دانشکده علوم ریاضی

گروه آمار

پایان‌نامه کارشناسی ارشد

مدل‌های فضایی-زمانی یک و چندمتغیره برای داده‌های زمین‌آماری حجیم

بهمن حمیدیان

استاد راهنما

دکتر حسین باغیشنی

بهمن ۱۳۹۴

تقدیم به:

پدرم با بوسه بر دستانش که وجودش مایه ی دلگرمی ام است.

و

مادرم با بوسه بر دستانش که وجودش برایم همه مهر است.

و

برادران عزیزم و خانواده هایشان که صفایشان مایه ی آرامشم است.

و

استاد راهنمای گرامیم، که برایم فراتر از یک استاد است و خواهد بود.

سپاس گزاری...

با سپاس فراوان از لطف خدای مهربان و سلام و صلوات بر پیامبر اکرم(ص).

سپاس خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت های او ندانند و کوشندگان، حق او را گزاردن نتوانند.

بدون شک جایگاه و منزلت معلم، اجّل از آن است که در مقام قدردانی از زحمات بی شائبه ی او، با زبان قاصر و دست ناتوان، چیزی بنگاریم. اما از آنجایی که تجلیل از معلم، سپاس از انسانی است که هدف و غایت آفرینش را تامین می کند و سلامت امانت هایی را که به دستش سپرده اند، تضمین؛ بر حسب وظیفه و از باب " من لم یشکر المنعم من المخلوقین لم یشکر اللّه عزّ و جلّ" :

از پدر عزیزم و مادر مهربانم این دو معلم بزرگوارم که همواره بر کوتاهی و درشتی من، قلم عفو کشیده و کریمانه از کنار غفلت هایم گذشته اند و در تمام عرصه های زندگی یار و یاوری بی چشم داشت برای من بوده اند؛

با تشکر از استاد بزرگوارم که شایسته ی هر نوع سپاس، تجلیل و تکریم اند؛ جناب آقای دکتر حسین باغیشنی؛ استاد راهنمای ارجمند که با ایجاد عشق به نوشتن، صبورانه، با ارائه رهنمودها، انتقادها و پیشنهادهایشان، در تمامی مراحل اجرای پایان نامه مرا حمایت و تشویق نمودند.

از استادان محترم؛ خانم دکتر اقبال، آقایان دکتر شاهسونی، دکتر نزاکتی، دکتر آرشی و دکتر ربیعی؛ که در طول دوران تحصیلی‌ام در دوره ی کارشناسی ارشد، جهت آموزش و ارتقای علمی بنده، زحمت کشیده اند سپاسگزارم.

تشکر و قدردانی فراوان خود را با بوسه بر دست‌های برادران عزیزم علی، محمود، احسان، وحید، بهرام و همسرانشان و همچنین برادرزاده هام محمد طاها، محمد یاسین و آرمین ابراز می‌نمایم.

از دوستان عزیزم، دانشجویان و همه ی خوبانی که با همکاری و همراهی شان این تحقیق به نتیجه رسید، تشکر می نمایم.

بهمن حمیدیان

بهمن ۱۳۹۴

تعهدنامه

اینجانب **بهمن حمیدیان** دانشجوی کارشناسی ارشد رشته **آمار** **دانشکده علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **مدل‌های فضایی-زمانی یک و چندمتغیره برای داده‌های زمین‌آماری حجیم** تحت راهنمایی دکتر **دکتر حسین باغیشنی** متعهد می‌شوم:

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

بهمن حمیدیان
بهمن ۱۳۹۴

مالکیت نتایج و حق نشر

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

چکیده

تحلیل داده‌های زمین‌آماری حجیم، معمولاً با محاسبات ماتریسی سنگین و هزینه‌بر مواجه است. این مساله باعث ایجاد مشکلات جدی در استنباط‌های بیزی مدل‌های سنتی شده است. علاوه بر این، محققین اغلب با مجموعه داده‌های فضایی چندمتغیره با ساختارهای وابستگی فضایی پیچیده روبرو می‌شوند که تحلیل آن‌ها بیش از پیش سخت خواهد بود. ساختار پیچیده (فضایی یا فضایی-زمانی) داده‌ها، حجم بالای آن‌ها و نقص مدل‌های موجود، کاربران آن‌ها را به سمت توسعه مدل‌های جدید و پویا سوق داده‌اند. در این پایان‌نامه به معرفی و استفاده از مدل‌هایی با عنوان مدل‌های بعد-پایین یا کم‌رتبه، به‌طور ویژه مدل فرآیند پیش‌گو، برای تحلیل داده‌های زمین‌آماری گاوسی حجیم می‌پردازیم که با کاهش فضای پارامتر موجب می‌شود تا نرخ همگرایی الگوریتم‌های نمونه‌گیری MCMC و در نتیجه سرعت محاسبات بهبود یابد. این بهبودی به دلیل پرهیز از محاسبات ماتریسی سنگین می‌باشد که باعث شده است بتوان از آن‌ها برای مجموعه داده‌های حجیم بهره گرفت. برای نمایش عملکرد این رده از مدل‌ها، داده‌های کیفیت آب منطقه وسیعی از استان گلستان را در بازه زمانی سال‌های ۱۳۸۲ تا ۱۳۹۲ مورد تحلیل قرار داده‌ایم.

واژه‌های کلیدی: استنباط بیزی، نمونه‌گیری MCMC، فرآیند پیش‌گو، فضایی-زمانی، زمین‌آمار

پیش گفتار

در مطالعات محیطی گاهی با داده‌هایی سر و کار داریم که مستقل از یکدیگر نیستند و به دلیل موقعیت مکانی داده‌ها در فضای (جغرافیایی) تحت مطالعه، نوعی وابستگی بین آن‌ها القا می‌شود. این‌گونه داده‌ها، داده‌های فضایی نامیده می‌شوند. از طرفی در مطالعات متعدد علمی نظیر هواشناسی، زمین‌شناسی و اقیانوس‌شناسی داده‌ها در طول زمان نیز به یکدیگر وابسته‌اند که در این صورت با داده‌های فضایی-زمانی سر و کار خواهیم داشت. تحلیل داده‌های فضایی-زمانی مستلزم تعیین دو نوع ساختار وابستگی هستند: وابستگی فضایی و وابستگی زمانی. مدل خطی کلاسیک که دارای پذیره اساسی استقلال متغیرهای پاسخ است، قادر به تحلیل این‌گونه داده‌ها نیست. بنابراین در این موارد از مدل‌های آمیخته خطی استفاده می‌شود که وابستگی داده‌ها با وارد شدن متغیرهای پنهان به مدل لحاظ می‌شود. استنباط مبتنی بر درستنمایی در این مدل‌ها، به دلیل وجود انتگرال‌های با بعد بالا برای محاسبه تابع درستنمایی، بسیار دشوار است. در مقابل، برازش این مدل‌ها از دیدگاه بیزی ساده‌تر است. این سادگی مرهون انقلاب روش‌های نمونه‌گیری MCMC است که در آن برای محاسبه توزیع پسین نه نیازی به حل انتگرال‌های با بعد بالاست و نه محاسبه مشتق‌های پیچیده. از طرفی، با توجه به ابزار مدرن جمع‌آوری داده‌ها، از جمله سیستم‌های سنجش از راه دور، این نوع داده‌ها به‌طور فزآینده‌ای حجیم و متنوع هستند. تحلیل داده‌های فضایی حجیم، معمولاً با محاسبات ماتریسی سنگین و هزینه‌بر مواجه است. این مساله باعث ایجاد مشکلات جدی در استنباط‌های بیزی مدل‌های سنتی شده است. اخیراً، مدل‌هایی با عنوان مدل‌های کم‌رتبه برای تحلیل داده‌های فضایی و فضایی-زمانی حجیم معرفی شده‌اند. در این پایان‌نامه، از یک زیررده از مدل‌های کم‌رتبه با عنوان مدل‌های فرآیند پیش‌گوی گاوسی، برای تحلیل داده‌های فضایی-زمانی حجیم گاوسی استفاده می‌کنیم. نتیجه استفاده از این مدل‌ها، کاهش فضای پارامتر است به‌طوری که نرخ همگرایی الگوریتم‌های نمونه‌گیری MCMC و در نتیجه سرعت محاسبات بهبود می‌یابند. با توجه به این مقدمه، ساختار این پایان‌نامه به‌صورت زیر است:

- در فصل اول، تعاریف و مفاهیم مورد نیاز برای ورود به موضوع پایان‌نامه را مطرح می‌کنیم.
- در فصل دوم، نحوه اجرای الگوریتم‌های MCMC و مدل‌های کم‌رتبه را معرفی خواهیم کرد.
- در فصل سوم، تحلیل مدل رگرسیون خطی فضایی یک و چندمتغیره، همراه با نسخه‌های تقریبی مدل‌های فرآیند پیش‌گو برای آن‌ها، مورد بررسی قرار می‌گیرد.
- در فصل چهارم، مدل‌های خطی فضایی-زمانی پویا همراه با مدل‌های فرآیند پیش‌گو مرتبط با آن را معرفی می‌کنیم و کاربرد مدل را با یک مثال واقعی در مورد کیفیت آب‌های زیرزمینی منطقه وسیعی از استان گلستان، نمایش می‌دهیم.
- در پیوست ب نیز کدهای نوشته‌شده در محیط R، برای مثال داده‌های اوزن، آورده شده‌اند.

فهرست مطالب

لیست تصاویر

لیست جداول

فصل ۱

مقدمه: تحلیل داده‌های فضایی و فضایی-زمانی

۱.۱ مقدمه

اغلب در مطالعات محیطی با داده‌هایی سر و کار داریم که مستقل از یکدیگر نیستند و وابستگی خاصی بین آن‌ها وجود دارد که ناشی از مکان (موقعیت جغرافیایی) آن در فضای تحت مطالعه است به این‌گونه داده‌ها، داده‌های فضایی[۱] می‌گویند. به‌عنوان مثال‌هایی از این نوع داده‌ها می‌توان میزان شیوع یک بیماری در استان‌های مختلف، درصد خلوص طلا در نقاط مختلف یک معدن و عوامل موثر بر کیفیت آب‌های زیرزمینی در نقاط مختلف یک استان را نام برد.

دو ویژگی عمده داده‌های فضایی، نمایش هر داده با موقعیت آن در فضای مورد مطالعه و وابستگی فضایی آن‌هاست. معمولاً داده‌های فضایی جمع‌آوری شده از موقعیت‌های مجاور، وابستگی بیشتری دارند و با افزایش فاصله بین موقعیت داده‌ها، وابستگی کاهش می‌یابد. آمار فضایی، شاخه نسبتاً جدیدی از آمار است که به تحلیل داده‌های فضایی می‌پردازد و در واقع بین مقادیر مختلف یک متغیر، فاصله و جهت قرارگیری آن‌ها ارتباطی برقرار می‌کند. این ارتباط فضایی، ساختار فضایی[۲] نامیده می‌شود. بررسی و مطالعه این ساختار و برآورد آن با استفاده از مشاهدات به‌دست آمده در بعضی از موقعیت‌های فضای مورد مطالعه، یکی از مسائل مهم آمار فضایی است. پیش‌گویی مقدار داده‌های فضایی در موقعیت‌های مشاهده‌نشده، به‌ویژه هنگامی که مشاهدات به دلایل مختلف از جمله خطای اندازه‌گیری نادقیق هستند، و تعیین خطای پیش‌گویی از موضوعات مهم دیگر در تحلیل داده‌های فضایی هستند.

اگرچه گسترش آمار فضایی عمدتاً توسط آماردانان صورت می‌گیرد، اما پیدایش و کاربرد عملی این شاخه از آمار به‌وسیله زمین‌شناسان و تحت عنوان زمین‌آمار[۳] صورت گرفت. نخستین بار به دنبال روند تکاملی ذخایر معدنی که از قبل از سال ۱۹۶۰ آغاز شده بود، ماترون با انتشار مقاله‌ای در سال ۱۹۶۲

---

[۱] Spatial data

[۲] Spatial structure

[۳] Geostatistics

پایه‌های زمین‌آمار را بنا نهاد. زمین‌آمار به شاخه‌ای از علم آمار گفته می‌شود که مبتنی بر تحلیل متغیرهایی است که دارای ساختار فضایی بوده و با داده‌ها یا متغیرهای فضایی سر و کار دارد. در این شاخه، ناحیه جغرافیایی یک ناحیه چگال[۴] است، به این معنی که بین هر دو موقعیت مکانی می‌توان یک مکان دیگر نیز در نظر گرفت.

در شکل‌گیری آمار فضایی افراد مختلفی سهیم بوده‌اند و به‌طور موازی با ماترون در سال ۱۹۶۲ فعالیت داشته‌اند. از جمله گاندین (۱۹۶۳) بحثی به نام تحلیل عینی[۵] را با مفاهیم بسیار نزدیک به تعاریف ماترون و دیگران در زمینه هواشناسی پیشنهاد کرد و از این طریق روش‌هایی برای تحلیل داده‌های فضایی ابداع نمود. دیوید (۱۹۷۷) از زمین‌آمار برای برآورد منابع ذخیره طلا استفاده کرد. یورنل و هیگبرتس (۱۹۷۸) به‌طور مفصل روش‌های زمین‌آماری را با اصطلاحات خاص خود شرح دادند. ریپلی (۱۹۸۱) در کتابی با عنوان آمار فضایی دریچه دیگری را گشود و با زبانی آماری سخن گفت. وی نظرات مربوط به پیش‌گویی فرآیندهای تصادفی را که ارتباط نزدیکی با سری‌های زمانی دارند، به کار گرفت. کرسی (۱۹۹۳) نیز با انتشار کتاب خود با عنوان "آمار برای داده‌های فضایی"، موجب پیشرفت گسترده این شاخه از علم آمار گردید.

داده‌های فضایی در مطالعات متعدد علمی مانند هواشناسی، محیط زیست، همه‌گیرشناسی، زمین‌شناسی و اقیانوس‌شناسی در طول زمان نیز به یکدیگر وابسته‌اند. مشاهداتی که هم از نظر موقعیت فضایی و هم از نظر موقعیت زمانی وابسته باشند، داده‌های فضایی-زمانی[۶] نامیده می‌شوند. تحلیل این‌گونه داده‌ها مستلزم تعیین دو نوع ساختار وابستگی هستند: وابستگی فضایی و وابستگی زمانی. به این ساختار وابستگی، ساختار وابستگی فضایی-زمانی[۷] گفته می‌شود و معمولاً از طریق تابع کوواریانس مدل‌بندی می‌شود. در سال‌های اخیر بررسی‌های زیادی برای تعیین این ساختار وابستگی، مدل‌سازی و تحلیل چنین داده‌هایی صورت پذیرفته‌اند. ساختار وابستگی این داده‌ها به وسیله کوواریانس فضایی-زمانی 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.

در سال‌های اخیر همراه با توسعه زمینه‌های مختلف آمار از جمله آمار بیزی، آمار فضایی نیز شاهد رشد و توسعه چشمگیری بوده است و به مدد این پیشرفت‌ها آمار فضایی به‌عنوان وسیله‌ای کارآمد برای مطالعه داده‌های فضایی و فضایی-زمانی مطرح گردیده است. نخستین بار کیتانیدیس (۱۹۸۶) با اشاره به مسائل و دشواری‌های موجود در روش‌های کلاسیک تحلیل داده‌های فضایی، به‌کارگیری روش‌های بیزی را پیشنهاد نمود. اما از آن‌جا که استفاده از رهیافت بیزی برای تحلیل داده‌های فضایی دشوار و نیازمند محاسبات پیچیده می‌باشد، در ابتدا چندان مورد توجه قرار نگرفت. اما با توسعه و گسترش آمار بیزی و ابداع روش‌های محاسبات بیزی پیشرفته نظیر روش‌های مونت کارلوی زنجیر مارکوفی[۸](MCMC)، تحلیل بیزی داده‌های فضایی نیز رشد و توسعه چشمگیری یافته است.

---

[۴] Dense

[۵] Objective analysis

[۶] Spatio-temporal data

[۷] Spatio-temporal structure

[۸] Markov chain Monte Carlo

گسترش داده‌های فضایی و فضایی-زمانی، موجب توسعه قابل توجهی در مدل‌سازی آماری شده است (کرسی، ۱۹۹۳؛ کرسی و ویکل، ۲۰۱۱). جامعه علمی در حال حرکت به دورانی است که دسترسی به محیط‌های غنی از داده‌ها فرصت‌های فوق‌العاده‌ای را فراهم می‌سازد تا به پیچیدگی‌ها و قابلیت‌های ساختارهای زمانی و مکانی فرآیندها در مقیاس گسترده دست یابیم. با توجه به ابزار مدرن جمع‌آوری داده‌ها، از جمله سیستم‌های سنجش از راه دور، این نوع داده‌ها به‌طور فزآینده‌ای حجیم و متنوع هستند. نتایج به‌دست آمده از تحلیل این داده‌ها، اغلب منجر به تصمیم‌گیری‌های مهمی در زمینه‌های مختلف اقتصادی، زیست‌محیطی، بهداشت و غیره می‌شوند. از طرفی، ساختارهای پیچیده و حجیم داده‌های حاصل از این فرآیندها، با مشکل انتخاب مدلی مناسب برای توصیف قابل قبول تغییرپذیری در سیستم‌های مورد علاقه محققین، مواجه هستند. اغلب مدل‌های موجود برای لحاظ کردن این پیچیدگی‌ها، ناقص هستند و توسعه چارچوب مدل‌هایی که قادر به در نظر گرفتن منابع مختلف عدم قطعیت باشد نیز کار ساده‌ای نیست.

همان‌طور که اشاره کردیم، یکی از شاخه‌های آمار فضایی که توسط زمین‌شناسان گسترش پیدا کرد، زمین‌آمار است. تحلیل داده‌های زمین‌آماری حجیم، معمولاً، با محاسبات ماتریسی سنگین و هزینه‌بر مواجه است. این مساله باعث ایجاد مشکلات جدی در استنباط‌های بیزی مدل‌های سنتی شده است. علاوه بر این، محققین اغلب با مجموعه داده‌های فضایی چندمتغیره با ساختارهای وابستگی فضایی پیچیده روبرو می‌شوند که تحلیل آن‌ها بیش از پیش سخت خواهد بود. ساختار پیچیده (فضایی یا فضایی-زمانی) داده‌ها، حجم بالای آن‌ها و نقص مدل‌های موجود، کاربران آن‌ها را به سمت توسعه مدل‌های جدید و پویا سوق داده‌اند. در حوزه آمار فضایی، مدل‌های سلسله‌مراتبی (کرسی و ویکل، ۲۰۱۱) کاربرد وسیعی در تحلیل داده‌های زمین‌آماری دارند. این مدل‌ها ساختار فضایی و زمانی داده‌ها را طی مراحل مختلف وارد مدل می‌کنند و قادر هستند عدم قطعیت این نوع ساختارها را به خوبی مدل‌بندی کنند. این مدل‌ها، به‌طور کلی، از چارچوب استنباط بیزی پیروی می‌کنند (گلمن و همکاران، ۲۰۱۳) و معمولا تحلیل آن‌ها بر اساس روش‌های نمونه‌گیری MCMC (رابرت و کسلا، ۲۰۰۴؛ بنرجی و همکاران، ۲۰۰۴)، صورت می‌پذیرد که موجب محبوبیت این مدل‌ها در طیف وسیعی از کاربردها شده است.

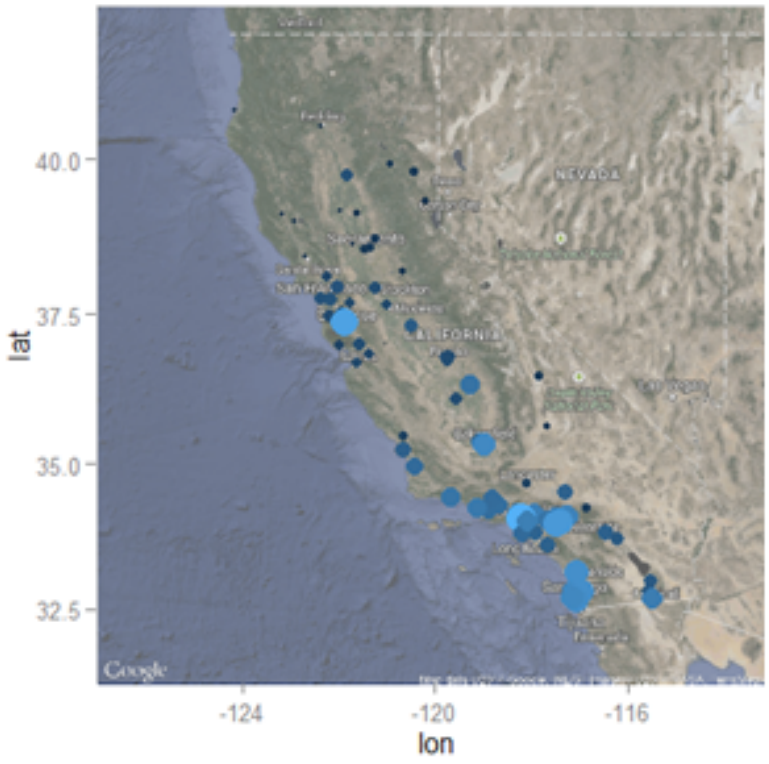
استفاده از مدل‌های سلسله‌مراتبی برای تحلیل داده‌های فضایی-زمانی منجر به افزایش کارایی محاسباتی، و انعطاف‌پذیری مدل‌های زمین‌آماری می‌شود. در این پایان‌نامه به معرفی و استفاده از مدل‌هایی با عنوان مدل‌های کم‌رتبه[۹]، برای تحلیل داده‌های زمین‌آماری گاوسی می‌پردازیم که با کاهش فضای پارامتر موجب می‌شود تا نرخ همگرایی الگوریتم‌های نمونه‌گیری MCMC و در نتیجه سرعت محاسبات بهبود یابد. این بهبودی به دلیل پرهیز از محاسبات ماتریسی سنگین می‌باشد که باعث شده است از آن‌ها برای مجموعه داده‌های حجیم نیز بتوان بهره گرفت. در ادامه، ابتدا به معرفی برخی از مفاهیم کلیدی آمار فضایی می‌پردازیم. بخش‌هایی از این فصل از کتاب دکتر محمدزاده (۱۳۸۹) با عنوان "آمار فضایی و کابردهای آن"، برداشت شده است.

[۹] Low-rank

۱.۱.۱ داده‌های فضایی

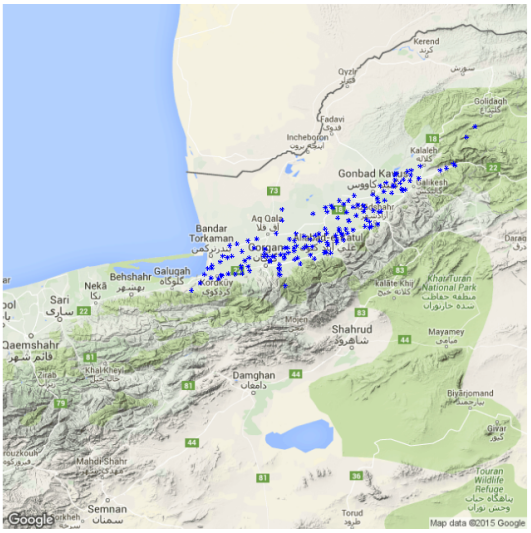
داده‌های فضایی به داده‌های وابسته‌ای گفته می‌شود که وابستگی آن‌ها ناشی از موقعیت و مکان قرار گرفتن آن‌ها در فضای (جغرافیایی) مورد مطالعه است. داده‌های فضایی بر اساس این که موقعیت‌های فضایی و متغیرهای تصادفی مرتبط به چه صورت باشند، به سه گروه داده‌های زمین‌آماری[۱۰]، داده‌های مشبکه‌ای[۱۱] و الگوهای نقطه‌ای[۱۲] تقسیم می‌شوند.

داده‌های زمین‌آماری: این نوع داده‌ها در موقعیت‌های ثابت و مشخص در ناحیه‌ای چگال مشاهده می‌شوند. یعنی بین دو موقعیت مفروض، امکان وجود موقعیت‌های دیگر هم هست و متغیر مورد مطالعه ممکن است پیوسته یا گسسته باشد. به عنوان مثال غلظت مواد معدنی در داخل یک معدن، تعداد نوعی جانور دریایی در یک مجموعه از مکان‌های نمونه‌گیری‌شده در طول یک ساحل (شکل ۱.۱) و کیفیت آب‌های زیرزمینی در ایستگاه‌های مختلف استان گلستان (شکل ۲.۱) را می‌توان نام برد. به‌طور کلی هدف اصلی در زمین‌آمار پیش‌گویی مقدار متغیر در یک مکان جدید مشاهده‌نشده می‌باشد.



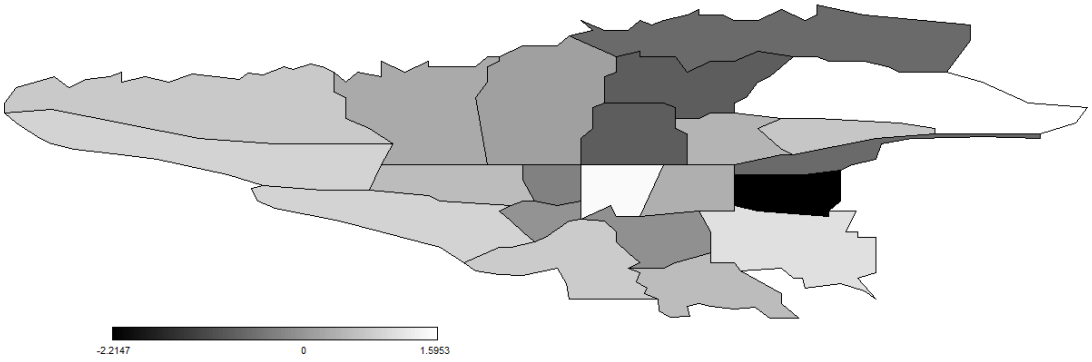
شکل ۱.۱: تعداد نوعی جانور دریایی در مکان‌های نمونه‌گیری‌شده در طول یک ساحل

[۱۰] Geostatistical data
[۱۱] Lattice data
[۱۲] Point patterns



شکل ۲.۱: پراکنش جغرافیایی ایستگاه‌های آب زیرزمینی استان گلستان

داده‌های مشبکه‌ای: این نوع داده‌ها مربوط به نواحی غیرچگال هستند، یعنی بین موقعیت‌های مشاهده‌شده امکان یافتن یک موقعیت جدید وجود ندارد. این مکان‌ها ممکن است منظم یا نامنظم باشند. تعداد حوادث رانندگی در مناطق ۲۲ گانه تهران (شکل ۳.۱)، تعداد اتباع خارجی غیرمجاز ساکن یا نرخ رشد اقتصادی هر استان، مثال‌هایی از این نوع داده‌هاست.



شکل ۳.۱: مناطق ۲۲ گانه تهران

الگوی نقطه‌ای: در این نوع داده‌ها موقعیت مشاهده‌شده، خود یک متغیر تصادفی است. این داده‌ها به سه دسته به‌طورکامل تصادفی فضایی[۱۳] (CSR)، منظم[۱۴] و خوشه‌ای[۱۵] تقسیم شده و اقدام به

[۱۳] Complete spatial randomness
[۱۴] Regular
[۱۵] Cluster

مدل‌بندی آن‌ها می‌شود. یک مثال برای این نوع مشاهدات، محل تخلفات رانندگی در یک مسابقه رالی (شکل ۴.۱) است.

شکل ۴.۱: محل تخلفات رانندگی در یک مسابقه رالی

۲.۱.۱ مدل‌های آماری فضایی: میدان تصادفی

تحلیل داده‌های فضایی نیازمند در نظر گرفتن یک مدل آماری است. معمولاً مدل‌بندی داده‌های فضایی با تعریف یک میدان تصادفی[۱۶] صورت می‌گیرد. بنابراین ابتدا میدان تصادفی و برخی از ویژگی‌های آن را که مورد نیاز ما است، تعریف می‌کنیم.

تعریف ۱.۱.۱ (میدان تصادفی). میدان تصادفی مجموعه‌ای از متغیرهای تصادفی مانند

$$Y(\cdot) = \{Y(\mathbf{s});\ \mathbf{s} \in D\}$$

است که در آن مجموعه اندیس‌گذار D یک زیرمجموعه از فضای اقلیدسی d بعدی، $d \geqslant ۱$، از R^d است.

تعریف ۲.۱.۱ (توابع میانگین، واریانس و کوواریانس میدان تصادفی). برای میدان تصادفی $Y(\cdot)$، میانگین در موقعیت $\mathbf{s}$ و کوواریانس در موقعیت‌های $\mathbf{s}_۱$ و $\mathbf{s}_۲$ به ترتیب به صورت

$$\begin{aligned} \mu(\mathbf{s}) &= E(Y(\mathbf{s})), \quad \mathbf{s} \in D \\ C(\mathbf{s}_۱, \mathbf{s}_۲) &= Cov(Y(\mathbf{s}_۱), Y(\mathbf{s}_۲)) = E\left[(Y(\mathbf{s}_۱) - \mu(\mathbf{s}_۱))(Y(\mathbf{s}_۲) - \mu(\mathbf{s}_۲))\right],\ \mathbf{s}_۱, \mathbf{s}_۱ \in D \end{aligned}$$

تعریف می‌شوند. برای $\mathbf{s}_۱ = \mathbf{s}_۲ = \mathbf{s}$، واریانس میدان تصادفی $Y(\cdot)$ در مکان $\mathbf{s}$ به صورت

$$Var\left[Y(\mathbf{s})\right] = E\left[(Y(\mathbf{s}_۱) - \mu(\mathbf{s}_۱)\right]^۲ = C(\mathbf{s}, \mathbf{s}) = C(\mathbf{s})$$

حاصل می‌شود.

هر میدان تصادفی را می‌توان به صورت

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \mathbf{s} \in D$$

[۱۶] Random field

تجزیه کرد، که در آن $\mu(\cdot)$ تغییرات بزرگ‌مقیاس[۱۷] یا روند[۱۸] و $\delta(\cdot)$ تغییرات کوچک‌مقیاس[۱۹] یا فرآیند خطا نامیده می‌شود. عبارت روند ممکن است ناشی از تغییرپذیری بین موقعیت‌های مشاهده‌شده باشد، اما فرآیند خطا ممکن است از خطای اندازه‌گیری یا تغییرپذیری در درون موقعیت مشاهده‌شده ناشی شود.

به‌طور معمول، تحلیل داده‌های فضایی بر اساس مشاهدات نمونه دشوار است، اما افزودن ویژگی‌هایی مانند مانایی[۲۰] و همسانگردی[۲۱] به ساختار همبستگی فضایی، مسأله را ساده‌تر می‌کند که در ادامه به معرفی آن‌ها می‌پردازیم.

تعریف ۳.۱.۱ (مانای ذاتی). میدان تصادفی $Y(\cdot)$ مانای ذاتی است، هرگاه

۱) $E\left[Y(\mathbf{s})\right] = \mu$

۲) $Var(Y(\mathbf{s}_1) - Y(\mathbf{s}_2)) = 2\gamma(\mathbf{s}_1 - \mathbf{s}_2),\ \mathbf{s}_1, \mathbf{s}_2 \in D \subset R^d.$

که در آن $\gamma(\cdot)$ تابعی مثبت است.

یعنی میانگین میدان تصادفی ثابت و مستقل از $\mathbf{s}$ و واریانس عبارت $(Y(\mathbf{s}_1) - Y(\mathbf{s}_2))$ فقط تابعی از فاصله موقعیت‌های $\mathbf{s}_1$ و $\mathbf{s}_2$ باشد.

تعریف ۴.۱.۱ (مانای مرتبه دوم). میدان تصادفی $Y(\cdot)$ مانای مرتبه دوم است، هرگاه

۱) $E\left[Y(\mathbf{s})\right] = \mu$

۲) $Cov(Y(\mathbf{s}_1), Y(\mathbf{s}_2)) = C(\mathbf{s}_1 - \mathbf{s}_2),\ \mathbf{s}_1, \mathbf{s}_2 \in D \subset R^d.$

یعنی میانگین میدان تصادفی ثابت و مستقل از $\mathbf{s}$ و کوواریانس عبارت $(Y(\mathbf{s}_1), Y(\mathbf{s}_2))$ فقط تابعی از فاصله موقعیت‌های $\mathbf{s}_1$ و $\mathbf{s}_2$ باشد.

تعریف ۵.۱.۱ (مانای قوی). میدان تصادفی $Y(\cdot)$ مانای قوی است، هرگاه به ازای هر $\mathbf{h} \in R^d$،

$$(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)) \overset{D}{=} (Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h}))\,.$$

به بیان دیگر در صورت انتقال موقعیت‌های $\mathbf{s}_1, \ldots, \mathbf{s}_n$ در راستای $\mathbf{h}$، توزیع توأم $(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))$ تغییر نمی‌کند.

تعریف ۶.۱.۱ (میدان تصادفی گاوسی). میدان تصادفی $Y(\cdot)$، گاوسی[۲۲] نامیده می‌شود، هرگاه برای هر $k \geq 1$ توزیع توأم $(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_k))$، نرمال چندمتغیره باشد.

قبل از ورود به بخش بعدی، با توجه به استفاده از نرم اقلیدسی در ادامه، ابتدا این تابع را تعریف می‌کنیم.

---

[۱۷] Large scale variation

[۱۸] Trend

[۱۹] Small scale variation

[۲۰] Stationary

[۲۱] Isotropic

[۲۲] Gaussian

تعریف ۷.۱.۱ (تابع نرم اقلیدسی). فرض کنید $1 \leqslant p < \infty$ باشد، تابع p–نرم برای بردار n بعدی $\mathbf{x}$ از مقادیر حقیقی به صورت زیر تعریف می‌شود:

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

اگر مقدار p برابر ۲ باشد، تابع نرم حاصل را اقلیدسی می‌نامند.

۳.۱.۱ ساختار وابستگی فضایی

در آمار فضایی، کمیت‌هایی مشابه واریانس و کوواریانس در آمار کلاسیک برای بیان ساختار وابستگی فضایی داده‌ها تعریف شده‌اند. یکی از این کمیت‌ها تغییرنگار[۲۳] است که به صورت زیر تعریف می‌شود.

تعریف ۸.۱.۱ (تغییرنگار). واریانس تفاضل بین مقادیر میدان تصادفی در دو موقعیت $\mathbf{s} \in R^d$ و $\mathbf{s}+\mathbf{h} \in R^d$ را تغییرنگار می‌نامند که به صورت

$$2\gamma(\mathbf{h}) = Var\left(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})\right)$$

تعریف می‌شود. تابع $\gamma(\mathbf{h})$ نیز نیم‌تغییرنگار[۲۴] نامیده می‌شود.

کوچک بودن تغییرنگار، نشانگر وابستگی زیاد و بزرگ بودن آن، نشانگر وابستگی کم میدان تصادفی است. تغییرنگار را می‌توان به‌عنوان معیاری برای نشان دادن تأثیرگذاری کمیت یک موقعیت بر کمیت موقعیت‌های دیگر در فضای مورد مطالعه به کار برد. یک حالت نمودار تابع تغییرنگار به‌عنوان تابعی از فاصله بین موقعیت‌ها در شکل ۵.۱ نشان داده شده است. اگر تغییرنگار برای تمام مقادیر $\mathbf{h}$ تقریباً ثابت باشد، داده‌ها همبستگی فضایی ندارند یا همبستگی فضایی آن‌ها بسیار ضعیف است. در واقع افزایش شیب نمودار تغییرنگار، بیانگر افزایش همبستگی فضایی داده‌هاست. تغییرنگار، سه پارامتر دامنه[۲۵]، ازاره[۲۶] و اثر قطعه‌ای[۲۷] دارد که به‌صورت زیر تعریف می‌شوند.

تعریف ۹.۱.۱ (دامنه). بازه‌ای که تغییرنگار در خارج آن، ثابت و به صورت افقی باشد.

تعریف ۱۰.۱.۱ (ازاره). در صورتی که تغییرنگار به کران بالایی منتهی شود، آن کران، ازاره تغییرنگار است و عبارت است از

$$\lim_{||\mathbf{h}|| \to \infty} \gamma(\mathbf{h}) = c$$

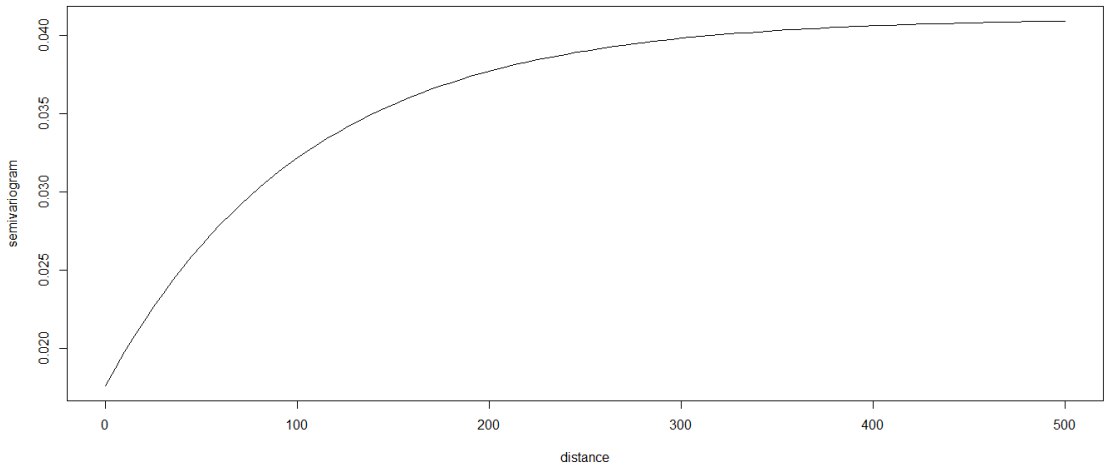
تعریف ۱۱.۱.۱ (اثر قطعه‌ای). مقدار تغییرنگار در مبدأ مختصات است که به صورت

$$\lim_{||\mathbf{h}|| \to 0} \gamma(\mathbf{h}) = c_0$$

تعریف می‌شود.

---

[۲۳] Variogram
[۲۴] Semi-variogram
[۲۵] Range
[۲۶] Sill
[۲۷] Nugget effect



شکل ۱.۵: یکی از حالت‌های تابع تغییرنگار

یکی از ویژگی‌های تغییرنگار، خاصیت همیشه منفی شرطی[۲۸] آن است، به این معنی که برای هر میدان تصادفی ذاتی و دنباله اعداد ثابت $\{a_i\}_{i=1}^m$ که $\sum_{i=1}^m a_i = 0$، رابطه $\sum_{i=1}^m \sum_{j=1}^m a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$ برقرار است.

تعریف ۱.۱.۱۲ (همبستگی‌نگار). تابع همبستگی‌نگار با شرط $C(0) > 0$ به صورت زیر تعریف می‌شود:

$$\rho(\mathbf{s}, \mathbf{s}+\mathbf{h}) = \frac{C(\mathbf{h})}{C(0)} = 1 - \frac{\gamma(\mathbf{h})}{C(0)}.$$

در حالت کلی توابع برداری تغییرنگار، کوواریانس و همبستگی‌نگار به اندازه و جهت بردار $\mathbf{h} \in R^d$ بستگی دارند، اما اگر این سه فقط تابعی از $||\mathbf{h}||$، یعنی نرم $\mathbf{h}$، در فضای اقلیدسی R^d باشند و به جهت $\mathbf{h}$ بستگی نداشته باشند، آن‌ها به‌ترتیب تغییرنگار، کوواریانس و همبستگی‌نگار همسانگرد هستند و در غیر این صورت ناهمسانگرد نام‌گذاری می‌شوند.

تعریف ۱.۱.۱۳ (میدان تصادفی همسانگرد). میدان تصادفی $Y(\cdot)$ همسانگرد است، هرگاه توابع تغییرنگار، کوواریانس یا همبستگی‌نگار آن همسانگرد باشند.

پذیره همسانگردی میدان تصادفی موجب آسان‌تر شدن مدل‌سازی و برآورد توابع تغییرنگار و کوواریانس می‌شود. زیرا بدون این پذیره، تابع تغییرنگار (یا هر کمیتی که وابستگی را اندازه می‌گیرد) تابعی از جهت میدان خواهد بود. به‌عبارت دیگر در هر جهت تابع تغییرنگار متفاوت خواهد بود.

مدل‌های معتبر تغییرنگار

مدل‌های معتبر تغییرنگار در شرط همیشه منفی شرطی صدق می‌کنند. مدل‌های پارامتری مختلفی در کرسی (۱۹۹۳) معرفی شده‌اند که از آن جمله می‌توان به مدل‌های تغییرنگار گاوسی، نمایی، کروی، توانی،

[۲۸] Conditional negative definite

لگاریتمی، موجی، سهمی‌گون و ماترن[۲۹] (شرمن، ۲۰۱۱) اشاره کرد. در ادامه دو مدل نمایی و ماترن را معرفی می‌کنیم.

۱. مدل نمایی: تابع نیم‌تغییرنگار نمایی به شکل

$$\gamma(\mathbf{h}) = c_\circ + c(1 - \exp\{-||\mathbf{h}||/a\}),\ \mathbf{h} \in R^d,\ d \geq 1$$

تعریف می‌شود که دارای سه پارامتر a، c و $c_\circ$ است به‌طوری که $a > 0$، $c > 0$ و $c_\circ > 0$.

۲. مدل ماترن: تابع نیم‌تغییرنگار ماترن به شکل

$$\gamma(\mathbf{h}) = c_\circ + c_1 \left[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}(\frac{\mathbf{h}}{\mathbf{h}a})^\nu K_\nu(\frac{\mathbf{h}}{\alpha})\right]$$

تعریف می‌شود که دارای دو پارامتر c و $c_\circ$ است و $K_\nu(t) = \frac{\Gamma(\alpha)}{2}(\frac{t}{2})^{-\nu}$ تابع بسل[۳۰] بهبودیافته نوع دوم مرتبه ν (آبراموویتز و استیگان، ۱۹۷۰) است.

برآورد تغییرنگار

برای برآورد تغییرنگار نیازمند برازش یک مدل معتبر به تغییرنگار تجربی داده‌ها هستیم.

تعریف ۱۴.۱.۱ (تغییرنگار تجربی). برای رده مشخصی از $\mathbf{h}$، میانگین $\gamma^*(\mathbf{h})$ که با گروه‌بندی تمام جفت نقاط به فاصله $\mathbf{h}$ محاسبه می‌شود، تغییرنگار تجربی نامیده می‌شود که در آن

$$\gamma^*(\mathbf{h}) = \frac{1}{2}\left[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})\right]^2.$$

برای برآورد تغییرنگار، ابتدا در فواصل مختلف تغییرنگار تجربی مشاهدات محاسبه و مقادیر حاصل به‌صورت نموداری برحسب فاصله رسم می‌شود. سپس با استفاده از روش‌های برازش، انواع مدل‌های تغییرنگار پارامتری معتبر به مقادیر تغییرنگار تجربی برازانده شده و پارامترهای بهترین مدل برآورد می‌شوند.

اگر میانگین میدان تصادفی $Y(\cdot)$ ثابت نباشد، برآورد تغییرنگار اریب است. کرسی (۱۹۹۳) ثابت کرد که اگر بتوان روند داده‌ها را به طریقی حذف کرد، برآورد تغییرنگار بر اساس داده‌های روند زدوده[۳۱] نسبت به برآورد آن بر اساس داده‌های اصلی، اریبی کمتری خواهد داشت. در این حالت، میدان تصادفی به‌صورت $Y(\mathbf{s}) = \Sigma_{j=1}^{p+1}\beta_j\varphi_{j-1}(\mathbf{s}) + \delta(\mathbf{s}),\ \mathbf{s} \in D$، تجزیه می‌شود، که در آن $\varphi(\mathbf{s}) = (\varphi_\circ(\mathbf{s}), \ldots, \varphi_p(\mathbf{s}))'$ بردار تابع‌هایی معلوم بر حسب موقعیت $\mathbf{s}$ هستند. آن‌گاه بردار ضرایب مجهول $\beta = (\beta_\circ, \ldots, \beta_p)'$ به یکی از روش‌های کمترین توان‌های دوم برآورد می‌شوند. در نتیجه با استفاده از داده‌های روند زدوده $\hat{\delta}(\mathbf{s}) = Y(\mathbf{s}) - \Sigma_{j=1}^{p+1}\hat{\beta}_j\varphi_{j-1}(\mathbf{s})$ تغییرنگار برآورد می‌شود.

روش‌های برازش مختلفی از جمله روش گشتاوری (ماترون، ۱۹۶۲)، درست‌نمایی ماکسیمم (ماترون، ۱۹۷۱؛ ماردیا و مارشال، ۱۹۸۴)، کمترین توان‌های دوم معمولی[۳۲] (OLS) (دیوید، ۱۹۷۷؛ یورنل و

[۲۹] Matern
[۳۰] Bessel function
[۳۱] Deterended data
[۳۲] Ordinary least square

هیگبرتس، ۱۹۷۸ و کلارک، ۱۹۷۹) کمترین توان‌های دوم تعمیم‌یافته[۳۳] (GLS) (یورنل و هیگبرتس، ۱۹۷۸) و کمترین توان‌های دوم وزنی[۳۴] (WLS) (کرسی، ۱۹۸۵) وجود دارند که تنها دو روش OLS و WLS را معرفی می‌کنیم. در روش OLS، تابع

$$\sum_{j=1}^{k} \{2\gamma^*(\mathbf{h}(j)) - 2\gamma(\mathbf{h}(j);\theta)\}^2 \qquad (1.1)$$

نسبت به θ می‌نیمم می‌شود، که در آن $2\gamma^*(\mathbf{h}(j))$ تغییرنگار تجربی و $2\gamma(\mathbf{h}(j);\theta)$ مدل تغییرنگار پارامتری، مانند نمایی، است. برای محاسبه (۱.۱)، لازم است تغییرنگار تجربی به ازای تمام فاصله‌های بین جفت مشاهدات حساب شود. در روش WLS بیش‌ترین وزن به فاصله‌های نزدیک و وزن‌های کم به فاصله‌های دور اختصاص داده می‌شوند. برای تعیین وزن‌ها روش‌های مختلفی معرفی شده‌اند که یکی از آن‌ها بر پایه واریانس است. در این حالت معکوس واریانس به‌عنوان وزن معرفی می‌شود. بنابراین در این روش θ به‌نحوی انتخاب می‌شود که عبارت

$$\sum_{j=1}^{k} \{Var\left[2\gamma^*(\mathbf{h}(j))\right]\}^{-1} \{2\gamma^*(\mathbf{h}(j)) - 2\gamma(\mathbf{h}(j);\theta)\}^2$$

می‌نیمم شود. مک‌براتنی و وبستر (۱۹۸۶) روش WLS را از لحاظ حجم محاسبات و دقت نتایج، بهتر از سایر روش‌ها معرفی کرده‌اند.

لازم به ذکر است که برای برآورد تغییرنگار، روش‌های ناپارامتری نیز به‌عنوان جانشینی برای مدل‌های پارامتری معتبر پیشنهاد شده‌اند که می‌توان به هال (۱۹۸۵) اشاره کرد که از روش‌های بازنمونه‌گیری بوت‌استرپ برای برآورد تغییرنگار استفاده کرد. در این پایان‌نامه، تنها از مدل‌های پارامتری استفاده می‌کنیم.

۴.۱.۱ پیش‌گویی فضایی

یک مساله مهم در آمار فضایی، پیش‌گویی مقدار نامعلوم میدان تصادفی در موقعیت $\mathbf{s}_0$ بر اساس مشاهدات است. معمولاً با پذیرفتن مانایی میدان تصادفی، معلوم بودن ساختار وابستگی فضایی و در نظر گرفتن تابع زیان توان دوم خطا، پیش‌گوی بهینه از می‌نیمم کردن میانگین شرطی

$$\hat{Y} = E(\mathbf{Y}(\mathbf{s}_0)|\mathbf{Y})$$

به‌دست می‌آید. بدون پذیره‌های ساده‌کننده، محاسبه این میانگین شرطی از یک توزیع $(n+1)$ متغیره مقدور نیست. یکی از این پذیره‌ها گاوسی بودن میدان تصادفی است. در آمار فضایی تعیین بهترین پیش‌گوی خطی نااریب برای $\mathbf{Y}(\mathbf{s}_0)$ کریگی[۳۵] نام دارد که اولین بار توسط ماترون (۱۹۶۳) به افتخار دی. جی. کریگ (۱۹۵۱) نام‌گذاری شده است. برای مشاهده جزییات کامل در مورد پیش‌گویی فضایی در زمین‌آمار به کرسی (۱۹۹۳) مراجعه کنید. پیش‌گویی فضایی برای مدل‌های فضایی-زمانی مورد نظر این پایان‌نامه در فصل‌های بعدی به تشریح مطرح شده است.

[۳۳] Generalized least square

[۳۴] Weighted least square

[۳۵] Kriging

۵.۱.۱ داده‌های فضایی-زمانی

اغلب داده‌های فضایی در بررسی‌های علوم محیطی مانند هواشناسی، محیط زیست، همه‌گیرشناسی، زمین‌شناسی و اقیانوس‌شناسی در طول زمان نیز به یکدیگر وابسته‌اند. مشاهداتی که هم از نظر موقعیت فضایی و هم از نظر موقعیت زمانی وابسته باشند، داده‌های فضایی-زمانی نامیده می‌شوند. با توجه به ماهیت این نوع داده‌ها، باید برخی از مفاهیم مربوط به آن‌ها را تعریف کنیم. مشابه مدل‌های معرفی‌شده برای داده‌های فضایی، مدل آماری برای داده‌های فضایی-زمانی نیز برحسب یک میدان تصادفی ارایه می‌شود که به صورت زیر آن را تعریف می‌کنیم.

تعریف ۱۵.۱.۱ (میدان تصادفی فضایی-زمانی). میدان تصادفی فضایی-زمانی مجموعه‌ای از متغیرهای تصادفی مانند $Y(\cdot,\cdot) = \{Y(\mathbf{s},\mathbf{t});\ (\mathbf{s},\mathbf{t}) \in D \times T\}$ است، که در آن $\mathbf{s}$ موقعیت فضایی در مجموعه $D \subseteq R^d$، $d \geq 1$، و $\mathbf{t}$ لحظه زمانی در مجموعه $T \subseteq R$ است.

هر میدان تصادفی فضایی-زمانی را می‌توان به صورت

$$Y(\mathbf{s},\mathbf{t}) = \mu(\mathbf{s},\mathbf{t}) + \delta(\mathbf{s},\mathbf{t})$$

تجزیه کرد، که در آن $\mu(\mathbf{s},\mathbf{t})$، $\mathbf{s} \in D \subseteq R^d,\ \mathbf{t} \in T \subseteq R$، تابع میانگین یا روند فضایی-زمانی میدان تصادفی است و $\delta(\mathbf{s},\mathbf{t})$ خطای فضایی-زمانی با میانگین صفر را نشان می‌دهد. به‌طور معمول برای مدل‌بندی روند داده‌های فضایی-زمانی روش‌های متعددی را می‌توان به کار گرفت. روش مرسوم، در نظر گرفتن روند به صورت ترکیبی خطی از موقعیت‌های فضایی و زمانی یا متغیرهای تبیینی است. ما در این پایان‌نامه مدل خطی فضایی-زمانی پویا را برای این نوع داده‌ها در نظر خواهیم گرفت. معرفی داده‌های فضایی-زمانی را با تعریف میدان تصادفی گاوسی، توابع کوواریانس و چگونگی ساخت آن به پایان می‌رسانیم.

تعریف ۱۶.۱.۱ (توزیع متناهی‌بعد میدان تصادفی). توزیع متناهی‌بعد میدان تصادفی $Y(\cdot,\cdot)$ برای $(\mathbf{s}_1,t_1),\ldots,(\mathbf{s}_k,t_k) \in D \times T$، به‌ازای $k \geq 1$، عبارت است از

$$F_{Y(\mathbf{s}_1,t_1),\ldots,Y(\mathbf{s}_k,t_k)}(y_1,\ldots,y_k) = P(Y(\mathbf{s}_1,t_1) \leq y_1,\ldots,Y(\mathbf{s}_k,t_k) \leq y_k)\,.$$

تابع چگالی متناظر با آن، در صورت وجود، تابع چگالی متناهی‌بعد میدان تصادفی نامیده می‌شود.

تعریف ۱۷.۱.۱ (میدان تصادفی فضایی-زمانی گاوسی). اگر به ازای $k \geq 1$، توزیع‌های متناهی‌بعد یک میدان تصادفی فضایی-زمانی، نرمال k متغیره باشند میدان تصادفی گاوسی نامیده می‌شود.

تعریف ۱۸.۱.۱. برای میدان تصادفی $Y(\cdot,\cdot)$ توابع میانگین، تغییرنگار و کوواریانس فضایی- زمانی به ترتیب به‌صورت زیر تعریف می‌شوند:

$$\begin{aligned}
\mu(\mathbf{s},\mathbf{t}) &= E(Y(\mathbf{s},\mathbf{t})), \qquad (\mathbf{s},\mathbf{t}) \in D \times T,\\
2\gamma(\mathbf{s},\mathbf{s}';\mathbf{t},\mathbf{t}') &= Var(Y(\mathbf{s},\mathbf{t}) - Y(\mathbf{s}',\mathbf{t}')), \qquad (\mathbf{s},\mathbf{t}),(\mathbf{s}',\mathbf{t}') \in D \times T,\\
C(\mathbf{s},\mathbf{s}';\mathbf{t},\mathbf{t}') &= Cov(Y(\mathbf{s},\mathbf{t}),Y(\mathbf{s}',\mathbf{t}')), \qquad (\mathbf{s},\mathbf{t}),(\mathbf{s}',\mathbf{t}') \in D \times T.
\end{aligned}$$

برای تحلیل داده‌های فضایی-زمانی نیازمند تعیین ساختار وابستگی آن‌ها از طریق تابع کوواریانس هستیم. در مورد تعیین این ساختار وابستگی تحقیقات گسترده‌ای انجام شده‌اند و هنوز یک زمینه تحقیقاتی پویا باقی مانده است. به عنوان چند نمونه می‌توان جونز و زانگ (۱۹۹۷)، کرسی و هوانگ (۱۹۹۹)، گنتینگ (۲۰۰۲) و استاین (۲۰۰۵) را نام برد. این تابع نقش به‌سزایی در پیش‌گویی موقعیت‌های فضایی یا زمانی فاقد مشاهده دارد و به‌طور معمول نامعلوم است و باید بر اساس مشاهدات برآورد شود. برازش مدل‌های معتبر به تابع کوواریانس فضایی-زمانی را پذیره‌هایی مانند مانایی، همسان‌گردی، تقارن[۳۶] و تفکیک‌پذیری[۳۷] آسان می‌کنند. به‌عنوان نمونه روحانی و هال (۱۹۸۹) توابع کوواریانس فضایی-زمانی تفکیک‌پذیر حاصل از جمع دو تابع کوواریانس فضایی و زمانی را تعریف کردند و برای حالت تفکیک‌ناپذیر، کرسی و هوانگ (۱۹۹۹) یک روش طیفی برای ساخت توابع کوواریانس فضایی-زمانی پیشنهاد دادند. یک راه دیگر استفاده از مدل‌های فضایی پویاست که با نحوه ساخت این توابع کوواریانس در فصل ۴ آشنا می‌شویم.

۶.۱.۱ مدل‌های آمیخته خطی

تقریباً در اغلب موارد استفاده از روش‌های آماری، مرکز توجه روی میانگین و تغییرات آن‌هاست که منجر به ایجاد روش‌های استنباط درباره میانگین‌ها یا درباره ماهیت آن‌ها می‌شود. در مدل‌های رگرسیونی نیز معمولاً مدل به صورت میانگین یک توزیع شرطی تعریف می‌شود. معمول‌ترین مدل رگرسیونی، مدل خطی[۳۸] (LM) است. در یک مدل خطی با متغیر پاسخ Y و ماتریس طرح X، تابع رگرسیونی، که همان میانگین شرطی $Y|X$ است، یک ترکیب خطی از X به صورت $E(Y|X) = X\beta$ در نظر گرفته می‌شود. در این مدل، بردار β، ضرایب رگرسیونی نامعلوم یا به عبارتی اثرات ثابت[۳۹] نامیده می‌شوند.

یکی از پذیره‌های مدل‌های خطی، استقلال بین مشاهدات است. زمانی که برای پاسخ‌ها این پذیره برقرار نیست، تعمیمی از این مدل‌ها معروف به مدل‌های آمیخته خطی[۴۰] (LMMs) مورد استفاده قرار می‌گیرد. در این مدل‌ها برای در نظر گرفتن وابستگی بین پاسخ‌ها از متغیرهای پنهان[۴۱] یا همان اثرات تصادفی[۴۲] استفاده می‌شود. میانگین پاسخ در یک مدل آمیخته خطی به صورت

$$E\left[Y|X, Z, \alpha\right] = X\beta + Z\alpha \qquad (۲.۱)$$

نوشته می‌شود که در آن X ماتریس طرح متناظر با اثرات ثابت β و Z ماتریس طرح متناظر با اثرات تصادفی α می‌باشند (مک‌کالاک و سیرل، ۲۰۰۱). اثرات تصادفی معمولاً به عنوان متغیرهای تصادفی نرمال (با میانگین صفر) در نظر گرفته می‌شوند و پارامترهایی که توزیع این اثرات را توصیف می‌کنند، به عنوان پارامترهای وابستگی، برآورد می‌شوند.

---

[۳۶] Symmetry
[۳۷] Separability
[۳۸] Linear model
[۳۹] Fixed effects
[۴۰] Linear mixed models
[۴۱] Latent variables
[۴۲] Random effects

۷.۱.۱ مدل‌های آمیخته خطی تعمیم‌یافته

در بسیاری از موقعیت‌های کاربردی که متغیر پاسخ پیوسته نیست، مانند پاسخ‌های دودویی و شمارشی، استفاده از توزیع نرمال به عنوان مدل خطا و مدل خطی مناسب نیست. زیرا مدل‌های خطی محدودیت موجود در میانگین پاسخ برای این داده‌ها را در نظر نمی‌گیرند. به عنوان مثال برای داده‌های دودویی، میانگین که همان احتمال موفقیت می‌باشد، در فاصله $(۰,۱)$ قرار می‌گیرد و چنان‌چه از مدل خطی استفاده کنیم ممکن است برآورد میانگین پاسخ هر مقداری در مجموعه اعداد حقیقی باشد. در نتیجه برای این داده‌ها، با فرض توزیع برنولی، معمولاً از رگرسیون پروبیت[۴۳] یا لجستیک[۴۴] استفاده می‌شود. برای مرتفع ساختن این مشکل نلدر و ودربرن (۱۹۷۲) مدل‌های خطی تعمیم‌یافته[۴۵] (GLMs) را معرفی کردند. این مدل‌ها تعمیمی از مدل‌های خطی هستند که برای تحلیل پاسخ‌های عضو خانواده توزیع‌های نمایی گسترش یافته‌اند. ذات این تعمیم در این‌گونه مدل‌ها دو چیز است: اول، متغیرهای پاسخ می‌توانند غیرنرمال باشند؛ دوم، میانگین پاسخ لزوماً به صورت ترکیبی خطی از متغیرهای تبیینی نیست، بلکه تابعی از آن به نام تابع پیوند[۴۶] به صورت ترکیب خطی از متغیرهای تبیینی در نظر گرفته می‌شود.

تعریف ۱.۱.۱۹. خانواده‌ای از توزیع‌ها را زیرده‌ای از خانواده توزیع‌های نمایی متعارف[۴۷] گوییم، هرگاه بتوان تابع چگالی آن را به شکل زیر نوشت:

$$f_Y(y) = \exp\left\{\frac{y\gamma - \psi(\gamma)}{a(\phi)} - c(y,\phi)\right\} \qquad (۳.۱)$$

که در آن ϕ و γ پارامترهای خانواده توزیع و $c(\cdot)$، $\psi(\cdot)$ و $a(\cdot)$ توابعی معلوم هستند. بنابراین، در یک خانواده نمایی، میانگین و واریانس به صورت

$$E(Y) = \frac{\partial\psi(\gamma)}{\partial(\gamma)} = \mu, \qquad Var(Y) = a(\phi)\frac{\partial^۲\psi(\gamma)}{\partial(\gamma)^۲} \qquad (۴.۱)$$

محاسبه می‌شوند.

در یک مدل خطی تعمیم‌یافته، اگر η ترکیبی خطی از متغیرهای تبیینی و ضرایب رگرسیونی، به نام پیش‌گوی خطی[۴۸] باشد، آن‌گاه

$$g(\mu) = \eta = X\beta$$

که $g(\cdot)$ یک تابع پیوند است. این تابع با پیوند میانگین پاسخ و پیش‌گوی خطی، میانگین را مدل‌بندی می‌کند. از جمله توابع پیوند مورد استفاده در GLMs، توابع پیوند لجیت، $g(\mu) = \ln\frac{\mu}{۱-\mu}$، پروبیت، $g(\mu) = \Phi^{-۱}(\mu)$ و لگاریتمی، $g(\mu) = \ln(\mu)$، می‌باشند. برای مقایسه این توابع به مک‌کالاک و نلدر (۱۹۸۹) مراجعه کنید.

---

[۴۳] Probit regression

[۴۴] Logistic

[۴۵] Generalized linear models

[۴۶] Link function

[۴۷] Canonical exponential distributions family

[۴۸] Linear predictor

از آن‌جایی که پذیره استقلال در بین مشاهدات غیرنرمال همیشه برقرار نیست، پس از مدل‌های آمیخته خطی تعمیم‌یافته[۴۹] (GLMMs) استفاده خواهیم کرد. این مدل‌ها ترکیبی طبیعی از دو رده LMMs و GLMs هستند. در واقع با افزودن اثرات تصادفی به پیش‌گوی خطی در یک GLM، یک خانواده از مدل‌های منعطف که در موقعیت‌های عملی مختلفی کاربرد دارند، تشکیل می‌شود.

متغیر پاسخ در یک GLMM مستقل شرطی بوده و دارای توزیعی از خانواده نمایی با تابع چگالی احتمال (۳.۱) می‌باشد و به‌طور مشابه، میانگین و واریانس آن از رابطه (۴.۱) به‌دست خواهند آمد. اگر میانگین شرطی برابر با $E[Y|X,Z,\alpha]=\mu$ باشد، آن‌گاه تابع پیوند روی این میانگین به عنوان ترکیبی خطی از هر دو عامل ثابت و تصادفی به مدل معرفی خواهد شد. یعنی

$$g(\mu)=X\beta+Z\alpha.$$

یک استفاده معمول از GLMMs در تحلیل داده‌های فضایی است.

۲.۱ دیدگاه‌های بسامدی و بیزی در استنباط آماری

روش‌های آماری را در یک دسته‌بندی کلی می‌توان به دو دسته بسامدی[۵۰] بر پایه درستنمایی و بیزی[۵۱] تقسیم کرد. هر کدام از این چارچوب‌ها منطق و هدف خاص خود را در استنباط آماری دنبال می‌کنند (رابرت، ۲۰۰۷). در حیطه مدل‌بندی، یکی از شاخه‌های پرطرفدار استنباط‌های بسامدی مبتنی بر تابع درستنمایی است. فرض کنید $y_1,\ldots,y_n$ یک نمونه تصادفی n تایی با تابع (چگالی) احتمال n متغیره $f(\cdot|\theta)$ است که در آن $\theta\in\theta$ بردار پارامتر مورد علاقه می‌باشد. در این صورت، تابع درستنمایی به صورت

$$L(\theta|\mathbf{y})=f(y_1,\ldots,y_n|\theta)$$

تعریف می‌شود. روش‌های معمول استنباط بسامدی، روش‌های مبتنی بر درستنمایی هستند که برای محاسبه تابع درستنمایی در رده مدل‌های آمیخته خطی و آمیخته خطی تعمیم‌یافته، حل عددی انتگرال‌های با بعد بالا را شامل می‌شوند که دارای محاسبات زمان‌بر و پیچیده‌ای می‌باشند. برای تشریح این پیچیدگی، مدل آمیخته خطی (۲.۱) را در نظر بگیرید. اگر فرض کنیم اثرات تصادفی $\alpha\in\Omega\subseteq R^r$ دارای تابع توزیع $G(\cdot|\nu)$ با تابع چگالی $g(\cdot|\nu)$ باشند، آن‌گاه تابع درستنمایی کناری پارامترها به صورت زیر تعریف می‌شود:

$$L(\theta,\nu|\mathbf{y})=\int_{R^r} f(\mathbf{y}|\alpha,\theta,\nu)g(\alpha|\nu)d\alpha.$$

به‌سادگی می‌توان درک کرد که با بزرگ بودن بعد اثرات تصادفی (که در مثال‌های فضایی بعد میدان تصادفی معمولاً با بعد مشاهدات برابر است) و پیچیده بودن تابع زیر انتگرال که از پیچیده بودن $g(\cdot|\nu)$ نتیجه می‌شود، محاسبه این انتگرال چقدر می‌تواند مشکل باشد.

[۴۹] Generalized linear mixed models

[۵۰] Frequecy

[۵۱] Bayesian

تقریب این نوع انتگرال‌ها معمولاً با دو رهیافت صورت می‌پذیرد. رهیافت اول، شامل روش‌هایی است که انتگرال را به صورت عددی تقریب می‌زنند، مانند روش‌های تربیع‌بندی[۵۲] (پینیر و بیتس، ۱۹۹۵). رهیافت دوم، روش‌هایی هستند که تابع زیر انتگرال را طوری تقریب می‌زنند که انتگرال آن صورت بسته داشته باشد، از جمله روش‌های تقریب لاپلاس[۵۳] (بریسلو و کلیتون، ۱۹۹۳؛ بریسلو و لین، ۱۹۹۵) و شبه‌درستنمایی تاوانیده[۵۴] (بریسلو و کلیتون، ۱۹۹۳).

هریک از روش‌های موجود در این دو رهیافت، دارای معایب قابل توجهی هستند. مشکل اساسی روش‌های انتگرال‌گیری عددی در شناسایی نواحی مهم (نواحی چگال‌تر) برای تابعی که باید انتگرال‌گیری شود، می‌باشد. انتگرال‌گیری‌های عددی برای نواحی با بعد حداکثر ۶ خوب عمل می‌کنند و در غیر این صورت با خطای زیادی مواجه می‌شوند، زیرا با افزایش بعد انتگرال با تجمیع خطاها مواجه می‌شویم. افزایش بعد انتگرال، افزایش محاسبات را نیز به همراه خواهد داشت (بریسلو و کلیتون، ۱۹۹۳). روش‌های تقریبی لاپلاس و شبه‌درستنمایی تاوانیده نیز معمولاً منجر به برآوردهای اریب به‌ویژه برای پاسخ‌های دودویی می‌شوند (مک‌کالاک، ۱۹۹۷).

مشکلات ذکرشده در مسیر استنباط مبتنی بر درستنمایی، در مدل‌های فضایی و فضایی-زمانی به‌دلیل بعدهای بالای اثرات تصادفی فضایی جدی‌تر نیز می‌شوند. بنابراین دیدگاه بیزی، که استنباط در آن مبتنی بر توزیع پسین است، می‌تواند یک جایگزین مناسب باشد. توزیع پسین در یک مدل بیزی، به‌روز شده باور و دیدگاه اولیه در مورد پارامتر با توجه به مشاهدات است. در یک مدل بیزی، فرض می‌شود مقدار واقعی پارامتر θ تحققی از یک توزیع احتمالی به نام توزیع پیشین است. اگر این توزیع را با $\pi(\theta)$ نشان دهیم، آن‌گاه بنابر قاعده بیز، توزیع پسین $\pi(\theta|\mathbf{y})$ معادل است با

$$\pi(\theta|\mathbf{y}) = \frac{L(\theta|\mathbf{y})\pi(\theta)}{\int L(\theta|\mathbf{y})\pi(\theta)d\theta} \propto c(\mathbf{y})L(\theta|\mathbf{y})\pi(\theta)$$

که در آن $c(\mathbf{y})$ ثابت نرمال‌ساز[۵۵] نامیده می‌شود و برابر است با

$$c(\mathbf{y}) = \frac{۱}{\int L(\theta|\mathbf{y})\pi(\theta)d\theta}.$$

رهیافت بیزی برای تقریب انتگرال‌های با بعد بالا از روش‌های مبتنی بر نمونه‌گیری (نمونه‌گیری از توزیع پسین) استفاده می‌کند. از جمله این روش‌ها می‌توان به روش‌های نمونه‌گیری نقاط مهم[۵۶] (هستربرگ، ۱۹۸۷) و MCMC (هستینگز، ۱۹۷۰) اشاره کرد. در این روش‌ها، قانون قوی اعداد بزرگ[۵۷] و قضیه ارگودیک[۵۸] (بیرخوف، ۱۹۴۲) ضامن همگرایی برآوردگرهای حاصل از نمونه‌ها به کمیت‌های مورد علاقه از توزیع پسین می‌باشند. روش نمونه‌گیری نقاط مهم برای انتگرال‌های با بعد کوچک مورد استفاده قرار می‌گیرد، در حالی که روش‌های MCMC برای مسائل انتگرال‌گیری با بعد بالا می‌توانند مفید باشند

[۵۲] Quadrature methods
[۵۳] Laplace approximation
[۵۴] Penalized quasi likelihood
[۵۵] Normalizing constant
[۵۶] Importance sampling
[۵۷] Strong law of large numbers
[۵۸] Ergodic theorem

به‌طوری که دیگر نیازی به حل انتگرال‌های با بعد بالا، ماکسیمم‌سازی و مشتق‌گیری از یک تابع پیچیده نیست (رابرت و کسلا، ۲۰۰۴).

۱.۲.۱ توزیع پیش‌گوی بیزی

در دیدگاه بیزی برای پیش‌گویی مقدار جدید پاسخ مانند y_0، از توزیع پیش‌گوی بیزی[۵۹] استفاده می‌شود. بدیهی است که اگر مدل بیزی پیچیده باشد، محاسبه توزیع پیش‌گو نیز مشکل بوده و نیاز به الگوریتم‌های تقریبی MCMC دارد.

تعریف ۱.۲.۱ (توزیع پیش‌گوی بیزی)**.** فرض کنید $\{Y(\mathbf{s})\,;\,\mathbf{s}\in D\}$ یک میدان تصادفی از توزیعی با تابع چگالی احتمال $f(\mathbf{y}|\phi)$ باشد، که در آن ϕ پارامترهای مدل و دارای توزیع پیشین $\pi(\phi)$ هستند. در این صورت توزیع پسین ϕ بر حسب مشاهدات $\mathbf{y}$، متناسب است با

$$\pi(\phi|\mathbf{y}) \propto f(\mathbf{y}|\phi)\pi(\phi).$$

اگر توزیع مشاهده جدید y_0 به صورت $f(y_0|\mathbf{y},\phi)$ باشد که می‌تواند به مشاهدات $\mathbf{y}$ و ϕ وابسته باشد، آن‌گاه توزیع پیش‌گو با میانگین‌گیری از این توزیع نسبت به توزیع پسین $\pi(\phi|\mathbf{y})$ به صورت

$$f(y_0|\mathbf{y}) = \int f(y_0|\mathbf{y},\phi)\pi(\phi|\mathbf{y})d\phi$$

حاصل می‌شود که توزیع پیش‌گوی بیزی نامیده می‌شود.

۳.۱ روش‌های مونت کارلوی زنجیر مارکوفی

در این بخش روش‌های مونت کارلوی زنجیر مارکوفی را به اختصار معرفی میکنیم. مونت کارلو نام بندر کوچکی در فرانسه است که در آن بازی‌های شانسی انجام می‌شده است. ساده‌ترین ابزار بازی‌ها چرخ رولت بوده که شامل توپی کوچک در محفظه‌ای است که روی چرخی گردان به دور محور خود قرار دارد. چرخ پس از مدتی متوقف شده و عددی به تصادف از محفظه خارج می‌شود. از این رو، دستگاه ذکرشده ساده‌ترین و قدیمی‌ترین روش تولید اعداد تصادفی بوده است. روش مونت کارلو شکل ویژه‌ای از انتگرال‌گیری عددی است که از اعداد تصادفی برای انتگرال‌گیری استفاده می‌کند. برای تشریح مسأله، فرض کنید محاسبه مقدار انتگرال

$$I = \int h(x)f(x)dx$$

مورد نظر است که در آن $f(\cdot)$ یک تابع چگالی احتمال است. با داشتن نمونه $x_1,\ldots,x_n$ از توزیع $f(x)$، تقریب مونت کارلوی I به صورت

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

[۵۹] Bayesian predictive distribution

محاسبه می‌شود. در حالت آرمانی، وقتی x_i ها مستقل باشند، با افزایش n دقت برآورد بیشتر می‌شود. ولی همواره این امر میسر نیست و نمی‌توان نمونه‌ای مستقل از توزیع مورد نظر تولید کرد. زیرا شکل تابع $f(\cdot)$ می‌تواند پیچیده و تولید نمونه مستقیم از آن مشکل باشد. یک راه حل معرفی‌شده تولید نمونه به صورت غیرمستقیم است. روش‌های تولید نمونه غیرمستقیم معمولاً یک توزیع پیشنهادی[۶۰] را در نظر می‌گیرند که تولید نمونه از آن ساده باشد. سپس با تعریف الگوریتمی، نمونه پیشنهادی به عنوان نمونه‌ای از توزیع هدف[۶۱] $f(\cdot)$ پذیرفته یا رد می‌شود. از جمله این روش‌ها می‌توان به نمونه‌گیری رد و پذیرش[۶۲] (کسلا و همکاران، ۲۰۰۴) و نمونه‌گیری نقاط مهم اشاره کرد. یکی از مشکلات این دو روش، عدم کارایی آن‌ها در ابعاد بالاست، زیرا معرفی توزیع پیشنهادی در این حالت خود یک مساله چالش‌برانگیز محسوب می‌شود.

یک رده کلی تولید نمونه غیرمستقیم (وابسته) که در بعد بالا نیز کارایی خود را حفظ می‌کند، روش‌های نمونه‌گیری از زنجیرهای مارکوف با توزیع مانای $f(\cdot)$ است (رابرتز و روزنتال، ۲۰۰۴). این رهیافت اولین بار توسط متروپولیس و همکاران (۱۹۵۳) معرفی شد و توسط هستینگز (۱۹۷۰) با جزییات نظری کامل تدوین شد.

زنجیرهای مارکوفی با توزیع مانای $f(\cdot)$، وقتی به این توزیع همگرا شوند، نمونه‌ای وابسته از $f(x)$ تولید می‌کنند. روش‌های MCMC به تولید زنجیر مارکوفی مانند $X_0, X_1, \ldots$ می‌پردازند که توزیع X_n به شرط $(X_0, \ldots, X_{n-1})$ با توزیع X_n به شرط X_{n-1} یکسان است و این زنجیر نمونه‌ای از توزیع هدف تولید می‌کند که با افزایش n زنجیر به توزیع هدف همگرا می‌شود. در این حالت، زنجیرهای تولیدشده نسبت به توزیع هدف

- برگشت‌پذیرند به این معنی که (X_n, X_{n+1}) و (X_n, X_{n-1}) هم‌توزیع هستند.
- تحویل‌ناپذیرند یعنی صرف‌نظر از نقطه شروع، زنجیر می‌تواند وارد هر وضعیتی بشود.

دارا بودن این دو ویژگی تضمین می‌کند که اگر زنجیر دارای یک توزیع حدی باشد، آن‌گاه این توزیع همان توزیع هدف خواهد بود. دو روش مهم که با کمک زنجیرهای مارکوفی از توزیع‌های پیچیده نمونه‌گیری می‌کنند، الگوریتم متروپولیس-هستینگز[۶۳] (MH) (متروپولیس و همکاران، ۱۹۵۳ و هستینگز، ۱۹۷۰) و نمونه‌گیری گیبز[۶۴] (GS) (گمان و گمان، ۱۹۸۴ و گلفاند و اسمیت، ۱۹۹۰) هستند. در ادامه سه الگوریتم متروپولیس-هستینگز، الگوریتم متروپولیس-هستینگز قدم زدن تصادفی[۶۵] (RWMH) و نمونه‌گیری گیبز را به اختصار معرفی می‌کنیم.

---

[۶۰] Proposal density
[۶۱] Target distribution
[۶۲] Accept-reject sampling
[۶۳] Metropolis-Hastings
[۶۴] Gibbs sampling
[۶۵] Random walk Metropolis-Hastings

۱.۳.۱ الگوریتم متروپولیس-هستینگز

فرض کنید X دارای تابع چگالی احتمال $f(x)$ باشد که نمونه‌گیری از آن به‌طور مستقیم سخت یا امکان‌ناپذیر است. در الگوریتم MH برای نمونه‌گیری از تابع هدف $f(x)$ به‌صورت زیر عمل می‌شود:
ابتدا تابع چگالی پیشنهادی $q(y|x)$ چنان انتخاب می‌شود که تکیه‌گاه آن با تکیه‌گاه $f(x)$ یکسان، دم‌های آن از دم‌های $f(x)$ کلفت‌تر و نمونه‌گیری از آن ساده باشد. سپس با در نظر گرفتن مقدار اولیه دلخواه x_0 و قرار دادن $i = 0$، برای تولید x_{i+1} گام‌های زیر طی می‌شوند:

۱) تولید مقدار y از توزیع پیشنهادی $q(y|x_i)$.

۲) محاسبه نرخ پذیرش $r_i = r(x_i, y) = min\left\{1, \frac{f(y)q(x_i|y)}{f(x_i)q(y|x_i)}\right\}$.

۳) تعیین مقدار جدید x_{i+1} به صورت زیر:

$$x_{i+1} = \begin{cases} y & \text{با احتمال } r_i \\ x_i & \text{با احتمال } 1 - r_i \end{cases}$$

راهی ساده برای اجرای گام سوم این است که مقدار u از توزیع یکنواخت استاندارد تولید و قرار داده شود

$$x_{i+1} = \begin{cases} y & \text{اگر } u < r_i \\ x_i & \text{اگر } u \geq r_i \end{cases}$$

با قراردادن $i = i + 1$ و تکرار گام‌های سه‌گانه بالا، مقادیر $x_0, x_1, \ldots$ یک زنجیر مارکوف تشکیل می‌دهند که توزیع مانای آن همان توزیع هدف $f(x)$ است.

عملکرد الگوریتم معمولاً بر حسب نرخ پذیرش آن ارزیابی می‌شود. نرخ پذیرش الگوریتم، میانگین احتمال پذیرش در تمام تکرارهاست که به صورت

$$\bar{\rho} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=0}^{T} r(x_i, y_i) = \int r(x, y) f(x) q(y|x) dy dx$$

تعریف می‌شود. کارایی الگوریتم MH به تولید نمونه از تابع چگالی پیشنهادی $q(y|x)$ وابسته است. در واقع یک نمونه خوب وقتی تولید می‌شود که تابع چگالی پیشنهادی یک تقریب مناسب برای تابع چگالی هدف باشد.

۲.۳.۱ الگوریتم متروپولیس-هستینگز قدم زدن تصادفی

اجرای یک الگوریتم MH با مشکلاتی مواجه است. در مواردی ممکن است تعیین تابع چگالی پیشنهادی دشوار باشد. یک روش جایگزین، کاوش در همسایگی مقدار جاری زنجیر می‌باشد. این ایده، برای تولید مقدار بعدی زنجیر، نمونه تولیدشده قبلی را به حساب می‌آورد. به این معنی که در همسایگی موقعیت جاری زنجیر، کاوش موضعی بیشتری را انجام می‌دهد. اجرای این ایده، شبیه‌سازی Y_i از رابطه

$$Y_i = X_i + \varepsilon_i$$

است، که در آن ε_i یک نوفه تصادفی مستقل از X_i با تابع چگالی $g(\cdot)$ است. به عنوان مثال، انتخاب توزیع یکنواخت برای نوفه به این معنی است که

$$Y_i \sim U(X_i - \delta, X_i + \delta).$$

یا توزیع پیشنهادی نرمال به معنی

$$Y_i \sim N(X_i, \tau^2)$$

است. در این الگوریتم توزیع پیشنهادی عبارتست از

$$q(y|x) = g(y - x).$$

اگر تابع چگالی $g(\cdot)$ حول صفر متقارن باشد، که معمولاً این طور است، زنجیر مارکوف متناظر با آن یک فرآیند قدم زدن تصادفی است. با این توضیحات، الگوریتم RWMH به صورت زیر اجرا می‌شود:

با شرط داشتن x_i

۱- $y_i \sim g(y - x_i)$ را تولید کن.

۲- قرار بده

$$x_{i+1} = \begin{cases} y_i & \text{با احتمال } r_i \\ x_i & \text{با احتمال } 1 - r_i \end{cases}$$

که در آن $r_i = min\left\{1, \frac{f(y_i)}{f(x_i)}\right\}$. با قراردادن $i = i + 1$ و تکرار گام‌های دو گانه بالا، مقادیر $x_0, x_1, \ldots$ یک زنجیر مارکوف خواهد بود که توزیع مانای آن همان توزیع هدف $f(x)$ است.

به دلیل وجود مرحله پذیرش MH، زنجیر $\{x_i\}$ یک زنجیر قدم زدن تصادفی نیست. احتمال پذیرش هم به $g(\cdot)$ وابسته نیست. به این معنی که، برای یک جفت مفروض (x_i, y_i)، احتمال پذیرش چه y_i از توزیع نرمال تولید شود چه از توزیع کوشی، یکسان خواهد بود. اما انتخاب $g(\cdot)$ های مختلف منجر به دامنه‌های مختلف مقادیر برای y_i ها و در نتیجه نرخ‌های پذیرش متفاوت خواهند شد. بنابراین نمی‌توان گفت انتخاب $g(\cdot)$ بر رفتار الگوریتم تأثیری ندارد.

پارامتر مقیاس فرآیند قدم زدن تصادفی، که وابسته به واریانس نوفه تصادفی است، باید به‌گونه‌ای انتخاب شود (کالبیده شود) تا کاوش به خوبی انجام شود. این کالبیدن[۶۶] برای رسیدن به یک تقریب مناسب از توزیع هدف در تعداد تکرار قابل قبول، خیلی مهم است. اگر واریانس جمله نوفه خیلی بزرگ باشد، اکثر نمونه‌های پیشنهادی رد می‌شوند و الگوریتم شدیداً ناکاراست. وقتی که این واریانس خیلی کوچک است، اکثر نمونه‌های پیشنهادی پذیرفته شده و زنجیری تولید می‌شود که خیلی شبیه به یک فرآیند قدم زدن تصادفی است و باز هم ناکاراست. یک راه برای انتخاب پارامتر مقیاس، بررسی نرخ‌های پذیرش است. رابرتز و روزنتال (۲۰۰۱) نشان دادند یک الگوریتم RWMH بهینه دارای نرخ پذیرش تقریبی ۰/۲۳۴ است. اگر نرخ پذیرش الگوریتم در دامنه [۰/۱۵, ۰/۵] قرار گیرد، مناسب است. برای آگاهی بیشتر و عمیق‌تر از جزییات این الگوریتم و الگوریتم‌های مشابه به رابرت و کسلا (۲۰۰۴) مراجعه کنید.

[۶۶] Calibration

۳.۳.۱ الگوریتم نمونه‌گیری گیبز

فرض کنید توزیع توأم $X = (X_1, \ldots, X_k)$ فرم پیچیده یا نامعلومی داشته باشد، اما چگالی‌های شرطی کامل $[X_r|X_{-r}]$ به ازای $r = 1, \ldots, k$ در دسترس باشند به‌طوری که

$$X_{-r} = (X_1, \ldots, X_{r-1}, X_{r+1}, \ldots, X_k).$$

فرض کنید علاقه‌مند به تولید نمونه از تابع چگالی توأم $f(x_1, \ldots, x_k)$ یا توابع چگالی کناری $f(x_j)$، $j = 1, \ldots, k$، هستیم. برای این منظور از روش GS به صورت زیر می‌توان استفاده کرد:

با در نظر گرفتن مقادیر اولیه دلخواه $x^{(0)} = (x_1^{(0)}, \ldots, x_k^{(0)})$ و قرار دادن $i = 0$، برای تولید $x^{(i+1)}$ گام‌های زیر اجرا می‌شود:

۱- تولید مقدار $x_1^{(i+1)}$ از توزیع $f(x_1|x_2^{(i)}, \ldots, x_k^{(i)})$.

۲- تولید مقدار $x_2^{(i+1)}$ از توزیع $f(x_2|x_1^{(i+1)}, x_3^{(i)} \ldots, x_k^{(i)})$.

.

.

.

k- تولید مقدار $x_k^{(i+1)}$ از توزیع $f(x_k|x_1^{(i+1)}, \ldots, x_{k-1}^{(i+1)})$.

با قرار دادن $i = i + 1$ و تکرار گام‌های k گانه بالا، نمونه $x^{(0)}, x^{(1)}, \ldots$ یک نمونه از توزیع هدف می‌باشد. وقتی $i \to \infty$، توزیع حدی $x^{(i)}$، همان $f(x)$ خواهد بود؛ یعنی برای i های بزرگ $x^{(i)}$ یک مشاهده تصادفی از توزیع $f(x)$ است. دنباله $x^{(0)}, x^{(1)}, \ldots$ تحققی از یک زنجیر مارکوف است. زیرا توزیع احتمال $x^{(i+1)}$ فقط به $x^{(i)}$ وابسته است و مستقل از $x^{(i-1)}, \ldots, x^{(1)}, x^{(0)}$ می‌باشد. به همین دلیل نمونه‌گیری گیبز یک روش MCMC است (کسلا و جرج، ۱۹۹۲).

۴.۱ سایر تعاریف

سایر کمیت‌های لازم برای ورود به مباحث فصل‌های آینده را در این بخش معرفی می‌کنیم.

۱.۴.۱ پیچیدگی محاسباتی الگوریتم‌ها

برای نشان دادن رابطه میان تعداد داده‌ها و منابع محاسباتی مورد نیاز برای حل یک مسأله با استفاده از یک الگوریتم، در متون یادگیری ماشین، از نماد $O(\cdot)$ استفاده می‌شود (گلوب و وان لون، ۱۹۹۶). از این نماد معمولاً برای بررسی زمان یا حافظه مورد نیاز برای حل مسأله‌ای با تعداد زیادی ورودی استفاده می‌شود. این کمیت را به صورت زیر می‌توان تعریف کرد:

اگر $f(x)$ و $g(x)$ توابعی باشند که روی زیرمجموعه‌ای از اعداد حقیقی تعریف شده باشند، آن‌گاه زمانی که $x \to \infty$

$$f(x) = O\left(g(x)\right)$$

اگر و تنها اگر عدد حقیقی مثبت M و عدد حقیقی x_0 موجود باشند به‌طوری که

$$|f(x)| \leq M|g(x)|, \quad x > x_0.$$

به بیان دیگر این نماد همان مفهوم کران بالا را دارد. مثلاً برای حافظه مصرفی، $O(\cdot)$ کران بالای مصرف حافظه را مشخص می‌کند. پس الگوریتمی مفید است که کران بالایی زمان اجرای آن پایین باشد. به‌عنوان مثال اگر تعداد عملیات اجرای الگوریتمی $2n^3 + n^2 + n$ باشد، برای n های بزرگ آن را با $O(n^3)$ نمایش می‌دهیم.

توجه داشته باشید، در متون آماری از این نماد برای نمایش نرخ‌های همگرایی برآوردگرها به مقادیر واقعی (سازگاری برآوردگرها) استفاده می‌شود.

تعریف ۱.۴.۱ (تابع trsolve). اگر $\mathbf{x}$ جواب دستگاه مثلثی $L\mathbf{x} = \mathbf{y}$ باشد که در آن L یک ماتریس بالا مثلثی $n \times n$، و $\mathbf{y}$ و $\mathbf{x}$ بردارهای n بعدی هستند، آنگاه تابع trsolve را به‌صورت

$$\mathbf{x} = trsolve(L, \mathbf{y})$$

نمایش می‌دهیم.

۲.۴.۱ توزیع‌ها

در این بخش، چند توزیع آماری را که در این پایان‌نامه از آن‌ها استفاده شده است، معرفی می‌کنیم.

تعریف ۲.۴.۱ (توزیع نرمال چندمتغیره). بردار $X = (X_1, \ldots, X_d)$ دارای توزیع نرمال d متغیره است که با نماد $N_d(\mu, \Sigma)$ نمایش می‌دهیم، هرگاه تابع چگالی آن به صورت

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}, \quad \mathbf{x} \in R^d$$

باشد، به‌طوری که $\mu = (\mu_1, \ldots, \mu_d)^T$ بردار میانگین و $\Sigma = (\sigma_{ij})$ ماتریس $d \times d$ متقارن و همیشه مثبت با درایه‌های $\sigma_{ij} = Cov(X_i, X_j)$ است. ماتریس Σ^{-1} که معکوس ماتریس کوواریانس است، به ماتریس دقت[۶۷] معروف می‌باشد.

تعریف ۳.۴.۱ (توزیع ویشارت). ماتریس $p \times p$ و همیشه مثبت X دارای توزیع ویشارت با n درجه آزادی و ماتریس مقیاس V است که با نماد $X \sim W_p(V, n)$ نمایش می‌دهیم، هرگاه تابع چگالی آن به صورت

$$\frac{|X|^{n-p-1/2}}{2^{np/2}|V|^{n/2}\Gamma_p(n/2)} \exp\left\{-\frac{1}{2}tr\left(V^{-1}X\right)\right\}$$

باشد، به‌طوری که $\Gamma_p(\cdot)$ تابع گاما و $tr(\cdot)$ تابع اثر[۶۸] ماتریس می‌باشد.

[۶۷] Precision matrix

[۶۸] Trace

تعریف ۴.۴.۱ (توزیع ویشارت معکوس). ماتریس همیشه مثبت X، دارای توزیع ویشارت معکوس است که با نماد $X \sim W_p^{-1}(V,n)$ نمایش می‌دهیم، هرگاه معکوس آن دارای توزیع ویشارت باشد؛ یعنی $X^{-1} \sim W_p(V^{-1},n)$.

تعریف ۵.۴.۱ (توزیع گامای معکوس). متغیر تصادفی Y دارای توزیع گامای معکوس است و با نماد $Y \sim IG(\alpha,\beta)$ نمایش می‌دهیم، هرگاه تابع چگالی آن به صورت

$$f(y) = \frac{\beta^\alpha y^{-\alpha-1}}{\Gamma(\alpha)} \exp\left(-\frac{\beta}{y}\right) \ , \ y \in (0,\infty) \, , \alpha > 0 , \beta > 0$$

باشد.

مشابه توزیع ویشارت، اگر متغیر تصادفی Y دارای توزیع گامای معکوس باشد، معکوس آن دارای توزیع گاما خواهد بود.

لم ۶.۴.۱ (لیندلی و اسمیت، ۱۹۷۲). *فرض کنید* $\mathbf{y}|\theta_1 \sim N_n(A_1\theta_1, C_1)$ و $\theta_1|\theta_2 \sim N_n(A_2\theta_2, C_2)$. *در این صورت*

الف) توزیع حاشیه‌ای $\mathbf{y}$ *عبارتست از*

$$\mathbf{y} \sim N_n(A_1A_2\theta_2, C_1 + A_1C_2A_1^T)$$

ب) توزیع شرطی $\theta_1|\mathbf{y}$ *برابر* $N_n(Bb,B)$ *است که در آن*

$$B^{-1} = A_1^TC_1^{-1}A_1 + C_2^{-1}$$

و

$$b = A_1^TC_1^{-1}\mathbf{y} + C_2^{-1}A_2\theta_2.$$

به شکلی مشابه لم ۶.۴.۱، لم زیر را داریم.

لم ۷.۴.۱. *فرض کنید* $\mathbf{y}|\theta_1 \sim N_n(A_1\theta_1, C_1)$، $\theta_1|\theta_2 \sim N_n(A_2\theta_2, C_2)$ و $\theta_1|\theta_3 \sim N_n(A_3\theta_3, C_3)$. *در این صورت، توزیع شرطی* $\theta_1|\mathbf{y}$ *برابر* $N_n(Bb,B)$ *است که در آن*

$$B^{-1} = A_1^TC_1^{-1}A_1 + C_2^{-1} + C_3^{-1}$$

و

$$b = A_1^TC_1^{-1}\mathbf{y} + C_2^{-1}A_2\theta_2 + C_3^{-1}A_3\theta_3.$$

برهان. با استفاده از قضیه بیز داریم

$$\pi(\theta_1|y) \propto p(y|\theta_1)\pi(\theta_1|\theta_2)\pi(\theta_1|\theta_3).$$

حاصل ضرب طرف راست رابطه فوق $\exp(-Q)$ است که در آن

$$\begin{aligned} Q &= (\mathbf{y} - A_1\theta_1)^TC_1^{-1}(\mathbf{y} - A_1\theta_1) + (\theta_1 - A_2\theta_2)^TC_2^{-1}(\theta_1 - A_2\theta_2) + \\ &\quad (\theta_1 - A_3\theta_3)^TC_3^{-1}(\theta_1 - A_3\theta_3) \\ &\propto \theta_1^TB^{-1}\theta_1 - 2b^T\theta_1 \\ &\propto (\theta_1 - Bb)^TB^{-1}(\theta_1 - Bb). \end{aligned}$$

بنابراین برهان تکمیل می‌شود. □

لم ۸.۴.۱ (ماردیا و همکاران، ۱۹۷۹). *اگر* $X = (X_1^T, X_2^T)^T \sim N_p(\mu, \Sigma)$، *آنگاه توزیع شرطی* $X_2|X_1$ *برابر است با*

$$X_2|X_1 \sim N_p(\mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1), \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12})$$

که در آن Σ_{11} *ماتریس کوواریانس متغیر* X_1، Σ_{22} *ماتریس کوواریانس متغیر* X_2 *و* Σ_{12} *ماتریس کوواریانس بین متغیرهای* X_1 *و* X_2 *است.*

۳.۴.۱ ضرب کرونکر

تعریف ۹.۴.۱ (ضرب کرونکر). ضرب کرونکر[۶۹] دو ماتریس A و B را با نماد $A \otimes B$ نشان می‌دهیم و به صورت ضرب هر درایه ماتریس A در همه ماتریس B است. به‌عبارتی اگر

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

و B یک ماتریس دلخواه باشد، آنگاه

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

۴.۴.۱ تولید نمونه از توزیع نرمال چندمتغیره

برای تولید یک نمونه n تایی از $N_d(\mu, \Sigma)$ باید

۱. یک ماتریس $n \times d$، Z، شامل nd متغیر نرمال استاندارد تولید کنیم.

۲. تجزیه $\Sigma = Q^T Q$ را انجام دهیم.

۳. تبدیل $X = ZQ + J\mu^T$ را اجرا کنیم، که در آن J یک بردار n بعدی از ۱ها است.

با اجرای این سه مرحله، سطرهای ماتریس X با بعد $n \times d$، نمونه‌های یک نرمال d متغیره است. دقت کنید در مرحله ۲ باید ماتریس Σ تجزیه شود. روش‌های مختلفی برای این تجزیه معرفی شده‌اند که می‌توان به روش‌های تجزیه چولسکی[۷۰] (گلوب و وان لون، ۱۹۹۶)، طیفی[۷۱] (استرانگ، ۱۹۹۸)

[۶۹] Kronecker product

[۷۰] Cholesky decomposition

[۷۱] Spectral decomposition

و مقدار ویژه[۷۲] (استوارت، ۱۹۹۸) اشاره کرد. با توجه به کاربرد زیاد تجزیه چولسکی، در ادامه آن را معرفی می‌کنیم.

تجزیه چولسکی

تجزیه چولسکی یک ماتریس متقارن همیشه مثبت به صورت $X = Q^T Q$ است، که در آن Q یک ماتریس بالا مثلثی است. تابع $chol(X)$ در R این تجزیه را انجام می‌دهد و خروجی آن یک ماتریس بالا مثلثی یعنی Q است به طوری که $Q^T Q = \Sigma$.

۵.۴.۱ معادله مثلثی و تعداد عملیات آن

فرض کنید A یک ماتریس پایین مثلثی $n \times n$، و $\mathbf{b}$ و $\mathbf{x}$ دو بردار n بعدی نامعلوم باشند، معادله زیر را در نظر بگیرید:

$$A\mathbf{x} = \mathbf{b}$$

که به‌طور معادل می‌توان نوشت

$$\begin{pmatrix} a_{۱,۱} & ۰ & \cdots & ۰ \\ a_{۲,۱} & a_{۲,۲} & \cdots & ۰ \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,۱} & a_{n,۲} & \cdots & a_{n,n} \end{pmatrix} \begin{pmatrix} x_۱ \\ x_۲ \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_۱ \\ b_۲ \\ \vdots \\ b_n \end{pmatrix}.$$

بردار $\mathbf{x}$ از حل معادله‌های زیر به‌دست می‌آید:

$$\begin{aligned} x_۱ &= b_۱/a_{۱,۱} \\ x_۲ &= (b_۲ - a_{۲,۱}x_۱)/a_{۲,۲} \\ x_۳ &= (b_۳ - a_{۳,۱}x_۱ - a_{۳,۲}x_۲)/a_{۳,۳} \\ &\vdots \\ x_n &= (b_n - a_{n,۱}x_۱ - a_{n,۲}x_۲ - \ldots - a_{n,n-۱}x_{n-۱})/a_{n,n}. \end{aligned}$$

همان‌طور که در بالا مشخص است، $x_۱$ فقط یک تقسیم نیاز دارد، $x_۲$ یک ضرب، یک جمع و یک تقسیم، $x_۳$ دو ضرب، دو جمع و یک تقسیم و به همین ترتیب x_n، n-1 ضرب، n-1 جمع و یک تقسیم برای محاسبه لازم دارند. بنابراین تعداد کل جمع‌ها

$$\sum_{i=۱}^{n-۱} i = \frac{n(n-۱)}{۲},$$

تعداد کل ضرب‌ها

$$\sum_{i=۱}^{n-۱} i = \frac{n(n-۱)}{۲},$$

[۷۲] Singular-value decomposition

و تعداد تقسیم‌ها n تاست. پس تعداد کل عملیات برابر با

$$\frac{n(n-۱)}{۲} + \frac{n(n-۱)}{۲} + n = n^۲$$

می‌باشد.

فصل ۲

مدل‌های خطی (گاوسی) فضایی

با انگیزه تحلیل داده‌های کیفیت آب‌های زیرزمینی استان گلستان و با توجه به شایع بودن حضور پاسخ‌های پیوسته در کاربردهای متعدد، در این پایان‌نامه تمرکز ما بر روی مدل‌های خطی است. از این رو، در این فصل به استنباط بیزی مدل‌های خطی فضایی می‌پردازیم. چارچوب کلی این مدل‌ها به صورت

$$\mathbf{y} = X\beta + Z\alpha + \epsilon \qquad (۱.۲)$$

است که در آن X ماتریس طرح $n \times p$ متناظر با بردار ضرایب رگرسیونی p بعدی β است و $Z = Z(\theta)$ ماتریس طرح $n \times r$ متناظر با بردار اثرات فضایی r بعدی α است. جمله ϵ مؤلفه خطاست که هم‌بعد با بردار مشاهدات n بعدی $\mathbf{y}$ است. همچنین θ بردار پارامترهای فرآیند فضایی α با بعد q است که به پارامترهای وابستگی معروف هستند.

در این پایان‌نامه، در مواردی که امکان برداشت نادرست وجود دارد، ماتریس‌ها را با حروف بزرگ لاتین، بردارها را با حروف کوچک لاتین مشکی و مقادیر عددی را با حروف کوچک لاتین نمایش می‌دهیم.

در حالتی که پذیره نرمال (گاوسی) بودن متغیر پاسخ را در نظر بگیریم، مدل (۱.۲) به شکل

$$\mathbf{y} \sim N_n(X\beta + Z(\theta)\alpha, D(\theta))$$

نتیجه می‌شود که در آن فرآیند خطا به صورت

$$\epsilon \sim N_n(\circ, D(\theta))$$

فرض شده، که در آن $D(\theta)$ ماتریس کوواریانس قطری است.

برای ورود به استنباط بیزی باید برای پارامترهای مدل (۱.۲)، $\zeta = \{\alpha, \beta, \theta\}$، توزیع پیشین تعریف کنیم. با فرض مستقل بودن مؤلفه‌های ζ داریم

$$\pi(\zeta) = \pi(\alpha)\pi(\beta)\pi(\theta).$$

معمولاً برای پارامترهای رگرسیونی β، توزیع پیشین نرمال در نظر گرفته می‌شود (بنرجی و همکاران، ۲۰۰۸). یعنی

$$\beta \sim N_p(\mu_\beta, \Sigma_\beta).$$

که در آن μ_β و Σ_β به ترتیب بردار میانگین و ماتریس کوواریانس معلوم هستند. در این پایان‌نامه، توزیع پیشین برای فرآیند فضایی α را نیز نرمال در نظر می‌گیریم. یعنی

$$\alpha \sim N_r(\circ, K(\theta)).$$

البته برخی از محققین از توزیع‌های دیگری نیز برای فرآیند فضایی استفاده کرده‌اند. به عنوان نمونه می‌توان به حسینی و همکاران (۲۰۱۱) و حسینی و محمدزاده (۲۰۱۲) اشاره کرد که از توزیع چوله نرمال برای میدان تصادفی فضایی استفاده کردند. برای پارامترهای وابستگی θ نیز، در متون مختلف، توزیع‌های پیشین مختلفی در نظر گرفته شده‌اند؛ معمولاً از توزیع‌های گامای معکوس برای آن‌ها استفاده می‌شود. همچنین با در نظر گرفتن دیدگاه بیز سلسله‌مراتبی (بنرجی و همکاران، ۲۰۰۴)، می‌توان برای ماتریس کوواریانس فرآیند فضایی، $K(\theta)$، نیز توزیع پیشین ماتریسی تعریف کرد که معمولاً از توزیع ویشارت معکوس استفاده می‌شود.

با تعریف توزیع‌های پیشین ذکرشده، توزیع پسین به صورت زیر نتیجه می‌شود:

$$\pi(\zeta|\mathbf{y}, X, Z) \propto \pi(\theta) \times N(\beta|\mu_\beta, \Sigma_\beta) \times N(\alpha|\circ, K(\theta)) \times N(\mathbf{y}|X\beta + Z(\theta)\alpha, D(\theta)) \quad (۲.۲)$$

که در آن $N(\mathbf{y}|\mu, \Sigma)$ تابع چگالی توزیع $N(\mu, \Sigma)$ است و $\pi(\theta)$ تابع چگالی احتمال توزیع پیشین پارامترهای θ است. شکل توزیع پسین (۲.۲) پیچیده است و صورت بسته ندارد. بنابراین برای انجام استنباط‌های بیزی مبتنی بر آن باید به‌نحوی این توزیع را تقریب زد. روش معمول، در این موارد، استفاده از روش‌های نمونه‌گیری MCMC است. در بخش بعدی نحوه اجرای الگوریتم‌های MCMC در این مدل را توضیح می‌دهیم.

۱.۲ نمونه‌گیری از توزیع پسین پارامترهای فرآیند خطی فضایی

برای تولید نمونه از توزیع پسین (۲.۲)، از روش نمونه‌گیری گیبز استفاده می‌کنیم. برای این کار ابتدا باید توزیع‌های شرطی کامل پارامترها را محاسبه کنیم. در واقع تولید نمونه در سه مرحله صورت می‌پذیرد:

۱. تولید نمونه از توزیع شرطی کامل $\theta|\mathbf{y}, \alpha, \beta$

۲. تولید نمونه از توزیع شرطی کامل $\beta|\mathbf{y}, \theta, \alpha$

۳. تولید نمونه از توزیع شرطی کامل $\alpha|\mathbf{y}, \theta, \beta$

با توجه به توزیع‌های پیشینی که روی پارامترهای مدل تعریف می‌شوند، ممکن است اجرای برخی از این مراحل نیاز به الگوریتم MH داشته باشد. در این صورت باید از یک الگوریتم ترکیبی معروف به الگوریتم MH درون گیبز[۱] استفاده کنیم. در ادامه هر مرحله را به تفکیک تشریح می‌کنیم.

[۱] Metropolis-Hastings-within-Gibbs

۱.۱.۲ تولید نمونه از پارامترهای وابستگی

برای نمونه‌گیری از بردار θ، به منظور سرعت بخشیدن به نرخ همگرایی، مؤلفه‌های α و β را با انتگرال‌گیری خارج می‌کنیم. یعنی تحقق‌های θ را از توزیع پسین کناری

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta) \times N(\mathbf{y}|X\mu_\beta, \Sigma_{\mathbf{y}|\theta}) \qquad (۳.۲)$$

تولید می‌کنیم، که در آن

$$\Sigma_{\mathbf{y}|\theta} = X\Sigma_\beta X^T + Z(\theta)K(\theta)Z(\theta)^T + D(\theta).$$

برای به‌روز رسانی θ (تولید نمونه)، باید در هر تکرار نمونه‌گیری ماتریس $\Sigma_{\mathbf{y}|\theta}$ محاسبه شود. معمولاً $D(\theta)$ ماتریس قطری و $X\Sigma_\beta X^T$ ثابت و مشخص است. بنابراین حجم محاسبات به ماتریس $Z(\theta)K(\theta)Z(\theta)^T$ معطوف می‌شود. برای حجم محاسبات لازم در محاسبه $Z(\theta)K(\theta)Z(\theta)^T$ ضرب ماتریس $Z(\theta)K(\theta)$ را به‌صورت زیر در نظر بگیرید:

$$\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1r} \\ z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nr} \end{pmatrix} \begin{pmatrix} k_{11} & k_{12} & \cdots & k_{1r} \\ k_{21} & k_{22} & \cdots & k_{2r} \\ \vdots & & \ddots & \vdots \\ k_{r1} & k_{r2} & \cdots & k_{rr} \end{pmatrix} = \begin{pmatrix} \bigcirc_{11} & \bigcirc_{12} & \cdots & \bigcirc_{1r} \\ \vdots & \cdots & \ddots & \vdots \\ \bigcirc_{n1} & \cdots & \bigcirc_{n(r-1)} & \bigcirc_{nr} \end{pmatrix}.$$

برای محاسبه مثلاً درآیه اول ماتریس حاصل که با $\bigcirc_{11}$ نشان داده شده است، سطر اول ماتریس $Z(\theta)$ در ستون اول $K(\theta)$ ضرب می‌شود، یعنی

$$\bigcirc_{11} = z_{11}k_{11} + z_{12}k_{21} + \ldots + z_{1r}k_{r1}.$$

بنابراین r ضرب و $r-1$ جمع برای هر درآیه لازم است و با توجه به nr درآیه موجود، تعداد عملیات برای آن $nr(2r-1)$ می‌باشد. در ضرب ماتریس حاصل از $Z(\theta)K(\theta)$ با ماتریس $Z(\theta)^T$، همانند بالا، داریم

$$\begin{pmatrix} \bigcirc_{11} & \cdots & \bigcirc_{1r} \\ \vdots & \ddots & \vdots \\ \bigcirc_{n1} & \cdots & \bigcirc_{nr} \end{pmatrix} \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{n1} \\ z_{12} & z_{22} & \cdots & z_{n2} \\ \vdots & & \ddots & \vdots \\ z_{1r} & z_{2r} & \cdots & z_{nr} \end{pmatrix} = \begin{pmatrix} \triangle_{11} & \cdots & \triangle_{n1} \\ \vdots & \ddots & \vdots \\ \triangle_{n1} & \cdots & \triangle_{nn} \end{pmatrix}.$$

با در نظر گرفتن درآیه اول ماتریس سمت راست، یعنی $\triangle_{11}$، داریم

$$\triangle_{11} = \bigcirc_{11}z_{11} + \bigcirc_{12}z_{12} + \ldots + \bigcirc_{1r}z_{1r}.$$

بنابراین r ضرب و $r-1$ جمع برای هر درآیه لازم است و با توجه به n^2 درآیه موجود، تعداد عملیات برای آن $n^2(2r-1)$ می‌باشد. در نتیجه تعداد عملیات مورد نیاز برای محاسبه $Z(\theta)K(\theta)Z(\theta)^T$ برابر با $(n^2+nr)(2r-1)$ است. به عبارت دیگر برای محاسبه $Z(\theta)K(\theta)Z(\theta)^T$، $O(rn^2)$ عملیات نیاز است.

در محاسبه رابطه (۳.۲)، با لگاریتم گرفتن از طرفین آن، داریم

$$\log \pi(\theta|\mathbf{y}) = const. + \log \pi(\theta) - \frac{۱}{۲}\log|\Sigma_{\mathbf{y}|\theta}| - \frac{۱}{۲}Q(\theta) \qquad (۴.۲)$$

که در آن

$$Q(\theta) = (\mathbf{y} - X\mu_\beta)^T \Sigma^{-۱}_{\mathbf{y}|\theta}(\mathbf{y} - X\mu_\beta).$$

در اولین مرحله نمونه‌گیری گیبز برای تولید نمونه از $\pi(\theta|\mathbf{y})$ یک الگوریتم RWMH با توزیع پیشنهادی نرمال چندمتغیره (هم‌بعد با θ) در نظر گرفته می‌شود. برای اجرای الگوریتم باید رابطه (۴.۲) محاسبه شود. محاسبه $Q(\theta)$ نیازمند محاسبه معکوس $\Sigma_{\mathbf{y}|\theta}$ است. پس باید ماتریس $\Sigma_{\mathbf{y}|\theta}$ تجزیه شود. اگر تجزیه چولسکی را به‌صورت $L = chol(\Sigma_{\mathbf{y}|\theta})$ در نظر بگیریم، تعداد عملیات برای محاسبه L که یک ماتریس بالا مثلثی است برابر $O(n^۳/۳)$ است (هانگر، ۲۰۰۵). در نتیجه برای محاسبه Q باید، با توجه به تعریف تابع trsolve در فصل اول، معادله $\mathbf{u} = trsolve(L, \mathbf{y} - X\mu_\beta)$ را حل کنیم، زیرا

$$Q(\theta) = \mathbf{u}^T\mathbf{u}.$$

تعداد عملیات حل این معادله برابر $O(n^۲)$ است (فصل اول را ببینید). از آن‌جا که $\mathbf{u}$ یک بردار n بعدی است و $\mathbf{u}^T\mathbf{u}$ به صورت

$$\begin{pmatrix} u_۱ & u_۲ & \dots & u_n \end{pmatrix} \begin{pmatrix} u_۱ \\ u_۲ \\ \dots \\ u_n \end{pmatrix} = u_۱^۲ + u_۲^۲ + \ldots + u_n^۲$$

است، $n-۱$ تا جمع، n تا ضرب و در کل $۲n-۱$ عملیات دیگر لازم است. همچنین

$$\log|\Sigma_{\mathbf{y}|\theta}| = ۲\sum_{i=۱}^{n}\log \ell_{ii}$$

که ℓ_{ii} عناصر روی قطر ماتریس L است، که قبلاً آن را به‌دست آورده‌ایم. پس تعداد عملیات محاسبه $\log|\Sigma_{\mathbf{y}|\theta}|$ برابر است با

$$\sum_{i=۱}^{n}\log \ell_{ii} = \log \ell_{۱۱} + \log \ell_{۲۲} + \ldots + \log \ell_{nn} = \log(\ell_{۱۱}\ell_{۲۲}\ldots\ell_{nn}).$$

بنابراین به $n-۱$ ضرب و یک لگاریتم نیازمندیم که در جمع n عملیات می‌شود. در نهایت، با توجه به غالب بودن مرتبه $n^۳$، تعداد کل عملیات برای محاسبه رابطه (۴.۲) از مرتبه $O(n^۳)$ می‌باشد.

بنابراین قابل تصور است که اگر n (حجم مشاهدات) بزرگ باشد، حجم عملیات لازم برای محاسبه توزیع پسین کناری θ چقدر می‌تواند بزرگ باشد. دقت کنید این حجم محاسبات حتی با پرهیز از معکوس کردن مستقیم ماتریس $\Sigma_{\mathbf{y}|\theta}$ با کمک تجزیه چولسکی و معادله‌های مثلثی است. به‌عنوان مثال اگر n ۱۰۰ باشد، این راه‌کار از نظر محاسباتی عملی است. در غیر این صورت، زمانی که n بزرگ باشد، محاسبات سخت و طاقت‌فرسا خواهد بود و نیاز به روش‌های جایگزین داریم.

حال اگر توزیع پیشین برای β تخت باشد، یعنی $\Sigma_\beta^{-۱} = ۰$، شبیه به رابطه (۴.۲) داریم

$$\log \pi(\theta|\mathbf{y}) = const. + \log \pi(\theta) - \frac{۱}{۲}\log|X^T\Sigma_{\mathbf{y}|\beta,\theta}X| - \frac{۱}{۲}\log|\Sigma_{\mathbf{y}|\beta,\theta}| - \frac{۱}{۲}Q(\theta) \quad (۵.۲)$$

که در آن

$$\Sigma_{\mathbf{y}|\beta,\theta} = Z(\theta)K(\theta)Z(\theta)^T + D(\theta) \qquad (۶.۲)$$

و

$$Q(\theta) = \mathbf{y}^T\Sigma^{-1}_{\mathbf{y}|\beta,\theta}\mathbf{y} - \mathbf{b}^T(X^T\Sigma^{-1}_{\mathbf{y}|\beta,\theta}X)^{-1}\mathbf{b}, \quad \mathbf{b} = X^T\Sigma^{-1}_{\mathbf{y}|\beta,\theta}\mathbf{y}.$$

محاسبات لازم شبیه به حالت قبلی می‌باشد، یعنی ابتدا

$$L = chol(\Sigma_{\mathbf{y}|\beta,\theta})$$

و

$$[\mathbf{v} : U] = trsolve(L, [\mathbf{y} : X])$$

را محاسبه می‌کنیم، سپس برای به دست آوردن $X^T\Sigma^{-1}_{\mathbf{y}|\beta,\theta}X$ باید معادله‌های

$$W = chol(U^TU),\ \mathbf{b} = U^T\mathbf{v}$$

و

$$\tilde{\mathbf{b}} = trsolve(w, \mathbf{b})$$

را حل کنیم. سرانجام رابطه (۵.۲) به شکل زیر حاصل می‌شود:

$$\log \pi(\theta) = \sum_{i=1}^{p} \log w_{i,i} - \sum_{i=1}^{n} l_{i,i} - \frac{1}{2}(\mathbf{v}^T\mathbf{v} - \tilde{\mathbf{b}}^T\tilde{\mathbf{b}})$$

که در آن $w_{i,i}$ و $l_{i,i}$ به ترتیب عناصر روی قطر ماتریس W و L هستند. تعداد عملیات در این حالت نیز $O(n^3)$ می‌باشد.

۲.۱.۲ نمونه‌گیری از بردار ضرایب رگرسیونی و اثرات تصادفی

در بخش قبل نحوه تولید نمونه از پسین کناری θ را تشریح کردیم. با اجرای مرحله اول می‌توان نمونه‌های α و β را با استفاده از یک نمونه‌گیری ترکیبی به شرح زیر به‌دست آورد. فرض کنید $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$، M نمونه از $\pi(\theta|\mathbf{y})$ باشد. برای هر $k = 1, ..., M$ نمونه‌های β و α را به ترتیب از توزیع‌های شرطی کامل $\beta^{(k)} \sim \pi(\beta|\theta^{(k)}, \mathbf{y})$ و $\alpha^{(k)} \sim \pi(\alpha|\theta^{(k)}, \mathbf{y})$ تولید می‌کنیم.

نمونه‌گیری از بردار ضرایب رگرسیونی

ابتدا توزیع شرطی کامل $\pi(\beta|\theta^{(k)}, \mathbf{y})$ را محاسبه می‌کنیم. با توجه به برقراری شرایط لم ۶.۴.۱ به صورت

$$\mathbf{y}|\theta_1 \sim N(A_1\theta_1, C_1) \equiv \mathbf{y}|\beta, \theta \sim N_p(X\beta, \Sigma_{\mathbf{y}|\beta,\theta})$$

و

$$\theta_1|\theta_2 \sim N(A_2\theta_2, C_2) \equiv \beta \sim N_p(\mu_\beta, \Sigma_\beta)$$

که در آن $A_1 = X$، $C_1 = \Sigma_{\mathbf{y}|\beta,\theta}$، $A_2\theta_2 = \mu_\beta$ و $C_2 = \Sigma_\beta$، بنا بر حکم ب لم ۶.۴.۱ داریم

$$\beta|\theta, \mathbf{y} \sim N_p(B\mathbf{b}, B)$$

که در آن

$$\mathbf{b} = C_۲^{-۱} A_۲ \theta_۲ + A_۱^T C_۱^{-۱} \mathbf{y} = \Sigma_\beta^{-۱} \mu_\beta + X^T \Sigma_{\mathbf{y}|\beta,\theta}^{-۱} \mathbf{y}$$

و

$$B^{-۱} = A_۱^T C_۱^{-۱} A_۱ + C_۲^{-۱} = \Sigma_\beta^{-۱} + X^T \Sigma_{\mathbf{y}|\beta,\theta}^{-۱} X.$$

پس توزیع شرطی کامل β نرمال با بردار میانگین $B\mathbf{b}$ و ماتریس کوواریانس B می‌باشد که در آن

$$\mathbf{b} = \Sigma_\beta^{-۱} \mu_\beta + X^T \Sigma_{\mathbf{y}|\beta,\theta}^{-۱} y, \;\; B = (\Sigma_\beta^{-۱} + X^T \Sigma_{\mathbf{y}|\beta,\theta}^{-۱} X)^{-۱}. \qquad (۷.۲)$$

برای تولید نمونه از $\beta^{(k)} \sim N_p(B\mathbf{b}, B)$، برای هر $k = ۱, ..., M$، باید $\mathbf{b}$ و B را به ازای هر $\theta^{(k)}$ محاسبه کنیم. برای نیل به این مقصد داریم

$$\mathbf{b} = \Sigma_\beta^{-۱} \mu_\beta + U^T \mathbf{v}$$

که در آن

$$[\mathbf{v} : U] = trsolve(L, [\mathbf{y} : X])$$

و

$$L = chol(\Sigma_{\mathbf{y}|\beta,\theta^{(k)}}).$$

سپس تجزیه چولسکی

$$L_B = chol(\Sigma_\beta^{-۱} + U^T U)$$

را انجام داده و با مشخص شدن $\mathbf{b}$ و B، $\beta^{(k)}$ را با استفاده از روش تولید نمونه از نرمال چندمتغیره به صورت زیر تولید می‌کنیم:
p متغیر نرمال استاندارد مستقل تولید می‌کنیم و آن p تا را در بردار $\mathbf{z}$ ذخیره می‌کنیم. سپس نمونه‌های $\{\beta^{(۱)}, \beta^{(۲)}, \ldots, \beta^{(M)}\}$ از $\pi(\beta|\mathbf{y})$ با رابطه

$$\beta^{(k)} = \mathbf{z}L + (L_B^T L_B)^{-۱} \mathbf{b}$$

به‌دست می‌آیند.

نمونه‌گیری از اثرات تصادفی

مشابه مرحله دوم، ابتدا توزیع شرطی کامل $\pi(\alpha|\theta^{(k)}, \mathbf{y})$ را محاسبه می‌کنیم. با توجه به برقراری شرایط لم ۶.۴.۱ به صورت

$$\mathbf{y}|\theta_۱ \sim N(B_۱\theta_۱, D_۱) \equiv (\mathbf{y} - X\mu_\beta)|\alpha, \theta \sim N_r(Z(\theta)\alpha, \Sigma_{\mathbf{y}|\alpha,\theta})$$

و

$$\theta_۱|\theta_۲ \sim N(B_۲\theta_۲, D_۲) \equiv \alpha \sim N_r(۰, K(\theta))$$

که در آن $B_۱ = Z(\theta)$، $D_۱ = \Sigma_{\mathbf{y}|\alpha,\theta}$، $B_۲\theta_۲ = ۰$ و $D_۲ = K(\theta)$. بنا بر حکم ب لم ۶.۴.۱ داریم

$$\alpha|\theta, \mathbf{y} \sim N_r(B_\alpha \mathbf{b}_\alpha, B_\alpha)$$

که در آن

$$\mathbf{b}_\alpha = D_2^{-1} B_2 \theta_2 + B_1^T D_1^{-1} \mathbf{y} = Z(\theta)^T \Sigma_{\mathbf{y}|\alpha,\theta}^{-1} (\mathbf{y} - X\mu_\beta)$$

و

$$B_\alpha^{-1} = B_1^T D_1^{-1} B_1 + D_2^{-1} = Z(\theta)^T \Sigma_{\mathbf{y}|\alpha,\theta}^{-1} Z(\theta) + K(\theta)^{-1}.$$

بنابراین توزیع شرطی کامل α نرمال با میانگین $B_\alpha \mathbf{b}_\alpha$ و ماتریس کوواریانس B_α می‌باشد که در آن

$$\mathbf{b}_\alpha = Z(\theta)^T \Sigma_{\mathbf{y}|\alpha,\theta}^{-1} (\mathbf{y} - X\mu_\beta),\ B_\alpha = \left(K(\theta)^{-1} + Z(\theta)^T \Sigma_{\mathbf{y}|\alpha,\theta}^{-1} Z(\theta) \right)^{-1} \quad (۸.۲)$$

به‌طوری که $\Sigma_{\mathbf{y}|\alpha,\theta} = X\Sigma_\beta X^T + D(\theta)$.

بردار $\mathbf{b}_\alpha$ مشابه β محاسبه می‌شود. کافی است تجزیه چولسکی

$$L = chol(\Sigma_{\mathbf{y}|\alpha,\theta^{(k)}})$$

انجام و معادله

$$[\mathbf{v} : U] = trsolve(L, [\mathbf{y} - X\mu_\beta : Z(\theta^{(k)})])$$

حل شود تا بردار $\mathbf{b}_\alpha$ به صورت $\mathbf{b}_\alpha = U(\theta^{(k)})\mathbf{v}$ نتیجه شود.

ماتریس B_α را می‌توان با عملیات مشابه محاسبه B به دست آورد ولی چون شامل تجزیه چولسکی $K(\theta)$ می‌باشد، ممکن است برای برخی توابع کوواریانس خاص (مثل توابع گاوسی یا ماترن با ν بزرگ) ناپایدار باشد (آرمسترانگ و دود، ۱۹۹۴). برای گذر از این مشکل با تعریف،

$$G(\theta)^{-1} = Z(\theta)^T \Sigma_{\mathbf{y}|\alpha,\theta}^{-1} Z(\theta)$$

و استفاده از فرمول هندرسون و سرل (۱۹۸۱) می‌توان نوشت

$$(K(\theta)^{-1} + G(\theta)^{-1})^{-1} = G(\theta) - G(\theta)^T (K(\theta) + G(\theta))^{-1} G(\theta).$$

با استفاده از تساوی بالا و حل آن، می‌توان به یک الگوریتم پایدار عددی برای محاسبه B_α دست یافت. برای هر $k = 1, ..., M$، معادله‌های

$$L = chol(K(\theta^{(k)}), G(\theta^{(k)})),$$

$$W = trsolve(L, G(\theta^{(k)}))$$

و

$$L_{B_\alpha} = chol(G(\theta^{(k)}) - W^T W)$$

را حل می‌کنیم. حال با معلوم شدن $\mathbf{b}_\alpha$ و B_α، $\alpha^{(k)}$ را با استفاده از روش تولید نمونه از نرمال چندمتغیره به صورت زیر تولید می‌کنیم:

p متغیر نرمال استاندارد مستقل تولید می‌کنیم و آن‌ها را در بردار $\mathbf{z}$ ذخیره می‌کنیم. سپس نمونه‌های $\{\alpha^{(1)}, \alpha^{(2)}, ..., \alpha^{(M)}\}$ را از توزیع $\pi(\alpha|\mathbf{y})$ با استفاده از رابطه

$$\alpha^{(k)} = L_{B_\alpha} \mathbf{z} + L_{B_\alpha}^T L_{B_\alpha} \mathbf{b}_\alpha$$

تولید می‌کنیم. توجه کنید که برآورد اثرات فضایی نیاز به انجام تجزیه چولسکی یک ماتریس همیشه مثبت $n \times n$ دارد. گام‌های بالا پایداری عددی را تثبیت می‌کنند اما اجرای آن‌ها از نظر محاسباتی می‌تواند سنگین باشد، به‌ویژه زمانی که n بزرگ باشد.

۲.۲ مدل‌های کم‌رتبه

مشکل محاسباتی اصلی در برآورد (۲.۲)، تجزیه یک ماتریس همیشه مثبت $n \times n$ است. تعداد عملیات از مرتبه $O(n^۳)$ است و برای هر تکرار بیشتر، این تعداد عملیات به کل زمان اجرای الگوریتم افزوده می‌شود. به عنوان مثال، فرض کنید $n = ۲۰۰۰$ مکان و $p = ۲$ متغیر پیشگو اندازه‌گیری شده باشند. اگر فرض شود هر تکرار MCMC، ۰/۳ ثانیه زمان CPU را در بر می‌گیرد و چنان‌چه ۱۰۰۰۰ تکرار برای همگرایی و استنباط کامل نیاز داشته باشیم، زمان مورد نیاز تقریباً ۵۰ دقیقه خواهد بود. واضح است که مجموعه داده‌های فضایی حجیم، مدل‌های خاص خود را برای تحلیل نیاز دارد. یک رده کارا برای داده‌های بزرگ، رده مدل‌های کم‌رتبه است (هایدون، ۲۰۰۲؛ کامن و واند، ۲۰۰۳؛ استاین، ۲۰۰۷، ۲۰۰۸؛ کرسی و یوهانسون، ۲۰۰۸؛ بنرجی و همکاران، ۲۰۰۸؛ کراینایکینو و همکاران؛ ۲۰۰۸).

در این رده، راه‌کار انتخاب $Z(\theta)$ با r های (بسیار) کوچکتر از n است که گزینش این انتخاب‌ها برای $Z(\theta)$ در فصل سوم، با معرفی مدل‌های فرآیند پیش‌گو، بررسی می‌شود. در این بخش، حالت خاصی از مدل‌های کم‌رتبه را معرفی می‌کنیم و کاهش زمان محاسبات نتیجه‌شده از آن در مقابل مدل کامل را تشریح می‌کنیم.

توزیع پسین مدل سلسله مراتبی (۲.۲) را بدون بردار α، در نظر بگیرید:

$$\pi(\beta, \theta|\mathbf{y}) \propto \pi(\theta) \times N(\beta|\mu_\beta, \Sigma_\beta) \times N(\mathbf{y}|X\beta, \Sigma_{\mathbf{y}|\beta,\theta})$$

که در آن $\Sigma_{\mathbf{y}|\beta,\theta}$ در رابطه (۶.۲) تعریف شده است. نمونه‌های β از توزیع نرمال چندمتغیره با میانگین $B\mathbf{b}$ و ماتریس کوواریانس B، که شبیه (۷.۲) هستند، تولید می‌شوند. اجرای الگوریتم نمونه‌گیری گیبز در بخش قبلی، برای n های بزرگ، هزینه‌بر است. زیرا همان‌طور که توضیح داده شد محاسبه ماتریس B با بعد p، شامل تجزیه چولسکی یک ماتریس $n \times n$ برای هر θ جدید است. رهیافت جانشین با هزینه محاسباتی کمتر، محاسبه $\Sigma_{\mathbf{y}|\beta,\theta}$ با استفاده از فرمول موریسون- وودباری- شرمن[۲] (دنگ، ۲۰۱۱)، به‌صورت زیر است:

$$\begin{aligned}\Sigma_{\mathbf{y}|\beta,\theta} &= D(\theta)^{-۱} - D(\theta)^{-۱}Z(\theta)\left(K(\theta)^{-۱} Z(\theta)^T D(\theta)^{-۱} Z(\theta)\right) Z(\theta)^T D(\theta)^{-۱} \\ &= D(\theta)^{-\frac{۱}{۲}}\left(I - H^T H\right) D(\theta)^{-\frac{۱}{۲}} \qquad (۹.۲)\end{aligned}$$

که در آن $H = trsolve(L, W^T)$، $W = D(\theta)^{-\frac{۱}{۲}} Z(\theta)$ و $L = chol\left(K(\theta)^{-۱} + W^T W\right)$. سپس دو معادله

$$[\mathbf{v} : V] = D(\theta)^{-\frac{۱}{۲}} Z(\theta)$$

و

$$\tilde{V} = HV$$

حل می‌شوند تا $\mathbf{b}$ و L_B به شکل زیر محاسبه شوند:

$$\mathbf{b} = \Sigma_\beta^{-۱}\mu_\beta + V^T\mathbf{y} - \tilde{V}^T H\mathbf{v},\ L_B = chol\left(\Sigma_\beta^{-۱} + V^T V - \tilde{V}^T\tilde{V}\right). \qquad (۱۰.۲)$$

[۲] Sherman-Woodbury-Morrison formula

بنابراین برای هر تکرار در نمونه‌گیری گیبز و هر θ جدید، نمونه‌های β از توزیع نرمال با مشخصات (۱۰.۲) تولید می‌شوند. پارامترهای بردار θ را نیز با استفاده از گام‌های RWMH به‌روز رسانی می‌کنیم. لگاریتم تابع چگالی پسین کناری θ را در نظر بگیرید:

$$\log \pi(\theta|\mathbf{y}) = const. + \log \pi(\theta) - \frac{۱}{۲} \log |\Sigma_{\mathbf{y}|\beta,\theta}| - \frac{۱}{۲} Q(\theta) \qquad (۱۱.۲)$$

که در آن

$$Q(\theta) = (\mathbf{y} - X\beta)^T \Sigma^{-۱}_{\mathbf{y}|\beta,\theta} (\mathbf{y} - X\beta).$$

در این حالت، ماتریس H مشابه ماتریس متناظر در رابطه (۹.۲) تعریف می‌شود که جزییات محاسبه آن را بیان کرده‌ایم. بنابراین

$$\mathbf{v} = D^{-\frac{۱}{۲}}(\mathbf{y} - X\beta),$$

$$w = H\mathbf{v}$$

و

$$T = chol(I_r - HH^T).$$

بنابراین تابع چگالی پسین کناری (۱۱.۲) به صورت زیر قابل بازنویسی است:

$$\log \pi(\theta) - \frac{۱}{۲} \sum_{i=۱}^{n} \log d_{i,i}(\theta) + \sum_{i=۱}^{n} \log t_{i,i}(\theta) - \frac{۱}{۲} (\mathbf{v}^T \mathbf{v} - w^T w)$$

که در آن $d_{i,i}(\theta)$ و $t_{i,i}(\theta)$ به ترتیب عناصر روی قطر ماتریس $D(\theta)$ و T هستند.

وقتی الگوریتم نمونه‌گیری گیبز همگرا و نمونه‌های پسین β و θ به‌دست آمدند، نمونه‌های پسین α را می‌توان با روش مشابه روش اجرایی در بخش قبل، تولید کرد. در واقع، از آن‌جا که نمونه‌های β و θ در دسترس هستند، نمونه‌های α را می‌توان از توزیع کامل شرطی متناظرش شبیه‌سازی کرد. با جایگزینی μ_β با β و $\Sigma_{\mathbf{y}|\alpha,\theta}$ با $D(\theta)$ در رابطه (۸.۲) و اجرای مرحله گیبز مربوط به α نمونه‌های α تولید خواهند شد. چون محاسبه معکوس ماتریس قطری $D(\theta)$ نسبت به $\Sigma_{\mathbf{y}|\alpha,\theta}$ خیلی راحت‌تر و سریع‌تر است، هزینه محاسباتی در این مدل به‌طور قابل ملاحظه‌ای کاهش می‌یابد.

۳.۲ پیش‌گویی فضایی

برای پیش‌گویی بردار تصادفی $\mathbf{y}_۰$، با بعد t، و ماتریس طرح $t \times p$ متناظر با آن، $X_۰$، فرض می‌کنیم

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_۰ \end{bmatrix} |\beta, \theta \sim N_{t+n} \left(\begin{bmatrix} X \\ X_۰ \end{bmatrix} \beta \, , \begin{pmatrix} C_{۱۱}(\theta) & C_{۱۲}(\theta) \\ C_{۱۲}(\theta)^T & C_{۲۲}(\theta) \end{pmatrix} \right) \qquad (۱۲.۲)$$

که در آن $C_{۱۱}(\theta) = \Sigma_{\mathbf{y}|\beta,\theta}$ ماتریس کوواریانس $\mathbf{y}$، $C_{۱۲}(\theta)$ ماتریس کوواریانس $n \times t$ بین $\mathbf{y}$ و $\mathbf{y}_۰$، و $C_{۲۲}(\theta)$ ماتریس کوواریانس $\mathbf{y}_۰$ هستند. همان‌طور که قبلا اشاره کردیم، برای پیش‌گویی مقادیر جدید $\mathbf{y}_۰$ باید ابتدا توزیع شرطی $\mathbf{y}_۰|\mathbf{y}, \beta, \theta$ را به دست آوریم. با توجه به این که فرض لم ۸.۴.۱ برای توزیع

(۱۲.۲) برقرار است، برای به دست آوردن توزیع شرطی $\mathbf{y}_\circ|\mathbf{y},\beta,\theta$ حکم لم را پیاده می‌کنیم. میانگین توزیع شرطی برابر است با

$$\mu_p = \mu_{\mathbf{y}_\circ} + \Sigma^T_{\mathbf{y},\mathbf{y}_\circ}\Sigma^{-1}_{\mathbf{y}}(\mathbf{y}-\mu_{\mathbf{y}}) = X_\circ\beta + C_{12}(\theta)^T C_{11}(\theta)^{-1}(\mathbf{y}-X\beta)$$

و ماتریس کوواریانس آن به صورت

$$\Sigma_p = \Sigma_{\mathbf{y}_\circ} - \Sigma^T_{\mathbf{y},\mathbf{y}_\circ}\Sigma^{-1}_{\mathbf{y}}\Sigma_{\mathbf{y},\mathbf{y}_\circ} = C_{22}(\theta) - C_{12}(\theta)^T C_{11}(\theta)^{-1}C_{12}(\theta)$$

است. بنابراین توزیع شرطی $\mathbf{y}_\circ|\mathbf{y},\beta,\theta$ نرمال با میانگین μ_p و ماتریس واریانس Σ_p است. برای انجام پیش‌گویی بیزی باید توزیع پیش‌گو را به دست آوریم. توزیع پیش‌گو در این حالت عبارتست از

$$f(\mathbf{y}_\circ|\mathbf{y}) = \int N_t(\mathbf{y}_\circ|\mu_p,\Sigma_p)\pi(\beta,\theta|\mathbf{y})d\beta d\theta.$$

مشابه توزیع پسین، این توزیع پیش‌گو که از روی آن ساخته می‌شود نیز پیچیده است و صورت بسته ندارد و باید با روش‌های MCMC تقریب زده شود. با توجه به نمونه‌های حاصل از توزیع پسین پارامترهای مدل، تقریب توزیع پیش‌گو ساده است. برای تولید نمونه از توزیع پیش‌گو کافیست برای هر نمونه پسین $\{\theta,\beta\}$، نمونه‌ای از توزیع $N_p(\mu_p,\Sigma_p)$ که در آن پارامترهای μ_p و Σ_p در نمونه‌های پسین مقداردهی شده‌اند، تولید شود. توجه کنید که تولید نمونه از توزیع پیش‌گو فقط شامل نمونه‌های MCMC بعد از همگرایی است. افزون بر آن، محاسبه μ_p و Σ_p همزمان با به‌روز رسانی پارامترهای مدل انجام می‌گیرد.

برای محاسبه μ_p و Σ_p، ابتدا باید معادله

$$[\mathbf{u}:V] = trsolve\left(L,\left[\mathbf{y}-X\beta^{(k)}:C_{12}(\theta^{(k)})\right]\right)$$

را حل کنیم که در آن $L = chol(C_{11}(\theta^{(k)}))$. سپس داریم

$$\mu_p = X_\circ\beta^{(k)} + V^T\mathbf{u},\ \ \Sigma_p = C_{22}(\theta^{(k)}) - V^TV.$$

مدل‌های با رتبه کم (با $r \ll n$) باعث کاهش هزینه‌های محاسباتی می‌شوند. در این مدل‌ها، قسمت اصلی عملیات محاسبه توزیع پسین به محاسبه

$$C_{12}(\theta)^T C_{11}(\theta)^{-1}C_{12}(\theta)$$

معطوف است که می‌تواند به صورت

$$U^TU - V^TV$$

نوشته شود، به‌طوری که $U = D^{-\frac{1}{2}}C_{12}(\theta)$ و $V = HU$ که H همانند رابطه (۹.۲) تعریف می‌شود. این عملیات از محاسبه مستقیم $C_{11}(\theta)^{-1}$ جلوگیری می‌کند.

به‌روز رسانی $\mathbf{y}_\circ^{(k)}$ نیازمند محاسبه تجزیه چولسکی Σ_p است، که یک ماتریس $t\times t$ بوده و برای t های بزرگ می‌تواند هزینه‌بر باشد. بنابراین، در بسیاری از کاربردها، بهتر است $t=1$ انتخاب شود و پیش‌گویی برای هر مکان به‌طور مستقل انجام شود. اگر از نمونه‌های پسینی α استفاده کنیم، گام‌های پیش‌گویی آسان‌تر هم می‌شود. در این حالت، نمونه‌های توزیع پیش‌گو از توزیع

$$N_p\left(X_\circ\beta^{(k)} + Z(\theta^{(k)})\alpha^{(k)}, D(\theta^{(k)})\right)$$

شبیه‌سازی می‌شوند. دلیل کاهش هزینه‌ها نیز قطری بودن ماتریس $D(\theta)$ است.

فصل ۳

مدل فرآیند پیش‌گوی گاوسی

در این فصل، مدل رگرسیون خطی را برای داده‌های زمین‌آماری در حالتی که متغیر پاسخ $\mathbf{y}(\mathbf{s})$ از توزیع گاوسی پیروی می‌کند، معرفی می‌کنیم. مساله را به تفکیک برای پاسخ‌های یک و چندمتغیره مطرح می‌کنیم. مدل رگرسیونی را برای دو حالت در نظر می‌گیریم: حالتی که همه مشاهدات در برازش مدل نقش داشته باشند (مدل پررتبه) و حالتی که قسمتی از داده‌ها برای برازش مدل انتخاب شوند (مدل کم‌رتبه). برای مدل‌های کم‌رتبه از یک رده خاص با نام مدل‌های فرآیند پیش‌گو (گاوسی) که توسط بنرجی و همکاران (۲۰۰۸) معرفی شد، استفاده می‌کنیم.

۱.۳ مدل خطی یک‌متغیره پررتبه

مدل رگرسیون خطی برای پاسخ $\mathbf{y}(\mathbf{s})$، $\mathbf{s} \in D \subseteq R^2$، به صورت

$$\mathbf{y}(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\beta + w(\mathbf{s}) + \varepsilon(\mathbf{s}) \qquad (۱.۳)$$

تعریف می‌شود، که در آن $\mathbf{x}(\mathbf{s})$ بردار متغیرهای تبیینی p بعدی، β بردار ضرایب رگرسیونی (اثرات ثابت) p بعدی و $w(\mathbf{s})$ فرآیند فضایی (یا همان اثرات تصادفی فضایی) تعریف‌شده در فصل قبل می‌باشد. جمله $\varepsilon(\mathbf{s})$ نیز فرآیند خطای اندازه‌گیری یا تغییرات کوچک‌مقیاس است. همچنین فرض می‌شود $\varepsilon(\mathbf{s}_i)$ ها، $i = 1, \ldots, n$، مستقل و هم‌توزیع با توزیع نرمال $N(0, \tau^2)$ هستند. معمولاً $w(\mathbf{s})$ یک فرآیند گاوسی با میانگین 0 و تابع کوواریانس $C(\mathbf{s}, \mathbf{s}')$ فرض می‌شود که از این به بعد این فرآیند گاوسی را با نماد $GP\{0, C(\mathbf{s}, \mathbf{s}')\}$ نمایش می‌دهیم.

در عمل اغلب تابع کوواریانس به صورت

$$C(\mathbf{s}, \mathbf{s}') = \sigma^2\rho(\mathbf{s}, \mathbf{s}'; \theta)$$

در نظر گرفته می‌شود، که در آن $\rho(\cdot; \theta)$ تابع همبستگی‌نگار و θ بردار q بعدی پارامترهای وابستگی فرآیند فضایی، شامل پارامترهای همواری، واریانس و دامنه، است. پارامتر همواری در تابع همبستگی‌نگار وظیفه کنترل میزان همواری همبستگی بین کمیت‌ها در موقعیت‌های فضایی را بر عهده دارد و پارامتر دامنه نیز نشان‌دهنده میزان کاهش همبستگی با افزایش فاصله است. پارامتر واریانس فرآیند فضایی،

$Var(w(\mathbf{s})) = \sigma^۲$، نیز واریانس فرآیند در هر نقطه، یعنی $C(\mathbf{s})$، را نشان می‌دهد. همان‌طور که در فصل اول عنوان کردیم، از توابع همبستگی‌نگار معتبر مختلف می‌توان استفاده کرد. یک رده منعطف از این توابع، تابع ماترن با ضابطه زیر است:

$$\rho(\|\mathbf{s}-\mathbf{s}'\|;\phi,\nu) = \frac{۱}{۲^{\nu-۱}\Gamma(\nu)} \left(\|\mathbf{s}-\mathbf{s}'\|\phi\right)^{\nu} k_{\nu}\left(\|\mathbf{s}-\mathbf{s}'\|\phi\right); \quad \phi > ۰, \quad \nu > ۰$$

که در آن ϕ پارامتر دامنه، ν پارامتر همواری، $\Gamma(\cdot)$ تابع گاما و $k_{\nu}(\cdot)$ تابع بسل بهبودیافته نوع دوم مرتبه ν است.

مدل سلسله‌مراتبی ساخته‌شده در (۱.۳)، یک عضو خاص از رده مدل‌های (۱.۲) است، که در آن $\mathbf{y}$ بردار n بعدی پاسخ با مولفه‌های $\mathbf{y}(\mathbf{s}_i)$، X ماتریس $n \times p$ با سطرهای $\mathbf{x}(\mathbf{s}_i)^T$، α بردار n بعدی با مولفه‌های $w(\mathbf{s}_i)$، $Z(\theta) = I_n$، $K(\theta)$ ماتریس $n \times n$ با درایه‌های $C(\mathbf{s}_i, \mathbf{s}_j; \theta)$ و $D(\theta) = \tau^۲ I_n$ است. همچنین بردار پارامترهای وابستگی فرآیند در $K(\theta)$ و $D(\theta)$ به صورت $\theta = \{\sigma^۲, \phi, \nu, \tau^۲\}$ تعریف می‌شود.

۲.۳ مدل‌های فرآیند پیش‌گو

مجموعه مکان‌های

$$S^* = \{\mathbf{s}^*_۱, \mathbf{s}^*_۲, .., \mathbf{s}^*_m\}$$

را در نظر بگیرید که ممکن است یک زیرمجموعه از مکان‌های مشاهده‌شده S باشد. این مجموعه را مجموعه گره[۱] می‌نامیم. تعریف می‌کنیم

$$\mathbf{w}^* = (w(\mathbf{s}^*_۱), \ldots, w(\mathbf{s}^*_m)) \sim GP\{۰, C^*(\theta)\}$$

که در آن $C^*(\theta) = \left(C(\mathbf{s}^*_i, \mathbf{s}^*_j; \theta)\right)$، $i, j = ۱, \ldots, m$، ماتریس کوواریانس $m \times m$ متناظر با مجموعه گره است. چون فرآیند تعریف‌شده در (۱.۳) گاوسی است، فرآیند $\mathbf{w}^*$ نیز یک فرآیند گاوسی خواهد بود. پیش‌گوی فضایی در مکان دلخواه $\mathbf{s}_۰$ با

$$\tilde{w}(\mathbf{s}_۰) = E\left[w(\mathbf{s}_۰)|\mathbf{w}^*\right]$$

تعریف می‌شود. از آن‌جایی‌که توزیع توأم $w(\mathbf{s}_۰)|\mathbf{w}^*$ نرمال است، بنا به لم ۸.۴.۱ به سادگی نتیجه می‌شود

$$\tilde{w}(\mathbf{s}_۰) = E\left[w(\mathbf{s}_۰)|\mathbf{w}^*\right] = \mu_۲ + \Sigma_{۲۱}\Sigma_{۱۱}^{-۱}(\mathbf{x}_۱ - \mu_۱) = c^T(\mathbf{s}_۰;\theta)C^{*-۱}(\theta)\mathbf{w}^*$$

که در آن $c^T(\mathbf{s}_۰;\theta) = \left(C(\mathbf{s}_۰, \mathbf{s}^*_۱;\theta), \ldots, C(\mathbf{s}_۰, \mathbf{s}^*_m;\theta)\right)$.

فرآیند $\tilde{w}(\mathbf{s})$ را فرآیند پیش‌گوی به‌دست آمده از فرآیند پدر[۲] یعنی $w(\mathbf{s})$ می‌نامند (بنرجی و همکاران، ۲۰۰۸). بنابراین

$$\tilde{w}(\mathbf{s}) \sim GP\left\{۰, \tilde{C}(\theta)\right\}$$

[۱] Knot

[۲] Parent process

که در آن

$$\begin{aligned}\tilde{C}(\theta) &= \tilde{C}(\mathbf{s},\mathbf{s}';\theta) = Cov(\tilde{w}(\mathbf{s}),\tilde{w}(\mathbf{s}')) \\ &= Cov(c^T(\mathbf{s};\theta)C^{*-1}(\theta)\mathbf{w}^*, c^T(\mathbf{s}';\theta)C^{*-1}(\theta)\mathbf{w}^*) \\ &= c^T(\mathbf{s};\theta)C^{*-1}(\theta)Var(\mathbf{w}^*)C^{*-1}(\theta)c(\mathbf{s}';\theta) = c^T(\mathbf{s};\theta)C^{*-1}(\theta)C^*(\theta)C^{*-1}(\theta)c(\mathbf{s}';\theta) \\ &= c^T(\mathbf{s};\theta)C^{*-1}(\theta)c(\mathbf{s}';\theta). \qquad (۲.۳)\end{aligned}$$

فرآیند پیش‌گو به وسیله تابع کوواریانس فرآیند پدر و مجموعه گره‌ها یعنی S^* تعیین می‌شود. بنابراین در واقع باید بنویسیم $\tilde{w}_{S^*}(\mathbf{s})$ ولی به اختصار آن را با $\tilde{w}(\mathbf{s})$ نمایش می‌دهیم. همچنین با توجه به تابع کوواریانس (۲.۳)، به‌وضوح در می‌یابیم که فرآیند پیش‌گو، صرف نظر از مانا بودن یا نبودن $w(\mathbf{s})$، مانا نیست.

با جایگزینی $w(\mathbf{s})$ با $\tilde{w}(\mathbf{s})$ در مدل (۱.۳)، مدل فرآیند پیش‌گو، به عنوان یک مدل کم‌رتبه، به صورت زیر به‌دست می‌آید:

$$\mathbf{y}(\mathbf{s}) = \mathbf{x}(\mathbf{s})^T\beta + \tilde{w}(\mathbf{s}) + \varepsilon(\mathbf{s}). \qquad (۳.۳)$$

از آن‌جایی که

$$\tilde{w}(\mathbf{s}) = c^T(\mathbf{s};\theta)C^{*-1}(\theta)\mathbf{w}^*$$

بنابراین فرآیند $\tilde{w}(\mathbf{s})$ یک تبدیل خطی فضایی از $\mathbf{w}^*$ است. در برازش مدل (۳.۳)، n اثر تصادفی $\{w(\mathbf{s}_i), i = 1,\dots n\}$ با m اثر تصادفی $\mathbf{w}^*$ جایگزین شده است و ما با یک توزیع توأم نرمال m بعدی شامل یک ماتریس کوواریانس $m \times m$ سر و کار داریم. بنابراین کاهش بعد به روشنی مشاهده می‌شود. این کاهش بعد باعث کاهش هزینه‌های محاسباتی و در نتیجه دریافت کارایی محاسباتی می‌شود. در مقابل بخشی از کارایی آماری مدل را از دست می‌دهیم. توجه کنید که اگرچه از پارامترهای مشابه در هر دو مدل استفاده می‌کنیم، ولی در واقع پارامترهای دو مدل با هم یکسان نیستند و مدل (۳.۳) با مدل (۱.۳) متفاوت است.

۱.۲.۳ ارتباط با سایر روش‌های کاهش بعد

تشکیل ترکیب خطی مبتنی بر مجموعه گره، شبیه سایر روش‌های تقریب فرآیند است. به عنوان چند نمونه می‌توان به موارد زیر اشاره کرد:

- هایدون (۲۰۰۱) یک تقریب متناهی از فرآیند پدر به شکل $\sum_{i=1}^{m} a_i(\mathbf{s};\theta)u_i$ ارائه کرد که در آن u_i ها متغیرهای مستقل و هم‌توزیع از نرمال استاندارد هستند و $a_i(\mathbf{s};\theta) = k(\mathbf{s},\mathbf{s}_i^*;\theta)$، $i = 1,\dots,m$، به‌طوری که $k(.;\theta)$ یک تابع هسته گاوسی[۳] است.

- کامن و واند (۲۰۰۳) با تجزیه چولسکی ماتریس کوواریانس $C^*(\theta)$، ابتدا متغیرهای $Z = c^T(\mathbf{s}_i;\theta)C^{*-\frac{1}{2}}(\theta)$

[۳] Gaussian kernel function

را تعریف کردند. سپس بردار $u \sim N(\circ, I)$ را تولید کرده و فرآیند $\tilde{w}$ را به صورت $\tilde{w} = Zu$ تعریف کردند که توزیع یکسانی با مدل فرآیند پیش‌گو در (۳.۳) دارد.

- یک روش کاهش بعد دیگر اخیراً توسط کرسی و یوهانسون (۲۰۰۸) معرفی شده است. در این روش فرض می‌شود $g(\mathbf{s})$ یک بردار m بعدی از توابع پایه تعریف‌شده در $R^{\mathbf{r}}$ باشد. در این صورت تابع کوواریانس به شکل $C(\mathbf{s}, \mathbf{s}') = g(\mathbf{s})^T K g(\mathbf{s}')$ تعریف می‌شود، که در آن K یک ماتریس $m \times m$ همیشه مثبت نامعلوم است که باید با استفاده از روش گشتاوری برآورد شود. اجرای این روش ممکن است برای مدل‌های سلسله‌مراتبی چالش برانگیز باشد زیرا برای برآورد تابع کوواریانس تجربی اطلاعاتی در دسترس نداریم که بتوان با روش گشتاورها آن را برآورد کرد.

فرآیند پیش‌گوی پیشنهادشده در این پایان‌نامه یک رقیب برای روش‌های ذکرشده محسوب می‌شود، در حالی‌که فواید محاسباتی زیادی هم دارد (بنرجی و همکاران، ۲۰۰۸). البته برای تعریف این فرآیند باید یک فرآیند پررتبه داشته باشیم زیرا فرآیند پیش‌گو از طریق آن القا می‌شود. در واقع، مدل فرآیند پیش‌گو یک تقریب بهینه از مدل در یک فضای n بعدی است که به یک فضای m بعدی تصویر می‌شود. بهینه بودن فرآیند پیش‌گو را در بخش بعدی شرح خواهیم داد. توجه کنید که فرآیند پیش‌گوی $\tilde{w}(\mathbf{s})$، یک تقریب گسسته از فرآیند اصلی نیست. آن‌چه که برای تعریف فرآیند پیش‌گو نیاز داریم، فقط یک تابع کوواریانس معتبر است، به این معنی که تابع کوواریانس باید یک ماتریس همیشه مثبت باشد.

۲.۲.۳ بهینگی فرآیند پیش‌گو

اشاره کردیم که فرآیند پیش‌گو بهترین تقریب و تصویر از فرآیند پدر است. این بهینگی را از چهار منظر می‌توان تشریح کرد.

امید ریاضی شرطی

به منظور یافتن بهترین تقریب برای $w(\mathbf{s})$، باید در بین تمام توابع حقیقی $f(\cdot)$، تابعی را بیابیم که $E\left[(w(\mathbf{s}) - f(w^*))^{\mathbf{r}} | \mathbf{w}^*\right]$ را می‌نیمم کند. چنین تابعی بهترین تصویر از فرآیند پدر $w(\mathbf{s})$ خواهد بود. با اضافه و کم کردن میانگین شرطی $E\left[w(\mathbf{s})|\mathbf{w}^*\right]$ و با توجه به صفر بودن امید ریاضی مولفه حاصلضرب، داریم

$$E\left[(w(\mathbf{s}) - f(\mathbf{w}^*))^{\mathbf{r}} | \mathbf{w}^*\right] = E\left\{(w(\mathbf{s}) - E\left[w(\mathbf{s})|\mathbf{w}^*\right])^{\mathbf{r}} | \mathbf{w}^*\right\} + \left\{E\left[w(\mathbf{s})|\mathbf{w}^*\right] - f(\mathbf{w}^*)\right\}^{\mathbf{r}}.$$

از آن‌جا که عبارت دوم سمت راست نامنفی است، بنابراین

$$E\left[(w(\mathbf{s}) - f(\mathbf{w}^*))^{\mathbf{r}} | \mathbf{w}^*\right] \geq E\left\{(w(\mathbf{s}) - E\left[w(\mathbf{s})|\mathbf{w}^*\right])^{\mathbf{r}} | \mathbf{w}^*\right\}.$$

بنابراین فرآیند پیش‌گوی

$$\tilde{w}(\mathbf{s}) = E\left[w(\mathbf{s})|\mathbf{w}^*\right]$$

بهترین تقریب برای فرآیند پدر می‌باشد.

تصویرسازی متعامد

برای نقطه مکانی دلخواه $\mathbf{s}_0$، فرآیند $\tilde{w}(\mathbf{s}_0)$ یک تصویر متعامد[۴] از $w(\mathbf{s}_0)$ بر روی یک زیرفضای خطی مشخص است (استاین، ۱۹۹۹). فرض کنید $\mathcal{H}_m$ فضای هیلبرت تولیدشده توسط m متغیر تصادفی $\mathbf{w}^*$ باشد و $\mathcal{H}_{m+1}$ نیز فضای هیلبرت تولیدشده توسط $w(\mathbf{s}_0)$ و $\mathbf{w}^*$. بنابراین، $\mathcal{H}$ تمام ترکیبات خطی حاصل از این $m+1$ متغیر تصادفی با میانگین صفر و واریانس متناهی همراه با نقاط حدی میانگین توان دوم آن‌ها را در بر دارد. یافتن بهترین تقریب بر حسب نرم حاصلضرب درونی، منتهی به می‌نیمم کردن $E\left[w(\mathbf{s})w(\mathbf{s}')\right]$ می‌شود. یافتن $\tilde{w}(\mathbf{s}_0) \in \mathcal{H}$ که نزدیکترین به $w(\mathbf{s}_0)$ باشد، حل دستگاه معادلات خطی

$$E\left[\{w(\mathbf{s}_0) - \tilde{w}(\mathbf{s}_0)\}\, w(\mathbf{s}_j^*)\right] = 0, \quad j = 1, \ldots, m$$

را نتیجه می‌دهد که جواب یکتای آن برابر

$$\tilde{w}(\mathbf{s}_0) = c^T(\mathbf{s}_0) C^{*-1}(\theta)\mathbf{w}^*$$

است.

معیار کولبک-لیبلر

بهینگی فرآیند پیش‌گو بر اساس معیار کولبک-لیبلر[۵] توسط ساتو (۲۰۰۲) و سیگر (۲۰۰۳) تشریح شده است. بنرجی و همکاران (۲۰۰۸) نیز این بهینگی را بر اساس جدا بودن دو مجموعه S و S^* نشان داده‌اند. برای اطلاع بیشتر به بنرجی و همکاران (۲۰۰۸) مراجعه کنید.

درونیابی قطعی

فرآیند پیش‌گوی $\tilde{w}(\mathbf{s}_0)$ به‌طور دقیق فرآیند $w(\mathbf{s})$ را روی S^* درونیابی می‌کند. برای توضیح بیشتر فرض کنید $\mathbf{s}_0 = \mathbf{s}_j^* \in S^*$. در این صورت

$$\tilde{w}(\mathbf{s}_j^*) = c^T(\mathbf{s}_j^*; \theta) C^{*-1}(\theta)\mathbf{w}^* = w(\mathbf{s}_j^*). \qquad (۴.۳)$$

زیرا $e_j^T C^*(\theta) = c^T(\mathbf{s}_j^*; \theta)$ که در آن e_j برداری است که مولفه j ام آن یک و بقیه صفر هستند. بنابراین تساوی (۴.۳)، با استفاده از یک ویژگی پیش‌گویی کریگی، نشان می‌دهد

$$E\left[\tilde{w}(\mathbf{s}_j^*)|\mathbf{w}^*\right] = w(\mathbf{s}_j^*), \quad Var\left[\tilde{w}(\mathbf{s}_j^*)|\mathbf{w}^*\right] = 0.$$

۳.۲.۳ انتخاب گره

مشابه هر روش مبتنی بر گره‌ها، انتخاب گره‌ها در فرآیندهای پیش‌گو یک مساله چالش برانگیز است. این مساله به این دلیل که گره‌ها دوبعدی هستند (نسبت به بعد یک در روش‌هایی مثل هموارسازی اسپلاین)، مشکل‌تر هم هست.

[۴] Orthogonal projection

[۵] Kullback-Leibler

ابتدا فرض کنید m معلوم است. در متون مربوط به روش‌های هموارسازی اسپلاین (و در اغلب متون مربوط به داده‌های تابعی[۶] یا مدل‌بندی رگرسیونی با استفاده از نمایش توابع پایه)، انتخاب مشاهدات به عنوان گره‌ها مرسوم است (رمزی و سیلورمن، ۲۰۰۵). اما این انتخاب در مدل فرآیند پیش‌گو به عنوان یک گزینه قابل دفاع نیست و این سوال مطرح می‌شود که آیا از یک زیرمجموعه از مکان‌های فضایی مشاهده‌شده استفاده کنیم یا یک مجموعه جدا از مکان‌های مشاهده‌شده؟

- اگر یک زیرمجموعه از مکان‌های نمونه‌گیری‌شده انتخاب کنیم، آیا این انتخاب باید تصادفی باشد؟

- اگر نخواهیم از زیرمجموعه‌ای از مکان‌های مشاهده‌شده استفاده کنیم، یک مساله طراحی نمونه‌گیری فضایی[۷] در حضور n نقطه نمونه‌گیری‌شده پیش می‌آید. البته در زمینه طراحی نمونه‌گیری فضایی تحقیقات گسترده‌ای صورت گرفته‌اند که در ژیا و همکاران (۲۰۰۶) خلاصه شده‌اند.

بنابراین به معیاری برای تصمیم‌گیری بین دو روش ۱) انتخاب گره‌ها بر اساس یک مشبکه منظم[۸] که در آن گره‌ها به‌صورت یک مشبکه با فواصل یکسان پخش شده‌اند یا ۲) انتخاب گره‌های بیشتر در محدوده‌هایی که نمونه‌گیری بیشتری انجام شده است، نیازمندیم. در عمل اگر داده‌ها به‌طور همگن در ناحیه تحت مطالعه پراکنده باشند، تفاوت ناچیزی در روش‌های انتخاب گره وجود دارد. اما تعداد گره‌ها تاثیر زیادی بر برآورد پارامترها و پیش‌گویی‌های بعدی دارد. بنابراین این تعداد باید بسته به حساسیت استنباط و هزینه محاسباتی، تعیین شود.

اخیراً دیگل و لوفاون (۲۰۰۶) طرح‌های مختلف نمونه‌گیری فضایی را مورد کنکاش قرار داده‌اند و استفاده از طرح‌های اصلاح‌شده مشبکه منظم را توصیه می‌کنند. در ادامه برخی از این روش‌ها را توضیح خواهیم داد. یک طرح نمونه‌گیری مشبکه منظم یک ماتریس $k \times f$ را شامل می‌شود که در آن مکان گره‌ها با فاصله یکسان از یکدیگر قرار گرفته‌اند. دو رده اصلاح‌شده از این نوع طرح‌ها، مشبکه منظم به همراه جفت‌های نزدیک[۹] و مشبکه منظم طرح پر[۱۰] نام دارند.

طرح نمونه‌گیری مشبکه منظم به همراه جفت‌های نزدیک

یک مشبکه منظم $k \times k$ را در نظر بگیرید. فرض کنید نقاط مشبکه با فاصله Δ از یکدیگر قرار داشته باشند. مشبکه را با انتخاب تصادفی m' نقطه و قرار دادن یک نقطه اضافی نزدیک به هر یک از این نقاط انتخابی، تقویت می‌کنیم. هر کدام از نقاط یا مکان‌های جدید ایجادشده، از یک دایره به مرکز m' و شعاع $\delta = \alpha\Delta$ به‌طور تصادفی انتخاب می‌شوند. ما برای نمایش آن از نماد $(k \times k, m', \alpha)$ استفاده می‌کنیم و انتخاب مقدار α اختیاری است.

[۶] Functional data

[۷] Spatial design

[۸] Regular grid

[۹] Lattice plus close pairs

[۱۰] Lattice plus infill

مشبکه منظم طرح پر

یک شبکه منظم $k \times k$ را در نظر بگیرید. مثل روش قبل m' نقطه از مشبکه را به‌طور تصادفی انتخاب می‌کنیم و در نقاط انتخابی یک مسبکه کوچکتر با فواصل نزدیک‌تر به هم طراحی می‌کنیم. برای نمایش آن از نماد $(k \times k, m', r \times r)$ استفاده می‌کنیم و انتخاب مقدار r اختیاری است. هر مشبکه جدید ایجادشده یک مشبکه $r \times r$ و شامل $r^۲ - ۴$ مکان اضافی است.

عملکرد و انتخاب m

یک ارزیابی مستقیم از عملکرد انتخاب گره‌ها، مقایسه بین توابع کوواریانس فرآیند پدر و فرآیند پیش‌گو است. در نهایت انتخاب m به هزینه محاسباتی و حساسیت استنباط بستگی دارد. بنابراین، در عمل باید تحلیل را با m های مختلف انجام داده و نتایج را با هم، بر اساس معیارهای تعریف‌شده، مقایسه کنیم. مقایسه باید نسبی باشد و زمان اجرای محاسبات در فرآیند انجام استنباط در نظر گرفته شود.

۴.۲.۳ مدل فرآیند پیش‌گوی اصلاح‌شده

افزودن $\tilde{\varepsilon}(\mathbf{s})$ به فرآیند پیشگو را بنرجی و همکاران (۲۰۰۹) فرآیند پیش‌گوی اصلاح‌شده[۱۱] نامیدند. دلیل معرفی و در بعضی موارد روی آوردن به این فرآیندها این است که فرآیند پیش‌گو واریانس فرآیند پدر در هر مکان دلخواه را کم‌برآورد[۱۲] می‌کند، زیرا

$$w(\mathbf{s})|\mathbf{w}^* \sim GP\left\{c^T(\mathbf{s};\theta)C^{*-۱}(\theta)\mathbf{w}^*, C(\mathbf{s},\mathbf{s}) - c^T(\mathbf{s};\theta)C^{*-۱}(\theta)c(\mathbf{s};\theta)\right\}$$

$$Var(w(\mathbf{s})) = C(\mathbf{s},\mathbf{s})$$

و

$$Var(\tilde{w}(\mathbf{s})) = c^T(\mathbf{s};\theta)C^{*-۱}(\theta)c(\mathbf{s};\theta).$$

بنابراین

$$۰ \leq Var(w(\mathbf{s})|\mathbf{w}^*) = C(\mathbf{s},\mathbf{s}) - c^T(\mathbf{s};\theta)C^{*-۱}(\theta)c(\mathbf{s};\theta).$$

در نتیجه

$$C(\mathbf{s},\mathbf{s}) \geq c^T(\mathbf{s};\theta)C^{*-۱}(\theta)c(\mathbf{s};\theta).$$

یعنی

$$Var(w(\mathbf{s})) \geq Var(\tilde{w}(\mathbf{s})).$$

برای اصلاح این مشکل، فرآیند پیش‌گوی اصلاح‌شده، به صورت زیر، پیشنهاد شد:

$$\ddot{w}(\mathbf{s}) = \tilde{w}(\mathbf{s}) + \tilde{\varepsilon}(\mathbf{s})$$

[۱۱] Modified predictive process model

[۱۲] Underestimate

که در آن $\tilde{\varepsilon}(\mathbf{s}) \overset{ind.}{\sim} GP\left\{۰, C(\mathbf{s},\mathbf{s}) - c^T(\mathbf{s};\theta)C^{*-۱}(\theta)c(\mathbf{s};\theta)\right\}$. به روشنی مشاهده می‌شود $Var(\ddot{w}(\mathbf{s})) = C(\mathbf{s},\mathbf{s}) = Var(w(\mathbf{s}))$. همچنین

$$\begin{aligned} E\left[\ddot{w}(\mathbf{s})|\mathbf{w}^*\right] &= E\left[\tilde{w}(\mathbf{s}) + \tilde{\varepsilon}(\mathbf{s})|\mathbf{w}^*\right] \\ &= E\left[\tilde{w}(\mathbf{s})|\mathbf{w}^*\right] + E\left[\tilde{\varepsilon}(\mathbf{s})|\mathbf{w}^*\right] = E\left[c^T(\mathbf{s};\theta)C^{*-۱}(\theta)\mathbf{w}^*|\mathbf{w}^*\right] + E\left[\tilde{\varepsilon}(\mathbf{s})\right] \\ &= c^T(\mathbf{s};\theta)C^{*-۱}(\theta)\mathbf{w}^* = \tilde{w}(\mathbf{s}). \end{aligned}$$

در نتیجه

$$E\left[\ddot{w}(\mathbf{s})|\mathbf{w}^*\right] = \tilde{w}(\mathbf{s})$$

که تضمین می‌کند فرآیند پیش‌گوی اصلاح‌شده، $\ddot{w}(\mathbf{s})$، بهینگی ذکرشده برای فرآیند پیش‌گو، $\tilde{w}(\mathbf{s})$، را دارا می‌باشد.

۳.۳ مثال شبیه‌سازی

با استفاده از یک مطالعه شبیه‌سازی، عملکرد دو مدل، یعنی مدل پررتبه و فرآیند پیش‌گوی آن، را مورد ارزیابی قرار دادیم. مطالعه شبیه‌سازی توسط توابع spLM و spRecover در بسته spBayes (فینلی و همکاران، ۲۰۱۵) در محیط نرم‌افزار R انجام شده است. در این شبیه‌سازی یک مجموعه داده به حجم ۲۰۰ از مدل (۱.۳) را در یک مربع واحد تولید کردیم. در این مدل فقط یک متغیر تبیینی به همراه پارامتر عرض از مبدا، با ضرایب رگرسیونی واقعی $\beta = (\beta_۰, \beta_۱)^T = (۱, ۵)$، در نظر گرفتیم. برای اجرای الگوریتم MCMC تشریح‌شده در فصل دوم باید

۱. تعداد نمونه‌های مونت کارلو

۲. توزیع پیشین پارامترها و ابرپارامترهای مدل

۳. مقادیر آغازین برای هر پارامتر

۴. مقادیر میزان‌ساز[۱۳] برای هر پارامتر

مشخص شوند. توزیع پیشین برای β را تخت، یعنی $\Sigma_\beta^{-۱} = ۰$، فرض کردیم. برای ساختار وابستگی فضایی نیز از تابع همبستگی نمایی به شکل

$$\rho(\mathbf{s},\mathbf{s}';\phi) = \exp(-\phi||\mathbf{s}-\mathbf{s}'||)$$

استفاده کردیم. مقادیر واقعی پارامترهای وابستگی را نیز برابر $\theta = (\sigma^۲, \phi, \tau)^T = (۲, ۶, ۱)$ در نظر گرفتیم. برای پارامترهای $\tau^۲$ و $\sigma^۲$ توزیع پیشین گامای معکوس، $IG(۲, ۱)$، و برای پارامتر دامنه، ϕ، توزیع پیشین یکنواخت

$$U(-log(۰/۰۵)/۱\ ,\ -log(۰/۰۵)/۰/۱) = U(۳, ۳۰)$$

در نظر گرفتیم. مقادیر میزان‌ساز نیز $\tau^۲ = ۰/۰۱$، $\sigma^۲ = ۰/۰۱$ و $\phi = ۰/۱$ در نظر گرفته شدند.

[۱۳] Tuning value

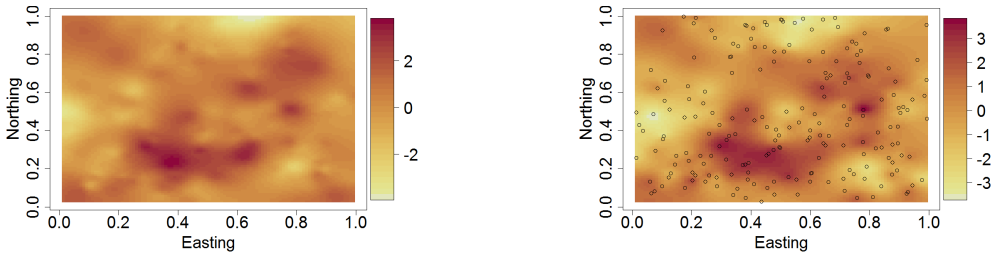
۱.۳.۳ مدل پررتبه

با مشخصات بالا، مدل پررتبه (۱.۳) را به داده‌های شبیه‌سازی شده از مدل واقعی، برازش دادیم. برآوردهای بیزی پارامترهای مدل (بر اساس میانه توزیع پسین) و فواصل اعتبار ۹۵٪ متناظر هر کدام در جدول ۱.۳ گزارش شده‌اند. با توجه به نتایج جدول می‌توان برازش نزدیک به واقعیت را مشاهده کرد. از طرفی فواصل اعتبار بیزی، مقادیر واقعی پارامترها را در بر دارند. به عبارت دیگر عدم قطعیت برآوردها به خوبی توسط مدل و روش برازش لحاظ شده است.

جدول ۱.۳: برآوردهای بیزی پارامترهای مدل (۱.۳) برای داده‌های شبیه‌سازی

پارامتر	برآورد (میانه)	چندک ۲٫۵٪	چندک ۹۷٫۵٪
σ^2	۲.۶۶	۱.۵۶	۶.۷۸
τ^2	۰.۸۵	۰.۴۳	۱.۲۸
ϕ	۷.۱۷	۳.۰۱	۱۴.۹۴
β_0	۰.۷۱	−۰.۷۸	۱.۷۷
β_1	۴.۹۶	۴.۷۹	۵.۱۷

شکل ۱.۳ رویه‌های واقعی درون‌یابی‌شده (آ) و برآوردشده پاسخ توسط مدل (ب) را نمایش می‌دهد. با توجه به مقیاس طیف رنگی دو شکل پهنه‌بندی، می‌توان به دقت پیش‌گویی حاصل از مدل پی برد.



(آ) رویه درون‌یابی‌شده مشاهدات (ب) برآورد میدان تصادفی فضایی

شکل ۱.۳: پهنه‌بندی مشاهدات واقعی و برآوردشده توسط مدل برازش‌شده در مثال شبیه‌سازی

۲.۳.۳ مدل فرآیند پیش‌گو

همان‌طور که در بخش‌های قبلی گفتیم، کاهش زمانی محاسبات یک نکته کلیدی در آمار فضایی برای داده‌های حجیم می‌باشد. برای تشریح مساله، عملکرد مدل فرآیند پیش‌گو را برای داده‌های شبیه‌سازی شده بالا مورد بررسی قرار می‌دهیم.

برای جدی شدن بحث هزینه محاسباتی، در این حالت، حجم مشاهدات شبیه‌سازی را برابر ۲۰۰۰ و چهار مدل مختلف را در نظر گرفتیم:

۱. فرآیند پیش‌گوی اصلاح‌نشده با ۲۵ گره

۲. فرآیند پیش‌گوی اصلاح‌شده با ۲۵ گره

۳. فرآیند پیش‌گوی اصلاح‌نشده با ۱۰۰ گره

۴. فرآیند پیش‌گوی اصلاح‌شده با ۱۰۰ گره

برای اجرای هر چهار مدل، توزیع‌های پیشین و مقادیر میزان‌ساز، همانند مدل پررتبه انتخاب شدند. گره‌ها را نیز از مشبکه‌های منظم ۵ × ۵ و ۱۰ × ۱۰ انتخاب کردیم. مکان قرار گرفتن این گره‌ها در شکل ۲.۳ (ب)-(ه)، نشان داده شده است.

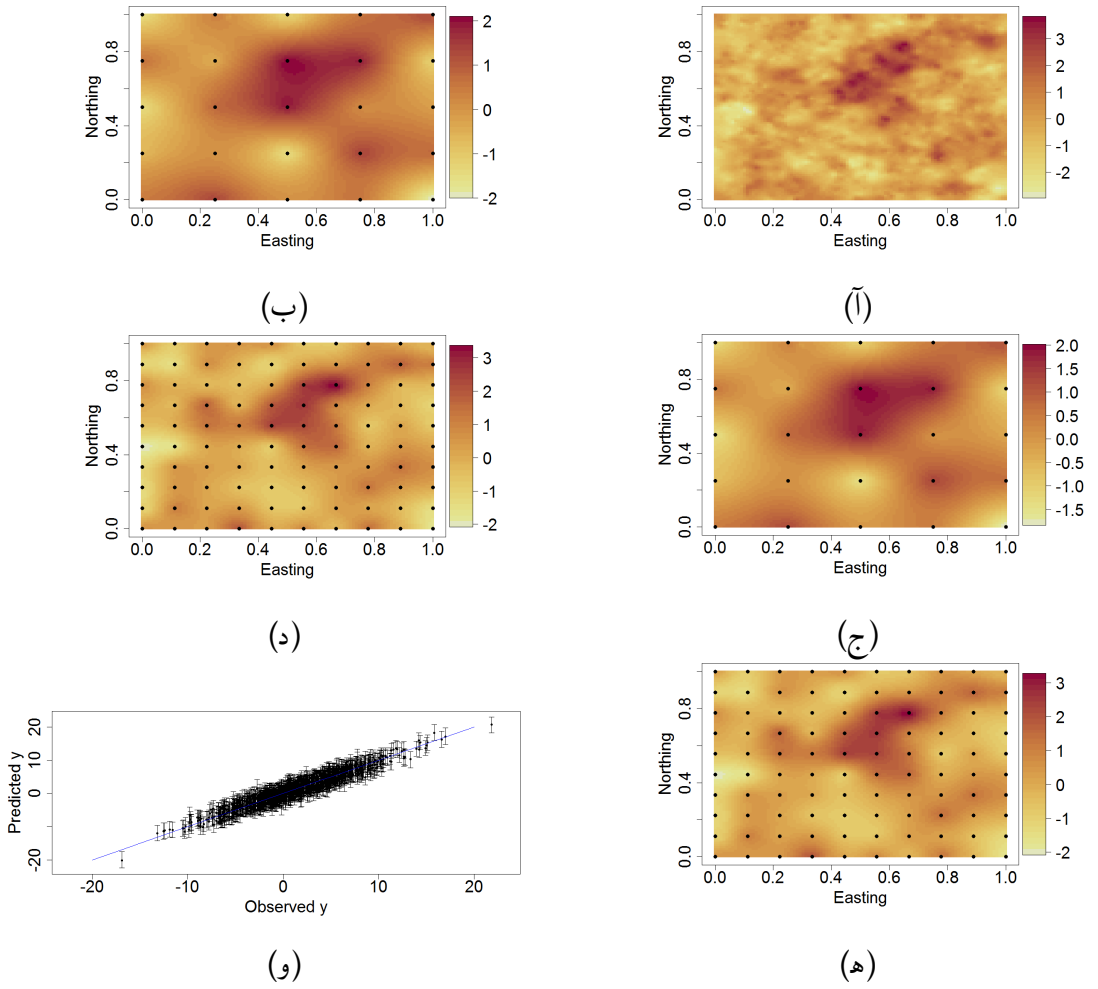
جدول ۲.۳: برآوردها (میانه توزیع پسین) و فواصل اعتبار بیزی ۹۵٪ پارامترهای مدل‌های پیش‌گوی انتخاب‌شده همراه با زمان اجرای آن‌ها.

پارامتر	مقدار واقعی	مدل ۱	مدل ۲	مدل ۳	مدل ۴
σ^2	۲	۲/۳	۱/۵۷	۱/۸۹	۱/۶۵
		(۱/۴۵, ۳/۴۸)	(۱/۰۴, ۲/۱۳)	(۱/۱۹, ۲/۶۰)	(۱/۲۳, ۳/۴۱)
τ^2	۱	۱/۷۲	۱/۱۹	۱/۴۱	۰/۸۴
		(۱/۶۰, ۱/۸۴)	(۰/۹۸, ۱/۴۲)	(۱/۳۳, ۱/۵۱)	(۰/۵۶, ۱/۰۳)
ϕ	۶	۳/۶۸	۳/۳۹	۸/۱۹	۷/۷۵
		(۳/۰۰, ۴/۸۶)	(۳/۰۳, ۴/۱۷)	(۵/۶۲, ۱۱/۲۳)	(۳/۹۳, ۱۱/۳)
β_0	۱	۰/۶۴	۰/۶۳	۰/۷۷	۰/۷۸
		(۰/۵۲, −۱/۸۳)	(۰/۳۷, −۱/۶۲)	(۰/۰۷, ۱/۴۰)	(۰/۰۳, ۱/۴۳)
β_1	۵	۴/۹۹	۴/۹۹	۴/۹۸	۴/۹۸
		(۴/۹۴, ۵/۰۵)	(۴/۹۴, ۵/۰۵)	(۴/۹۳, ۵/۰۳)	(۴/۹۳, ۵/۰۳)
زمان		۰/۱۳	۰/۱۴	۰/۷۰	۰/۷۷

جدول ۲.۳ نتایج استنباط بیزی مدل‌های فرآیند پیش‌گو را به همراه زمان لازم برای برازش آن‌ها، نشان می‌دهد. همان‌طور که در بخش فرآیندهای اصلاح‌شده اشاره شد، از نتایج دو مدل ۱ و ۳ می‌توان دریافت که یک اریبی به سمت بالا یا همان بیش اریبی در برآورد پارامتر τ^2 وجود دارد. این اریبی توسط فرآیند پیش‌گوی اصلاح‌شده، در دو مدل ۲ و ۴، به‌خوبی حذف شده است. هر چه تعداد گره‌ها بیشتر می‌شود، نتایج به مدل پررتبه نزدیک‌تر می‌شوند. در مقابل زمان برازش مدل نیز افزایش می‌یابد. به عنوان مثال زمان لازم برای برازش مدل ۳، تقریبا ۵ برابر بیشتر از زمان لازم برای برازش مدل ۱ است.

پهنه‌بندی میدان تصادفی در هر ۴ مدل، در شکل ۲.۳، (ب)-(ه)، به همراه پهنه‌بندی میدان تصادفی فضایی واقعی، (آ)، گزارش شده‌اند. روشن است که برای تعداد گره‌های کوچک، برخی از اطلاعات فرآیند فضایی نادیده گرفته شده‌اند. این نکته را می‌توان به‌راحتی از مقایسه رویه واقعی میدان تصادفی فضایی با رویه‌های برآوردشده از مدل‌های فرآیند پیش‌گوی ۱ تا ۴، دریافت. با این حال شکل ۲.۳ (و)، که نمودار پراکنش مقادیر واقعی مشاهدات در مقابل برآوردشده حاصل از مدل ۴، به همراه کران‌های اعتبار

برای هر نقطه، را نشان می‌دهد، بیانگر دقت پیش‌گویی مدل تقریبی فرآیند پیش‌گو و مدل‌بندی خوب عدم قطعیت موجود در داده‌هاست.



شکل ۲.۳: مکان گره‌های انتخابی چهار مدل فرآیند پیش‌گو به همراه پهنه‌بندی رویه‌های برآوردشده آن‌ها: (ب)-(ه)؛ پهنه‌بندی رویه واقعی میدان تصادفی فضایی: (آ)؛ و پراکنش مقادیر مشاهده‌شده واقعی در مقابل برآوردشده حاصل از مدل با تعداد ۱۰۰ گره: (و).

۴.۳ مدل خطی چندمتغیره پررتبه

فرض کنید $D \subseteq R^d$ یک زیرمجموعه از فضای اقلیدسی d بعدی و $\mathbf{s} \in D$ یک نقطه مکانی موجود در D باشد. در هر مکان $\mathbf{s}$ بردار k بعدی متغیر پاسخ را به صورت

$$\mathbf{y}(\mathbf{s}) = (y_1(\mathbf{s}), \ldots, y_k(\mathbf{s}))$$

نمایش می‌دهیم و

$$S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$$

را مجموعه n مکان در D تعریف می‌کنیم که متغیرهای پاسخ در آن‌ها مشاهده شده‌اند. مدل رگرسیون فضایی چندمتغیره برای هر مکان $\mathbf{s} \in S$ به شکل زیر تعریف می‌شود:

$$\mathbf{y}_j(\mathbf{s}) = \mathbf{x}_j^T(\mathbf{s})\beta_j + w_j(\mathbf{s}) + \varepsilon_j(\mathbf{s}), \quad j = ۱, \ldots, k \qquad (۵.۳)$$

که در آن $\mathbf{x}_j(\mathbf{s})$ بردار p_j بعدی متغیرهای تبیینی، β_j بردار p_j بعدی ضرایب رگرسیونی، و $w_j(\mathbf{s})$ و $\varepsilon_j(\mathbf{s})$ به ترتیب فرآیند فضایی و فرآیند خطای مرتبط با متغیر پاسخ $\mathbf{y}_j(\mathbf{s})$ هستند. باقیمانده‌ها یا خطای اندازه‌گیری $\varepsilon(\mathbf{s}) = (\varepsilon_۱(\mathbf{s}), \ldots, \varepsilon_k(\mathbf{s}))^T$ از یک توزیع نرمال چندمتغیره با میانگین صفر و ماتریس کوواریانس Ψ، یعنی $\varepsilon(\mathbf{s}) \sim N_k(۰, \Psi)$، پیروی می‌کنند که معمولا فرض می‌شود Ψ یک ماتریس قطری با عناصر روی قطر $\tau_j^۲$ است. اثرات یا تغییرات فضایی $\mathbf{w}(\mathbf{s}) = (w_۱(\mathbf{s}), \ldots, w_k(\mathbf{s}))$ با استفاده از یک فرآیند گاوسی k بعدی مدل‌بندی می‌شود که دارای میانگین صفر و ماتریس کوواریانس $C_w(\mathbf{s}, \mathbf{s}')$ است، که درایه‌های این ماتریس، مقدار کوواریانس بین $w_i(\mathbf{s})$ و $w_j(\mathbf{s}')$، $i, j = ۱, \ldots, k$، یعنی $Cov(w_i(\mathbf{s}), w_j(\mathbf{s}'))$ را اندازه می‌گیرند.

۱.۴.۳ روش‌های ساخت تابع کوواریانس

یکی از بخش‌های مهم در مدل (۵.۳)، تعیین فرآیند فضایی $\mathbf{w}(\mathbf{s})$ است. این فرآیند به‌طور کامل توسط تابع کوواریانس $C_{\mathbf{w}}(\mathbf{s}, \mathbf{s}')$، مشخص می‌شود که یک ماتریس بلوکی $kn \times kn$ است به‌طوری که ℓ امین بلوک آن، $\ell = ۱, \ldots, k$، شامل درایه‌های $C_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}'_j)$، $i, j = ۱, \ldots, n$، می‌باشد. برای حالت خاص $k = ۱$، این تابع کوواریانس همان ماتریس کوواریانس مدل یک‌متغیره است.

ماتریس کوواریانس یک فرآیند چندمتغیره معتبر باید همیشه مثبت باشد. بنابراین باید متقارن هم باشد. این دو الزام، دو شرط زیر را نتیجه می‌دهند:

$$\begin{aligned} &i) \;\; C_{\mathbf{w}}(\mathbf{s}, \mathbf{s}') = C_{\mathbf{w}}(\mathbf{s}', \mathbf{s})^T \\ &ii) \;\; \sum_{i=۱}^{n}\sum_{j=۱}^{n} u_i^T C_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}'_j) u_j > ۰; \;\; \forall u_i, u_j \in R^m / \{۰\}. \end{aligned} \qquad (۶.۳)$$

شرط اول ضمانت می‌کند که $C_{\mathbf{w}}(\cdot, \cdot)$ متقارن است، هرچند ماتریس کوواریانس نیازی به متقارن بودن ندارد. شرط دوم نیز تضمین می‌کند $C_{\mathbf{w}}(\cdot, \cdot)$ همیشه مثبت می‌باشد. این دو شرط باید به ازای هر عدد صحیح n و هر مجموعه متناهی از مکان‌های $S = \{\mathbf{s}_۱, \ldots, \mathbf{s}_n\} \subset D$ برقرار باشند.

مشخص کردن توابع کوواریانس معتبر که دو شرط (۶.۳) را داشته باشند، کار ساده‌ای نیست. یک روش، استفاده از یک مدل خطی هم‌ناحیه‌ساز[۱۴] (LMC) (واکرنل، ۲۰۰۶؛ گلفاند و اسمیت، ۲۰۰۴) می‌باشد. در این روش فرض می‌شود ماتریس کوواریانس به صورت

$$C_{\mathbf{w}}(\mathbf{s}, \mathbf{s}') = A(\mathbf{s}) M(\mathbf{s}, \mathbf{s}') A^T(\mathbf{s}')$$

است که در آن $M(\mathbf{s}, \mathbf{s}')$ یک ماتریس قطری $k \times k$ است به‌طوری که هر درایه‌اش یک تابع همبستگی‌نگار (با پارامترهای مخصوص به خود) می‌باشد و $A(\mathbf{s})$ یک ماتریس پایین مثلثی $k \times k$ است که از روش‌های

[۱۴] Linear model of co-regionalization

مختلفی برای تعیین آن استفاده می‌کنند. سپس فرآیند فضایی $\mathbf{w}(\mathbf{s})$ به صورت ترکیب خطی

$$\mathbf{w}(\mathbf{s}) = A(\mathbf{s})\mathbf{v}(\mathbf{s})$$

در نظر گرفته می‌شود، که در آن $\mathbf{v}(\mathbf{s}) = (v_1(\mathbf{s}), \ldots, v_k(\mathbf{s}))^T$ و هر $v_i(\mathbf{s})$ یک فرآیند گاوسی فضایی با میانگین صفر و واریانس یک است. به عبارت دیگر

$$v_i(\mathbf{s}) \sim GP(0, \rho_i(\cdot;\theta_i))$$

به‌طوری که $Var(v_i(\mathbf{s})) = 1$ و $\rho_i(\cdot;\theta_i)$ تابع همبستگی‌نگار متناظر با فرآیند $v_i(\mathbf{s})$ است. بنابراین

$$\Gamma_{\mathbf{v}}(\mathbf{s},\mathbf{s}') = Cov(v_i(\mathbf{s}), v_j(\mathbf{s}')) = \begin{cases} \rho_i(\mathbf{s},\mathbf{s}';\theta_i) & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \quad (۷.۳)$$

بنا به (۷.۳)، ماتریس کوواریانس $\Gamma_{\mathbf{v}}(\mathbf{s},\mathbf{s}')$ یک ماتریس قطری با عناصر روی قطر $\rho_i(\mathbf{s},\mathbf{s}';\theta_i)$ است. با توجه به $\mathbf{w}(\mathbf{s}) = A(\mathbf{s})\mathbf{v}(\mathbf{s})$، داریم

$$\begin{aligned} C_{\mathbf{w}}(\mathbf{s},\mathbf{s}') &= Cov(\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}')) = Cov(A(\mathbf{s})\mathbf{v}(\mathbf{s}), A(\mathbf{s}')\mathbf{v}(\mathbf{s}')) \\ &= A(\mathbf{s})Cov(\mathbf{v}(\mathbf{s}), \mathbf{v}(\mathbf{s}'))A(\mathbf{s}')^T = A(\mathbf{s})\Gamma_{\mathbf{v}}(\mathbf{s},\mathbf{s}')A(\mathbf{s}')^T. \end{aligned}$$

در نتیجه یک تابع کوواریانس معتبر برای فرآیند پدر ساخته شد. یک انتخاب برای $\rho_i(\cdot;\theta_i)$ تابع همبستگی‌نگار ماترن است. چندین انتخاب معتبر دیگر برای تابع همبستگی‌نگار توسط بنرجی و همکاران (۲۰۰۴) ارایه شده‌اند.

ماتریس $A(\mathbf{s})$ نیز نامعلوم است و باید به گونه‌ای آن را تعیین کنیم. یکی از گزینه‌ها برای ساختن آن، توجه به رابطه

$$C_{\mathbf{w}}(\mathbf{s},\mathbf{s}) = A(\mathbf{s})\Gamma_{\mathbf{v}}(\mathbf{s},\mathbf{s})A(\mathbf{s})^T = A(\mathbf{s})Var(\mathbf{v}(\mathbf{s}))A(\mathbf{s})^T = A(\mathbf{s})A(\mathbf{s})^T$$

است. با توجه به این تساوی، یک راه برای تعیین $A(\mathbf{s})$، محاسبه ماتریس پایین مثلثی حاصل از تجزیه چولسکی ماتریس $C_{\mathbf{w}}(\mathbf{s},\mathbf{s})$ است. یک گزینه دیگر این است که فرض کنیم عناصر $A(\mathbf{s})$ خود از یک فرآیند فضایی تولید می‌شوند. به عنوان مثال، گلفاند و همکاران (۲۰۰۴) از یک فرآیند فضایی ویشارت معکوس برای ایجاد $A(\mathbf{s})A(\mathbf{s})^T$ استفاده کردند.

۲.۴.۳ نمایش مدل چندمتغیره برای استنباط

فرض کنید در هر کدام از b مکان فضایی، متغیر پاسخ k بعدی مشاهده شده باشد. قرار می‌دهیم $\mathbf{y} = (\mathbf{y}(\mathbf{s}_1)^T, \ldots, \mathbf{y}(\mathbf{s}_b)^T)^T$ بردار n بعدی پاسخ باشد که در آن $n = mb$. همچنین فرض کنید X ماتریس $n \times p$ متغیرهای تبیینی متناظر با $\mathbf{y}$ باشد، به‌طوری که $p = \sum_{j=1}^{k} p_j$ و $\beta = (\beta_1^T, \ldots, \beta_k^T)^T$ بردار p بعدی حاصل از تجمیع همه β_j ها، $j = 1, \ldots, k$، باشد. بنابراین مدل رگرسیون فضایی چندمتغیره، از مدل سلسله‌مراتبی (۲.۲) با مشخصات زیر حاصل می‌شود: $D(\theta) = I_b \otimes \Psi$ که در آن $\otimes$ ضرب کرونه‌کر است، α برداری n بعدی حاصل از تجمیع $\mathbf{w}(\mathbf{s}_i)$ هاست، و $K(\theta)$ یک ماتریس بلوکی $n \times n$ با بلوک‌های $k \times k$ به صورت $AM(\mathbf{s}_i,\mathbf{s}_j)A^T$ است.

۵.۳ مدل‌های چندمتغیره فرآیند پیش‌گو

در بخش قبل اشاره کردیم که تحقق‌های $\mathbf{w}(\mathbf{s})$ روی مجموعه متناهی S از یک توزیع نرمال چندمتغیره با ماتریس کوواریانس $C_{\mathbf{w}}(\cdot,\cdot;\theta)$ پیروی می‌کنند. برآورد (۵.۳) نیازمند تجزیه چولسکی یک ماتریس $kn \times kn$ خواهد بود که به‌دست آوردن آن نیازمند محاسبات پیچیده از مرتبه $O(n^۳k^۳)$ است. بنابراین وقتی kn بزرگ باشد، این محاسبات پیچیده‌تر هم خواهند شد. راه حل آسان و موثر برای حل این مشکل، استفاده از نسخه چندمتغیره مدل‌های پیش‌گو است.

مشابه حالت تک متغیره، فرض می‌کنیم $\mathbf{w}(\mathbf{s}) \sim MVGP_k(\circ, C_{\mathbf{w}}(\cdot,\cdot;\theta))$ که یک فرآیند فضایی چندمتغیره گاوسی با میانگین صفر و ماتریس کوواریانس $C_{\mathbf{w}}(\cdot,\cdot;\theta)$ می‌باشد. فرآیند $\mathbf{w}(\mathbf{s})$ را فرآیند پدر می‌نامیم و مجموعه گره‌ها را به صورت $S^* = \{\mathbf{s}^*_۱, \ldots, \mathbf{s}^*_m\}$ که $m \ll n$، در نظر می‌گیریم. فرآیند $\mathbf{w}^*$ را تحقق‌های فرآیند پدر روی مجموعه S^* تعریف می‌کنیم. یعنی

$$\mathbf{w}^* = (\mathbf{w}(\mathbf{s}^*_۱), \mathbf{w}(\mathbf{s}^*_۲) \ldots, \mathbf{w}(\mathbf{s}^*_m)).$$

بنابراین فرآیند پیش‌گوی چندمتغیره برای $\mathbf{w}(\mathbf{s})$ به صورت

$$\tilde{\mathbf{w}}(\mathbf{s}) = E[\mathbf{w}(\mathbf{s})|\mathbf{w}^*] = Cov(\mathbf{w}(\mathbf{s}), \mathbf{w}^*)Var(\mathbf{w}^*)^{-۱}\mathbf{w}^* = c^T(\mathbf{s};\theta)C^{*-۱}(\theta)\mathbf{w}^* \quad (۸.۳)$$

تعریف می‌شود که در آن

$$c^T(\mathbf{s};\theta) = (C_{\mathbf{w}}(\mathbf{s}, \mathbf{s}^*_۱;\theta), \ldots, C_{\mathbf{w}}(\mathbf{s}, \mathbf{s}^*_m;\theta))$$

یعنی $c^T(\mathbf{s};\theta)$ یک ماتریس بلوکی $k \times mk$ شامل m بلوک $k \times k$ با درایه‌های $C_{\mathbf{w}}(\mathbf{s}, \mathbf{s}^*_j;\theta)$ است. $C^*(\theta)$ نیز ماتریس کوواریانس بلوکی $mk \times mk$ برای فرآیند $\mathbf{w}^*$ با درایه‌های $C_{\mathbf{w}}(\mathbf{s}^*_i, \mathbf{s}^*_j;\theta)$ است. رابطه (۸.۳) نشان می‌دهد فرآیند پیش‌گو دارای میانگین صفر و ماتریس کوواریانس زیر است:

$$\begin{aligned}\Gamma_{\tilde{\mathbf{w}}}(\mathbf{s},\mathbf{s}') &= Cov(c^T(\mathbf{s};\theta)C^{*-۱}(\theta)\mathbf{w}^*, c^T(\mathbf{s}';\theta)C^{*-۱}(\theta)\mathbf{w}^*)\\ &= c^T(\mathbf{s};\theta)C^{*-۱}(\theta)Cov(\mathbf{w}^*,\mathbf{w}^*)C^{*-۱}(\theta)c(\mathbf{s}';\theta)\\ &= c^T(\mathbf{s};\theta)C^{*-۱}(\theta)C^*(\theta)C^{*-۱}(\theta)c(\mathbf{s}';\theta) = c^T(\mathbf{s};\theta)C^{*-۱}(\theta)c(\mathbf{s}';\theta).\end{aligned}$$

فرآیند پیش‌گو در حالت چندمتغیره خواصی مشابه حالت یک متغیره دارد. به عنوان مثال، می‌توان نشان داد فرآیند پیش‌گوی چندمتغیره یک درونیاب دقیق است؛ یعنی برای هر $\mathbf{s}^*_j \in S^*$ داریم

$$\tilde{\mathbf{w}}(\mathbf{s}^*_j) = c^T(\mathbf{s}^*_j;\theta)C^{*-۱}(\theta)\mathbf{w}^* = e_j^TC^*(\theta)C^{*-۱}(\theta)\mathbf{w}^* = \mathbf{w}(\mathbf{s}^*_j).$$

تساوی سوم برقرار است، زیرا $e_j^TC^*(\theta) = c^T(\mathbf{s}^*_j;\theta)$ که در آن e_j برداری است که j امین عضو آن یک و سایر عناصر صفر هستند. سایر خواص بهینگی، مشابه حالت یک متغیره، نیز برقرار هستند.

ساخت فرآیند چندمتغیره پیش‌گو فقط نیازمند شناخت $C_{\mathbf{w}}(\mathbf{s},\mathbf{s}')$ و تعیین مجموعه گره S^* است که نحوه ساختن $C_{\mathbf{w}}(\mathbf{s},\mathbf{s}')$ را در بخش قبل و S^* را در بخش یک متغیره شرح دادیم. همانند قبل فرآیند پیش‌گو با جایگزینی $\mathbf{w}(\mathbf{s})$ با $\tilde{\mathbf{w}}(\mathbf{s})$ در مدل (۵.۳) اجرا می‌شود.

۶.۳ مدل‌های آمیخته خطی تعمیم‌یافته

در بسیاری از مسایل زمین‌آماری، با پاسخ‌هایی مواجه می‌شویم که برای آن‌ها نمی‌توان پذیره توزیع نرمال را در نظر گرفت. به‌ویژه زمانی که پاسخ‌ها گسسته هستند، ماهیت آن‌ها قابل مدل‌بندی با توزیع نرمال نیست. به عنوان مثال متغیر پاسخ $\mathbf{y(s)}$ ممکن است نشان‌دهنده این باشد که در مکان $\mathbf{s}$، در ۲۴ ساعت گذشته، باران باریده است یا خیر. یا ممکن است تحلیل یک متغیر شمارشی مانند تعداد بیمه‌شدگان در پنج سال گذشته در یک خانواده، با مکان $\mathbf{s}$ مورد علاقه باشد. در فصل اول عنوان کردیم که برای مدل‌بندی این نوع پاسخ‌ها از مدل‌های آمیخته خطی تعمیم‌یافته فضایی[۱۵] (SGLMMs) استفاده می‌کنیم. این مدل‌ها معمولاً برای مدل‌بندی داده‌های فضایی گسسته روی یک ناحیه جغرافیایی پیوسته به کار می‌روند.

دیگل و همکاران در سال ۱۹۹۸ استفاده از SGLMM برای تحلیل داده‌های فضایی را پیشنهاد کردند. محققین مختلفی در این حوزه مطالعات مختلفی را انجام داده‌اند. به عنوان چند نمونه می‌توان به لین و همکاران (۲۰۰۰)، کامن و واند (۲۰۰۳)، و بنرجی و همکاران (۲۰۰۴) اشاره کرد. در این رده از مدل‌ها فرض می‌شود میانگین پاسخ به کمک تابع پیوند $g(\cdot)$ به پیش‌گوی خطی به صورت زیر مرتبط می‌شود:

$$\eta(\mathbf{s}) \equiv g\left\{E\left[Y(\mathbf{s})\right]\right\} = \mathbf{x}^T(\mathbf{s})\beta + w(\mathbf{s}).$$

با فرض $Z(\theta) = I_n$، $K(\theta) = C(\mathbf{s}_i, \mathbf{s}_j; \theta)$ و $\alpha = w(\mathbf{s}_i)$، شبیه به مدل سلسله‌مراتبی (۲.۲) داریم

$$\pi(\zeta|\mathbf{y}, X) \propto \pi(\theta) \times N(\beta|\mu_\beta, \Sigma_\beta) \times N(w(\mathbf{s})|\circ, C(\mathbf{s}_i, \mathbf{s}_j; \theta)) \times \prod_{i=1}^{n} f(\mathbf{y}(\mathbf{s}_i)|\eta(\mathbf{s}_i))$$

که در آن $f(.)$ تابع (چگالی) احتمال متغیر پاسخ است که عضوی از خانواده توزیع‌های نمایی می‌باشد.

مانند موارد قبلی برای n های بزرگ، محاسبات حجیم و طاقت‌فرسا خواهد بود. مشکلات محاسباتی برای این مدل‌ها، به دلیل خارج شدن از رده مدل‌های گاوسی، جدی‌تر نیز می‌شوند. برای مرتفع ساختن این مشکل، می‌توان با استفاده از مدل فرآیند پیش‌گوی تعریف‌شده از $\mathbf{w}^*$ به‌جای w استفاده کرد که فقط نیازمند به‌روز رسانی یک بردار m بعدی می‌باشد. محاسبات با اضافه کردن مؤلفه خطا می‌تواند ساده‌تر هم بشود. یعنی

$$\eta(\mathbf{s}) \equiv g\left\{E\left[Y(\mathbf{s})\right]\right\} = \mathbf{x}^T(\mathbf{s})\beta + w(\mathbf{s}) + \varepsilon(\mathbf{s})$$

که در آن $\varepsilon(\mathbf{s}) \overset{i.i.d.}{\sim} N(\circ, \tau^2)$ به‌طوری که τ^2 معلوم و خیلی کوچک است. دلیل سادگی محاسبات آن است که توزیع w و در نتیجه $\mathbf{w}^*$ نرمال خواهد شد.

[۱۵] Spatial GLMMs

فصل ۴

مدل فضایی-زمانی پویا برای داده‌های زمین‌آماری حجیم

۱.۴ مقدمه

مدل‌های فضایی-زمانی خطی در سال‌های اخیر محبوبیت چشمگیری در زمینه‌های مهندسی، اقتصاد، ژنتیک و غیره پیدا کرده‌اند. گلفاند و همکاران (۲۰۰۵) چارچوب مدل‌های پویا با نام مدل‌های فضایی-زمانی با ضرایب وابسته فضایی را ساختند. رده‌های دیگری از مدل‌های خطی پویا برای داده‌های فضایی-زمانی را می‌توان در استروند و همکاران (۲۰۰۱) و تونلیتو (۱۹۹۷) یافت. اما ما به دنبال مدل‌هایی هستیم که به ما اجازه بدهد مجموعه داده‌های فضایی-زمانی بزرگ را بررسی کنیم. برای این منظور از رده مدل‌های کم‌رتبه و نمایش مدل‌های فرآیند پیش‌گو که توسط بنرجی و همکاران (۲۰۰۸) معرفی شدند، استفاده می‌کنیم.

۲.۴ مدل خطی فضایی-زمانی پویا

فرض کنید $y_t(\mathbf{s})$ متغیر پاسخ در مکان $\mathbf{s}$ و زمان t باشد، به طوری که $t = ۱, \ldots, N_t$ و $\mathbf{s} \in D \subseteq R^۲$.
مدل خطی فضایی-زمانی پویا به صورت زیر نوشته می‌شود:

$$
\begin{aligned}
y_t(\mathbf{s}) &= \mathbf{x}_t(\mathbf{s})^T \beta_t + u_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}), \qquad \varepsilon_t(\mathbf{s}) \overset{ind.}{\sim} N_n(۰, \tau^۲); \\
\beta_t &= \beta_{t-۱} + \eta_t, \qquad \eta_t \overset{i.i.d.}{\sim} N_p(۰, \Sigma_\eta); \\
u_t(\mathbf{s}) &= u_{t-۱}(\mathbf{s}) + w_t(\mathbf{s}), \qquad w_t(\mathbf{s}) \overset{ind.}{\sim} GP(۰, C_t(\cdot, \theta_t)), \quad t = ۱, ۲, \ldots, N_t
\end{aligned} \tag{۱.۴}
$$

که اختصارات $ind.$ و $i.i.d.$ به ترتیب نشان‌دهنده مستقل و مستقل و هم‌توزیع هستند. این مدل حالت خاصی از یک مدل فضای حالت[۱] است که در قالب معادله مشاهده (تساوی اول (۱.۴)) و معادله‌های

[۱] State space

حالت (دو تساوی دوم و سوم رابطه (۱.۴)) بیان می‌شوند. همچنین $\mathbf{x}_t(\mathbf{s})$ بردار p بعدی متغیرهای تبیینی و β_t بردار p بعدی ضرایب رگرسیونی وابسته به زمان متناظر با متغیرهای تبیینی می‌باشد.

مشابه قبل، معمولاً، شکل تابع کوواریانس به صورت $C_t(\mathbf{s}_1, \mathbf{s}_2; \theta_t) = \sigma_t^2 \rho(\mathbf{s}_1, \mathbf{s}_2; \phi_t)$ است، که در آن $\theta_t = \{\sigma_t^2, \phi_t\}$، $\rho(\cdot; \phi_t)$ تابع همبستگی با پارامتر ϕ_t، و σ_t^2 مولفه واریانس فضایی می‌باشد. یک انتخاب برای تعریف تابع همبستگی فضایی، تابع نمایی است که به صورت

$$C_t(\mathbf{s}_1, \mathbf{s}_2; \theta_t) = \sigma_t^2 \exp\{-\phi_t ||\mathbf{s}_1 - \mathbf{s}_2||\}$$

نمایش داده می‌شود و در آن $||\mathbf{s}_1 - \mathbf{s}_2||$، فاصله اقلیدسی بین مکان‌های $\mathbf{s}_1$ و $\mathbf{s}_2$ است. با این وجود، هر تابع همبستگی فضایی معتبر را نیز می‌توان به کار برد؛ به طور مثال کرسی (۱۹۹۳)، چایلز و دلفینر (۱۹۹۹) و بنرجی و همکاران (۲۰۰۴) را ببینید.

با پذیرفتن $\beta_0 \sim N_p(\mathbf{m}_0, \Sigma_0)$ که در آن $\mathbf{m}_0$ و Σ_0 به ترتیب بردار میانگین و ماتریس کوواریانس معلوم هستند و $u_0(\mathbf{s}) = 0$، برای $t = 0$، مدل سلسله‌مراتبی بیزی (۱.۴) کامل می‌شود. برای هر نقطه زمانی، تعداد مکان‌های یکسانی را در نظر می‌گیریم. بنابراین در ماتریس مکانی-زمانی داده‌ها، سطرها نمایانگر مکان و ستون‌ها نمایشگر زمان هستند؛ به عبارتی (i, j)امین عنصر $y_j(\mathbf{s}_i)$ می‌باشد.

در مواردی که تعداد مکان‌ها، به اندازه کافی بزرگ (یعنی سطرهای زیاد برای ماتریس فضایی-زمانی داده‌ها) و تعداد نقطه‌های زمانی، تحت کنترل یا محدود باشند، انجام استنباط بیزی کامل برای مدل (۱.۴)، از نظر محاسباتی، طاقت‌فرسا است. برای تشریح این سختی محاسباتی، فرض کنید از روش نمونه‌گیری گیبز برای تولید نمونه از توزیع پسین پارامترها (شامل اثرات تصادفی فضایی-زمانی) استفاده کنیم. اجرای این الگوریتم نیازمند تجزیه ماتریس کوواریانس فضایی-زمانی است که برای هر نقطه زمانی، تعداد عملیات مورد نیاز از مرتبه سوم تعداد مکان‌ها است. یعنی برای مثلا n مکان، تعداد عملیات لازم برابر $O(n^3)$ است. این تعداد عملیات برای تمام مراحل الگوریتم باید در تعداد تکرارهای نمونه‌گیری نیز ضرب شود.

۳.۴ فرآیند پیش‌گو

برای دوری از تنگنای محاسباتی مذکور، فرآیند فضایی-زمانی در (۱.۴) را با یک فرآیند پیش‌گو جایگزین می‌کنیم (بنرجی و همکاران، ۲۰۰۸). برای این منظور، مجموعه گره

$$S^* = \{\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_m^*\}$$

را، به‌طوری که $m \ll n$، در نظر می‌گیریم. تعریف می‌کنیم

$$\tilde{w}_t(\mathbf{s}) = E\left[w_t(\mathbf{s})|\mathbf{w}_t^*\right]$$

که در آن $\mathbf{w}_t^* = \left(w_t(\mathbf{s}_1^*), w_t(\mathbf{s}_2^*), \ldots, w_t(\mathbf{s}_m^*)\right)^T$. صورت فرآیند پیش‌گو برای مدل (۱.۴) با جایگزینی $u_t(\mathbf{s})$ در مدل (۱.۴) با $\tilde{u}_t(\mathbf{s}) = \sum_{k=1}^{t} \left[\tilde{w}_k(\mathbf{s}) + \tilde{\varepsilon}_k(\mathbf{s})\right]$ حاصل می‌شود، که در آن $\tilde{\varepsilon}_k(\mathbf{s})$ مؤلفه‌ای برای کاهش کم‌برآوردی تغییرات فضایی حاصل از فرآیند پیش‌گو است. اضافه کردن $\tilde{\varepsilon}_k(\mathbf{s})$ به فرآیند، فرآیند پیش‌گوی اصلاح‌شده بنرجی و همکاران (۲۰۰۹) را در پی خواهد داشت. مشابه فصل

قبل، کم برآوردی واریانس فرآیند پدر در هر مکان و زمان دلخواه، مشهود است؛ زیرا با توجه به

$$w_k(\mathbf{s})|w_k^* \sim N_n\left(\mathbf{c}_k(\mathbf{s};\theta_k)^T C_k^{*-1}(\theta_k) w_k^*, C_k(\mathbf{s},\mathbf{s}) - \mathbf{c}_k(\mathbf{s};\theta_k)^T C_k^{*-1}(\theta_k)\mathbf{c}_k(\mathbf{s};\theta_k)\right)$$

نتیجه می‌شود

$$0 \leq Var(w_k(\mathbf{s})|w_k^*) = C_k(\mathbf{s},\mathbf{s}) - \mathbf{c}_k(\mathbf{s};\theta_k)^T C_k^{*-1}(\theta_k)\mathbf{c}_k(\mathbf{s};\theta_k).$$

بنابراین

$$C_k(\mathbf{s},\mathbf{s}) \geq \mathbf{c}_k(\mathbf{s};\theta_k)^T C_k^{*-1}(\theta_k)\mathbf{c}_k(\mathbf{s};\theta_k).$$

از طرفی چون

$$Var(\tilde{w}_k(\mathbf{s})) = \mathbf{c}_k(\mathbf{s};\theta_k)^T C_k^{*-1}(\theta_k)\mathbf{c}_k(\mathbf{s};\theta_k)$$

و

$$Var(w_k(\mathbf{s})) = C_k(\mathbf{s},\mathbf{s})$$

پس

$$Var(w_k(\mathbf{s})) \geq Var(\tilde{w}_k(\mathbf{s})).$$

فرآیند پیش‌گوی اصلاح‌شده، مشابه فصل قبل، به صورت زیر پیشنهاد می‌شود:

$$\ddot{w}_k(\mathbf{s}) = \tilde{w}_k(\mathbf{s}) + \tilde{\varepsilon}_k(\mathbf{s})$$

که در آن

$$\tilde{\varepsilon}_k(\mathbf{s}) \overset{ind.}{\sim} N_n(0, C_k(\mathbf{s},\mathbf{s}) - \mathbf{c}_k(\mathbf{s};\theta_k)^T C_k^{*-1}(\theta_k)\mathbf{c}_k(\mathbf{s};\theta_k)).$$

با اجرای این فرآیند، پیچیدگی محاسباتی (و در نتیجه زمان اجرا)، برای هر نقطه زمانی دلخواه، از $O(n^3)$ به $O(m^3)$ کاهش می‌یابد. مشابه مدل‌های فضایی، نکته کلیدی در این مدل‌ها، انتخاب تعداد و مکان گره‌هاست.

۱.۳.۴ برازش مدل

فرض کنید مجموعه $S = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ شامل مکان‌هایی است که متغیرهای پاسخ و تبیینی مدل در آن‌ها مشاهده شده‌اند و برای هر مکان در N_t زمان مختلف، مشاهده داشته باشیم. همچنین فرض کنید $(y_t(\mathbf{s}_1), y_t(\mathbf{s}_2), \ldots, y_t(\mathbf{s}_n))^T$ بردار n بعدی پاسخ و X_t ماتریس $n \times p$، در زمان t، است که iامین سطر آن بردار $x_t(\mathbf{s}_i)^T$ می‌باشد. مدل فرآیند پیش‌گو را می‌توان به صورت زیر نوشت:

$$y_t(\mathbf{s}) = X_t(\mathbf{s})^T\beta_t + \tilde{\mathbf{u}}_t(\mathbf{s}) + \varepsilon_t(\mathbf{s}); \qquad \varepsilon_t(\mathbf{s}) \overset{ind.}{\sim} N_n(0, \tau_t^2 I_n), \quad t = 1, 2, \ldots, N_t \quad (۲.۴)$$

که در آن $\tilde{\mathbf{u}}_t = (\tilde{u}_t(\mathbf{s}_1), \tilde{u}_t(\mathbf{s}_2), \ldots, \tilde{u}_t(\mathbf{s}_n))^T$ تحقق‌های فرآیند پیش‌گو است. همچنین می‌توان نوشت

$$\tilde{\mathbf{u}}_t = \sum_{k=1}^{t} \left[C_k(\theta_k)^T C_k^*(\theta_k)^{-1} w_k^* + \tilde{\varepsilon}_k\right], \quad t = 1, 2, \ldots, N_t$$

که در آن $C_k(\theta_k)^T$ یک ماتریس $n \times m$ است که (i,j)امین مولفه آن کوواریانس بین $w_k(\mathbf{s}_i)$ و $w_k(\mathbf{s}_j^*)$، $C_k(\mathbf{s}_i, \mathbf{s}_j^*; \theta_k)$، است و $C_k^*(\theta_k)$ ماتریسی $m \times m$ با مولفه (i,j)ام $C_k(\mathbf{s}_i^*, \mathbf{s}_j^*; \theta_k)$ است. همچنین $\tilde{\varepsilon}_k \overset{ind.}{\sim} N(\circ, \tilde{D}_k)$، به‌طوری که $\tilde{D}_k$ یک ماتریس قطری $n \times n$ با iامین عنصر روی قطر

$$C_k(\mathbf{s}_i, \mathbf{s}_i) - \mathbf{c}_k(\mathbf{s}_i)^T C_k^*(\theta)^{-1} \mathbf{c}_k(\mathbf{s}_i)$$

است که در آن $\mathbf{c}_k(\mathbf{s}_i)^T$، iامین سطر ماتریس $C_k(\theta_k)^T$ می‌باشد.

برای انجام استنباط بیزی، باید توزیع پسین پارامترهای

$$\zeta = \{\beta_\circ, \Sigma_\eta, \{\theta_t\}, \{\beta_t\}, \{\tilde{\mathbf{u}}_t\}, \{\mathbf{w}_t^*\}, \{\tau_t^2\}\}$$

را، تحت توزیع پیشین $\pi(\zeta)$، به‌دست آوریم. با فرض استقلال بین مولفه‌های ζ و در نظرگرفتن توزیع‌های پیشین نرمال برای β، $\tilde{u}_t$ و $\mathbf{w}^*$، ویشارت معکوس برای Σ_η، و گامای معکوس برای پارامترهای وابستگی τ^2 و $\theta = (\phi, \sigma^2)$، توزیع پسین کامل مدل سلسله‌مراتبی متناسب است با

$$\begin{aligned}
\pi(\zeta|y_1, \ldots, y_{N_t}, X_1, \ldots, X_{N_t}) \propto & \prod_{t=1}^{N_t} \pi(\theta_t) \prod_{t=1}^{N_t} IG(\tau_t^2|a_\tau, b_\tau) N(\beta_\circ|m_\circ, \Sigma_\circ) IW(\Sigma_\eta|r_\eta, \gamma_\eta) \\
& \times \prod_{t=1}^{N_t} N(\beta_t|\beta_{t-1}, \Sigma_\eta) \prod_{t=1}^{N_t} N\left(\mathbf{w}_t^*|\circ, C_t^*(\theta_t)\right) \\
& \times \prod_{t=1}^{N_t} N\left(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t-1} + C_t(\theta_t)^T C_t^*(\theta_t)^{-1} \mathbf{w}_t^*, \tilde{D}_t\right) \\
& \times \prod_{t=1}^{N_t} N\left(\mathbf{y}_t|X_t\beta_t + \tilde{\mathbf{u}}_t, \tau_t^2 I_n\right).
\end{aligned}$$

تولید نمونه از این توزیع پسین با استفاده از الگوریتم نمونه‌گیری گیبز، نیازمند محاسبه توزیع‌های شرطی کامل است که جزییات محاسبه آن‌ها همراه با الگوریتم نمونه‌گیری گیبز را در ادامه، ارایه می‌دهیم.

توزیع شرطی کامل $\beta_\circ$

اولین مرحله الگوریتم، محاسبه توزیع شرطی کامل $\beta_\circ$ و تولید نمونه از آن است.

قضیه ۱.۳.۴. *توزیع شرطی کامل $\beta_\circ$ به صورت*

$$\beta_\circ|\mathbf{y}, \zeta_{-\beta_\circ} \sim N\left(\mu_{\beta_\circ|.}, \Sigma_{\beta_\circ|.}\right)$$

است، به‌طوری که منظور از $\zeta_{-\beta_\circ}$ بردار پارامترهای توزیع پسین به‌جز $\beta_\circ$ است و

$$\Sigma_{\beta_\circ|.} = \left(\Sigma_\circ^{-1} + \Sigma_\eta^{-1}\right)^{-1}$$

$$\mu_{\beta_\circ|.} = \Sigma_{\beta_\circ|.} \left[\Sigma_\circ^{-1} m_\circ + \Sigma_\eta^{-1} \beta_1\right].$$

برهان. برای محاسبه توزیع شرطی کامل $\beta_\circ$، با حذف مولفه‌هایی که به پارامتر $\beta_\circ$ بستگی ندارند، این توزیع تنها به

$$N(\beta_\circ|m_\circ, \Sigma_\circ) \times N(\beta_1|\beta_\circ, \Sigma_\eta)$$

وابسته خواهد بود. حال برای به‌دست آوردن توزیع شرطی کامل $\beta_\circ$، با به‌کارگیری لم لیندلی اسمیت، لم ۶.۴.۱، داریم

$$\mathbf{y}|\theta_1 \sim N(A_1\theta_1, C_1) \equiv \beta_1|\beta_\circ \sim N(\beta_\circ, \Sigma_\eta)$$

و

$$\theta_1|\theta_2 \sim N(A_2\theta_2, C_2) \equiv \beta_\circ \sim N(\mu_\circ, \Sigma_\circ)$$

که در آن $A_1 = 1$، $C_1 = \Sigma_\eta$، $A_2\theta_2 = \mu_\circ$ و $C_2 = \Sigma_\circ$. بنابراین، با توجه به حکم ب لم ۶.۴.۱

$$\beta_\circ|\mathbf{y}, \zeta_{-\beta_\circ} \sim N_p(Bb, B)$$

که در آن

$$B^{-1} = C_2^{-1} + A_1^T C_1^{-1} A_1 = \Sigma_\circ^{-1} + \Sigma_\eta^{-1}$$

و

$$b = C_2^{-1} A_2\theta_2 + A_1^T C_1^{-1}\mathbf{y} = \Sigma_\circ^{-1}\mu_\circ + \Sigma_\eta^{-1}\beta_1.$$

□

توزیع شرطی کامل ضرایب رگرسیونی

با توجه به وابستگی ضرایب رگرسیونی به زمان، برای هر $t = 1, \ldots, N_t$ باید پارامترهای توزیع شرطی به‌روز شوند.

قضیه ۲.۳.۴. *توزیع شرطی کامل β_t نرمال است. یعنی*

$$\beta_t|\mathbf{y}, \zeta_{-\beta_t} \sim N\left(\mu_{\beta_t|.}, \Sigma_{\beta_t|.}\right)$$

که در آن

$$\Sigma_{\beta_t|.} = \left(2\Sigma_\eta^{-1} + \frac{1}{\tau_t^2} X_t^T X_t\right)^{-1}$$

$$\mu_{\beta_t|.} = \Sigma_{\beta_t|.}\left[\Sigma_\eta^{-1}\left(\beta_{t-1} + \beta_{t+1}\right) + X_t^T \frac{(\mathbf{y}_t - \tilde{\mathbf{u}}_t)}{\tau_t^2}\right].$$

البته از آن‌جا که برای $t = N_t$، $\beta_{t+1} = \circ$ است، در این حالت

$$\Sigma_{\beta_t|.} = \left(\Sigma_\eta^{-1} + \frac{1}{\tau_t^2} X_t^T X_t\right)^{-1}.$$

برهان. با حذف مولفه‌هایی که به پارامتر β_t، در زمان t، وابسته نیستند، توزیع شرطی کامل متناسب است با

$$N((\mathbf{y}_t - \tilde{\mathbf{u}}_t)|X_t\beta_t, \tau_t^2 I_n) \times N(\beta_t|\beta_{t-1}, \Sigma_\eta) \times N(\beta_t|\beta_{t+1}, \Sigma_\eta).$$

برای به‌دست آوردن توزیع شرطی کامل β_t، با به‌کارگیری لم ۷.۴.۱، داریم

$$\mathbf{y}|\theta_1 \sim N(A_1\theta_1, C_1) \equiv (\mathbf{y}_t - \tilde{\mathbf{u}}_t)|\beta_t \sim N(X_t\beta_t, \tau_t^2 I_n)$$

$$\theta_1|\theta_2 \sim N(A_2\theta_2, C_2) \equiv \beta_t|\beta_{t-1} \sim N(\beta_{t-1}, \Sigma_\eta)$$

و

$$\theta_1|\theta_3 \sim N(A_3\theta_3, C_3) \equiv \beta_t|\beta_{t+1} \sim N(\beta_{t+1}, \Sigma_\eta)$$

که در آن $A_1 = X_t$، $C_1 = \tau_t^2 I_n$، $A_2\theta_2 = \beta_{t-1}$، $C_2 = \Sigma_\eta$، $A_3\theta_3 = \beta_{t+1}$ و $C_3 = \Sigma_\eta$. در نتیجه، بنا به حکم لم ۷.۴.۱

$$\beta|\mathbf{y}, \zeta_{-\beta_t} \sim N_p(Bb, B)$$

که در آن

$$B^{-1} = C_2^{-1} + C_3^{-1} + A_1^T C_1^{-1} A_1 = 2\Sigma_\eta^{-1} + \frac{1}{\tau_t^2} X_t^T X_t$$

و

$$b = C_2^{-1} A_2\theta_2 + C_3^{-1} A_3\theta_3 + A_1^T C_1^{-1}\mathbf{y} = \Sigma_\eta^{-1}(\beta_{t-1} + \beta_{t+1}) + X_t^T \frac{(\mathbf{y}_t - \tilde{\mathbf{u}}_t)}{\tau_t^2}.$$

برای $t = N_t$ واضح است که زمان بعدی را نداریم. در این حالت، با لحاظ نکردن $N(\beta_t|\beta_{t+1}, \Sigma_\eta)$ در درستنمایی به نتیجه مورد نظر خواهیم رسید. □

توزیع شرطی کامل $\mathbf{w}_t^*$

قضیه ۳.۳.۴. *با تعریف $F_t(\theta_t) = C_t(\theta_t)^T C_t^*(\theta_t)^{-1}$، توزیع شرطی کامل $\mathbf{w}_t^*$ عبارتست از*

$$\mathbf{w}_t^*|\mathbf{y}, \zeta_{-w_t^*} \sim N\left(\mu_{\mathbf{w}_t^*|.}, \Sigma_{\mathbf{w}_t^*|.}\right)$$

که در آن

$$\Sigma_{\mathbf{w}_t^*|.} = \left(C_t^*(\theta_t)^{-1} + F_t(\theta_t)^T \tilde{D}_t^{-1} F_t(\theta_t)\right)^{-1}$$

و

$$\mu_{\mathbf{w}_t^*|.} = \Sigma_{\mathbf{w}_t^*|.} F_t(\theta_t)^T \tilde{D}_t^{-1}(\tilde{\mathbf{u}}_t - \tilde{\mathbf{u}}_{t-1}).$$

برهان. با حذف مولفه‌هایی که مستقل از پارامتر $\mathbf{w}_t^*$ هستند، توزیع شرطی مورد نظر متناسب است با

$$N(\mathbf{w}_t^*|0, C_t^*(\theta_t)) \times N\left(\tilde{\mathbf{u}}_t - \tilde{\mathbf{u}}_{t-1}|C_t(\theta_t)^T C_t^*(\theta_t)^{-1}\mathbf{w}_t^*, \tilde{D}_t\right).$$

در این حالت نیز با استفاده از لم ۶.۴.۱، داریم

$$\mathbf{y}|\theta_1 \sim N(A_1\theta_1, C_1) \equiv (\tilde{\mathbf{u}}_t - \tilde{\mathbf{u}}_{t-1})|\mathbf{w}_t^* \sim N\left(C_t(\theta_t)^T C_t^*(\theta_t)^{-1}\mathbf{w}_t^*, \tilde{D}_t\right)$$

و

$$\theta_1|\theta_2 \sim N(A_2\theta_2, C_2) \equiv \mathbf{w}_t^* \sim N(0, C_t^*(\theta_t))$$

که در آن $A_1 = F_t(\theta_t)$، $C_1 = \tilde{D}_t$، $A_2\theta_2 = 0$ و $C_2 = C_t^*(\theta_t)$. بنابراین، بنا به بند ب لم ۶.۴.۱

$$w_t^*|\mathbf{y}, \zeta_{-w_t^*} \sim N_p(Bb, B)$$

که در آن

$$B^{-1} = C_2^{-1} + A_1^T C_1^{-1} A_1 = \left(C_t^*(\theta_t)^{-1} + F_t(\theta_t)^T \tilde{D}_t^{-1} F_t(\theta_t)\right)$$

و

$$b = C_2^{-1} A_2\theta_2 + A_1^T C_1^{-1}\mathbf{y} = F_t(\theta_t)^T \tilde{D}_t^{-1}(\tilde{\mathbf{u}}_t - \tilde{\mathbf{u}}_{t-1}).$$

□

توزیع شرطی کامل $\tilde{\mathbf{u}}_t$

قضیه ۴.۳.۴. *توزیع شرطی کامل مولفه $\tilde{\mathbf{u}}_t$ عبارتست از $\tilde{\mathbf{u}}_t \sim N\left(\mu_{\tilde{\mathbf{u}}_t|.}, \Sigma_{\tilde{\mathbf{u}}_t|.}\right)$ که در آن*

$$\Sigma_{\tilde{\mathbf{u}}_t|.} = \left(\tilde{D}_t^{-1} + \tilde{D}_{t+1}^{-1} + \frac{1}{\tau_t^2} I_n\right)^{-1}$$

و

$$\mu_{\tilde{\mathbf{u}}_t|.} = \Sigma_{\tilde{\mathbf{u}}_t|.} \left[\tilde{D}_t^{-1}\left(\tilde{\mathbf{u}}_{t-1} - F_t(\theta_t)\mathbf{w}_t^*\right) + \tilde{D}_{t+1}^{-1}\left(\tilde{\mathbf{u}}_{t+1} - F_{t+1}(\theta_{t+1})\mathbf{w}_{t+1}^*\right) + \frac{\mathbf{y}_t - X_t\beta_t}{\tau_t^2}\right].$$

همچنین برای $t = N_t$، $\tilde{D}_{t+1} = 0$ می‌باشد.

برهان. توزیع شرطی کامل $\tilde{\mathbf{u}}_t$ متناسب است با

$$N\left(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t-1} + F_t(\theta_t)\mathbf{w}_t^*, \tilde{D}_t\right) \times N\left(\tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1} - F_{t+1}(\theta_{t+1})\mathbf{w}_{t+1}^*, \tilde{D}_{t+1}\right)$$
$$\times N\left((\mathbf{y}_t - X_t\beta_t)|\tilde{\mathbf{u}}_t, \tau_t^2 I_n\right).$$

با استفاده از لم ۷.۴.۱، نتیجه می‌شود

$$\mathbf{y}|\theta_1 \sim N(A_1\theta_1, C_1) \equiv (\mathbf{y}_t - X_t\beta_t)|\tilde{\mathbf{u}}_t \sim N\left(\tilde{\mathbf{u}}_t, \tau_t^2 I_n\right)$$

$$\theta_1|\theta_2 \sim N(A_2\theta_2, C_2) \equiv \tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t-1} \sim N\left(\tilde{\mathbf{u}}_{t-1} + F_t(\theta_t)\mathbf{w}_t^*, \tilde{D}_t\right)$$

و

$$\theta_1|\theta_3 \sim N(A_3\theta_3, C_3) \equiv \tilde{\mathbf{u}}_t|\tilde{\mathbf{u}}_{t+1} \sim N\left(\tilde{\mathbf{u}}_{t+1} - F_{t+1}(\theta_{t+1})\mathbf{w}_{t+1}^*, \tilde{D}_{t+1}\right)$$

که در آن $A_1 = I_n$، $C_1 = \tau_t^2$، $A_2\theta_2 = \tilde{\mathbf{u}}_{t-1} + F_t(\theta_t)\mathbf{w}_t^*$، $C_2 = \tilde{D}_t$، $C_3 = \tilde{D}_{t+1}$ و

$$A_3\theta_3 = \tilde{\mathbf{u}}_{t+1} - F_{t+1}(\theta_{t+1})\mathbf{w}_{t+1}^*.$$

بنابراین، بنا به حکم لم ۷.۴.۱

$$\tilde{\mathbf{u}}_t|\mathbf{y}, \zeta_{-\tilde{\mathbf{u}}_t} \sim N_p(Bb, B)$$

که در آن

$$B^{-1} = C_2^{-1} + C_3^{-1} + A_1^T C_1^{-1} A_1 = \left(\tilde{D}_t^{-1} + \tilde{D}_{t+1}^{-1} + \frac{1}{\tau_t^2} I_n\right)$$

و

$$b = C_2^{-1} A_2\theta_2 + C_3^{-1} A_3\theta_3 + A_1^T C_1^{-1}\mathbf{y}$$
$$= \tilde{D}_t^{-1}\left(\tilde{\mathbf{u}}_{t-1} + F_t(\theta_t)\mathbf{w}_t^*\right) + \tilde{D}_{t+1}^{-1}\left(\tilde{\mathbf{u}}_{t+1} - F_{t+1}(\theta_{t+1})\mathbf{w}_{t+1}^*\right) + \frac{\mathbf{y}_t - X_t\beta_t}{\tau_t^2}.$$

□

توزیع شرطی کامل $\tau_t^۲$

قضیه ۵.۳.۴. *توزیع شرطی کامل $\tau_t^۲$ یک توزیع گاما به صورت $IG\left(a_{\tau|.}, b_{\tau|.}\right)$ است که در آن*

$$a_{\tau|.} = a_\tau + \frac{n}{۲}$$

و

$$b_{\tau|.} = b_\tau + \frac{۱}{۲}\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right)^T\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right).$$

برهان. توزیع شرطی کامل برای $\tau_t^۲$ متناسب است با

$$IG\left(\tau_t^۲|a_\tau, b_\tau\right) \times N\left((\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t)|_\circ, \tau_t^۲ I_n\right).$$

بنابراین

$$\begin{aligned}
\pi(\tau_t^۲|\mathbf{y}, \zeta_{-\tau_t^۲}) &\propto \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)}\left(\frac{۱}{\tau_t^۲}\right)^{a_\tau+۱}\exp(-\frac{b_\tau}{\tau_t^۲}) \times \frac{۱}{\left(۲\pi\tau_t^۲\right)^{\frac{n}{۲}}|I_n|^{\frac{۱}{۲}}} \\
&\times \exp\left\{-\frac{۱}{۲}\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right)^T\frac{۱}{\tau_t^۲ I_n}\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right)\right\} \\
&\propto \left(\frac{۱}{\tau_t^۲}\right)^{a_\tau+\frac{n}{۲}+۱}\exp\left\{-\frac{b_\tau}{\tau_t^۲} - \frac{\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right)^T\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right)}{۲\tau_t^۲}\right\} \\
&\propto IG\left(a_\tau + \frac{n}{۲}, b_\tau + \frac{۱}{۲}\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right)^T\left(\mathbf{y}_t - X_t\beta_t - \tilde{\mathbf{u}}_t\right)\right).
\end{aligned}$$

□

توزیع شرطی کامل θ_t

پارامترهای وابستگی فرآیند، مجموعه $\theta_t = \{\sigma_t^۲, \theta_{۱t}, \theta_{۲t}\}$ است. اگر فرض کنیم توزیع پیشین این پارامترها به صورت $IG(\sigma_t^۲|a_\sigma, b_\sigma)\pi(\theta_{۱t}, \theta_{۲t})$ است، آن‌گاه به‌راحتی، مشابه $\tau_t^۲$، می‌توان نشان داد توزیع شرطی کامل $\sigma_t^۲$ یک توزیع گامای معکوس به صورت $IG(a_{\sigma|.}, b_{\sigma|.})$ است.

قضیه ۶.۳.۴. *توزیع شرطی کامل $\sigma_t^۲$ یک توزیع گامای معکوس به صورت $IG(a_{\sigma|.}, b_{\sigma|.})$ است به‌طوری که*

$$a_{\sigma|.} = a_\sigma + \frac{n^*}{۲}$$

و

$$b_{\sigma|.} = b_\sigma + \frac{۱}{۲}\mathbf{w}_t^{*T}R_t^*(\theta_t)^{-۱}\mathbf{w}_t^*$$

که در آن $C_t^*(\theta_t) = \sigma_t^۲ R_t^*(\theta_t)$.

برهان. مشابه برهان $\tau_t^۲$، است. □

توزیع شرطی کامل بقیه مولفه‌های θ_t صورت بسته ندارد و مرحله نمونه‌گیری از آن‌ها باید به کمک یک الگوریتم متروپولیس-هستینگز قدم زدن تصادفی اجرا شود. توزیع پیشنهادی برای اجرای الگوریتم RWMH به صورت زیر تعریف می‌شود:

$$\pi(\theta_{1t}, \theta_{2t}) N(\mathbf{w}_t^* | 0, C_t^*(\theta_t)) N(\tilde{\mathbf{u}}_t | \tilde{\mathbf{u}}_{t-1} + F_t(\theta_t)\mathbf{w}_t^*, \tilde{D}_t)$$

که $F_t(\theta_t)$ در توزیع شرطی کامل $\mathbf{w}_t^*$ تعریف شده است.

توزیع شرطی کامل Σ_η

قضیه ۷.۳.۴. *با در نظرگرفتن توزیع ویشارت معکوس برای پارامتر Σ_η، یعنی*

$$\pi(\Sigma_\eta) = \frac{1}{|\Sigma_\eta|^{r_\eta+(p+1)/2}} \exp\left\{-\frac{1}{2} tr\left(\Upsilon_\eta^{-1}\Sigma_\eta^{-1}\right)\right\}$$

توزیع شرطی کامل نیز ویشارت معکوس خواهد بود که تنها پارامترهای آن به‌روز می‌شوند. در واقع، نمونه‌های این پارامتر از توزیع شرطی کامل $IW(r_{\eta|.}, \Upsilon_{\eta|.})$ تولید می‌شوند که در آن

$$r_{\eta|.} = r_\eta + N_t$$

و

$$\Upsilon_{\eta|.} = \left[\Upsilon_\eta^{-1} + \Sigma_{t=1}^{N_t} (\beta_t - \beta_{t-1})(\beta_t - \beta_{t-1})^T\right]^{-1}.$$

برهان. توزیع شرطی کامل پارامتر Σ_η متناسب است با

$$IW(r_\eta, \Upsilon_\eta) \times \prod_{t=1}^{N_t} N(\beta_t | \beta_{t-1}, \Sigma_\eta).$$

بنابراین

$$\begin{aligned}\pi(\Sigma_\eta | \mathbf{y}, \zeta_{-\Sigma_\eta}) \propto & \frac{1}{|\Sigma_\eta|^{r_\eta+(p+1)/2}} \exp\left\{-\frac{1}{2} tr\left(\Upsilon_\eta^{-1}\Sigma_\eta^{-1}\right)\right\} \\ & \times \frac{1}{(2\pi)^{\frac{N_t}{2}} |\Sigma_\eta|^{\frac{N_t}{2}}} \exp\left\{-\frac{1}{2}\Sigma_{t=1}^{N_t} (\beta_t - \beta_{t-1})^T \Sigma_\eta^{-1} (\beta_t - \beta_{t-1})\right\}.\end{aligned}$$

برای تکمیل برهان، از خواص دترمینان و اثر ماتریس استفاده می‌کنیم. در ابتدا، خاصیت تفکیک‌پذیری دترمینان دو ماتریس، به ما اجازه می‌دهد که $|\Sigma_\eta|^{-\frac{N_t}{2}}$ را از تابع چگالی احتمال نرمال چندمتغیره با $|\Sigma_\eta|^{-\frac{r_\eta+(p+1)}{2}}$ از تابع چگالی احتمال ویشارت معکوس، ترکیب کنیم. همچنین

$$\sum_{t=1}^{N_t} \left((\beta_t - \beta_{t-1})^T \Sigma_\eta^{-1} (\beta_t - \beta_{t-1})\right) = tr\left[\sum_{t=1}^{N_t} \left((\beta_t - \beta_{t-1})^T \Sigma_\eta^{-1} (\beta_t - \beta_{t-1})\right)\right].$$

سپس جابجایی‌های زیر را می‌توانیم داشته باشیم

$$\begin{aligned} tr\left[\sum_{t=1}^{N_t} \left((\beta_t - \beta_{t-1})^T \Sigma_\eta^{-1} (\beta_t - \beta_{t-1})\right)\right] &= tr\left[\sum_{t=1}^{N_t} \left((\beta_t - \beta_{t-1})(\beta_t - \beta_{t-1})^T \Sigma_\eta^{-1}\right)\right] \\ &= tr\left[\left(\sum_{t=1}^{N_t} (\beta_t - \beta_{t-1})(\beta_t - \beta_{t-1})^T\right) \Sigma_\eta^{-1}\right]. \end{aligned}$$

سرانجام، عبارت زیر

$$tr\left[\left(\sum_{t=۱}^{N_t}(\beta_t-\beta_{t-۱})(\beta_t-\beta_{t-۱})^T\right)\Sigma_\eta^{-۱}\right]$$

را با $tr\left(\Upsilon_\eta^{-۱}\Sigma_\eta^{-۱}\right)$ می‌توان جمع کرد. بنابراین

$$\begin{aligned}
\pi(\Sigma_\eta|\mathbf{y},\zeta_{-\Sigma_\eta}) &\propto \frac{۱}{|\Sigma_\eta|^{r_\eta+(p+۱)/۲}}\exp\left\{-\frac{۱}{۲}tr\left(\Upsilon_\eta^{-۱}\Sigma_\eta^{-۱}\right)\right\}\\
&\times\frac{۱}{|\Sigma_\eta|^{\frac{N_t}{۲}}}\exp\left\{-\frac{۱}{۲}tr\left(\left(\sum_{t=۱}^{N_t}\left(\beta_t-\beta_{t-۱}\right)\left(\beta_t-\beta_{t-۱}\right)^T\right)\Sigma_\eta^{-۱}\right)\right\}\\
&=\frac{۱}{|\Sigma_\eta|^{r_\eta+N_t+(p+۱)/۲}}\exp\left[-\frac{۱}{۲}tr\left(\Upsilon_\eta^{-۱}\Sigma_\eta^{-۱}\right)\right.\\
&\left.-\frac{۱}{۲}tr\left(\left(\sum_{t=۱}^{N_t}\left(\beta_t-\beta_{t-۱}\right)\left(\beta_t-\beta_{t-۱}\right)^T\right)\Sigma_\eta^{-۱}\right]\right.\\
&=\frac{۱}{|\Sigma_\eta|^{r_\eta+N_t+(p+۱)/۲}}\\
&\exp\left[-\frac{۱}{۲}tr\left(\left(\Upsilon_\eta^{-۱}+\sum_{t=۱}^{N_t}\left(\beta_t-\beta_{t-۱}\right)\left(\beta_t-\beta_{t-۱}\right)^T\right)\Sigma_\eta^{-۱}\right)\right].
\end{aligned}$$

در نتیجه

$$\Sigma_\eta|\mathbf{y},\zeta_{-\Sigma_\eta}\sim IW\left(r_\eta+N_t,\Upsilon_\eta^{-۱}+\sum_{t=۱}^{N_t}(\beta_t-\beta_{t-۱})(\beta_t-\beta_{t-۱})^T\right)$$

و برهان کامل می‌شود. □

۲.۳.۴ ارزیابی عملکرد مدل

برای ارزیابی عملکرد مدل می‌توان از روش شبیه‌سازی مشاهدات بر اساس مدل برازش‌شده و مقایسه آن‌ها با مقادیر واقعی، استفاده کرد. برای این منظور، باید توزیع پسین پیش‌گو یعنی

$$f(\mathbf{y}_{rep,t}(\mathbf{s}_i)|\mathbf{y})=\int N(\mathbf{y}_{rep,t}(\mathbf{s}_i);\mathbf{x}_t(\mathbf{s}_i)^T\beta_t+\tilde{\mathbf{u}}_t(\mathbf{s}_i),\tau_t^۲)\pi(\beta_t,\tilde{\mathbf{u}}_t(\mathbf{s}_i),\tau_t^۲|\mathbf{y})$$

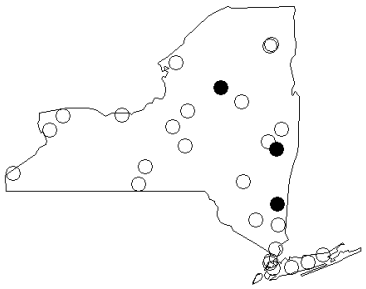
را برای هر نقطه زمانی t محاسبه کنیم که در آن $\pi\left(\beta_t,\tilde{\mathbf{u}}_t(\mathbf{s}_i),\tau_t^۲|\mathbf{y}\right)$ توزیع پسین کناری پارامترهای $\{\beta_t,\tilde{\mathbf{u}}_t,\tau_t^۲\}$ است. شبیه‌سازی از توزیع پسین پیش‌گو با وجود نمونه‌های MCMC تولیدشده کار سرراستی است. کافی است به ازای نمونه‌های تولیدشده از توزیع پسین کناری $\left\{\beta_t,\tilde{\mathbf{u}}_t(\mathbf{s}_i),\tau_t^۲\right\}$، مشاهدات از توزیع $N(\mathbf{x}_t(\mathbf{s}_i)^T\beta_t+\tilde{\mathbf{u}}_t(\mathbf{s}_i),\tau_t^۲)$ تولید شوند.

همانند آن‌چه در فصل قبل گفته شد، وقتی متغیر پاسخ در مدل‌های فضایی-زمانی گسسته است می‌توانیم با تعمیمی مشابه فصل ۳، با انتخاب تابع پیوند مناسب، از چارچوب مدل‌های خطی تعمیم‌یافته برای مدل‌بندی داده‌ها استفاده کنیم. در این صورت مدل حاصل را، مدل آمیخته خطی تعمیم‌یافته فضایی-زمانی می‌نامیم.

۴.۴ مثال شبیه‌سازی

مدل پویای (۱.۴) و فرآیند پیش‌گوی آن توسط تابع spDynLM در نرم‌افزار R قابل اجراست. در این بخش، مدل پررتبه را با استفاده از مجموعه داده حاصل از ایستگاه‌های نظارت اوزن که توسط ساهو و باکر (۲۰۱۱) گزارش و تحلیل شده‌اند، تحلیل کرده‌ایم. این مجموعه داده به نسبت کوچک هستند و به کاهش بعد احتیاج ندارند.

مجموعه داده‌ها، ۲۸ ایستگاه نظارت را شامل می‌شوند که سازمان حفاظت محیط زیست آمریکا آن را از ۱ جولای تا ۳۱ آگوست ۲۰۰۶ جمع‌آوری کرده است. متغیر پاسخ، حداکثر غلظت اوزن در ۸ ساعت از یک روز (0.38HRMAX) است و متغیرهای تبیینی شامل حداکثر دمای هوا (cMAXTMP)، سرعت باد (WDSP) و رطوبت نسبی (RM) هستند. تعداد کل داده‌ها بر اساس $n = ۲۸$ مکان و $N_t = ۶۲$ روز، برابر ۱۷۳۶ مشاهده است که ۱۱۴ تای آن‌ها از دست رفته یا گمشده[۲] هستند. در این مطالعه شبیه‌سازی از متغیرهای تبیینی که دارای ساختار باقیمانده فضایی و زمانی هستند، برای پیش‌گویی مقادیر گمشده متغیر پاسخ استفاده می‌کنیم. برای به‌دست آوردن درک بهتر از عملکرد مدل‌های فضایی-زمانی معرفی‌شده و توان پیش‌گویی، نیمی از مشاهدات ذخیره‌شده در سه ایستگاه را برای اعتبارسنجی مدل کنار گذاشتیم؛ به عبارت دیگر این داده‌های کنار گذاشته‌شده در برازش مدل حضور ندارند. شکل ۱.۴ مکان ایستگاه‌های نظارت و سه ایستگاهی که داده‌های آن‌ها لحاظ نشده‌اند را نشان می‌دهد.



شکل ۱.۴: دایره‌ها نشان‌دهنده مکان ایستگاه‌های نظارت اوزن در ایالت نیویورک است. دایره‌های توپر ایستگاه‌هایی هستند که نیمی از داده‌های آن برای ارزیابی عملکرد پیش‌گویی مدل لحاظ نشده‌اند.

برای برازش مدل باید مختصات ایستگاه‌ها، مقادیر میزان‌ساز و توزیع‌های پیشین پارامترهای مدل ϕ_t، $\sigma_t^۲$ و $\tau_t^۲$ تعیین شوند. تابع کوواریانس فرآیند فضایی با استفاده از تابع همبستگی نمایی

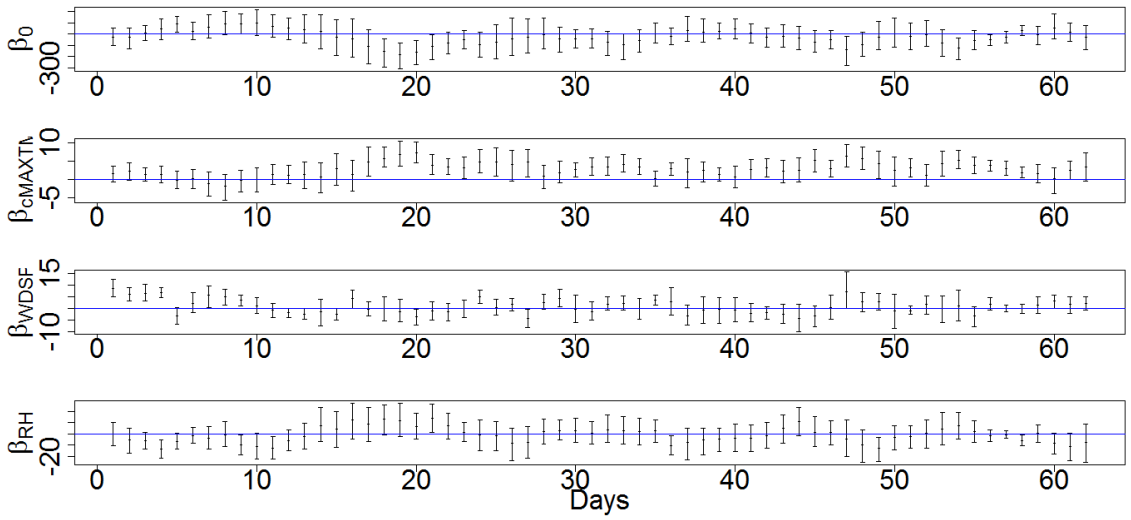
$$\rho(\mathbf{s}, \mathbf{s}^T; \phi) = \exp(-\phi||\mathbf{s} - \mathbf{s}^T||)$$

[۲] Missing data

ساخته شده است. بنابراین

$$C(\mathbf{s}, \mathbf{s}^T; \phi) = \sigma^2 \exp(-\phi||\mathbf{s} - \mathbf{s}^T||).$$

برای این مثال، فرض کردیم β دارای توزیع پیشین نرمال باشد؛ یعنی $\beta \sim N_4(0, I_4)$ که I_4 ماتریس همانی با بعد ۴ است. برای پارامترهای τ^2 و σ^2 توزیع پیشین گامای معکوس، $IG(2, 25)$، و برای پارامتر فضایی ϕ توزیع پیشین یکنواخت، $U(0.006, 0.100)$، را اختصاص دادیم. تعداد نمونه‌های MCMC نیز ۵۰۰۰ در نظر گرفته شد. نمودار زمانی برآوردهای پارامترها برای تعبیر آن‌ها در زمان‌های

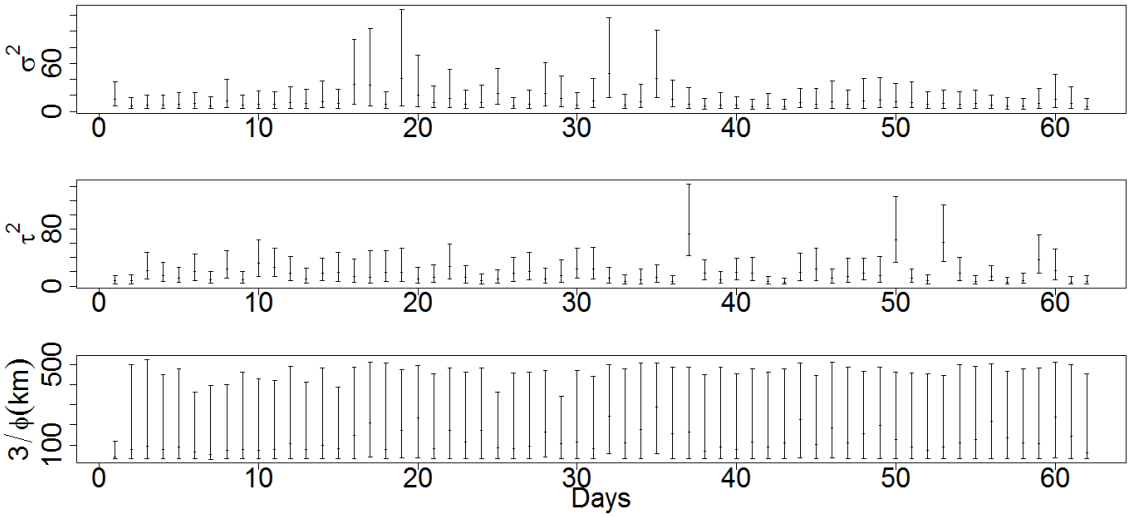


شکل ۲.۴: برآوردهای بیزی (میانه توزیع پسین) اثرات متغیرهای تبیینی به همراه فاصله اعتبار ۹۵ درصد آن‌ها

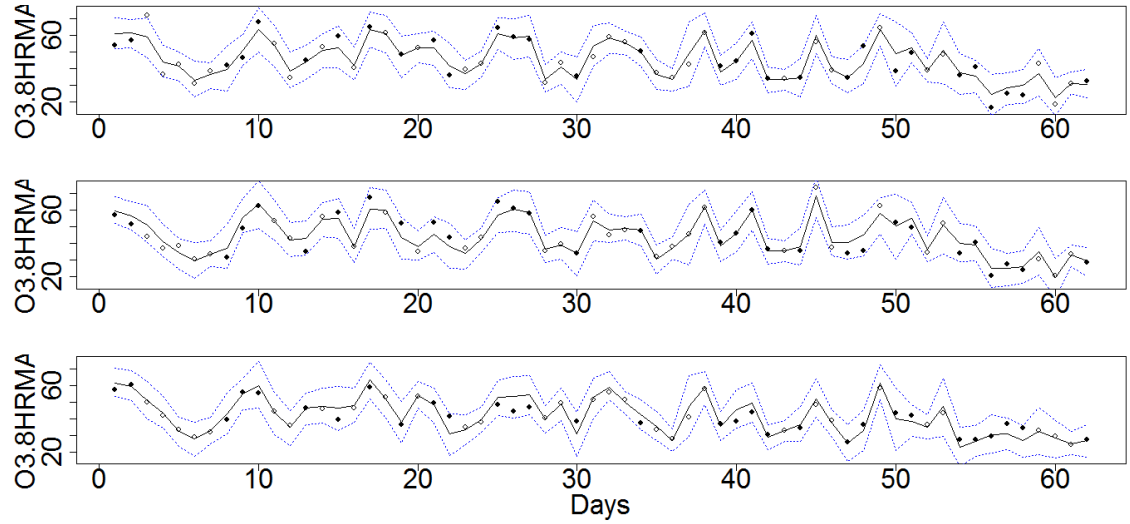
مختلف، مفید هستند. این نمودارها برای ضرایب رگرسیونی و پارامترهای وابستگی فرآیند به ترتیب در دو شکل ۲.۴ و ۳.۴ نمایش داده شده‌اند. خط میانی این دو شکل، مقدار صفر را مشخص می‌کند. در مورد ضرایب رگرسیونی، این نمودارها روند تغییرات زمانی در تأثیر متغیرهای تبیینی را بهتر شرح می‌دهند. برای نمونه، نمودار الگوی سینوسی برای ضریب ثابت β_0، یک وابستگی قوی با دو متغیر cMAXTMP و RM و وابستگی کوچکی با WDSP را نشان می‌دهد. ماهیت پویای اثر متغیرهای تبیینی بر روی متغیر پاسخ از روی شکل ۲.۴ کاملا مشخص است. البته، مثلا، فاصله اعتبار متغیر RM در تمامی زمان‌ها مقدار صفر را در بر دارد که می توان معنی‌دار نبودن اثر این متغیر را نتیجه گرفت. تاثیر متغیرها نیز در برخی زمان‌ها معکوس (منفی) و در زمان‌های دیگر مستقیم (مثبت) است.

با فقط حداکثر ۲۸ مشاهده در هر زمان، اطلاعات کافی برای برآورد پارامترهای وابستگی θ وجود ندارد. این واقعیت در شکل ۳.۴، در فاصله اعتبارهای مبهم برای ϕها و طول‌های کوتاه نواحی اعتبار برای σ^2 و τ^2 منعکس شده است. با این حال یک روند غیر قابل اغماض در مولفه واریانس، در طول زمان، مشهود است.

شکل ۴.۴ مقادیر مشاهده‌شده و پیش‌گویی‌شده برای سه ایستگاه مورد استفاده در اعتبارسنجی مدل را نمایش می‌دهد. در این شکل، دایره خالی نشانگر مشاهدات مورد استفاده برای برازش مدل و دایره پر نشانگر آن مشاهدات است که در برازش مدل لحاظ نشده و برای اعتبارسنجی مدل کنار گذاشته



شکل ۳.۴: برآوردهای بیزی (میانه توزیع پسین) پارامترهای وابستگی به همراه فاصله اعتبار ۹۵ درصد آن‌ها



شکل ۴.۴: میانه توزیع پیش‌گو و فاصله اعتبار ۹۵% آن‌ها برای سه ایستگاه حذف‌شده که به ترتیب با خطوط غیرمنقطع و خطوط منقطع نمایش داده شده‌اند. دایره‌های خالی نشان‌گر مشاهدات مورد استفاده برای برازش مدل و دایره‌های پر نشان‌گر مشاهدات آزمون (حذف‌شده) برای اعتبارسنجی مدل هستند.

شده‌اند. میانه توزیع پیش‌گو و فاصله اعتبار ۹۵% برای سه ایستگاه به ترتیب با خطوط غیرمنقطع و خطوط منقطع نمایش داده شده‌اند. مقادیر پیش‌گویی به مقادیر واقعی نزدیک هستند و همه آن‌ها درون فاصله اعتبار پیش‌گویی‌شده قرار دارند. بنابراین می‌توان به‌خوبی توانایی پیش‌گویی مدل فضایی-زمانی پیشنهادی را دریافت. همانند نتیجه ساهو و باکر (۲۰۱۱)، کاهش قابل توجهی در سطوح اوزن در ۱۰ روز آخر مشهود است.

۵.۴ عوامل موثر بر کیفیت آب‌های زیر زمینی استان گلستان

آب زیرزمینی به طور کامل خالص و تمیز نیست، و همیشه دارای مقداری مواد محلول است. مقادیر این مواد محلول، غلظت و نوع آن‌ها، شایستگی کاربرد آب برای اهداف مختلف را تعیین خواهند کرد. از این‌رو، مطالعات ژئوشیمیایی آب زیرزمینی، با توجه به کاربرد آب، مهم است. این مطالعات، به درک و فهم عمیق‌تر و بهتر فرآیندهای تکاملی و کیفی آب در منطقه مورد مطالعه، کمک خواهد کرد و می‌تواند اطلاعاتی در مورد محدودیت‌های کلی توسعه یا برنامه‌ریزی برای تصفیه مناسب آب فراهم آورد. این اطلاعات با توجه به تغییرات آتی در کیفیت منابع آبی، مورد نیاز است. زمانی‌که آب زیرزمینی در محیط زیر زمین جریان پیدا می‌کند، تغییرات شیمیایی آن با افزایش مواد جامد محلول و یون‌های اصلی، همراه است. هرچه مدت ماندگاری آب در زیر زمین زیاد باشد، کیفیت آن نامناسب‌تر خواهد بود. همچنین سنجش و مطالعه خواص شیمیایی آبخوان‌های ساحلی و حاشیه دریاچه، به دلیل تهدید آلودگی آب زیرزمینی با آب دریا و دریاچه، پیچیده است.

آب‌های زیرزمینی، در منطقه مورد مطالعه، برای اهداف مختلفی مانند شرب، کشاورزی و صنعت مورد استفاده قرار می‌گیرند. به دلیل اهمیتی که آب‌های زیرزمینی در منطقه مورد مطالعه دارد، پایش، مطالعه و شناخت فرآیندها و پارامترهای حاکم و موثر بر کیفیت آب زیرزمینی، از نظر شرب، کشاورزی، و صنعت، ضروری است. کیفیت آب زیرزمینی اساساً به کیفیت کاتیون‌ها و آنیون‌های موجود در آن بستگی دارد. تحلیل‌های شیمیایی بر اساس تفسیر کیفیت آب در ارتباط با منشا، زمین‌شناسی، اقلیم و استفاده آب انجام می‌گیرند.

توانایی اندازه‌گیری آب برای عبور جریان الکتریکی، هدایت الکتریکی[۳] (EC) نامیده می‌شود. هدایت الکتریکی تابعی از غلظت آنیون‌ها، کاتیون‌ها و دمای محلول بوده و با مقدار کل جامدات محلول[۴] (TDS) نیز وابستگی دارد و یکی از سریع‌ترین روش‌های ارزیابی کلی کیفیت آب‌های زیرزمینی می‌باشد. قابلیت هدایت الکتریکی یا عکس آن، یعنی مقاومت الکتریکی، یکی از عوامل مهم در تحقیقات هیدروژئولوژی می‌باشد و در اندازه‌گیری آن احتیاجی به دستگاه‌های پچیده نبوده و مشکلات نمونه‌برداری وجود ندارد. همچنین اندازه‌گیری آن سریع و دقیق است. مقدار استاندارد تعیین‌شده برای EC توسط سازمان بهداشت جهانی ۵۰۰ میکروموس است. با توجه به این ویژگی‌های EC، بر آن شدیم تا کیفیت آب‌های زیرزمینی بخشی از استان گلستان را مورد مطالعه قرار دهیم.

همان‌طور که در مثال قبلی اشاره کردیم، مدل فرآیند پررتبه فضایی-زمانی و فرآیند پیش‌گوی مرتبط با آن، توسط تابع spDynLM در نرم افزار R اجرا می‌شود. در این مطالعه، ابتدا مدل پررتبه و سپس مدل پیش‌گو را ارزیابی خواهیم کرد.

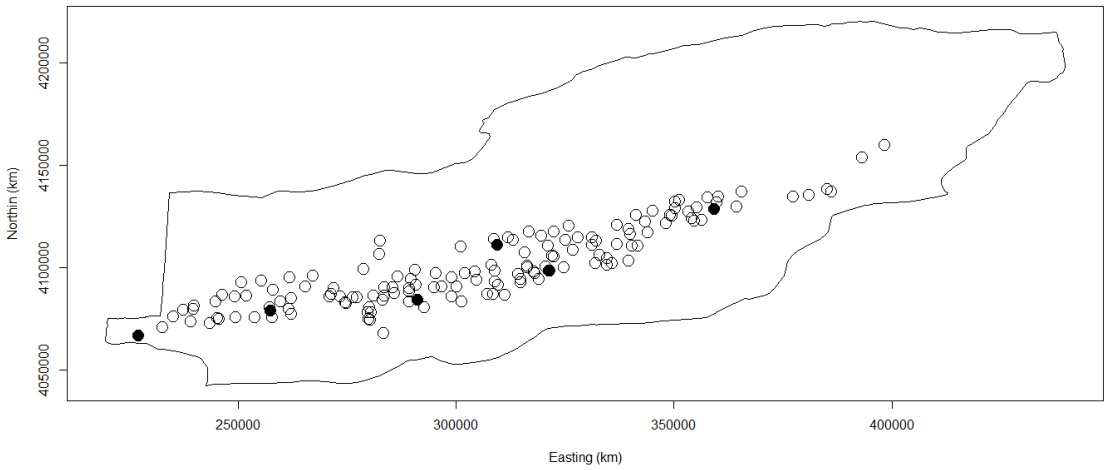
داده‌های موجود شامل اندازه‌های ثبت‌شده در ۱۵۰ ایستگاه از آب‌های زیرزمینی استان گلستان است که از آبان‌ماه سال ۸۲ تا آبان‌ماه سال ۹۲ گردآوری شده‌اند. منظور از ایستگاه‌های آب زیرزمینی، چشمه‌ها و چاه‌های عمیق و نیمه‌عمیق است و زمان‌های نمونه‌گیری نیز در دو ماه سال یعنی اردیبهشت و آبان

[۳] Electric conductivity

[۴] Total dissolved solution

انجام گرفته‌اند. بنابراین مجموعه داده‌ها شامل ۱۵۰ نقطه مکانی (سطر) و ۲۱ نقطه زمانی (ستون) و در مجموع ۳۱۵۰ مشاهده فضایی-زمانی است. متغیر پاسخ EC و متغیرهای تبیینی شامل سختی کل[۵] (th)، پتاسیم (k)، سدیم (na)، سولفات (so4)، کلر (cl) ، بیکربنات (hco3)، ph و کل جامدات محلول هستند. این داده‌ها را با هر دو مدل پررتبه (۱.۴) و فرآیند پیش‌گو (۲.۴) با انتخاب ۲۵، ۵۰، ۷۵ و ۱۰۰ گره، مورد تحلیل قرار دادیم.

برای ارزیابی عملکرد مدل پررتبه، در ۶ ایستگاه (علامت‌های دایره توپر در شکل ۵.۴) ۱۰ نقطه زمانی را کنار گذاشته و مدل را با بقیه داده‌ها برازش دادیم. به‌طور مشابه، برای مدل فرآیندهای پیش‌گو با انتخاب گره‌های مختلف، ۲ ایستگاه برای ۲۵ گره و ۳ ایستگاه برای ۵۰، ۷۵ و ۱۰۰ گره (شکل ۱۵.۴)، حذف شدند. مکان‌های جغرافیایی ایستگاه‌های نمونه‌گیری در شکل ۵.۴ نمایش داده شده‌اند. بر اساس تحلیل اکتشافی داده‌ها بر اساس داده‌های روند زدوده (شکل ۶.۴)، برای شش زمان ۱، ۲، ۵،



شکل ۵.۴: پراکنش جغرافیایی ایستگاه‌های آب زیرزمینی

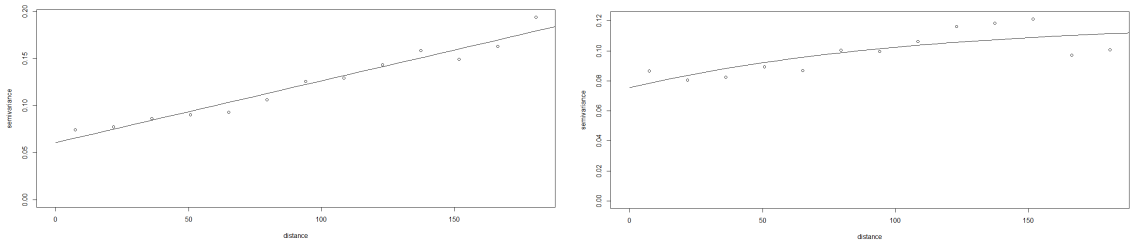
۱۰، ۱۵و ۲۰ به عنوان نمونه، تابع تغییرنگار نمایی برازش خوبی به تغییرنگار تجربی داده‌ها دارد. در نتیجه تابع کوواریانس نمایی را برای ساختار وابستگی فضایی انتخاب کردیم. برای ضرایب رگرسیونی، توزیع پیشین نرمال $\beta \sim N(۰, I_۸)$ را انتخاب کردیم. همچنین برای ضرایب وابستگی فضایی، توزیع گامای معکوس

$$\phi_t, \sigma_t^۲, \tau_t^۲ \sim IG(۲, ۲۵)$$

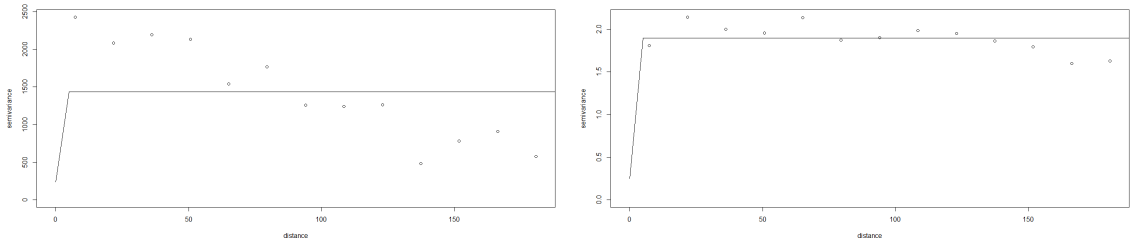
انتخاب شد. افزون بر این، فرض کردیم $\Sigma_\eta \sim IW(۲, diag(۰/۰۰۱, ۸))$ که در آن $diag(۰/۰۰۱, ۸)$ یک ماتریس قطری با بعد ۸ با درآیه‌های روی قطر ۰/۰۰۱ است. با این مشخصات، مدل پررتبه برازش داده شد.

برآوردهای بیزی ضرایب رگرسیونی وابسته به زمان به همراه فواصل اعتبار ۹۵ درصد آن‌ها در شکل‌های ۷.۴ و ۸.۴ در طول ۲۱ زمان نمایش داده شده‌اند. این شکل‌ها می‌توانند در توضیح تغییرپذیری

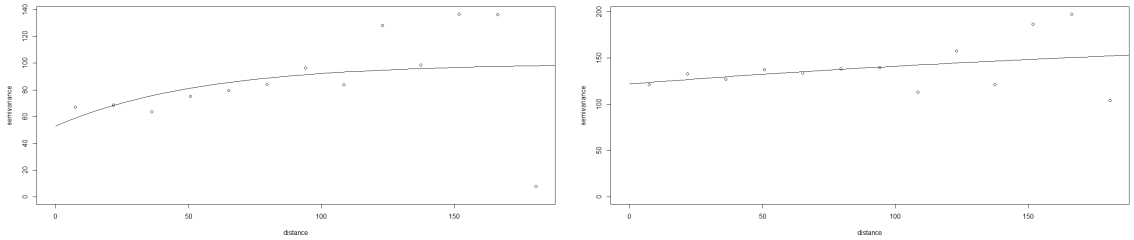
[۵] Total hardness



(آ) زمان اول (ب) زمان دوم



(ج) زمان پنجم (د) زمان دهم

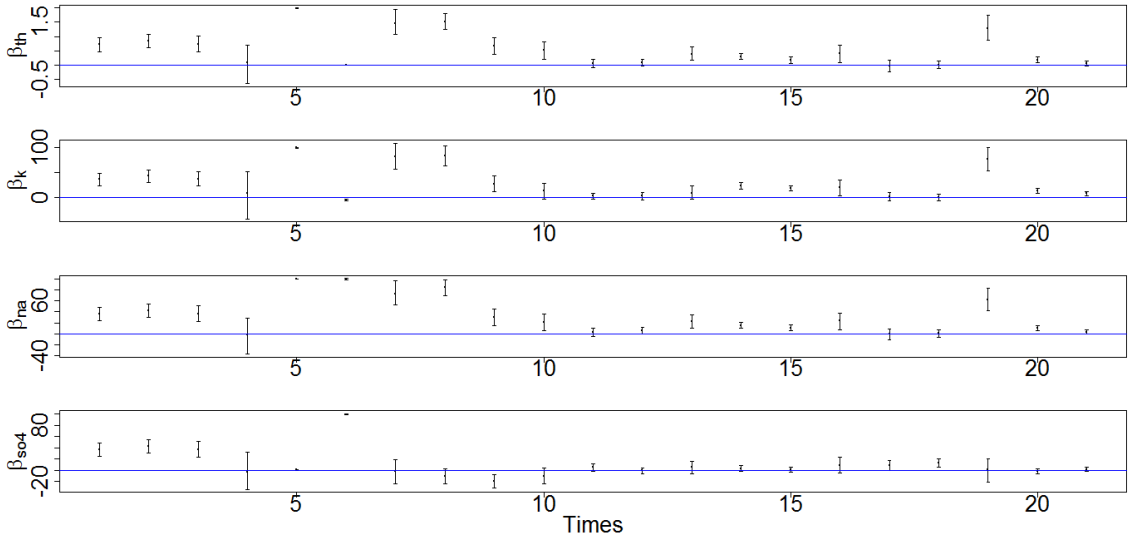


(ه) زمان پانزدهم (و) زمان بیستم

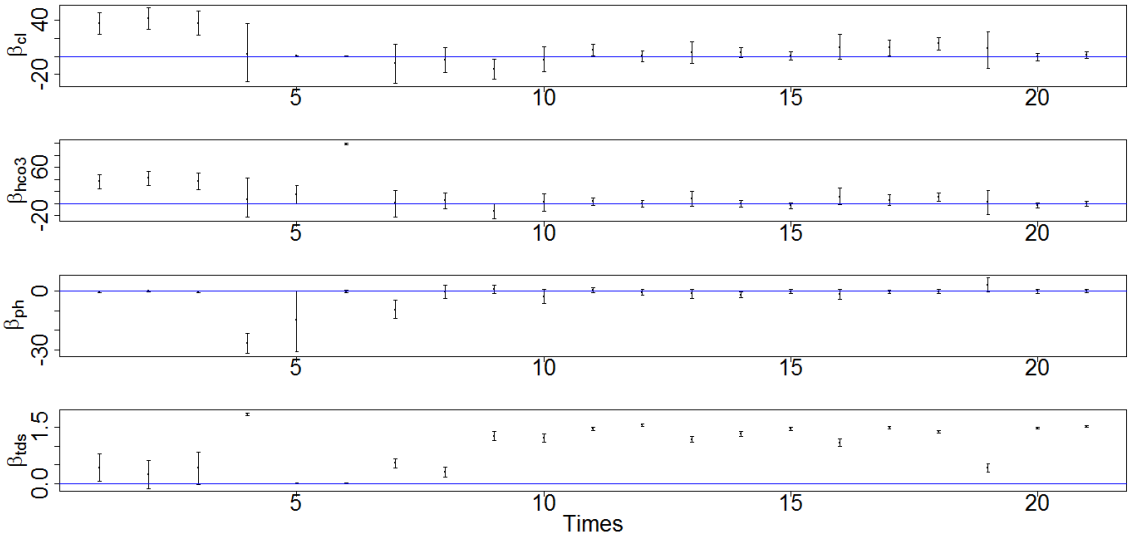
شکل ۶.۴: تغییرنگار برازش‌شده به تغییرنگار تجربی داده‌ها در شش زمان ۱، ۲، ۵، ۱۰، ۱۵و ۲۰

زمانی ضرایب مفید باشند. از روی شکل‌ها می‌توان تاثیر زمانی هر کدام از متغیرهای تبیینی را بر روی EC تعبیر کرد. همان‌طور که در شکل ۷.۴ می‌بینیم، متغیر سختی کل، پتاسیم و سدیم در اکثر زمان‌ها تاثیر مثبتی بر متغیر پاسخ، هدایت الکتریکی، دارند و بین همین سه متغیر هم رفتار مشابهی در زمان‌های مختلف دیده می‌شود. اما متغیر سولفات، در زمان‌های مختلف، یک موج کسینوسی دارد و دارای هر دو اثر مثبت و منفی نسبت به متغیر پاسخ است. همچنین با توجه به شکل ۸.۴، می‌توان گفت متغیرهای تبیینی کلر و بیکربنات، به‌جز زمان‌های ۷ تا ۱۰، ۹ و ۲۰، تاثیر مثبتی بر متغیر پاسخ دارد. همچنین متغیر ph، در اکثر زمان‌ها، تاثیری بی‌معنی دارد و حتی در مواردی تاثیر عکس بر رفتار متغیر پاسخ دارا می‌باشد. متغیر کل جامدات محلول نیز بیشترین تاثیر مثبت بر هدایت الکتریکی را دارد.

بنابراین، با توجه به نتایج به‌دست آمده، هر چه متغیرهای سختی کل، پتاسیم، سدیم و کل جامدات محلول مقادیر بزرگتری اختیار کنند، EC نیز افزایش می‌یابد و در نتیجه کیفیت آب کاهش می‌یابد. بنابراین برای جلوگیری از کاستن کیفیت آب باید جلوی افزایش این مقادیر را در آب‌های زیرزمینی گرفت. متغیر ph تاثیر معنی‌داری بر هدایت الکتریکی ندارد ولی در بعضی از زمان‌ها بیشتر بودن آن به افزایش کیفیت آب کمک می‌کند.



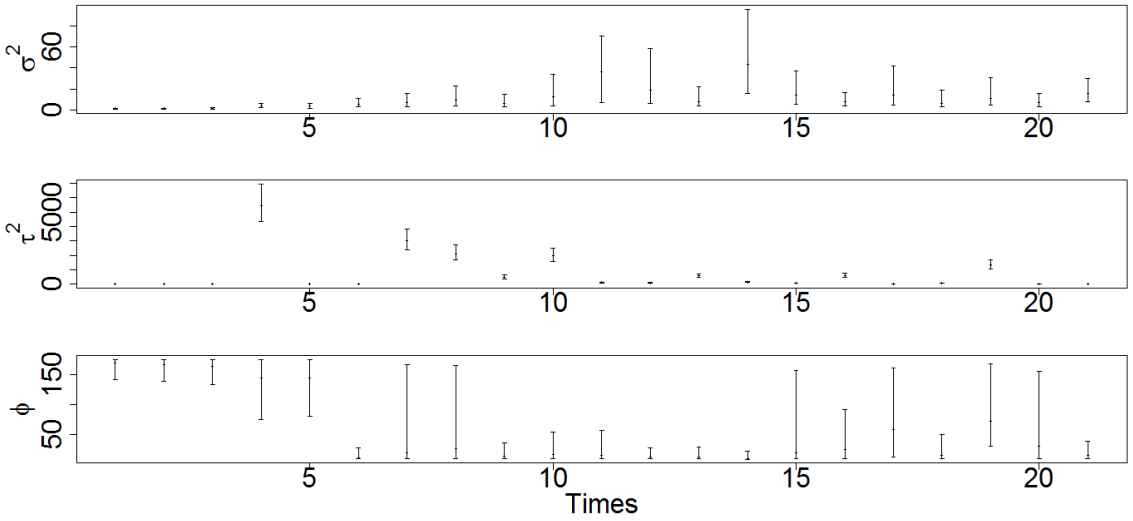
شکل ۷.۴: برآورد (میانه) و فواصل اعتبار ۹۵% بیزی برای اثرات رگرسیونی وابسته به زمان در مدل پررتبه (۱.۴)



شکل ۸.۴: برآورد (میانه) و فواصل اعتبار ۹۵% بیزی برای اثرات رگرسیونی وابسته به زمان در مدل پررتبه (۱.۴)

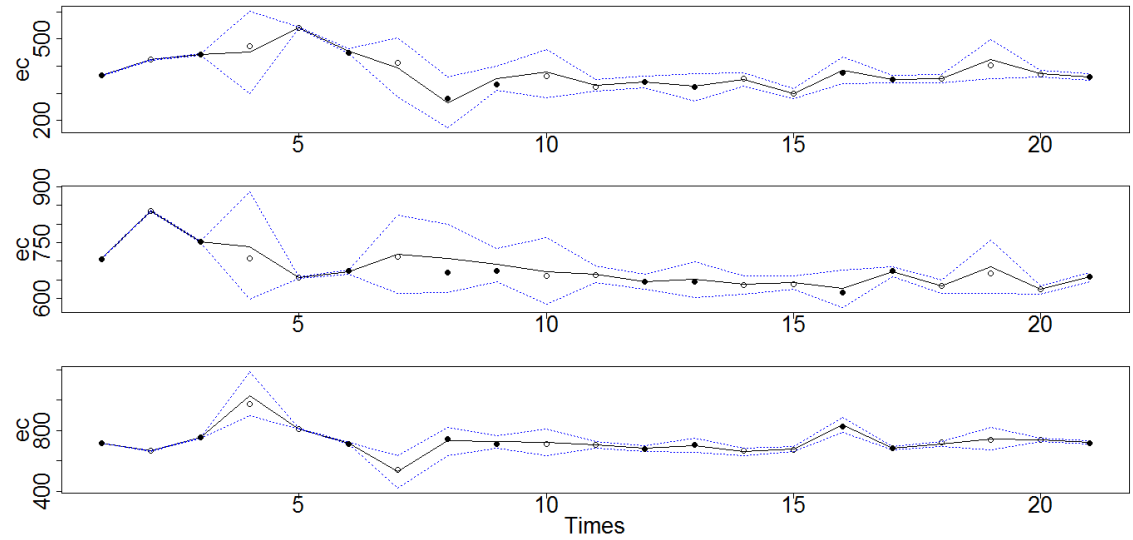
شکل ۹.۴، مشابه شکل‌های ۷.۴ و ۸.۴، نتایج استنباط بیزی را برای پارامترهای وابستگی مدل نمایش می‌دهند. پارامتر τ^2 جز چند زمان معدود، اغلب حول صفر یا صفر گزارش شده است. علاوه بر این، فواصل اعتبار گزارش‌شده کوتاه هستند که حاکی از کمبود اطلاعات برای برآورد مناسب این پارامتر است. پارامتر σ^2 وضعیت مشابهی با τ^2 دارد. برآورد پارامتر ϕ در طول زمان ماهیت تغییرپذیری بیشتری دارد و مدل پویا نشان می‌دهد که در نظر گرفتن این نوع ماهیت لازم است. در مجموع، با توجه به وجود ۱۵۰ ایستگاه، برای هر زمان، مدل تقریباً توانسته است پارامترهای فرآیند و تغییرپذیری آن‌ها را به‌طور قابل قبول برآورد کند.

شکل‌های ۱۰.۴ و ۱۱.۴ مقادیر مشاهده‌شده و برآوردشده ۶ ایستگاه حذف‌شده برای اعتبارسنجی



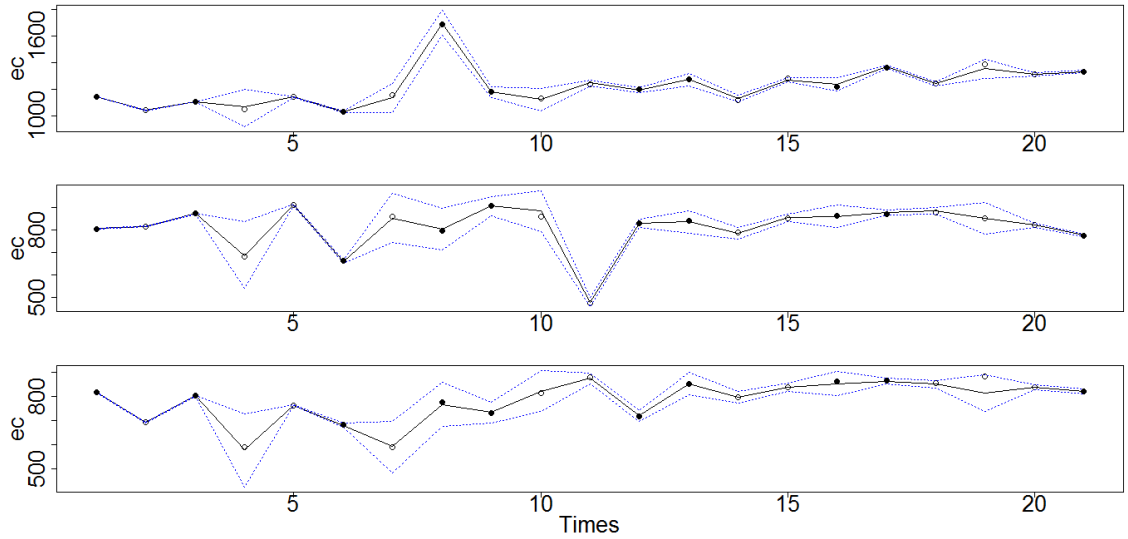
شکل ۹.۴: برآورد (میانه) و فواصل اعتبار ۹۵% بیزی برای پارامترهای ساختار وابستگی مدل پررتبه (۱.۴)

مدل را نمایش می‌دهند. در این شکل‌ها، دایره‌های تو خالی نشان‌دهنده مشاهدات مورد استفاده در برازش مدل و دایره‌های تو پر نشان‌دهنده مشاهدات حذف‌شده هستند. مقادیر پیش‌گویی بیزی (مبتنی بر میانه توزیع پیش‌گو) و فواصل اعتبار به ترتیب با خطوط غیرمنقطع و منقطع نمایش داده شده‌اند. همان‌طور که مشهود است برآورد فواصل اعتبار در این ایستگاه‌ها کوتاه است که نشان از عملکرد پیش‌گویی خوب مدل دارد. همچنین داده‌های حذف‌شده همگی در داخل فاصله اعتبار قرار گرفته‌اند که حاکی از سازگاری مدل برازش‌شده در تشریح روند فضایی-زمانی مدل است.



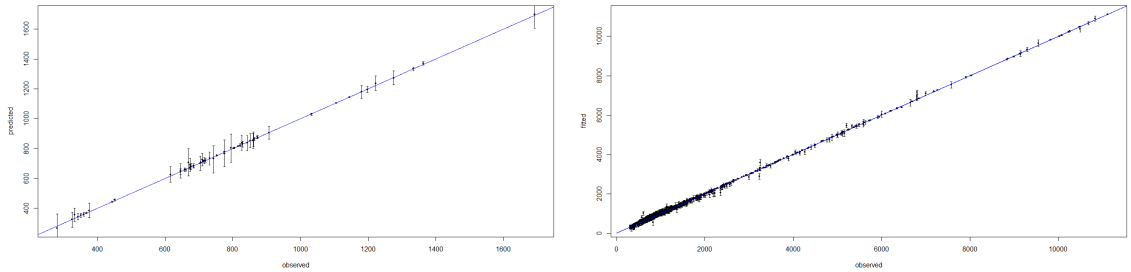
شکل ۱۰.۴: میانه توزیع پیش‌گو و فواصل اعتبار بیزی ۹۵% که به ترتیب با خطوط غیرمنقطع و منقطع برای ۳ ایستگاه حذف‌شده در مدل پررتبه (۱.۴) نشان داده شده‌اند.

علاوه بر این، شکل ۱۲.۴ (آ)، پراکنش مقادیر مشاهده‌شده در مقابل مقادیر برازش‌شده توسط مدل را نشان می‌دهد. پراکنش این جفت مقادیر حول نیمساز ربع اول و سوم، بیانگر برازش بسیار خوب



شکل ۱۱.۴: میانه توزیع پیش‌گو و فواصل اعتبار بیزی ۹۵% که به ترتیب با خطوط غیرمنقطع و منقطع برای ۳ ایستگاه حذف‌شده در مدل پررتبه (۱.۴) نشان داده شده‌اند.

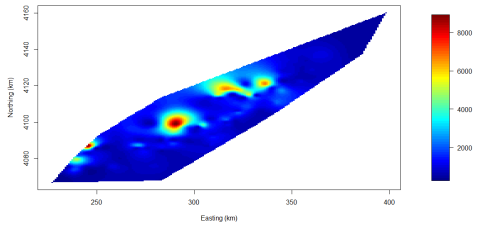
مشاهدات مدل‌ساز است. همچنین، شکل ۱۲.۴ (ب)، پراکنش مقادیر مشاهده‌شده در مقابل مقادیر پیش‌گویی‌شده توسط مدل را برای داده‌های آزمون (اعتبارسنجی مدل) نشان می‌دهد. برای هر نقطه پیش‌گویی‌شده نیز فاصله اعتبار پیش‌گویی ترسیم شده است. با نگاهی به این نمودار پراکنش، می‌توان به راحتی توانایی بالای مدل در پیش‌گویی را مشاهده کرد.



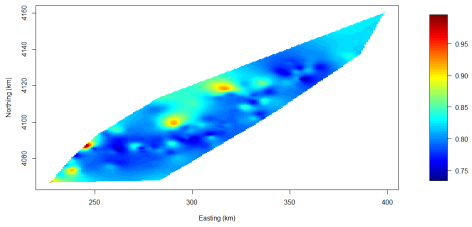
(ب) (آ)

شکل ۱۲.۴: مقادیر مشاهده‌شده در مقابل برازش‌شده: (آ)؛ مقادیر مشاهده‌شده در مقابل پیش‌گویی‌شده: (ب)

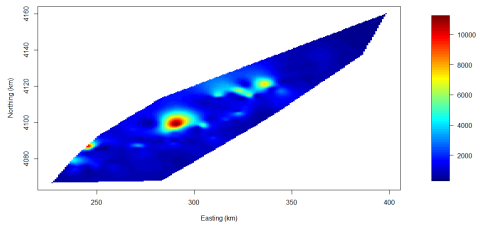
در نهایت، شکل ۱۳.۴ رویه‌های پهنه‌بندی بیزی (بر اساس میانه توزیع پسین) را برای متغیر پاسخ، EC، به همراه انحراف معیار آن‌ها، در ۵ نقطه زمانی مختلف، به عنوان نمونه نمایش می‌دهد. با توجه به شکل‌ها و مقیاس کنار آن‌ها، کیفیت بسیار پایین آب را در قسمت‌های خاصی از استان که با رنگ قرمز نشان داده شده‌اند، می‌توان مشاهده کرد. یک قسمت کوچکی نیز در گوشه نزدیک به دریا بیانگر کیفیت آب پایین مشابه است که امکان نفوذ آب دریا به آب‌های زیرزمینی در این ناحیه دلیل اصلی این کیفیت پایین است. به‌علاوه، هر چقدر به شرق و جنوب استان نزدیک می‌شویم، کیفیت آب بهتر می‌شود.



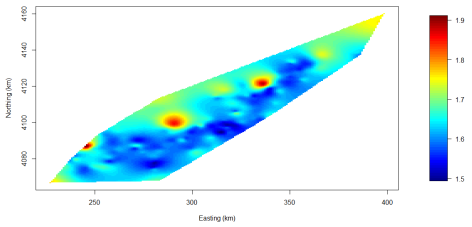
(آ) زمان اول



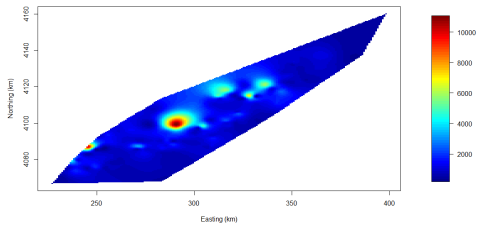
(ب) زمان اول



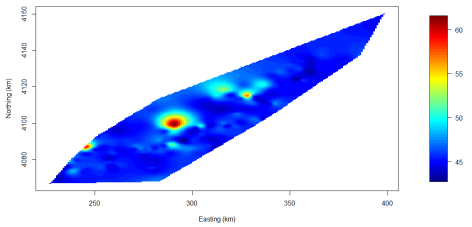
(ج) زمان پنجم



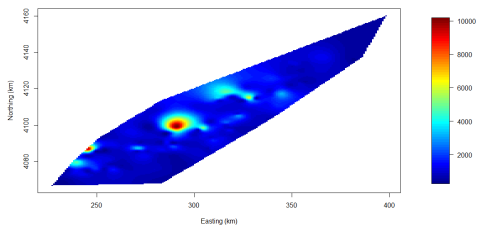
(د) زمان پنجم



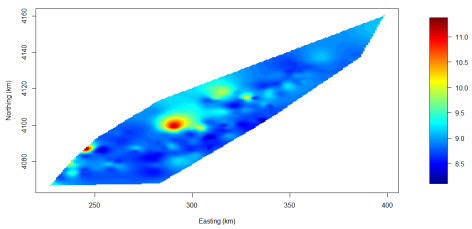
(ه) زمان دهم



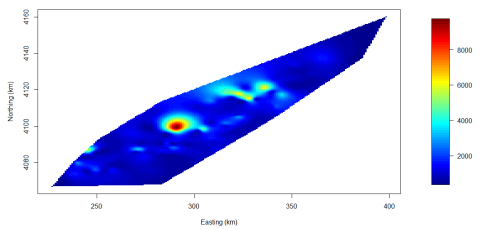
(و) زمان دهم



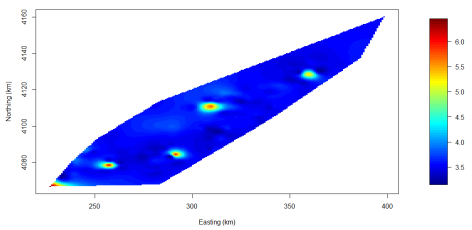
(ز) زمان پانزدهم



(ح) زمان پانزدهم



(ط) زمان بیست ویکم

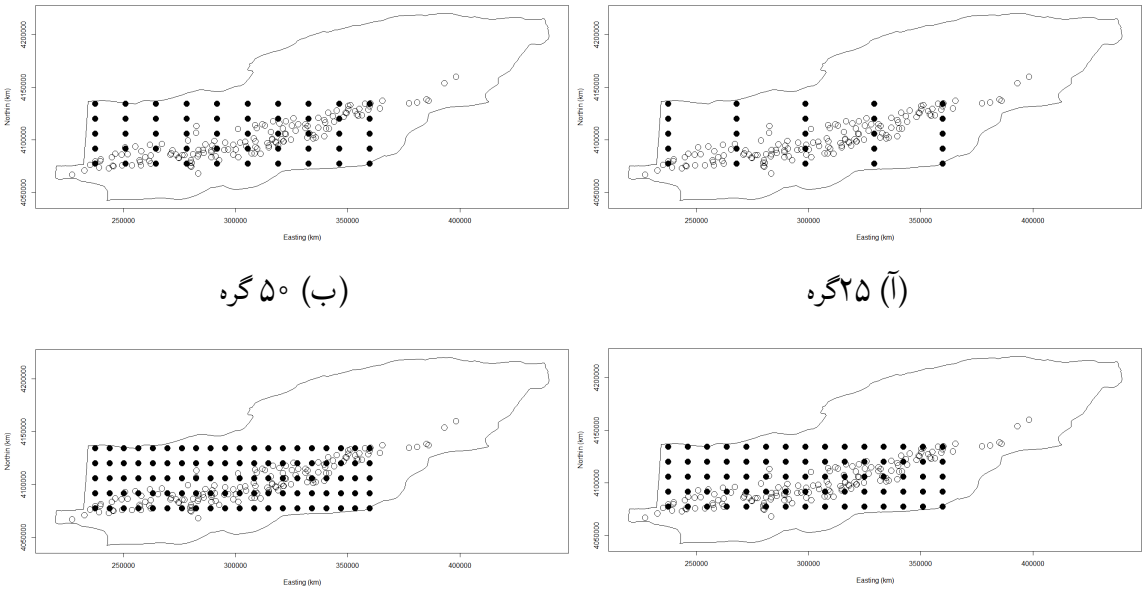


(ی) زمان بیست ویکم

شکل ۱۳.۴: رویه‌های پهنه‌بندی برآوردشده (ستون سمت راست) و انحراف معیار متناظر آن‌ها (ستون سمت چپ) در ۵ نقطه زمانی متفاوت

۱.۵.۴ فرآیند پیش‌گو

برای ارزیابی عملکرد مدل فرآیند پیش‌گوی (۲.۴)، تعداد گره‌های ۲۵، ۵۰، ۷۵ و ۱۰۰ انتخاب شدند. همچنین مکان گره‌ها به روش مشبکه منظم، مطابق شکل ۱۴.۴ طراحی شدند. برای مدل فرآیند پیش‌گو



(آ) ۲۵گره (ب) ۵۰ گره

(ج) ۷۵ گره (د) ۱۰۰ گره

شکل ۱۴.۴: مکان گره‌ها روی نقشه استان گلستان که با علامت دایره تو پر مشخص شده‌اند

نیز از مدل کوواریانس نمایی، پیشین‌ها و مقادیر آغازین مشابه مدل پررتبه استفاده کردیم. مشابه مدل پررتبه، نتایج استنباط‌های بیزی در شکل‌های ۱۶.۴ تا ۲۱.۴، نمایش داده شده‌اند.

برآوردهای ضرایب مدل فرآیند پیش‌گو با گره‌های مختلف، کم و بیش مشابه مدل پررتبه بوده و از تفسیر مشابهی برخوردار هستند. به همین دلیل از تشریح بیشتر آن‌ها پرهیز می‌کنیم.

برای مدل پیش‌گو با ۲۵ گره (شکل ۲۰.۴)، برآورد پارامترهای وابستگی چندان کارا نیست. به عنوان مثال، با مشاهده فواصل اعتبار پارامتر σ^2 در زمان‌ها مختلف، بزرگ بودن طول آن‌ها قابل تشخیص است که حاکی از ناکافی بودن تعداد گره‌های انتخاب‌شده در برآورد کارای این پارامترها است. اما این مشکل برای مدل‌های با گره‌های ۵۰ به بالا به شدت مدل با ۲۵ گره نیست و نتایج، تقریبا مشابه مدل پررتبه هستند. اما، به‌طور کلی، کمبود اطلاعات در مدل‌های فرآیند پیش‌گو با گره‌های مختلف (در برآورد پارامترهای وابستگی مدل) ویژگی است که از مدل پررتبه متناظر خود به ارث برده شده است.

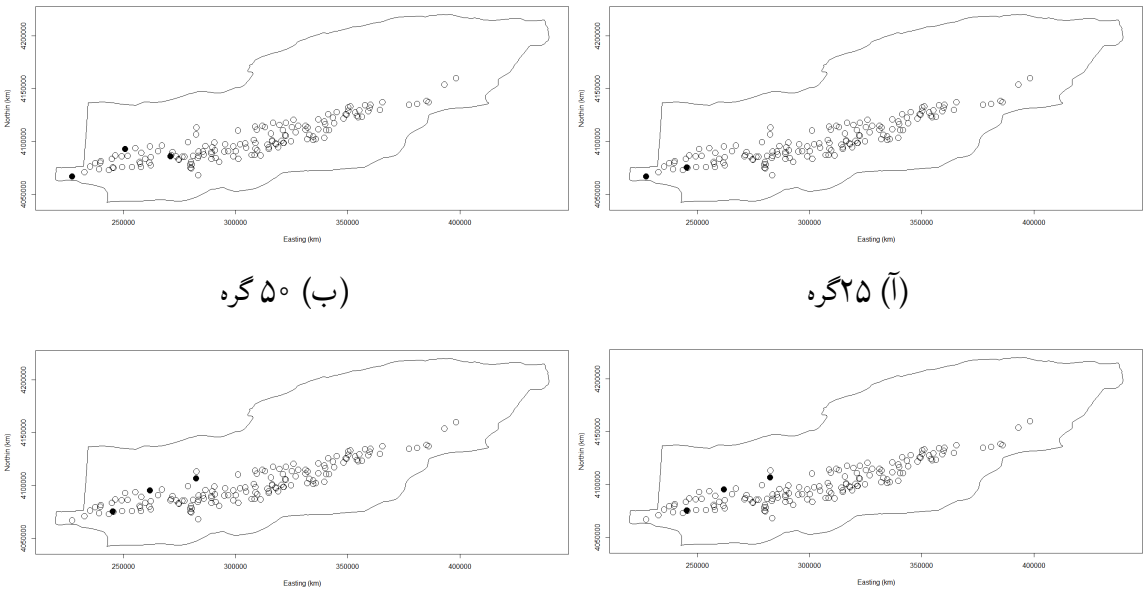
شکل ۲۴.۴ مقادیر مشاهده‌شده و پیش‌گویی برای ۲ ایستگاهی که برای اعتبارسنجی مدل (در مدل فرآیند پیش‌گو با ۲۵ گره) حذف شده بودند را نمایش می‌دهد. همچنین شکل‌های ۲۵.۴، ۲۶.۴ و ۲۷.۴، نمودارهای مشابه برای مدل‌های با به ترتیب ۵۰، ۷۵ و ۱۰۰ گره را برای ۳ ایستگاهی نشان می‌دهند که ۱۰ مشاهده آن‌ها برای اعتبارسنجی مدل کنار گذاشته شده‌اند. همان‌طور که در شکل‌ها می‌بینیم، برآورد فواصل اعتبار در این ایستگاه‌ها کوتاه است که نشان از مدل‌بندی خوب عدم قطعیت موجود در پیش‌گویی

داده‌ها است. همچنین داده‌های کنارگذاشته شده همگی در داخل فواصل اعتبار برآوردشده برای مدل قرار گرفته‌اند که حاکی از توان پیش‌گویی بالای مدل برازش‌شده است.

مشابه مدل پررتبه، نمودارهای پراکنش مقادیر مشاهده‌شده در مقابل برازش‌شده و مقادیر مشاهده‌شده (حذف‌شده) در مقابل پیش‌گویی‌شده، برای مدل‌های فرآیند پیش‌گو با گره‌های مختلف، در شکل‌های ۲۸.۴ تا ۳۱.۴ ترسیم شده‌اند. نمودارهای قسمت (الف) در این چهار شکل، همگی قدرت خوب پیش‌گویی مدل را، حتی برای مدل فرآیند پیش‌گو با ۲۵ گره، در برآورد مقادیر مشاهده‌شده‌ای که در برازش مدل نقش دارند، نشان می‌دهند. اما نتیجه مشابه برای پیش‌گویی مقادیر پاسخ جدید برقرار نیست. در شکل ۲۸.۴ (ب) فواصل اعتبار عریض نشان از عدم قطعیت شدید در پیش‌گویی مقادیر جدید دارد، هرچند که مقادیر پیش‌گویی به واقعیت خیلی نزدیک هستند. در مقابل، بر خلاف این نتیجه، برای مدل‌های با گره‌های ۵۰ تا به بالا، فواصل اعتبار کوتاهتر در پیش‌گویی پاسخ‌های ایستگاه‌های حذف‌شده تاییدکننده پیش‌گویی دقیق‌تر هستند.

نقشه‌های پهنه‌بندی پاسخ‌های برآوردشده، EC، برای مدل فرآیند پیش‌گو با ۵۰ گره، به عنوان نمونه، در شکل ۳۲.۴ به همراه انحراف معیار متناظر، در ۵ نقطه زمانی مختلف، نمایش داده شده‌اند. برای مدل‌های با سایر تعداد گره، این نتایج مشابه هستند. به همین دلیل از گزارش آن‌ها صرفنظر کردیم. برآورد رویه پاسخ در این شکل، میانه توزیع پسین است. تفسیر این نقشه‌های پهنه‌بندی مشابه مدل پررتبه است.

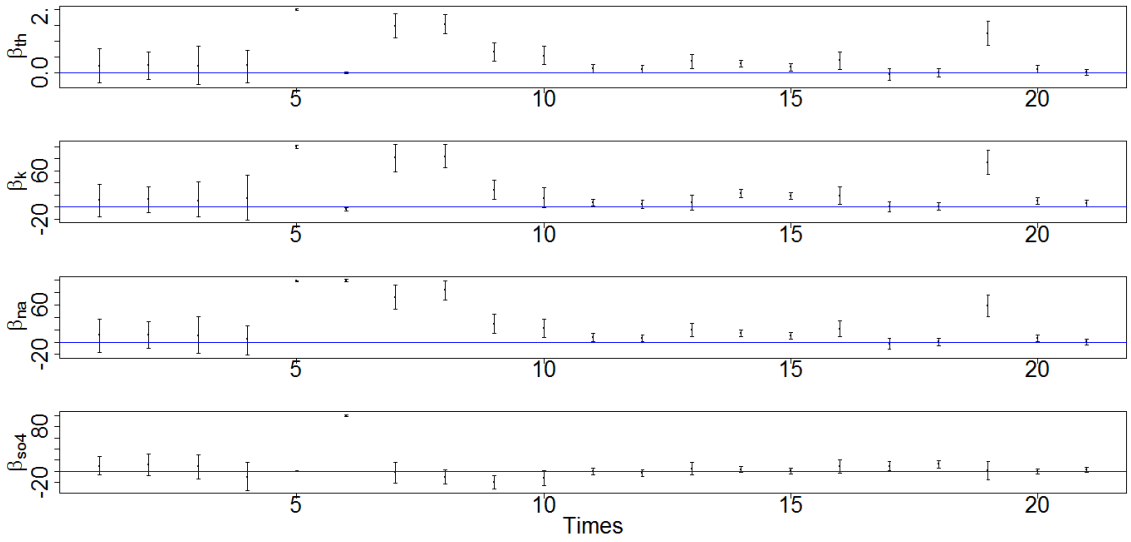
هزینه‌های محاسباتی لازم برای برازش مدل‌های پررتبه و کم‌رتبه با گره‌های مختلف، در جدول ۱.۴ گزارش شده‌اند. واضح است که زمان برازش مدل‌های فرآیند پیش‌گو، در مقایسه با مدل پررتبه، با کاهش



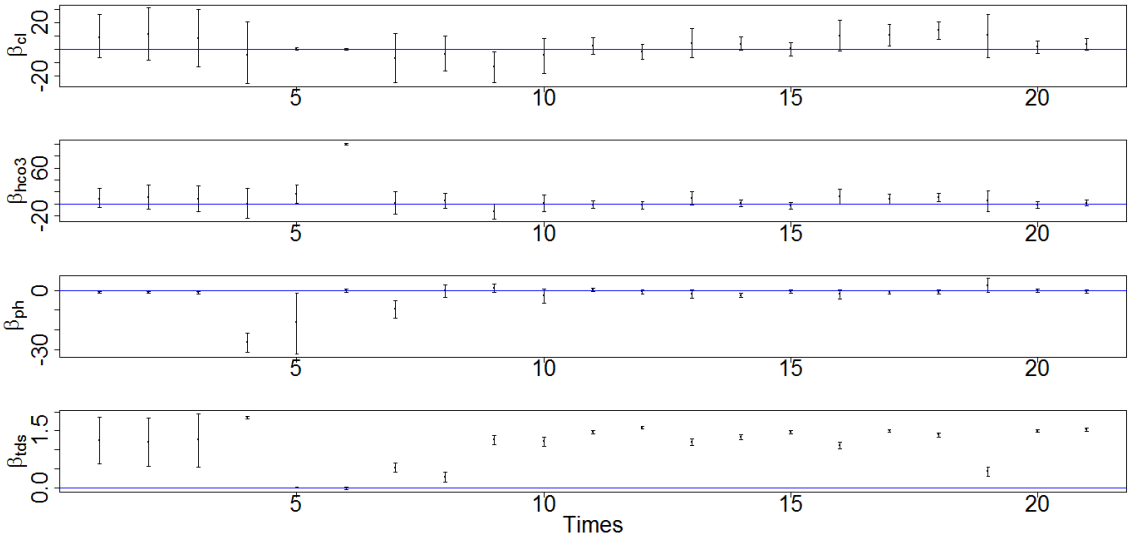
(آ) ۲۵گره (ب) ۵۰ گره

(ج) ۷۵ گره (د) ۱۰۰ گره

شکل ۱۵.۴: مکان ایستگاه‌های کنارگذاشته شده روی نقشه استان گلستان که با علامت دایره تو پر مشخص شده‌اند

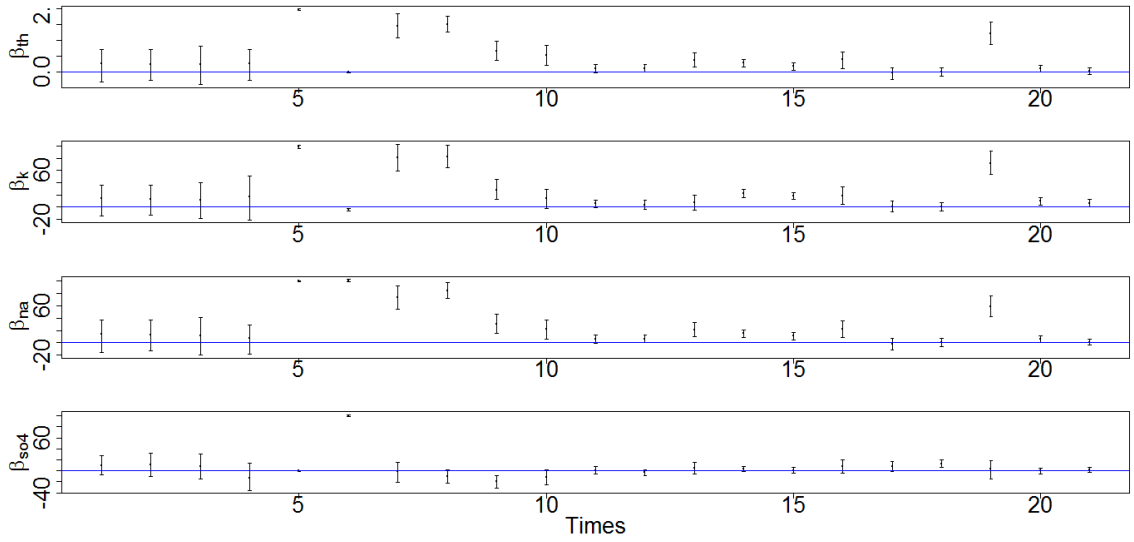


(آ)

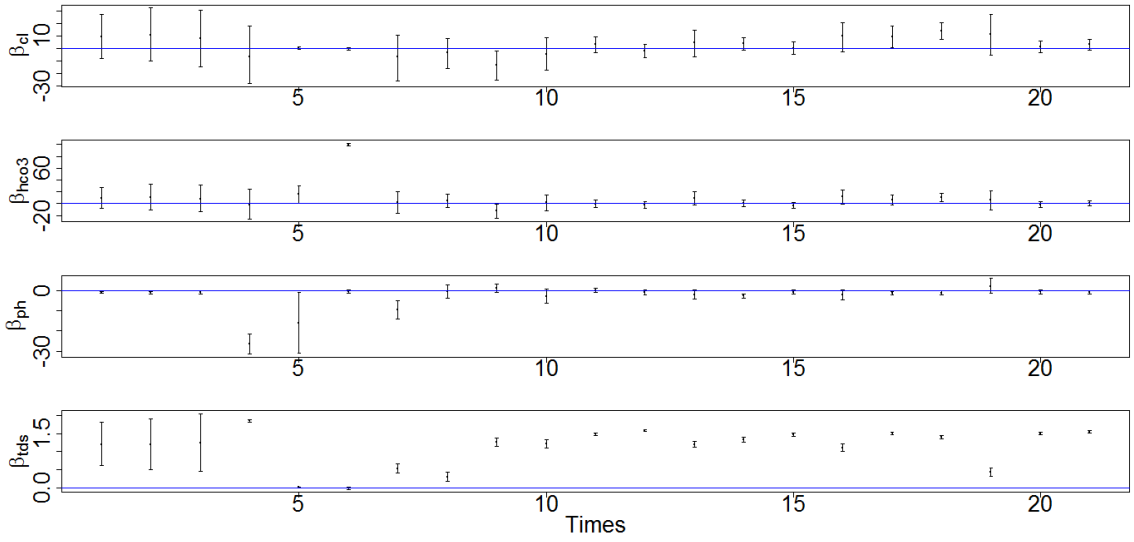


(ب)

شکل ۱۶.۴: برآورد (میانه توزیع پسین) و فواصل اعتبار ۹۵% بیزی برای اثرات متغیرهای تبیینی مدل فرآیند پیش‌گوی (۲.۴) با انتخاب ۲۵ گره

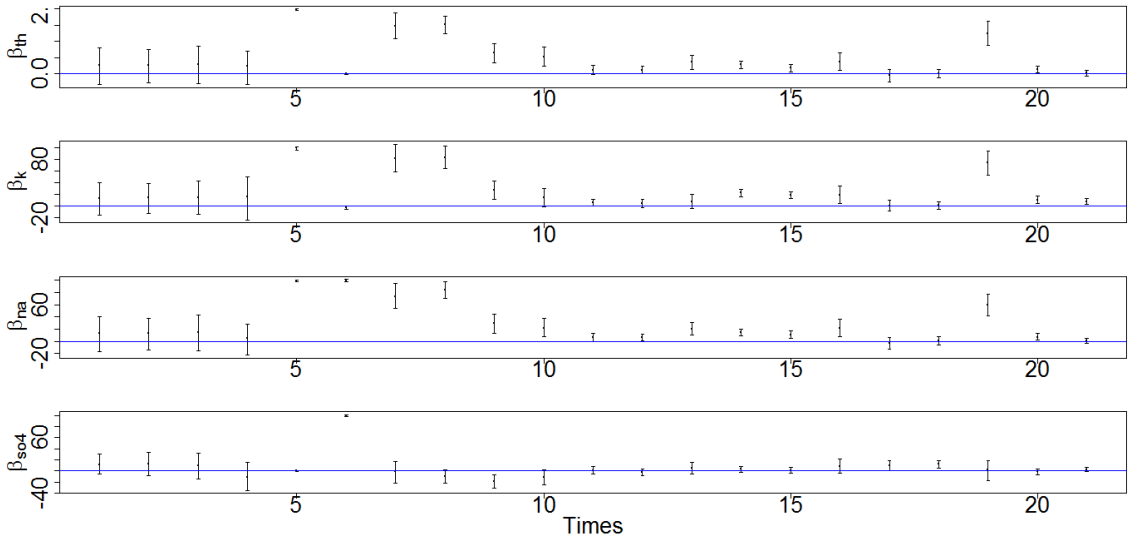


(آ)

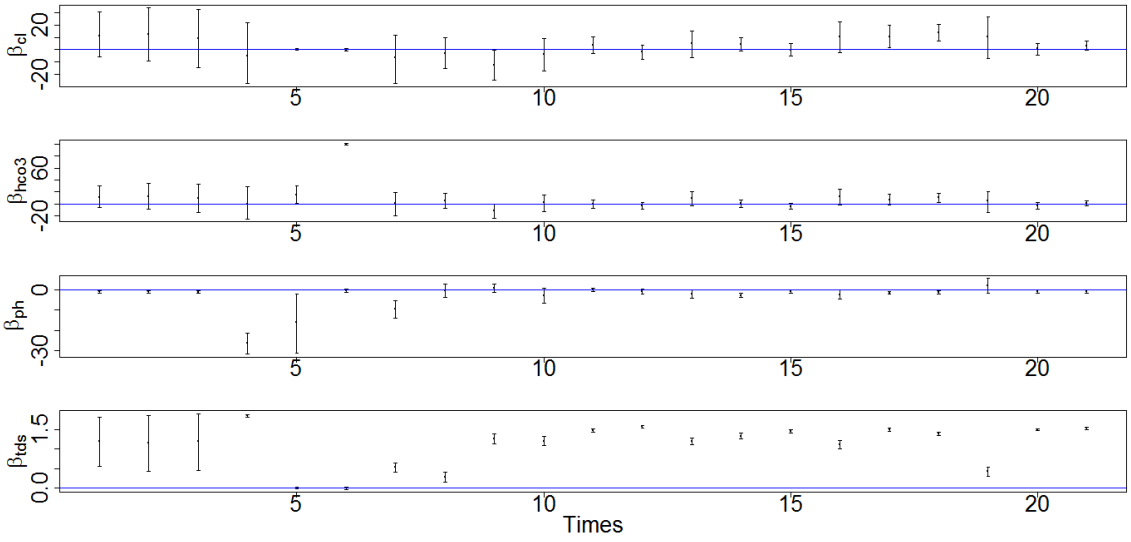


(ب)

شکل ۱۷.۴: برآورد (میانه توزیع پسین) و فواصل اعتبار ۹۵% بیزی برای اثرات متغیرهای تبیینی مدل فرآیند پیش‌گوی (۲.۴) با انتخاب ۵۰ گره

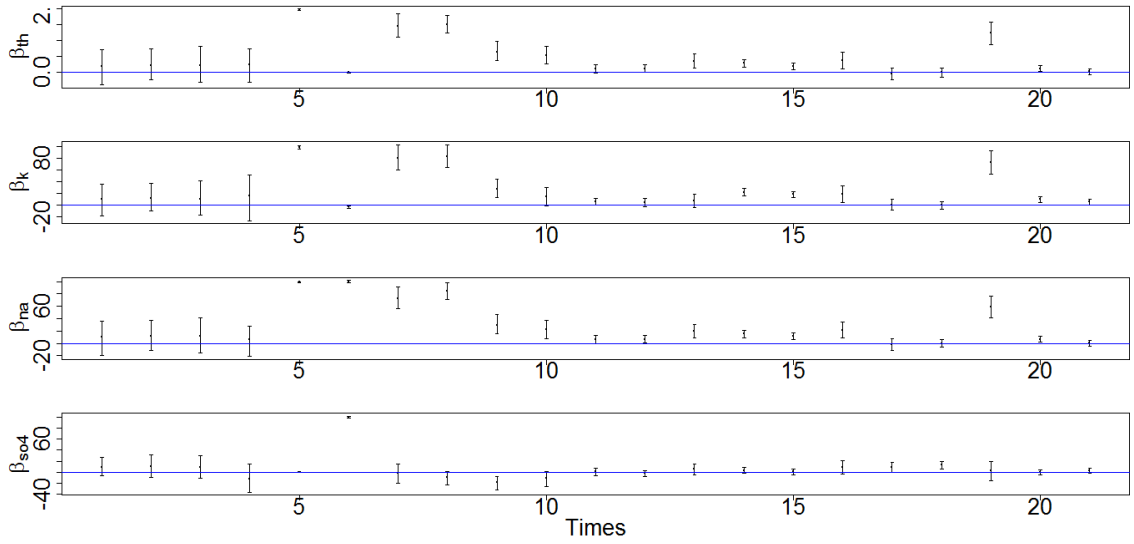


(آ)

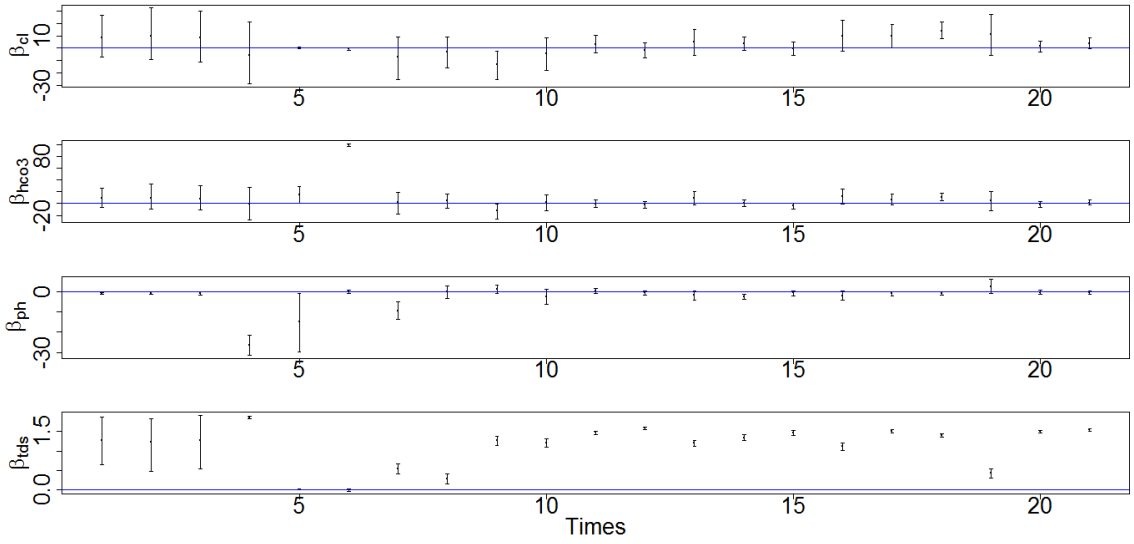


(ب)

شکل ۱۸.۴: برآورد (میانه توزیع پسین) و فواصل اعتبار ۹۵% بیزی برای اثرات متغیرهای تبیینی مدل فرآیند پیش‌گوی (۲.۴) با انتخاب ۷۵ گره

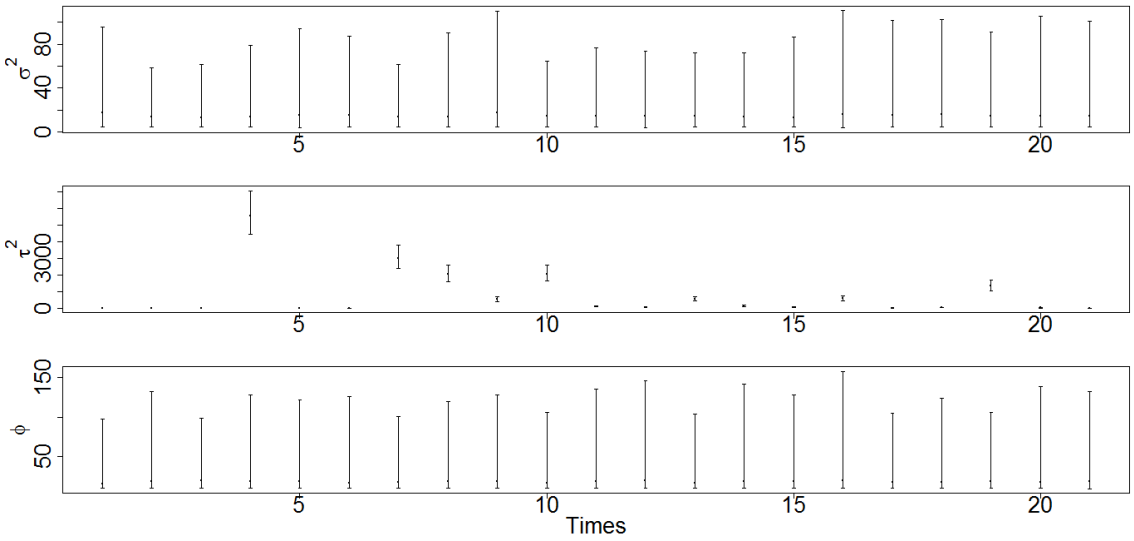


(آ)

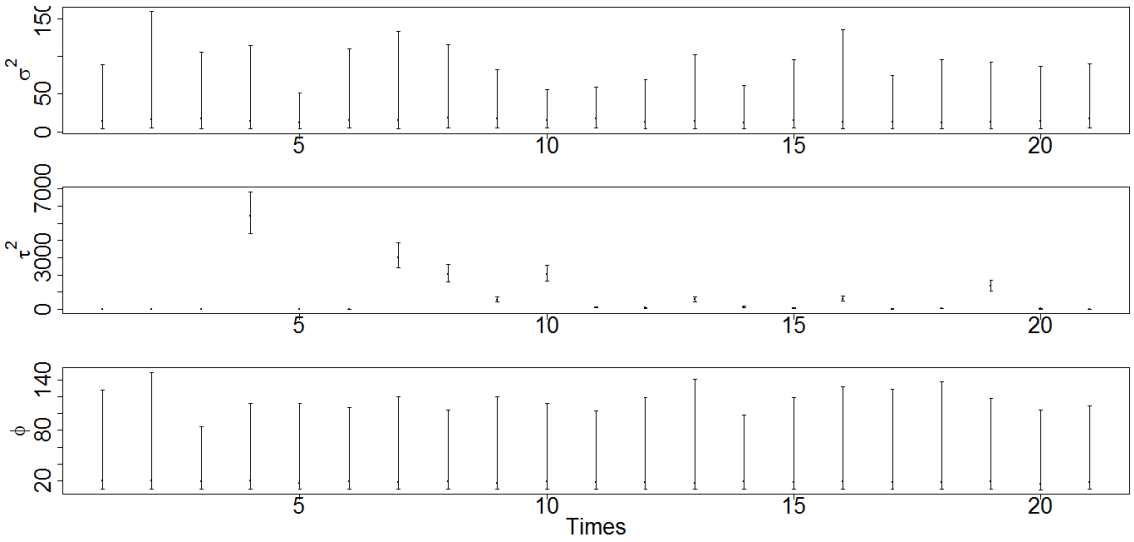


(ب)

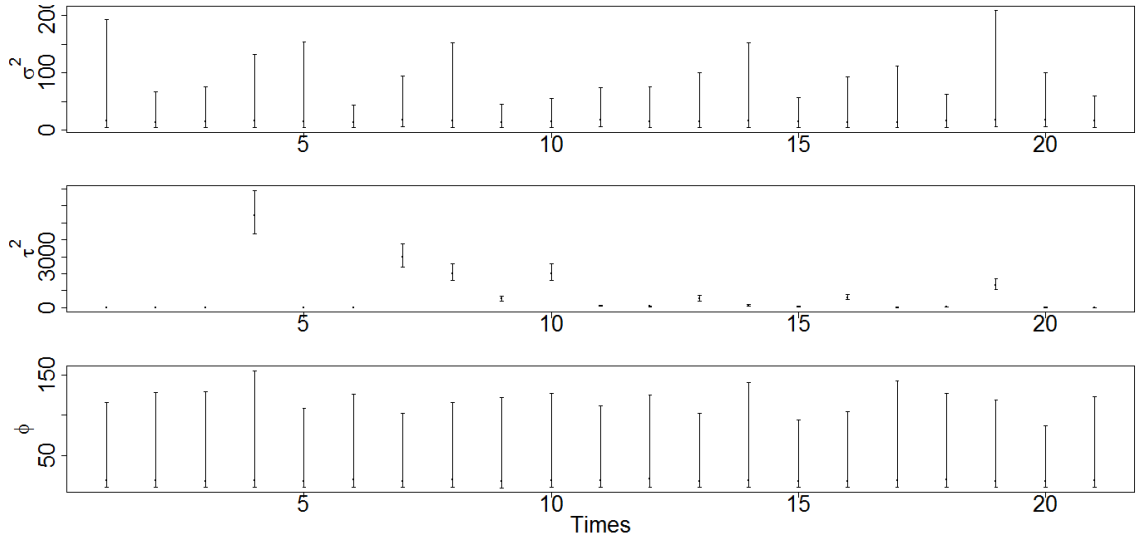
شکل ۱۹.۴: میانه و فواصل اعتبار بیزی ۹۵% برای اثرات متغیرهای تبیینی مدل فرآیند پیش‌گو (۲.۴) با انتخاب ۱۰۰ گره



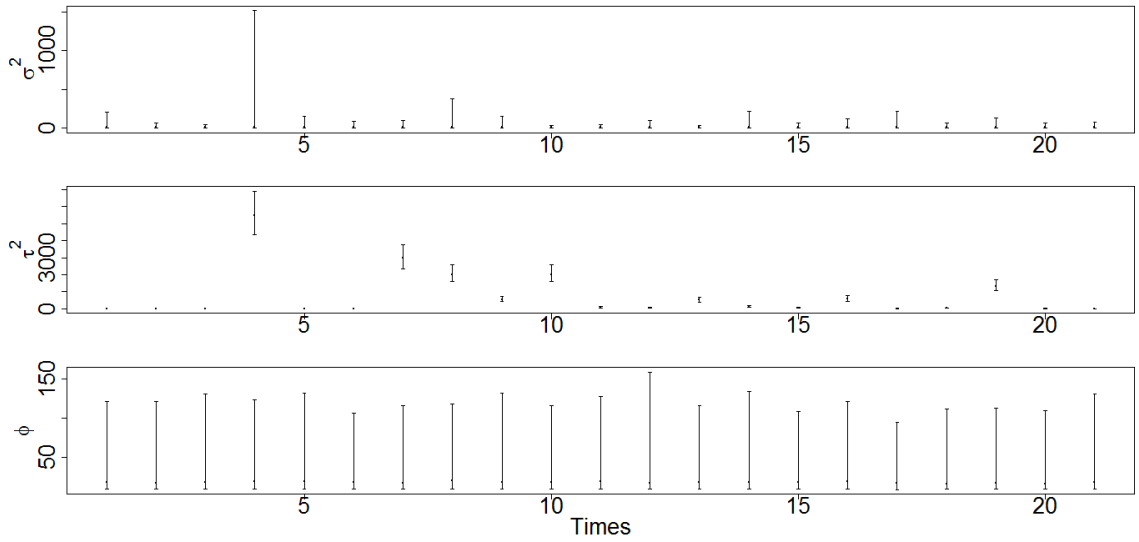
شکل ۲۰.۴: برآورد (میانه توزیع پسین) و فواصل اعتبار ۹۵% بیزی برای پارامترهای وابستگی فضایی در مدل فرآیند پیش‌گوی (۲.۴) با ۲۵ گره



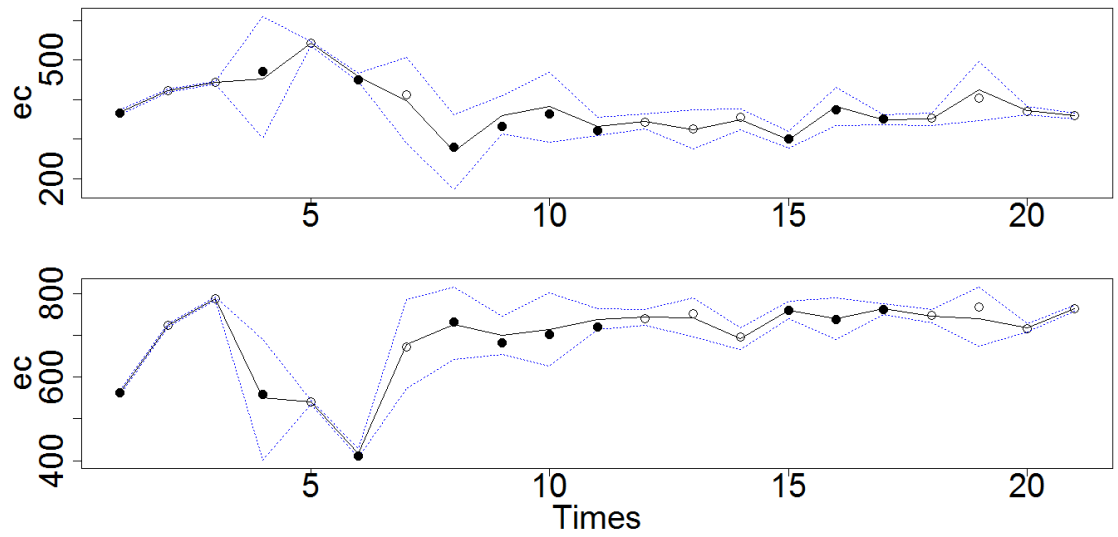
شکل ۲۱.۴: برآورد (میانه توزیع پسین) و فواصل اعتبار ۹۵% بیزی برای پارامترهای وابستگی فضایی در مدل فرآیند پیش‌گوی (۲.۴) با ۵۰ گره



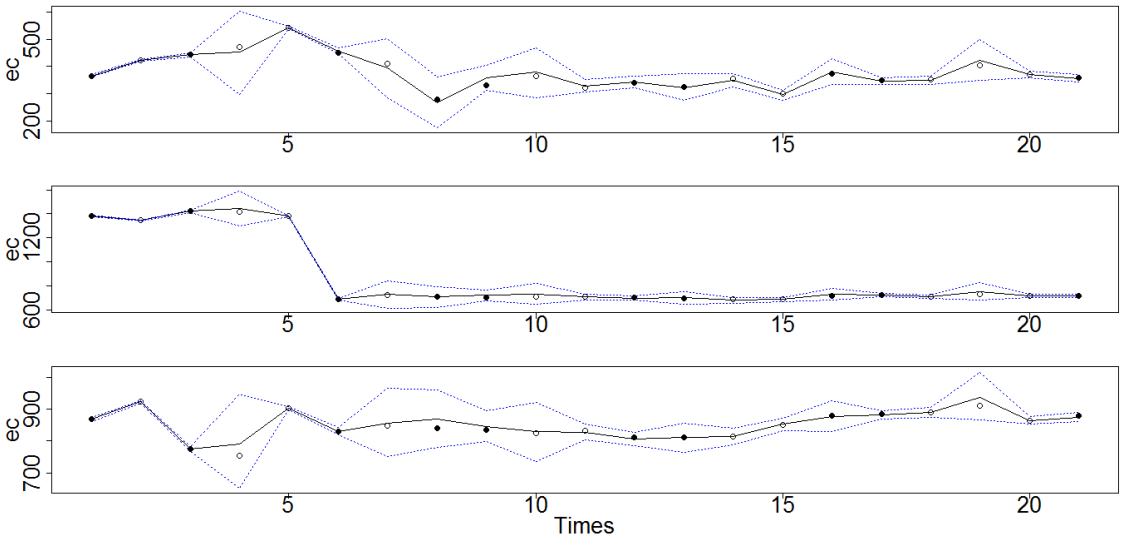
شکل ۲۲.۴: برآورد (میانه توزیع پسین) و فواصل اعتبار ۹۵% بیزی برای پارامترهای وابستگی فضایی در مدل فرآیند پیش‌گوی (۲.۴) با ۷۵ گره



شکل ۲۳.۴: برآورد (میانه توزیع پسین) و فواصل اعتبار ۹۵% بیزی برای پارامترهای وابستگی فضایی در مدل فرآیند پیش‌گوی (۲.۴) با ۱۰۰ گره



شکل ۲۴.۴: میانه توزیع پیش‌گو و فواصل اعتبار ۹۵% که به ترتیب با خطوط غیرمنقطع و منقطع برای دو ایستگاه حذف‌شده در مدل فرآیند پیش‌گوی (۲.۴) با ۲۵ گره نشان داده شده‌اند. دایره‌های تو خالی نشان‌گر مشاهدات مورد استفاده برای برازش مدل و دایره‌های تو پر نشان‌گر مشاهدات مورد استفاده برای اعتبارسنجی مدل هستند.



شکل ۲۵.۴: میانه توزیع پیش‌گو و فواصل اعتبار ۹۵% که به ترتیب با خطوط غیرمنقطع و منقطع برای سه ایستگاه حذف‌شده در مدل فرآیند پیش‌گوی (۲.۴) با ۵۰ گره نشان داده شده‌اند. دایره‌های تو خالی نشان‌گر مشاهدات مورد استفاده برای برازش مدل و دایره‌های تو پر نشان‌گر مشاهدات مورد استفاده برای اعتبارسنجی مدل هستند.

زیادی همراه است. به عنوان نمونه، برازش مدل فرآیند پیش‌گو با ۵۰ گره نسبت به مدل پررتبه، نزدیک به ۶ برابر سریعتر است و این زمان قابل توجهی را برای کاربر ذخیره می‌کند. این کاهش هزینه محاسباتی در حالی است که نتایج استنباط بیزی به‌دست آمده هم تفاوت چشمگیری با مدل پررتبه ندارند.

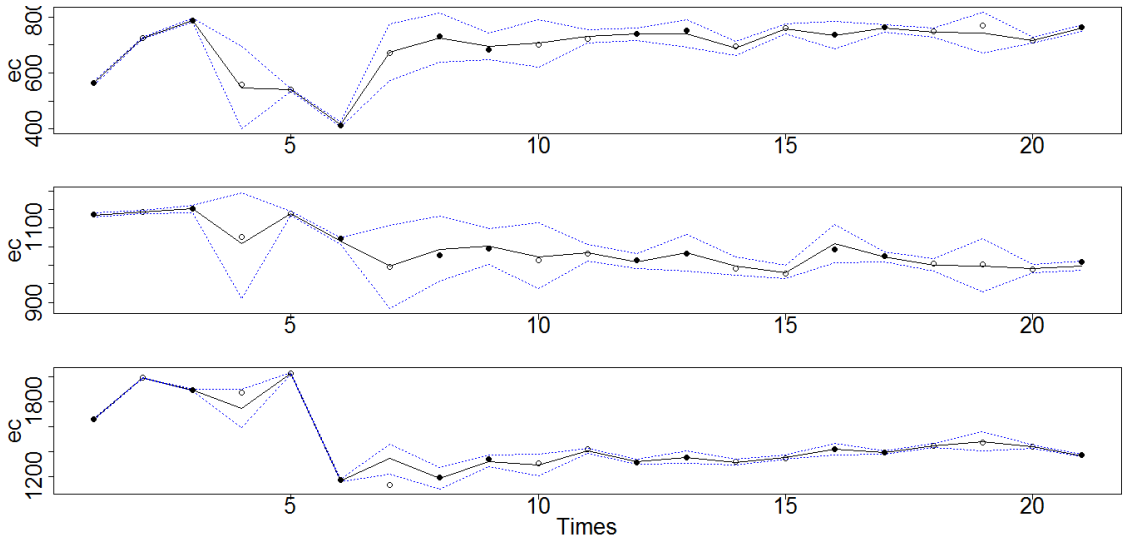
جدول ۱.۴: زمان اجرای مدل‌های پررتبه و فرآیند پیش‌گو با تعداد گره‌های مختلف

مدل	پررتبه	۲۵ گره	۵۰ گره	۷۵ گره	۱۰۰ گره
زمان اجرا	۱۵۷۷٫۹۰	۱۱۲٫۵۳	۳۳۹٫۰۲	۷۳۶٫۳۷	۱۳۲۰٫۱۴

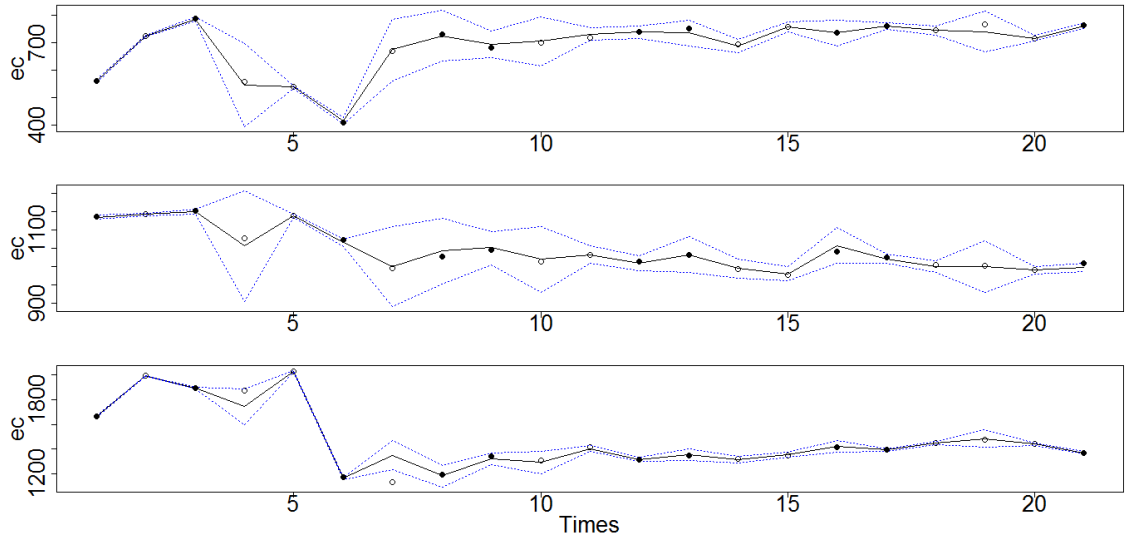
۶.۴ نتیجه‌گیری و آینده تحقیق

در این پایان‌نامه، به معرفی و مدل‌بندی داده‌های فضایی-زمانی حجیم پرداختیم. این نوع داده‌ها، با گسترش ابزار گردآوری داده‌ها، به شدت رو به افزایش هستند و توجه به مدل‌بندی و تحلیل آن‌ها حایز اهمیت است. مدل‌بندی این نوع داده‌ها، با توجه به ساختارهای وابستگی فضایی و زمانی پیچیده آن‌ها، در کنار حجم معمولا بالا، یک مساله مهم و داغ تحقیقاتی محسوب می‌شود. برخی از محققان وارد این حوزه شده‌اند و مدل‌هایی نیز معرفی کرده‌اند. اما حجم بالای داده‌ها، باعث ناکارآمدی محاسباتی در برازش آن‌ها شده است. به همین دلیل آماردان‌های مختلفی از روش‌ها و مدل‌های تقریبی برای برازش داده‌ها استفاده می‌کنند که برخی از آن‌ها را در فصل سوم معرفی کردیم.

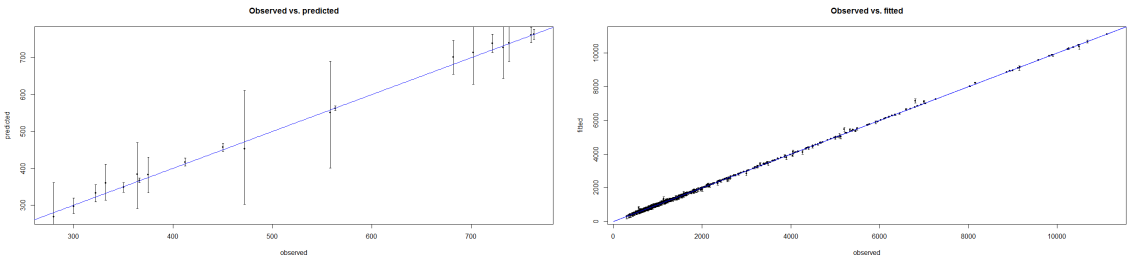
یک رده از مدل‌ها که استفاده از آن‌ها رو به گسترش است، رده مدل‌های فرآیند پیش‌گو است که توسط



شکل ۲۶.۴: میانه توزیع پیش‌گو و فواصل اعتبار ۹۵% که به ترتیب با خطوط غیرمنقطع و منقطع برای سه ایستگاه حذف‌شده در مدل فرآیند پیش‌گوی (۲.۴) با ۷۵ گره نشان داده شده‌اند. دایره‌های تو خالی نشان‌گر مشاهدات مورد استفاده برای برازش مدل و دایره‌های تو پر نشان‌گر مشاهدات مورد استفاده برای اعتبارسنجی مدل هستند.

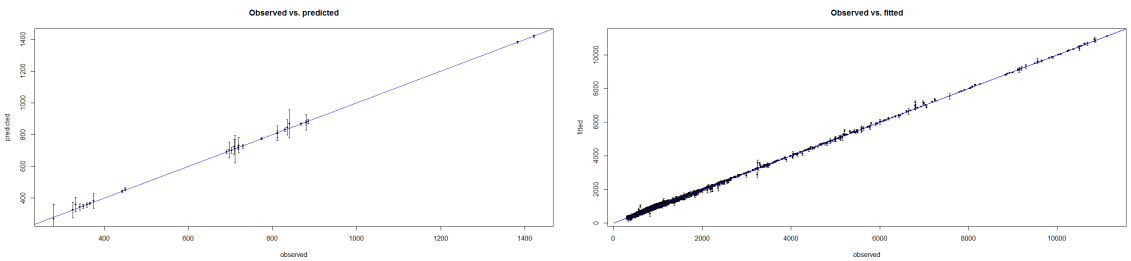


شکل ۲۷.۴: میانه توزیع پیش‌گو و فواصل اعتبار ۹۵% که به ترتیب با خطوط غیرمنقطع و منقطع برای سه ایستگاه حذف‌شده در مدل فرآیند پیش‌گوی (۲.۴) با ۱۰۰ گره نشان داده شده‌اند. دایره‌های تو خالی نشان‌گر مشاهدات مورد استفاده برای برازش مدل و دایره‌های تو پر نشان‌گر مشاهدات مورد استفاده برای اعتبارسنجی مدل هستند.



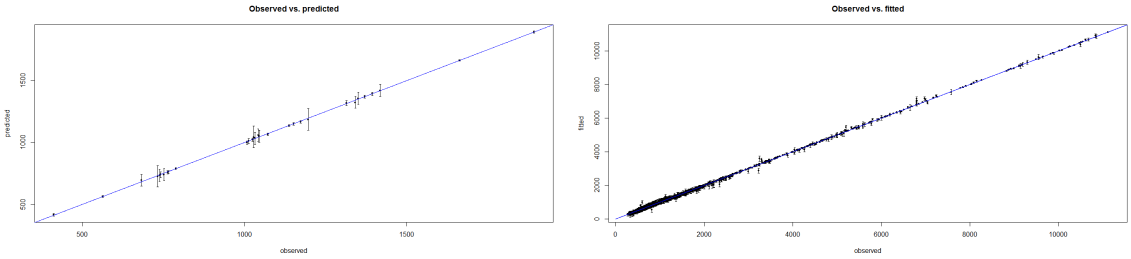
(ب) (آ)

شکل ۲۸.۴: مدل فرآیند پیش‌گوی (۲.۴) با ۲۵ گره. مقادیر مشاهده‌شده در مقابل برازش‌شده: (آ)؛ مقادیر مشاهده‌شده در مقابل پیش‌گویی‌شده: (ب)



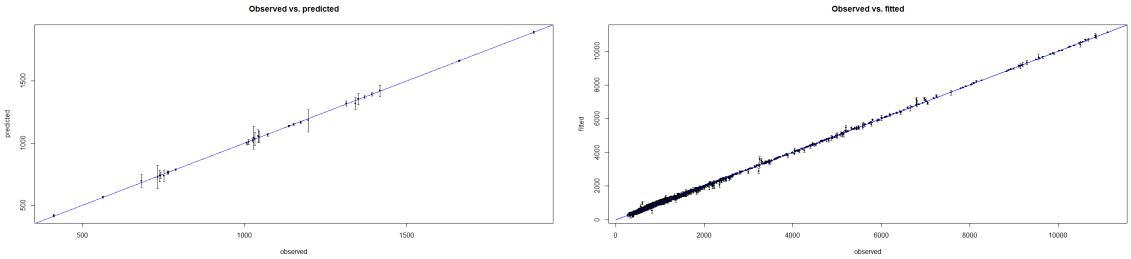
(ب) (آ)

شکل ۲۹.۴: مدل فرآیند پیش‌گوی (۲.۴) با ۵۰ گره. مقادیر مشاهده‌شده در مقابل برازش‌شده: (آ)؛ مقادیر مشاهده‌شده در مقابل پیش‌گویی‌شده: (ب)



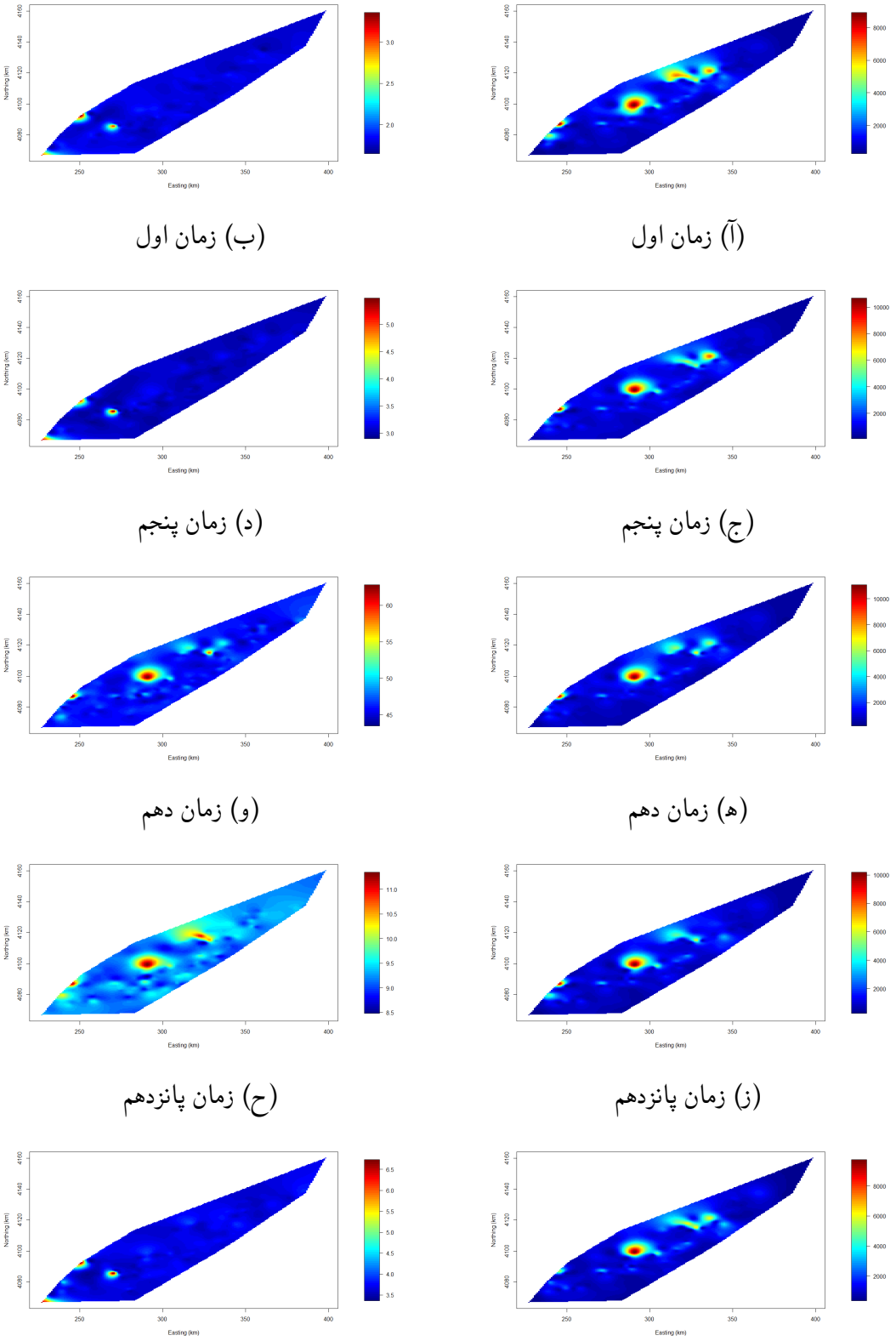
(آ) (ب)

شکل ۳۰.۴: مدل فرآیند پیش‌گوی (۲.۴) با ۷۵ گره. مقادیر مشاهده‌شده در مقابل برازش‌شده: (آ)؛ مقادیر مشاهده‌شده در مقابل پیش‌گویی‌شده: (ب)



(آ) (ب)

شکل ۳۱.۴: مدل فرآیند پیش‌گوی (۲.۴) با ۱۰۰ گره. مقادیر مشاهده‌شده در مقابل برازش‌شده: (آ)؛ مقادیر مشاهده‌شده در مقابل پیش‌گویی‌شده: (ب)



(آ) زمان اول (ب) زمان اول

(ج) زمان پنجم (د) زمان پنجم

(ه) زمان دهم (و) زمان دهم

(ز) زمان پانزدهم (ح) زمان پانزدهم

(ط) زمان بیست ویکم (ی) زمان بیست ویکم

شکل ۳۲.۴: نقشه‌های پهنه‌بندی برآوردشده از نمونه‌های زنجیر مارکوف MCMC در قالب میانه (ستون سمت راست) و انحراف معیار (ستون سمت چپ) در ۵ زمان مختلف

بنرجی و همکاران (۲۰۰۸) معرفی شد. یکی از دلایل استقبال از این مدل‌ها، وجود بسته نرم‌افزاری spBayes در R برای برازش آن‌هاست. ما در این پایان‌نامه، از یک مدل فضایی-زمانی پویا (گاوسی) برای مدل‌بندی داده‌های فضایی-زمانی گاوسی استفاده کردیم. با در نظر گرفتن حجم بالای این داده‌ها نیز از مدل‌های فرآیند پیش‌گوی خطی برای تقریب مدل استفاده کردیم که از نظر محاسباتی کارا باشد. عملکرد مدل و فرآیند پیش‌گو را در یک مثال شبیه‌سازی و یک مثال واقعی مربوط به داده‌های کیفیت آب ایستگاه‌های آب استان گلستان، مورد ارزیابی قرار دادیم.

نتایج نشان دادند که در کنار به دست آوردن کارایی محاسباتی قابل توجه، کارایی آماری از دست رفته در نتایج استنباط (با انتخاب مناسب تعداد گره‌ها) قابل صرفنظر کردن است. این نتیجه در مسایل کاربردی واقعی، که با داده‌های حجیم سر و کار دارند، می‌تواند بسیار مهم باشد. بنابراین، در مجموعه داده‌هایی که ویژگی‌های آن‌ها مشابه مواردی است که در این پایان‌نامه به آن‌ها پرداخته شد، استفاده از چارچوب مدل‌های فرآیند پیش‌گو را به عنوان یک انتخاب توصیه می‌کنیم.

با توجه به مطالبی که در این پایان‌نامه به آن‌ها، به تشریح، پرداخته شد و با توجه به هدف آن، برخی از مواردی که در آینده می‌توانند مورد علاقه و به عنوان موضوع‌های تحقیقاتی مطرح باشند، در زیر فهرست شده‌اند:

- استفاده عملی از رده مدل‌های فرآیند پیش‌گو در مسایلی که داده‌های فضایی-زمانی حجیم غیرگاوسی دارند؛ در این موارد، چارچوب مدل‌بندی یک زیررده از GLMMs خواهد بود. به دلیل پیچیده‌تر بودن برازش این مدل‌ها نسبت به مدل‌های گاوسی، که ناشی از مشکلات جدی همگرایی زنجیرهای MCMC در این مدل‌هاست، انتظار می‌رود برازش مدل‌های فرآیند پیش‌گو در این رده، مشکلات اجرایی و استنباطی بیشتری نسبت به رده مدل‌های گاوسی داشته باشد.

- بررسی امکان انجام استنباط‌های مبتنی بر درستنمایی برای داده‌های فضایی-زمانی حجیم: انجام این نوع استنباط به دلیل ساختارهای وابستگی پیچیده داده‌ها خیلی دشوار خواهد بود. زیرا انتگرال‌های رام‌نشدنی با ابعاد خیلی بالا و سپس مشتق‌گیری‌های سخت از تابع درستنمایی حاصل از انتگرال‌گیری، دو وظیفه خیلی مشکل و گاهی نشدنی هستند که استنباط‌های مبتنی بر درستنمایی را برای این داده‌ها محدود کرده‌اند. اما شاید بتوان از روش‌هایی مانند همسانه‌سازی داده‌ها[۶] (له‌له و همکاران، ۲۰۰۷؛ باغیشنی و محمدزاده، ۲۰۱۱) برای این منظور استفاده کرد. این روش از الگوریتم‌های نمونه‌گیری MCMC برای به دست آوردن برآوردگرهای ماکسیمم درستنمایی در مدل‌های پیچیده بهره می‌برد که نتیجه آن پرهیز از محاسبه انتگرال‌های رام‌نشدنی و مشتق‌گیری از توابع پیچیده است.

- انتخاب مدل و معرفی معیارهای انتخاب مدل، همیشه یک بخش مهم در فرآیند مدل‌بندی محسوب می‌شود. معرفی چنین معیارهایی، به‌ویژه در مدل‌های آمیخته، همیشه یک چالش بوده است. بررسی امکان معرفی یک معیار انتخاب مدل مبتنی بر مدل‌های فرآیند پیش‌گو می‌تواند به عنوان یک موضوع جالب نیز مطرح باشد.

[۶] Data cloning

پیوست آ

آ.۱ فضای هیلبرت و مقدمات مرتبط با آن

تعریف آ.۱.۱ (فضای برداری)**.** یک فضای برداری[1] عبارت است از:

۱. یک میدان F از اسکالرها

۲. یک مجموعه V از اشیایی به نام بردارها

۳. یک قاعده به نام جمع برداری که به هر جفت از بردارهای α و β از V بردار $\alpha+\beta$ را که مجموع α و β نامیده می‌شود وابسته می‌سازد با این شرایط که

- جمع جابجایی است؛ یعنی $\alpha+\beta=\beta+\alpha$
- جمع شرکت‌پذیر است؛ یعنی $(\alpha+\beta)+\gamma=\alpha+(\beta+\gamma)$
- بردار یکتای ۰ به نام بردار صفر در V موجود است به‌طوری که به‌ازای هر α در V، $\alpha+۰=\alpha$
- به‌ازای هر بردار α در V، بردار یکتای $-\alpha$ در V موجود است به‌طوری که $\alpha+-\alpha=۰$

۴. یک قاعده به نام ضرب اسکالری که به هر اسکالر c از F و هر بردار α از V بردار $c\alpha$ در V را که حاصل ضرب α و c نامیده می‌شود وابسته سازد با این شرایط که

- به‌ازای هر α در V، $۱\alpha=\alpha$
- $(c_۱c_۲)\alpha=c_۱(c_۲\alpha)$
- $c(\alpha+\beta)=c\alpha+c\beta$
- $(c_۱c_۲)\alpha=c_۱\alpha+c_۲\alpha$

[1] Vector space

تعریف آ.۲.۱ (فضای باناخ[۲]). اگر یک نرم $||\cdot||$ روی X وجود داشته باشد، آنگاه $(X, ||\cdot||)$ را یک فضای نرم‌دار[۳] گویند. فضای نرم‌دار X با متر $d(x,y) = ||x-y||$ یک فضای متریک است. حال اگر این فضای متریک کامل باشد، یعنی هر دنباله کوشی[۴] در آن همگرا باشد، آن را یک فضای باناخ گویند (رودین، ۱۳۶۲).

تعریف آ.۳.۱ (ضرب داخلی[۵]). فرض کنید $\mathcal{H}$ یک فضای برداری روی میدان F باشد، تابع $<\cdot,\cdot>: \mathcal{H}\times\mathcal{H}\to F$ را یک ضرب داخلی روی $\mathcal{H}$ گویند هرگاه برای هر $x,y,z\in\mathcal{H}$ و $\alpha\in F$ داشته باشیم:

الف) $<x,x>\geq 0$

ب) $<x,x>=0$ اگر و فقط اگر $x=0$،

ج) $<x,y>=\hat{<y,x>}$،

د) $<\alpha x,x>=\alpha<x,y>$،

ه) $<x+y,z>=<x,z>+<y,z>$.

در این صورت $(\mathcal{H}, <\cdot,\cdot>)$ را یک فضای ضرب داخلی[۶] گویند.

قضیه آ.۴.۱. *فرض کنید $(\mathcal{H}, <\cdot,\cdot>)$ یک فضای ضرب داخلی باشد، اگر تابع*

$$||\cdot||: \mathcal{H}\to[0,\inf)$$

را با ضابطه $||x|| = <x,x>^{1/2}$ تعریف کنیم آنگاه $(\mathcal{H}, ||\cdot||)$ یک فضای نرم‌دار خواهد بود.

تعریف آ.۵.۱ (فضای هیلبرت[۷]). فرض کنید $(\mathcal{H}, <\cdot,\cdot>)$ یک فضای ضرب داخلی باشد، اگر فضای نرم‌دار $(\mathcal{H}, ||\cdot||)$ که $||.|| = <\cdot,\cdot>^{1/2}$، یک فضای باناخ باشد، $\mathcal{H}$ را فضای هیلبرت گویند.

آ.۲ معیار کولبک لیبلر

معیار اطلاع کولبک-لیبلر (KL) یک معیار برای ارزیابی میزان نزدیکی مدل‌های آماری به توزیع درست جامعه است. معیار اطلاع KL در واقع میزان ریسک احتمالی انتخاب یک مدل رقابتی به جای توزیع درست است. هرچه این معیار کوچکتر باشد مدل انتخابی مدلی مناسب‌تر خواهد بود.

فرض می‌کنیم که $f(x)$ یک مدل آماری ساخته شده بر اساس مشاهدات باشد. می‌خواهیم نزدیکی این مدل را به تابع چگالی احتمال (تابع جرم احتمال) درست $h(x)$ بر اساس معیار اطلاع KL ارزیابی کنیم. این معیار به صورت

$$KL(h,f) = E_h\left[\log\left\{\frac{h(x)}{f(x)}\right\}\right]$$

که در آن E_h امید ریاضی تحت تابع چگالی یا تابع جرم احتمال h است.

---

[۲] Banach space

[۳] Normed space

[۴] Cauchy sequence

[۵] Inner product

[۶] Inner product space

[۷] Hilbert space

آ.۳ برخی از خواص دترمینان و اثر ماتریس

در این بخش چهار خاصیت از دترمینان و اثر ماتریس که در این پایان نامه مورد استفاده قرار گرفته‌اند، را بیان می‌کنیم. اگر A و B دو ماتریس دلخواه و a یک عدد باشند، آن‌گاه این خواص را برای دترمینان و اثر ماتریس داریم:

۱. $|AB| = |A||B|$

۲. $tr(a) = a$

۳. $tr(AB) = tr(BA)$

۴. $tr(X) + tr(Y) = tr(X + Y)$

پیوست ب

دستورات نرم‌افزار R برای مثال شبیه‌سازی داده‌های اوزن

در این پیوست دستورات مربوط به برازش مدل‌های موجود در فصل چهارم، مربوط به مثال شبیه‌سازی داده‌های اوزن، ارایه شده است.

۱) فراخوانی بسته‌های لازم

```
library(spBayes)
library(MBA)
library(fields)
library(xtable)
library(colorspace)
library(maps)
```

۲) قراخوانی داده‌های اوزن

```
data("NYOzone.dat")
```

۳) رسم ناحیه جغرافیایی مورد مطالعه و تعیین داده‌های آزمون برای اعتبارسنجی مدل

```
N.t <- 62
n <- 28

## hold 31 days out of a stations 1, 5, and 10
hold.out <- c(1,5,10)

par(mar=c(0,0,0,0))
```

```
map(database="state",regions="new york")
points(NYOzone.dat[,c("Longitude","Latitude")], cex=2)
points(NYOzone.dat[hold.out,c("Longitude","Latitude")], pch=19, cex=2)

missing.obs <- sample(1:N.t, 31)
NYOzone.dat[,paste("O3.8HRMAX.",1:N.t,sep="")][hold.out,missing.obs] <- NA
```

۴) مشخص کردن مدل پررتبه و اجرای الگوریتم نمونه‌گیری

```
mods <- lapply(paste("O3.8HRMAX.",1:N.t, "~cMAXTMP.",1:N.t,
                  "+WDSP.",1:N.t, "+RH.",1:N.t, sep=""), as.formula)

p <- 4 ## number of predictors

coords <- NYOzone.dat[,c("X.UTM","Y.UTM")]/1000
max.d <- max(iDist(coords))

starting <- list("beta"=rep(0,N.t*p), "phi"=rep(3/(0.5*max.d), N.t),
                 "sigma.sq"=rep(2,N.t), "tau.sq"=rep(1, N.t),
                 "sigma.eta"=diag(rep(0.01, p)))

tuning <- list("phi"=rep(2, N.t))

priors <- list("beta.0.Norm"=list(rep(0,p), diag(100000,p)),
               "phi.Unif"=list(rep(3/(0.9*max.d), N.t),
               rep(3/(0.05*max.d), N.t)),
               "sigma.sq.IG"=list(rep(2,N.t), rep(25,N.t)),
               "tau.sq.IG"=list(rep(2,N.t), rep(25,N.t)),
               "sigma.eta.IW"=list(2, diag(0.001,p)))

n.samples <- 5000

m.i <- spDynLM(mods, data=NYOzone.dat, coords=as.matrix(coords),
          starting=starting, tuning=tuning, priors=priors, get.fitted=TRUE,
          cov.model="exponential", n.samples=n.samples, n.report=2500)
```

۵) محاسبه نتایج و رسم نمودارهای مربوط برای اثرات رگرسیونی و وابستگی فضایی

```
ts.params.plot <- function(x, xlab="", ylab="", cex.lab=2.5,
 cex.axis=2.5, prob=c(0.5, 0.025, 0.975),
                           mar=c(5,5.5,0.1,1), zero.line=TRUE){
  N.t <- ncol(x)
  q <- apply(x, 2, function(x){quantile(x, prob=prob)})

  par(mar=mar)##bottom, left, top and right
  plot(1:N.t, q[1,], pch=19, cex=0.5, xlab=parse(text=xlab),
 ylab=parse(text=ylab), ylim=range(q), cex.lab=cex.lab, cex.axis=cex.axis)
  arrows(1:N.t, q[1,], 1:N.t, q[3,], length=0.02, angle=90)
  arrows(1:N.t, q[1,], 1:N.t, q[2,], length=0.02, angle=90)
  if(zero.line){abline(h=0, col="blue")}
}

burn.in <- floor(0.75*n.samples)
beta <- m.i$p.beta.samples[burn.in:n.samples,]

par(mfrow=c(4,1))
beta.0 <- beta[,grep("Intercept", colnames(beta))]
ts.params.plot(beta.0, ylab="beta[0]")

beta.1 <- beta[,grep("cMAXTMP", colnames(beta))]
ts.params.plot(beta.1, ylab="beta[cMAXTMP]")

beta.2 <- beta[,grep("WDSP", colnames(beta))]
ts.params.plot(beta.2, ylab="beta[WDSP]")

beta.3 <- beta[,grep("RH", colnames(beta))]
ts.params.plot(beta.3, ylab="beta[RH]", xlab="Days")
###

theta <- m.i$p.theta.samples[burn.in:n.samples,]

par(mfrow=c(3,1))
```

```
sigma.sq <- theta[,grep("sigma.sq", colnames(theta))]
ts.params.plot(sigma.sq, ylab="sigma^2", zero.line=FALSE)

tau.sq <- theta[,grep("tau.sq", colnames(theta))]
ts.params.plot(tau.sq, ylab="tau^2", zero.line=FALSE)

phi <- theta[,grep("phi", colnames(theta))]
ts.params.plot(3/phi, ylab="3/phi (km)", xlab="Days",
 zero.line=FALSE)
```

۶) محاسبه مقادیر پیش‌گویی (میانه توزیع پیش‌گو) و فواصل اعتبار بیزی برای داده‌های آزمون

```
y.hat <- apply(m.i$p.y.samples[,burn.in:n.samples], 1, quants)

data(NYOzone.dat)

obs.O3 <- NYOzone.dat[,paste("O3.8HRMAX.",1:N.t,sep="")]

ylim <- c(15,75)
ylab <- "O3.8HRMAX"

par(mfrow=c(3,1),mar=c(5,5.5,0.1,1))
station.1 <- y.hat[,seq(hold.out[1],ncol(y.hat),n)]

plot(1:N.t, obs.O3[hold.out[1],], ylim=ylim, ylab=ylab, cex.lab=2.5,
cex.axis=2.5, xlab="")
lines(1:N.t, station.1[1,1:N.t])
lines(1:N.t, station.1[2,1:N.t], col="blue", lty=3)
lines(1:N.t, station.1[3,1:N.t], col="blue", lty=3)
points(missing.obs, obs.O3[hold.out[1],missing.obs], pch=19)

station.2 <- y.hat[,seq(hold.out[2],ncol(y.hat),n)]

plot(1:N.t, obs.O3[hold.out[2],], ylim=ylim, ylab=ylab, cex.lab=2.5,
cex.axis=2.5, xlab="")
lines(1:N.t, station.2[1,1:N.t])
```

```
lines(1:N.t, station.2[2,1:N.t], col="blue", lty=3)
lines(1:N.t, station.2[3,1:N.t], col="blue", lty=3)
points(missing.obs, obs.O3[hold.out[2],missing.obs], pch=19)

station.3 <- y.hat[,seq(hold.out[3],ncol(y.hat),n)]

plot(1:N.t, obs.O3[hold.out[3],], ylim=ylim, ylab=ylab, cex.lab=2.5,
 cex.axis=2.5, xlab="Days")
lines(1:N.t, station.3[1,1:N.t])
lines(1:N.t, station.3[2,1:N.t], col="blue", lty=3)
lines(1:N.t, station.3[3,1:N.t], col="blue", lty=3)
points(missing.obs, obs.O3[hold.out[3],missing.obs], pch=19)
```

مراجع

[۱] محمدزاده، م. (۱۳۸۹). *آمار فضایی و کاربردهای آن*، چاپ اول، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

[۲] رودین، و. (۱۳۶۲). *اصول آنالیز ریاضی*، (ترجمه علی اکبر عالم زاده)، چاپ اول، انتشارات مبتکران، تهران.

[3] Abramowitz, M. and Stegun, I. (1970), *Handbook of Mathematical Functions*, Dover, New York.

[4] Armstrong, M., Dowd, P. A. (1994), *Geostatistical Simulations*, Kluver Academic Publishers, France.

[5] Baghishani, H. and Mohammadzadeh, M. (2011), A data cloning algorithm for computing maximum likelihood estimates in spatial generalized linear mixed models, *Computational Statistics & Data Analysis*, **55**(4), 1748-1759.

[6] Banerjee, S., Carlin, C. P. and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman Hall/CRC, Boca Raton, Florida.

[7] Banerjee, S., Gelfand, A. E., Finely, A.O. and Sang, H. (2008), Gaussian predictive process models for large spatial data sets, *Journal of the Royal Statistical Society*, **70**(4), 825-848.

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

[9] Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**(421), 9-25.

[10] Breslow, N. E. and Lin, X. (1995), Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika*, **82**(1), 81-91.

[11] Casella, G. and George, E. (1992), Explaining the Gibbs sampler, *Journal of the American Statistical Association*, **46**(3), 167-174.

[12] Casella, G. Robert, C. and Martin Wells, M. (2004), Generalized Accept–Reject sampling schemes, *Lecture Notes-Monograph Series*, **45**, 342–347.

[13] Csato, L. (2002), *Gaussian Processes—Iterative Sparse Approximation*, PhDThesis, AstonUniversity, Birmingham.

[14] Chiles, J.P. and Delfiner, P. (1999), *Geostatistics; Modeling Spatial Uncertainty*, Wiley, New York.

[15] Clark, I. (1979), Does Geostatistics Work?, *Proceedings of the 16th APCOM International Symposium*, 213-225.

[16] Crainiceanu, C. M., Diggle, P. J. and Rowlingson, B. (2008), Bivariate binomial spatial modeling of Loa loa Prevalence in tropical Africa (with discussion), *Journal of the American Statistical Association*, **103**(481), 21–37.

[17] Cressie, N. (1985), Fitting variogram models by Weighted Least Squares, *Journal of International Association for Mathematical Geology*, **17**(5), 563-586.

[18] Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley, New York.

[19] Cressie, N. and Huang, H. C. (1999), Classes of nonseparable, spatio-temporal stationary covariance functions, *Journal of the American Statistical Association*, **94**(448), 1330–1340.

[20] Cressie, N. and Johannesson, G. (2008), Fixed rank kriging for very large data sets, *Journal of the Royal Statistical Society: Series B*, **70**(1), 209–226.

[21] Cressie, N. and Wikle, C. (2011), *Statistics for Spatio-Temporal Data*, John Wiley, London.

[22] David, M. (1977), *Geostatistical ore Reserve Estimation*, Elsevier, Amsterdam.

[23] Deng, C. Y. (2011), A generalization of the Sherman–Morrison–Woodbury formula, *Elsevier: Applied Mathematics Letters*, **24**(9), 1561–1564.

[24] Diggle, P. and Lophaven, S. (2006), Bayesian geostatistical design, *Scandinavian Journal of Statistics*, **33**(1), 53–64.

[25] Diggle, P. J., Moyeed, R. A. and Tawn, J. A. (1998), Model-Based geostatistics, *Journal of the Royal Statistical Society*, **47**(3), 299-350.

[26] Finely, A. O., Sang, H., Banerjee, S. and Gelfand, A. E. (2009), Improving the performance of predictive process modeling for large datasets, *Computational Statistics & Data Analysis*, **53**(8), 2873-2884.

[27] Finely, A. O., Banerjee, S. and Gelfand, A.E. (2012), Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes, *Journal of Geographical Systems*, **14**(1), 29-47.

[28] Finely, A. O., Banerjee, S. and Gelfand, A. E. (2015), SpBayes for large univariate and multivariate Point-Referenced Spatio-Temporal data models, *Journal of Statistical Software*, **63**(13), 1-28.

[29] Gandin, L. S. (1963), *Objective Analysis of Meteorological Fields*, Gidrometeor, Leningrad.

[30] Gelfand, A. E., Banerjee, S. and Gamerman, D. (2005), Spatial process modelling for univariate and multivariate dynamic spatial data, *Environmetrics*, **16**(5), 465–479.

[31] Gelfand, A. E., Schmidt, A., Banerjee, S. and Sirmans, C. F. (2004), Nonstationary multivariate process modelling through spatially varying coregionalization, *Test*, **13**(2), 263-312.

[32] Gelfand, A. and Smith, A. (1990), Sampling based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**(410), 398-409.

[33] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013), *Bayesian Data Analysis*, Third edition, Chapman Hall/CRC, New York.

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

[35] Gneiting, T. (2002), Nonseparable, stationary covariance functions for space-time data, *Journal of the American Statistical Association*, **97**(458), 590–600.

[36] Golub, G. and Van Loan, C. (1996), *Matrix Computations*, Third edition, Johns Hopkins University Press, London.

[37] Hall, P. (1985), Resampling a coverage process, *Stochastic Processes and Applications*, **20**(2), 231–246.

[38] Hastings, W. (1970), Monte Carlo sampling methods using Markov chains and their application, *Biometrika*, **57**(1), 97-109.

[39] Henderson, H. V. and Searle, S. R. (1981), On deriving the inverse of a sum of matrices, *Society for Industrial and Applied Mathematics*, **23**(1), 53-60.

[40] Hesterberg, T. C. (1987), Importance sampling in multivariate problems, *In Proceedings of the Statistical Computing Section, American Statistical Association 1987 Meeting*, 412-417.

[41] Higdon, D. (2002), Space and space-time modeling using process convolutions. In: Anderson, C., Barnett, C., Chatwin, P. C. and El-Shaarawi, A. H.: *Quantitative methods for current environmental issues*, Springer, Berlin, 37–56.

[42] Hosseini, F., Eidsvik, J. and Mohammadzadeh, M. (2011), Approximate Bayesian inference in spatial GLMM with skew normal latent variables, *Computational Statistics & Data Analysis*, **55**(4), 1791-1806.

[43] Hosseini, F. and Mohammadzadeh, M. (2012), Bayesian prediction for spatial GLMM's with closed skew normal latent variables, *Australian & New Zealand Journal of Statistics*, **54**(1), 43-62.

[44] Hunger, R. (2005), *Floating Point Operations in Matrix-Vector Calculus*, Technical Report, Munich University of Technology.

[45] Jones, R. H. and Zhang, Y. (1997), Models for continuous stationary spacetime processes. In: Gregoire, T., Brillinger, D., Diggle, P. and et al.: *Modelling Longitudinal and Spatially Correlated Data*, Springer, New York, 289–298.

[46] Journel, A. G. and Huijbregts, C. J. (1978), *Mining Geostatistics*, Academic Press, London.

[47] Kammann, E. E. and Wand, M. P. (2003), Geoadditive models, *Journal of the Royal Statistical Society: Series C* , **52**(1), 1–18.

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

[49] Krig, D. G. (1951), A Statistical approach to some basic mine valuation problem on the witwatersrand, *Journal of the Chemical, Metallurgical and Minining Society of South Africa*, **52**(6), 119-139.

[50] Lele, S. R., Dennis, B. and Lutscher, F. (2007), Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods, *Ecology Letters*, **10**(7), 551–563.

[51] Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000), Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV, *The Annals of Statistics*, **28**(6), 1570-1600.

[52] Lindley, D. V. and Smith, A. F. M. (1972), Bayes estimates for the linear models, *Journal of the Royal Statistical Society: Series B*, **34**(1), 1-41.

[53] Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London.

[54] Mardia, K. V. and Marshall, R. J. (1984), Maximum Likelihood Estimation of models for residual covariance in spatial regression, *Biometrika*, **71**(1), 135-146.

[55] Matheron, G. (1962), *Traite de Geostatistique Appliquee, Tome I*, Memoires du Bureau de Recherches Geologiques et Minieres, **14**, Paris.

[56] Matheron, G. (1963), Principles of geostatistics, *Economic Geology*, **58**(8), 1246-1266.

[57] Matheron, G. (1971), *The Theory of Regionalized Variables and its Applications*, Chaiers du Center de Morphologie Mathematique, Fontainebleau, France.

[58] McBratney, A. B. and Webster, R. (1986), Choosing functions for Semi-Variograms of soil properties and fitting them to sampling estimates, *Journal of Soil Science*, **37**(4), 617-639.

[59] McCulloch, C. E. (1997), Maximum Likelihood algorithms for Generalized Linear Mixed Models, *Journal of the American Statistical Association*, **92**(437), 162-170.

[60] McCulloch, C. E. and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, John Wiley, New York.

[61] McCullach, P. and Nelder, J. A. (1989), *Generalized Linear Models*, Second edition, Chapman & Hall, London.

[62] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**(6), 1087-1092.

[63] Nelder, J. A. and Wedderburn, R. W. M. (1972), Generalized linear models, *Journal of the Royal Statistical Society, Series A*, **135**(3), 370-384.

[64] Pinheiro, J. C. and Bates, D. M. (1995), Approximations to the log-likelihood function in the nonlinear mixed-effects models, *Journal of Computational and Graphical Statistics*, **4**(1), 12-35.

[65] Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Second edition, Springer, New York.

[66] Ripley, B. D. (1981), *Spatial Statistics*, John Wiley, New York.

[67] Robert, C. (2007), *The Bayesian Choice*, Second edition, Springer, New York.

[68] Robert, C. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Second edition, Springer, New York.

[69] Roberts, G. O. and Rosenthal, J. S. (2001), Optimal scaling for various Metropolis–Hastings algorithms, *Statistical Science*, **16**(4), 351-367.

[70] Roberts, G. O. and Rosenthal, J. S. (2004), General state space Markov chains and MCMC algorithms, *Probability Surveys*, **1**, 20-71.

[71] Rouhani, S. and Hall, T. J. (1989), Space-Time kriging of Ground-water data. in: Armstrong, M.: *Geostatistics*, Kluwer Academic Publishers, Dordrecht, **2**, 639-650.

[72] Sahu, S. K. and Bakar, K. S. (2011), Comparison of Bayesian models for daily Ozone concentration levels, *Statistical Methodology*, **9**(1-2), 144-157.

[73] Seeger, M., Williams, C. K. I. and Lawrence, N. (2003), Fast forward selection to speed up sparse Gaussian process regression. in: Bishop, C. M. and Frey, B. J.: *Proceedings of the 9th International Workshop Artificial Intelligence and Statistics*, KeyWest: Society for Artificial Intelligence and Statistics.

[74] Sherman, M. (2011), *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*, First edition, Wiley Series in Probability and Statistics, New York.

[75] Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging*, Springer Series in Statistics, New York.

[76] Stein, M. L. (2005), Space-time covariance functions, *Journal of the American Statistical Association*, **100**(469), 310–321.

[77] Stewart, G. W. (1998), Perturbation theory for the singular value decomposition, *DRUM: Digital Repository at the University of Maryland*, University of Maryland.

[78] Strang, G. (1998), *Introduction to Linear Algebra*, Third edition, Wellesley-Cambridge Press.

[79] Stroud, R. J., Muller, P. and Sanso, B. (2001), Dynamic models for spatio-temporal data, *Journal of the Royal Statistical Society, Series B*, **63**(4), 673-689.

[80] Tonellato, S. (1997), Bayesian dynamic linear models for spatial time series, *Tech report, Rapporto di riceria 5/1997*, Dipartimento di Statistica - Universita CaFoscari di Venezia, Venice Italy.

[81] Wackernagel, H. (2006), *Multivariate Geostatistics: An Introduction with Applications*, Third edition, Springer, New York.

[82] Xia, G., Miranda, M. L. and Gelfand, A. E. (2006), Approximately optimal spatial design approaches for environmental health data, *Environmetrics*, **17**(4), 363–385.

واژه‌نامه فارسی به انگلیسی

آمیخته ..... Mixed
ابر پیشین ..... Hyper prior
اثر تصادفی ..... Random effect
اثر فضایی ..... Spatial effect
اسپلاین ..... Spline
اعتبارسنجی متقابل ..... Cross validation
الگوریتم گیبز ..... Gibbs algorithm
الگوریتم متروپولیس- هستینگز ..... Metropolis- Hastings algorithm
آمار فضایی ..... Spatial statistics
الگوی نقطه‌ای ..... Point pattern
اثر قطعه‌ای ..... Nugget effect
ازاره ..... Sill
بازنمونه‌گیری ..... Resampling
بازه اطمینان ..... Confidence interval
برآوردگر ماکسیمم درستنمایی ..... Maximum likelihood estimator
برازش ..... Fit
بهترین پیش‌گوی خطی نااریب ..... Best linear unbiased predictor
بیز سلسله مراتبی ..... Hierachical Bayes
بهینه‌سازی ..... Optimization
پارامتر ..... Parameter
پارامتر شکل ..... Shape parameter
پارامتر مقیاس ..... Scale parameter
پارامتر هموارى ..... Smothing parameter
پیش‌گو ..... Predictor
پیش‌گوی بهینه ..... Optimal predictor

پیش‌گویی ..... Prediction
پیش‌گویی فضایی ..... Spatial prediction
تابع بسل ..... Bessel function
تابع پیوند ..... Link function
تابع چگالی ..... Density function
تابع درستنمایی ..... Likelihood function
تجزیه چولسکی ..... Cholesky decomposition
تحلیل بیزی ..... Bayesian analysis
تحلیل بسامدی ..... Frequentist analysis
تحلیل فضایی ..... Spatial analysis
تحلیل فضایی- زمانی ..... Spatio- temporal analysis
تحلیل کاوش‌گرانه داده‌های فضایی ..... Spatial exploration data analysis
ترکیب خطی ..... Linear combination
تغییرات بزرگ‌مقیاس ..... Large scale variation
تغییرات کوچک‌مقیاس ..... Small scale variation
تغییرنگار ..... Variogram
تغییرنگار تجربی ..... Empirical variogram
تغییرنگار نمایی ..... Exponential variogram
توزیع نامزد ..... Proposal distribution
توزیع شرطی کامل ..... Full conditional distribution
توزیع گاوسی ..... Gaussian distribution
توزیع وارون گاما ..... Inverse Gamma distribution
توزیع وارون ویشارت ..... Inverse wishart distribution
توزیع پیش‌گو ..... Predictive distribution
توزیع پیش‌گوی بیزی ..... Bayesian predictive distribution
توزیع پیشینی ..... Prior distribution
توزیع پسینی ..... Posterior distribution
چگالی نامزد ..... Proposal density
چگالی هدف ..... Target density
خطای اندازه‌گیری ..... Measurement error
خوش آمیزنده ..... Mixing well
داده‌های روندزدوده ..... Deterended data

داده‌های زمین‌آماری Geostatistical data
داده‌های فضایی Spatial data
داده‌های فضایی- زمانی Spatio- temporal data
داده‌های مشبکه‌ای Lattice data
دامنه Range
روند Trend
ساختار فضایی Spatial structure
شبیه‌سازی Simulation
ضرایب پویا Dynamic cofficients
ضرب ماتریسی کرونه‌کر Kronecker matrix product
کالبیده Calibrated
کم برآورد Under estimate
کوواریانس فضایی Spatial covariance
کوواریانس فضایی- زمانی Spatio- temporal covariance
ماکسیمم‌سازی امیدریاضی Expectation maximization
مانای ذاتی Intrinsically stationary
مانای قوی Strong stationary
مانای مرتبه دوم Second order stationary
متعادل Balanced
متغیر پنهان Latent variable
متغیر تبیینی Covariate
مدل آمیخته خطی تعمیم‌یافته Generalized linear mixed model
مدل خطی پویا Dynamic linear model
مدل نمایی Exponential model
مشبکه منظم Regular lattice
معادله مشاهداتی Observation equation
معادله فرگشتی Evolution equation
موقعیت Location
میدان تصادفی Random field
میدان تصادفی فضایی- زمانی Spatio- temporal random field
نمونه‌گیری گیبز Gibbs sampling
نیم‌تغییرنگار Semi-variogram

وابستگی فضایی ..... Spatial dependence
هم‌بستگی‌نگار ..... Correlogram
هم‌تغییرنگار ..... Covariogram
همسان‌گرد ..... Isotropic
همیشه مثبت ..... Positive definite
همیشه منفی ..... Negative definite

واژه‌نامه انگلیسی به فارسی

متعادل ..... Balanced
تحلیل بیزی ..... Bayesian analysis
توزیع پیش‌گوی بیزی ..... Bayesian predictive distribution
تابع بسل ..... Bessel function
بهترین پیش‌گوی خطی نااریب ..... Best linear unbiased predictor
کالبیده ..... Calibrated
تجزیه چولسکی ..... Cholesky decomposition
بازه اطمینان ..... Confidence interval
هم‌بستگی‌نگار ..... Correlogram
متغیر تبیینی ..... Covariate
هم‌تغییرنگار ..... Covariogram
اعتبارسنجی متقابل ..... Cross validation
تابع چگالی ..... Density function
داده‌های روندزدوده ..... Deterended data
ضرایب پویا ..... Dynamic cofficients
مدل خطی پویا ..... Dynamic linear model
تغییرنگار تجربی ..... Empirical variogram
معادله فرگشتی ..... Evolution equation
ماکسیمم‌سازی امیدریاضی ..... Expectation maximization
تحلیل کاوش‌گرانه داده‌ها ..... Exploration data analysis
مدل نمایی ..... Exponential model
تغییرنگار نمایی ..... Exponential variogram
برازش ..... Fit
تحلیل بسامدی ..... Frequentist analysis
توزیع شرطی کامل ..... Full conditional distribution

توزیع گاوسی ..... Gaussian distribution
مدل آمیخته خطی تعمیم‌یافته ..... Generalized linear mixed model
داده‌های زمین‌آماری ..... Geostatistical data
الگوریتم گیبز ..... Gibbs algorithm
نمونه‌گیری گیبز ..... Gibbs sampling
ابر پیشین ..... Hyper prior
مانای ذاتی ..... Intrinsically stationary
توزیع وارون گاما ..... Inverse Gamma distribution
توزیع وارون ویشارت ..... Inverse wishart distribution
همسان‌گرد ..... Isotropic
ضرب ماتریسی کرونه‌کر ..... Kronecker matrix product
تغییرات بزرگ‌مقیاس ..... Large scale variation
متغیر پنهان ..... Latent variable
داده‌های مشبکه‌ای ..... Lattice data
تابع درستنمایی ..... Likelihood function
ترکیب خطی ..... Linear combination
تابع پیوند ..... Link function
موقعیت ..... Location
برآوردگر ماکسیمم درستنمایی ..... Maximum likelihood estimator
خطای اندازه‌گیری ..... Measurement error
الگوریتم متروپولیس- هستینگز ..... Metropolis- Hastings algorithm
خوش آمیزنده ..... Mixing well
همیشه منفی ..... Negative definite
اثر قطعه‌ای ..... Nugget effect
معادله مشاهداتی ..... Observation equation
بهینه‌سازی ..... Optimization
پارامتر ..... Parameter
الگوی نقطه‌ای ..... Point pattern
همیشه مثبت ..... Positive definite
توزیع پسینی ..... Posterior distribution
پیش‌گویی ..... Prediction
توزیع پیش‌گو ..... Predictive distribution

پیش‌گو ..... Predictor
توزیع پیشینی ..... Prior distribution
چگالی نامزد ..... Proposal density
توزیع نامزد ..... Proposal distribution
اثر تصادفی ..... Random effect
میدان تصادفی ..... Random field
دامنه ..... Range
مشبکه منظم ..... Regular lattice
بازنمونه‌گیری ..... Resampling
پارامتر مقیاس ..... Scale parameter
مانای مرتبه دوم ..... Second order stationary
نیم‌تغییرنگار ..... Semi-variogram
پارامتر شکل ..... Shape parameter
تغییرنگار تجربی ..... Empirical variogram
ازاره ..... Sill
شبیه‌سازی ..... Simulation
تغییرات کوچک‌مقیاس ..... Small scale variation
پارامتر همواری ..... Smothing parameter
تحلیل فضایی ..... Spatial analysis
کوواریانس فضایی ..... Spatial covariance
داده‌های فضایی ..... Spatial data
وابستگی فضایی ..... Spatial dependence
تحلیل کاوش‌گرانه داده‌های فضایی ..... Spatial exploration data analysis
پیش‌گویی فضایی ..... Spatial prediction
آمار فضایی ..... Spatial statistics
ساختار فضایی ..... Spatial structure
داده‌های فضایی- زمانی ..... Spatio- temporal data
تحلیل فضایی- زمانی ..... Spatio- temporal analysis
کوواریانس فضایی- زمانی ..... Spatio- temporal covariance
میدان تصادفی فضایی- زمانی ..... Spatio- temporal random field
اسپلاین ..... Spline
مانای قوی ..... Strong stationary

چگالی هدف ..... Target density
روند ..... Trend
کم برآورد ..... Under estimate
تغییرنگار ..... variogram

نمایه

Abstract

Analysis of large geostatistical data sets, usually, entail the expensive matrix computations. This problem creates challenges in implementing statistical inferences of traditional Bayesian models. In addition, researchers often face with multiple spatial data sets with complex spatial dependence structures that their analysis is difficult. Complex structures of (spatial or spatio-temporal) data, their sizes and available incomplete models, have been pushed users to develope new and dynamic models. In this thesis, we use low-rank models, particularly predictive process models, to analyze Gaussian geostatistical data. This models improve MCMC sampler convergence rate and decrease sampler run-time by reducing parameter space. This improvement is due to the avoidance of expensive matrix computations that also can be used for large data sets. To evaluate the performance of this class of models, we conduct a simulation study as well as analysis of a real data set regarding the quality of underground mineral water of a large area in Golestan province, Iran.

keywords: Inference Bayesian; MCMC Sampling; Predictive process; Spatio-Temporal; Geostatistical

Shahrood University of Technology

Faculty of Mathematical Sciences

Gaussian univariate and multivariate spatio-temporal models for large geostatistical data

Bahman Hamidian

Supervisor

Dr. Hossein Baghishani

February 2016