیلی کوچک به قطر اصلی ماتریس دقت**: در این حالت با افزودن یک مقدار کوچک $\gamma > 0$ به تمام اعضای قطر اصلی ماتریس دقت Q آن را به یک ماتریس معین مثبت تبدیل می‌کنند. بنابراین کافی است نشان دهیم که اختلاف این تقریب با Q تابعی از γ است؛ به عبارتی این اختلاف با γ کنترل می‌شود. زیرا با کمی کندوکاو بیشتر دریافتیم که افزودن قید مجموع مساوی با صفر به مدل به اندازه γ به قطر اصلی ماتریس دقت اضافه می‌کند. بنابراین به سراغ روشی رفتیم که در آن به اختیار خودمان به اندازه γ به قطر اصلی اضافه کردیم و کارایی آن را سنجیدیم.

قضیه ۱.۱.۴. فرض کنید $\boldsymbol{X}$ یک میدان تصادفی گوسی مارکوفی با ماتریس دقت $\boldsymbol{Q}$ با بعد n باشد که ذاتی هم هست. تفاوت بین نرم قوی[۴] و نرم ضعیف[۵] دو ماتریس دقت برابر است با $\|\mathrm{Prec}(\boldsymbol{X}|\mathbf{1}'\boldsymbol{X}) - \boldsymbol{Q}\| = c\gamma$ که در آن c یک مقدار ثابت است.

لازم به ذکر است که $\boldsymbol{X}|\mathbf{1}'\boldsymbol{X}$ معادل با افزودن قید مجموع مساوی با صفر به مدل است که به اندازه γ به مدل اضافه می‌کند. برای اثبات قضیه، در ابتدا نرم قوی و نرم ضعیف را از رو و هلد (۲۰۰۵) تعریف می‌کنیم.

تعریف ۱.۱.۴. فرض کنید $\boldsymbol{M}$ یک ماتریس $n \times n$ است. نرم قوی $\boldsymbol{M}$ برابر است با

$$\|\boldsymbol{M}\|_s = \max_{x:\ x'x=1} (x'\boldsymbol{M}'\boldsymbol{M}x)^{1/2}.$$

همچنین، نرم ضعیف $\boldsymbol{M}$ برابر است با

$$\|\boldsymbol{M}\|_w = \left(\frac{1}{n}\sum_{ij} \boldsymbol{M}_{ij}^2\right)^{1/2}.$$

برهان. فرض کنید $\{(\lambda_i, e_i),\ i = 1, \ldots, n\}$ جفت‌های مقادیر ویژه و بردار ویژه ماتریس $\boldsymbol{Q}$ هستند. بنابراین رابطه

$$\boldsymbol{Q} = \sum_i \lambda_i e_i e_i' = \boldsymbol{v}\boldsymbol{\Lambda}\boldsymbol{v}',$$

برقرار است، که در آن $\boldsymbol{v}' = (e_1, \ldots, e_n)$، $\boldsymbol{v}\boldsymbol{v}' = 1$، $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ و $\lambda_1 = 0$، $e_1 = 1$. توجه داشته باشید که بردارهای ویژه یکسان باقی می‌مانند. جفت‌های مقادیر ویژه و بردار ویژه ماتریس شرطی $\boldsymbol{X}$ به شرط $\mathbf{1}'\boldsymbol{X}$ برابرند با

$$(0, 1), (\lambda_2 + \gamma, e_2), \ldots, (\lambda_n + \gamma, e_n)$$

زیرا با استفاده از رو و هلد (۲۰۰۵)

$$\mathrm{Prec}(\boldsymbol{X}|\mathbf{1}'\boldsymbol{X}) = \boldsymbol{v}\tilde{\boldsymbol{\Lambda}}\boldsymbol{v}'$$

که در آن $\tilde{\boldsymbol{\Lambda}} = \mathrm{diag}(0, \lambda_2 + \gamma, \ldots, \lambda_n + \gamma)$ و $\mathrm{Prec}(\boldsymbol{X})$ ماتریس دقت $\boldsymbol{X}$ است. بنابراین نتیجه

$$||\mathrm{Prec}(\boldsymbol{X}|\mathbf{1}'\boldsymbol{X}) - \boldsymbol{Q}|| = c\gamma,$$

حاصل می‌شود، که در آن مقدار c برای نرم قوی برابر با یک و در غیر این صورت برابر با $\sqrt{\frac{n-1}{n}}$ است. □

[۴] Strong norm
[۵] Weak norm

یادآور می‌شویم که نرم قوی و نرم ضعیف را می‌توان با مقادیر ویژه نیز حساب کرد به‌طوری که برای یک ماتریس دقت با مقادیر ویژه $\lambda_1, \ldots, \lambda_n$ نرم قوی و نرم ضعیف به ترتیب برابر است با

$$\|\boldsymbol{M}\|_s^2 = \max_k \lambda_k, \quad k = 1, \ldots, n$$

و

$$\|\boldsymbol{M}\|_w^2 = \frac{1}{n}\text{trace}(M'M) = \frac{1}{n}\sum_k \lambda_k, \quad k = 1, \ldots, n.$$

با کمی تحقیق بیشتر دریافتیم اگر کسی توزیع پیشین با دقت $\tilde{\boldsymbol{Q}} = \boldsymbol{Q} + \gamma \boldsymbol{I}$ را برای مدل فضایی در نظر بگیرد، اگرچه اضافه کردن این مقدار به قطر ماتریس، مدل را سره[۶] می‌سازد، اما فضای صفر[۷] در ماتریس دقت همچنان باقی است. در نتیجه، راه حل اول محکم‌تر و مفیدتر در برطرف شدن مشکل به‌نظر می‌رسد.

در این رساله، برخلاف زمان زیادی که صرف پیدا کردن و برهان قضیه در راه حل دوم شد، از روش اول برای رسیدن به هدف استفاده می‌کنیم چرا که سعی کردیم تمرکز خود را در استفاده از یک روش با اجرای ساده‌تر و در عین حال کاراتر بالا ببریم.

مقیاس‌بندی مدل بی‌سیگ

در این بخش درباره یک عقیده کلیدی صحبت می‌کنیم که به‌کار گرفتن آن موجب می‌شود تا مدل‌هایی که در واقع یک میدان تصادفی گوسی ذاتی[۸] هستند، از نظر تفسیر پارامتر راحت‌تر و قابل مقایسه با سایر مدل‌ها شوند.

تا این جا دریافتیم که در مدل رگرسیونی سلسله‌مراتبی بیزی، یک روش متداول در مدل‌بندی فضایی استفاده از مدل‌های IGMRF است. دقت کنید که منظور از IGMRF همان مدل ICAR است که میدان تصادفی آن گوسی فرض می‌شود. بنابراین، در این رساله که فقط از پذیره گوسی بودن میدان تصادفی استفاده می‌کنیم، این دو عبارت را به‌جای هم به کار می‌بریم. در مدل IGMRF یک ماتریس دقت مقیاس‌بندی‌شده وجود دارد که ساختار همسایگی مدل را منعکس می‌کند و پارامتر مقیاس در آن به عنوان پارامتر دقت تصادفی گزارش می‌شود. پیشین انتخاب‌شده برای پارامتر دقت روی درجه هموارسازی تاثیر می‌گذارد و در نتیجه تأثیر قوی بر روی نتایج پسین دارد (کرشیانکی، ۲۰۱۸). در واقع هر IGMRF یک پارامتر دقت دارد که تفاوت‌های بین همسایه‌ها را کنترل می‌کند. مقادیر بزرگ نشان از رابطه‌ای قوی (همواری بیشتر) دارد، در حالی‌که مقادیر کوچکتر نشان از رابطه ضعیف (همواری کمتر) دارد. برای

[۶] Proper
[۷] Null space
[۸] Intrinsic Gaussian Markov Random Fields (IGMRF)

مثال، برناردلی و همکاران (۱۹۹۵) را ببینید. بنابراین، یکی از چالش‌های پیش رو این هست که τ_u تفسیرپذیر نیست زیرا به گراف تحت مطالعه بستگی دارد.

اولین مرحله برای بخشیدن قابلیت تفسیرپذیری به τ_u این هست که مدل بی‌سیگ را از ذاتی بودن در بیاوریم (که در بخش قبل به شرح کامل آن پرداختیم). دومین مرحله این است که پارامتر τ_u نسبت به تمام ناحیه‌ها و مجموعه یال‌ها، پایدار[۹] شود. برای این که ناپایداری τ_u را دریابیم، تعریف انحراف استاندارد کناری

$$\mathrm{sd}(u_i) = \frac{\sqrt{\boldsymbol{Q}^-_{ii}}}{\sqrt{\tau_u}}$$

را ارایه می‌کنیم، که در آن $\boldsymbol{Q}^-$ واریانس تعمیم‌یافته $\boldsymbol{Q}$ است. انحراف استاندارد کناری روی i (یعنی روی نواحی) ثابت نیست و ماکسیمم و هر نوع میانگین آن‌ها به شدت به n وابسته است. بنابراین، سریای و رو (۲۰۱۴) استفاده از توزیع پیشین بر اساس نوع IGMRF را پیشنهاد کردند. در روش پیشنهادی آن‌ها، ماتریس ساختار همسایگی که به نوع گراف وابسته است، به‌گونه‌ای مقیاس‌بندی می‌شود که میانگین هندسی مقادیر واریانس‌های کناری برابر با یک شود. به این ترتیب، مدل بی‌سیگ با مدل‌های دیگر قابل مقایسه می‌شود. این کار با نقش بستن پارامتر دقت تصادفی بر انحراف استاندارد کناری مدل IGMRF شکل می‌گیرد و در نهایت توزیع پیشین روی τ_u مجدد محاسبه می‌شود. برای توضیح بیشتر، فرض کنید اثر تصادفی $\boldsymbol{u}$ یک IGMRF با میانگین صفر و ماتریس دقت $\tau_u\boldsymbol{R}$ باشد. ماتریس دقت مقیاس‌بندی‌شده را با $\tau_u\boldsymbol{R}.\sigma^2_{GV}(\boldsymbol{R})$ نشان می‌دهیم که $\sigma^2_{GV}(\boldsymbol{R})$ واریانس تعمیم‌یافته $\boldsymbol{R}$ است که توسط سربای و رو (۲۰۱۴) به صورت

$$\sigma^2_{GV}(\boldsymbol{R}) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(\mathrm{diag}(\boldsymbol{R}^-))\right),$$

پیشنهاد شده است. در این‌جا $\boldsymbol{R}^-$ ماتریس تعمیم‌یافته $\boldsymbol{R}$ است. با انجام این کار واریانس کناری مربوط به هر یک از عنصرهای بردار تصادفی $\boldsymbol{u}$ برابر یک می‌شود.

۳.۱.۴ شناسایی‌پذیری در مدل‌های چندجمله‌ای

از آن جایی که قید $\sum_j \pi_{ij} = 1$ در مدل پیشنهادی موجب می‌شود تا پارامترهای مدل منحصر به فرد نباشند، بنابراین پارامترهای مدل قابل شناسایی نیستند. برای مثال در برآورد $\boldsymbol{\beta}_1$ در معادله (۳.۴)، اگر هر مقدار ثابتی را به همه J پارامتر $(\beta_{11},\ldots,\beta_{1J})$ اضافه کنیم، تابع درستنمایی یکسان باقی می‌ماند، به‌طوری که داده‌ها توانایی شناسایی پارامترها را ندارند (جین و همکاران، ۲۰۱۳). در این‌جا، ما به کمک دو قید رده گوشه و مجموع مساوی با صفر مشکل ناشناسا بودن مدل را مرتفع می‌کنیم.

[۹] Stable

قید رده گوشه برای مدل چندجمله‌ای

یک از متداول‌ترین روش‌ها برای برطرف کردن مشکل شناسایی‌پذیری در مدل چندجمله‌ای، استفاده از قید رده گوشه است که در آن پارامترهای مربوط به یکی از رده‌ها را برابر با صفر در نظر می‌گیرند. در نتیجه استفاده از این روش، افراد در عین روبرو بودن با J رده قادرند تنها پارامترهای مربوط به $J-1$ رده را برآورد کنند. برای مثال مدل (۳.۴) را در نظر بگیرید. اگر رده J، رده گوشه باشد، بنابراین برای همه iها، $\eta_{iJ}=0$ یا β_{kJ}ها و α_J مساوی با صفر است. بنابراین بقیه پارامترها باید نسبت به رده J تفسیر شوند. به‌کار گرفتن این روش در استنباط بیزی به این معنی است که توزیع پیشین مربوط به رده گوشه با رده‌های دیگر متفاوت است. بنابراین، پیشین قابلیت تغییرپذیری ندارد. در نتیجه نتایج مربوط به توزیع پسین تحت تاثیر انتخاب رده گوشه است (کانگدان، ۲۰۰۵). مشکل مشابهی را می‌توان در توزیع‌های با پاسخ عاملی نیز یافت (اگرستی، ۲۰۱۳) که در فصل سوم به شرح مفصل آن پرداختیم. لازم به ذکر است اگر یک رده با مشاهدات کمتر به‌عنوان رده پایه انتخاب شود، نتایج حاصل پایدار نیست (سید، ۲۰۲۰).

قید مجموع مساوی با صفر برای مدل چندجمله‌ای

قید مجموع مساوی با صفر از نظر مفهومی مشابه مرکزی کردن متغیر پیوسته است، که در آن میانگین را از متغیر پیشگو کم می‌کنند. برای مثال، در مدل (۳.۴) قید مجموع مساوی با صفر به صورت $\sum_j \beta_{kj}=0$ و $\sum_j \alpha_j=0$ خواهد بود. در این حالت، رده‌ها تغییرپذیر هستند. در این فصل از رساله به تحلیل و مقایسه قید پیشنهادی با حالت متداول یعنی قید رده گوشه می‌پردازیم.

مدل‌های بیزی سلسله‌مراتبی یکی از ابزارهای مهم روز در مدل‌بندی با ساختارهای پیچیده در فرآیندهای فضایی محسوب می‌شوند. در این حالت، معمولا تابع درستنمایی مدل صورت بسته‌ای ندارد و آماردانان از روش‌های شبیه‌سازی مونت کارلویی برای تقریب توزیع پسین استفاده می‌کنند. اما همان‌گونه که پیش‌تر اشاره شد، متاسفانه این روش از نظر محاسباتی گران است (ارمرود و وند، ۲۰۱۰؛ مینانسی و همکاران، ۲۰۱۱؛ میلدج و همکاران، ۲۰۱۲؛ بلی و همکاران، ۲۰۱۷). همچنین، به‌طور ویژه، برای مدل‌های چندجمله‌ای علاوه بر هزینه محاسباتی بالا، حتی ممکن است به همگرایی نرسد (دپراتره و وندبروک، ۲۰۱۷؛ بانسال و همکاران، ۲۰۲۰). در مقابل می‌توان از روش INLA، که در فصل دوم معرفی شد، بهره گرفت. با این وجود INLA تا کنون برای مدل پیشنهادی چندجمله‌ای به‌کار گرفته نشده است، زیرا در این مدل، متغیر پاسخ به صورت ناخطی به عناصر پیشگوی خطی وابسته است[۱۰]. برای رهایی از این مشکل از تبدیل چندجمله‌ای‌ـ‌پوآسن[۱۱] که اولین بار توسط بیکر (۱۹۹۴) معرفی شد،

[۱۰] https://www.r-inla.org/faq
[۱۱] Multinomial-Poisson transformation (MPT)

بهره می‌گیریم تا بتوانیم هسته درستنمایی مدل چندجمله‌ای را به هسته یک مدل پواسون تبدیل و از قابلیت روش INLA استفاده کنیم. البته این تبدیل هزینه هم دارد که در مقابل کارایی محاسباتی روش INLA ناچیز است.

۴.۱.۴ تبدیل چندجمله‌ای‌ـ پواسن

فرض کنید برای هر آزمایش i، $\{y_{ij},\ \ i=۱,\ldots,n,\ j=۱,\ldots,J\}$ تحقق‌هایی از Y_{ij}ها است که دارای توزیع چندجمله‌ای با پارامترهای احتمال $\{\pi_{ij},\ \ i=۱,\ldots,n,\ j=۱,\ldots,J\}$ هستند و توسط مدل (۲.۴) به متغیرهای تبیینی $\boldsymbol{x}$ متصل می‌شوند. تابع درستنمایی مدل به صورت

$$\mathrm{L}_M=\prod_{i=۱}^{n}\prod_{j=۱}^{J}\left(\frac{\exp(\eta_{ij})}{\sum_{s=۱}^{J}\exp(\eta_{is})}\right)^{y_{ij}},$$

قابل نمایش است. چالش اصلی هنگام کار کردن با L_M به دلیل وجود $\Gamma_i=\sum_{s=۱}^{J}\exp(\eta_{is})$ رخ می‌دهد، زیرا موجب می‌شود تا استفاده از لگاریتم بی‌اثر شود. با استفاده از تبدیل MP، Γ_i به‌عنوان یک پارامتر اضافی قابل برآورد جایگزین می‌شود. تبدیل حاصل به صورت

$$\mathrm{L}_P=\prod_{i=۱}^{n}\prod_{j=۱}^{J}\Big\{\exp(\eta_{ij}+\phi_i)\Big\}^{y_{ij}}\exp\{-\exp(\eta_{ij}+\phi_i)\},$$

است، که در آن $\boldsymbol{\phi}=(\phi_۱,\ldots,\phi_n)'$ بردار پارامترهای اضافه است که موجب می‌شود تا تبدیل MP موثر و قابل استفاده شود. تابع درستنمایی L_P تابعی از پارامترهای مدل و بردار $\boldsymbol{\phi}$ است. با استفاده از سرافینی (۲۰۱۸)، برآورد ماکسیمم درستنمایی پارامترهای ϕ_i در نسبت با Γ_i تعریف می‌شود، به‌طوری که $\exp(\phi_i)\propto\frac{۱}{\Gamma_i}$. بنابراین، نتایج مشاهده‌شده بر وجود رابطه $\mathrm{L}_M\propto\mathrm{L}_P$ تاکید می‌کند. قبل از تعبیه این نتیجه برای توسعه مدل بیزی چندجمله‌ای، برای تشریح مفید بودن این تبدیل، ابتدا چند مثال که تبدیل MP در آن‌ها مفید است را بیان می‌کنیم.

مثال ۱.۱.۴. مدل لگ‌خطی. پالمگرن (۱۹۸۱) حالت خاصی از تبدیل چندجمله‌ای‌ـ پواسن را برای مدل لگ‌خطی یافت. فرض کنید y_{ij} فراوانی مشاهده‌شده خانه ijام از یک جدول توافقی $I\times J$ را نشان دهد. فرض کنید $\{Y_{ij}\}$ دارای توزیع چندجمله‌ای با پارامترهای $\{\frac{\eta_{ij}}{\sum_j\eta_{ij}}\}$ است، که در آن $\eta_{ij}=\exp(\lambda^i+\lambda^j)$ شامل اثرات متغیرهای سطری و ستونی است. هسته تابع درستنمایی و تبدیل MP متناظر آن به صورت زیر هستند:

$$\mathrm{L}_{\mathrm{M}}=\prod_{i=۱}^{I}\prod_{j=۱}^{J}\left(\frac{\exp(\lambda^i+\lambda^j)}{\sum_j\exp(\lambda^i+\lambda^j)}\right)$$

$$Y_{ij}\sim\mathrm{Poi}(\exp(\phi_i+\lambda^i+\lambda^j)).$$

مثال ۲.۱.۴. وایتد (۱۹۸۰) تبدیل MP را برای یک مدل کاکس (کاکس، ۱۹۷۲) با متغیرهای تبیینی رسته‌ای پیشنهاد کرد. فرض کنید i اندیس مجموعه افراد در معرض خطر باشد که برای

آن $i = ۱, \ldots, n$، و j اندیس مربوط به سطوح متغیر تبیینی باشد، به‌گونه‌ای که $j = ۱, \ldots, J$. قرار می‌دهیم $y_{ij} = ۱$ اگر فرد i که در مجموعه خطر قرار دارد، دارای متغیر تبیینی سطح j باشد؛ در غیر این صورت $y_{ij} = ۰$. فرض کنید N_{ij} نشان‌دهنده افرادی با سطح jام متغیر تبیینی است که در مجموعه خطر i قرار دارند. در این صورت، هسته تابع درستنمایی به شکل

$$\mathrm{L} = \prod_{i=۱}^{n} \prod_{j=۱}^{J} \left(\frac{\exp(\boldsymbol{x_{ij}\beta})}{\sum_{u=۱}^{J} \exp N_{ij}(\boldsymbol{x_{iu}\beta})} \right),$$

است. برای این که بتوانیم از تبدیل MP استفاده کنیم، باید درستنمایی را در مقدار ثابت $\prod_i \prod_j N_{ij}^{y_{ij}}$ ضرب کنیم. لازم به ذکر است این عمل تغییری در برآورد ماکسیمم درستنمایی و واریانس مجانبی ایجاد نمی‌کند. بنابراین هسته درستنمایی و تبدیل MP به صورت زیر خواهند بود:

$$\mathrm{L_M} = \prod_{i=۱}^{n} \prod_{j=۱}^{J} \left(N_{ij} \frac{\exp(\boldsymbol{x_{ij}\beta})}{\sum_j \exp(\boldsymbol{x_{ij}\beta})} \right),$$
$$Y_{ij} \sim \mathrm{Poi}(\exp(\phi_i + \boldsymbol{x_{ij}\beta})).$$

مثال ۳.۱.۴. مدل راش تعمیم‌یافته: داروچ (۱۹۸۱)، مک کولاح (۱۹۸۲)، کرسی و هلاند (۱۹۸۳)، کانوی (۱۹۸۹، ۱۹۹۲) و اگرستی (۱۹۹۳) از تبدیل MP برای تعمیم مدل راش[۱۲] استفاده کردند. مدل راش تعمیم‌یافته برای داده‌های چندرده‌ای، احتمال این که فرد i به سوال k پاسخ j را بدهد، به صورت

$$\mathrm{P}(j \text{ رده}|i \text{ شخص}, k \text{ سوال}) = \frac{\exp(\alpha_{ij} + x_k\beta_j)}{۱ + \sum_{j=۲}^{J} \exp(\alpha_{ij} + x_k\beta_j)},$$

در نظر بگیرید (کانوی، ۱۹۸۹). شرطی کردن روی تعداد پاسخ‌ها در هر سطح باعث می‌شود تا پارامترهای مربوط به α حذف شوند. برای مثال، سه سطح از پاسخ را بر روی سه سوال در نظر بگیرید. فرض کنید فرد ۱ به ترتیب سطوح پاسخ ۲، ۳ و ۲ را برای سوال‌های ۱، ۲ و ۳ انتخاب کند. در این صورت، هسته درستنمایی برای فرد ۱ به صورت

$$\mathrm{L}^{(۱)} = \frac{\exp(\alpha_{۱۲} + x_۱\beta_۲)}{۱ + \sum_{j=۲}^{۳} \exp(\alpha_{۱j} + x_۱\beta_j)} \frac{\exp(\alpha_{۱۳} + x_۲\beta_۳)}{۱ + \sum_{j=۲}^{۳} \exp(\alpha_{۱j} + x_۲\beta_j)} \frac{\exp(\alpha_{۱۲} + x_۳\beta_۲)}{۱ + \sum_{j=۲}^{۳} \exp(\alpha_{۱j} + x_۳\beta_j)}$$

است. زیرا سه مجموعه ممکن از دو پاسخ در ۲ سطح و یک پاسخ در سطح ۳ به ترتیب $\{۲,۲,۳\}$، $\{۲,۳,۲\}$ و $\{۳,۲,۲\}$ وجود دارند. بنابراین، هسته درستنمایی شرطی برای فرد ۱ به صورت

$$\mathrm{L}_M^{(۱)} = \frac{\exp(x_۱\beta_۲ + x_۲\beta_۳ + x_۳\beta_۲)}{\exp(x_۱\beta_۲ + x_۲\beta_۲ + x_۳\beta_۳) + \exp(x_۱\beta_۲ + x_۲\beta_۳ + x_۳\beta_۲) + \exp(x_۱\beta_۳ + x_۲\beta_۲ + x_۳\beta_۲)},$$

است. هسته درستنمایی شرطی برای همه مشاهدات برابر است با

$$\mathrm{L}_M = \prod_i \mathrm{L}_M^{(i)},$$

[۱۲] Rasch

که با گروه‌بندی افراد می‌توان آن را به شکل ساده‌تری نوشت. فرض کنید s نشان‌دهنده یک مجموعه به شکل {تعداد پاسخ در سطح ۱، تعداد پاسخ در سطح ۲، تعداد پاسخ در سطح ۳} باشد. همچنین فرض کنید J_i زیرمجموعه‌های (اندیس‌گذاری‌شده با j) از سطوح پاسخ برای $k-1$ سوال که به مجموعه i مربوط می‌شود را نشان دهد. به‌عنوان نمونه، در مثال بالا $i=\{0,2,1\}$ و $J_{\{0,2,1\}}=\{\{2,2,3\},\{2,3,2\},\{3,2,2\}\}$. فرض کنید y_{ij} تعداد افرادی را نشان دهد که به زیرمجموعه $j\in J_i$ تعلق دارند. فرض کنید برای $k\in j$، $\delta_{kc}=1$، یعنی پاسخ برای k از زیرمجموعه j برابر c است؛ در غیر این صورت $\delta_{kc}=0$. هسته درستنمایی شرطی و تبدیل MP متناظر به ترتیب زیر هستند:

$$\mathrm{L}_M=\prod_{i=1}^{n}\prod_{j\in J_i}\left(\frac{\exp(\sum_{k\in g}\sum_c x_k\beta_c\delta_{kc})}{\exp(\sum_{u\in J_i}(\sum_{k\in g}\sum_c x_k\beta_c\delta_{kc})}\right)^{y_{ij}}$$

$$Y_{ij}\sim \mathrm{Pois}(\exp(\phi_i+\sum_{k\in j}\sum_c x_{kc}\boldsymbol{\beta}\delta_{kc})).$$

۵.۱.۴ تبدیل چندجمله‌ای‌ـ پواسن بیزی

بیکر (۱۹۹۴) نتایج کلی را در رابطه با معادل بودن برآورد ماکسیمم درستنمایی حاصل از درستنمایی پواسن با درستنمایی چندجمله‌ای و همچنین ویژگی‌های مجانبی واریانس‌های مربوط به آن‌ها ارائه داد. این نتایج به مدل‌های آماری متعددی قابل تعمیم است و توانست سال‌ها برای محققان سادگی محاسبات را به همراه بیاورد.

عقیده بیکر در سال ۲۰۰۴ توسط افرادی همچون سیمان و ریچاردسون در قالب مدل‌های کنترل پا به عرصه مدل‌بندی بیزی گذاشت. هدف آن‌ها رسیدن به نتایجی مشابه با نتایج بیکر اما در پیکره بیزی بود. قبل از آن در سال ۲۰۰۲، فردی به نام گاستافسون نتایج مشابهی را ارائه داده بود. سرانجام گاش و همکاران (۲۰۰۶) توانستند نشان دهند که در استنباط بیزی هم می‌توان از این تبدیل استفاده کرد. نتایج حاصل از آن در قالب قضیه زیر منتشر شد.

قضیه ۲.۱.۴. فرض کنید برای هر $i=1,\ldots,n$، $\sum_j y_{ij}\geq 1$. اگر برای ϕ_iها پیشین‌های مستقل ناسره، $\pi(\phi_i)\propto 1$، و برای سایر عناصر پیشگو پیشین‌های سره مستقل از $\boldsymbol{\phi}$ انتخاب شوند، آن‌گاه توزیع پسین حاصل از درستنمایی تبدیل چندجمله‌ای‌ـ پوآسن با توزیع پسین حاصل از درستنمایی چندجمله‌ای معادل است.

برهان. فرض کنید $\alpha_i=\exp(\phi_i)$، $i=1,\ldots,n$. در نتیجه $\pi(\alpha_i)\propto\alpha_i^{-1}$. همچنین فرض کنید پیشین $\pi(\boldsymbol{\beta})$ برای زمانی باشد که پیشگو تابعی از $\boldsymbol{\beta}$ است. توزیع پسین کناری $\boldsymbol{\beta}$ حاصل از

$\mathrm{L}_p(\boldsymbol{\phi},\boldsymbol{\beta})$ برابر با

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\beta})\prod_{i=1}^{n}\int_{0}^{\infty}\left(\alpha_i^{-1}\prod_j(\alpha_i g_{ij}(\boldsymbol{\beta}))^{y_{ij}}\exp(-\alpha_i g_{ij}(\boldsymbol{\beta}))d\alpha_i\right)\\
&= \pi(\boldsymbol{\beta})\prod_{i=1}^{n}\left(\prod_j(g_{ij}(\boldsymbol{\beta}))^{y_{ij}}\int_{0}^{\infty}\alpha_i^{\sum_j y_{ij}-1}\exp(-\alpha_i G_i(\boldsymbol{\beta}))d\alpha_i\right)\\
&\propto \pi(\boldsymbol{\beta})\prod_{i=1}^{n}\prod_j\left(\frac{g_{ij}(\boldsymbol{\beta})}{G_i(\boldsymbol{\beta})}\right)^{y_{ij}}\\
&= \pi(\boldsymbol{\beta})\mathrm{L}_{\mathrm{M}}(\boldsymbol{\beta})
\end{aligned}$$

است، که در آن $g_{ij}(\boldsymbol{\beta})$ برحسب توابعی از $\boldsymbol{\beta}$ و $G_i(\boldsymbol{\beta})=\sum_j g_{ij}(\boldsymbol{\beta})$ است و به وضوح با پسین $\boldsymbol{\beta}$ حاصل از $\mathrm{L}_{\mathrm{M}}(\boldsymbol{\beta})$ یکسان است. □

قضیه زیر سره بودن پسین بالا را در شرایط عمومی‌تر نشان می‌دهد.

قضیه ۳.۱.۴. فرض کنید $\sum_j y_{ij}\geq 1$، $i=1,\ldots,n$. اگر $\pi(\boldsymbol{\beta})$ سره باشد، آن‌گاه $\pi(\boldsymbol{\beta}|\boldsymbol{y})$ نیز سره است.

برهان. فرض کنید Θ فضای پارامتر $\boldsymbol{\beta}$ باشد، بنابراین با استفاده از قضیه بالا داریم

$$\int_{\boldsymbol{\beta}\in\Theta}\pi(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta}\propto\int_{\boldsymbol{\beta}\in\Theta}\pi(\boldsymbol{\beta})\mathrm{L}_{\mathrm{M}}(\boldsymbol{\beta})d\boldsymbol{\beta}<\infty$$

□

توجه داشته باشید اگر از پیشین $\pi(\alpha_i)\propto\alpha_i^{\alpha_i-1}(\alpha_i>0)$ در قضیه‌های بالا استفاده کنیم، آن‌گاه پسین حاصل دیگر معادل با $\pi(\boldsymbol{\beta})\mathrm{L}_{\mathrm{M}}(\boldsymbol{\beta})$ نخواهد بود، زیرا در این حالت در توان $G_i(\boldsymbol{\beta})$ عبارت $\sum_j y_{ij}+\alpha_i$ به جای $\sum_j y_{ij}$ قرار می‌گیرد.

در این رساله با در نظر گرفتن $\mathrm{L}_M\propto\mathrm{L}_P$ با پیشگوی

$$\eta_{ij}=\alpha_j+\sum_{k=1}^{l}x_{kij}\beta_{kj}+f(u_{ij}).$$

میانگین مدل پواسن بعد از تبدیل MP به صورت $\gamma_{ij}=\exp(\alpha_j+\sum_{k=1}^{l}x_{kij}\beta_{kj}+f(u_{ij})+\phi_i)$ است. همان‌گونه که قبلا هم بیان شد، برای مقایسه کاملتر و رسیدن به نتایج بهتر، مدل پیشنهادی را با مدل چندجمله‌ای کسری (لی، ۲۰۱۸) مقایسه می‌کنیم. برای این کار ابتدا مدل کسری را معرفی می‌کنیم.

۶.۱.۴ مدل چندجمله‌ای کسری

در این مدل متغیر پاسخ به‌عنوان کسری از وقایع اتفاق‌افتاده مورد نظر نسبت به همه حوادث رخداده تعریف می‌شود. فرض کنید w_{ij} نسبت وقایع اتفاق‌افتاده از رده jام برای آزمایش i

باشد، به‌گونه‌ای که $0 < w_{ij} < 1$ و $\sum_{j=1}^{J} w_{ij} = 1$. فرض کنید تابع w_{ij} کسری از (۴.۴) باشد به‌گونه‌ای که

$$\mathrm{E}(w_{ij}|\eta_{ij}) = \frac{\exp(\eta_{ij})}{\sum_{s=1}^{J}\exp(\eta_{is})} = G_j(\eta_{ij}).$$

به روشنی معلوم است که $0 < G_j(\cdot) < 1$ و $\sum_{j=1}^{J} G_j(\cdot) = 1$. برای توسعه یک نسخه بیزی از مدل چندجمله‌ای کسری به شکلی که با مدل چندجمله‌ای پیشنهادی در بخش قبل قابل مقایسه باشد، برای این مدل توزیع‌های پیشین یکسانی را در نظر می‌گیریم. برای رفع مشکل شناسایی‌پذیری در این مدل نیز از قید رده گوشه و برای برازش مدل به داده‌ها به علت عدم امکان اجرای برازش با روش INLA از الگوریتم‌های MCMC استفاده می‌کنیم.

۲.۴ مطالعه شبیه‌سازی

در این بخش عملکرد مدل پیشنهادی را با یک مثال شبیه‌سازی با تولید اثر فضایی و احتمال‌های مربوط به رده‌ها مورد بررسی قرار می‌دهیم. به همین منظور نواحی $\{i = 1, \ldots, n\}$ را بر روی یک مشبکه منظم 10×10 تعریف کردیم تا به کمک آن یک مدل چندجمله‌ای فضایی را شبیه‌سازی کنیم. پیشگوی خطی

$$\eta_{ij} = \alpha_j + x_{ij}\beta_j + f(u_{ij}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, J, \tag{۷.۴}$$

را در نظر گرفتیم، که در آن $n = 100$ تعداد نواحی و $J = 3$ تعداد رده‌های پاسخ هستند. اثر فضایی، $f(u_{ij})$، برای هر رده از مدل (۵.۴) با $\tau_u = 0/1$ پیروی می‌کند. برای تولید مشاهدات پاسخ از مدل لجیت چندجمله‌ای با پیشگوی خطی (۷.۴) استفاده کردیم که در آن β_jها از توزیع نرمال استاندارد تولید و به گونه‌ای مقیاس‌بندی شدند تا جمع آن‌ها برابر صفر شود. همچنین $\boldsymbol{\alpha} = (-1, -1, 2)$ تعیین شد. متغیر تبیینی x_j از توزیع یکنواخت استاندارد برای $j = 1, 2, 3$ شبیه‌سازی شد. بعد از استفاده از تبدیل چندجمله‌ای‌ـ‌پواسن، میانگین توزیع پواسن به صورت

$$\gamma_{ij} = \exp(\alpha_j + x_{ij}\beta_j + f(u_{ij}) + \phi_i),$$

است. از توزیع پیشین تعریف‌شده گاما در روش INLA برای τ_u و از توزیع‌های پیشین گوسی مبهم مستقل با میانگین صفر و واریانس بزرگ 10^5 برای $\{\alpha_j\}$، $\{\beta_j\}$ و $\{\phi_i\}$ استفاده کردیم. برای قید رده گوشه، $\alpha_1 = \beta_1 = 0$ و $f(u_{i1}) = 0$ تعیین شدند. همانند مثال مدل نرمال عاملی در فصل سوم، با داشتن برآوردهای سایر رده‌ها و با انجام محاسبات معمولی، می‌توان برآورد رده گوشه را نیز به‌دست آورد. فرض کنید β_2^* و β_3^* برآوردهای ضرایب رگرسیونی برای رده‌های دوم و سوم به ازای $\beta_1 = 0$ باشند. در این صورت

$$\beta_1 = \frac{-1}{3}\beta_2^* + \frac{-1}{3}\beta_3^*, \quad \beta_2 = \frac{2}{3}\beta_2^* + \frac{-1}{3}\beta_3^*, \quad \beta_3 = \frac{-1}{3}\beta_2^* + \frac{2}{3}\beta_3^*. \tag{۸.۴}$$

همچنین می‌توان روابط مشابهی را برای α_j و $f(u_{ij})$ یافت.

برای قید مجموع مساوی با صفر قرار می‌دهیم

$$\sum_j f(u_{ij}) = ۰, \quad \sum_j \beta_j = ۰, \quad \sum_j \alpha_j = ۰.$$

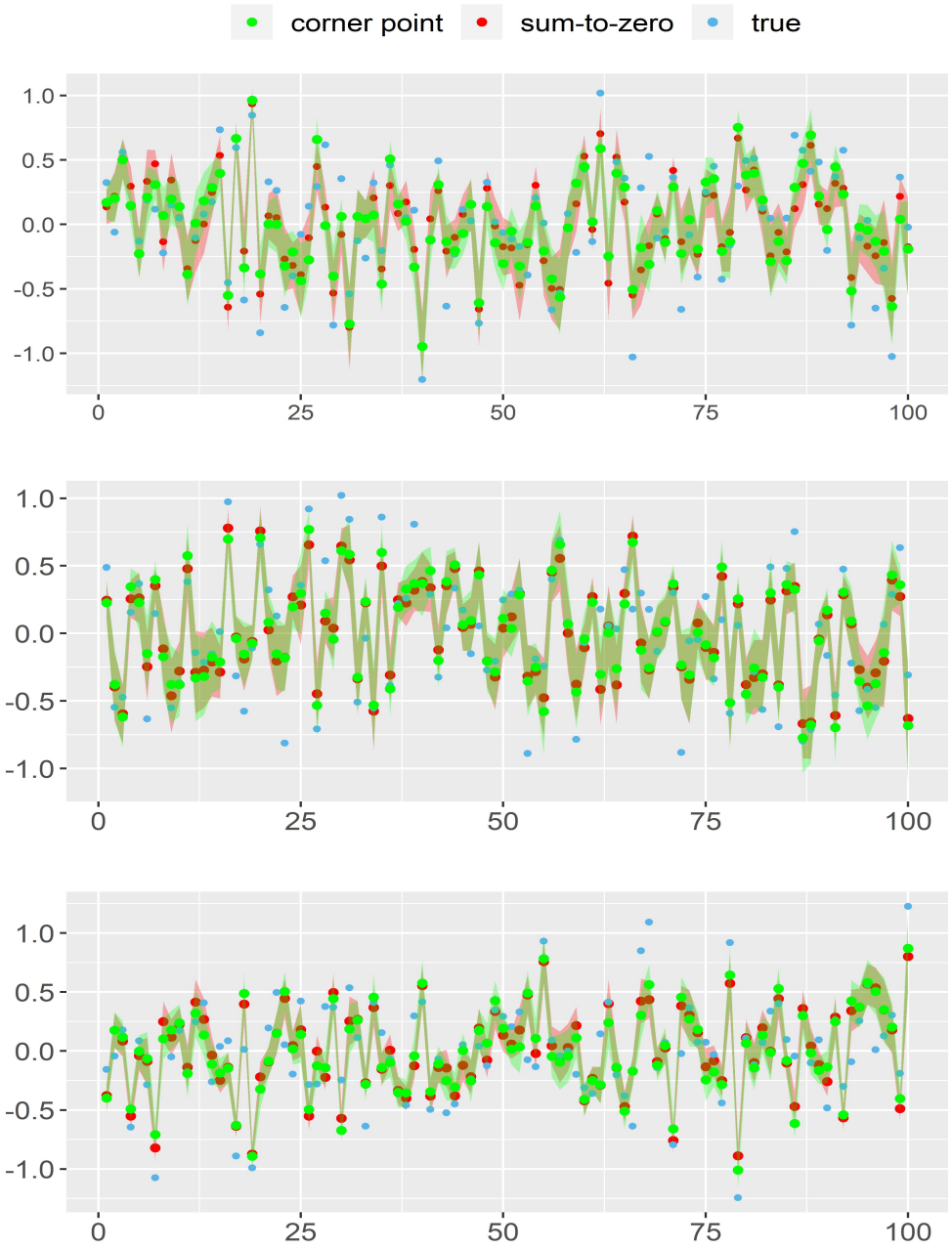
در هر دو این موارد، قید $\sum_i f(u_{ij}) = ۰$ نیز به‌کارگرفته شد تا مدل CAR قابل شناسایی باشد. نتایج مربوط به برآورد اثر فضایی (شکل ۳.۴) در اکثر نواحی، تفاوت کمی را بین مقادیر برآوردشده و مقدار واقعی نشان می‌دهد. احتمال‌های برآوردشده در هر رده در شکل ۴.۴ برای هر دو قید شناسایی‌پذیری مدل چندجمله‌ای، مقادیر نزدیک به مقادیر شبیه‌سازی‌شده را نشان می‌دهد. چگالی‌های کناری پسین پارامتر τ_u برای هر دو نوع قید نیز در شکل ۵.۴ نمایش داده شده‌اند. برآورد τ_u برای قید مجموع مساوی با صفر تقریبا دقیق است، در حالی‌که برای قید رده گوشه بیش‌برآورد و حتی ناسازگار برآورد شده است.

جدول ۱.۴: برآورد میانگین (انحراف استاندارد) پسینی برای پارامترهای مدل با استفاده از دو قید شناسایی. از روابط (۸.۴) استفاده شده است تا برآورد پارامترهای رده گوشه بازیابی شوند.

پارامتر	مقدار واقعی	مجموع مساوی با صفر	رده گوشه
$\alpha_۱$	−۱	−۰/۹۰۴ (۰/۰۰۸)	−۱/۱۱۷ (۰/۰۰۹)
$\alpha_۲$	−۱	−۱/۰۹۷ (۰/۰۰۸)	−۰/۹۰۱ (۰/۰۰۹)
$\alpha_۳$	۲	۱/۹۷۲ (۰/۰۰۲)	۲/۰۲۰ (۰/۰۰۳)
$\beta_۱$	−۰/۱۴	−۰/۱۳۱ (۰/۰۰۲)	−۰/۱۰۰ (۰/۰۰۳)
$\beta_۲$	۱/۲۶	۱/۲۵۳ (۰/۰۰۵)	۱/۳۹۷ (۰/۰۰۶)
$\beta_۳$	−۱/۱۱	−۱/۱۱۸ (۰/۰۰۶)	−۰/۹۸۶ (۰/۰۰۶)
τ_u	۰/۱۰	۰/۱۲۱ (۰/۰۱۱)	۲/۷۰۰ (۰/۲۸۰)

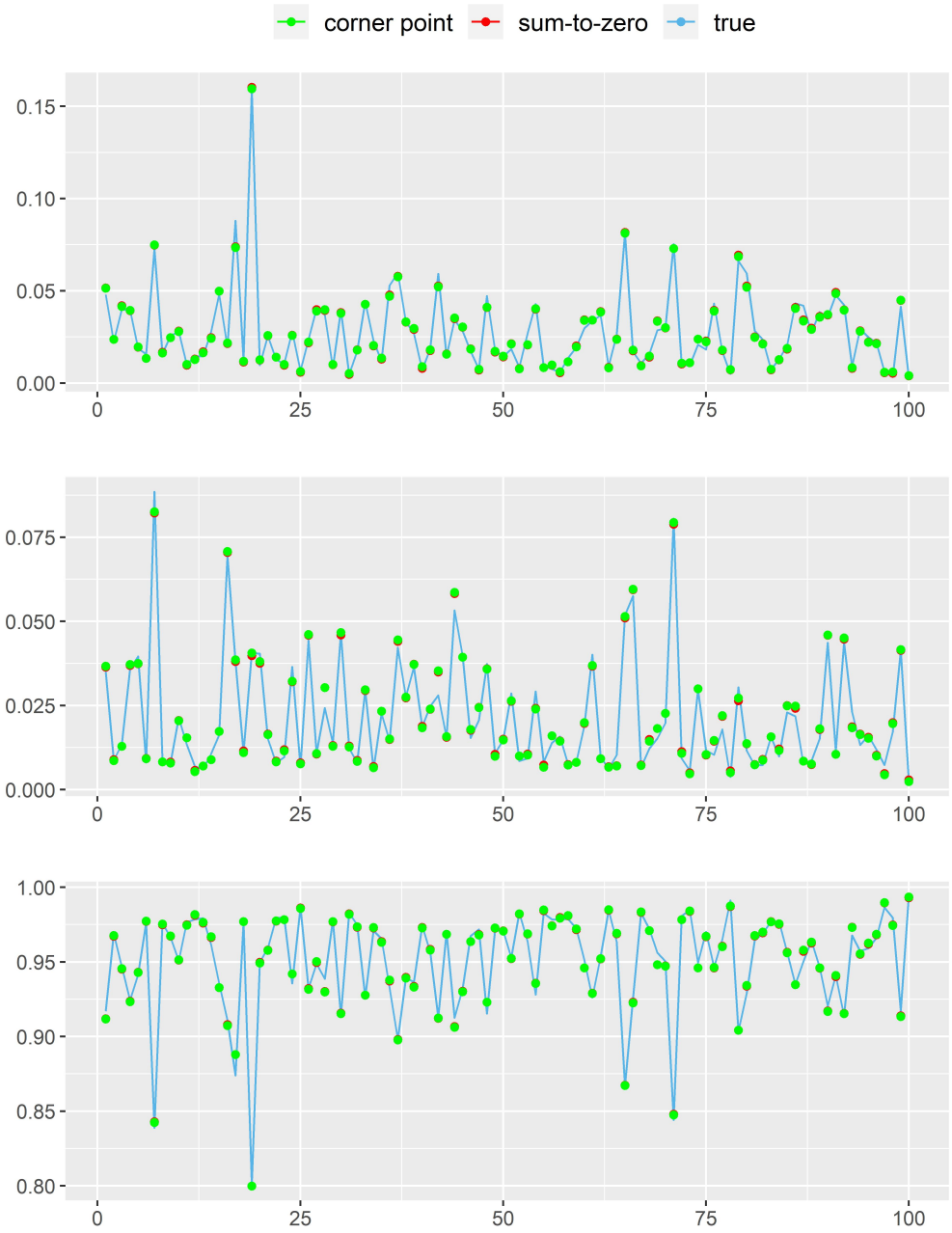
جدول ۱.۴ برآوردهای پسینی پارامترهای مدل شامل ضرایب رگرسیونی و پارامتر دقت اثر فضایی، τ_u، را نشان می‌دهد. برآوردها برای قید مجموع مساوی با صفر تقریبا به مقادیر واقعی پارامترها نزدیک است. در مقابل، تفاوت بیشتری بین برآوردهای حاصل از به‌کارگیری قید رده گوشه با مقادیر واقعی به چشم می‌خورد.

در تفسیر قید مجموع مساوی با صفر همه پارامترها در تناسب با میانگین تفسیر می‌شوند که تقریباً دقیق برآورد شده‌اند. بنابراین انحراف استاندارد پسینی برای برآوردها کوچکتر خواهد بود. اثرات در رده گوشه با توجه به رده‌ای که برای این کار انتخاب می‌شود، برآورد می‌شوند. در واقع محققان یا آماردانان باید باورهای خود را نسبت به تفاوت‌های بین رده‌ها تعریف و تعیین کنند. در مقابل، روش مجموع مساوی با صفر هیچ رده خاصی را به‌طور ویژه در نظر نمی‌گیرد که خود موجب بهبود بخشیدن تفسیر می‌شود، زیرا مقدار متغیر پاسخ را در میانگین متغیرهای تبیینی بیان می‌دارد. در رده گوشه برای مثلا رده j، در تفسیر یک متغیر تبیینی با ثابت نگه داشتن سایر متغیرهای تبیینی، انتظار می‌رود با برآورد پارامتر مربوط عوض شود. در قید مجموع مساوی با صفر می‌توان برای هر رده توزیع پیشین تعریف کرد تا از مشکل تغییرپذیری در قید رده گوشه جلوگیری کرد.



شکل ۳.۴: میانگین پسین اثر فضایی در هر رده: نواحی سایه‌زده‌شده کران اعتبار بیزی ۹۵% را نشان می‌دهد. از بالا به پایین به ترتیب متعلق به رده‌های اول تا سوم است.

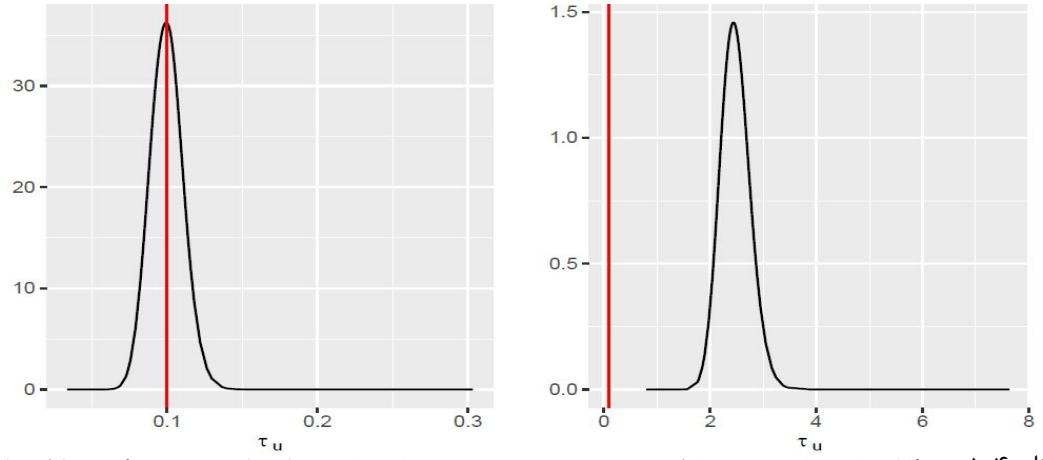
یافته‌ها در این قسمت با نتایج به‌دست آمده از شبیه‌سازی مدل نرمال عاملی در فصل سوم یکسان هستند.



شکل ۴.۴: برآورد احتمال هر رده: از بالا به پایین به ترتیب متعلق به رده‌های اول تا سوم است.

۳.۴ مثال کاربردی: شدت تصادفات جاده‌ای

امنیت در جاده‌ها یکی از بحث‌های کلیدی در بسیاری از کشورها است. تصادفات، علاوه بر خسارت‌های جانی و مالی خواه با مرگ همراه باشد یا نباشد، عواقب و هزینه‌های بعضاً جبران‌ناپذیری برای مردم و حکومت‌ها در پی دارند. ضمن این که ضربه‌های اجتماعی و روانی ناشی از تصادفات که حتی ممکن است سال‌ها بعد از آن ادامه یابد را نمی‌توان نادیده گرفت. بنابراین شناسایی عوامل ایجاد تصادف به‌ویژه در کشورهای پیشرفته توجه زیادی را به خود



شکل ۵.۴: چگالی‌های کناری پسین پارامتر دقت τ_u: قید مجموع مساوی با صفر (چپ)، و قید رده گوشه (راست). خط عمودی مقادیر واقعی را نشان می‌دهد.

جلب کرده است (سروری، ۲۰۲۰). این در حالی است که به بیان بانک جهانی تقریبا ۹۳ درصد از تصادفات در کشورهای در حال توسعه اتفاق می‌افتد (آمبو و همکاران، ۲۰۲۰). با این وجود، با یک فقدان بزرگ تحقیقاتی در این دسته از کشورها روبرو هستیم.

یک بخش کلیدی در تحلیل داده‌های مربوط به تصادف، کشف الگوهای مربوط به شدت جراحات و آسیب‌ها است (لیو و همکاران، ۲۰۲۰). پژوهشگران در بسیاری از کاربردهای مربوط به تحلیل داده‌های تصادف، مدل‌های گسسته را برای تحلیل شدت جراحات و آسیب‌ها به‌کار می‌گیرند (پنمتسا و همکاران، ۲۰۱۷). برای مثال، فارمر و همکاران (۱۹۹۷) اثرات مربوط به نوع ماشین و ویژگی‌های تصادف بر روی شدت جراحات را با کمک توزیع دوجمله‌ای مدل‌بندی کردند. دنل و همکاران (۱۹۹۶) یک پاسخ با چهار رده شدت جراحات را به کمک مدل لجیت ترتیبی و پروبیت ترتیبی تحلیل کردند. کالکمن و همکاران (۲۰۰۲) تحلیل تصادف‌های حاصل از برخورد دو ماشین به کمک مدل‌های احتمالی ترتیبی را بررسی کردند. همچنین، سای و همکاران (۲۰۰۹) از مدل بیزی پروبیت ترتیبی برای تحلیل داده‌های مربوط به جراحات ناشی از تصادف استفاده کردند. تعدادی از پژوهشگران از مدل لجیت چندجمله‌ای برای تحلیل این دسته از داده‌ها استفاده کردند که از جمله آن‌ها می‌توان به کارسون و همکاران (۲۰۰۱)، بریتمن و همکاران (۲۰۰۷)، شهید و همکاران (۲۰۱۳)، دانگ و همکاران (۲۰۱۳)، پنمتسا و همکاران (۲۰۱۷)، رضاپور و همکاران (۲۰۱۹) و وداگاما و همکاران (۲۰۲۰) اشاره کرد. لی و همکاران (۲۰۱۹) با در نظر گرفتن داده‌های پاسخ به‌صورت کسری از ماشین‌های تصادفی، مدل لوجیت کسری را به‌کار گرفتند. از تحقیقات مربوط به این زمینه در سال‌های اخیر هم می‌توان به زنگ و همکاران (۲۰۱۹) اشاره کرد که در آن مدل لجیت ترتیبی تعمیم‌یافته برای مطالعه فراوانی تصادفات در آزادراه‌ها به‌کار گرفته شد.

مرور پژوهش‌های ذکرشده در بالا و مدل‌های به‌کار گرفته‌شده در آن‌ها به خوبی نشان می‌دهد که سطح‌بندی یا رده‌بندی میزان جراحات و آسیب‌های ناشی از تصادف همیشه مد نظر محققان بوده است، به‌طوری که تحقیقات فنگ و همکارن (۲۰۱۶) و بهنود و همکاران

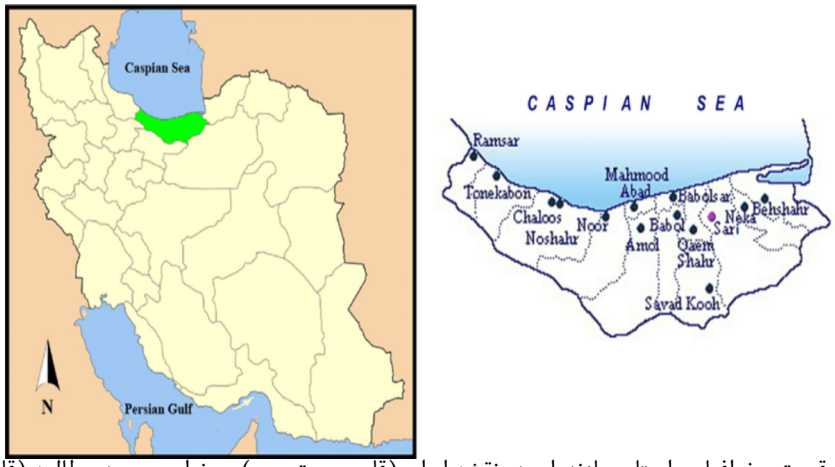
(۲۰۱۶) تاثیرات متفاوت ناشی از جراحات را خاطرنشان می‌کند. مدل لوجیت چندجمله‌ای خواه با رده‌های ترتیبی و خواه بدون ترتیب و همچنین مدل پروبیت ترتیبی برای مدل‌بندی سطوح مختلف شدت جراحات در داده‌های تصادف به دفعات فراوان مورد استفاده، نقد و بررسی قرار گرفته‌اند. ترتیب طبیعی شدت جراحات (از سطح بدون آسیب تا سطح آسیب جدی و مرگ) موجب شده است تا تحلیل‌گران متعددی از مدل لوجیت یا پروبیت ترتیبی به‌عنوان یک مدل احتمالی ترتیبی برای این دسته از داده‌ها استفاده کنند. یاماموتو و همکاران (۲۰۰۴)، ابدل (۲۰۰۳)، وئون و همکاران (۲۰۰۳)، کاتک و همکاران (۲۰۰۲)، کاتک و همکاران (۲۰۰۱)، کالک و همکاران (۲۰۰۲)، دانل و همکاران (۱۹۹۶)، الرو و همکاران (۲۰۱۲)، هائو و همکاران (۲۰۱۴) و زنگ و همکاران (۲۰۱۹)، تعدادی از محققان در این زمینه می‌باشند که از مدل‌های احتمالی ترتیبی استفاده کردند. با این وجود و برخلاف این محبوبیت، تعدادی محدودیت جدی بر این مدل‌ها حکمرانی می‌کند (ساولین، ۲۰۱۷). به‌طور مثال، تصادفات با شدت جراحات کمتر یا بدون آسیب‌دیدگی معمولا گزارش نمی‌شوند که خود موجب اریبی در برآورد مدل‌های ترتیبی می‌شود (سالوم و همکاران، ۲۰۱۹). در مقابل، این مسئله در مدل‌های غیر ترتیبی ممکن است اثر ناچیزی را از خود به جای بگذارد (جانز و همکارن، ۲۰۱۳؛ سالوم و همکاران، ۲۰۱۹). علاوه بر این، برآوردهای رگرسیونی حاصل از مدل لجیت غیرترتیبی[۱۳] سازگار هستند (مکفادن، ۱۹۸۱ و واشنگتن، ۲۰۰۳). بنابراین مدل چندجمله‌ای غیرترتیبی (اسمی) می‌تواند وجود برآوردهای سازگار از ضرایب رگرسیونی را برخلاف این عدم گزارش کامل تصادفات برای محققان تضمین کند. محدودیت دومی که محققان در استفاده از مدل‌های ترتیبی با آن روبرو هستند، تاثیری است که بر روی متغیرهای تبیینی دارد (سولینن و منرینگ، ۲۰۰۷). به این معنی که متغیری که باعث افزایش احتمال در رده با بیشترین جراحت می‌شود، باعث کاهش احتمال مربوطه در رده با کمترین جراحت می‌شود (سولینن و منرینگ، ۲۰۰۷؛ جیدایپالی و همکاران، ۲۰۱۱). بنابراین، استفاده از مدل‌های غیرترتیبی مثل مدل چندجمله‌ای برای تحلیل شدت جراحت در داده‌های تصادف پیشنهادی مناسب محسوب می‌شود اگرچه ذات این دسته از داده‌ها ترتیبی باشد (سالوم و همکاران، ۲۰۱۹). همچنین سالینن و همکاران (۲۰۱۱)، مدل‌های مختلفی را مطالعه کرده و سرانجام یافتند که مفیدترین مدل برای مدل‌بندی گسسته داده‌های مربوط به شدت جراحات در تصادف، مدل لوجیت غیرترتیبی است.

برای اطمینان از این که داده‌ها قابل برازش به مدل چندجمله‌ای باشد، از آزمون استقلال جانشین‌های نامرتبط[۱۴] (سالوم و همکاران، ۲۰۱۹) بهره گرفتیم. این ویژگی نشان می‌دهد نسبت احتمالات برای هر دو رده از شدت جراحات یکسان است اگرچه رده‌های دیگر در دسترس باشند یا خیر. در واقع این ویژگی یکی از مفروضات مهم توزیع چندجمله‌ای محسوب می‌شود. بنابراین یافتن این که ویژگی IIA برای داده‌های ما برقرار است یا نه امری ضروری

[۱۳] Non-ordred multinomial
[۱۴] Independence of irrelevant alternatives (IIA)

محسوب می‌شود که در این بخش مورد بررسی قرار خواهد گرفت.

داده‌های تصادف به‌گونه‌ای اثر فضایی را از خود نشان می‌دهند که با کمک متغیرهای تبینی قابل توصیف نیست. علاوه بر این، اثرات مربوط به متغیرهای تبیینی خود ممکن است به‌طور فضایی تغییر کنند. بنابراین، مدل‌های فضایی انتخاب طبیعی برای این دسته از داده‌ها محسوب می‌شوند (زیاکوپولوس و یانیس، ۲۰۲۰). در این رساله، مدل چندجمله‌ای بیزی سلسله‌مراتبی و از نوع فضایی را برای تحلیل داده‌های رده‌ای تصادف به‌کار می‌گیریم. با این هدف و با در نظر گرفتن همه جنبه‌های محل حادثه، شامل آسیب دیدن افراد و وسایل، سه رده از شدت آسیب و جراحات در تصادف را در نظر می‌گیریم. فهمیدن بهتر داده‌های تصادف به سیاستمداران این عرصه کمک می‌کند تا سیاست‌های مناسبی را برای کاهش روند تصادفات جاده‌ای انتخاب کنند. برای این کار داده‌های تصادف استان مازندران را مورد مطالعه قرار می‌دهیم (شکل ۶.۴ از پیردشتی و همکاران، ۲۰۱۵).



شکل ۶.۴: موقعیت جغرافیایی استان مازندران در نقشه ایران (قاب سمت چپ)، و نواحی مورد مطالعه (قاب سمت راست)

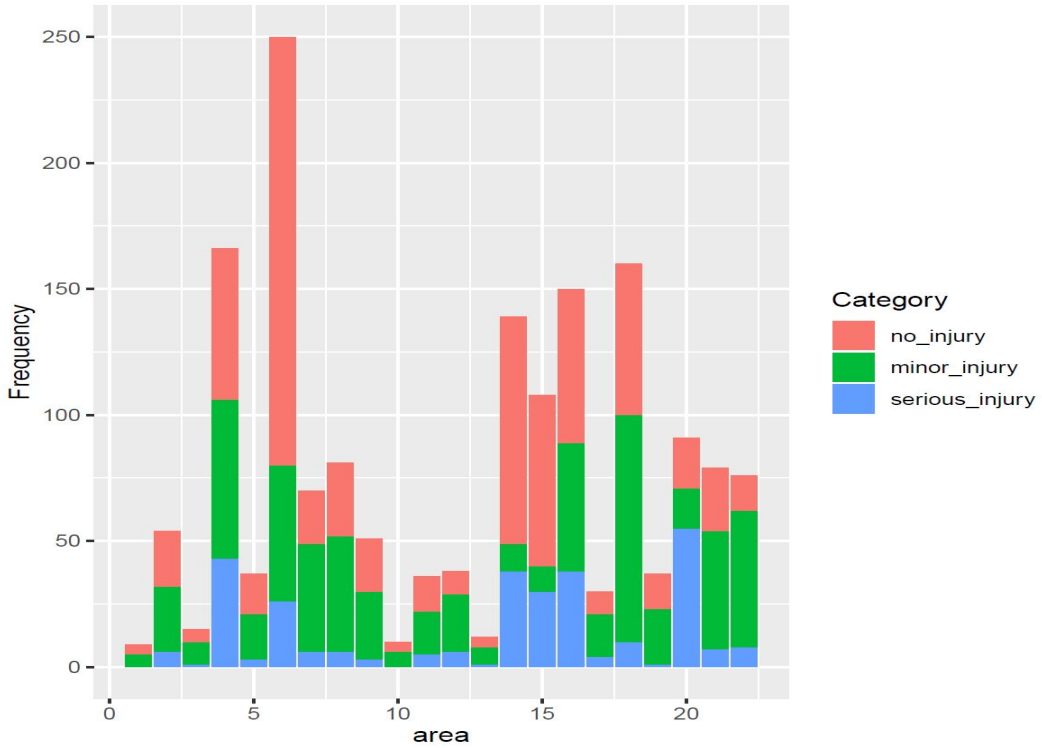
مازندران از نظر زمین‌شناسی و جغرافیایی، به دو بخش نواحی کوهی و جنگل‌های ساحلی تقسیم می‌شود. مساحت این استان ۲۳۸۴۲ کیلومتر مربع است. مازندران دارای ۲۲ مرکز فرمانداری است که به‌عنوان نواحی تصادف در نظر گرفته شده‌اند. مجموعه داده مورد نظر مربوط به تصادفات ثبت‌شده طی سال‌های ۲۰۱۶ تا ۲۰۱۸ در این مراکز است. در این مجموعه فقط تصادف‌های با اتومبیل در نظر گرفته شده‌اند. پس از چندی تلاش و پیگیری همه جانبه، به علت برخی مسایل امنیتی و نبودن اطلاعات ثبت‌شده کافی و جامع به تنها سه متغیر کمکی سرعت رانندگی، سن راننده (عامل تصادف) و نسبت تصادفات در روز نسبت به شب محدود شدیم. میزان جراحات افراد و خرابی ماشین‌ها و محیط پیرامون، به‌عنوان منبع تقسیم‌بندی شدت جراحات و رده‌بندی مورد استفاده قرار گرفتند. در نهایت سه دسته تصادفات به کمک مرکز پلیس به شرح زیر رده‌بندی شد:

۱. بدون آسیب (no injury): به تصادفی اشاره دارد که در آن افراد آسیب ندیده‌اند و خسارتی کمتر از ۳ میلیون تومان در پی داشته باشد.

۲. آسیب کم (minor injury): به تصادفی اشاره دارد که در آن افراد آسیب جدی ندیده یا با جراحات کمی مانند شکستگی‌های ساده روبرو شده‌اند. همچنین خسارات مالی ناشی از آن بین ۳ تا ۹ میلیون تومان باشد.

۳. آسیب جدی (serious injury): به تصادفی اشاره دارد که در آن افراد به‌طور جدی آسیب دیده یا با مرگ همراه شده‌اند. همچنین خسارت‌های مالی ناشی از آن بیشتر از ۹ میلیون تومان باشد.

از بین داده‌هایی که ما برای تحلیل و بررسی جدا کردیم، ۱۹٪ مربوط به آسیب‌دیدگی جدی، ۳۸٪ مربوط به آسیب کم و ۴۳٪ مربوط به تصادفات بدون آسیب است. فراوانی‌های مربوط به این رده‌ها در شکل ۷.۴ نمایش داده شده‌اند.



شکل ۷.۴: توزیع فراوانی مربوط به رده‌های جراحت برای همه ۲۲ ناحیه استان مازندران.

میانگین و انحراف استاندارد متغیرهای تبیینی انتخاب‌شده برای هر رده در جدول ۲.۴ گزارش شده‌اند. بنابراین با یک متغیر پاسخ سه رده‌ای و سه متغیر تبیینی روبرو هستیم.

جدول ۲.۴: میانگین (انحراف استاندارد) متغیرهای تبیینی در هر رده از پاسخ

متغیرهای کمکی	بدون آسیب	آسیب کم	آسیب دیدگی جدی
میانگین سرعت ماشین(km/h)	۵۶/۹۷ (۱۶/۶۸)	۶۷/۸۷ (۱۷/۱۹)	۸۹/۶۱ (۱۷/۶۵)
میانگین سن	۳۳/۰۶ (۷/۹۸)	۳۳/۹۹ (۹/۶۰)	۳۵/۷۰ (۱۰/۱۸)
نسبت تصادفات روز به شب	۰/۵۸ (۰/۳۰)	۰/۶۴ (۰/۲۵)	۰/۵۷ (۰/۲۳)

برای رسیدن به اطمینان از مفید بودن مدل پیشنهادی برای مجموعه داده مد نظر، برقرار

بودن ویژگی IIA را با کمک آزمون هاوسمن‌ـ‌مک‌فادن[۱۵] (هاسمن و مک‌فادن، ۱۹۸۴) بررسی کردیم. وقتی که این ویژگی برقرار باشد، برآوردهایی که با در نظر گرفتن تعدادی از رده‌ها به‌دست می آید، تفاوت معناداری با برآوردی که از همه رده‌ها با هم به‌دست آید، ندارد. بنابراین آزمون HM برآورد سازگار و کارای حاصل از مدل کامل، $\hat{\beta}^f$ را با برآورد سازگار اما ناکارا از مدل محدودشده، $\hat{\beta}^r$ را با استفاده از آماره

$$\chi^2_{\text{HM}} = (\hat{\beta}^r - \hat{\beta}^f)'\{\hat{V}(\hat{\beta}^f) - \hat{V}(\hat{\beta}^r)\}^{-1}(\hat{\beta}^r - \hat{\beta}^f),$$

بررسی می‌کند، که در آن $\hat{V}(\hat{\beta}^f)$ و $\hat{V}(\hat{\beta}^r)$ به ترتیب ماتریس کوواریانس‌های برآوردشده برای مدل کامل و مدل محدودشده هستند. در این‌جا χ^2_{HM} فرض صفر این که پارامترهای مدل محدودشده با پارامترهای مدل کامل یکسان هستند را بررسی می‌کند که از توزیع کی‌دو با درجه آزادی برابر با متغیرهای برآوردشده در مدل محدودشده پیروی می‌کند. اگر فرض صفر رد نشود، یعنی وقتی که سطح معنی‌داری بیشتر از سطح آزمون (مثلا ۵٪) باشد یا آماره آزمون χ^2_{HM} منفی باشد، آن‌گاه ویژگی IIA برقرار است و مدل لجیت چندجمله‌ای مناسب است (زنگ، ۲۰۱۱). نتایج حاصل از آزمون HM در جدول ۳.۴ گزارش شده‌اند که به مناسب بودن مدل چندجمله‌ای برای مجموعه داده شدت تصادف‌های مازندران اشاره دارد.

جدول ۳.۴: نتایج آزمون HM برای بررسی ویژگی IIA در مجموعه داده تصادف‌های مازندران. علامت * سطح معنی‌داری برابر ۱ را نشان می‌دهد.

رده حذف‌شده	χ^2_{HM}	فرضیه صفر	ویژگی IIA
آسیب کم	−۵/۳۲	رد نشد	برقرار است
آسیب کم	۳/۰۶*	رد نشد	برقرار است
آسیب دیدگی جدی	−۲/۵۶	رد نشد	برقرار است

در مطالعات گذشته (برای مثال به رگرو و همکاران (۲۰۱۸) رجوع کنید)، نشان داده شده است سن اثر ناخطی با تصادفات دارد. علاوه بر این، رابطه همبستگی پیرسن بین سن و سرعت حدود ۲۱% برآورد شد که به نظر می‌رسد قابل اغماض باشد. به همین دلایل، به کمک فرآیند قدم زدن تصادفی، اثر متغیر تبیینی سن را با دو هدف به صورت ناپارامتری مدل‌بندی کردیم. هدف اول، دوری از هم‌خطی احتمالی بین سن و سرعت است. هدف دوم، لحاظ کردن رابطه ناخطی ممکن بین متغیر سن و متغیر پاسخ است. در آغاز، مدل را با همه متغیرهای تبیینی موجود در داده‌ها برازش دادیم. متغیر نسبت تصادفات روز به شب اثر معنی‌داری را از خود نشان نداد. بنابراین مدل را بدون حضور این متغیر تبیینی دوباره برازش دادیم. نتیجه دور از هم انتظار نیست، زیرا امروزه به‌دلیل پیشرفت در صنعت روشنایی جاده‌ای و به‌خصوص لامپ‌های مه‌شکن، لامپ‌های ایستاده با نور مناسب در خیابان‌ها به خوبی نصب شده‌اند.

[۱۵] Hausman-McFadden (HM)

سرانجام، پیشگوی مدل

$$\eta_{ij} = \alpha_j + \text{speed}_{ij}\beta_j + f(\text{age}_{ij}) + f(u_{ij}), \quad i = ۱, \ldots, ۲۲ \quad j = ۱, ۲, ۳,$$

برای تحلیل نهایی به‌کار گرفته شد، که در آن $f(u_{ij})$ مدل فضایی ICAR و $f(\text{age}_{ij})$ اثر هموار سن را مدل‌بندی می‌کند. برای برآورد $f(\text{age})$، از فرآیند قدم زدن تصادفی مرتبه دوم استفاده کردیم. این فرآیند در عمل و کاربرد با یک مدل اسپلاین معادل است (رو و هلد، ۲۰۰۵). برای به‌کار گرفتن قید رده گوشه، رده بدون آسیب را به‌عنوان رده گوشه یا پایه انتخاب کردیم. توزیع‌های پیشین پارامترهای مدل را مشابه مثال شبیه‌سازی در نظر گرفتیم. برای تقریب توزیع پسین در مدل چندجمله‌ای کسری نیز از یک الگوریتم نمونه‌گیری متروپولیس‌ـ‌هستینگز (هستینگز، ۱۹۷۰) استفاده و با استفاده از بسته `mcmcpack` (مارتین و همکاران، ۲۰۱۱) در نرم‌افزار `R` (تیم هسته مرکزی `R`، ۲۰۲۱) آن را اجرا کردیم. برازش مدل بیزی چندجمله‌ای کسری، حدود دو ساعت و نیم به طول انجامید، در حالی‌که اجرای مدل چندجمله‌ای با ساختار STAR با روش INLA کمتر از ۴ دقیقه طول کشید. همگرایی زنجیر MCMC تولیدشده برای مدل کسری نیز به کمک بسته نرم‌افزاری `coda` (کولس و کارلین، ۱۹۹۶) مورد بررسی قرار گرفت. نتایج برآورد مدل در جدول ۴.۴ گزارش شده‌اند. نتایج حاصل برای قید مجموع مساوی با صفر و مدل چندجمله‌ای کسری بیشتر قابل مقایسه هستند؛ به‌ویژه در برآورد پارامتر دقت τ_u. مشابه مطالعه شبیه‌سازی، قید رده گوشه عملکرد متفاوتی دارد که قابل اتکا نیست.

۱.۳.۴ بررسی مدل‌ها از نظر قدرت پیشگویی

توزیع پیشگو برای مشاهده جدید z برابر است با $\pi(z|\mathbf{y}) = \int_{\boldsymbol{\theta}} \text{p}(z|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$. توزیع پیشگو برای ارزیابی مدل از نظر قدرت پیشگویی مناسب است. برای ارزیابی توانایی پیشگویی مدل‌های ارائه‌شده، مقادیر مولفه پیشگوی شرطی[۱۶] (پتیت، ۱۹۹۰) را برای مشاهدات می‌توان حساب کرد که بر اساس توزیع پیشگو محاسبه می‌شوند. برای مشاهده iام مقدار CPO برابر

$$\text{CPO}_i = \pi(y_i|\boldsymbol{y}_{-i}),$$

است، که در آن $\boldsymbol{y}_{-i}$ بردار تمام مشاهدات بدون $\boldsymbol{y}_i$ است. در واقع، CPO مقدار احتمال پسین مشاهده y_i را بر اساس مدلی که بر روی همه مشاهدات بدون حضور مشاهده y_i برازش شده است، محاسبه می‌کند. مقادیر بزرگتر CPO نشان‌دهنده برازش بهتر مدل هستند. برای تجمیع همه اطلاعات CPO ها در یک معیار عددی، می‌توان امتیاز لگاریتمی[۱۷] را به‌کار برد. امتیاز لگاریتمی به صورت

$$\text{log.score} = -\sum_i \log(\text{CPO}_i),$$

[۱۶] Conditional predictive ordinate (CPO)
[۱۷] Logarithmic score

جدول ۴.۴: خلاصه‌ای از آماره‌های پسینی با استفاده از مدل‌های چندجمله‌ای STAR و چندجمله‌ای کسری برای داده‌های تصادف مازندران.

	مدل چندجمله‌ای STAR						مدل چندجمله‌ای کسری		
	رده گوشه			مجموع مساوی با صفر					
	میانگین	انحراف استاندارد	CI ۹۵%	میانگین	انحراف استاندارد	CI ۹۵%	میانگین	انحراف استاندارد	CI ۹۵%
عرض از مبدأ									
بدون آسیب	۰/۲۱	۰/۰۵	(۰/۱۴, ۰/۲۸)	۰/۱۹	۰/۰۳	(۰/۱۱, ۰/۲۶)	۰/۱۶	۰/۰۲	(۰/۱۰, ۰/۲۱)
آسیب کم	−۰/۰۴	۰/۰۹	(−۰/۰۹, ۰/۰۱)	−۰/۰۵	۰/۰۲	(−۰/۱۰, ۰/۰۰۱)	−۰/۰۵	۰/۰۶	(−۰/۰۹, ۰/۰۱)
آسیب جدی	−۰/۱۷	۰/۰۸	(−۰/۲۳, ۰/۱۶)	−۰/۱۴	۰/۰۴	(−۰/۲۱, ۰/۱۲)	−۰/۱۶	۰/۰۶	(−۰/۲۳, ۰/۱۴)
سرعت									
بدون آسیب	−۰/۸۸	۰/۰۵	(−۱/۱۱, ۰/۰۱)	−۱/۰۵	۰/۰۲	(−۱/۳۳, ۰/۱۳)	−۰/۹۳	۰/۰۶	(−۱/۱۱, ۰/۰۵)
آسیب کم	۰/۰۳	۰/۰۶	(۰/۰۰۸, ۰/۰۷)	۰/۰۲	۰/۰۵	(۰/۰۰۶, ۰/۰۴)	۰/۰۳	۰/۰۲۷	(۰/۰۰۱, ۰/۰۵)
آسیب جدی	۰/۹۴	۰/۰۶	(۰/۵۶, ۱/۳۰)	۰/۹۱	۰/۰۱	(۰/۲۴, ۱/۵۱)	۱/۰۱	۰/۰۶	(۰/۴۶, ۱/۶۰)
τ_u	۰/۶۳	۰/۱۱	(۰/۲۲, ۱/۰۳)	۱/۰۱	۰/۳۲	(۰/۸۴, ۲/۷۳)	۰/۹۳	۰/۱۴	(۰/۷۸, ۱/۳۵)
log.score		۲۲۸/۳۲			۲۲۱/۱۰			۳۰۴/۰۲	
DIC		۲۱۳/۱۰			۱۹۰/۷۸			۲۴۰/۲۳	

محاسبه می‌شود. مقادیر کوچکتر این معیار کیفیت پیشگویی بهتری از مدل را نشان می‌دهند. امتیاز لگاریتمی مدل‌های برازش‌شده بر داده‌های تصادف‌های مازندران در جدول ۴.۴ گزارش شده‌اند. همان‌طور که انتظار داشتیم، مدل STAR با قید مجموع مساوی صفر از بقیه مدل‌ها برتر است و توانایی پیشگویی بالاتری دارد. همچنین در مقام مقایسه با مدل‌های دیگر، مقدار بزرگتر امتیاز لگاریتمی برای مدل چندجمله‌ای کسری، طبیعت ضعیف این مدل را در پیشگویی توصیف می‌کند.

۲.۳.۴ انتخاب مدل

برخی از اوقات ممکن است مدل‌های متعددی برای برازش و تحلیل داده‌ها در دسترس باشند به‌طوری که تشخیص واقعی مدل صحیح و مفید را با کمی پیچیدگی و چالش روبرو کند. در این‌گونه موارد به معیاری نیازمندیم که ما را در تعیین مدلی که مناسب داده‌های ما باشد، یاری رساند. به عبارتی، معیاری نیاز است که میزان سازگاری داده‌ها با مدل پیشنهادی را تعیین کند.

در مدل‌های بیزی ، معمولا افراد از معیار اطلاع انحراف[۱۸] برای این کار استفاده می‌کنند. این معیار در ابتدا به صورت

$$\mathrm{D}(\boldsymbol{\theta}) = -۲\log(\mathrm{p}(y|\boldsymbol{\theta})),$$

تعریف می‌شود. اما در مدل‌های بیزی، معیار تعریف‌شده یک متغیر تصادفی محسوب می‌شود و از امید ریاضی آن تحت توزیع پسین به‌عنوان معیار برازش استفاده می‌کنند. با کمک ایده شمارش تعداد پارامترهای تأثیرگذار، می‌توان معیار

$$\mathrm{p_D} = \mathrm{E}(\mathrm{D}(\boldsymbol{\theta}|y)) - \mathrm{D}(\mathrm{E}(\boldsymbol{\theta}|y)) = \bar{\mathrm{D}} - \mathrm{D}(\bar{\boldsymbol{\theta}}),$$

را تعریف کرد. بنابراین

$$\mathrm{DIC} = \bar{\mathrm{D}} - \mathrm{p_D}.$$

مقادیر کوچکتر DIC نشان‌دهنده مدل‌های بهتر هستند. برای داده‌های تصادف، در مقایسه با مدل‌های دیگر، مقدار کوچکتر DIC برای مدل با قید مجموع مساوی با صفر شاهد مدعای ما در برازش بهتر این مدل با این قید است.

همان‌گونه که از نتایج نشان داده‌شده در جدول ۴.۴ پیداست، متغیر تبیینی سرعت برای رده‌های آسیب کم و آسیب جدی، معنی‌داری است. بنابراین، سرعت زیاد احتمال تصادف با آسیب‌دیدگی را افزایش می‌دهد. این نتیجه با زنگ و همکاران (۲۰۱۶ و ۲۰۱۹) هماهنگ است. همچنین برآورد میانگین پسینی اثر ناپارامتری (ناخطی) سن با کران اعتبار بیزی ۹۵٪ در شکل ۸.۴، برای هر سه مدل و هر سه رده آسیب، نمایش داده شده است. نمودارها نشان از تغییر معنی‌دار اثر سن در رده‌های مختلف مربوط به شدت جراحات دارد. نتایج حاصل از

[۱۸] Deviance information criterion (DIC)

هر سه مدل به‌طور تقریبی شبیه به هم هستند. شکل‌های مربوط به برآورد اثر سن، مناسب بودن مدل ناخطی ناپارامتری را برای برازش این اثر نشان می‌دهد.

به نظر می‌رسد افراد جوان‌تر باعث افزایش احتمال تصادف می‌شوند. اما افراد با سنین بالاتر تصادفات با خسارت کمتری را رقم می‌زنند که می‌توان علت را تجربه کافی درکنترل ماشین در شرایط سخت و ناگهانی به‌وجود آمده دانست. آن‌ها همچنین با جاده‌ها آشنایی بیشتری دارند. این نتایج با نتایج به‌دست آمده از محامد (۲۰۱۹) منطبق است. بنا بر پورتر (۲۰۱۱) در هر مایل مسافت پیموده‌شده، راننده‌های جوان با نرخ بالاتری تصادفات با مرگ و میر را موجب می‌شوند. بعد از دامنه سن جوانی، شیب مثبتی را برای بدون آسیب در سن‌های ۴۵ و ۶۰ سالگی می‌بینیم. به نظر می‌رسد این افراد تجربه لازم را برای کنترل ماشین در شرایط سخت دارند. آن‌ها، همچنین، قوانین مربوط به حداکثر سرعت در جاده‌ها را رعایت می‌کنند. بنابراین هنگامی که تصادف برای رانندگان بین ۴۵ تا ۶۰ اتفاق می‌افتد، احتمال این‌که خسارت ناشی از آن در رده بدون آسیب باشد، بیشتر است. از سویی دیگر، برای رانندگان با سن بیشتر از این، کاهش توانایی نگران‌کننده است زیرا ممکن است آن‌ها را به موقعیت‌های دشواری سوق دهد. اگرچه در داده‌های مربوط به استان مازندران رانندگان مسن‌تر از ۶۰ سالگی نداشتیم.

نتایج استنباط پسینی برای پارامتر دقت مدل ICAR، حضور اثر فضایی در مدل را تاکید می‌کند. عدم در نظر گرفتن آن موجب اریبی در برآوردها خواهد شد. میانگین و انحراف استاندارد مربوط به اثر فضایی به ترتیب در شکل‌های ۹.۴ و ۱۰.۴ نمایش داده شده‌اند. برآورد اثر فضایی برای هر سه مدل تقریباً مشابه است. همچنین، نتایج نرخ تصادف الگوی فضایی متفاوتی را در سرتاسر استان مازندران به رخ می‌کشد. بنابراین، یک نتیجه اساسی در این رابطه در نظر گرفتن اثر فضایی در تحلیل داده‌های تصادف مربوط به شدت جراحات است.

نقشه‌های مربوط به احتمال‌های برآوردشده در هر رده برای هر سه مدل، در شکل ۱۱.۴ نمایش داده شده است. نقشه‌ها نشان می‌دهند که احتمال تصادف در شهرهای توسعه‌یافته کم است. احتمال تصادف در شهرهایی که در مسیر اصلی ترددهای بین استانی هستند، بیشتر است. بعضی از شهرها الگوهای مشابهی را از خود به نمایش گذاشته‌اند، مانند بابل، قائم‌شهر و ساری، که در آن‌ها احتمال داشتن تصادف با آسیب کم رو به افزایش است. تفاوت در کاربری زمین‌ها اثرات متفاوتی در میزان شدت تصادف دارند. در بعضی از شهرها افزایش فاصله از دریا، نزدیکترین جنگل و کوه احتمال تصادف را کاهش می‌دهد، به‌طوری که شهرهای با فاصله کمتر از دریا تصادفات بیشتری را تجربه می‌کنند. همچنین حضور گردشگران ناشی و نابلد در جاده‌های روستایی، احتمال تصادف با ماشین‌آلات کشاورزی را افزایش می‌دهد.

بنابراین در شهرهای با جاده روستایی بیشتر، احتمال تصادف نیز بیشتر است. عواملی مانند شرایط آب و هوایی ممکن است با قرار گرفتن در مدل، اثر معنی‌داری داشته باشند اما متاسفانه به دلیل در دسترس نبودن اطلاعات مربوط به آن‌ها در مدل مورد استفاده قرار نگرفتند. یافته‌ها ناهمگنی قابل توجهی را در داده‌های ما نشان می‌دهد و به نظر می‌رسد که مدل پیشنهادی آن را به خوبی بر عهده گرفته است. پس لازم است تا داده‌های مربوط به

تصادف برای تحلیل شدت جراحات را با مدل چندجمله ای برازش داده و اثر فضایی را در آن در نظر بگیریم. حضور توریست یا همان صنعت گردشگری در استان مازندران یکی از اساسی‌ترین صنایع در این استان محسوب می‌شود. با در نظر گرفتن اثر فضایی و توزیع احتمال در هر رده دریافتیم نواحی با احتمال حضور توریست بیشتر، تعداد تصادف بیشتری دارد. بنابراین در نظر گرفتن حضور توریست در تصمیمات مربوط به سیاست‌های پیشگیرانه تصادف امری ضروری است. تعطیلات تابستان احتمال حضور توریست را افزایش می‌دهد. از طرفی دیگر، عدم وجود آگاهی‌های لازم و نداشتن آشنایی کافی با جاده‌ها، به تصادفات با خسارت‌های مالی زیاد یا حتی مرگ منجر می‌شوند.

۴.۴ خلاصه فصل و نتیجه‌گیری

امروزه یادگیری ماشین[۱۹]، هوش مصنوعی[۲۰] و برنامه‌ریزی‌های صنعتی در علومی همچون اقتصاد، علوم سیاسی و ژئوفیزیک نیاز به پیش‌بینی رده‌ها دارد که مدل لجیت چندجمله‌ای با اثر فضایی نقش مهمی را در این میان می‌تواند ایفا می‌کند. همچنین داده‌های رده‌ای فضایی مشبکه‌ای به علت پیشرفت در صنعت ثبت داده‌ها به‌طور گسترده‌ای در دسترس هستند که داده‌های رده‌ای تصادف که روی شدت جراحات رده‌بندی شده باشد، یکی از آن‌ها محسوب می‌شود. به همین دلیل، ما روش تحلیل داده‌های چندجمله‌ای فضایی با ساختار جمعی ـ رگرسیونی را انتخاب کردیم و از روش INLA به جای روش متداول نمونه‌گیری MCMC برای برازش مدل‌ها استفاده کردیم. برای آن‌که بتوان مدل را با INLA برازش داد، از تبدیل چندجمله‌ای ـ پواسن بهره‌مند شدیم. با استفاده از این تبدیل می‌توان به راحتی متغیرهایی را که به صورت یک تابع در مدل ظاهر می‌شوند، مانند اثرات ناخطی، اثر فضایی و اثر فضایی ـ زمانی افزون بر اثرات ثابت در پیشگوی جمعی ـ ساختاری مربوط به مدل چندجمله‌ای به‌کار گرفت.

انجام استنباط بیزی در مدل چندجمله‌ای به دلیل وجود مشکل شناسایی‌پذیری، چالشی است. برای رفع این مشکل، قید مجموع مساوی با صفر را معرفی کرده و کارایی آن را در برآورد با قید متداول رده گوشه مقایسه کردیم. در همین حال مدل چندجمله‌ای کسری را نیز به علت مشابهت با مدل معرفی‌شده، به داده‌های واقعی برازش داده و نتایج حاصل از آن را مقایسه کردیم. خلاصه نتایج به شرح زیر است:

۱. با در نظر گرفتن دو معیار کارایی محاسباتی و توانایی پیشگویی، مدل چندجمله‌ای فضایی با ساختار جمعی ـ رگرسیونی و قید مجموع مساوی با صفر بر مدل چندجمله‌ای کسری برتری دارد.

---

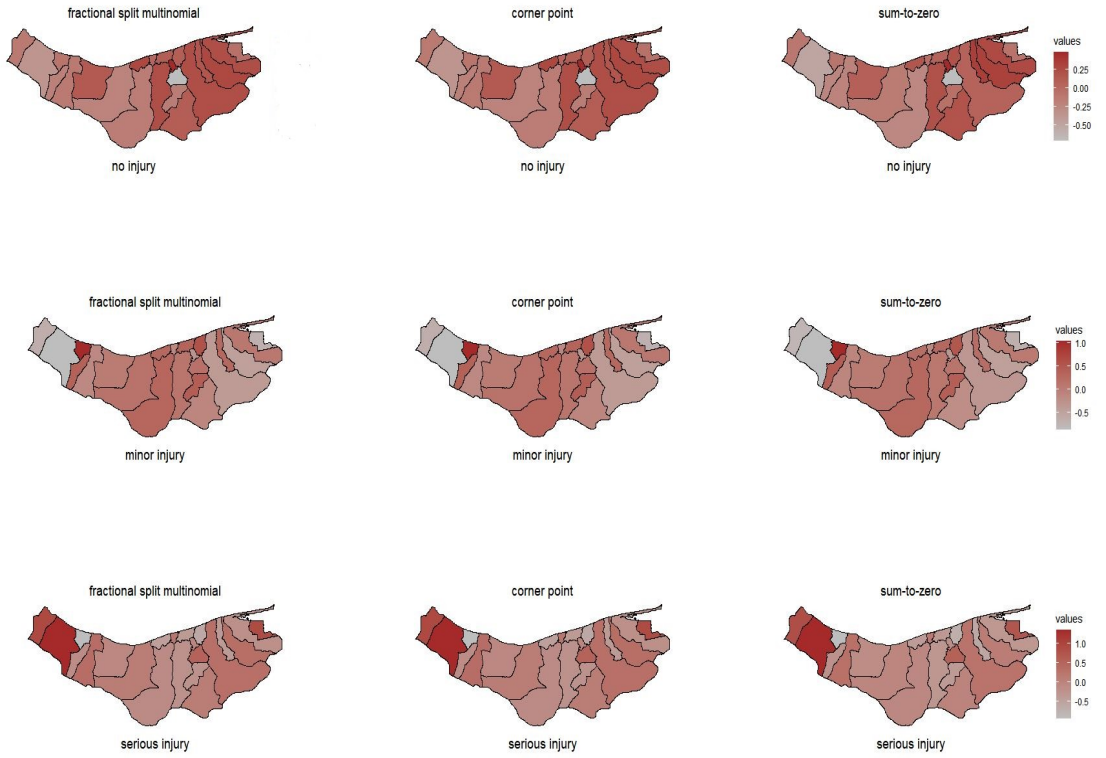
[۱۹] machine learning
[۲۰] artificial intelligence

۲. مدل با قید مجموع مساوی با صفر برآوردهای کاراتری برای پارامترها به‌ویژه پارامتر دقت مدل بی‌سیگ فراهم آورده است.

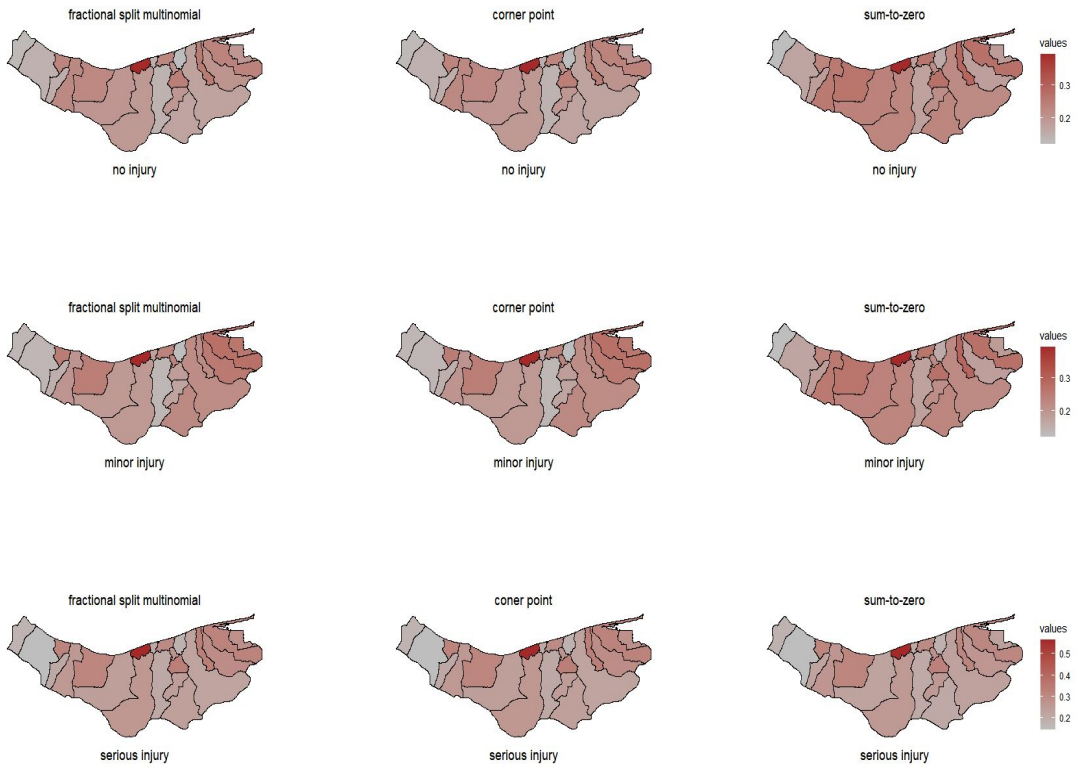
۳. در حالت کلی تفاوت چشمگیری بین برآورد احتمال‌های رده‌ها دیده نمی‌شود و برآوردها تفاوت کمی با مقادیر واقعی دارند.

۴. اگر به آسانی تفسیر یک مدل باور داشته باشیم، باید مدل با قید مجموع مساوی با صفر را برگزینیم. زیرا این قید به بعضی از پارامترهای مدل نگاهی خاص ندارد؛ در مقابل، همه بر حسب تفاوت بین یک پارامتر و میانگین همان پارامتر در رده‌ها تفسیر می‌شود. بنابراین وقتی یک توزیع پیشین انتخاب می‌شود، می‌توان تفاوت بین پارامترها را حدس زد. در مقابل برای قید رده گوشه محقق باید برآورد را نسبت به رده انتخاب‌شده به‌عنوان رده گوشه تفسیر کند. از طرف دیگر، در قید مجموع مساوی با صفر توزیع پیشین را به همه عناصر موجود در رده‌ها نسبت می‌دهیم که این کار خود از مشکل تغییرپذیری توزیع پیشین جلوگیری می‌کند.

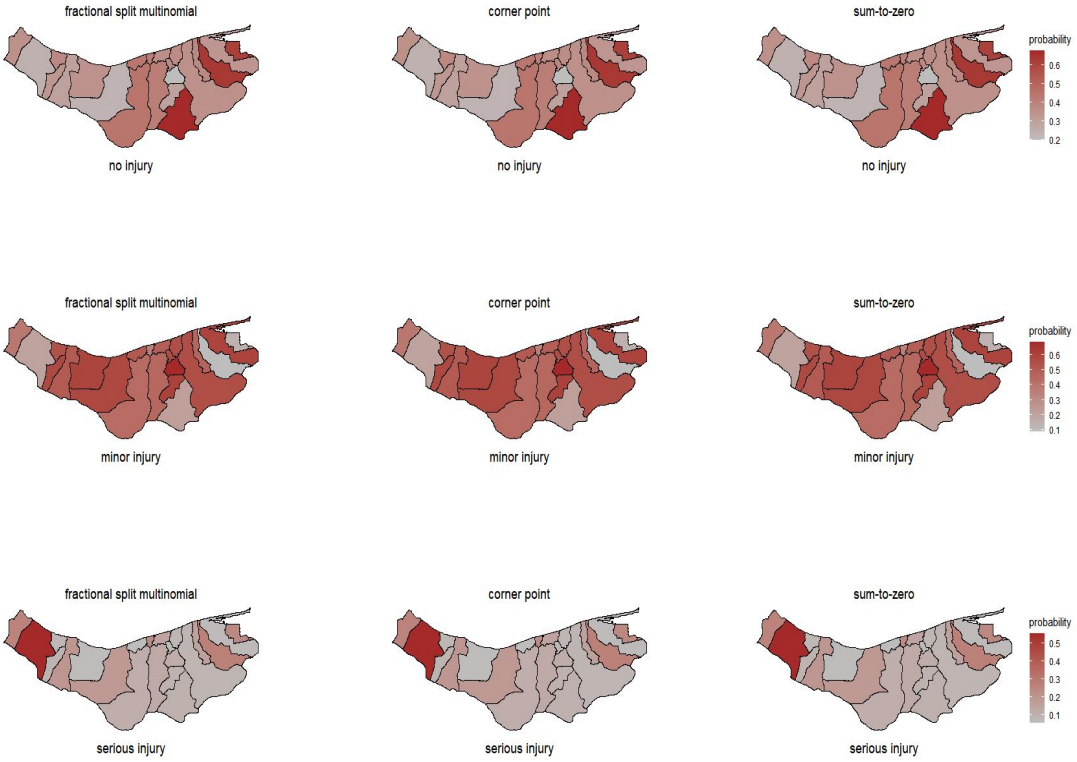
شکل ۸.۴: برآوردهای میانگین پسینی اثر سن با کران اعتبار بیزی ۹۵٪ در داده‌های تصادفات مازندران



شکل ۹.۴: برآوردهای میانگین پسینی اثر فضایی در هر رده و برای هر مدل: از بالا به پایین مربوط به رده‌های با شدت جراحت کم تا بالا.



شکل ۱۰.۴: برآورد انحراف استاندارد پسینی اثر فضایی در هر رده برای هر مدل: از بالا به پایین مربوط به رده‌های با شدت جراحت کم تا بالا.



شکل ۱۱.۴: توزیع احتمال برآوردشده در هر رده برای هر مدل: از بالا به پایین مربوط به رده‌های با شدت جراحت کم تا بالا.

فصل ۵

مدل لوجیت تکی‌شده برای داده‌های چندجمله‌ای فضایی زمین‌آماری

در فصل قبل حالتی را بررسی کردیم که در آن متغیر پاسخ در هر رده دارای فراوانی بیشتر از یک است که برای داده‌های مشبکه‌ای یک پذیره قابل قبول و استاندارد است. در این فصل داده‌های چندجمله‌ای را در نظر می‌گیریم که متغیر پاسخ فراوانی ندارد. در واقع، اثر فضایی را از نوع زمین‌آماری در نظر می‌گیریم، زیرا نبود فراوانی بیشتر از یک در این حالت، در اکثر کاربردها، معمول و پذیرفتنی است. برای رهایی از مشکل شناسایی‌پذیری و داشتن یک مدل با برازش آسان و در عین حال تا حد قابل قبولی کارا، مدل لوجیت تکی‌شده[۱] (بگ و گری، ۱۹۸۴) را پیشنهاد می‌کنیم. به کمک شبیه‌سازی و یک مثال واقعی، کارایی برآوردهای حاصل از مدل پیشنهادی را با مدل چندجمله‌ای مقایسه می‌کنیم.

۱.۵ مدل لوجیت تکی‌شده

مدل (۱.۴) از فصل چهارم را در نظر بگیرید. فرض کنید متغیر Y_{ij} مقدار یک را بگیرد اگر پاسخ متعلق به رده j است و در غیر این صورت صفر. بر اساس مدل (۲.۴)، با این فرض که

[۱] Individualized logit

متغیرهای تبیینی برای همه رده‌ها یکسان هستند، می‌توان رابطه

$$\log(\frac{\pi_{ij}}{\pi_{i۱}}) = \boldsymbol{x}_i'\boldsymbol{\beta}_j, \quad i = ۱,\ldots,n, \quad j = ۲,\ldots,J \tag{۱.۵}$$

را نوشت، که در آن $\boldsymbol{x}_i = (x_{۱i},\ldots,x_{li})'$. برای وارد کردن عرض از مبدا، می‌توان قرار داد $x_۱ = ۱$. با پیروی از بگ و گری (۱۹۸۴) و اگرستی (۲۰۰۲)، اگر از یک مدل لوجیت چندجمله‌ای برای مقایسه رده j با رده اول استفاده کنیم، آن‌گاه رابطه

$$\log(\frac{\phi_{ij}}{\phi_{i۱}}) = \boldsymbol{x}_i'\boldsymbol{\gamma_j}, \quad j = ۲,\ldots,J \tag{۲.۵}$$

برقرار است، که در آن

$$\phi_{ij} = \mathrm{P}(Y_{ij} = ۱|\boldsymbol{x}_i, y_{i۱} + y_{ij} = ۱)$$

و

$$\phi_{i۱} = \mathrm{P}(Y_{ij} = ۰|\boldsymbol{x}_i, y_{i۱} + y_{ij} = ۱).$$

در این حالت تعداد $J-۱$ مدل تکی‌شده تعریف می‌شود. در حقیقت در مدل لوژستیک تکی‌شده، مشکل تشخیص و شناسایی پارامترها به کمک یک مجموعه از مدل‌های لوجیت ساده به جای یک مدل چندجمله‌ای برطرف شده است. بنا بر بگ و گری (۱۹۸۴) می‌توان نشان داد دو مدل (۱.۵) و (۲.۵) از نظر پارامتربندی معادل هستند و برای $j = ۲,\ldots,J$، $\gamma_j = \beta_j$. به‌طور دقیق‌تر، با استفاده از قضیه بیز، می‌توان نوشت $\phi_{ij} = \frac{\pi_{ij}}{\pi_{i۱}+\pi_{ij}}$. بنابراین

$$\frac{\pi_{ij}}{\pi_{i۱}} = \frac{\phi_{ij}}{(۱-\phi_{ij})}.$$

در نتیجه، اگر با روش تکی‌شده $J-۱$ رگرسیون لوژستیک برازش و پارامترها با استفاده از مدل (۲.۵) برآورد شوند، می‌توان از پارامترهای برآوردشده استفاده و با تعبیه آن‌ها در مدل (۱.۵) احتمال‌های رده‌ها را برآورد کرد.

به روشنی می‌توان درک کرد که روش تکی‌شده یا همان لوجیت تکی‌شده به صورت تحلیلی انعطاف‌پذیر است (رم و همکاران، ۱۹۹۵). در حالت کلی، روش تکی‌شده از کارایی بالایی برخوردار است (بگ و گری، ۱۹۸۴) و به‌ویژه برای مجموعه داده‌های بزرگ و تنک مفید است (رم و همکاران، ۱۹۹۵). این روش، همچنین برای توسعه یک چارچوب استنباطی بیزی با روش INLA نیز مناسب است.

۱.۱.۵ توزیع مجانبی برآورد درستنمایی ماکسیمم

تابع درستنمایی

$$\mathrm{L} = \prod_{i=۱}^{n}\prod_{j=۱}^{J}\pi_{ij}^{y_{ij}},$$

را در نظر بگیرید. برآوردهای ماکسیمم درستنمایی حاصل از مدل چندجمله‌ای از $l(J-۱)$ معادله

$$\sum_{i=۱}^{n} x_{ki}(y_{ij}-\pi_{ij})=۰,\quad k=۱,\ldots,l,\ j=۲,\ldots,J$$

به‌دست می‌آیند، که در آن

$$\pi_{ij}=\frac{\exp(\eta_{ij})}{\sum_{s=۱}^{J}\exp(\eta_{is})},$$

برآوردگرهای $\hat{\boldsymbol{\beta}}_۱,\ldots,\hat{\boldsymbol{\beta}}_J$ دارای توزیع نرمال مجانبی با میانگین $\boldsymbol{\beta}_۱,\ldots,\boldsymbol{\beta}_J$ و ماتریس کوواریانس $n^{-۱}\boldsymbol{V}^{-۱}$ است، به‌گونه‌ای که $\boldsymbol{V}=((\boldsymbol{V}_{jj'}))$ یک ماتریس $l(J-۱)\times l(J-۱)$ است که متشکل از زیرماتریس‌های $\boldsymbol{V}_{jj'}$ با بعد $l\times l$ است و مولفه (k,t)ام آن‌ها عبارتند از

$$\lim_{n\to\infty} n^{-۱}\sum_{i=۱}^{n} x_{ki}x_{ti}\pi_{ij}(\delta(j,j')-\pi_{ij'})$$

که در آن $\delta(j,j')$ دلتای کرونه‌کر است. در استفاده از مدل تکی‌شده، $J-۱$ تابع درستنمایی شرطی ماکسیمم می‌شوند که هر درستنمایی به صورت

$$\mathrm{L}_j=\prod_{C_j}\phi_{ij}^{y_{ij}}(۱-\phi_{ij})^{y_{i۱}},\ j=۲,\ldots,J,$$

است، که در آن $C_j=\{i: y_{i۱}+y_{ij}=۱\}$. با قرار دادن $\ell_j=\log\mathrm{L}_j$، می‌توان معادلات امتیاز متناظر را به صورت

$$\frac{\partial \ell_j}{\partial \beta_{kj}}=\sum_{i=۱}^{n} x_{ik}(y_{ij}-(y_{ij}+y_{i۱})\phi_{ij})=۰,\quad k=۱,\ldots,l,\quad j=۲,\ldots,J,$$

نوشت. برآوردهای $\check{\boldsymbol{\beta}}_j$ از حل معادلات امتیاز به‌دست می‌آیند. چون $\check{\boldsymbol{\beta}}_j$ برآورد ماکسیمم درستنمایی است، بنا بر تعریف همگرایی در احتمال، می‌توان رابطه

$$\{\sqrt{m_j}(\check{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)-m_j^{-۱/۲}\boldsymbol{H}_j^{*-۱}\frac{\partial \ell_j}{\partial \boldsymbol{\beta}_j}\}\to ۰$$

را نوشت، که در آن m_j اندازه (تعداد اعضای) مجموعه C_j را نشان می‌دهد و مولفه (k,t)ام ماتریس $\boldsymbol{H}_j^*$ عبارتست از

$$\lim_{m_j\to\infty} -m_j^{-۱}\frac{\partial^۲ \ell_j}{\partial\beta_{kj}\partial\beta_{tj}}.$$

در این جا می‌پذیریم که، نسبت برآمدها در هر رده به‌طور مجانبی، غیر قابل چشم‌پوشی است. چنین پذیره‌ای نامعمول نیست و به‌عنوان نمونه وقتی که متغیرهای تبیینی به‌طور یکنواخت کراندار باشند، برقرار است. با این پذیره، چون $m_j/n\to K_j>۰$، نتیجه

$$\{\sqrt{n}(\check{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)-n^{-۱/۲}\boldsymbol{H}_j^{-۱}\frac{\partial \ell_j}{\partial \boldsymbol{\beta}_j}\}\to ۰,$$

به‌دست می‌آید، که در آن $\boldsymbol{H}_j = K_j \boldsymbol{H}_j^*$. چون این نتیجه برای همه jها برقرار است، پس می‌توان نوشت

$$\{\sqrt{n}(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{H}^{-1}\boldsymbol{S}_n\} \to 0 \qquad (۳.۵)$$

که در آن

$$(\check{\boldsymbol{\beta}} - \boldsymbol{\beta}) = ((\check{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)', \ldots, (\check{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J)'), \quad \boldsymbol{S}_n = n^{-1/2}((\partial \ell_2/\partial \boldsymbol{\beta}_2)', \ldots, (\partial \ell_J/\partial \boldsymbol{\beta}_J)')'$$

و $\boldsymbol{H}$ یک ماتریس قطری بلوکی با بلوک‌های $\boldsymbol{H}_2, \ldots, \boldsymbol{H}_J$ است. با به‌کار گرفتن قضیه حد مرکزی روی $\boldsymbol{S}_n$، می‌توان همگرایی در توزیع زیر را نتیجه گرفت:

$$\boldsymbol{S}_n \xrightarrow{\mathrm{D}} \mathrm{N}(0, \boldsymbol{W}) \qquad (۴.۵)$$

که در آن $\boldsymbol{W}$ یک ماتریس متشکل از زیرماتریس‌های $\boldsymbol{W}_{jj'}$ با بعد $l \times l$ است که مولفه (k,t)ام آن‌ها عبارتست از

$$\lim_{n\to\infty} n^{-1}\mathrm{cov}(\frac{\partial \ell_j}{\partial \beta_{kj}}, \frac{\partial \ell_{j'}}{\partial \beta_{tj'}}),$$

عناصر $\boldsymbol{W}$ را می‌توان با کمک رابطه

$$\mathrm{cov}(\frac{\partial \ell_j}{\partial \beta_{kj}}, \frac{\partial \ell_{j'}}{\partial \beta_{tj'}}) = \sum_{i=1}^{n} x_{ki}x_{ti}\pi_{i1}\phi_{ij}\phi_{ij'}^{1-\delta(j,j')},$$

محاسبه کرد (بگ و گری، ۱۹۸۴). با استفاده از (۳.۵) و (۴.۵) می‌توان نتیجه گرفت که $\sqrt{n}(\check{\boldsymbol{\beta}}-\boldsymbol{\beta})$ در توزیع به $\mathrm{N}(0, \boldsymbol{H}^{-1}\boldsymbol{W}\boldsymbol{H}^{-1})$ همگرا است. بگ و گری تاکید کردند که $\boldsymbol{H}_j = \boldsymbol{W}_{jj}$ و بنابراین بلوک‌های قطری $\boldsymbol{H}^{-1}\boldsymbol{W}\boldsymbol{H}^{-1}$ برابر هستند با $\boldsymbol{H}_j^{-1} = \boldsymbol{W}_{jj}^{-1}$.

به‌طور خلاصه برآوردهای ماکسیمم درستنمایی حاصل از مدل چندجمله‌ای و مدل‌های لوجیت تکی‌شده، دارای توزیع مجانبی نرمال با بردار میانگین صفر و ماتریس‌های کوواریانس به ترتیب $n^{-1}\boldsymbol{V}^{-1}$ و $\boldsymbol{H}^{-1}\boldsymbol{W}\boldsymbol{H}^{-1}$ هستند. در حالت کلی مقایسه دو ماتریس کوواریانس راحت نیست، اما روشن است که عناصر خارج از قطر اصلی ماتریس $\boldsymbol{H}^{-1}\boldsymbol{W}\boldsymbol{H}^{-1}$ که متناظر با کواریانس‌های بین پارامترهای مدل‌های لوجیت جدا هستند، در حالت کلی ناصفر هستند.

۲.۱.۵ همگرایی برآوردگر بیزی مدل تکی‌شده

در این بخش، نشان می‌دهیم که برآوردگر بیزی پارامترهای رگرسیونی در مدل لوجیت‌های تکی‌شده به مقدار واقعی پارامترها همگرا است. برای این کار همگرایی در احتمال را با $\xrightarrow{\mathrm{P}}$ نشان می‌دهیم.

قضیه ۱.۱.۵. در مدل لوجیت‌های تکی‌شده، چنان‌چه توزیع پیشین پارامترهای $\boldsymbol{\beta}$ روی فضای پارامتر Θ دارای چگالی مثبت و پیوسته باشد، آن‌گاه برآوردگر بیزی حاصل به مقدار واقعی پارامترها همگرا است.

برای اثبات این نتیجه به قضیه برنشتاین–وون‌میسس[۲] نیاز داریم که در زیر بیان شده است.

قضیه ۲.۱.۵. (باتاچاریا و همکاران، ۲۰۱۶). اگر توزیع پیشین π برای $\boldsymbol{\beta}$ دارای چگالی مثبت و پیوسته روی فضای پارامتر Θ باشد، آن‌گاه تحت π، فاصله تغییرات کل[۳] بین توزیع پسین $\boldsymbol{\beta}$ و توزیع نرمال $\mathrm{N}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}, -\boldsymbol{H}^{-1})$، چنان‌چه $n \longrightarrow \infty$، به صفر همگرا است، که در آن $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ برآوردگر درستنمایی ماکسیمم پارامترها است.

برهان. با استفاده از نامساوی مثلثی و قضیه برنشتاین–وون‌میسس، برهان روشن است. فرض کنید $\hat{\boldsymbol{\beta}}_{\mathbf{ML}}$ و $\hat{\boldsymbol{\beta}}_{\mathbf{B}}$ به ترتیب برآوردگرهای ماکسیمم درستنمایی و بیزی پارامتر $\boldsymbol{\beta}$ باشند. بنابر نابرابری مثلثی می‌توان نوشت

$$|\hat{\boldsymbol{\beta}}_{\mathbf{B}} - \boldsymbol{\beta}| \leqslant |\hat{\boldsymbol{\beta}}_{\mathbf{ML}} - \boldsymbol{\beta}| + |\hat{\boldsymbol{\beta}}_{\mathbf{B}} - \hat{\boldsymbol{\beta}}_{\mathbf{ML}}|.$$

برای نشان دادن همگرایی مد نظر، باید نشان دهیم که سمت راست نابرابری در احتمال به صفر میل می‌کند. یکی از نتایج قضیه برنشتاین–وون‌میسس، همگرایی میانگین توزیع پسین در احتمال به برآوردگر ماکسیمم درستنمایی است. از طرفی دیگر، در بخش قبل دیدیم که برآوردگر ماکسیمم درستنمایی پارامترهای حاصل از مدل لوجیت‌های تکی‌شده برای $\boldsymbol{\beta}$ سازگار است. در نتیجه به سادگی نتیجه قضیه برهان می‌شود و می‌توان گفت $\hat{\boldsymbol{\beta}}_{\mathbf{B}} \xrightarrow{\mathrm{P}} \boldsymbol{\beta}$. □

۲.۵ مدل‌بندی وابستگی فضایی زمین‌آماری

تعریف داده‌های زمین‌آماری و میدان تصادفی گوسی را به خاطر بیاورید. علی‌رغم اهمیت و کاربردهای متنوع مدل‌بندی فضایی داده‌های زمین‌آماری با میدان‌های تصادفی گوسی، استفاده از آن‌ها برای موقعیت‌های کاربردی با حجم بالایی از داده‌ها، همیشه از نظر محاسباتی چالش‌برانگیز بوده است. چالش اصلی نیز مربوط به تجزیه ماتریس کوواریانس چگال میدان تصادفی است که هزینه محاسباتی آن از مرتبه $\mathrm{O}(n^3)$ است. راهکارهای مختلفی برای حل این چالش محاسباتی پیشنهاد شده‌اند. برای دسترسی به یک مقاله مروری جدید در این حوزه و آشنایی با این راهکارها می‌توانید به هیتان و همکاران (۲۰۱۸) مراجعه کنید.

برای رهایی از این مشکل، لیندگرن و همکاران (۲۰۱۱) استفاده از روش معادلات دیفرانسیل جزئی تصادفی[۴] را پیشنهاد کردند. این روش بر اساس ابزارهای پیشرفته فرآیندهای تصادفی پایه‌ریزی شده است که از نظر محاسباتی کاراست و تاکنون استفاده‌های فراوانی از آن شده است.

---

[۲] Bernstein–von Mises theorem
[۳] Total variation distance
[۴] Stochastic partial differential equation (SPDE)

میدان تصادفی گوسی کاملاً به وسیله میانگین و کوواریانس (ماتریس یا تابع) میدان شناسایی می‌شود. از نظر محاسباتی، کار کردن با یک میدان تصادفی گوسی، با بردارها و ماتریس‌های جبری روبرو است که برای آن می‌توان از بسته‌های استاندارد رایانه‌ای بهره گرفت. در مدل‌بندی آماری، افراد معمولا از میدان تصادفی گوسی با ساختار کوواریانس پارامتربندی‌شده استفاده می‌کنند که اغلب زیرمجموعه‌ای از خانواده ساختار کوواریانس مترن است. روش SPDE از نمایش مارکوفی گوسی برای میدان تصادفی گوسی استفاده می‌کند که در آن بهره گرفتن از ماتریس دقت تنک، کارایی محاسباتی در روش‌های عددی را به‌ویژه در تجزیه ماتریس‌ها، به طرز قابل ملاحظه‌ای افزایش می‌دهد (رو و هلد، ۲۰۰۵). در این رساله از یک میدان تصادفی گوسی با میانگین صفر و تابع کوواریانس مترن بهره‌مند خواهیم شد. برای یادآوری، ضابطه تابع کوواریانس مترن در زیر نوشته شده است:

$$\mathrm{Cov}(Y(s_i), Y(s_j)) = \frac{\sigma^2}{\Gamma(\lambda)2^{\lambda-1}}(\kappa||s_i - s_j||)^{\lambda} K_{\lambda}(\kappa||s_i - s_j||)$$

که در آن، $||s_i - s_j||$ فاصله اقلیدسی بین موقعیت‌های $s_i,\ s_j \in D$، σ انحراف استاندارد کناری میدان مترن، $K_{\lambda}(\cdot)$ تابع بسل نوع دوم با مرتبه $\lambda > 0$ و $\Gamma(\cdot)$ تابع گاما است. پارامتر λ نیز درجه همواری فرآیند را اندازه‌گیری می‌کند و معمولا ثابت نگه داشته می‌شود. همچنین κ پارامتر مقیاس است. به پارامترهای مدل مترن عنوان ابرپارامترهای مدل می‌دهیم و به میدان تصادفی گوسی پارامتربندی‌شده حاصل از آن، اثر تصادفی فضایی یا عنصر مدل فضایی زمین‌آماری نسبت می‌دهیم. از فواید استفاده از تابع مترن می‌توان به توانایی تفسیر فیزیکی پارامترهای آن اشاره کرد (باکا و همکاران، ۲۰۱۸).

راه‌های مختلفی برای انجام استنباط با ساختار کوواریانس میدان تصادفی گوسی وجود دارند. یک راه سنتی مرسوم ساختن ماتریس کوواریانس بر اساس مکان‌های مشاهدات مستقیماً از تابع کوواریانس و بعد ترکیب آن با ماتریس کوواریانس برای سایر عناصر مدل است تا یک ماتریس پررتبه برای مشاهدات یا مدل پنهان به‌وجود آورند. این روش برای مشاهدات تا اندازه صد موقعیت مکانی خوب کار می‌کند، اما برای داده‌های حجیم به اندازه صد هزار از نظر محاسباتی شدنی نیست، زیرا محاسبات با ماتریس کوواریانس Σ به شدت زمان‌گیر است. از طرف دیگر و فراتر از مسائل محاسباتی، ساختن تابع کوواریانس برای فضاهای هندسی به شکل کره (مثل سطح زمین) برای معرفی نامانایی در تابع کوواریانس و توسعه تابع کوواریانس فضایی در مدل های فضایی‌ـ زمانی جداناپذیر، چالش‌برانگیز است (باکا، ۲۰۱۸).

همان‌گونه که در فصل دوم مشاهده کردید، تمرکز روش INLA به‌جای ماتریس کوواریانس بر روی ماتریس دقت $Q = \Sigma^{-1}$ است، زیرا می‌توان نشان داد که مدل‌های مختلف با ساختار ماتریس کوواریانس پیچیده، ماتریس دقت تنک دارند (رو و هلد، ۲۰۰۵). ماتریس‌های دقت اکثراً صفر دارند و این صفرها در رایانه ذخیره نمی‌شوند. از طرفی در توزیع گوسی چندمتغیره، تنکی ساختار ماتریس دقت به استقلال شرطی بین متغیرهای تصادفی مربوط می‌شود (رو و هلد، ۲۰۰۵). فرض کنید بعد یک ماتریس دقت در یک مدل فضایی با ساختار فضایی تنک،

$n \times n$ باشد. استخراج نمونه، حساب کردن ثابت نرمال‌ساز و انجام محاسبات با ماتریس دقت به محاسباتی از مرتبه $O(n^{۳/۲})$ نیاز دارد و این در حالی است که هزینه انجام محاسباتی مشابه با ماتریس کوواریانس از مرتبه $O(n^۳)$ است. از طرف دیگر، پژوهش‌های کاربردی که در آن‌ها صدها هزار مکان داشته باشیم با ماتریس دقت شدنی هستند و انجام دادن کارهایی پژوهشی فضایی با صدها مکان فقط به تعدادی ثانیه زمان نیاز دارد.

روش SPDE یک نمایش از ماتریس کوواریانس مترن به‌عنوان پاسخی از معادله دیفرانسیل تصادفی

$$(\kappa - \Delta)^{\alpha/۲} \tau u(s) = \Omega(s), \quad s \in D \subset \mathbb{R}^d \tag{۵.۵}$$

است، که در آن Δ عملگر لاپلاسی[۵] را نشان می‌دهد، τ واریانس را کنترل می‌کند، α درجه همواری را کنترل می‌کند، $\kappa > ۰$ پارامتر مقیاس و $\Omega(s)$ یک فرآیند نوفه سفید گوسی فضایی با واریانس واحد است. ارتباط با پارامتر همواری ساختار مترن، یعنی λ، و واریانس $\sigma^۲$ از طریق $\alpha = \lambda + d/۲$ و

$$\sigma^۲ = \frac{\Gamma(\lambda)}{\Gamma(\alpha)(۴\pi)^{d/۲}\kappa^{۲\lambda}\tau^۲}$$

است، که در آن d بعد میدان را نشان می‌دهد. پاسخ معادله را می‌توان با به‌کار گرفتن روش عناصر متناهی[۶] با توابع پایه و با نمایش مثلثی با دامنه D به صورت

$$u(s) = \sum_{g=۱}^{G} \phi_g(s)\tilde{u}_g \tag{۶.۵}$$

تقریب زد، که در آن G تعداد کل رأس‌های مثلث، $\{\phi_g\}$ مجموعه توابع پایه و $\tilde{u}_g$ وزن‌های با توزیع گوسی با میانگین صفر هستند.

بر اساس روش SPDE، می‌توان تقریب پیوسته‌ای از پاسخ معادله (۵.۵) که تقریبا ساختار کوواریانس مترن با دامنه کراندار محدودشده را دارد، ساخت. اگرچه پارامترهای معادله (۵.۵) در حالت استاندارد اما متفاوت از معادله (۶.۵) است، ولی رابطه یک به یک در میان آن‌ها وجود دارد. همچنین می‌توان از رابطه یک به یک بین پارامترهای تابع کوواریانس مترن و پارامترهای SPDE استفاده کرد تا با یک تقریب از نظر محاسباتی کارا به برآورد مدل پرداخت و در نهایت با پارامترهای معلوم تابع کوواریانس مترن نتایج حاصل را تفسیر کرد. برای مقادیر صحیح α، پاسخ معادله SPDE مارکوفی است که در نتیجه ماتریس دقت تنک را برای تقریب‌های گسسته منعکس می‌کند. با در نظر گرفتن $\lambda = ۱$، برای $d = ۲$، خواهیم داشت $\alpha = ۲$ و $\sigma^۲ = \frac{۱}{۴\pi\kappa^۲\tau^۲}$. روشن است که دو پارامتر κ و τ اثری توام بر واریانس کناری میدان تصادفی دارند.

سرعت رو به بالا اجرای مدل‌های فضایی در R-INLA، راحتی استفاده از مدل فضایی (به‌طور خاص داده‌های زمین‌آماری) همزمان با دیگر عناصر مدل و توانایی به‌کار گرفتن بخش وسیعی

[۵] Laplacian operator
[۶] Finite elements

از درستنمایی‌های مشاهدات برای فرآیندهای پنهان، موجب شده است تا INLA به ابزار خیلی مفیدی در استفاده از مدل‌های آماری کاربردی تبدیل شود. برازش مدل‌های خیلی پیچیده با روش INLA با رهیافت بیزی تنها ممکن است یک روز به طول انجامد. این امر موجب شده است تا پژوهشگران قادر باشند تا تعداد بیشتری از مدل‌ها را برازش دهند، به فهم بیشتری از داده‌ها برسند، درباره حساسیت پیشین‌ها بحث کنند، درباره حساسیت انتخاب درستنمایی مشاهدات نظر دهند و به مقایسه پیشگویی‌ها بپردازند.

کاربردهای متنوعی از به‌کارگیری مدل‌های فضایی برای داده‌های زمین‌آماری در R-INLA وجود دارند که در زیر تنها به برخی از آن‌ها اشاره می‌کنیم. نور و همکاران (۲۰۱۴) در مجله لنست تحلیل فضایی‌ـزمانی شدت انتقال مالاریا را مورد بررسی قرار دادند. گلدینگ و همکاران (۲۰۱۷) مرگ و میر زیر پنج سال و مرگ و میر نوزادان در چند کشور مختلف با یک مدل فضایی ـ زمانی جداپذیر در گروه‌های سنی مختلف را مدل‌بندی و تحلیل کردند. در مجله Science، جاسینو و همکاران (۲۰۱۷) اثرات تکه تکه شده در بیماری‌های عفونی پویا را مطالعه کردند. بات و همکاران (۲۰۱۵) اثر روش‌های مختلف در کنترل مالاریا را در آفریقا تحلیل کردند. همچنین، ظرفیت‌های INLA در مدل‌بندی فضایی یک ابزار کلیدی بود تا شادیک و همکاران (۲۰۱۸) برآورد سراسری قرار گرفتن در معرض محیط با ذرات بسیار ریز با قطر کمتر از ۲٫۵ میکرومتر را ارائه دهند که به $PM_{۲٫۵}$ معروف است. نتایج مربوط به این پژوهش در سال ۲۰۱۶ در مطالعه جهانی بیماری (گاکیدو و همکاران، ۲۰۱۷) و ارزیابی سازمان بهداشت جهانی از خطر سلامتی ناشی از آلودگی محیط (سازمان جهانی سلامت، ۲۰۱۶) مورد استفاده قرار گرفت.

یک امتیاز بزرگ میدان تصادفی گوسی در روش SPDE با عناصر خطی، در مقایسه با مدل دقیق مترن، این است که انتگرال‌هایی از میدان تصادفی گوسی در یک زیرناحیه دلخواه از فضای مطالعه D را می‌تواند به‌عنوان ترکیب خطی از ضرایب نمایش دهد که به‌عنوان ردیف‌هایی از یک ماتریس بیان می‌شود. این به آن معنی است که می‌توان هم‌زمان، به صورت توام، هم مشاهدات نقطه‌ای (زمین‌آماری) و هم مشاهدات مشبکه‌ای را با هم برازش داد بدون این‌که در محاسبات مربوط به کوواریانس به مشکل برخورد (موراگا، ۲۰۱۷). از این ویژگی می‌توان، برای مثال، در مدل‌بندی همزمان میزان بارش باران در ایستگاه هواشناسی و میزان آب روان از حوضه‌های آبریز رودخانه‌ها استفاده کرد. این روش همچنین، محاسبه عدم قطعیت را برای نقشه‌های کانتور مقدور می‌سازد، همان‌طوری که بولین و لیندگرن (۲۰۱۵) و بولین و لیندگرن (۲۰۱۷) نشان دادند. راحتی افزودن مشاهدات مشبکه‌ای به این معنی است که در بعضی موارد (در صورت نیاز) می‌توان از این مدل به جای مدل بی‌سیگ بهره گرفت تا در مدل‌های فضایی با اندازه‌های در حال تغییر و مدل‌های فضایی‌ـزمانی که مرزهای مربوط به نواحی در طول بازه زمانی مد نظر تغییر می‌کند، ساختار وابستگی واقعی‌تری را به نمایش گذاشت.

۱.۲.۵ پارامتربندی و پیشین‌ها

متأسفانه پارامترهای تابع مترن تفسیرپذیری مشکلی دارند. به همین علت لیندگرن و همکاران (۲۰۱۱) تعریف توزیع پیشین بر روی انحراف استاندارد کناری σ و دامنه تجربی $r = \sqrt{8\nu}/\kappa$ را پیشنهاد و توصیه کردند. این توصیه در بسته نرم‌افزاری R-INLA نیز پیاده شده است. تفسیر این دو پارامتر راحت است. اگر بتوان چند نمونه از پارامترهای فضایی استخراج کرد، σ تغییرات میدان فضایی را نشان می‌دهد و r فواصل بین نواحی کوچک و بزرگ را متصل می‌کند. پیشین توأم برای r و σ به‌وسیله فلاگستمن و همکاران (۲۰۱۷) بر اساس رهیافت پیشین‌های با پیچیدگی جریمه‌شده[۷] (سیمپسون و همکاران، ۲۰۱۷) توسعه یافته است. توزیع پیشین پیشنهادی در عمل و کاربرد موفق بوده است (واکفید و همکاران، ۲۰۱۸). برای اطلاعات بیشتر در مورد این پیشین‌ها به سیمپسون و همکاران (۲۰۱۷) مراجعه کنید.

۳.۵ مقایسه روش‌های لوجیت تکی‌شده و چندجمله‌ای به کمک شبیه‌سازی و مثال واقعی

۱.۳.۵ مطالعه شبیه‌سازی

در این بخش عملکرد مدل پیشنهادی را در مقایسه با مدل معمولی چندجمله‌ای در چارچوب تحلیل داده‌های زمین‌آماری به کمک شبیه‌سازی مطالعه و بررسی می‌کنیم. برای این کار، ۲۰۰ نقطه را در یک ناحیه $[0,1] \times [0,1]$ به صورت تصادفی تولید کردیم. پیشگوی جمعی به صورت

$$\mu_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \beta_{2j} x_{2ij} + u(s_{ij}), \quad i = 1, \ldots, 200, \quad j = 1, 2, 3$$

در نظر گرفته شد، که در آن $J = 3$ تعداد رده‌ها و $u(s_{ij})$ اثر فضایی است که تحقق‌های آن از یک میدان مترن با پارامترهای واقعی $\sigma^2 = 5$، $\nu = 1$ و $\kappa = 7$ تولید شدند. متغیرهای تبیینی x_1 و x_2 نیز از توزیع نرمال استاندارد تولید شدند. دقت کنید که تنظیمات شبیه‌سازی برای هر سه رده یکسان در نظر گرفته شد. رده اول را به‌عنوان رده گوشه انتخاب کردیم. بنابراین

$$\beta_{01} = \beta_{11} = \beta_{21} = 0, \; u(s_{i1}) = 0.$$

با این قید مقادیر واقعی ضرایب رگرسیونی را به صورت $\boldsymbol{\beta}_0 = (0, 1, 1), \boldsymbol{\beta}_1 = (0, -1, 1)$ و $\boldsymbol{\beta}_2 = (0, 1, 1)$ در نظر گرفتیم. عملکرد روش تکی‌شده را با ارزیابی کارایی نسبی برآوردهای حاصل از آن نسبت به برآورد ماکسیمم درستنمایی محدودشده[۸] مدل چندجمله‌ای (وود،

[۷] Penalized complexity (PC)
[۸] Restricted maximum likelihood estimate

۲۰۱۱) که در بسته نرم‌افزاری mgcv قابل محاسبه است، مورد بررسی قرار دادیم. در روش تکی‌شده، دو مدل لوجیت شامل مقایسه رده دوم با اول و رده سوم با اول برازش می‌شوند. توزیع‌های پیشین برای پارامترها و ابرپارامترهای مدل‌ها را همان پیشین‌های پیش‌فرض بسته R-INLA در نظر گرفتیم: یعنی پیشین‌های نرمال برای پارامترهای رگرسیونی با میانگین صفر و واریانس‌های بزرگ 10^3 و پیشین‌های PC برای ابرپارامترهای مدل مترن. برای ارزیابی کامل‌تر کارایی مدل، دو سناریو مختلف را مد نظر قرار دادیم. در سناریو اول حالتی را در نظر گرفتیم که در آن فراوانی مربوط به رده‌ها نامتعادل با نسبت ۰/۳۴ است. در سناریوی دوم، فراوانی مربوط به رده‌ها متعادل انتخاب شدند. جدول‌های ۱.۵ و ۲.۵ خلاصه‌ای از استنباط‌های پسینی مربوط به ضرایب رگرسیونی و پارامترهای تابع کوواریانس مترن را، برای دو سناریوی شبیه‌سازی، نمایش می‌دهند. علاوه بر این، میزان کارایی از دست‌رفته برآوردهای حاصل از مدل لجیت تکی‌شده را نسبت به مدل چندجمله‌ای حساب کردیم که در این جدول‌ها گزارش شده‌اند. برای محاسبه میزان کارایی از دست‌رفته از فرمول‌های

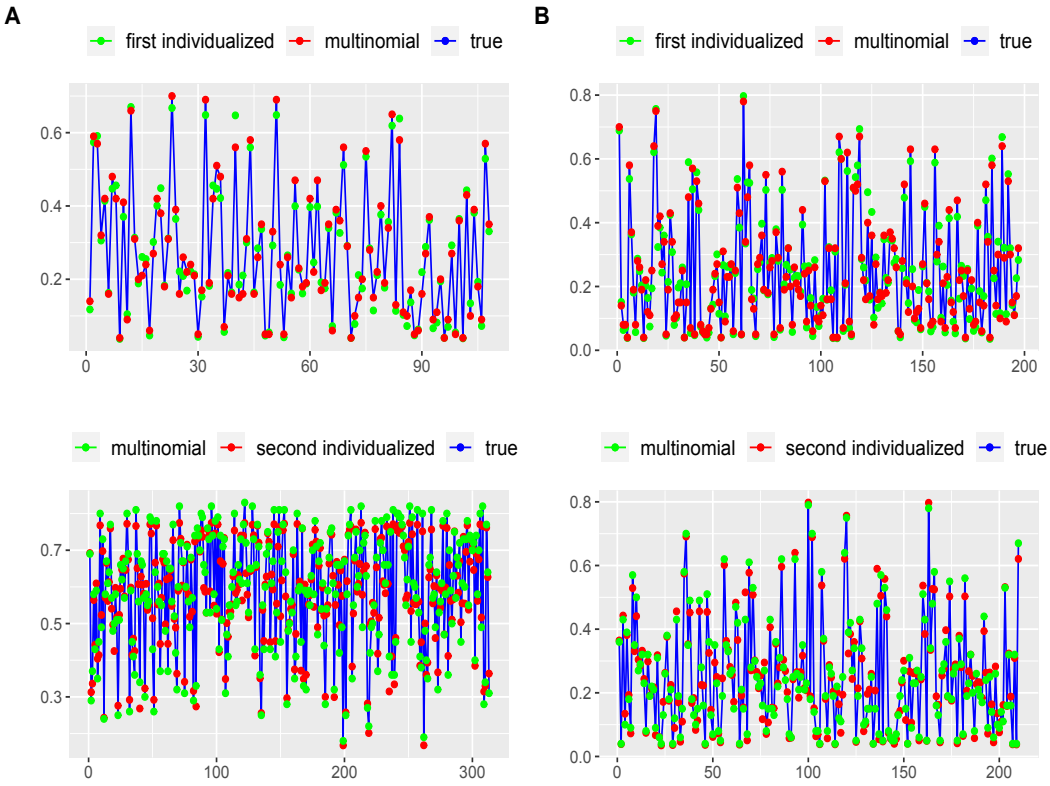
$$\mathrm{MSE}(\beta) = \sum_{k=1}^{K} \frac{(\beta_{\ell j} - \hat{\beta}_{\ell j}^{(k)})^2}{K}, \quad \ell = 0, 1, 2,$$

$$\mathrm{MSE}(\sigma) = \sum_{k=1}^{K} \frac{(\sigma - \hat{\sigma}^{(k)})^2}{,} K$$

$$\mathrm{MSE}(r) = \sum_{k=1}^{K} \frac{(r - \hat{r}^{(k)})^2}{K},$$

می‌توان استفاده کرد، که در آن β، σ و r مقادیر واقعی هستند و $K = 100$ تعداد تکرارهای شبیه‌سازی است. همچنین $\hat{\beta}$، $\hat{\sigma}$ و $\hat{r}$ برآوردهای ماکسیمم درستنمایی محدودشده هستند که برای محاسبه MSE برآوردهای روش تکی‌شده باید آن‌ها را با برآوردهای بیزی (میانگین پسینی) جایگزین کرد.

برآوردهای حاصل از مدل بیزی لوجیت‌های تکی‌شده و مدل کلاسیک چندجمله‌ای تقریبا به مقادیر واقعی نزدیک هستند. در حالت کلی، مقادیر مربوط به کارایی همه پارامترها بالا هست. اگرچه در بعضی موارد کارایی از دست‌رفته به حدود ۳۴٪ هم می‌رسد، اما دسترسی و آسانی برازش مدل لجیت‌های تکی‌شده و دوری از مشکل شناسایی‌پذیری ذاتی مدل چندجمله‌ای، انگیزه کافی را برای نادیده گرفتن این میزان کارایی از دست‌رفته به خوبی ایجاد می‌کند. در مقایسه با دو سناریوی فراوانی متعادل و نامتعادل، روشن است که اگر فراوانی رده‌ها متعادل باشند، میزان از دست دادن کارایی برآوردهای مدل لوجیت‌های تکی‌شده کمتر از حالت نامتعادل است. روند مشابهی در برآورد احتمال‌های رده‌ها دیده می‌شود. برآورد احتمال‌های رده‌ها برای دو سناریو در شکل ۱.۵ نمایش داده شده است که تقریبا به مقادیر واقعی نزدیک هستند.



شکل ۱.۵: برآورد احتمال رده‌های دوم و سوم در سناریوهای نامتعادل (چپ) و متعادل (راست). ردیف بالا مربوط به رده دوم و ردیف پایین مربوط به رده سوم است.

جدول ۱.۵: میانگین (انحراف استاندارد) پسینی ضرایب رگرسیونی در مدل‌های لوجیت‌های تکی‌شده و چندجمله‌ای برای سناریوی نامتعادل.

			چندجمله‌ای		
پارامتر	مقدار واقعی	لوجیت تکی اول–دوم	لوجیت تکی اول–سوم	رده دوم	رده سوم	کارایی از دست‌رفته
β_{01}						
β_{02}	۱	۰/۸۵ (۰/۰۸)		۰/۹۵ (۰/۰۹)		۰/۱۶
β_{03}	۱		۱/۱۹ (۰/۱۱)		۱/۰۱ (۰/۰۴)	۰/۲۴
β_{11}						
β_{12}	−۱	−۰/۷۴ (۰/۰۵)		−۰/۹۱ (۰/۰۳)		۰/۳۴
β_{13}	۱		۰/۸۲ (۰/۱۱)		۰/۹۶ (۰/۰۸)	۰/۲۶
β_{21}						
β_{22}	۱	۱/۲۰ (۰/۱۶)		۰/۹۵ (۰/۱۰)		۰/۲۸
β_{23}	۱		۱/۱۴ (۰/۱۲)		۰/۹۳ (۰/۰۵)	۰/۲۵
σ	۲/۲۳	۲/۲۱۹	۲/۵۴۰	۲/۱۲۲	۲/۱۴۴	۰/۱۵
r	۰/۴	۰/۳۵	۰/۴۳	۰/۴۳	۰/۴۱	۰/۲۲

۴.۵ مثال کاربردی: کیفیت آب‌های زیرزمینی

در این بخش، مدل‌بندی کیفیت منابع آب‌های زیرزمینی در استان گلستان مورد نظر است. استان گلستان در نقشه ایران در شکل ۲.۵ نشان داده شده است. مسئله کیفیت آب در سرتاسر

جدول ۲.۵: میانگین (انحراف استاندارد) پسینی ضرایب رگرسیونی در مدل‌های لجیت‌های تکی‌شده و چندجمله‌ای برای سناریوی متعادل.

			چندجمله‌ای		
پارامتر	مقدار واقعی	لوجیت تکی اول–دوم	لوجیت تکی اول–سوم	رده دوم	رده سوم	کارایی از دست‌رفته
β_{01}						
β_{02}	۱	۰/۹۱ (۰/۱۰)		۱/۰۱ (۰/۰۴)		۰/۱۷
β_{03}	۱		۰/۹۸ (۰/۰۹)		۱/۰۵ (۰/۰۸)	۰/۱۰
β_{11}						
β_{12}	−۱	−۰/۹۴ (۰/۰۷)		−۱ (۰/۰۳)		۰/۰۹
β_{13}	۱		۰/۹۲ (۰/۱۱)		۱/۰۱ (۰/۰۸)	۰/۲۳
β_{21}						
β_{22}	۱	۱/۰۹ (۰/۱۱)		۰/۹۳ (۰/۱۰)		۰/۲۶
β_{23}	۱		۱/۰۰ (۰/۱۲)		۰/۹۵ (۰/۰۳)	۰/۱۵
σ	۲/۲۳	۲/۱۱	۲/۲۹	۲/۰۱	۲/۱۵	۰/۲۱
r	۰/۴	۰/۳۸	۰/۴۱	۰/۳۹	۰/۳۸	۰/۱۲

دنیا به‌ویژه در کشور عزیزمان ایران به‌دلیل کمبود آب و برخی از سایر مسائل جغرافیایی از اهمیت بالایی برخوردار است و همیشه نیاز به وجود آب تازه موجب شده است تا مردم گرایش فراوانی به سمت آب‌های زیرزمینی داشته باشند. هدف اولیه از تحلیل و مدل‌بندی این داده‌ها، پیدا کردن رابطه‌ای بین عناصر موجود در آب و کیفیت آن است تا بتوان اطلاعات مفیدی را برای مدیریت و برنامه‌ریزی آینده آب فراهم کرد.

شکل ۲.۵: استان گلستان: ناحیه هاشورخورده در نقشه ایران.

داده‌ها از ۱۵۸ موقعیت مکانی شامل چشمه‌ها و چاه‌های آب در بخش بزرگی از استان گلستان گردآوری شده‌اند. این موقعیت‌ها، در واقع، شامل ایستگاه‌های سنجش کیفیت آب هستند. استان گلستان به مساحت ۲۰۳۶۷ کیلومتر مربع در کنار دریای خزر واقع شده است. این استان هم از نظر کشاورزی و هم صنعتی از مناطق مهم کشور است.

برای اندازه‌گیری و سنجش کیفیت آب‌های زیرزمینی یک شاخص معمول و استاندارد، رسانایی الکتریکی[۹] است که با عبور جریان الکتریکی بین دو صفحه فلزی (الکترود) اندازه‌گیری می‌شود. این متغیر بر حسب میکرومز[۱۰] در هر سانتی‌متر اندازه‌گیری می‌شود و تحت تأثیر

[۹] Electrical Conductivity (EC)
[۱۰] Micromhos

مواد جامد مانند کلوراید، سولفات، منیزیم و سدیم است. در واقع هرچه مقدار این مواد در آب بیشتر باشند، مقدار EC بیشتر و در نتیجه کیفیت آب پایین‌تر است. بنا بر یک استاندارد پذیرفته‌شده جهانی، آب مناسب آشامیدن دارای مقدار EC کمتر از ۸۰۰ میکرومز است (لیوز، ۲۰۰۳). به‌دلیل طبیعت آب‌های زیرزمینی و تغییر مقدار EC آن بر اساس اقلیم‌های متفاوت، وجود مواد جامد و املاح متفاوت در آن و تأثیر رفتار انسان‌ها بر آن، در نظر گرفتن تغییرات فضایی در تحلیل کیفیت آب امری ضروری است. اهمیت این ضرورت بیشتر هم درک می‌شود وقتی که بدانیم مدل‌بندی و تحلیل درست کیفیت آب بر اساس مدل‌های فضایی به نتایج زیر منتهی می‌شود:

- اطلاع‌رسانی و آگاهی‌بخشی به عموم مردم در استفاده از نقاط آب‌خیز برای آشامیدن یا مصارف دیگر.
- تولید اطلاعات مفید برای مدیریت آب و برنامه‌ریزی‌های مربوط به آن
- شناسایی مواد جامد با تأثیر زیاد بر میزان کیفیت آب

با اهداف مطرح‌شده، برای رده‌بندی کیفیت آب از EC استفاده کردیم که در آن رده مشاهدات (متغیر پاسخ)، بر اساس لیوز (۲۰۰۳)، برای سه رده به‌صورت زیر تعریف شدند:

۱. آب قابل شرب: $۰ < EC < ۸۰۰$

۲. کیفیت نامناسب برای شرب ولی مناسب برای کشاورزی: $۸۰۰ < EC < ۲۵۰۰$

۳. کیفیت مناسب فقط برای اهداف صنعتی: $EC > ۲۵۰۰$

برای کنترل و مدیریت کیفیت آب باید در ابتدا عناصر موجود در آن را شناسایی کرد. بنابراین، باید اثرات مربوط به عناصری که در خاک اندازه‌گیری شده‌اند را مورد بررسی قرار دهیم که خلاصه‌ای از این عناصر در جدول ۳.۵ گزارش شده‌اند. مقادیر ثبت‌شده برای هر متغیر میانگین در سال است.

جدول ۳.۵: متغیرهای تبیینی موجود در مجموعه داده آب استان گلستان

متغیر تبیینی	نماد	واحد
پتاسیم	k	$L^{۱}$ mg
بی‌کربنات	$hco_{۳}$	$L^{۱}$mg
کلوراید	cl	$L^{۱}$mg
سدیم	na	$L^{۱}$mg
سختی کل	th	$L^{۱}$mg
کل جامدات محلول‌شده	tds	$L^{۱}$mg

۱.۴.۵ مدل‌بندی و برآورد پارامتر

برای تحلیل داده‌ها یک مدل چندجمله‌ای فضایی داریم که پیشگوی جمعی در آن به صورت

$$\eta_{ij} = \beta_{\circ j} + \sum_{k=۱}^{۶} \beta_{kj} x_{kij} + u(s_{ij}), \quad i = ۱, \ldots, ۱۵۸, \quad j = ۱, ۲, ۳$$

است، که در آن x_kها، برای همه رده‌ها، همان متغیرهای تبیینی نوشته‌شده در جدول ۳.۵ هستند. رده سوم را به‌عنوان رده گوشه در نظر گرفتیم، زیرا این رده بیانگر کیفیت پایین آب است که نه برای آشامیدن مناسب است و نه برای کشاورزی. به همین دلیل مایل هستیم تا مقایسه‌ای که توسط مدل لجیت‌های تکی‌شده انجام می‌شود بین دو سطح کیفیت قابل شرب و قابل کشاورزی با رده‌ای باشد که فقط برای مصارف صنعتی مناسب است. توزیع‌های پیشین مدل لوجیت‌های تکی‌شده را مشابه با مثال شبیه‌سازی انتخاب کردیم. با یک تحلیل مقدماتی در مورد انتخاب متغیرهای تبیینی موثر، تنها حضور دو متغیر سدیم و سختی کل در مدل نهایی تایید شد. نتایج برازش مدل کلاسیک چندجمله‌ای و مدل بیزی لوجیت‌های تکی‌شده در جدول ۴.۵ گزارش شده‌اند. نتایج برازش در دو مدل تقریبا شبیه هم هستند؛ مقادیر انحراف استاندارد پسینی در مدل لجیت‌های تکی‌شده در اکثر موارد کوچکتر از انحراف استاندارد (مبتنی بر درستنمایی محدودشده) متناظر در مدل چندجمله‌ای است. در نتیجه فواصل اطمینان کلاسیک ۹۵٪ در مدل چندجمله‌ای کمی پهن‌تر از فواصل اعتبار بیزی مدل لوجیت‌های تکی‌شده است. همان‌گونه که انتظار داشتیم ضرایب دو متغیر تبیینی سدیم و سختی کل در هر دو رده کیفیت قابل شرب و مناسب برای کشاورزی (نسبت به مصارف فقط صنعتی) منفی هستند که نشان‌دهنده تاثر معکوس این دو عامل بر کیفیت آب هستند؛ یعنی هر چه مقادیر این دو متغیر در آب بیشتر شوند، کیفیت آب کاهش پیدا خواهد کرد. شکل‌های ۳.۵ و ۴.۵، به ترتیب، نقشه‌های پهنه‌بندی اثرات فضایی و انحراف معیار متناظر را برای رده‌های اول و دوم در ناحیه مورد مطالعه نشان می‌دهند. نقشه احتمال وقوع هر رده نیز در شکل ۵.۵ نمایش داده شده است. نتایج حاصل از دو مدل قابل رقابت و شبیه هستند.

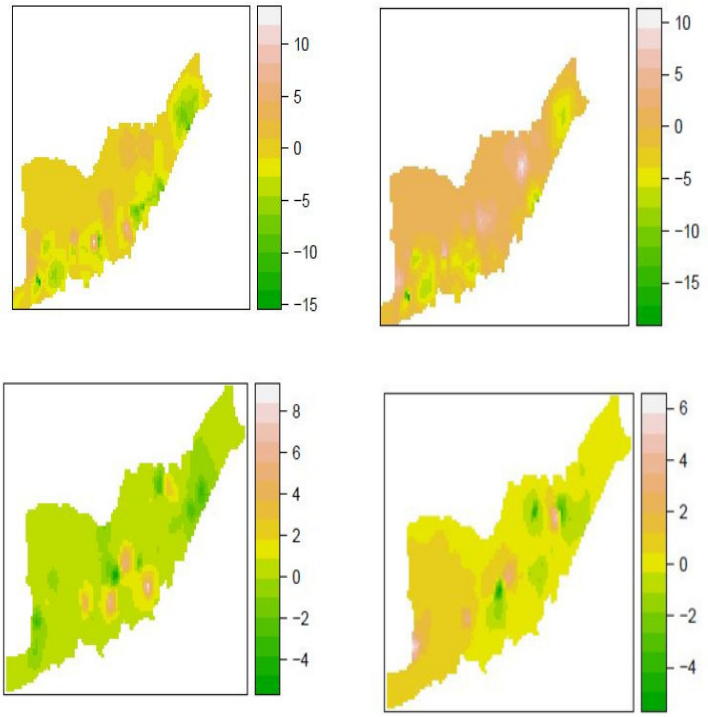
با در نظر گرفتن اثر فضایی در شکل ۳.۵ در می‌یابیم که در نواحی شمال شرقی و جنوبی اثر فضایی منفی است و کیفیت آب این مناطق برای شرب پایین است. در نواحی مرکزی و شمالی، آب با کیفیت قابل شرب را بیشتر می‌توان یافت زیرا این نواحی دارای اثر فضایی بزرگ و مثبت هستند. برآوردهای احتمال در شکل ۵.۵ نیز این نتایج را تایید می‌کنند.

۵.۵ خلاصه و نتیجه‌گیری

در این فصل تحلیل بیزی داده‌های چندجمله‌ای فضایی از نوع زمین‌آماری با مدل لوجیت‌های تکی‌شده مد نظر قرار گرفت و عملکرد مدل پیشنهادی با مدل چندجمله‌ای کلاسیک مقایسه

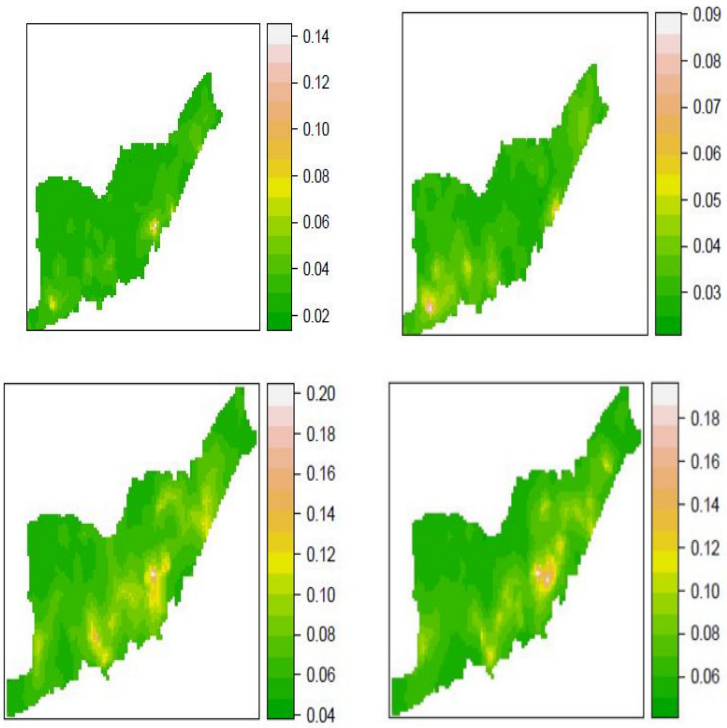
جدول ۴.۵: خلاصه‌ای از برآوردهای ضرایب رگرسیونی و پارامترهای کوواریانس مترن دو مدل برای مجموعه داده کیفیت آب استان گلستان.

		رده اول			رده دوم		
مدل		میانگین	انحراف استاندارد	CI ۹۵%	میانگین	انحراف استاندارد	CI ۹۵%
لوجیت تکی‌شده							
	عرض از مبدأ	۳/۱۹	۰/۰۹	(۲/۱۳, ۴/۲۷)	۲/۶۴	۰/۰۶	(۰/۱۰, ۴/۲۲)
	na	−۰/۰۳	۰/۰۸	(−۱/۰۱, −۰/۰۰۸)	−۰/۰۵۱	۰/۰۷	(−۰/۱۱, −۰/۰۰۲)
	th	−۰/۶۲	۰/۱۶	(−۰/۸۱, −۰/۳۱)	−۰/۴۳	۰/۱۵	(−۰/۹۶, −۰/۰۱)
	σ	۶/۲۳	۰/۲۱	(۲/۲۱, ۸/۵۹)	۷/۳۲	۰/۱۵	(۱/۲۳, ۱۳/۵۶)
	r	۰/۷۱	۰/۱۵	(۰/۰۱, ۱/۳۹)	۱/۲۱	۰/۱۶	(۰/۶۳, ۲/۳۱)
چندجمله‌ای							
	عرض از مبدأ	۴/۶۶	۰/۱۲	(۲/۵۵, ۶/۹۳)	۱/۸۲	۰/۱۱	(۰/۳۳, ۲/۵۱)
	na	−۰/۰۱۱	۰/۱۱	(−۰/۱۵, −۰/۰۰۷)	−۰/۰۷	۰/۱۳	(−۰/۱۸, −۰/۰۳)
	th	−۰/۶۹	۰/۱۹	(−۰/۹۱, −۰/۳۱)	−۰/۳۵	۰/۱۴	(−۰/۵۱, −۰/۱۲)
	σ	۵/۲۴	۰/۱۲	(۱/۲۱, ۹/۴۱)	۶/۳۳	۰/۱۳	(۰/۰۱, ۱۲/۹۶)
	r	۰/۶۸	۰/۳۱	(۰/۰۷, ۱/۳۱)	۰/۷۵	۰/۲۳	(۰/۱۶, ۱/۵۲)



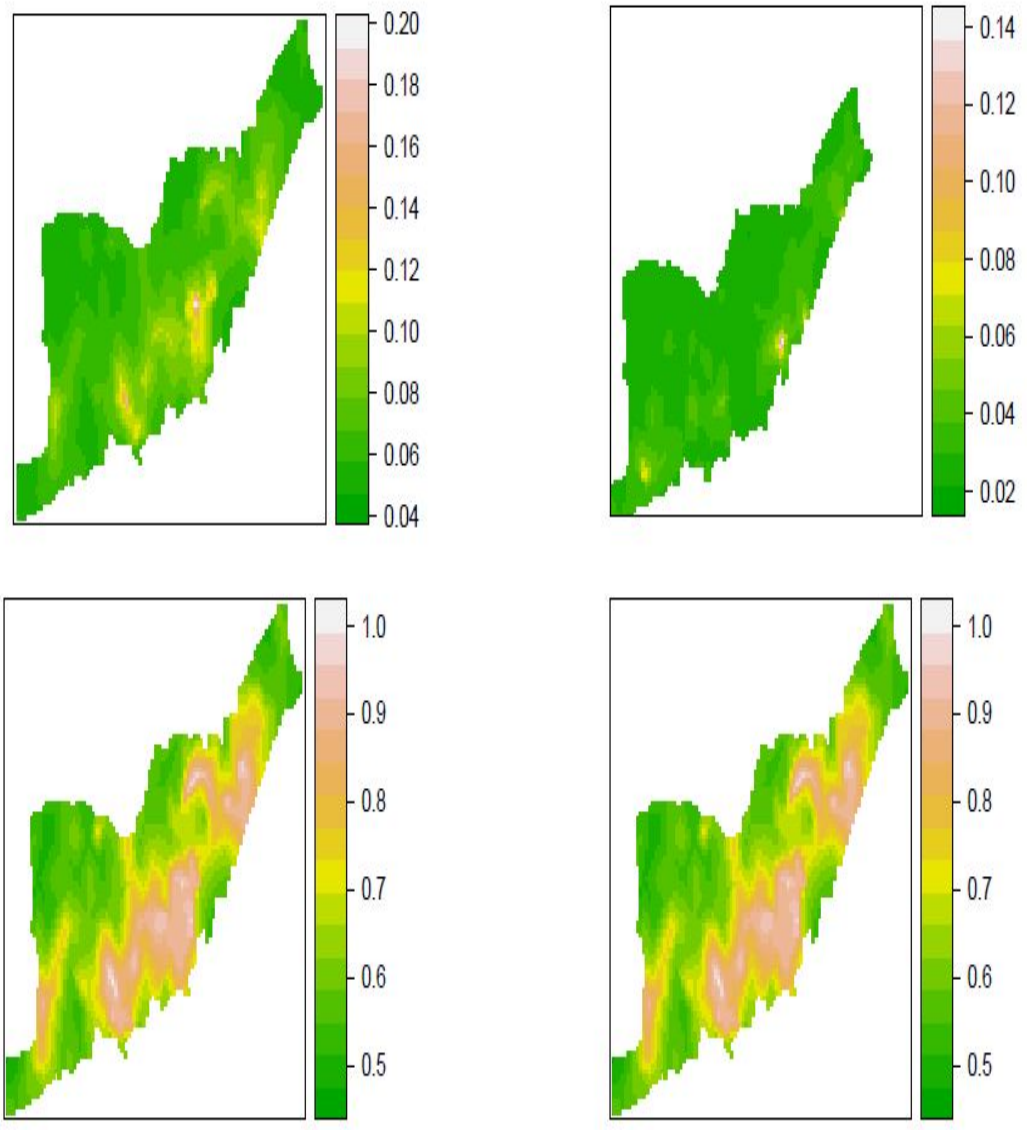
شکل ۳.۵: میانگین پسینی اثر فضایی برای رده اول (ردیف اول) و رده دوم (ردیف دوم) در دو مدل لوجیت‌های تکی‌شده (سمت چپ) و چندجمله‌ای (سمت راست).

شد. در واقع برای دوری از ساختار پیچیده مدل و داشتن یک مدل کارا با برازش آسان، مدل لوجیت‌های تکی‌شده را پیشنهاد کردیم. برای برازش مدل از روش INLA استفاده کردیم. با استفاده از این مدل‌بندی می‌توان حتی اثرات ناخطی را وارد مدل کرد و همچنان با سادگی برازش روبرو شد. از آن‌جایی که خیلی از بسته‌های نرم‌افزاری مربوط به روش بیزی قابلیت بهره‌مندی از مدل چندجمله‌ای با این ساختار را ندارند، یا زمان محاسباتی و حافظه مورد نیاز در آن‌ها حجیم است، یا قابلیت به‌کارگیری روش SPDE را ندارند به ناچار برای مقایسه



شکل ۴.۵: انحراف استاندارد پسینی اثر فضایی برای رده اول (ردیف اول) و رده دوم (ردیف دوم) در دو مدل لوجیت‌های تکی‌شده (سمت چپ) و چندجمله‌ای (سمت راست).

مدل پیشنهادی با مدل چندجمله‌ای از روش برآورد محدودشده در بسته نرم‌افزاری mgcv بهره گرفتیم. برای مقایسه جزئی‌تر نتایج مشاهده‌شده، کارایی برآوردها را نیز محاسبه کردیم. نتایج حاصل از دو مدل را نزدیک به یکدیگر و نزدیک به مقادیر واقعی یافتیم. میزان کارایی از دست‌رفته در مدل لوجیت‌های تکی‌شده با در نظر گرفتن سختی‌های پیش‌رو در برازش مدل چندجمله‌ای و دسترس نبودن بسته‌های نرم‌افزاری مناسب، قابل چشم‌پوشی است. از طرفی مشکل شناسایی‌پذیری ذاتی مدل چندجمله‌ای نیز وجود ندارد. در باور ما امروزه با پیشرفت در علومی همچون یادگیری ماشین و علم داده‌ها بسیاری از افرادی که لزوماً در رشته آمار نیستند، مثل متخصصان رشته‌های علوم اجتماعی، اقتصاد، و ژئوفیزیک، و دانش کدنویسی و ریاضی محدودی دارند، و حتی شاید توانایی لازم در به‌کارگیری مستقیم الگوریتم‌ها را ندارند، نیاز به تحلیل داده‌ها با این ساختار دارند که با به‌کارگیری مدل لجیت‌های تکی‌شده و افزایش کارایی آن به کمک روش INLA قادر خواهند بود تا بستری مناسب در این زمینه را فراهم آورند. در پایان نیز، برای ارزیابی بیشتر کارایی مدل، کاربست آن را برای تحلیل مجموعه داده واقعی مربوط به کیفیت آب در استان گلستان نشان دادیم.



شکل ۵.۵: برآوردهای احتمال برای رده اول (ردیف اول) و رده دوم (ردیف دوم) در دو مدل لوجیت‌های تکی‌شده (سمت چپ) و چندجمله‌ای (سمت راست).

فصل ۶

مدل‌بندی داده‌های چندجمله‌ای فضایی–زمانی

در این فصل سعی کردیم تا ساختار وابستگی فضایی–زمانی را در یک مدل چندجمله‌ای با ساختار جمعی–رگرسیونی تعبیه کنیم و داده‌های چندجمله‌ای فضایی–زمانی را با مدل پیشنهادی تحلیل کنیم. برای برازش مدل از روش تقریبی INLA استفاده می‌کنیم. این نگاه برای مدل چندجمله‌ای، تا جایی که خبر داریم، تاکنون توسط سایرین استفاده نشده است. در این‌جا نیز برای رفع مشکل شناسایی‌پذیری مدل از قیدهای رده گوشه و مجموع برابر با صفر استفاده می‌کنیم. در ادامه، ابتدا نحوه وارد کردن انواع اثر متقابل فضایی–زمانی را تشریح می‌کنیم.

۱.۶ اثر فضایی–زمانی

امروزه پیشرفت ابزار و فنون ثبت داده‌ها و همچنین کارایی این ابزار در به‌روز رسانی و ذخیره کردن داده‌های حجیم با فاصله زمانی طولانی، موجب شده است تا افراد برای بررسی و تحلیل آماری پدیده‌هایی در علوم مختلف مانند پزشکی، بوم‌شناسی، زمین‌شناسی و علوم وابسته به آب و خاک که هم به زمان و هم به مکان وابسته‌اند، علاقه بیشتری از خود نشان دهند. در نتیجه، آن‌ها می‌توانند از انتشار دست‌آوردهای آماری برای برنامه‌ریزی‌های آینده کره زمین

بهره گیرند.

با توجه به جایگاه ویژه و اهمیت مدل چندجمله‌ای در تحلیل داده‌ها و در علوم مختلف، بر آن شدیم تا تحلیل کاراتر داده‌های چندجمله‌ای فضایی‌ـ‌زمانی را با در نظر گرفتن وابستگی فضایی‌ـ‌زمانی آن‌ها مد نظر قرار دهیم. برای اثر فضایی، مشابه فصل‌های قبل از مدل ICAR در یک ناحیه گسسته و مدل SPDE برای یک ناحیه پیوسته چگال می‌توان استفاده کرد.

ساختار فضایی‌ـ‌زمانی در داده‌ها در فصل اول و بخش ۳.۶.۱ معرفی شد. برای یک مرور ساده می‌توان گفت، اثر فضایی را می‌توان به راحتی و با افزودن اثر زمانی به اثر فضایی‌ـ‌زمانی توسعه داد. یک فرآیند فضایی‌ـ‌زمانی به صورت

$$Y(s,t) \equiv \{y(s,t), \quad (s,t) \in D \subset \mathbb{R}^۲ \times \mathbb{R}^+\}$$

است، که در آن فرض می‌کنیم در n مکان فضایی (نقطه فضایی) و T نقطه زمانی مشاهده شده است. معمولا افراد میدان تصادفی گوسی را برای این فرآیند در نظر می‌گیرند که دارای ماتریس کوواریانس $C((s,t),(s',t'))$ است. در این حالت، داده‌ها می‌توانند از نوع زمین‌آماری، مشبکه‌ای یا نوع سوم یعنی الگوهای نقطه‌ای فضایی باشند. در این فصل، به‌طور خاص، داده‌های مشبکه‌ای را در نظر می‌گیریم. در این حالت، مدل‌بندی دو ماتریس دقت مربوط به ساختار همسایگی مکان‌ها و ماتریس دقت مربوط به اثر زمانی مطرح هستند. مدل‌های مختلفی برای مدل‌بندی پدیده‌های فضایی‌ـ‌زمانی معرفی شده‌اند (ویکل و کرسی، ۲۰۱۱). راحت‌ترین راه، در نظر گرفتن اثر متقابل بین دو اثر فضایی و زمانی است. یکی از راه‌ها در نظر گرفتن اثر متقابل بین فضا و اتورگرسیو مرتبه اول در طول زمان است (مارتینز بنتیو و همکاران، ۲۰۰۸؛ ویوار و فرینو، ۲۰۰۹؛ راشورس و همکاران، ۲۰۱۴؛ رویز و همکاران، ۲۰۱۲؛ بلانجیاردو و همکاران، ۲۰۱۵). با این وجود، هیچ‌کدام از این مطالعات به فضا و زمان به‌عنوان اثر اصلی توجه نکرده‌اند. نور‌ـ‌هلد (۲۰۰۰)، مدلی را معرفی کرد که در آن اثرهای اصلی به قوت حضور دارند و برای مولفه فضایی‌ـ‌زمانی نیز چهار نوع اثر متقابل لحاظ می‌کند. در زیر جزئیات بیشتری از آن‌ها را تشریح می‌کنیم.

۱.۱.۶ اثرات متقابل ذاتی فضایی‌ـ‌زمانی

برای توضیح بیشتر درباره اثرات متقابل زمان و فضا، پیشگوی خطی را به صورت

$$\eta_{it} = \sum_{k=۱}^{l} x_{kit}\beta_k + u_{it} + \nu_{it} + \delta_{it}, \quad i = ۱,\ldots,n, \quad t = ۱,\ldots,T, \tag{۱.۶}$$

در نظر می‌گیریم. همانند نور‌ـ‌هلد (۲۰۰۰) و بر اساس نیاز داده‌های در دسترس، مدل فضایی را مدل بی‌سیگ در نظر می‌گیریم. پیشگوی جمعی‌ـ‌ساختاری معرفی‌شده در (۱.۶) را در نظر بگیرید. مولفه u بیانگر اثر فضایی، ν نشان‌دهنده اثر زمانی و δ اثر فضایی‌ـ‌زمانی است. فرض می‌کنیم همه اثرات تصادفی دارای توزیع گوسی با میانگین صفر و ماتریس‌های دقت $\tau_u\mathbf{R}_u$

برای $\mathbf{u}$، $\tau_\nu \mathbf{R}_\nu$ برای $\boldsymbol{\nu}$ و $\tau_\delta \mathbf{R}_\delta$ برای $\boldsymbol{\delta}$ است، که در آن τ_u، τ_ν، τ_δ و $\mathbf{R}_u$، $\mathbf{R}_\nu$، $\mathbf{R}_\delta$ به ترتیب پارامترهای دقت و ماتریس‌های دقت متناظر با اثرات فضایی، زمانی و فضایی‌ـ زمانی هستند. در این‌جا، $\mathbf{R}_\delta$ را می‌توان به‌عنوان حاصلضرب کرونه‌کر ماتریس دقت اثر زمانی و ماتریس دقت اثر فضایی نوشت.

۲.۱.۶ اثرات اصلی

ماتریس ساختار برای اثر زمانی، $\boldsymbol{\nu}$، را می‌توان به صورت $\mathbf{R}_\nu = \tilde{\mathbf{H}} - \mathbf{H}$ نوشت، که در آن $\mathbf{H}$ ماتریس ساختار همسایگی اثر زمانی و $\tilde{\mathbf{H}}$ یک ماتریس قطری با عناصر روی قطر برابر با جمع ردیف ماتریس $\mathbf{H}$ است. در این‌جا، مدل قدم زدن تصادفی مرتبه اول را برای مدل‌بندی اثر زمانی در نظر می‌گیریم که ماتریس دقت آن برابر است با

$$\mathbf{R}_\nu = \begin{bmatrix} ۱ & -۱ & & & \\ ۱ & ۲ & -۱ & & \\ & \ddots & \ddots & \ddots & \\ & & -۱ & ۲ & -۱ \\ & & & -۱ & ۱ \end{bmatrix} = \begin{bmatrix} ۱ & & & \\ & ۲ & & \\ & & \ddots & \\ & & & ۲ \\ & & & & ۱ \end{bmatrix} - \begin{bmatrix} & ۱ & & & \\ ۱ & & ۱ & & \\ & \ddots & & \ddots & \\ & & ۱ & & ۱ \\ & & & ۱ & \end{bmatrix}.$$

به‌طور مشابه، ماتریس دقت برای اثر فضایی را نیز می‌توان به صورت $\mathbf{R}_u = \tilde{\mathbf{G}} - \mathbf{G}$ نوشت که در آن $\mathbf{G}$ ماتریس ساختار همسایگی فضایی با عناصر

$$G_{ij} = \begin{cases} ۱ & j \sim i \\ ۰ & \text{بقیه جاها}, \end{cases}$$

است و $\tilde{\mathbf{G}}$ یک ماتریس قطری با عناصر روی قطر برابر با جمع ردیف ماتریس $\mathbf{G}$ است.

۳.۱.۶ اثرات متقابل فضایی‌ـ زمانی

در تحلیل داده‌های با اثر فضایی مشبکه‌ای بر اساس کلیتون و همکاران (۱۹۹۶)، نور‌ـ هلد (۲۰۰۰) و بلانجیاردو و همکاران (۲۰۱۵)، ماتریس کوواریانس فضایی‌ـ زمانی را طوری در نظر می‌گیریم که جداپذیر باشد و از حاصلضرب کرونه‌کر ماتریس دقت اثر فضایی و ماتریس دقت اثر زمانی به‌دست آید. در حالت کلی، چهار نوع از این اثر به صورت زیر قابل تعریف هستند.

۱. **اثر فضایی‌ـ زمانی نوع اول**: اگر دو اثر فضایی و زمانی غیرساختاری باشند، آن‌گاه

$$\mathbf{R}_\delta = \mathbf{R}_u \otimes \mathbf{R}_\nu = \mathbf{I} \otimes \mathbf{I} = \mathbf{I}.$$

۲. **اثر فضایی‌ـ زمانی نوع دوم**: در این حالت، اثر فضایی غیرساختاری را با اثر زمانی ساختاری ترکیب می‌کنند. بنابراین، ماتریس ساختاری اثر فضایی‌ـ زمانی به صورت

$$\mathbf{R}_\delta = \mathbf{R}_u \otimes \mathbf{R}_\nu = \mathbf{I} \otimes \mathbf{R}_\nu,$$

بازنویسی می‌شود. در این نوع، ساختاری در فضا وجود ندارد اما روند زمانی از یک ناحیه به ناحیه دیگر فرق می‌کند.

۳. **اثر فضایی‌ـ‌زمانی نوع سوم:** در این نوع اثر متقابل، ساختار در اثر فضایی وجود دارد اما اثر زمانی غیرساختاری است. پس

$$\mathbf{R}_\delta = \mathbf{R}_\nu \otimes \mathbf{R}_u$$

که در آن $\mathbf{R}_u$ ماتریس ساختار همسایگی مدل بی‌سیگ است و $\mathbf{R}_\nu = \mathbf{I}$. در حقیقت ناهمگنی فضایی از یک نقطه زمانی تا نقطه زمانی دیگر متفاوت است بدون این‌که اثر زمانی ساختاری باشد.

۴. **اثر فضایی نوع چهارم:** در این نوع، هر دو ماتریس دقت اثر زمانی و اثر فضایی از نوع ساختاری فرض می‌شوند. بنابراین، ماتریس ساختاری اثر فضایی‌ـ‌زمانی برابر است با $\mathbf{R}_\delta = \mathbf{R}_u \otimes \mathbf{R}_\nu$. نوع چهارم پیچیده‌ترین و پرکاربردترین نوع اثر متقابل فضایی‌ـ‌زمانی است. ماتریس دقت در این نوع به صورت

$$\begin{aligned} \tau_\delta \mathbf{R}_\nu \otimes \mathbf{R}_u &= \tau_\delta(\tilde{\mathbf{H}} - \mathbf{H}) \otimes (\tilde{\mathbf{G}} - \mathbf{G}) \\ &= \tau_\delta(\tilde{\mathbf{H}} \otimes \tilde{\mathbf{G}} - \tilde{\mathbf{H}} \otimes \mathbf{G} - \mathbf{H} \otimes \tilde{\mathbf{G}} + \mathbf{H} \otimes \mathbf{G}), \end{aligned} \qquad (۲.۶)$$

به‌دست می‌آید. اولین جمله (۲.۶) یک ماتریس قطری است که مولفه‌هایش از ضرب تعداد همسایه‌های فضایی در تعداد همسایه‌های زمانی نتیجه می‌شوند. جمله دوم شامل همسایه‌های فضایی هستند که با تعداد همسایه‌های زمانی وزن‌دار شده‌اند. همچنین، جمله سوم همسایه‌های زمانی هستند که با تعداد همسایه‌های فضایی وزن‌دار شده‌اند. آخرین جمله هم ساختار همسایگی فضایی‌ـ‌زمانی است.

برای پیاده کردن اثر فضایی‌ـ‌زمانی، بنا بر نور‌ـ‌هلد (۲۰۰۰) باید قید مجموع برابر با صفر را برای این اثر نیز در نظر بگیریم. زیرا مدل ذاتی است و برای شناسایی نیاز به قید دارد. برای شرح بیشتر بخش زیر را دنبال کنید.

۴.۱.۶ قیدها و فضای صفر

از توجیح قیدهای مربوط به اثرات اصلی شروع می‌کنیم و به قیدهای مربوط به اثرات متقابل نوع دوم، سوم و چهارم می‌رسیم. بر اساس رو و هلد (۲۰۰۵)، ماتریس دقت یک IGMRF مرتبه اول را با ضرب پارامتر دقت در ماتریس ساختاری نشان دادیم. این ماتریس پررتبه نیست و کمبود رتبه‌ای برابر با یک دارد؛ به این معنی که فضای صفر در آن یک بعدی است. در نتیجه، ماتریس دقت آن یک مقدار ویژه برابر با صفر دارد. همه عناصر بردار ویژه متناظر هم مقدار یکسانی دارند. بنا بر رو و هلد (۲۰۰۵) (فصل سوم) مقداری که عناصر بردار ویژه به

خود اختصاص می‌دهد، می‌تواند یک باشد. بنابراین در این‌جا بردار $\frac{۱}{n}\mathbf{۱}$ را در نظر می‌گیریم، که در آن منظور از $\mathbf{۱}$ برداری از ۱ هاست. همان‌گونه که در فصل چهارم نیز اشاره شد، یک راه حل کابردی این است که مقدار کوچکی به عناصر ماتریس دقت فرایند IGMRF اضافه کنیم بدون این‌که چگالی آن عوض شود. راه دیگر این است که به کمک قید مجموع برابر با صفر، فرآیند را شناسایی‌پذیر کنیم.

برای اثر متقابل نوع دوم، فرآیند زمانی در δ یک IGMRF مرتبه اول مانند ν است. ماتریس ساختاری نیز $\mathbf{R}_\nu \otimes \mathbf{I}$ است، که در آن $\mathbf{I}$ یک ماتریس همانی با بعد n است. بنابراین، همه کمبود رتبه برابر با n است. فضای صفر برابر است با ضرب کرونه‌کر فضای صفر ماتریس دقت اثر زمانی در ماتریس همانی با بعد n؛ یعنی

$$\mathbf{W}_۲ = \frac{۱}{n}\mathbf{۱}_{T\times ۱} \otimes \mathbf{I}_n.$$

بنابراین به n قید مجموع برابر با صفر، هر کدام برای یک ناحیه یا هر کدام از n فرآیند زمانی نیاز داریم.

اثر متقابل نوع سوم به هر نقطه زمانی t یک میدان فضایی اختصاص می‌دهد. بنابراین δ یک IGMRF مرتبه اول فضایی مانند $\mathbf{u}$ است. کمبود رتبه $\mathbf{I}_T \otimes \mathbf{R}_u$ نیز برابر T است و فضای صفر،

$$\mathbf{W}_۳ = \mathbf{I}_T \otimes \frac{۱}{T}\mathbf{۱}_{n\times ۱}$$

است. بنابراین، قید مجموع برابر با صفر برای هر کدام از T میدان فضایی است.

اثر متقابل نوع چهارم یک IGMRF در دو جهت فضایی و زمانی است. کمبود رتبه در این حالت برابر با $n+T-۱$ است؛ زیرا به بیان ساده، برای n فرآیند زمانی نیاز به قید مجموع برابر با صفر داریم و همین‌طور برای میدان فضایی در هر یک از $T-۱$ نقطه زمانی. به طریقی دیگر، می‌توان گفت $T-۱$ قید برای اثر نوع سوم و nتا برای اثر نوع دوم نیاز داریم که در مجموع برابر است با $n+T-۱$. برای وارد کردن اثر فضایی–زمانی در مدل چندجمله‌ای و سپس برازش آن به کمک روش INLA، دوباره از تبدیل MP استفاده می‌کنیم. رده‌ها را طوری در نظر می‌گیریم که پاسخ همانند فصل چهارم دارای فراوانی باشد؛ چرا که در تحلیل داده‌های فضایی–زمانی با داده‌های چندجمله‌ای روبرو هستیم که اغلب از فراوانی برخوردارند. راه حل ما برای برطرف شدن مشکل شناسایی‌پذیری در حالت فضایی–زمانی، اضافه کردن قید مجموع برابر با صفر به مدل است. در این صورت، قیدهای زیادی در مدل داریم که توانایی نسبتاً محدود INLA در به‌کارگیری قیدها ما را با چالش روبرو می‌کند.

در روش INLA می‌توان قیدهای خطی به صورت $\mathbf{A}\boldsymbol{\chi} = \boldsymbol{e}$ را در مدل به‌کار گرفت، که در آن $\boldsymbol{\chi}$ بردار مولفه‌های پنهان با یک میدان گوسی است. طول بردار $\boldsymbol{e}$ و ردیف ماتریس $\mathbf{A}$ برابر با طول قیدهای مورد نیاز است. هزینه مثلا k قید در INLA برابر است با $\mathrm{O}(nk^۲)$ که n در آن اندازه نمونه و در محاسبات ما برابر با اندازه نقاط فضایی–زمانی است. این هزینه در هر مرحله از الگوریتم INLA باید تکرار شود. بدیهی است هرچه تعداد قیدها افزایش یابد، پیچیدگی و

هزینه محاسباتی بالاتر خواهد رفت. بنابراین، متأسفانه برای به‌کارگیری قیدها در INLA با مشکل مواجه هستیم. در حالت کلی دو راه برای دوری از این مشکل وجود دارد:

۱. پارامتربندی متفاوت

۲. پیاده کردن قیدها بر روی نمونه‌های توزیع پسین.

۲.۶ راه‌های دوری از قید مجموع مساوی با صفر در اثرات متقابل

در این بخش راه حل موجود و راه حل پیشنهادی خود را برای رفع این مشکل بیان می‌کنیم. راه حل اول استفاده از یک نوع پارامتربندی متفاوت است که اولین بار توسط کلیتون (۱۹۹۶) معرفی شد. راه حل دوم، در نظر گرفتن قید بعد از برازش مدل است. در زیر جزئیات بیشتری از این دو روش را تشریح می‌کنیم.

۱.۲.۶ پارامتربندی متفاوت

این روش معمولا در رگرسیون خطی زمانی که با متغیرهای عاملی روبرو باشیم، به‌کار می‌رود. در این روش یک سطح به‌عنوان پایه در نظر گرفته می‌شود و بقیه سطوح با این سطح پایه مقایسه و تفسیر می‌شوند. کلیتون (۱۹۹۶) آن را برای دوری از ذاتی بودن IGMRF پیشنهاد کرد.

فرض کنید $\boldsymbol{x}$ یک بردار تصادفی از یک IGMRF مرتبه اول با ماتریس دقت $\mathbf{Q}_{\boldsymbol{x}}$ و کمبود رتبه k است. به دنبال استفاده از پارامتربندی متفاوت، k عنصر از بردار $\boldsymbol{x}$ برابر صفر خواهند شد. فرض کنید I یک بردار به طول k برای شناسایی عناصری از $\boldsymbol{x}$ که صفر شدند، است. وقتی که عنصری از $\boldsymbol{x}$ صفر می‌شود، آن‌گاه ماتریس دقت توزیع شرطی عناصر باقیمانده تحت شرط $\boldsymbol{x}|\boldsymbol{x}I = ۰$ همان ماتریس دقت قبلی است بدون سطر I و ستون I. حال فرض کنید $\boldsymbol{x}^{(I)}$ یک بردار هم طول با بردار $\boldsymbol{x}$ باشد که تعداد I عنصر ورودی از آن برابر صفر است. ماتریس دقت عناصر باقیمانده برابر با ماتریس دقت $\boldsymbol{x}|\boldsymbol{x}I = ۰$ است. فضای صفر $\mathbf{Q}_{\boldsymbol{x}}$ یعنی $\mathbf{N}_{\boldsymbol{x}}$ را می‌توان به‌عنوان ترکیب خطی در محاسبه کلی قید به صورت $\boldsymbol{x}|\mathbf{N}_{\boldsymbol{x}}\boldsymbol{x} = ۰$ به‌کار گرفت. در ادامه یک کاربرد از این ایده را برای مدل قدم زدن تصادفی مرتبه اول نشان می‌دهیم.

پارامتربندی متفاوت در مدل قدم زدن تصادفی

مدل قدم زدن تصادفی مرتبه اول را با بردار $\boldsymbol{\nu} = (\nu_۱, \ldots, \nu_T)'$ و ماتریس دقت $\tau_\nu \mathbf{R}_\nu$ در نظر بگیرید. به جای استفاده از قید مجموع مساوی با صفر یعنی $\boldsymbol{\nu}'\mathbf{۱} = ۰$، یکی از عناصر $\boldsymbol{\nu}$، مثلا

ν_l را مساوی صفر قرار می‌دهیم. توزیع شرطی $\boldsymbol{\nu}|\nu_l = ۰$، یک توزیع گوسی است که ماتریس دقت آن همان ماتریس دقت فرآیند قدم زدن تصادفی مرتبه اول اصلی است که ردیف l و ستون l آن حذف شده باشد (کلیتون، ۱۹۹۶). پیامد این کار این است که مدل پارامتربندی‌شده $\boldsymbol{\nu}^{(l)} = (\nu_۱^{(l)}, \ldots, \nu_T^{(l)})'$، تفاوت‌های $\nu_۱ - \nu_l, \ldots, \nu_T - \nu_l$ را مدل‌بندی می‌کند. در حقیقت اثرات اصلی با ν_l به $\nu_۱^l, \ldots, \nu_{l-۱}^{(l)}, \nu_{l+۱}^{(l)}, \ldots, \nu_T^{(l)}$ منتقل می‌شوند، در حالی‌که ν_l خودش نیست. با این وجود، می‌توان اثرات اصلی ν_t را به‌وسیله اثرات منتقل‌شده $\nu_t^{(l)}$ به صورت

$$\nu_t = \nu_t^{(l)} - \frac{۱}{T}\sum_{t=۱}^{T} \nu_t^{(l)}$$

بازیابی کرد، که در واقع ترکیب خطی تعریف‌شده روی $\boldsymbol{\nu}^{(l)}$ است.

پارامتربندی متفاوت در اثرات متقابل

برای این‌که بتوانیم روش پارامتربندی متفاوت را برای اثرات متقابل فضایی‌ـ‌زمانی پیاده کنیم، باید یک نقطه زمانی l برای اثر متقابل نوع دوم و یک نقطه فضایی a برای اثر متقابل نوع سوم و همزمان یک نقطه زمانی l و یک نقطه فضایی a را برای اثر متقابل نوع چهارم به‌عنوان سطح پایه انتخاب کنیم. بنابراین ترکیبات خطی برای اثر متقابل نوع دوم همانند حالت اثر زمانی به صورت

$$\delta_{it} = \delta_{it}^{(l)} - \frac{۱}{T}\sum_{t=۱}^{T} \delta_{it}^{(l)},$$

بازنویسی می‌شود. اگر $t = l$ باشد، آن‌گاه $\delta_{it}^{(l)} = ۰$. به‌طور مشابه برای اثر نوع سوم می‌توان نوشت

$$\delta_{it} = \delta_{it}^{(a)} - \frac{۱}{n}\sum_{i=۱}^{n} \delta_{it}^{(a)}.$$

برای اثر نوع چهارم باید هر دو جهت فضا و زمان را در نظر بگیریم. بنابراین

$$\delta_{it} = \delta_{it}^{(a,l)} - \frac{۱}{n}\sum_{i=۱}^{n} \delta_{it}^{(a,l)} - \frac{۱}{T}\sum_{t=۱}^{T} \delta_{it}^{(a,l)} + \frac{۱}{nT}\sum_{i=۱}^{n}\sum_{t=۱}^{T} \delta_{it}^{(a,l)}.$$

توجه داشته باشید که روابط بالا از روی فضای صفر ماتریس دقت اثرات اصلی و بر اساس کلیتون (۱۹۹۶) به‌دست آمده‌اند.

متأسفانه به‌کارگیری این ایده در کار ما همچنان دست و بال ما را بسته نگه داشت، زیرا همچنان با قیدهای زیادی رو به رو هستیم اگر چه نوع ترکیب آن عوض شده باشد. افزون بر این، همه این قیدها در مدل چندجمله‌ای برای سه رده تکرار می‌شود. این یعنی اگر فردی داده‌ای به حجم مثلا ۱۰۰۰ داشته باشد، در عمل مدل داده‌ای به حجم ۳۰۰۰ یعنی سه برابر را برازش می‌دهد که همه این قیدها در آن تکرار می‌شود. این حجم با افزایش تعداد رده‌ها نیز

افزایش می‌یابد. چیزی که به قول دوست نروژی‌ام، ماریای عزیز، به کابوس شب شبیه است. لازم به ذکر است که اجرای آن با حجم داده شبیه‌سازی‌شده پایین نیز ما را به نتایج دلخواه نرساند!

۲.۲.۶ پیاده کردن قیدها بر روی نمونه‌های توزیع پسین

در تشریح این ایده فرض کنید $\boldsymbol{Y}$ دارای توزیع گوسی با ماتریس دقت $\mathbf{Q}$ است. همچنین فرض کنید $\boldsymbol{A}$ ماتریس تعریف قیدها است. قیدها می‌توانند برای رفع ذاتی بودن مدل یا رفع مشکل شناسایی‌پذیری پارامترهای مدل باشند. با فرض وارون‌پذیری $\mathbf{Q}$، مراحل زیر را انجام می‌دهیم:

۱. برازش مدل بدون قیدها و تولید نمونه از توزیع پسین.

۲. پیاده کردن قیدها بر روی نمونه‌های تولیدشده پسینی.

ابتدا این روش را بر روی یک مدل زمانی کلاسیک به‌کار می‌گیریم.

مثال ۱.۲.۶. فرض کنید یک مدل کلاسیک با معادله

$$y_i = \alpha + x_i + \epsilon_i,$$

داریم که در آن x از یک مدل قدم‌زدن تصادفی مرتبه اول پیروی می‌کند. مؤلفه ϵ_i نیز دارای توزیع نرمال با میانگین صفر و دقت متفاوت به ازای هر مشاهده است. برای دوری از ترکیب شدن x با عرض از مبدأ α باید قید $\sum_i x_i = ۰$ را در نظر بگیریم. از این مدل، نمونه‌ای به حجم ۴۰ تولید کردیم. دقت جمله ϵ_i را به صورت $\frac{۰/۱}{\sqrt{\exp(\mathrm{N}(n,۱,۳))}}$ در نظر گرفتیم. مقدار α را نیز برابر با یک گرفتیم. بنابراین ابتدا مدل را با در نظر گفتن کمبود رتبه در ماتریس دقت و پیشین‌های پیش‌فرض بسته `R-INLA` برازش دادیم و در خروجی از نمونه‌های پسین حاصل به اندازه ۱۰۰۰ نمونه جدا کردیم. سپس ماتریس $\boldsymbol{A}$ را گونه‌ای طراحی کردیم تا با پیاده کردن آن به قید دلخواه برسیم. خروجی حاصل تا چهار رقم اعشار برای چند نمونه اول در جدول ۱.۶ گزارش شده است. همان‌طور که مشهود است، مقادیر برآوردشده به مقادیر شبیه‌سازی‌شده خیلی نزدیک هستند. بنابراین، به مدل اصلی برمی‌گردیم و اجرای این روش را در آن پی می‌گیریم.

توجه داشته باشید که اهمیت به‌کارگیری این روش به‌خاطر اهمیت به‌کارگیری اثرات متقابل در تحلیل داده‌های چندجمله‌ای است که تاکنون افراد (حتی به کمک روش‌های MCMC) فقط موفق به استفاده از اثر نوع اول شده‌اند که آن هم ساختاری ندارد و از این رو مورد توجه نیست!

جدول ۱.۶: مقایسه میانگین و انحراف معیار برآوردشده با مقادیر شبیه‌سازی‌شده در مثال ۱.۲.۶.

میانگین شبیه‌سازی‌شده	میانگین برآوردشده	انحراف استاندارد شبیه‌سازی‌شده	انحراف استاندارد برآوردشده
−۱۶/۷۹۹۵	−۱۶/۷۹۹۷	۱/۴۹۷۵	۱/۵۱۵۸
−۱۴/۴۸۶۱	−۱۴/۴۸۶۲	۱/۴۹۷۶	۱/۵۱۵۸
۱۳/۷۱۱۷	−۱۳/۷۱۱۹	۱/۴۹۷۷	۱/۵۱۵۸
−۱۲/۶۵۶۳	−۱۲/۶۵۶۴	۱/۴۹۷۵	۱/۵۱۵۷
−۱۲/۵۸۵۸	−۱۲/۵۸۰۸	۱/۲۳۹۱	۱/۲۳۳۳
−۱۱/۴۰۳۰	−۱۱/۳۹۸۳	۱/۲۴۰۱	۱/۲۳۳۸
−۱۰/۵۶۴۳	−۱۰/۵۵۹۲	۱/۲۳۹۲	۱/۲۳۳۳
−۹/۵۵۶۹	−۹/۵۵۱۹	۱/۲۳۹۲	۱/۲۳۳۳
−۸/۶۵۷۴	−۸/۶۵۲۳	۱/۲۳۹۳	۱/۲۳۳۲
−۸/۵۷۰۸	−۸/۵۶۹۷	۱/۰۰۰۳	۰/۹۹۹۰
−۷/۵۲۲۵	−۷/۵۲۱۶	۰/۹۹۹۵	۰/۹۹۸۳
−۶/۶۴۵۶	−۶/۶۴۴۷	۰/۹۹۹۶	۰/۹۹۸۳
−۵/۵۳۳۷	−۵/۵۳۲۹	۰/۹۹۹۴	۰/۹۹۸۳
−۴/۴۵۲۳	−۴/۴۵۱۴	۰/۹۹۹۵	۰/۹۹۸۳
−۴/۴۲۶۳	−۴/۴۱۰۸	۰/۸۵۰۰	۰/۸۵۸۲
−۳/۵۶۶۳	۳/۵۵۱۴	۰/۸۵۰۸	۰/۸۵۸۷
−۲/۰۱۵۲	−۱/۹۹۹۹	۰/۸۴۹۹	۰/۸۵۸۰
−۰/۹۰۲۰	−۰/۸۸۶۳	۰/۸۵۰۱	۰/۸۵۸۲
۰/۰۰۴۹	۰/۰۲۰۳	۰/۸۴۹۹	۰/۸۵۸۰
۰/۱۰۰۷	۰/۱۰۱۲	۰/۸۵۴۴	۰/۸۵۰۱
۱/۰۵۲۹	۱/۰۵۳۴	۰/۸۵۴۵	۰/۸۵۰۱
۱/۹۷۶۶	۱/۹۷۷۱	۰/۸۵۴۵	۰/۸۵۰۱
۳/۲۲۶۷	۳/۲۲۷۲	۰/۸۵۵۱	۰/۸۵۰۶
۳/۸۳۱۲	۳/۸۳۱۹	۰/۸۵۴۶	۰/۸۵۰۲
۴/۷۳۵۷	۴/۷۳۳۴	۱/۰۱۶۲	۱/۰۲۳۳

۳.۶ مطالعه شبیه‌سازی

در مطالعه شبیه‌سازی ارایه‌شده در این بخش، داده‌ها را از مدل (۲.۴) با تابع درستنمایی (۱.۴) با در نظر گرفتن پیشگوی جمعی

$$\eta_{ijt} = x_{ijt}\beta_j + u_{ijt} + \nu_{ijt} + \delta_{ijt},\ i = ۱,\ldots,n,\ \ t = ۱,\ldots,T,\ \ j = ۱,\ldots,J, \qquad (۳.۶)$$

تولید کردیم، که در آن $T = ۵$، $n = ۱۵$ و $J = ۳$ انتخاب شدند. برای سبکی مدل و تمرکز بیشتر بر روی اثر متقابل، عرض از مبدأ را حذف کردیم. اثر زمان برای سه رده را به ترتیب زیر شبیه‌سازی کردیم.

$$t_۱ = \sin(۲\pi(۱:T)/۵۰),\ \ t_۲ = \log(۱:T),\ \ t_۳ = ((۱:T)/T)^۲,$$

و در فرآیند مدل‌بندی و برازش، اثر زمانی ν را به کمک یک فرآیند قدم زدن تصادفی مرتبه اول مدل کردیم. برای تولید اثر فضایی، u، نواحی $i = ۱, \ldots, n$ را بر روی یک مشبکه منظم ۱۰×۱۰ تعریف کردیم و تحقق‌های آن را از یک مدل بی‌سیگ به کمک تابع `qsample` در بسته `R-INLA` شبیه‌سازی کردیم. اثر متقابل δ را نیز از نوع چهارم شبیه‌سازی کردیم. برای محاسبه ماتریس دقت اثر متقابل از ضرب کرونه‌کر ماتریس دقت اثر زمانی و ماتریس دقت فضایی مدل بی‌سیگ، استفاده کردیم. متغیر تبیینی x برای همه رده‌ها از توزیع نرمال استاندارد شبیه‌سازی شد و ضرایب β_j نیز از یک توزیع نرمال استاندارد تولید و به گونه‌ای مقیاس‌بندی شدند که جمع آن‌ها برابر صفر شود.

لازم به ذکر است برای جلوگیری از ترکیب نشدن اثرها با هم از مقیاس‌بندی مدل (تعریف‌شده در فصل چهارم) بهره‌مند شدیم. در این‌جا مقیاس‌بندی پیشین‌های تعریف‌شده در مدل یا به‌عبارتی همان IGMRFها نه تنها به ما کمک می‌کند تا اثرها با هم ترکیب نشوند و برآوردهای حاصل از پسین برای ابرپارامترها قابل مقایسه باشند، بلکه این امکان را می‌دهد تا بتوانیم برای ابرپارامترها نیز توزیع پیشین در نظر بگیریم. این به آن علت است که پارامتر دقت IGMRF مقیاس‌بندی‌شده، دقت کناری است. بنابراین ریشه دوم معکوسش انحراف استاندارد کناری است که اندازه اثر را در مقیاس پیشگوی خطی اندازه‌گیری می‌کند. برازش مدل نیز بدون در نظر گرفتن قیدها اجرا شد.

برای برازش مدل از توزیع پیشین PC برای ابرپارامترهای مدل استفاده کردیم. ایده اصلی این توزیع پیشین بر پایه جریمه کردن اختلاف یک مدل (ساده) پایه و یک مدل با ساختار پیچیده‌تر است. در این ایده، پارامتر طوری در نظر گرفته می‌شود که انحراف از مدل پایه را اندازه‌گیری و جریمه می‌کند. برای اثر تصادفی، مدل پایه یک خط مستقیم است که واریانس صفر دارد. بنابراین واریانس (انحراف معیار) اختلاف مدل پایه از مدل پیچیده‌تر را اندازه‌گیری می‌کند.

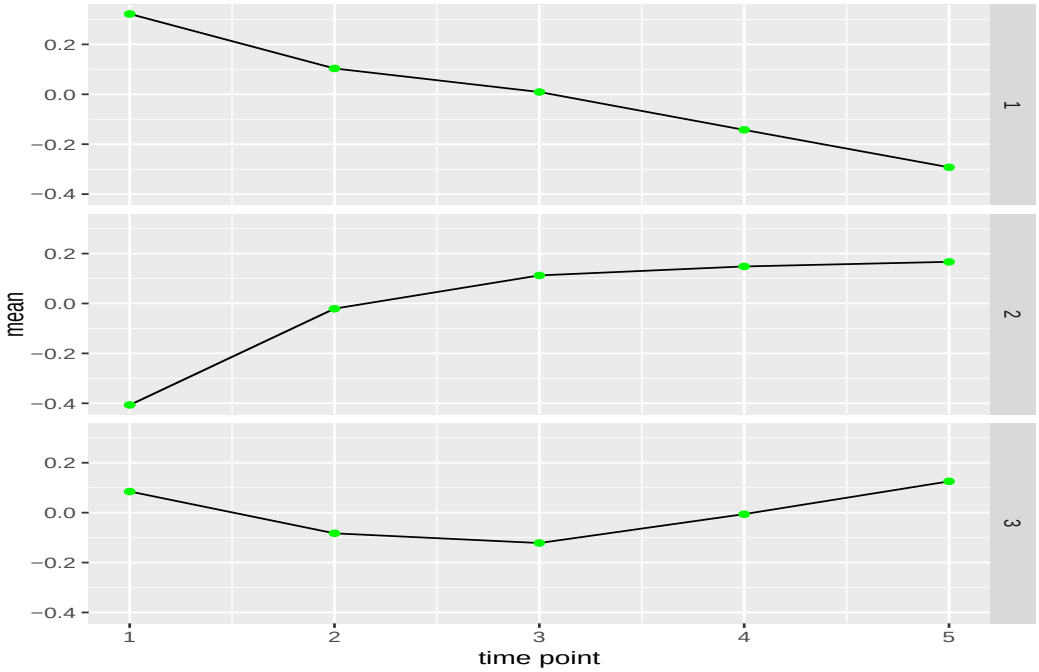
فرض کنید $\upsilon > ۰$ پارامتری باشد که مدل را کنترل می‌کند. هرچه υ بزرگتر باشد، مدل پیچیده‌تر است. ایده پیشین PC، تعریف توزیع‌هایی است که در آن‌ها $\mathrm{P}(\upsilon > q) = \alpha$ که در آن مقادیر q و α توسط کاربر به گونه‌ای تعیین می‌شوند که باور پیشینی او را بر توزیع پیشین القا کنند. در این مثال، به‌طور تجربی، مقادیر مختلف را بررسی کردیم و در نهایت $q = ۰/۵$ و $\alpha = ۰/۵$ را انتخاب کردیم. تفسیر این مقادیر به سادگی قابل درک هستند: به‌عنوان نمونه توزیع پیشین PC برای پارامتر دقت مدل بی‌سیگ با این مقادیر منتخب به این معنی است که انتظار داریم ۵۰% تغییرات با ساختار فضایی توضیح داده بشود.

نتایج حاصل از برازش مدل در جدول ۲.۶ و شکل‌های ۱.۶، ۲.۶ و ۳.۶ نمایش داده شده‌اند. با توجه به نتایج می‌توان گفت برآوردهای پسینی هم ضرایب رگرسیونی و هم اثرات فضایی، زمانی و متقابل آن‌ها از دقت قابل قبول برخوردارند. مطالعه شبیه‌سازی حاوی نتایج راضی‌کننده‌ای است. با این حال، نرسیدن به نتایج نظری تاییدکننده ما را مجبور کرد، تا زمان یافتن مبانی نظری برای اثبات ریاضی آن، از روش دیگری برای ادامه کار استفاده کنیم.

از آن‌جایی که با روش‌های دیگر در مدل فضایی–زمانی به نتیجه دلخواه نرسیدیم به ناچار، در ادامه، مدل را با قید رده گوشه برازش می‌دهیم. اگرچه که در این حالت برآورد مدل از دقت پایین‌تری برخوردار است، اما در حال حاضر تنها راه برازش مدل فضایی–زمانی چندجمله‌ای به شمار می‌رود.

جدول ۲.۶: برآورد میانگین (انحراف استاندارد) پسینی اثرات رگرسیونی ثابت.

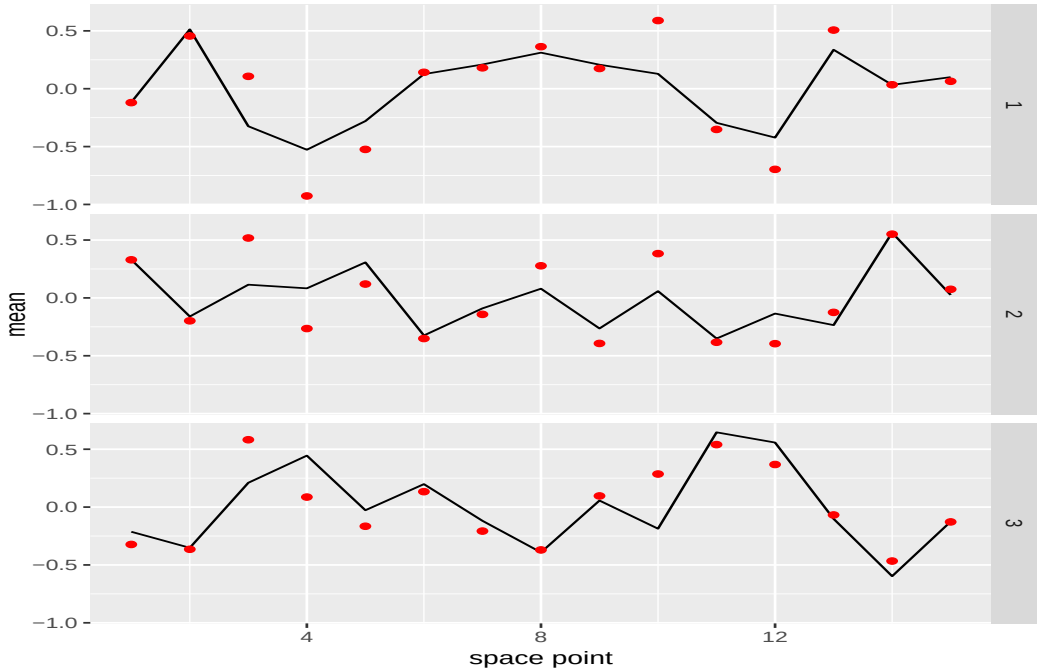
پارامتر	مقدار واقعی	برآورد
β_1	−۰/۱۷۳	−۰/۱۲۱ (۰/۰۰۶)
β_2	۱/۱۷۳	۱/۵۳۱ (۰/۰۰۳)
β_3	−۱/۰۰	−۱/۱۰۸ (۰/۰۰۹)



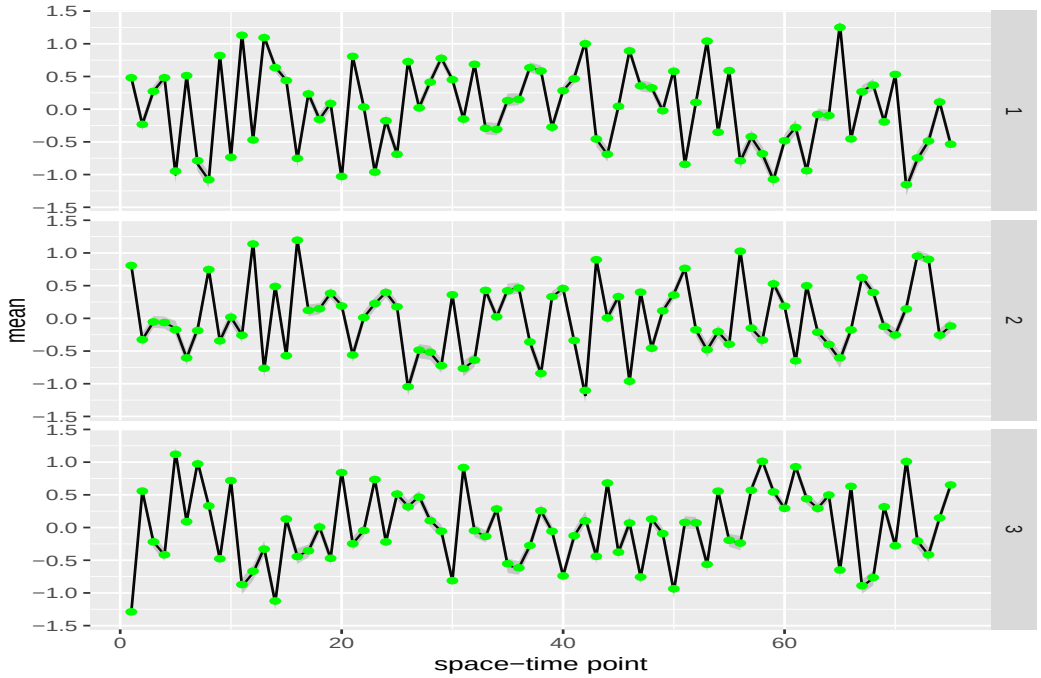
شکل ۱.۶: میانگین پسین اثر زمانی (دایره‌های سبزرنگ) به همراه اثر زمانی شبیه‌سازی‌شده واقعی (خط مشکی) برای هر سه رده.

۴.۶ برازش مدل با اثر فضایی–زمانی با قید رده گوشه

در این بخش مدل چندجمله‌ای جمعی–ساختاری با اثر فضایی–زمانی را با در نظر گرفتن قید رده گوشه برای شناسایی‌پذیری مطالعه می‌کنیم. اگرچه نتایج مربوط به همه اثرات متقابل خوب بودند، اما به دلیل محدودیت فضای رساله از آوردن همه نتایج خودداری کرده و فقط نتایج مربوط به اثر متقابل نوع چهارم را ذکر می‌کنیم که در باور بسیاری از کاربران علوم دیگر مهمتر و پرثمرتر جلوه کرده است.



شکل ۲.۶: میانگین پسین اثر فضایی (دایره‌های قرمزرنگ) به همراه اثر فضایی شبیه‌سازی‌شده واقعی (خط مشکی) برای هر سه رده.



شکل ۳.۶: میانگین پسین اثر متقابل فضایی–زمانی (دایره‌های سبزرنگ) به همراه اثر متقابل شبیه‌سازی‌شده واقعی (خط مشکی) برای هر سه رده.

۱.۴.۶ مطالعه شبیه‌سازی

در این مثال، داده‌ها را بر اساس پیشگوی

$$\eta_{ijt} = \alpha_j + x_{ijt}\beta_j + u_{ijt} + \nu_{ijt} + \delta_{ijt},\ i = ۱,\dots,n,\quad t = ۱,\dots,T,\quad j = ۱,\dots,J$$

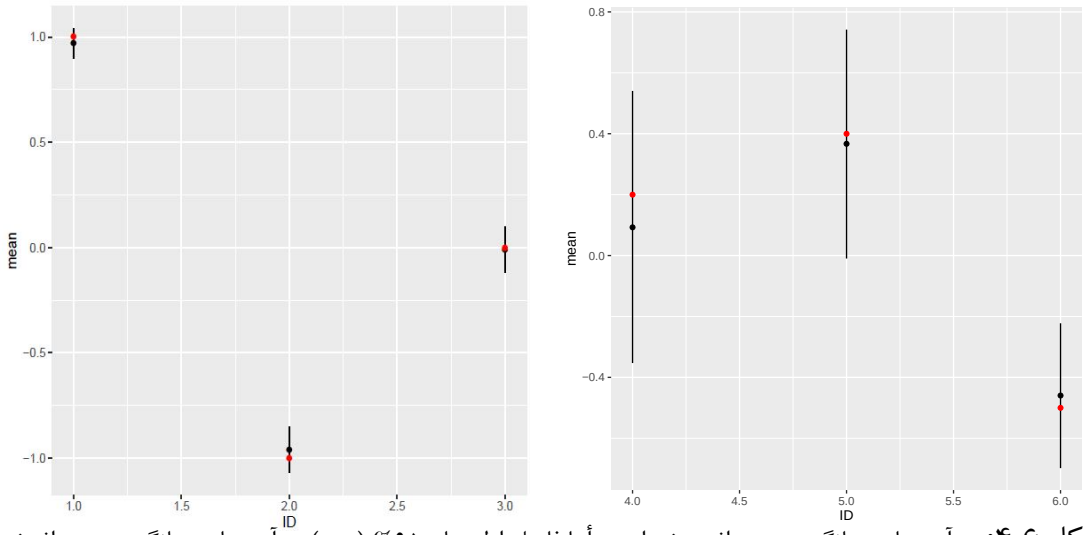
با $n = ۳۰$ ناحیه فضایی، $T = ۵$ نقطه زمانی و $J = ۳$ رده شبیه‌سازی کردیم. در این شبیه‌سازی، رده اول را به‌عنوان رده گوشه در نظر گرفتیم. بنابراین قرار دادیم

$$\alpha_۱ = ۰, \quad u_{i۱t} = ۰, \quad \delta_{i۱t} = ۰, \quad \nu_{i۱t} = ۰, \quad \beta_۱ = ۰.$$

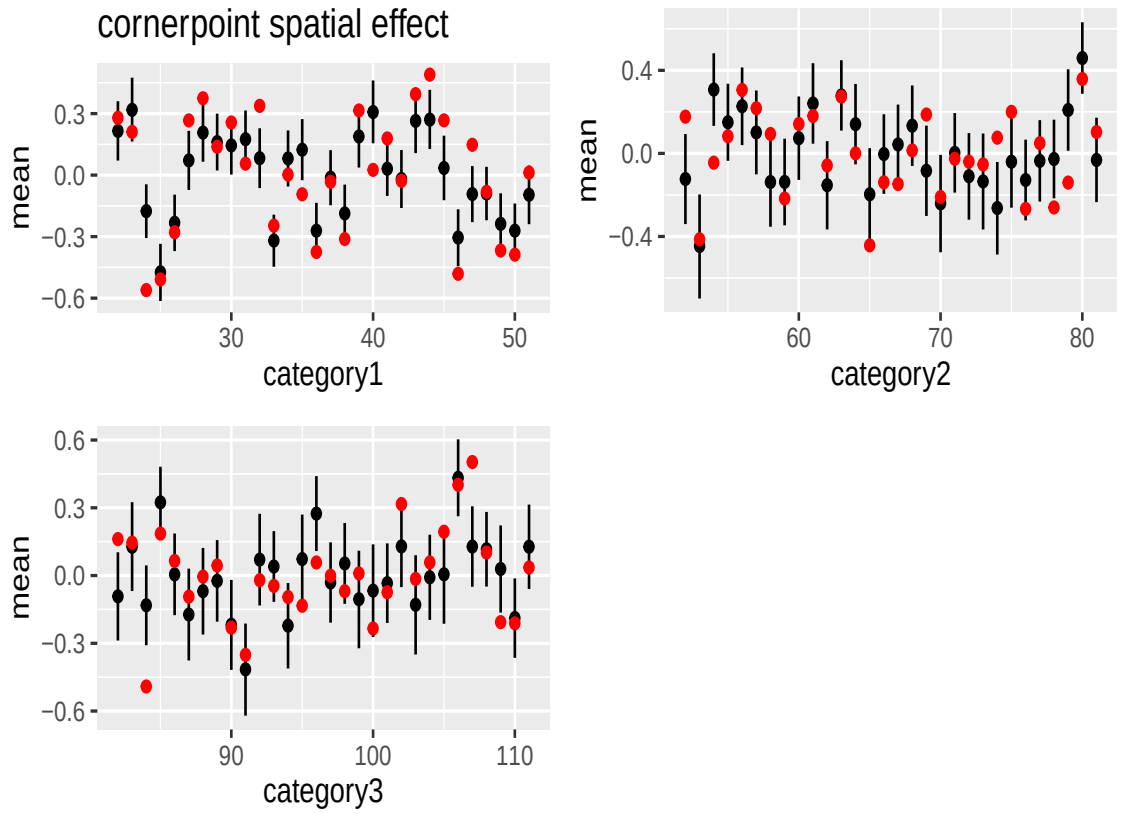
متغیر تبیینی و پارامترهای عرض از مبدأ و ضرایب رگرسیونی را از توزیع نرمال استاندارد تولید کردیم. مشابه با مثال شبیه‌سازی قبل از مقیاس‌بندی مدل و توزیع پیشین PC برای پارامترهای دقت اثرات پنهان استفاده کردیم. اثر زمانی را به کمک توابع

$$t_۱ = \sin(t/۰/۳), \quad t_۲ = \cos(t/۰/۶), \quad t_۳ = \cos(t/۰/۴),$$

تولید کردیم. سایر جزئیات مشابه مثال قبل است. نتایج برازش مدل برای پارامترهای رگرسیونی در شکل ۴.۶ و برآوردهای اثرات فضایی، زمانی و فضایی–زمانی به ترتیب در شکل‌های ۵.۶، ۶.۶ و ۷.۶ نمایش داده شده‌اند. در این شکل‌ها به ترتیب از چپ به راست مربوط به رده‌های اول تا سوم است. پسین کناری مربوط به دقت‌ها نیز در شکل ۱.۴.۶ و احتمال پسین برآوردشده هر رده برای ۱۰۰ نقطه فضایی در شکل ۹.۶ رسم شده‌اند. لازم به ذکر است قید $\sum_i u_{ij} = ۰$ نیز به‌کار گرفته شد تا مدل بی‌سیگ قابل شناسایی باشد. نتایج مربوط به برآوردهای هر سه نوع اثر فضایی، زمانی و فضایی–زمانی در اکثر نواحی، تفاوت نسبتاً کمی را بین مقادیر برآوردشده و مقدار واقعی نشان می‌دهد. برآورد میانگین پسینی مربوط به اثر زمان بهتر از برآورد مربوط به سایر اثرها است. احتمال‌های برآوردشده در هر رده، مقادیر نزدیک به مقادیر شبیه‌سازی‌شده را نشان می‌دهد. اما چگالی‌های کناری پسین پارامترهای دقت به‌غیر از اثر زمانی، نادقیق هستند و کم‌برآوردی را نشان می‌دهند.



شکل ۴.۶: برآوردهای میانگین پسین اثر عرض از مبدأ با فاصله اطمینان ۹۵% (چپ)، برآوردهای میانگین پسین اثر ضریب رگرسیونی با فاصله اطمینان ۹۵% (راست). برآوردها به رنگ قرمز نمایش داده شده‌اند و مقادیر واقعی به رنگ مشکی.



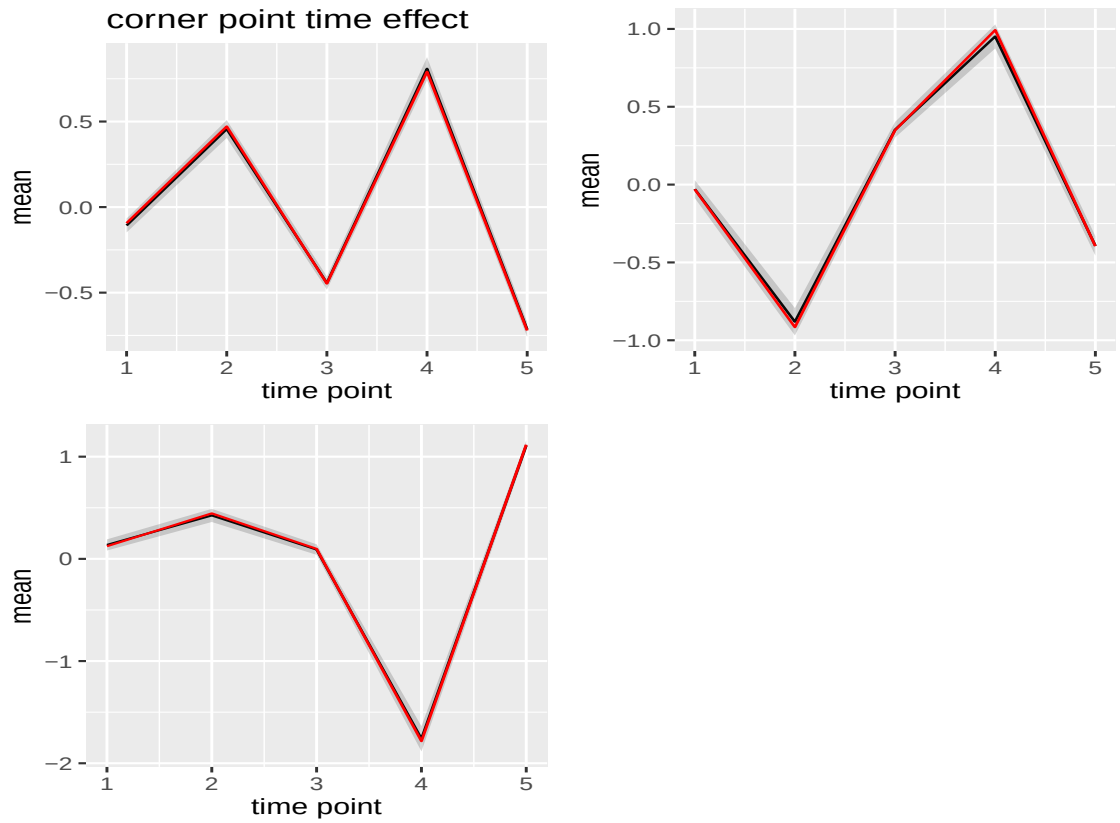
شکل ۵.۶: برآوردهای میانگین پسین اثر فضایی در هر رده با فاصله اطمینان ۹۵%. رده‌ها در شکل از سمت چپ به‌ترتیب اول تا سوم هستند. مقادیر واقعی به رنگ مشکی و برآوردها به رنگ قرمز هستند.

۵.۶ خلاصه فصل و نتیجه‌گیری

می‌دانیم با پیشرفت در فن‌آوری ثبت و ذخیره‌سازی داده‌ها و دسترسی به داده‌هایی که هم‌زمان در زمان و مکان دچار تحول می‌شوند و فرآیند تغییر را طی می‌کنند، بسیار آسانتر از گذشته است. همین امر موجب شده است تا افراد در پی کشف فرآیند موجود در داده‌ها باشند و مدل‌های مختلف و روش‌های آماری را به‌کار گیرند. در این میان داده‌های چندجمله‌ای که در ذات خود طبیعت هستند، از این امر مستثنی نیستند و توجه به آن‌ها بیشتر از قبل ضروری است. در این فصل مدل چندجمله‌ای جمعی–ساختاری با اثر فضایی–زمانی را مورد بررسی و مطالعه قرار دادیم. مشکلات پیش روی شناسایی‌پذیری و راه حل‌ها را بحث و ارزیابی کردیم. سرانجام قید رده گوشه را به‌عنوان توجیه‌پذیرترین راه حل موجود که پشتوانه نظری ریاضی دارد، پذیرفتیم و عملکرد قابل قبول آن را در یک مطالعه شبیه‌سازی نشان دادیم.

۶.۶ آینده پژوهش

با توجه به مطالعاتی که در طول سال‌های گذشته در باب شناسایی‌پذیری و مدل‌بندی داده‌های رسته‌ای به‌ویژه از نوع چندجمله‌ای انجام داده‌ایم و مشکلات موجود را شناسایی کردیم، در



شکل ۶.۶: برآوردهای میانگین پسین اثر زمانی در هر رده با فاصله اطمینان ۹۵٪. رده‌ها در شکل از سمت چپ به‌ترتیب اول تا سوم هستند. مقادیر واقعی به رنگ مشکی و برآوردها به رنگ قرمز هستند.

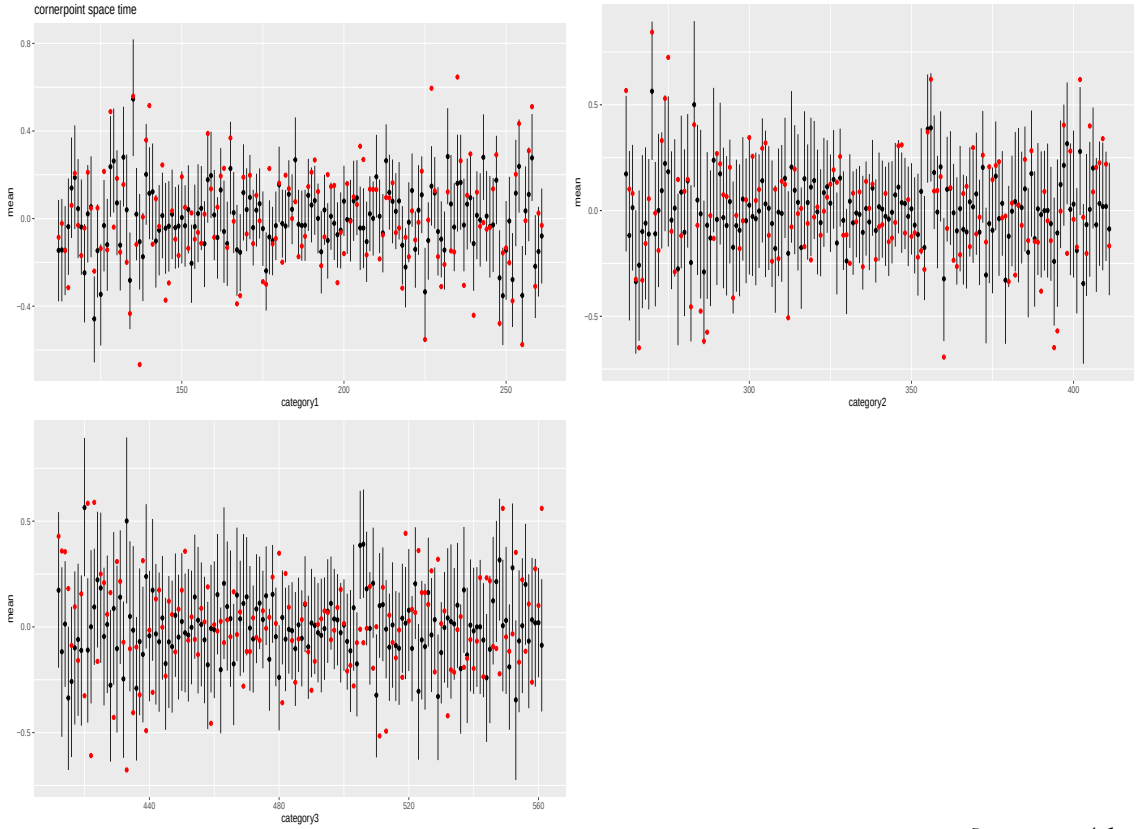
حالت کلی سه راه را برای ادامه کار به شرح زیر بررسی می‌توان کرد:

۱. **عوض کردن نوع قید** به‌جای بررسی قید به‌صورت مجموع مساوی با صفر از حالت کلی‌تر $A\mathbf{x} \leq b$ استفاده نماییم. نوع دیگر از قیدها وجود دارد که می‌تواند بر روی نمونه‌ها و احتمال‌ها تعریف شود. مثلا بررسی قید $n_i \geq (۲J - ۱)$ یعنی مجموع آزمایش‌ها برای هر فرد بشتر از دو برابر تفاضل تعداد کل رده‌ها از یک باشد.

۲. **عوض کردن رهیافت به‌کارگیری مدل** همان‌گونه که قبلا نیز گفته شد، INLA در به کارگیری قیدهای غیر خطی ناتوان است و در به‌کارگیری قیدها خطی توانایی محدودی دارد. بنابراین باید به دنبال روش دیگری باشیم که کارایی برابر با روش INLA داشته باشد و یا حتی از آن کاراتر هم باشد. این روش می‌تواند استنباط تغییراتی [۱] (بلی و همکاران، ۲۰۱۷) با توسعه بیزی و واردکردن اثر فضایی و فضایی–زمانی باشد که باید در ابتدا آن را به‌کار بگیریم و کارایی آن را با INLA مورد مقایسه قرار دهیم که قطعا پیاده‌کردن آن به سادگی گفتن آن نیست.

۳. **تغییر مدل** به جای استفاده از مدل‌های جمعی–ساختاری چندجمله‌ای از مدل‌های جمعی–ساختاری رگرسیون درختی بیزی استفاده کنیم که یک LGM است و برای

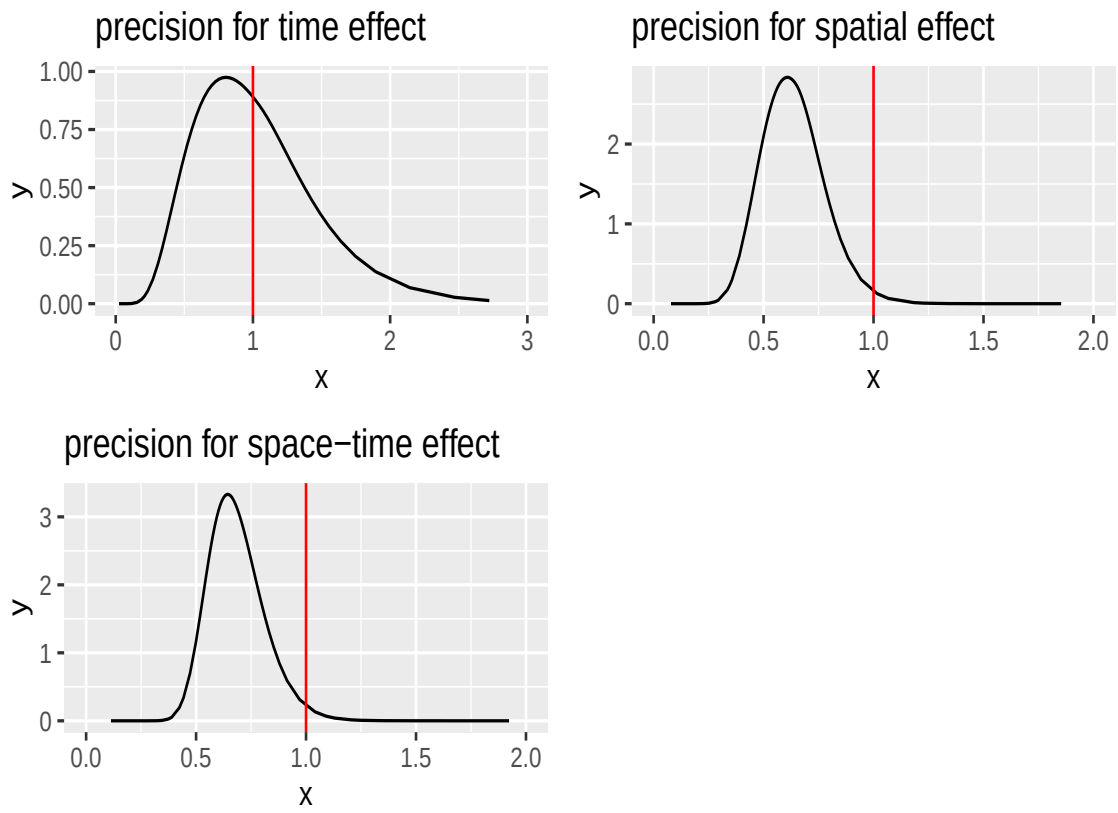
[۱] Variational inference



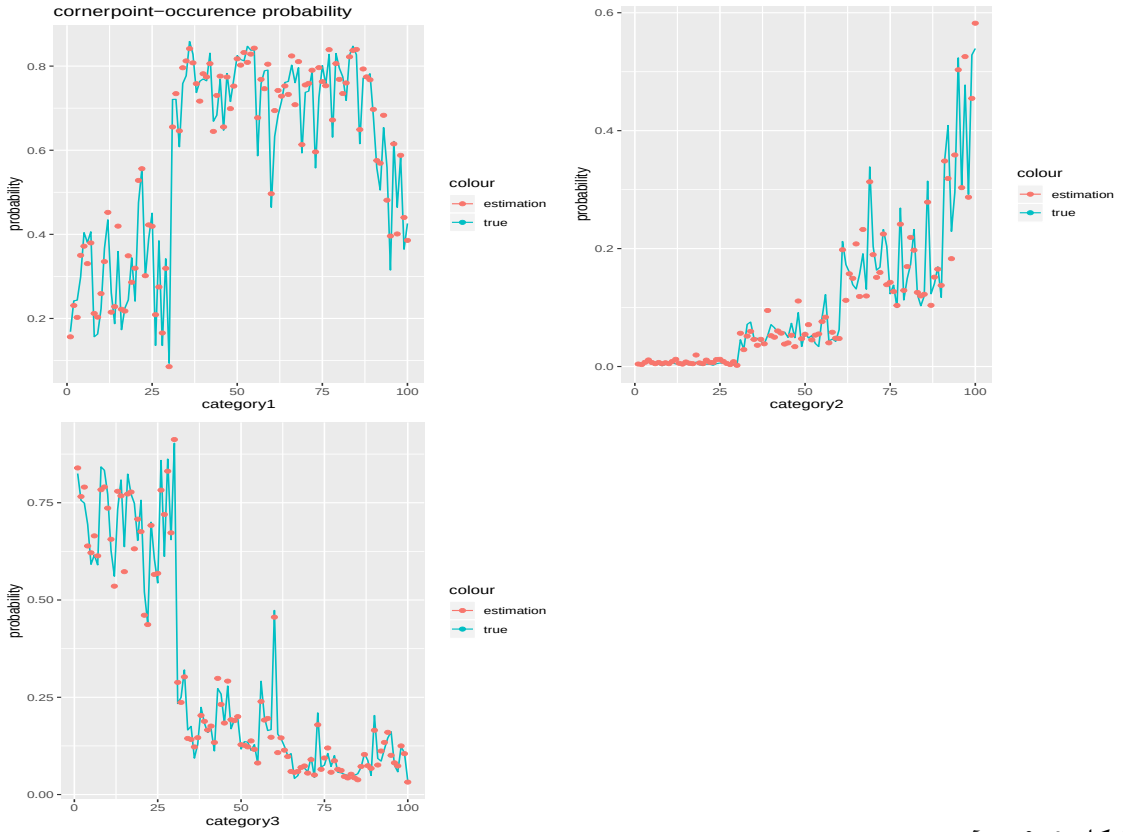
شکل ۷.۶: برآوردهای میانگین پسین اثر فضایی_زمانی در هر رده با فاصله اطمینان ۹۵%. رده‌ها در شکل از سمت چپ به‌ترتیب اول تا سوم هستند مقادیر واقعی به رنگ مشکی و برآوردها به رنگ قرمز هستند..

شناسایی نیاز به قید ندارد. بنابراین همچنان قادر خواهیم بود تا از روش کارای INLA استفاده کنیم که البته پیاده‌کردن آن به سادگی گفتن آن نیست و نیاز به زمان زیاد و پیداکردن روشی برای واردکردن وابستگی‌های خاص به مدل دارد.

روش‌های بیان شده در بالا که در طول تحقیقات خود در یک سال اخیر به آن‌ها دست‌یافته‌ایم را می‌توان با صرف مدت زمان بیشتر به کار گرفت و نتایج آن را در صورت قبولی ارائه داد تا این‌گونه کمکی به همه کاربران مدل چندجمله‌ای به‌خصوص در رشته‌های غیر آمار که تعداد آن‌ها کم نیست، کرده باشیم. از طرفی از مشکلات مربوط به قید رده گوشه نیز دوری کرده‌ایم.



شکل ۸.۶: برآوردهای چگالی‌های کناری دقت برای هر اثر.



شکل ۹.۶: برآورد احتمال پسین مربوط به هر رده برای ۱۰۰ نقطه فضایی_زمانی. رده‌ها در شکل از سمت چپ به‌ترتیب اول تا سوم هستند.

مراجع

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

[2] Abramowitz, M. and Stegun, I. A. (1972). "Bessel functions J and Y". in Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, 9th printing. **New York: Dover**. 358-364.

[3] Agresti, A. (1993). "Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters".**Scandinavian Journal of Statistic**. 20, 63-71.

[4] Agresti, A. (2019). "Introduction to Categorical Data Analysis". **Wiley, New Jersey**. USA.

[5] Agresti, A. (2002). "Categorical data analysis". second Edition. **New York: Wiley**. USA.

[6] Agresti, A. (2013). "Categorical data analysis". 3d Edition. **New York: Wiley**. USA.

[7] Alvarez, R.M. and Nagler, J., (1995). "Economics, issue and the Perot candidacy: voter choice in the 1992 presidential election". **American Journal of Political Science**. 39, 714–744.

[8] Alvarez, R.M. and Nagler, J., (1998). "When politics and models collide: estimating models of multiparty competition". **American Journal of Political Science**. 42, 55–96.

[9] Alvarez, R.M. and Nagler, J., (2001). "Correlated disturbances in discrete choice models: a comparison of multinomial probit and logit models". **Political Analysis**. Forthcoming. 914.

[10] Alvarez, R.M. and Nagler, J., Bowler, S., (2000). "Issues, economics, and the dynamics of multiparty elections: the British 1987 general election". **American Political Science Review**. 94, 131–149.

[11] Ambo, B.T., Ma, J. and Fu, C. (2020). "Investigating influence factors of traffic violation using multinomial logit method". **International Journal of Injuries Control Safety Promotion**. 28(1), 78-85.

[12] Bakka , H., Rue., H., Arne Fuglstad, G., Riebler., A., Bolin, D., Illian., J., Krainski., E., Simpson., D. and Lindgren., F. (2018). "Spatial modelling with R-INLA: A review". **computational statistics**. 1400-1443.

[13] Baker, S.G. (1994). "The multinomial-Poisson transformation". **Journal of Royal Stattistical Society**. Ser D (The Statistician). 43(4), 495-504.

[14] Bansal, P., Krueger, R., Bierlaire, M., Daziano, R.A. and Rashidi, T.H. (2020). "Bayesian estimation of mixed multinomial logit models: advances and simulation-based evaluations". **ransportation Research Part B: Methodological**. 131, 124–142.

[15] Barmi, H. E. and Dykstra. R. L. (1998). "Maximum likelihood estimates via duality for log convex models when cell probabilities are subject to convex constraints". **The Annals of Statistics**. 26, 1878-1893.

[16] Bartlett, M. S. (1937). "Some examples of statistical methods of research in agriculture and applied biology". **Journal of Royal Statistics Society Supplementary**. 4, 137183.

[17] Basse, G. and Bojinov, I. (2020). "A general theory of identification". **Harvard Business School**. 20-86.

[18] Basu, P.A. (1983). "Identifiability". **In Encyclopedia of Statistical Sciences** (S. Kotz and N. L. Johnson, eds.), Wiley Interscience. 4, 2–6.

[19] Bhattacharya, L., Li, L. and Victor Patrangenaru. (2016). **A Course in Mathematical Statistics and Large Sample Theory**. Springer-Verlag New York.

[20] Behnood, A. and Mannering, F.L. (2016). "An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models". **Analytic Methods in Accident Research**. 12, 1–17.

[21] Berman, S. M. (1963). "Note on extreme values, competing risks and semi-Markov processes". **The Annals of Mathematical Statistics**. 34, 1104-1106.

[22] Bernardinelli, L., Clayton, D. G., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). "Bayesian analysis of space-time variation in disease risk". **Statistics in Medicine**. 21–22(14), 2433–2443.

[23] Besag J. (1974). "Spatial interaction and the statistical analysis of lattice systems (with discussion)".**Journal of Royal Statistical Society**. Ser B. 36, 192–236.

[24] Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckho, P. A., Wenger, E. A., Briet, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Grin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., and Gething, P. W. (2015). "The effect of malaria control on plasmodium falciparum in Africa between 2000 and 2015". **Nature.** (526), 207-211.

[25] Bivand, R., Hauke, J., and Kossowski, T. (2013a). "Computing the Jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods". **Geographical Analysis**. 45(2), 150-179.

[26] Bivand, R. and Lewin-Koh, N. (2017). "maptools: Tools for Reading and Handling Spatial Objects". **R package version 0.9-2**.

[27] Bivand, R. and Piras, G. (2015). "Comparing implementations of estimation methods for spatial econometrics". **Journal of Statistical Software**. Articles, 63(18).

[28] Blangiardo, M. and Cameletti, M. (2015). Spatial and Spatio-temporal Bayesian Models with R-INLA. **John Wiley and Sons**.

[29] Blei, D.M., Kucukelbir, A. and McAuliffe, J.D. (2017). "Variational inference: a review for statisticians". **Journal of the American Statistical Association**. 112(518), 859–877.

[30] Bolin, D. and Lindgren, F. (2015). "Excursion and contour uncertainty regions for latent Gaussian models". **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**. 77(1), 85-106.

[31] Bolin, D. and Lindgren, F. (2017). "Quantifying the uncertainty of contour maps". **Journal of Computational and Graphical Statistics**. 26(3), 513-524.

[32] Bosco, C., Alegana, V., Bird, T., Pezzulo, C., Bengtsson, L., Sorichetta, A., Steele, J., Hornby, G., Ruktanonchai, C. and Ruktanonchai, N. (2017). "Exploring the high-resolution mapping of gender-disaggregated development indicators". **Journal of The Royal Society Interface**. 14(129), 2016-0825.

[8] Boudreau, S. A., Shackell, N. L., Carson, S., and den Heyer, C. E. (2017). "Connectivity, persistence, and loss of high abundance areas of a recovering marine fish population in the north west atlantic ocean". **Ecology and evolution**. 7(22), 9739-9749.

[33] Box, G. E. P. and Tiao, G. C. (1973). "Bayesian Inference in Statistical Analysis". **Addison-Wesley Publishing Company**. Reading, Mass.-London-Don Mills.

[34] Box, G., Jenkins, G., M. and Reinsel, G., C. (1994). "Time Series Analysis: Forecasting and Control". third edition. **Prentice-Hall**.

[35] Braitman, K.A., Kirley, B.B., Ferguson, S. and Chaudhary, N.K. (2007). "Factors leading to older drivers' intersection crashes". **Traffic Injuries Prevention**. 8(3), 267-274.

[36] Burgette. F.L., Puelz, D. and Hahn, R.P. (2020). "A symmetric prior for multinomial probit models". **Bayesian Analysis**. 1-18.

[37] Carson, J. and Mannering, F. (2001). "The effect of ice warning signs on ice-accident frequencies and severities". **Accident Analysis and Prevention**. 33(1), 99-109.

[38] Chen, Z. and Lynn, K. (2001). "Note on the Estimation of the Multinomial Logit Model with Random Effects". **Biometrica.**55(2), 89-95.

[39] Cressie, Noela. C., (1993). "Statistics for spatial data". **John wiley and Sons, INC**. New York.

[40] Clayton, D. G. (1996). "Markov chain Monte Carlo in practice". **Chapman and Hall**.

[41] Congdon, P. (2005). "Bayesian Models for Categorical Data". **Queen Mary, University of London**. UK.

[42] Conaway, M. R. (1989). "Analysis of repeated categorical measurements with conditional likelihood methods". **Journal of American Statistic Association**. 84, 53-62.

[43] Conaway, M. R.(1992). "The analysis of repeated categorical measurements subject to nonignorable nonresponse". **Journal of American Statistic Association**. 87, 817-824.

[44] Cowles, MK. and Carlin, BP. (1996). "Markov chain Monte Carlo convergence diagnostics: a comparative review". **Journal of American Statistic Association**. 91, 883–904.

[45] Cox, D. R. (1972). "Regression models and life-tables (with discussion)". **Journal of Royal Statistical Society**. B, 34, 187-220.

[46] Cressie, N. and Holland, P. (1983). "Characterizing the manifest probabilities of latent trait models". **Psychometrika**. 48, 129-141.

[47] Cressie, N. and Christopher K. Wikle. (2011). "Statistics for Spatio-Temporal Data". **John willey and son**. New Jercy.

[48] Cross, Manski, C. (2002). "Regression, short and long". **Econometrica**. 70.

[49] Czepiel, S. A. (2019). "Maximum likelihood estimation of logistic regression models: Theory and implementation".**https://czep.net/stat**.

[50] Daneil A.powers, and Y. Xie (1999). "Statistical methods for categorical data analysis". **Academic Press**.

[51] Darroch, J. N. (1981). "The Mantel-Haenszel test and tests of marginal symmetry, fixed-effects and mixed models for a categorical response". **International Statistics Review**. 49, 285-307.

[52] Depraetere, N. and Vandebroek, M. (2017). "A comparison of variational approximations for fast inference in mixed logit models". **Computational Statistics**. 32(1), 93–125.

[53] Dong, C., Richards, S.H., Huang, B. and Jiang, X. (2013). "Identifying the factors contributing to the severity of truck-involved crashes". **International Journal of Injeries Control Safety Promotion**. 22(2), 116-126.

[54] Dow, J. K. and Endersby, J. W. (2004). "Multinomial probit and multinomial logit: a comparison of choice models for voting research". **Electoral Studies**. 23(1), 107-222.

[55] Eluru, N., Bagheri, M., Miranda-Moreno, L.F. and Fu, L. (2012). "A latent class modeling approach for identifying vehicle driver injury severity factors at highway railway crossings". **Accident Analysis and Prevention**. 47, 119–127.

[56] Farmer, C.M., Braver, E.R. and Mitter, E.L. (1997). "Two-vehicle side impact crashes: the relationship of vehicle and crash characteristics to injury severity". **Accident Analysis and Prevention**. 29(3), 399-406.

[57] Fahrmeir L and Tutz G (2001). "Multivariate Statistical Modelling Based on Generalized Linear Models". **Springer-Verlag**. New York.

[58] Feng, S., Li, Z., Ci, Y. and Zhang, G. (2016). "Risk factors affecting fatal bus accident severity: their impact on different types of bus drivers". **Accident Analysis and Prevention**. 86, 29-39.

[59] Freni-Sterrantino, A., Ventrucci, M. and Rue, H. (2017). "Note on intrinsic conditional autoregressive models for disconnected graphs". **Spatial and Spatio-temporal Epidemiology**. 26, 25-34.

[60] Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2017). "Constructing priors that penalize the complexity of Gaussian random fields". **Journal of the American Statistical Association**. arXiv:1503.00256.

[61] Gakidou, E., Afshin, A., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulle, A. M., Abera, S. F. and Aboyans, V. (2017). "Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2016: a systematic analysis for the global burden of disease study (2016)". **Lancet**. 390(10100), 1345-1422.

[62] Ghosh M., Zhang L. and Mukherjee B. (2006). "Equivalence of posteriors in the Bayesian analysis of the multinomial Poisson transformation". **Metron-International Journal of Statistics**. 64, 19-28.

[63] Geedipally S., Turner P. and Patil S. (2011). "Analysis of motorcycle crashes in Texas with multinomial logit model". **Transportation Research Record: Journal of the Transportation Research Board**. 2265, 62–69.

[64] Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E. and Casey, D. C. (2017). "Mapping under-5 and neonatal mortality in Africa, 200015: a baseline analysis for the sustainable development goals". **The Lancet**. 390(10108), 2171-2182.

[65] Goodman, L. A. (1959). "Some alternatives to ecological correlation". **American Journal of Sociology**. 64 (6), 610-625.

[66] Guse, B., Kiesel, J., Pfannerstill, M. and Fohrer, N. (2020). "Assessing parameter identifiability for multiple performance criteria to constrain model parameters". **Hydrological Sciences Journal**. 2150-3435.

[67] Gustafson, P. (2004). "On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables". **Statistical Science**. 20(2), 111-140.

[68] Hall, D. B. and Prstgaard, J. T. (2001). "Order restricted score tests for homogeneity in generalised linear and nonlinear mixed models". **Biometrika**. 88, 739-751.

[69] Hao, W. and Daniel, J. (2014). "Motor vehicle driver injury severity study under various traffic control at highway-rail grade crossings in the United States". **Journal of Safety Research**. 51, 41–48.

[70] Hasti, T., J. and Tibshirani, R., J. (1990). Generalized additive models. **Chpman and Hall**.

[71] Hastings, W.K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". **Biometrika**. 57 (1), 97–109.

[72] Hausman J and McFadden D. (1984). "Specification tests for the multinomial logit model". **Econometrica**. 52, 1219–1240.

[73] Heaton, M. J., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M. and Lindgren, F. (2017). "Methods for analyzing large spatial data: A review and comparison". **arXiv preprint arXiv:1710.05013**.

[74] Held, L. and Rue, H. (2010). "Conditional and intrinsic autoregressions". Handbook of Spatial Statistics. **CRC/Chapman and Hall**. M., and Guttorp, P., editors, Raton, FL.

[75] Hsiao, C. (1983). "Identification". **Handbook of econometrics 1**. 223-283.

[76] Hosmer D., Jr., Lemeshow, S. and Sturdivant, R. (2013). "Applied Logistic Regression". **John Wiley Sons Inc**. Hoboken New Jersey.

[77] Hurwicz, L. (1950). "Generalization of the concept of identification". Statistical Inference in Dynamic Economic Models. **John Wiley and Sons, Inc**.

[78] Jacquez, J. A. and T. Perry (1990). "Parameter estimation: local identifiability of parameters". **American Journal of Physiology-Endocrinology and Metabolism**. 258 (4), E727-E736.

[79] Johnson, W. O., Gastwirth, J. L. and Pearson, L. M. (2001). "Screening without a gold standard: The Hui–Walter paradigm revisited". **American Journal of Epidemiology**. 153, 921-924.

[80] Jain, A., Mukhopadhyay, S., Tolety, S. and Bhattacharjee, T. (2013). "Multinomial logit models, Econometrics II term paper". **http://www.igidr.ac.in/faculty**.

[81] Jones, S., Gurupackiam, S. and Walsh, J. (2013). "Factors influencing the severity of crashes caused by motorcyclists: analysis of data from Alabama". **Journal of Transportation Engineering, Part B: Pavements**. 139, 949–956.

[82] Jousimo, J., Tack, A. J. M., Ovaskainen, O., Mononen, T., Susi, H., Tollenaere, C., and Laine, A.-L. (2014). "Ecological and evolutionary effects of fragmentation on infectious disease dynamics". **Science**. 344(6189), 1289-1293.

[83] Khattak, A. (2001). "Injury severity in multivehicle rear-end crashes". **Transportation Research Record: Journal of the Transportation Research Board**. 1746, 59–68.

[84] Knorr-Held, L. (2000). "Bayesian modelling of inseparable space-time variation in disease risk". **Statistics in Medicine**. 19(17-18), 2555-2567.

[85] Kockelman, K. and Kweon, J.Y. (2002). "Driver injury severity: an application of ordered probit models". **Accident Analysis and Prevention**. 34(3), 313-321.

[86] Koopmans, T. C. (1949). "Identification problems in economic model construction". **Econometrica, Journal of the Econometric Society**. 125-144.

[87] Koopmans, T. C. and O. Reiersol (1950). "The identification of structural characteristics". **The Annals of Mathematical Statistics**. 21 (2), 165-181.

[88] Krainski, E.T. (2018). **Statistical Analysis of Space-time Data: New Models and Applications**. PhD thesis, Norwegian University of Science and Technology.

[89] Kudo, A. (1963). "A multivariate analogue of the one-sided test". **Biometrika**. 50, 403-418.

[90] Kweon, Y.J. and Kockelman, K.M. (2003). "Overall injury risk to different drivers: combining exposure, frequency, and severity models". **Accident Analysis and Prevention**. 35(4), 441-450.

[91] Laplace, P S. (1774). "Mémoires de mathématique et de physique, Tome Sixième" Memoir on the probability of causes of events". **Statistical Science**. 1(3), 366-367.

[92] Lee, J. and Mannering, F. (2002). "Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis". **Accident Analysis Prevention**. 34, 149–161.

[93] Lee J., Yasmin S., Eluru N., Abdel-Aty M. and Cai, Q. (2018). "Analysis of crash proportion by vehicle type at traffic analysis zone level: a mixed fractional split multinomial logit modeling approach with spatial effects". **Accident Analysis Prevention**. 111, 12-22.

[94] Lee L.Y.J., Green J.P. and Ryan M.L. (2017). "On the "Poisson Trick" and its extensions for fitting multinomial regression models". **arXivpreprint arXiv:1706.09523**.

[95] Lewis, B. (2003). "Small Dams". CRC Press/Balkema.

[96] Lin. X. (1997). "Variance component testing in generalised linear models with random effects". **Biometrika**. 84, 309-326.

[97] Liu, J., Khattak, J.A., Li, X., Nie, Q. and Ling, Z. (2020). "Bicyclist injury severity in traffic crashes: a spatial approach for geo-referenced crash data to uncover non-stationary correlates". **Journal of Safety Research**. 73, 25-35.

[98] Luce, R. D. (1959). "Individual choice behaviour". **John Wiley**. New York.

[99] Maddala, G.S., (1983). "Limited-Dependent and Qualitative Variables in Econometrics". **Cambridge University Press**. New York.

[100] Martin, AD., Quinn. KM. and Park, JH. (2011). "MCMCpack: Markov Chain Monte Carlo in R". **Journal of Statistical Software**. 42(9), 22.

[101] Matzkin, R. L. (2007). "Nonparametric identification". **Handbook of Econometrics**. 6, 5307-5368.

[102] Manski, C. F. (2009). Identification for prediction and decision. **Harvard University Press**.

[103] Martinez-Beneito, M. A., López-Quilez, A., and Botella-Rocamora, P. (2008). "An autoregressive approach to spatio-temporal disease mapping". **Statistics in Medicine**. 27(10), 2874-2889.

[104] McCullagh, P. (1982). "Some applications of quasisymmetry". **Biometrika**. 69, 303-308.

[105] McCullagh, P., Nelder, J. (1989). "Generalized Linear Models". Second edition. **Chapman and Hall**.

[106] McFadden, D. (1981). "Econometric models of probabilistic choice". **Structure Analysis of Discrete Data with Econometric Applications**. MIT Press, Cambridge MA.

[107] Mervyn. J. S. (1994). "On tests against one-sided hypotheses in some generalized linear models". **Biometrics**. 50, 853-858.

[108] Miaou, S.P. (1994). "Relationship between truck accidents and geometric design of road sections: Poisson vs. negative binomial regressions". **Accident Analysis and prevention**. 26, 471-482.

[109] Minasny, B., Vrugt, J.A. and McBratney, A.B. (2011). "Confronting uncertainty in model-based geostatistics using Markov Chain Monte Carlo simulation". **Geoderma.** 163, 150-162.

[110] Milledge, D.G., Lane, S.N., Heathwaite, A.L. and Reaney, S.M. (2012). "A Monte Carlo approach to the inverse problem of diffuse pollution risk in agricultural catchments". **Science of the Total Environment**. 433, 434-449.

[111] Mohammed, A.A., Ambak, K., Mosa, A.M. and Syamsunur, D. (2019). "A review of the traffic accidents and related practices worldwide". **Open Transportation Journal**. 13, 65-83.

[112] Moraga, P., Cramb, S. M., Mengersen, K. L., and Pagano, M. (2017). "A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE". **Spatial Statistics**. 21, 27-41.

[113] Mullahy, J. and Robert, S. A. (2010). "No time to lose: Time constraints and physical activity in the production of health". **Review of the Economics of the Household**. 8, 409-432.

[114] Noor, A. M., Kinyoki, D. K., Mundia, C. W., Kabaria, C. W., Mutua, J. W., Alegana, V. A., Fall, I. S., and Snow, R. W. (2014). "The changing risk of Plasmodium falciparum malaria infection in Africa: 2000-10: a spatial and temporal analysis of transmission intensity". **Lancet**. 383(9930), 1739-1747.

[115] O'Donnell, C.J. and Connor, D.H. (1996). "Predicting the severity using models of ordered multiple choice". **Accident Analysis and Prevention**. 28(6), 739 - 753.

[116] Ormerod, J.T. and Wand, M.P. (2010). "Explaining variational approximations". **Journal of the American Statistical Association**. 64(2), 140-153.

[117] Pahlavan-Rad, M.R.; Khormali, F.; Toomanian, N.; Brungard, C.W.; Kiani, F.; Komaki, C.B. and Bogaert, P. (2016). "Legacy soil maps as a covariate in digital soil mapping: A case study from Northern Iran". **Geoderma**. 279, 141-148.

[118] Palmgren, J. (1981). "The Fisher information matrix for log linear models arguing conditionally on observed explanatory variables". **Biometrika**. 68, 563-566.

[119] Palmgren, J. and Ekholm, A. (1987). "Exponential family non-linear models for categorical data with errors of observation". **Applied Stochastic Models and Data Analysis**. 3, 111-124.

[120] Paulino, C. D. M. and C. A. de Braganca Pereira (1994). "On identifiability of parametric statistical models". **Journal of the Italian Statistical Society**. 3(1), 125-151.

[121] Penmetsa, P. and Pulugurtha, S.S. (2017). "Modeling crash injury severity by road feature to improve safety". **Traffic Injury Prevention**. 19, 102-109.

[122] Pettit, L.I. (1990). "The conditional predictive ordinate for the normal distribution". **Journal of the Royal Statistical Society**. Series B (Methodological). 52(1), 175-84.

[123] Pirdashti, H., Pirdashti, M., Mohammadi, M., Gharavi Baigi, M. and Movagharnejad, K. (2015). "Efficient use of energy through organic rice-duck mutualism system". **Agronomy for Sustainable Development**. 35, 1489-1497.

[124] Plummer, M. (2016). "rjags: Bayesian Graphical Models using MCMC". **R package version 4-6**.

[125] Porter, B.E. (2011). "Handbook of Traffic Psychology". **Old Dominion University Norfolk VA**. USA, Elsevier.

[126] Quinn, K.M., Martin, A.D. and Whitford, A.B., (1999). "Voter choice in multiparty democracies: a test of competing theories and models". **American Journal of Political Science**. 43, 1231-1247.

[127] R Core Team. (2021). "R: A language and environment for statistical computing, R Foundation for Statistical Computing". **Vienna, Austria. URL https://www.R-project.org/**.

[128] Regev, S., Rolison, J.J. and Moutari, S. (2018). "Crash risk by driver age, gender, and time of day using a new exposure methodology". **Journal of Safety Research**. 66, 131-140.

[129] Rezapour, M., Molan, M.A. and Ksaibati, K. (2019). "Application of multinomial regression model to identify parameters impacting traffic barrier crash severity". **The Open Transportation Journal**. 13, 57-64.

[130] Robert, C. P. and Casella, G. (1999). "Monte Carlo Statistical Methods". **Springer-Verlag**. New York.

[131] Rothenberg, T. J. (1971). "Identification in parametric models". **Econometrica: Journal of the Econometric Society.** 577-591.

[132] Rue, H., Held, L. (2005). "Gaussian Markov Random Fields: Theory and Applications, Monographs on Statistics and Applied Probability". **Chapman and Hall**. Boca Raton.

[133] Rue, H., Martino, S., Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". **Journal of Royal Statistical Society.** Series B (Statistical Methodology). 71, 319-392.

[134] Rue, H., Riebler, A., Sigrunn, H., Sørbye, J., Illian, B., Simpson, D.P. and Lindgren. F.K. (2017). "Bayesian computing with INLA: a review". **Annual Review of Statistics and Its Application**. 4, 395-421.

[135] Ruiz-Cárdenas, R., Krainski, E. T. and Rue, H. (2012). "Direct fitting of dynamic models using integrated nested Laplace approximations-INLA". **Computational Statistics and Data Analysis**. 56(6), 1808-1828.

[136] Rushworth, A., Lee, D. and Mitchell, R. (2014). "A spatio-temporal model for estimating the long- term effects of air pollution on respiratory hospital admissions in Greater London". **Spatial and Spatio-temporal Epidemiology**. 1(10), 29-38.

[137] Rom., M., t. and Cohen, A., (1995). "Estimation in the polytomous logistic regression model". **Journal of Statistical Planning and Inference**. 43, 341-353.

[138] Sadeghi-Bazargani, H., Sharifian, S., Khorasani-Zavareh, D., Zakeri, R., Sadigh, M., Golestani, M., Masoudifar, R., Rahmani, F., Mikaeeli, N., Namvaran, J., Pour-Ebrahim, K., Rezaei, M., Arabzadeh, B., Samadirad, B., Seyffarshad, A., Mirza-Mohammadi-Teimorloue, F., Kazemnezhad, S., Marin, S., Sheikhi, S. and Mohammadi, R. (2020). "Road safety data collection systems in Iran: a comparison based on relevant organizations". **Chinese Journal of Traumatology**. 23, 265-270.

[139] Said, F. (2020). **Constrained Statistical Inference for Categorical Data**. PhD thesis. Carleton University, Canada.

[140] Salum, J.H., Kitali, A.E., Bwire, H., Sando, T. and Alluri, P. (2019). "Severity of motorcycle crashes in Dar es Salaam, Tanzania".**Traffic Injuries Prevention**. 20(2), 189-195.

[141] Savolainen, P., Mannering, F. (2007). "Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes". **Accident Analysis and Prevention**. 39(5), 955-963.

[142] Savolainen, P.T., Mannering, F.L., Lord, D. and Quddus, M.A. (2011). "The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives". **Accident Analysis and Prevention**. 43(5), 1666-1676.

[143] Schott., J., R. (2016). "Matrix Analysis for statistics". third edition. **Wiley and Sons, Inc.** Hoboken, New Jersey.

[144] Schrodle, B., Held, L., and Rue, H. (2012). "Assessing the impact of a movement network on the spatio-temporal spread of infectious diseases". **Biometrics.** 68(3), 736-744.

[145] Seaman, S. R. and Richardson, S. (2004). "Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies". **Biometrika**. 91, 15-25.

[146] Serafini, F. (2018). "Multinomial logit models with INLA: vignettes for INLA". https://rdrr.io/github/inbo/INLA/f/vignettes/multinomial.Rmd.

[147] Shaddick, G., Thomas, M. L., Green, A., Brauer, M., Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Dingenen, R. V. and Dora, C. (2018). "Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution". **Journal of the Royal Statistical Society**, Series C (Applied Statistics), 67(1), 231-253.

[148] Shaheed, M., Gkritza, K., Zhang, W. and Hans, Z. (2013). "A mixed logit analysis of two-vehicle crash severities involving a motorcycle". **Accident Analysis and Prevention**. 61, 119-28.

[149] Simpson, D. P., Rue, H., Riebler, A., Martins, T. G., and Srbye, S. H. (2017). "Penalising model component complexity: A principled, practical approach to constructing priors (with discussion)". **Statistical Science.** 32(1), 1-28.

[150] Slessarev E. W., Lin Y., Bingham N. L., Johnson J. E., Dai Y. and Schimel J. P. (2016). "Water balance creates a threshold in soil pH at the global scale". **Nature.** 540(7634): 567.

[151] Sorbye, S. and Rue, H. (2014). "Scaling intrinsic Gaussian Markov random field priors in spatial modelling". **Spatial Statistics.** 8, 39-51.

[152] Soroori, E., Moghaddam, M., A. and Salehi, M. (2020). "Modelling spatial non-stationary and overdispersed crash data: development and comparative analysis of global and geographically weighted regression models applied to macrolevel injury crash data". **Journal of Transportation Safety and Security**. 13, 1000-1024.

[153] Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995). "BUGS: Bayesian inference using Gibbs sampling". **Version 0.50, MRC Biostatistics Unit, Cambridge**.

[154] Stan Development Team (2015). "Stan Modelling Language User's Guide and Reference Manual". **Version 2.6.1**.

[155] Sun, Y., Wang, Y., Yuan, K., Chan, O., T. and Huang, Y. (2020). "Discovering spatio-temporal clusters of road collisions using the method of fast Bayesian model-based cluster detection'.**Sustainability**. 12(20), 81-86.

[156] Swartz, T., Haitovsky, Y., vexler., A., yang, T. (2004). "Bayesian identifiability and misclassification in multinomial data". **The Canadian Journal of Statistics**. 32, 1-18.

[157] Theil, H. (1969). "A multinomial extension of the linear logit model". **International Economic Review**. 10, 251–259.

[158] Ugarte, M. D., Adin, A., Goicoa, T., and Militino, A. F. (2014). "On fitting spatio-temporal disease mapping models using approximate Bayesian inference". **Statistical methods in medical research**. 23(6), 507-530.

[159] Vivar, J. C. and Ferreira, M. A. R. (2009). "Spatio-temporal models for Gaussian areal data". **Journal of Computational and Graphical Statistics**. 18(3), 658-674.

[160] Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2018). "Estimating under five mortality in space and time in a developing world context". **Statistical Methods in Medical Research**. 28(9), 2614-2634.

[161] Washington, S., Karlaftis, M. and Mannering, F. (2003). "Statistical and Econometric Methods for Transportation Data Analysis". **Chapman and Hall/CRC.**

[162] Wedagama, D. and Dissanayake, D. (2020). "A model of latent class multinomial logit to investigate motorcycle accident injuries". **Engineering and Applied Science Research**. 47(4), 422-429.

[163] Whitehead, J. (1980). "Fitting Cox's regression model to survival data using GLIM". **Journal of Applied Statistics**. 29, 268-275.

[164] Wieland, F., Hauber, L., A., Rosenblatt, M., Tönsing, C. and Timmer, J. (2021). "On structural and practical identifiability". **Current Opinion in Systems Biology**. 25, 60-69.

[165] World Health Organization (2016). **Ambient air pollution: A global assessment of exposure and burden of disease**.

[166] Xie, Y., Zhang, Y. and Liang, F. (2009). "Crash injury severity analysis using Bayesian ordered probit models". **Journal of Transportation Engineering**. 135(1), 18-25.

[167] X. Lin. (1997). "Variance component testing in generalised linear models with random effects". **Biometrika**. 84, 309-326.

[168] Yamamoto, T. and Shankar, V. (2004). "Bivariate ordered response probit model of driver's and passenger's injury severities in collisions with fixed object". **Accident Analysis Prevention**. 36(5), 869-876.

[169] Zeng, Q., Wen, H. and Huang, H. (2016). "The interactive effect on injury severity of driver vehicle units in two-vehicle crashes". **Journal of Safety Research**. 59, 105-111.

[170] Zeng, Q., Gu, W., Zhang, X., Wen, H., Lee, J. and Hao, W. (2019). "Analyzing freeway crash severity using a Bayesian spatial generalized ordered logit model with conditional autoregressive priors". **National Library of Medicine**. 127, 87-95.

[171] Zhang, H. (2004). "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics".**Journal of the American Statistical Association**. 99(465), 250-261.

[172] Zhang Y. (2011). "Exploring single vehicle crash severity on Rural, two-lane highways with crash-level and occupant-level multinomial logit models, doctoral dissertation". **Salt Lake City, UT: Department of Civil and Environmental Engineering**. University of Utah.

[173] Ziakopoulos, A. and Yannis, G. (2020). "A review of spatial approaches in road safety". **Accident Analysis and Prevention**. 135, 105-323.

واژه‌نامه فارسی به انگلیسی

ابرپارامتر ..... Hyper parameter
اثر ..... Effect
اثر پنهان ..... Latent effect
اثرثابت ..... Fixe effect
اثرفضایی ..... Spatail effect
احتمال ..... Probability
احتمال تجمعی ..... Cumulative probability
استنباط ..... Inferance
اسمی ..... Nominal
استنباط بیزی ..... Bayesian Inferance
آزمایش ..... Trail
بیز ..... Nested
برآورد ..... Estimate
برآوردگر ..... Estimator
بردار تصادفی ..... Random vector
بخت‌ها ..... Odds
بعد بالا ..... High dimentional
بیز ..... Bayes
بیز تجربی ..... Associated prime ideal
پارامتر ..... Parameter
پیوسته ..... Continues
پیشگو ..... Predictor
پروبیت ..... Probit
پیشین ..... Prior
پوآسن ..... Poisson

تابع پیوند ..... Link function
تابع درستنمایی ..... Likelihood function
تبدیل ..... Transformation
تجمعی ..... Cumulative
تقریب ..... Approximation
تقریب ..... Laplace Approximation
تصادفی ..... Random
ترتیبی ..... Ordinal
تحلیل ..... Analysis
توزیع ..... Distribution
توزیع بتا ..... Beta Distribution
توزیع پیشین ..... Prior distribution
توزیع پسین ..... Posterior distribution
توزیع پسین حاشیه‌ای ..... Marginal posterior distribution
توزیع مانا ..... Stationary distribution
جدول احتمالی ..... Contegency table
چگالی ..... Density
چندجمله‌ای ..... Multinomial
چند رده‌ای ..... Multi-category
چند متغیره ..... Multi-variable
خانواده نمایی ..... Exponentail family
خطی ..... Linear
داده مشبکه‌ای ..... Lattice data
درستنمایی ..... Likelihood
دوجمله‌ای ..... Binomial
دودویی ..... Binary
رده‌ای ..... Categorical
طولی ..... Longitudinal
عرض از مبدا ..... Intercept
غیر خطی ..... Non-linear
گاما ..... Gamma
فضایی ..... Spatial

فضایی_زمانی ..... Spatio-temporal
فرم بسته ..... Closed form
کوواریانس ..... Covariance
لجیت تجمعی ..... Cumulative logit
لگاریتم درستنمایی ..... Log-likelihood
لگ_خطی ..... Log-linear
مارکوفی ..... Markovian
ماکزیمم ..... Maximum
ماکزیمم درستنمایی ..... Maximum likelihood
مدل احتمال خطی ..... Linear probability model
متروپولیس ..... Metropolis-Hasting
متغیر ..... variable
متغیر پاسخ ..... Response variable
متغیر پنهان ..... Latent variable
متغیر تصادفی ..... Random variable
متغیر توضیحی ..... explanatory variable
مدل پارامتری ..... Parametric model
مدل ناپارامتری ..... Non-Parametric model
مدل خطی تعمیم یافته ..... Generalized linear model
مدل سلسله مراتبی ..... Higerarical model
مستقل ..... Independance
مقدار حقیقی ..... Real value
میدان تصادفی گوسی مارکوفی ..... Gaussian markov random field
میدان تصادفی ..... Random field
نرمال ..... Normal
نسبت بخت ..... Odds ratio
نمونه ..... Sample
نمونه‌گیری ..... sampling

واژه‌نامه انگلیسی به فارسی

تحلیل ..... Analysis
تقریب ..... Approximation
بیز ..... Bayes
استنباط بیزی ..... Bayesian inferance
توزیع بتا ..... Beta distribution
دودویی ..... Binary
دوجمله‌ای ..... Binomial
رده‌ای ..... Categorical
جدول احتمالی ..... Contigency table
پیوسته ..... Continues
کوواریانس ..... Covariance
پیوند کانونی ..... Cononical link
فرم بسته ..... Closed form
تجمعی ..... Cumulative
لجیت تجمعی ..... Cumulative logit
احتمال تجمعی ..... Cumulative probability
چگالی ..... Density
توزیع ..... Distribution
بیز تجربی ..... Empirical bayes
برآورد ..... Estimate
برآوردگر ..... Estimator
متغیر توضیحی ..... Explanatory variable
تابع ..... Function
گاما ..... Gamma
میدان تصادفی مارکوفی گوسی ..... Gaussian markov random field

مدل خطی تعمیم‌یافته ..... Genralized linear model
نمونه‌گیری گیبس ..... Gibbs sampling
بعد بالا ..... High dimetional
مدل سلسله مراتبی ..... Hierarical model
ابر پارامتر ..... Hyper parameter
مشاهدات مستقل ..... Independance observations
عرض از مبدا ..... Intercept
اثر پنهان ..... Latent effect
داده‌های مشبکه‌ای ..... Lattice data
لگاریتم درستنمایی ..... Log-likelihood
لگ ـ خطی ..... Log-linear
طولی ..... Lognitudinal
تابع پیوند ..... Link function
پاسخ حاشیه‌ای ..... Miancearginal response
مدل حاشیه‌ای ..... Marginal model
چندجمله‌ای ..... Multinomial
مدل پارامتری ..... Parametric model
متغیر تصادفی ..... Random variable
پیشین ..... Prior
پسین ..... Posterior
نمونه ..... Sample
مونه‌گیرین ..... Sampling

Abstract

A categorical variable is a random variable that groups observations. For example, classifying people with different characteristics according to their blood type results in a categorical variable. Nowadays, the widespread of various sciences has made analyzing categorical data very essential. Hence, different scientific disciplines pursue more accurate methods for analyzing such data. In classical statistics, several methods have been proposed to analyze categorical data. These existing methods and relevant software packages are not efficient enough for analyzing data with complex spatial and spatio-temporal structures. Bayesian analysis and inference of categorical data require sampling methods such as the MCMC algorithms, while they are not also efficient for categorical data with complex structures. Our proposed approach, in this thesis, is to use an approximate Bayesian method, called integrated nested Laplace approximation (INLA), which is not accompanied by the significant problems seen in the MCMC algorithms for implementing the Bayesian analysis of spatial models. To deal with the inherent identifiability problem of the multinomial models, we consider two different types of identifiability constraints and compare their performances. We use the Poisson-multinomial transformation to develop the model in the class of latent Gaussian models applicable for INLA. We also utilize the individualized logit model as another type of modeling depending on the response variable type. Finally, we assess and compare the proposed models through both simulated and real data examples.

Keywords: Bayesian inference; fractional split multinomial model; identifiability; individualized logit; INLA method; multinomial-Poisson transformation; spatial multinomial model; spatio-temporal dependency structure.

Faculty of Mathematical Sciences

PhD Thesis in Statistics

Bayesian Categorical Data Analysis with Spatial and Spatio-temporal Structures

By: Leila Barmoudeh

Supervisor:

Dr. Hossein Baghishani

February 2022