
بسم الله الرحمن الرحيم

دانشکده علوم ریاضی

پایان‌نامه کارشناسی ارشد آمار ریاضی

مدل‌بندی سه بعدی ژئومکانیکی بیزی با استفاده از مدل‌های زمین‌آماری: مطالعه موردی یکی از مخازن نفتی ایران

نگارنده: فتانه فخری

استاد راهنما
دکتر حسین باغیشنی

استاد مشاور
دکتر بهزاد تخم‌چی

مهر ۱۳۹۹

شماره:

تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

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

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد **خانم فتانه فخری** با شماره دانشجویی ۹۶۳۵۰۴۴ رشته آمار ریاضی گرایش ----- تحت عنوان **مدل بندی سه بعدی ژئومکانیکی بیزی با استفاده از مدلهای زمین آماری مطالعه موردی یکی از مخازن نفتی ایران** که در تاریخ **۱۳۹۹/۰۷/۲۹** با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☑ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☐

ج) درجه خوب: نمره۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐

ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐

نوع تحقیق: **نظری** ☑ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر حسین باغیشنی	استادیار	
۲- استادراهنمای دوم	-----	-----	-----
۳- استاد مشاور	دکتر بهزاد تخم چی	دانشیار	
۴- نماینده تحصیلات تکمیلی	دکتر محمدهادی نوری اسکندری	استادیار	
۵- استاد ممتحن اول	دکتر محمدرضا ربیعی	استادیار	
۶ـ استاد ممتحن دوم	دکتر داود شاهسونی	دانشیار	

نام و نام خانوادگی رئیس دانشکده: دکتر مهدی قوتمنه

تاریخ و امضاء و مهر دانشکده:

با تمام کاستی‌ها تقدیم به

خدای مهربان
و خانواده‌ی عزیزم

سپاس‌گزاری

سپاس بی‌کران پروردگار یکتا را که هستی‌مان بخشید، به طریق علم و دانش رهنمون‌مان شد، به هم‌نشینی رهروان دانش مفتخرمان نمود و خوشه‌چینی از علم و معرفت را روزی‌مان ساخت.

با امتنان بیکران از مساعدت‌های بی‌شائبه‌ی اساتید بزرگوار خود، دکتر حسین باغیشنی و دکتر بهزاد تخم‌چی صمیمانه کمال تشکر را دارم، که نه تنها در نوشتن این پایان‌نامه و فراگیری علم با سعه‌صدر مرا یاری دادند و از هیچ کمکی دریغ ننمودند، بلکه درس اخلاق، فروتنی، استقامت و امید را نیز به من دادند.

از استاد بزرگوار جناب آقای دکتر داوود شاهسونی که یک ترم سرکلاس ایشان حضور داشتم و در فراگیری علوم جدید آماری به بنده بسیار کمک کردند و زحمت داوری این پایان‌نامه را نیز بر عهده گرفتند صمیمانه تشکر میکنم.

از استاد محترم، جناب آقای دکتر محمدرضا ربیعی که لطف کردند و داوری این پایان‌نامه را بر عهده گرفتند صمیمانه تشکر مینمایم.

و در نهایت از خانواده عزیزم و دوستانی که همیشه کنارم بوده‌اند، سپاسگزارم.

از خداوند مهربانم براي شما عزیزان، آرزوی سلامتی و توفیق روز افزون را مسألت دارم.

فتانه فخری

مهر ۱۳۹۹

تعهد نامه

اینجانب **فتانه فخری** دانشجوی کارشناسی ارشد رشته **آمار ریاضی** دانشکده **علوم ریاضی** دانشگاه صنعتی شاهرود، نویسنده پایان‌نامه با عنوان **مدل‌بندی سه بعدی ژئومکانیکی بیزی با استفاده از مدل‌های زمین‌آماری: مطالعه موردی یکی از مخازن نفتی ایران** ، تحت راهنمایی دکتر **حسین باغیشنی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود’’ یا “ Shahrood University of Technology ” به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌شود.
- 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.

فتانه فخری
مهر ۱۳۹۹

مالکیت نتایج و حق نشر

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

چکیده

استفاده از ویژگی‌های مکانیکی سنگ برای مطالعه فضاهای زیرزمینی، ژئومکانیک نامیده می‌شود. ژئومکانیک یکی از شاخه‌های علمی مهم در حوزه اکتشاف و حفاری مخازن زیرزمینی، به‌ویژه مخازن نفتی، است. اغلب داده‌های حاصل در مطالعات ژئومکانیکی، از نوع زمین‌آماری هستند که مدل‌بندی و تحلیل آن‌ها نیازمند علم آمار فضایی است. از طرفی، بسیاری از داده‌های زمین‌آماری در مطالعات ژئومکانیکی سه‌بعدی و حجیم هستند که توسعه مدل‌های آمار فضایی و روش‌های محاسباتی کارا برای تحلیل آن‌ها را با چالش مواجه می‌کند. با نگاهی به این واقعیت و برای مقابله با دو چالش ذکرشده، در این پایان‌نامه، برای اولین بار، دو روش پیشنهادی جدید را در حوزه مدل‌بندی سه بعدی فضایی ارایه می‌کنیم. اول، یک روش ناپارامتری به نام میانه‌پالایش سه بعدی را برای برآورد تابع روند یک میدان فضایی سه بعدی معرفی می‌کنیم. با استفاده از این روش، پیشگویی فضایی کریگینگ میانه‌پالایش را نیز توسعه می‌دهیم. همچنین کارایی کریگینگ میانه‌پالایش پیشنهادی را با یک کریگینگ عام که تابع روند در آن با روش اسپلاین برآورد می‌شود، در مثال‌های شبیه‌سازی و واقعی مربوط به داده‌های مقاومت فشاری تک‌محوری سنگ در یکی از مخازن نفتی ایران، مقایسه می‌کنیم. دوم، با استفاده از یک رویکرد بیزی، روش تقریبی بیزی لاپلاس آشیانه‌ای جمع‌بسته (INLA) و ترکیب آن با روش معادلات دیفرانسیل جزیی تصادفی (SPDE) را به حالت سه بعدی توسعه می‌دهیم.

کلمات کلیدی: زمین‌آمار، استنباط بیزی، کریگینگ میانه‌پالایش سه بُعدی، تقریب لاپلاس آشیانه‌ای جمع‌بسته، رهیافت INLA+SPDE.

پیش‌گفتار

از گذشته تا به امروز یکی از مسائل اساسی که افکار بسیاری از پژوهشگران را به خود معطوف کرده است، کشف رفتار مواد و موضوع مورد بررسی در مکان‌های فاقد نمونه است. در واقع، محققان ترجیح می‌دهند قبل از نمونه‌برداری، ذهنیتی نسبی از رفتار مواد در آن مکان داشته باشند. این ترجیح به‌ویژه در علوم زمین که نمونه‌برداری هزینه‌های گزافی را بر عهده‌ی پژوشگر متحمل می‌سازد، بارزتر است. روشن است که آگاهی بر رفتار مواد در ناحیه تحت مطالعه، می‌تواند موجب کاهش هزینه‌های اقتصادی و صرفه‌جویی در زمان شود. بنابراین، پیشگویی نقش بسزایی در علوم مختلف دارد. در علوم زمین و نفت اطلاعات مورد نیاز برای مدل‌بندی و پیشگویی رفتار سنگ، در یک مکان از پیش تعریف‌شده، با نمونه‌برداری از زمین به‌صورت مستقیم و در مواردی با استفاده از روابطی معلوم، اطلاعات مورد نیاز محاسبه می‌شوند. این نوع از داده‌ها که از لحاظ مفهومی پیوسته هستند، مثالی از داده‌های زمین‌آماری است. برای پیشگویی داده‌های زمین‌آماری باید روند داده‌ها را تشخیص داد. اگر نتوان ضابطه‌ای پارامتری برای تابع روند تشخیص داد، از روش‌های ناپارامتری برای مدل‌بندی آن استفاده می‌شود. در تحقیقات متعددی این نوع از مدل‌سازی‌ها به‌خوبی توسعه داده شده‌اند، اما حجم بالا و سه بعدی بودن داده‌ها در مطالعات ژئومکانیکی، پژوهشگران این حیطه را با مشکلات زیادی روبرو می‌سازد.

یکی از روش‌های ناپارامتری در مدل‌بندی تابع روند که کمتر به آن توجه شده است، روش میانه‌پالایش است که برای داده‌های سه بعدی با حجم بالا از پیچیدگی محاسباتی در اجرا برخوردار است. از طرفی، مشابه همین مشکل را می‌توان در تحلیل بیزی این نوع از داده‌ها دریافت. روش‌های مبتنی بر نمونه‌گیری مانند الگوریتم‌های MCMC روش معمول برای محاسبه‌ی توزیع توام پسین مدل‌های آماری است. اما این روش‌ها از محدودیت‌های محاسباتی سنگین و مشکلات همگرایی برخوردار است. به همین دلیل، رو و همکارانش در سال ۲۰۰۹ یک روش تقریبی معروف به روش تقریب لاپلاس آشیانی جمع‌بسته (INLA) را ارائه کردند. در سال ۲۰۱۵ لیندگرین و همکارانش برای مدل‌بندی کارای داده‌های زمین‌آماری با استفاده از INLA، روش معادلات دیفرانسیل جزیی تصادفی (SPDE) را پیشنهاد دادند. تا قبل از این پژوهش به‌ندرت از رویکرد بیزی با استفاده از تقریب INLA برای داده‌های سه بعدی استفاده شده بود. چالش‌های این مسائل موجب شد که نگارنده را برای بررسی‌های بیشتر در این زمینه وادار سازد. بر اساس مقدمات گفته‌شده، ساختار این پایان‌نامه به‌صورت زیر است:

- فصل اول مروری بر تعاریف و مقدمات آمار فضایی است.
- در فصل دوم مدل‌های سه بعدی زمین‌آماری را معرفی می‌کنیم و ابزار برآورد ساختار وابستگی داده‌ها در یک میدان سه بعدی را تشریح می‌کنیم. همچنین روش کریگینگ معمولی سه بعدی را بر روی داده‌های مقاومت فشاری تک‌محوری سنگ در یکی از مخازن نفتی پیاده می‌کنیم.

- در فصل سوم روش میانه‌پالایش سه بعدی و الگوریتم اجرای آن را معرفی می‌کنیم. سپس با استفاده از مثال‌های شبیه‌سازی و داده‌های مقاومت فشاری تک‌محوری سازند، عملکرد آن را با روش ناپارامتری اسپلاین در پیشگویی فضایی به همراه کریگینگ عام مقایسه می‌کنیم.
- در فصل چهارم، جزییات روش بیزی تقریبی INLA را معرفی می‌کنیم و با استفاده از یک مثال شبیه‌سازی روش INLA + SPDE سه بعدی پیشنهادی را اجرا و عملکرد آن را ارزیابی می‌کنیم. همه کدهای کامپیوتری برای اجرای روش‌های این پایان‌نامه در محیط نرم‌افزار R نوشته شده‌اند.

فهرست مقاله‌های مستخرج از پایان‌نامه

۱. فخری، ف.، باغیشنی، ح.، تخم‌چی، ب. (۱۳۹۷)، ناهمسانگردی هندسی در پیشگویی فضایی، اولین کنفرانس دانشجویی آمار، دانشگاه تربیت مدرس تهران، ایران.

۲. فخری، ف.، باغیشنی، ح.، تخم‌چی، ب. (۱۳۹۸)، پیشگویی فضایی سه بعدی به منظور برآورد مقاومت فشاری تک محوری در یکی از مخازن نفتی ایران، سومین کنفرانس آمار فضایی، زنجان، ایران، ۱۳۲ – ۱۲۴.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

آمار فضایی

ظهور علم آمار در داده‌هایی که به موقعیت مکانی و جهت قرارگیری از هم وابسته هستند، اولین بار در تحلیل نقشه‌ها رخ داد. برای مثال هالی (۱۶۸۶) جهت بادهای تجاری و موسمی مناطق استوایی را بر روی بخش‌هایی از نقشه‌ی زمین بررسی کرد. وی سعی کرد جهت این دو نوع باد را به یک عامل فیزیکی نسبت دهد. استیودنت (۱۹۰۷) به بررسی توزیع ذرات معلق درون مایعات علاقه‌مند بود. او به‌جای تحلیل مکانی ذرات، تعداد داده‌ها در هر ناحیه را به‌صورت گلبول‌شمار[۱] در هر میلی‌مترمربع شمارش و هر ناحیه را به چهارصد مربع تقسیم کرد و برای تحلیل از تعداد سلول‌های مخمر استفاده کرد. استیودنت متوجه شد که تعداد سلول‌های مخمر در هر مربع از توزیع پواسن تبعیت می‌کند. فیشر به وابستگی مکانی داده‌ها در آزمایش‌های میدانی کشاورزی واقف بود چرا که وی برای حذف این وابستگی‌ها، کرت‌های آزمایشی را در فواصل دور از هم برمی‌گزید. وی قوانین تصادفی‌سازی، بلوک‌بندی و تکراری را بنا کرد (فیشر، ۱۹۳۵). همچنین یتز (۱۹۳۸) با نشر یافته‌های پژوهش خود ادعا کرد که تصادفی‌سازی علاوه بر کنترل اریبی ناخواسته، اثر همبستگی مکانی را نیز کاهش می‌دهد، ولی به‌طور کامل از بین نمی‌برد. اسمیت (۱۹۳۸) به دنبال انتخاب ابعاد کرت‌های کشاورزی بود و دریافت که هرگونه افزایش در اندازه‌ی کرت، کاهش واریانس خطا را نتیجه می‌دهد. اگرچه تحلیل‌های او تجربی بودند و فرمول‌بندی آن‌ها دشوار اما وجود همبستگی مکانی را در آزمایشات میدانی‌اش به رسمیت می‌شناخت. همچنین از روش‌های نزدیکترین

[۱] Hemocytometer

همسایگی[۲] برای تحلیل داده‌ها در علوم کشاورزی با در نظر گرفتن همبستگی مکانی به‌طور غیرمستقیم، یا با استفاده از باقیمانده‌های همسایگی‌های کرت‌های مجاور به‌عنوان متغیرهای مورد بررسی استفاده می‌شد (پاپدکیس، ۱۹۳۷؛ بارتلت، ۱۹۷۸؛ ویلکینسون و همکاران، ۱۹۸۳؛ بساگ، ۱۹۸۶). در حال حاضر کاربرد آمار را تقریباً در هر شاخه علمی می‌توان یافت. در حیطه‌ی زمین‌شناسی، زیست‌شناسی و علوم زیست محیطی، غالباً (و البته نه همیشه) داده‌ها را نمی‌توان به‌صورت تصادفی در نظر گرفت یا بلوک‌بندی کرد و همچنین امکان تکرار داده‌ها وجود ندارد. پس نیاز به مدل‌های آماری بر اساس رویکردهای جدید برگرفته از تکنولوژی‌های قدیمی و جدید، وجود دارد. حل بسیاری از مسائل موجود از قبیل ارزیابی منابع، نظارت بر محیط زیست (به‌عنوان مثال، گرم شدن هوای کره زمین) و تصویربرداری پزشکی، تابع اطلاعات برگرفته از داده‌هایی است که معمولا به موقعیت مکانی و جهت قرارگیری آن‌ها در ناحیه جغرافیایی مورد مطالعه بستگی دارد. برای تحلیل این نوع داده‌ها، که به داده‌های فضایی معروف هستند، روش‌های کلاسیک استنباط آماری جوابگو نیستند. وارد کردن ساختار وابستگی القاشده توسط موقعیت‌های مکانی داده‌ها (ساختار وابستگی فضایی) در تحلیل داده‌ها توسط روش‌های آمار فضایی ممکن شده است (کرسی، ۱۹۹۳).

در این بخش ابتدا انواع داده‌های فضایی را معرفی می‌کنیم و سپس مفاهیم اولیه آمار فضایی، پیشگوی فضایی و استنباط بیزی را ارائه می‌دهیم.

۱.۱ انواع داده‌های فضایی

در توصیف ماهیت داده‌های فضایی، تمایز بین گسستگی و پیوستگی فضای موقعیت‌ها و نیز مقادیر اندازه‌گیری شده برای هر متغیر مهم است. طبقه‌بندی داده‌های فضایی با نوع مفهوم فضا و سطح اندازه‌گیری اولین گام ضروری در انتخاب روش آماری مناسب برای حل یک سوال است. اما این طبقه‌بندی به تنهایی کافی نیست، چرا که ممکن است یک شی فضایی نمایان‌گر فضاهای جغرافیایی کاملا متفاوتی باشد. کرسی (۲۰۱۵) داده‌های فضایی را بر اساس تقسیم‌بندی سنتی و بسته به نوع مسئله به سه نوع داده‌های مشبکه‌ای[۳]، الگوی نقطه‌ای[۴] و زمین‌آماری[۵] تقسیم کرد. گاهی زمان به‌عنوان یک پارامتر در داده‌های فضایی تاثیرگذار است و با گذشت زمان داده‌ها اندازه‌گیری می‌شوند، از مدل‌های فضایی–زمانی[۶] می‌توان برای مدل‌بندی این نوع از داده‌ها، استفاده کرد (کرسی و ویکل، ۲۰۱۵). علاوه بر این فیشر و وانگ (۲۰۱۱) بر اساس یک نوع‌شناسی، داده‌های فضایی را به چهار گروه داده‌های

[۲] Nearest-neighbor methods
[۳] Lattice data
[۴] Point pattern
[۵] Geostatistical data
[۶] spatio-temporal

زمین‌آماری، الگوی نقطه‌ای، ناحیه‌ای[۷] و داده‌های متقابل فضایی[۸] تقسیم کردند. متغیرهای اندازه‌گیری شده در آمار فضایی ممکن است گسسته یا پیوسته باشند. در ادامه هر کدام از انواع داده‌های فضایی را معرفی می‌کنیم.

داده‌های زمین‌آماری: این نوع از داده‌ها که داده‌های میدانی[۹] نیز نامیده می‌شوند، مربوط به متغیرهایی هستند که از لحاظ مفهومی پیوسته (دید میدانی) و مشاهدات آن‌ها از یک مجموعه ثابت و از پیش تعریف‌شده، نمونه‌برداری می‌شوند. در این نوع از داده‌ها متغیر مورد بررسی می‌تواند گسسته یا پیوسته باشد. برای حالت پیوسته می‌توان اندازه‌گیری مقدار مقاومت فشاری تک‌محوری سازند را در چاه مثال زد و برای حالت گسسته می‌توان تعداد کرم‌های خاکی درون نمونه خاک‌های برداشت‌شده در مکان‌های یک زمین کشاورزی حاصلخیز را نام برد.

داده‌های الگوی نقطه‌ای: به داده‌های متناهی در ناحیه‌ی مورد مطالعه، متشکل از مجموعه مکان‌های نقطه‌ای که رخداد مورد نظر در آن نقاط روی داده است، داده‌های الگوی نقطه‌ای می‌گویند. در این نوع از داده‌ها مکان خود یک متغیر است. داده‌های الگوی نقطه‌ای در سه دسته کاملاً فضایی تصادفی[۱۰]، منظم[۱۱] و خوشه‌ای[۱۲] تقسیم‌بندی می‌شوند (کرسی، ۱۹۹۳). وجود یک نوع بیماری خاص در منطقه‌ی تحت مطالعه یا بروز یک نوع جرم را می‌توان مثالی برای این نوع داده‌ها دانست.

داده‌های ناحیه‌ای: داده‌های ناحیه‌ای یا مشبکه‌ای[۱۳] داده‌هایی متناظر با مشاهدات با تعداد ثابت از یک فضای واحد (شی فضایی) هستند که ممکن است یک مشبکه منظم مانند تصاویر سنجش از دور یا مجموعه‌ی نامنظم منطقه یا محدوده، مانند بخش و شهرستان‌های محدوده‌ی سرشماری یا حتی کشورها باشد. هدف اصلی از تحلیل داده‌های فضایی ناحیه‌ای مدل‌بندی احتمالاتی مشاهدات است، در حالی‌که در زمین‌آمار هدف اصلی پیش‌گویی مقدار متغیر در یک مکان بدون مشاهده است.

داده‌های متقابل فضایی: این نوع از داده‌ها به داده‌های روند مبدا ـ مقصد یا داده‌های ربطی شهرت دارند و مقدار اندازه‌گیری شده مربوط به پیوند بین دو نقطه یا دو ناحیه را شامل می‌شود. داده‌های مربوط به تجارت بین‌المللی یا مهاجرت از جمله داده‌های متقابل فضایی هستند.

---

[۷] Areal
[۸] Spatial interaction data
[۹] Field data
[۱۰] Complete spatial randomness
[۱۱] Requlority
[۱۲] Clustring
[۱۳] Lattice data

۲.۱ مدل‌بندی آماری فضایی

برای تحلیل داده‌های فضایی، یک مدل آماری در نظر گرفته می‌شود. در آمار فضایی معمولاً یک میدان تصادفی که مجموعه‌ای از متغیرهای تصادفی مانند $\{Z(\mathbf{s}); \mathbf{s} \in D\}$ است به‌عنوان مدل آماری داده‌های فضایی در نظر گرفته می‌شود، که در آن مجموعه اندیس‌گذار D یک زیر مجموعه از فضای اقلیدسی d بعدی، $d \geqslant ۱$، از $\mathbb{R}^d$ است. در مورد میدان تصادفی $Z(\cdot)$، میانگین در موقعیت $\mathbf{s}$ و کوواریانس در موقعیت‌های $\mathbf{s}_۱$ و $\mathbf{s}_۲$ به ترتیب به‌صورت

$$\begin{aligned} \mu(\mathbf{s}) &= E[Z(\mathbf{s})], \mathbf{s} \in D \\ C(\mathbf{s}_۱, \mathbf{s}_۲) &= Cov[Z(\mathbf{s}_۱), Z(\mathbf{s}_۲)] \\ &= E[(Z(\mathbf{s}_۱) - \mu(\mathbf{s}_۱))(Z(\mathbf{s}_۲) - \mu(\mathbf{s}_۲))], \quad \mathbf{s}_۱, \mathbf{s}_۲ \in D \end{aligned} \tag{۱.۱}$$

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

$$Var[Z(\mathbf{s})] = E[(Z(\mathbf{s}) - \mu(\mathbf{s})]^۲ = C(\mathbf{s}, \mathbf{s}) \tag{۲.۱}$$

حاصل می‌شود. هر میدان تصادفی $Z(\cdot)$ را می‌توان به‌صورت

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in D \tag{۳.۱}$$

تجزیه کرد، که در آن $\mu(\cdot)$ تغییرات بزرگ‌مقیاس[۱۴] یا روند[۱۵] و $\delta(\cdot)$ فرآیند خطا یا تغییرات کوچک‌مقیاس[۱۶] میدان نامیده می‌شوند. تغییرات کوچک‌مقیاس ممکن است ناشی از خطای اندازه‌گیری یا تغییرپذیری در درون موقعیت مشاهده‌شده باشد و تغییرات بزرگ‌مقیاس ممکن است ناشی از تغییرپذیری بین موقعیت‌های مشاهده‌شده باشند. به‌طور معمول تحلیل داده‌های فضایی بر اساس مشاهدات نمونه بسیار دشوار است، اما برخی ویژگی‌های میدان تصادفی موجب ساده‌سازی مسأله خواهند شد، که در ادامه معرفی می‌شوند (محمدزاده، ۱۳۹۸).

تعریف ۱.۲.۱. میدان تصادفی $Z(\cdot)$، مانای ذاتی[۱۷] نامیده می‌شود، اگر

۱. میانگین میدان تصادفی ثابت یا مستقل از مکان $(\mathbf{s})$ باشد، یعنی $E(Z(\mathbf{s})) = \mu$.

۲. برای عبارت $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))$ واریانس باید تنها تابعی از موقعیت‌های $\mathbf{s}_i$ و $\mathbf{s}_j$ باشد، یعنی

$$Var(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) = ۲\gamma(\mathbf{s}_i - \mathbf{s}_j), \quad \mathbf{s}_i, \mathbf{s}_j \in D \subset \mathbb{R}^d.$$

بنابراین واریانس به مکان قرارگیری مشاهدات از هم بستگی ندارد و در همه جای میدان ثابت است.

[۱۴] Large scale variation
[۱۵] Trend
[۱۶] Small scale variation
[۱۷] Intrinsic stationary

تعریف ۱.۲.۱. میدان تصادفی $Z(\cdot)$، مانای مرتبه دوم[۱۸] یا مانای ضعیف نامیده می‌شود، اگر

۱. میانگین میدان تصادفی $Z(\cdot)$، مستقل از مکان $\mathbf{s}$ یا ثابت باشد. یعنی

$$E(Z(\mathbf{s})) = \mu, \quad \mathbf{s} \in D \subset \mathbb{R}^d.$$

۲. کوواریانس $Z(\mathbf{s}_i)$ و $Z(\mathbf{s}_j)$ فقط تابعی از موقعیت‌های $\mathbf{s}_i$ و $\mathbf{s}_j$ باشند، یعنی

$$Cov(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)) = C(\mathbf{s}_i - \mathbf{s}_j), \quad \mathbf{s}_i, \mathbf{s}_j \in D \subset \mathbb{R}^d.$$

واضح است که تحت مانایی مرتبه دوم

$$Var(Z(\mathbf{s})) = Cov(Z(\mathbf{s}) - Z(\mathbf{s})) = C(\circ) = \sigma^2.$$

تعریف ۱.۲.۱. میدان تصادفی $Z(\cdot)$، مانای قوی[۱۹] نامیده می‌شود، هرگاه برای همه‌ی موقعیت‌های $\mathbf{s}_1, \ldots, \mathbf{s}_n$ و گام $\mathbf{h}$ توزیع توأم $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ با توزیع توأم $Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_n + \mathbf{h})$ یکی باشد، یعنی

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

به عبارتی، در صورت انتقال موقعیت‌های $\mathbf{s}_1, \ldots, \mathbf{s}_n$ در راستای $\mathbf{h} \in \mathbb{R}^d$ توزیع توام $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ تغییر نمی‌کند.

می‌توان ثابت کرد که مانای قوی شرط کافی برای مانای مرتبه دوم و مانای مرتبه دوم شرط کافی برای مانای ذاتی است، ولی در حالت کلی عکس آن برقرار نیست (محمدزاده، ۱۳۹۸).

۳.۱ ساختار همبستگی فضایی

در آمار فضایی از تغییرنگار[۲۰] و هم‌تغییرنگار[۲۱] برای مدل‌بندی ساختار همبستگی داده‌های فضایی استفاده می‌شود. در اکثر موارد داده‌های نزدیک به هم شباهت بیشتری دارند. پس متوسط تفاضل مقادیر $Z(\mathbf{s})$ و $Z(\mathbf{s}+\mathbf{h})$ که به فاصله‌ی $\mathbf{h}$ از یکدیگر قرار دارند، می‌تواند معیاری مناسب برای بیان اندازه‌ی این تشابه و وابستگی متقابل مقادیر میدان تصادفی در دو موقعیت $\mathbf{s}$ و $\mathbf{s}+\mathbf{h}$ باشد. هر چه متوسط این تفاضل کوچک باشد، واریانس میدان تصادفی کوچک است و می‌توان گفت $Z(\mathbf{s})$ و $Z(\mathbf{s}+\mathbf{h})$ به یکدیگر نزدیکتر و به‌عبارتی شباهت بیش‌تری به هم دارند. به همین صورت هر چه متوسط تفاضل‌ها بیشتر باشد، واریانس مقادیر میدان تصادفی بزرگتر بوده و تشابه آن‌ها کمتر است.

[۱۸] Second order stationary
[۱۹] Strong stationary
[۲۰] Variogram
[۲۱] Covariogram

۱.۳.۱ تغییرنگار

تغییرنگار نظری، $۲\gamma(h)$، واریانس تفاضل بین دو مقدار مشاهده‌شده میدان تصادفی در دو مکان فضایی $\mathbf{s}$ و $\mathbf{s}+\mathbf{h}$ است که به فاصله $\mathbf{h}$ از هم قرار دارند و به‌صورت زیر تعریف می‌شود:

$$۲\gamma(\mathbf{h}) = Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})). \qquad (۴.۱)$$

تابع $\gamma(\mathbf{h})$ نیم‌تغییرنگار[۲۲] نامیده می‌شود. این تابع برای توصیف میزان وابستگی مقادیر میدان تصادفی و برای پیش‌گویی قابل اعتماد توسط پیشگوی فضایی معروف به کریگینگ[۲۳] مورد نیاز است. تغییرنگار تجربی (نمونه) معمولا توسط روش‌های گشتاوری در دنباله‌ای از گام‌ها محاسبه می‌شود و یک یا چند مدل معتبر (مجاز[۲۴]) تغییرنگار بر روی آن برازش داده می‌شود. یک تغییرنگار ممکن است در طول یک برش یا توری در فواصل منظم یا در فواصل پراکنده نامنظم محاسبه شود. برآوردگرهای تغییرنگار ممکن است به‌صورت گشتاوری محاسبه شوند یا این‌که برای محاسباتشان از میانه و مد استفاده شود که برآوردگرهای تنومند[۲۵] نامیده می‌شوند و نسبت به مقادیر دورافتاده[۲۶] حساس نیستند. تغییرنگار ممکن است کراندار (در یک میدان مانای مرتبه دوم) یا بدون کران (در یک میدان مانای ذاتی) باشد. پارامترهای تغییرنگار که توسط مدل‌های معتبر تغییرنگار برآورد می‌شوند، خلاصه‌ای از تغییرات فضایی را که برای پیشگویی فضایی نیاز است به ما می‌دهند. میانگین کمترین توان‌های دوم باقی‌مانده‌ها[۲۷] و معیار آکاییک[۲۸] می‌تواند در انتخاب بهترین مدل برازش داده شده به ما کمک کند (اولیور و وستر، ۲۰۱۵). برآورد تغییرنگار معمولا براساس روش گشتاوری ماترون[۲۹] (MOM) به‌صورت

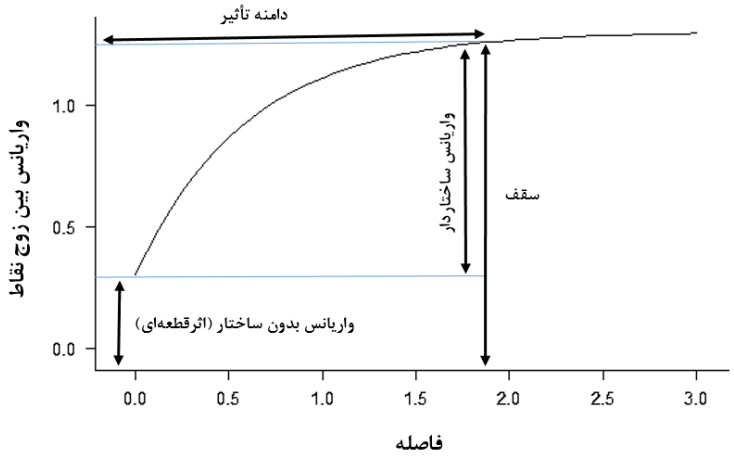
$$\hat{\gamma}(\mathbf{h}) = \frac{۱}{۲m(\mathbf{h})} \sum_{i=۱}^{m(\mathbf{h})} \{z(\mathbf{s}_i) - z(\mathbf{s}_i+\mathbf{h})\}^۲ \qquad (۵.۱)$$

محاسبه می‌شود، که در آن $z(\mathbf{s}_i)$ و $z(\mathbf{s}_i+\mathbf{h})$ مقادیر مشاهده‌شده میدان $Z(\cdot)$ در موقعیت‌های $\mathbf{s}_i$ و $\mathbf{s}_i+\mathbf{h}$ هستند و $m(\mathbf{h})$ تعداد جفت نمونه‌هایی است که در فاصله $\mathbf{h}$ از هم قرار دارند. اگر تغییرنگار برای تمام مقادیر $\mathbf{h}$ تقریباً ثابت باشد، داده‌ها همبستگی فضایی ندارند یا بسیار ضعیف است. برعکس، شیب نمودار تغییرنگار بیانگر همبستگی فضایی داده‌ها است.

از عوامل موثر بر تغییرنگار تجربی می‌توان اندازه نمونه، فاصله‌ی نمونه‌گیری و مقیاس فضایی، فاصله‌ی گام‌ها از هم، توزیع داده‌ها، همسان‌گردی[۳۰] و روند نام برد (اولیور و وستر،

[۲۲] Semivariogram
[۲۳] Kriging
[۲۴] Authorized
[۲۵] Robust estimators
[۲۶] Outliers
[۲۷] Mean squared residuals
[۲۸] Akaike
[۲۹] Matheron's method of moments
[۳۰] Isotropy

۲۰۱۵). تغییرنگار دارای سه پارامتر اثرقطعه‌ای[۳۱] (واریانس بدون ساختار)، سقف[۳۲] (آستانه یا ازاره) و دامنه[۳۳] است که به‌صورت زیر تعریف می‌شوند.



شکل ۱.۱: اجزاء تشکیل دهنده‌ی تغییرنگار

- **دامنه:** در آمار فضایی قاعدتاً مشاهدات نزدیک به هم تشابه بیشتری نسبت به نقاط دور از هم دارند، لذا با افزایش فاصله بین دو مشاهده، مقدار تغییرنگار افزایش می‌یابد. افزایش تغییرنگار تا یک فاصله‌ی معین ادامه دارد و سپس به ازاء مقادیر بزرگتر فاصله مقدار تغییرنگار ثابت می‌ماند. به‌عبارتی نمونه‌ها تا فاصله‌ی معینی بر هم تأثیر می‌گذارند و دارای وابستگی مکانی هستند. فاصله‌ای که در آن تغییرنگار به حد ثابتی می‌رسد و به خط افقی نزدیک می‌شود را دامنه یا شعاع تأثیر می‌نامند. دامنه تاثیر بزرگتر بر وابستگی مکانی گسترده‌تر دلالت دارد. خارج از دامنه‌ی تاثیر مشاهدات بر هم اثری ندارند و با استفاده از آمار کلاسیک می‌توان این داده‌ها را تحلیل کرد.

- **سقف:** به مقدار ثابتی که تغییرنگار در دامنه‌ی تأثیر به آن می‌رسد، سقف یا آستانه گفته می‌شود. مقدار سقف برابر با واریانس تمام نمونه‌هایی است که در محاسبه‌ی تغییرنگار به کار رفته است. تغییرنگارهایی که به سقف مشخص می‌رسند به اصطلاح کراندار هستند و از اهمیت بیشتری در پیشگویی برخوردارند. در مواردی ممکن است که تغییرنگار کراندار نباشد که می‌تواند نشان دهنده‌ی وجود روند یا عدم ناایستایی داده‌ها باشد.

- **اثرقطعه‌ای:** به لحاظ نظری مقدار تغییرنگار در مبداء مختصات، $\mathbf{h} = 0$، باید برابر با صفر باشد. ولی در عمل، تغییرنگارهای واقعی که از تجربه حاصل می‌شوند از چنین دستوری تبعیت نمی‌کنند. به مقدار تغییرنگار به ازاء $\mathbf{h} = 0$ اثرقطعه‌ای می‌گویند که با C_0 نمایش

[۳۱] Nugget effect
[۳۲] Sill
[۳۳] Range

می‌دهند. کرسی (۱۹۹۳) دو خطای نمونه‌گیری و خطای اندازه‌گیری یا تغییرات شدید کمیت را دلایل صفر نبودن اثر قطعه‌ای می‌شمارد و به‌صورت کلی اثرقطعه‌ای را جمع دو خطای اندازه‌گیری[۳۴] (c_{ME}) و خطای ریزمقیاس[۳۵] (c_{MS}) می‌داند. در منابع زمین‌آماری عوامل تاثیرگذار بر مقدار بیشتر از صفر اثرقطعه‌ای را به دو دسته تقسیم کرده‌اند:

۱. وجود مولفه‌های تصادفی در توزیع متغیر، که در واقع به تصادفی بودن متغییرها برمی‌گردد.
۲. خطاهای نمونه‌برداری، آماده‌سازی و تحلیل آزمایشگاهی (حسنی‌پاک، ۱۳۷۷).

تفاضل مقدار سقف و اثرقطعه‌ای را واریانس ساختاردار، C، می‌نامند که تابعی از موقعیت مکانی و جهت قرارگیری داده‌ها است. واریانس ساختاردار تغییراتی است که می‌توان دلیل آن را در خصوصیات خود متغیر فضایی یافت. از نسبت $\frac{C}{C_0}$ می‌توان برای تشخیص وجود ساختار فضایی در داده‌ها استفاده کرد، که برابر است با نسبت مولفه‌ی ساختاردار به مولفه‌ی بدون ساختار تغییرنگار. نسبت دیگری نیز بدین منظور وجود دارد که برابر است با $\frac{C_0}{C+C_0}$ و معرف آن است که چه مقدار از تغییرپذیری را اثرقطعه‌ای توجیه می‌کند. درصورتی‌که $\frac{C}{C+C_0} > \frac{1}{2}$ باشد، نقش مولفه غیرساختارمند از مولفه ساختاردار کمتر است و می‌توان نتیجه گرفت ساختار فضایی داده‌ها قابل ملاحظه است. هر چه مقدار قدرنسبت فوق به یک نزدیکتر باشد استفاده از آمارفضایی ضرورت بیشتری پیدا می‌کند.

۲.۳.۱ هم‌تغییرنگار

کوواریانس دو مقدار $Z(\mathbf{s})$ و $Z(\mathbf{s}+\mathbf{h})$ به‌صورت

$$C(\mathbf{h}) = \mathrm{Cov}(\mathrm{Z}(\mathbf{s}), \mathrm{Z}(\mathbf{s}+\mathbf{h}) \qquad (۶.۱)$$

تعریف می‌شود و آن‌را هم‌تغییرنگار[۳۶] می‌نامند.
اگر میدان تصادفی، مانای مرتبه‌ی دوم باشد آن‌گاه رابطه‌ی (۶.۱) به‌صورت

$$C(\mathbf{h}) = E[Z(\mathbf{s})Z(\mathbf{s}+\mathbf{h})] - \mu^2$$

ساده می‌شود. تابع هم‌تغییرنگار، تابعی همیشه مثبت[۳۷] است (کرسی، ۱۹۹۳). یعنی برای هر دنباله از اعداد حقیقی ثابت $\{a_i\}_{i=1}^m$، $\sum_{i=1}^m \sum_{j=1}^m a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geqslant 0$. هم‌تغییرنگار نمایشگر میزان شباهت تغییرپذیری دو متغیر $Z(\mathbf{s})$ و $Z(\mathbf{s}+\mathbf{h})$ است.

[۳۴] Measurment error
[۳۵] Micro scale error
[۳۶] Covariogram
[۳۷] Positive definite

۳.۳.۱ همبستگی‌نگار

ضریب همبستگی بین $Z(\mathbf{s})$ و $Z(\mathbf{s}+\mathbf{h})$ همبستگی‌نگار[۳۸] نام دارد و با شرایط $C(\circ) > \circ$ به‌صورت

$$\rho(\mathbf{s}, \mathbf{s}+\mathbf{h}) = \rho(\mathbf{h}) = \frac{C(\mathbf{h})}{C(\circ)} = ۱ - \frac{\gamma(\mathbf{h})}{C(\circ)}$$

تعریف می‌شود.

با توجه به این‌که تغییرنگار تابعی از فاصله است، همبستگی‌نگار نیز تابعی از فاصله خواهد بود. بنابراین با افزایش فاصله بین موقعیت‌ها، مقدار همبستگی‌نگار کم و با کاهش فاصله مقدار آن افزایش پیدا می‌کند.

۴.۳.۱ ناهمسان‌گردی

تابع تغییرنگار ممکن است به موقعیت مکانی و به‌طور نسبی به جهت جغرافیایی قرارگیری داده‌ها در ناحیه مورد مطالعه وابسته باشد. اگر این تابع فقط به فاصله مکانی داده‌ها از یکدیگر بستگی داشته باشد و به خود مکان و جهت جغرافیایی وابسته نباشد، میدان تصادفی فضایی حاصل را، که برای مدل‌بندی داده‌ها استفاده می‌شود، همسان‌گرد[۳۹] و در غیر این‌صورت ناهمسان‌گرد[۴۰] می‌نامند (زیمرمن، ۱۹۹۳؛ مونترو و همکاران، ۲۰۱۵). پذیره همسان‌گردی میدان تصادفی موجب آسان‌تر شدن مدل‌بندی داده‌ها، برآورد تغییرنگار و هم‌تغییرنگار و در نهایت افزایش دقت در نتایج حاصل از تحلیل داده‌های فضایی مثل پیش‌گویی می‌شود. بنابراین، بررسی همسان‌گردی ساختار همبستگی داده‌ها قبل از تحلیل آن‌ها یک ضرورت محسوب می‌شود (محمدزاده، ۱۳۹۸). حسنی‌پاک (۱۳۷۷) و سارما (۲۰۱۰) دو نوع ناهمسان‌گردی هندسی[۴۱] و ناهمسان‌گردی ناحیه‌ای[۴۲] را مورد بحث قرار داده‌اند. همچنین زیمرمن (۱۹۹۳) به بررسی ناهمسان‌گردی دامنه‌ای[۴۳] و ناهمسان‌گردی اثر قطعه‌ای[۴۴] پرداخت. علاوه بر این‌ها، اریکسون و سیسکا (۲۰۰۰) ناهمسان‌گردی در شیب و توان را در توابع تغییرنگار مورد بررسی قرار داده‌اند. در این‌جا سعی کرده‌ایم سه نوع ناهمسان‌گردی دامنه‌ای، ازاره‌ای و اثر قطعه‌ای را به اختصار تشریح کنیم.

- **ناهمسان‌گردی هندسی:** مهم‌ترین نوع ناهمسان‌گردی است که به آن ناهمسان‌گردی دامنه‌ای یا ناهمسان‌گردی بیضوی نیز گفته می‌شود (مونترو و همکاران، ۲۰۱۵). اگر

---

[۳۸] Correlogram
[۳۹] Isotropic
[۴۰] Anisotropic
[۴۱] Geometric Anisotropy
[۴۲] Zonal Anisotropy
[۴۳] Range Anisotrapy
[۴۴] Nugget Anisotrapy

تغییرنگار میدان در جهت‌های مختلف دارای ازاره یکسان ولی دامنه تاثیر متفاوتی باشد، ناهمسان‌گردی هندسی رخ داده است.

- **ناهمسان‌گردی اثر قطعه‌ای:** اگر مقدار اثر قطعه‌ای تغییرنگار در جهت‌های مختلف یکسان نباشد، این نوع ناهمسان‌گردی را ناهمسان‌گردی اثر قطعه‌ای می‌نامند.
- **ناهمسان‌گردی ازاره‌ای:** اگر مقدار ازاره (سقف) تغییرنگار در جهت‌های مختلف یکسان نباشد، ناهمسان‌گردی ازاره‌ای رخ داده است. در منابع زمین‌آماری این نوع ناهمسان‌گردی را ناهمسان‌گردی ناحیه‌ای نیز می‌نامند (حسنی‌پاک، ۱۳۷۷).

تعریف ۱.۳.۱. میدان تصادفی‌ای که مانای مرتبه دوم و همسان‌گرد باشد، همگن[۴۵] نامیده می‌شود.

۵.۳.۱ خانواده‌های توابع همبستگی فضایی

این بحث را با این سوال مطرح می‌کنیم که چه ویژگی‌هایی را می‌توان برای خانواده‌های توابع همبستگی فضایی انتظار داشت؟ در ابتدا اگر ما به قانون جغرافیایی توبلر[۴۶] باور داشته باشیم، انتظار داریم که با افزایش فاصله مکانی $\mathbf{s}_i$ و $\mathbf{s}_j$ همبستگی فضایی بین $Z(\mathbf{s}_i)$ و $Z(\mathbf{s}_j)$ کاهش یابد. به‌عبارتی $\rho(\mathbf{h})$ یک تابع غیرافزایشی از $\mathbf{h}$ است. همچنین، با استفاده از واحدهای مختلف می‌توان فاصله را اندازه‌گیری کرد، پس $\rho(\mathbf{h})$ باید دارای یک پارامتر مقیاس‌گذاری واحد باشد (دیگل و جورجی، ۲۰۱۹). با این مقدمه برخی از خانواده‌های پارامتری که این دو ویژگی را دارند، عبارتند از خانواده‌های نمایی، مترن و کروی که در بخش بعدی مورد بررسی قرار گرفته‌اند.

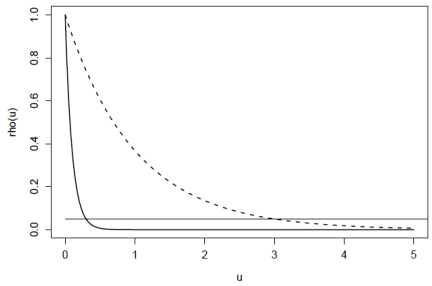
خانواده نمایی

تابع همبستگی معتبر نمایی به‌صورت (۷.۱) تعریف می‌شود:

$$\rho(\mathbf{h},\phi) = \exp(-\mathbf{h}/\phi): \quad \mathbf{h} \geq 0, \tag{۷.۱}$$

که در آن $\phi > 0$ تنها پارامتر مدل است. شکل ۲.۱ مقدار $\rho(\mathbf{h},\phi)$ را به ازاء $\phi = 0/1$ و $\phi = 1/0$ نشان می‌دهد. پارامتر مقیاس‌بندی ϕ را پارامتر دامنه می‌نامند. برای واحد مشخصی از فاصله، مقدار بزرگتری از ϕ نشان می‌دهد که مقدار همبستگی فضایی از دامنه بیشتر است. یک تعریف مرسوم از دامنه عملی[۴۷] در همبستگی فضایی فاصله‌ای است که به ازای آن $\rho(\mathbf{h}) = 0/05$ شود. برای تابع همبستگی نمایی (۷.۱) دامنه عملی تقریبا سه برابر ϕ است.

[۴۵] Homogenuos
[۴۶] Tobler
[۴۷] Pratical range



شکل ۲.۱: تابع همبستگی نمایی به ازاء $\phi = ۰/۱$ (خط ممتد) و $\phi = ۱$ (خط‌چین). محل تقاطع خط افقی باریک $(\rho(\mathbf{h}) = ۰/۰۵)$ با منحنی مقدار دامنه‌ی عملی را نشان می‌دهد.

خانواده مترن

پارامتر دامنه ϕ در مدل نمایی کنترل مرتبه‌ی سرعت میل کردن همبستگی فضایی به سمت صفر را با استفاده از جداسازی ممکن می‌سازد. خانواده مترن این انعطاف‌پذیری را به شکل تابع همبستگی می‌افزاید و به نوبه‌ی خود تاثیر مستقیمی بر همواری فرآیند فضایی دارد. مترن (۱۹۱۷-۲۰۰۷) آماردان سوئدی شاغل در دانشگاه سلطنتی جنگل‌داری استکهلم بود. او در رساله‌ی دکتری خود در سال ۱۹۶۰ نگاهی ژرف به روش‌های توسعه‌یافته آمار فضایی دارد که ماحصل آن را می‌توان معرفی خانواده دو پارامتری از توابع همبستگی دانست که با نام وی به ثبت رسیده است.

تابع همبستگی مترن به‌صورت

$$\rho(\mathbf{h};\phi,k) = \{۲^{k-۱}\Gamma(k)\}^{-۱}(\mathbf{h}/\phi)^k \mathrm{K}_k(\mathbf{h}/\phi)$$

است که در آن $\phi > ۰$ و $k > ۰$ پارامترهای مدل هستند، $\Gamma(\cdot)$ تابع گاما و $K_k(\cdot)$ تابع بسل تعدیل شده[۴۸] نوع سوم از مرتبه‌ی k است. شکل ۳.۱ برای $k = (۰/۵, ۱/۵, ۲/۵)$ مقدار $\rho(\mathbf{h};\phi,k)$ را نشان می‌دهد. پارامتر ϕ طوری انتخاب شده است که دامنه‌ی عملی برابر با سه باشد.

تابع کوواریانس مترن به صورت

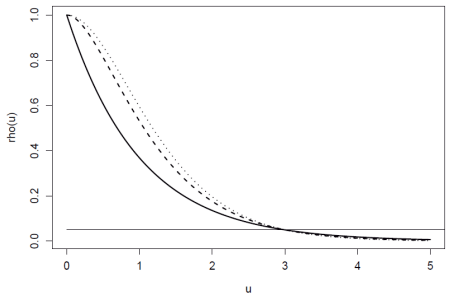
$$\mathrm{Cov}(\mathbf{h}) = \frac{\sigma^۲}{۲^{k-۱}\Gamma(k)}(\phi||\mathbf{h}||)^k \mathrm{K}_k(\phi||\mathbf{h}||) \qquad (۸.۱)$$

که در آن $||\mathbf{h}||$ فاصله‌ی اقلیدسی بین $\mathbf{s}_i$ و $\mathbf{s}_j$ است، به‌طوری که $||\mathbf{h}|| = \mathbf{s}_i - \mathbf{s}_j$، $K_k(\cdot)$ تابع بسل اصلاح‌شده‌ی نوع دوم از مرتبه‌ی k است و k درجه همواری تابع را کنترل می‌کند که اغلب مقدار ثابتی است.

خانواده کروی

تجربه نشان داده است که خانواده مترن به دلیل انعطاف‌پذیری و دامنه‌ی وسیعی که دارد می‌تواند برای بسیاری از مدل‌سازی‌های همبستگی فضایی کافی باشد. ولی در زمین‌شناسی

[۴۸] Modified Bessel function



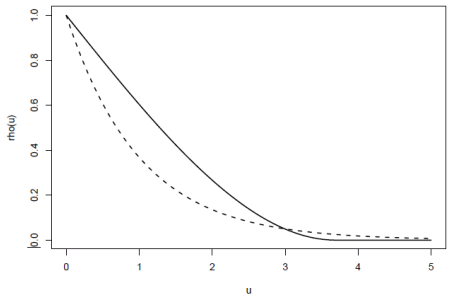
شکل ۳.۱: خانواده همبستگی مترن. برای $k = (۰/۵, ۱/۵, ۲/۵)$ (خط ممتد، خط‌چین و نقطه‌چین به ترتیب) مقدار ϕ طوری انتخاب شده که در هر سه مورد مقدار دامنه‌ی عملی برابر با ۳ شود و این مقدار از تقاطع خط نازک افقی $(\rho = ۰/۰۵)$ با منحنی به‌دست می‌آید.

کلاسیک از مدل‌های دیگری نیز استفاده می‌شود که در اینجا فقط به شرح خانواده کروی که کاربرد بیشتری دارد، می‌پردازیم.

تابع همبستگی کروی از خانواده تک پارامتری است و به شکل

$$\rho(\mathbf{h}) = \begin{cases} ۱ - \frac{۳}{۲}(\mathbf{h}/\phi) + \frac{۱}{۲}(\mathbf{h}/\phi)^۳ & : h \leq \phi \\ ۰ & : h > \phi \end{cases} \qquad (۹.۱)$$

تعریف می‌شود. در شکل ۴.۱ منحنی دو تابع همبستگی کروی و مترن برای پارامتر هموار $k = ۱/۵$ رسم شده‌اند.



شکل ۴.۱: تابع همبستگی کروی (خط ممتد) و مترن (خط‌چین) برای مقدار $k = ۱/۵$. در هر دو مورد مقدار ϕ طوری انتخاب شده است که دامنه‌ی عملی برابر با ۳ شود.

۴.۱ استنباط آماری

استنباط آماری فرآیند نتیجه‌گیری از داده‌ها است. به‌طور کلی مباحث استنباط شامل آزمایش کردن[۴۹]، برآورد[۵۰] و پیش‌گویی[۵۱] است. استنباط و مدل‌بندی آماری رشد چشم‌گیری به‌ویژه برای مقابله با تغییرات و عدم قطعیت‌ها داشته‌اند. این موضوع یک امر بلندپروازانه به نظر می‌رسد، زیرا هر کسی که زندگی را پشت‌سر بگذارد حتی با کمترین تنش‌ها، عدم قطعیت‌های همه‌جانبه زندگی را به‌طور کامل درک می‌کند. بدیهی نیست که در مواجه شدن با عدم قطعیت چیزی سخت‌گیرانه، علمی یا حتی منطقی بگوییم.

مکاتب مختلف فکری رشته آمار در واکنش به عدم قطعیت‌ها پدید آمده‌اند. دو مکتب غالب در دنیای امروز، استنباط مبتنی بر درستنمایی و استنباط بیزی هستند. در دیدگاه بیزی، تمامی عدم قطعیت‌ها می‌توانند از طریق قوانین استاندارد شده احتمال[۵۲] مدل‌سازی و پردازش شوند (پاوتن، ۲۰۰۱). در ادامه به‌طور خلاصه این دو دیدگاه استنباطی را معرفی می‌کنیم.

۱.۴.۱ استنباط مبتنی بر درستنمایی

تعریف ۱.۴.۱. فرض کنید $f(\mathbf{z}|\boldsymbol{\theta})$ تابع چگالی (جرم) احتمال نمونه $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ باشد، به‌طوری که $\mathbf{Z} = \mathbf{z}$ مشاهده شده باشد. تابعی از $\boldsymbol{\theta}$ که به‌صورت $L(\boldsymbol{\theta}|\mathbf{z}) = f(\mathbf{z}|\boldsymbol{\theta})$ باشد را تابع درستنمایی[۵۳] می‌نامند. برای مشاهدات مستقل، تابع درستنمایی برابر است با

$$L(\boldsymbol{\theta}; \mathbf{z}) = \prod_{i=1}^{n} f(\mathbf{z}_i|\theta_1, \ldots \theta_k) \qquad (۱۰.۱)$$

که در آن $\boldsymbol{\theta} = (\theta_1, \ldots \theta_k)^T$ بردار پارامترهای نامعلوم است. در تابع چگالی کمیت $\mathbf{Z}$ متغیر و مقدار $\boldsymbol{\theta}$ ثابت است، اما در تابع درستنمایی مقدار $\boldsymbol{\theta}$ می‌تواند تمامی مقادیر روی پارامتر را اختیار کند و $\mathbf{z}$ یک مقدار مشاهده شده است.

قضیه ۱.۴.۱. فرض کنید $\mathbf{z}$ و $\mathbf{y}$ دو مشاهده از فضای نمونه باشند، به‌طوری‌که تابع درستنمایی $\mathbf{z}$ متناسب با تابع درستنمایی $\mathbf{y}$ باشد، یعنی ثابتی مانند $C(\mathbf{z}, \mathbf{y})$ داشته باشیم به‌طوری‌که برای تمام $\boldsymbol{\theta}$ها داشته باشیم

$$L(\boldsymbol{\theta}|\mathbf{z}) = C(\mathbf{z}, \mathbf{y}) L(\boldsymbol{\theta}|\mathbf{y}).$$

---

[۴۹] Testing
[۵۰] Estimation
[۵۱] Prediction
[۵۲] Standard rules of probability
[۵۳] Likelihood function

در این‌صورت نتایج استنباط برای پارامتر $\boldsymbol{\theta}$ با استفاده از $\mathbf{z}$ مشابه نتایج حاصل با استفاده از $\mathbf{y}$ است. به این قضیه اصل درستنمایی[۵۴] می‌گویند که در آن مقدار ثابت $C(\mathbf{z},\mathbf{y})$ به پارامتر $\boldsymbol{\theta}$ بستگی ندارد.

نتیجه عملی استفاده از اصل درستنمایی، پایه‌ریزی استنباط‌های آماری بر اساس تابع درستنمایی است. یکی از محصولات این نتیجه، محاسبه برآوردگرهای ماکسیمم درستنمایی برای پارامترهای نامعلوم است.

تعریف ۲.۴.۱. فرض کنید $L(\boldsymbol{\theta}|\mathbf{z})$ تابع درستنمایی برای متغیرهای تصادفی با مشاهدات $\mathbf{z}_1,\ldots,\mathbf{z}_n$ باشد. اگر $\hat{\boldsymbol{\theta}}$ مقدار تابع درستنمایی را بیشینه کند، آن‌گاه $\hat{\boldsymbol{\theta}}$ برآوردگر درستنمایی ماکسیمم[۵۵] $\boldsymbol{\theta}$ تعریف می‌شود (مود و همکاران، ۲۰۱۷). برآورد درستنمایی ماکسیمم مقداری از فضای نمونه $\boldsymbol{\theta}$ را اختیار می‌کند که در آن نقطه، $\mathbf{z}$ بیشترین شانس مشاهده شدن را داشته باشد. اگر تابع درستنمایی بر حسب پارامتر مشتق‌پذیر باشد، برآوردگر درستنمایی ماکسیمم در صورت وجود از حل معادله‌ی زیر به‌دست می‌آید:

$$\frac{\partial L(\boldsymbol{\theta}|\mathbf{z})}{\partial \theta_i} = 0, \qquad i = 1,\ldots,k.$$

همچنین $L(\boldsymbol{\theta}|\mathbf{z})$ و $l(\boldsymbol{\theta}|\mathbf{z}) = \ln L(\boldsymbol{\theta}|\mathbf{z})$ در یک نقطه از $\boldsymbol{\theta}$ ماکسیمم می‌شوند و در بعضی موارد، پیدا کردن ماکسیمم لگاریتم درستنمایی ساده‌تر است.

در مدل‌های واقعی ولی پیچده آماری معمولاً محاسبه تابع درستنمایی کار مشکل و از نظر محاسباتی دارای پیچیدگی‌های مضاعفی است. معمولاً این مسأله، آماردانان را به استفاده از دیدگاه‌های استنباطی دیگر سوق می‌دهد.

۲.۴.۱ استنباط بیزی

رویکرد بیزی[۵۶] اساساً با رویکرد کلاسیک و از آن جمله مبتنی بر درستنمایی متفاوت است. برخی جنبه‌های آمار بیزی می‌تواند بر سایر رویکردهای آماری مفید واقع شود. در آمار کلاسیک پارامتر نامعلوم و یک کمیت ثابت در نظر گرفته می‌شود، به‌طوری‌که نمونه‌های تصادفی $\mathbf{Z}_1,\ldots,\mathbf{Z}_n$ از جامعه با شاخص $\boldsymbol{\theta}$ انتخاب شده و براساس مقادیر مشاهده شده در نمونه، اطلاعاتی درباره‌ی پارامتر به ما می‌دهند. در رویکرد بیزی $\boldsymbol{\theta}$ یک کمیت است که می‌تواند دارای توزیع احتمال (که توزیع پیشین[۵۷] نامیده می‌شود) باشد. سپس اطلاعات نمونه برگرفته از جامعه‌ای با شاخص $\boldsymbol{\theta}$ و توزیع پیشین باهم ترکیب شده و به اصطلاح به‌روز می‌شوند. این

[۵۴] Likelihood principle
[۵۵] Maximum likelihood estimation
[۵۶] Bayesian approach
[۵۷] Prior distribution

پیشین به‌روز شده را توزیع پسین[۵۸] می‌نامند که با استفاده از قاعده‌ی بیز[۵۹] انجام می‌شود. از این رو نام آمار بیزی را بر آن نهاده‌اند (کسلا و برگر، ۲۰۰۲). هر چه در توزیع پسین از اطلاعات پیشین پارامتر بیشتر استفاده شود، تاثیر پیشین در توزیع پسین بیشتر شده و به اصطلاح توزیع پیشین آگاهی‌بخش‌تر[۶۰] خواهد بود. با این وجود اگر اطلاعات توزیع پیشین در دسترس نبوده یا مبهم باشد، از توزیع پیشین ناآگاهی‌بخش[۶۱] یا مبهم[۶۲] برای تحلیل بیزی استفاده می‌شود. توزیع‌های پیشین ناآگاهی‌بخش اغلب ناسره[۶۳] هستند، یعنی

$$\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty.$$

از طرفی چون توزیع پیشین ناسره توزیع‌احتمال نیستند، لذا ممکن است توزیع پسین سره را نتیجه ندهد و باید توجه داشت که تنها در صورت سره بودن توزیع پسین استنباط بیزی ممکن خواهد شد.

تعریف ۳.۴.۱. فرض کنید $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ یک نمونه‌ی تصادفی با توزیع معلوم $\pi(\mathbf{z}|\boldsymbol{\theta})$ باشد، به‌طوری که $\boldsymbol{\theta} \in \Theta$ پارامتر مجهول باشد. اگر توزیع پیشین $\pi(\boldsymbol{\theta})$ مشخص باشد، آن‌گاه توزیع پسین، توزیع شرطی $\boldsymbol{\theta}$ به شرط نمونه مشاهده شده $\mathbf{Z} = \mathbf{z}$ است و به‌صورت

$$\pi(\boldsymbol{\theta}|\mathbf{z}) = \frac{\pi(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\mathbf{z})} \qquad (۱۱.۱)$$

تعریف می‌شود، که در آن $m(\mathbf{z})$ توزیع کناری $\mathbf{z}$ و برابر است با

$$m(\mathbf{z}) = \int \pi(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \qquad (۱۲.۱)$$

نکته ۱.۴.۱. توزیع پسین بر روی مشاهدات نمونه شرطی شده است و از اطلاعات و دانسته‌های مربوط به $\boldsymbol{\theta}$ ساخته می‌شود، که به‌صورت یک کمیت و متغیر در نظر گرفته می‌شود. در دیدگاه بیزی تمام استنباط‌ها بر اساس توزیع پسین طراحی و ساخته می‌شوند. برای مثال، میانگین توزیع پسین می‌تواند به‌عنوان برآوردی برای پارامتر $\boldsymbol{\theta}$ مورد استفاده قرار گیرد.

استنباط بیزی تقریبی

اغلب مسائل استنباطی در دیدگاه بیزی به محاسبه انتگرال‌هایی بر اساس توزیع پسین منتهی می‌شوند. فرض کنید محاسبه‌ی امیدریاضی توزیع پسین تابعی از پارامتر، یعنی

$$E[g(\boldsymbol{\theta})|\mathbf{z}] = \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{z})d\boldsymbol{\theta} = \frac{\int_{\boldsymbol{\theta}} g(\boldsymbol{\theta})f(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} f(\mathbf{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \qquad (۱۳.۱)$$

[۵۸] Posterior distribution
[۵۹] Bayes' rule
[۶۰] Informative
[۶۱] Noninformative
[۶۲] Vague
[۶۳] Imporper

مد نظر باشد. به‌طور مثال، برای محاسبه‌ی احتمال پسین یک مجموعه یا آزمون فرضیه، $g(\cdot)$ به‌صورت یک تابع نشان‌گر در نظر گرفته می‌شود. در مدل‌های پیچیده با توزیع‌های پسین پیچیده، معمولاً محاسبه رابطه‌ی (۱۳.۱) دشوار خواهد بود و نمی‌توان آن را به شکل بسته حساب کرد. اگر توزیع پیشین و توزیع پسین هر دو متعلق به یک خانواده باشند، اصطلاحاً این خانواده از توزیع‌ها را خانواده‌ی توزیع‌های مزودج می‌نامند. انتخاب توزیع پیشین از خانواده توزیع‌های مزدوج می‌تواند منجر به محاسبات ساده برای توزیع پسین و رابطه (۱۳.۱) شود. اما در حالت کلی پیچیدگی محاسباتی یکی از مشکلات اساسی دیدگاه بیزی است که برای برخورد با آن از روش‌های تقریب توزیع پسین، مانند روش‌های شبیه‌سازی MCMC[۶۴] و تقریب لاپلاس آشیانه‌ای جمع‌بسته[۶۵] (INLA) استفاده می‌کنند (رو و همکاران، ۲۰۰۹؛ متروپولیس و همکاران، ۱۹۵۳؛ هستینگز، ۱۹۷۰).

۵.۱ پیشگویی فضایی

یکی از اهداف اصلی در تحلیل داده‌های فضایی، پیش‌گویی پاسخ در مکان‌های فاقد مشاهده است که با استفاده از روش کریگینگ انجام می‌شود (کرسی، ۱۹۹۳). اصطلاح کریگینگ به افتخار دانیال کریگ[۶۶] (۲۰۱۳ – ۱۹۱۹) نام‌گذاری شده است. کریگ در استفاده از روش‌های آماری برای ارزیابی و تحلیل مخازن معدنی پیشگام بود و بعدها این روش‌ها توسط ماترون[۶۷] و دیگر همکارانش در دانشکده‌ی معدن پاریس فرمول‌بندی، توسعه و نشر داده شدند (دیگل و جورجی، ۲۰۱۹).

کریگینگ بهترین پیش‌گوی خطی نقطه‌ای و بلوکی است، به‌طوری که بهترین در این‌جا به معنی پیش‌گو با کمترین واریانس است. پیش‌گوی کریگینگ یک میانگین وزنی است که به نقاط نزدیکتر وزن بیشتر و به نقاط دورتر وزن کمتری اختصاص می‌دهد (اولیور و وستر، ۲۰۱۵). یکی از پذیره‌های روش کریگینگ، مانایی میدان تصادفی است. با معلوم بودن ساختار همبستگی فضایی و در نظر گرفتن تابع زیان توان دوم خطا، پیش‌گوی بهینه (کریگینگ) از می‌نیمم کردن میانگین توان دوم خطای پیش‌گو به دست می‌آید. در روش‌های کلاسیک، میدان تصادفی را گوسی در نظر می‌گیرند. بر حسب ساختار فضایی میدان تصادفی، سه نوع کریگینگ تعریف می‌شوند: ۱) کریگینگ ساده[۶۸] که در آن فرض می‌شود میانگین میدان معلوم است؛ ۲) کریگینگ معمولی[۶۹] که در آن فرض می‌شود میانگین ثابت ولی نامعلوم است؛ و ۳)

[۶۴] Markov chain Monte Carlo
[۶۵] Integrated nested Laplace approximation
[۶۶] Daniel G. Krige
[۶۷] Matheron
[۶۸] Simple kriging
[۶۹] Ordinary kriging

کریگینگ عام[۷۰] که در آن میانگین به‌عنوان تابعی از موقعیت‌ها (و سایر متغیرهای تبیینی) در نظر گرفته می‌شود و به آن روند می‌گویند. در ادامه، به‌طور مختصر دو نوع کریگینگ معمولی و عام را معرفی می‌کنیم.

۱.۵.۱ کریگینگ معمولی

فرض کنید مقادیر متغیر پاسخ Z را در n مکان ناحیه تحت مطالعه، $D \subset \mathbb{R}^d$، مثل $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$، مشاهده کرده باشیم و هدف پیش‌گویی مقدار پاسخ در مکان جدید $\mathbf{s}_0$، یعنی $Z(\mathbf{s}_0)$، باشد. در کریگینگ معمولی تغییرات مشاهدات پاسخ از نظر فضایی به هم وابسته، تصادفی و تنها در یک مقدار نامعلوم μ، مشترک هستند. در این حالت می‌توان میدان تصادفی گوسی $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ را به‌صورت زیر تجزیه کرد:

$$Z(\mathbf{s}) = \mu + \delta(\mathbf{s}) \qquad (۱۴.۱)$$

که در آن μ ثابت و نامعلوم و $\delta(\cdot)$ میدان تصادفی مانای گوسی با میانگین صفر است. کریگینگ معمولی، به‌عنوان یک میانگین وزنی از مشاهدات، به‌صورت زیر تعریف می‌شود

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{s}_i)$$

ضرایب λ_i وزن‌های کریگینگ هستند که سهم هر کدام از مشاهدات را (بسته به فاصله آن‌ها تا مکان جدید $\mathbf{s}_0$) در پیش‌گویی پاسخ یدک می‌کشند. چنان‌چه شرط $\sum_{i=1}^{n} \lambda_i = 1$ را در نظر بگیریم، کریگینگ یک پیش‌گوی نااریب خواهد بود، یعنی $E[\hat{Z}(\mathbf{s}_0)] = \mu$. مقادیر بهینه ضرایب $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$ از حل معادله

$$\boldsymbol{\lambda}^T = (\boldsymbol{\gamma} + \mathbf{1}_n \frac{(1 - \mathbf{1}_n^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma})}{\mathbf{1}_n^T \boldsymbol{\Gamma}^{-1} \mathbf{1}_n})^T \boldsymbol{\Gamma}^{-1} \qquad (۱۵.۱)$$

به‌دست می‌آیند، که در آن $\boldsymbol{\Gamma}$ یک ماتریس $n \times n$ با درایه (i, j)

$$\gamma(\mathbf{s}_i - \mathbf{s}_j) = \frac{1}{2} \mathrm{Var}(Z(\mathbf{s}_i) - Z(s_j))$$

و $\boldsymbol{\gamma} = (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \gamma(\mathbf{s}_0 - \mathbf{s}_2), \ldots, \gamma(\mathbf{s}_0 - \mathbf{s}_0))^T$ هستند، به‌طوری که $\gamma(\cdot)$ تابع نیم‌تغییرنگار است. همچنین $\mathbf{1}_n$ برداری به طول n از ۱ها است. در مورد چگونگی رسیدن به این ضرایب و نحوه محاسبه واریانس کریگینگ معمولی می‌توانید به (کرسی، ۱۹۹۳) مراجعه کنید.

کریگینگ ساده حالت خاصی از کریگینگ معمولی است که در آن فرض می‌شود، بین مشاهدات روندی وجود ندارد و میانگین معلوم و مستقل از مختصات فضایی است. در این حالت شرط ناریبی صفر بودن مجموع ضرایب لازم نیست و معادله محاسبه ضرایب کریگینگ، مشابه (۱۵.۱)، با کمی تغییرات جزیی، نتیجه می‌شود.

[۷۰] Universal kriging

۲.۵.۱ کریگینگ عام

کریگینگ عام نسخه تعمیم‌یافته روش کریگینگ معمولی است. در واقعیت، در کاربردهای متعدد، مشاهدات یک میدان تصادفی دارای روند هستند و میانگین در موقعیت‌های مختلف متغیر است و این مهم باید لحاظ شود. کریگینگ عام به‌عنوان یک پیش‌گوی نااریب برای وقتی که روند میدان تصادفی یک ترکیب خطی نامعلوم از توابع معلوم است، استفاده می‌شود. در حالتی که میدان تصادفی $Z(\mathbf{s})$ دارای روند باشد، مدل به‌صورت

$$Z(\mathbf{s}) = \sum_{j=1}^{p+1} \varphi_{j-1}(\mathbf{s})\beta_{j-1} + \delta(\mathbf{s}), \qquad \mathbf{s} \in D \tag{۱۶.۱}$$

در نظر گرفته می‌شود، که در آن $\varphi_j(\mathbf{s})$ را تابع بنیادی[۷۱] می‌نامند که بر حسب ماهیت روند در داده‌ها تعیین می‌شود. بردار ضرایب $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ برداری نامعلوم از پارامترها، $\{\varphi_0(\mathbf{s}), \ldots, \varphi_n(\mathbf{s})\}$ مجموعه‌ای معلوم از توابع و $\delta(\cdot)$ یک میدان تصادفی مانای گوسی با میانگین صفر است. در این حالت، تابع روند $\mu(\mathbf{s})$ چندجمله‌ای

$$\mu(\mathbf{s}) = \sum_{j=1}^{p+1} \varphi_{j-1}(\mathbf{s})\beta_{j-1}, \qquad \mathbf{s} \in D$$

است، با این فرض که $\varphi_0(\mathbf{s}) = 1$، ضرایب $\boldsymbol{\lambda}$ به‌صورت

$$\boldsymbol{\lambda}^T = \{\boldsymbol{\gamma} + \mathbf{X}(\mathbf{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})^{-1}(\mathbf{s} - \mathbf{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})\}^T\boldsymbol{\Gamma}^{-1}$$

به‌دست می‌آیند، که در آن $\boldsymbol{\gamma}$ و $\boldsymbol{\Gamma}$ مشابه قبل تعریف می‌شوند و $\mathbf{X}$ ماتریسی $n \times (p+1)$ با درایه‌های

$$\mathbf{x}_{ij} = \varphi_{j-1}(\mathbf{s}_i), \quad j = 1, \ldots, p+1, \quad i = 1, \ldots, n$$

است (کرسی، ۱۹۹۳).

[۷۱] Basis function

فصل ۲

مدل‌های سه بعدی زمین‌آماری

۱.۲ مقدمه

دنیایی که ما در آن زندگی می‌کنیم مملو از وابستگی‌هاست. در تحقیقات نادیده گرفتن این وابستگی می‌تواند نتایج غلط و زیان‌های گزافی را بر پژوهشگر و کاربران آن عرصه متحمل سازد. در علوم مختلف از جمله نفت، زمین‌شناسی، بوم‌شناسی گیاهی و حیوانی، سنجش از دور[۱] و پزشکی، اغلب داده‌هایی که استفاده می‌شوند، به موقعیت مکانی و جهت قرارگیری آن‌ها بستگی دارد. به‌طور معمول در اکثر موارد هر چه فاصله بین مشاهدات کمتر باشد شباهت بین آن‌ها بیشتر و در فاصله بیشتر شباهت کمتر است. تحلیل این نوع داده‌ها که به داده‌های فضایی شهرت دارند با روش‌های استنباطی آمار کلاسیک ناممکن است. وارد کردن ساختار وابستگی القاشده توسط موقعیت‌های مکانی داده‌ها (ساختار وابستگی فضایی) در تحلیل داده‌ها توسط روش‌های آمار فضایی ممکن شده است. یکی از زیرشاخه‌های آمار فضایی، زمین‌آمار است و به مدل‌ها و روش‌هایی اشاره دارد که از دو مشخصه پیروی می‌کنند. اول اینکه Z_iها، $i = ۱, \ldots, n$، داده‌هایی هستند که در یک مجموعه گسسته مشتمل بر مکان‌های $\mathbf{s}_i$ در برخی از مناطق فضایی مانند D مشاهده شده‌اند. دوم هر مقدار مشاهده‌شده‌ی $\mathbf{Z}$ با اندازه‌گیری مستقیم یا با روابط آماری برآورد شده باشند (دیگل و ریبریو، ۲۰۰۷). داده‌های زمین‌آماری در موقعیت‌های ثابت و در ناحیه‌ای پیوسته مشاهده می‌شوند و متغیر مورد بررسی می‌تواند پیوسته

[۱] Remote Sensing

یا گسسته باشد. در حقیقت زمین‌آمار از ابتدای دهه ۱۹۵۰ به بعد در صنعت معدن توسعه یافت که در حال حاضر به‌طور گسترده‌ای در سایر علوم زیست محیطی برای نقشه‌برداری، نظارت و مدیریت استفاده می‌شود. همچنین با استفاده از مدل‌سازی همبستگی فضایی، زمین‌آمار می‌تواند پیش‌بینی‌های نااریب با کمترین واریانس را ارائه دهد (اولیور و وستر، ۲۰۱۵). روش‌های مدل‌سازی مبتنی بر زمین‌آمار نقش مهمی را در صنعت اکتشاف (معدن و نفت) ایفا می‌کند. چرا که در این صنایع با در نظر گرفتن هزینه‌های زیادی که صرف حفاری می‌شود، شناخت ناحیه مورد مطالعه و پیشگویی سه بعدی آن می‌تواند به کم کردن هزینه‌ها و کاهش خطا کمک کند. پیش‌بینی ویژگی‌های الاستیک ژئومکانیکی مخازن نفتی، به‌طور گسترده، نقش مهمی در بهینه‌سازی و کاهش خطرات و افزایش بهره‌وری حفاری در یک مخزن ایفا می‌کند (نظری استاد و همکاران، ۲۰۱۸). هر چاه دارای طول و عرض جغرافیایی است و از طرفی هزینه‌های حفاری با افزایش عمق به‌صورت نمایی رشد می‌کند (عبدآل و ال‌سهلاوی، ۲۰۱۴). مدل‌سازی سه بعدی متغیرهای تاثیرگذار در حفاری می‌تواند علاوه بر افزایش سرعت و کاهش هزینه، منجر به برنامه‌ریزی بهتر فرآیند حفاری و کاهش مخاطرات طبیعی شود. با این حال، ساخت مدل‌های سه بعدی قابل اعتماد برای این ویژگی‌ها، همچنان با سوال‌ها و گمانه‌هایی همراه است که پاسخ‌گویی به آن‌ها مستلزم تحقیق و توسعه مدل‌های کارا در میدان‌های تصادفی سه بعدی است.

در این فصل، ابتدا مسأله کاربردی مورد نظر پایان‌نامه در مورد مدل‌بندی سه بعدی مقاومت فشاری تک‌محوری سنگ در یکی از مخازن نفتی ایران را تشریح و هدف اصلی را بیان می‌کنیم. سپس تحلیل‌های موجود معمول در این حوزه را ارائه می‌دهیم.

۲.۲ تشریح مسأله

مطالعه رفتارهای سنگ مخزن با توجه به ویژگی‌های فیزیکی و مکانیکی آن، می‌تواند الگویی از تعادل سنگ را نشان دهد. دخالت‌های انسانی مانند حفاری، تولید یا تزریق ممکن است موجب برهم خوردن تعادل موجود در ساختار سنگ شود که نتایج این تغییرات را در حوزه مکانیک سنگ می‌توان مطالعه کرد (زوباک، ۲۰۰۷). استفاده از مکانیک سنگ برای محیط‌های زیرزمینی، ژئومکانیک نامیده می‌شود. اولین گام در هر فرآیند مربوط به ژئومکانیک، ایجاد یک مدل مکانیکی زمین[۲] (MEM) است. در واقع MEM یک نمایش عددی از خواص مکانیکی زمین است. مدل‌سازی سه بعدی ژئومکانیکی به بررسی پارامترهای ژئومکانیکی مخزن در سه بعد می‌پردازد و از این پارامترها در مقیاس حجمی استفاده می‌شود. پیش‌بینی ویژگی‌های الاستیک ژئومکانیکی مخازن نفتی، به‌طور گسترده، نقش مهمی در بهینه‌سازی و کاهش خطرات و افزایش بهره‌وری حفاری در یک مخزن ایفا می‌کند (نظری استاد و همکاران، ۲۰۱۸).

[۲] Mechanical earth model

با این حال، ساخت مدل‌های سه بعدی قابل اعتماد برای این ویژگی‌ها همچنان با سوال‌ها و گمانه‌هایی همراه است که پاسخ‌گویی به آن‌ها مستلزم تحقیق و توسعه مدل‌های کارا است. روش‌های متنوعی برای مدل‌سازی ژئومکانیک وجود دارد که از جمله می‌توان به زمین‌آمار اشاره کرد.

داده‌های استفاده شده در این پژوهش مربوط به مقاومت فشاری تک محوری سنگ[۳] (UCS) است. مقاومت فشاری نهایی سنگ، مقداری است که تنش‌های فشاری تک‌محوری سبب شکست کامل سنگ می‌شود. یکی از پارامترهای مهم در ارزیابی مقاومت سنگ و ساخت مدل ژئومکانیک، مقاومت تک محوری سنگ بکر UCS است. معمولاً مقدار مقاومت فشاری از طریق آزمایش به‌دست می‌آید. از طرفی دسترسی به مغزه حفاری مشکل و هزینه‌بر است، به همین دلیل محققان صنعت نفت روابطی را برای برآورد پارامترهای مکانیکی از روی نگاره‌های پتروفیزیکی ارائه داده‌اند. در این پژوهش برای محاسبه‌ی مقدار مقاومت فشاری تک محوری از رابطه‌ی پیشنهادی توسط شرکت ملی مناطق نفت‌خیز جنوب ایران که برای مخازن نفتی کشور ارائه شده، استفاده کردیم. مقدار UCS به صورت

$$\mathrm{UCS} = ۲/۲۷\mathrm{E}_{sta} + ۴/۷۴$$

محاسبه می‌شود که در آن E_{sta} مدول یانگ استاتیک است و از رابطه‌ی $\mathrm{E}_{sta} = ۰/۷\mathrm{E}_{\mathrm{Dyn}}$ به‌دست می‌آید و $\mathrm{E}_{\mathrm{Dyn}}$ مدول یانگ دینامیک است و از رابطه‌ی

$$\mathrm{E}_{Dyn} = \frac{\rho \mathrm{V}_s^۲(۳\mathrm{V}_p^۲ - ۴\mathrm{V}_s^۲)}{۳\mathrm{V}_p^۲ - ۴\mathrm{V}_s^۲}$$

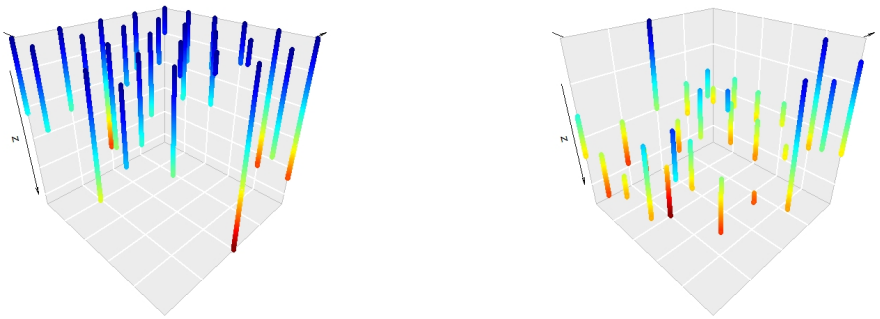
به‌دست می‌آید، که در آن $\mathrm{V_s^2}$ و $\mathrm{V_p^2}$ به ترتیب سرعت موج برشی و فشاری بر حسب $\frac{Km}{sec}$ و ρ دانسیته سنگ بر حسب $\frac{gm}{cc}$ است.

منطقه مورد مطالعه در این پایان‌نامه، یکی از مخازن نفتی ایران شامل ۲۷ حلقه چاه با طول و عرض جغرافیایی و عمق‌های مشخص است که مقدار UCS با استفاده از متغیرهای ثبت‌شده‌ی مرتبط با این متغیر به‌طور تقریبی محاسبه شده است. لازم به ذکر است که چاه‌ها دارای ضخامت‌های متفاوتی هستند. از طرفی این مخزن بر روی یک چین زمین‌شناسی قرار دارد، بنابراین از رابطه‌ی (۱.۲) برای یکسان‌سازی عمق چاه‌ها در همان مختصات طول و عرض جغرافیایی استفاده کردیم.

$$\mathrm{s}_{(new)i} = \mathrm{s}_{(old)i} - |\min(\mathrm{s}_{ij}) - \min(\mathrm{s}_{(old)i})|, \quad j = ۱\ldots۱۰, \quad i = ۱, \ldots, m_j \qquad (۱.۲)$$

که $\mathrm{s}_{(new)i}$ و $\mathrm{s}_{(old)i}$ به ترتیب مختصات عمق حقیقی و اصلاح شده برای مکان iام در هر چاه است، همچنین $\min(\mathrm{s}_{ij})$ کمترین عمق اندازه‌گیری‌شده در تمام چاه‌ها و m_j نشان‌دهنده‌ی تعداد مشاهدات در هر چاه است. در شکل ۱.۲ نمودار سه بعدی چاه‌ها با عمق حقیقی و عمق اصلاحی نمایش داده شده‌اند. از بین چاه‌ها پنج چاه که شامل ۱۸۱۴۰ مشاهده است را

[۳] Unconfined compressive strength



(آ) نمودار سه بعدی چاه‌ها با عمق حقیقی (ب) نمودار سه بعدی چاه‌ها بعد از یکسان‌سازی عمق

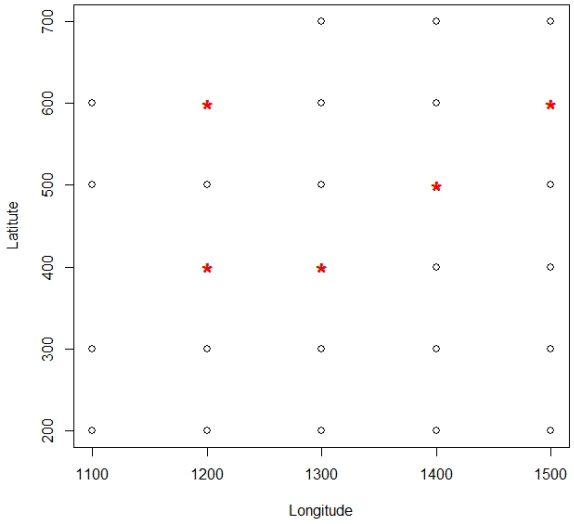
شکل ۱.۲: نمودار سه بعدی چاه‌ها

به‌طور تصادفی برای مجموعه داده‌های آزمون[۴] انتخاب کردیم تا نتایج را بر اساس اعتبارسنجی متقابل[۵] مورد ارزیابی قرار دهیم و از بقیه چاه‌ها که ۵۸۴۶۱ مشاهده را شامل می‌شود، به‌عنوان مجموعه آموزشی[۶] برای مدل‌سازی سه بعدی استفاده کردیم. شکل ۲.۲ مربوط به محل قرارگیری چاه‌ها در موقعیت‌های مکانی طول و عرض جغرافیایی به تفکیک مجموعه‌های آموزشی و آزمون است.

۳.۲ ساختار همبستگی فضایی

همان‌طور که در فصل اول بیان کردیم، در آمار فضایی از تغییرنگار و هم‌تغییرنگار برای بررسی پراکندگی و شباهت داده‌ها استفاده می‌شود و می‌توان گفت که مفاهیمی شبیه به واریانس و کوواریانس در آمار کلاسیک دارند. در زمین‌آمار توصیف نحوه‌ی تغییرپذیری مکانی داده‌ها با استفاده از تابع تغییرنگار را واریوگرافی[۷] می‌نامند که شامل رسم تغییرنگار تجربی، برازش مدل مناسب به آن و تحلیل تغییرنگار مشخص‌شده برای ساختار تغییرات داده‌ها است. اهمیت واریوگرافی به این دلیل است که تمامی تحلیل‌ها و استنباط‌های آماری از جمله پیشگویی، با تکیه بر پارامترهای حاصل از این عملیات انجام می‌شوند.

[۴] Test set
[۵] Cross validation
[۶] Train set
[۷] Variography



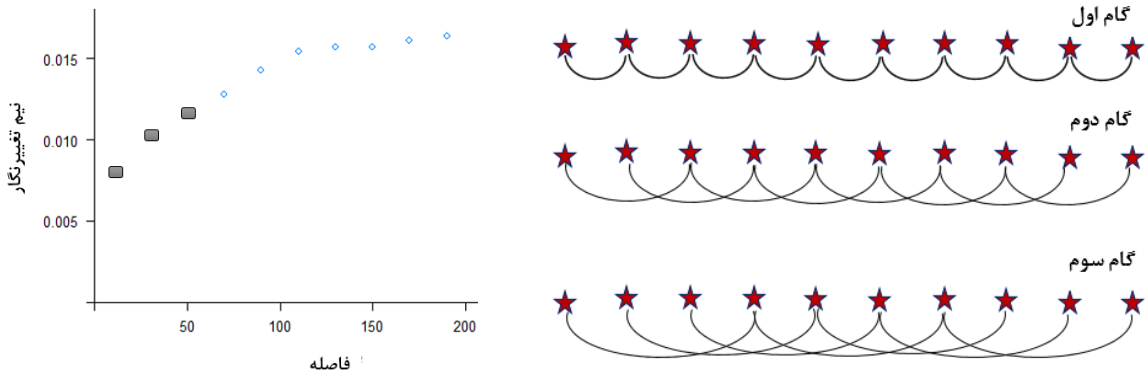
شکل ۲.۲: محل قرارگیری چاه‌های مجموعه آموزشی (توخالی) و مجموعه آزمون (ستاره‌ای) بر حسب طول و عرض جغرافیایی و محل قرارگیری آن‌ها

۱.۳.۲ برآورد تغییرنگار مخزن نفتی

برآورد دقیق تغییرنگار برای پیش‌گویی قابل اعتماد توسط کریگینگ مورد نیاز است. قبلاً ذکر شده که تغییرنگار تجربی (نمونه) معمولاً توسط روش‌های گشتاوری در دنباله‌ای از گام‌ها محاسبه می‌شود و یک یا چند مدل معتبر تغییرنگار بر روی آن برازش داده می‌شود. با توجه به ماهیت سه بعدی میدان تصادفی مخزن نفتی و اهمیت برآورد ساختار وابستگی مقاومت فشاری تک‌محوری توسط تغییرنگار، روند برآورد تغییرنگار را از فضای یک بعدی تا سه بعدی تشریح می‌کنیم.

برآورد تغییرنگار یک‌بعدی

نمونه‌گیری در یک بعد ممکن است به‌صورت برش‌های عمودی یا افقی باشد. گام $\mathbf{h} = h$ به ازاء $|\mathbf{h}|$ مقادیر عددی را ارائه می‌دهد و جایگزین مقدار $\mathbf{h}$ در معادله‌ی (۵.۱) می‌شود. نیم‌تغییرنگار، $\hat{\gamma}(\mathbf{h})$، می‌تواند فقط به‌صورت چندگانه در بازه‌ی نمونه محاسبه شود. شکل ۳.۲ (آ) به‌صورت مقایسه‌ای، سه گام $\mathbf{h} = ۱, ۲, ۳$ بین جفت نقاط را نشان می‌دهد. نتایج به‌صورت مجموعه‌ای از نیم تغییرنگارهای $\hat{\gamma}(۱), \hat{\gamma}(۲), \hat{\gamma}(۳), \cdots$ است. نمودار نیم‌تغییرنگار یک‌بعدی تجربی ده گام اول در شکل ۳.۲ (ب) آورده شده است.



(آ) نمایش سه گام $\mathbf{h} = ۱,۲,۳$ بین جفت نقاط (ب) نمودار نیم‌تغییرنگار تجربی برای ۱۰ گام اول

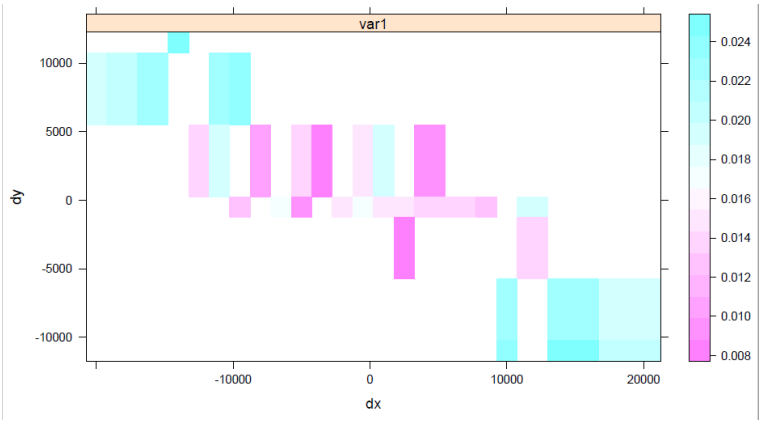
شکل ۳.۲: تغییرنگار یک بعدی

برآورد تغییرنگار دو بعدی

برآورد تغییرنگار در دو بعد طی سه مرحله انجام می‌شود. ابتدا باید ناحیه را به شبکه‌های دو بعدی تقسیم کنیم. این کار می‌تواند موجب تشخیص همسانگردی میدان با استفاده از چرخش در جهت‌های مختلف تغییرات شود. می‌توان تغییرنگار را در شبکه‌ها در جهت‌های مختلف محاسبه کرد. به‌عنوان مثال در جهت طول، عرض یا در جهت قطر. دوم، تغییرنگار دو بعدی به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned}\hat{\gamma}(p,q) &= \frac{۱}{۲(m-p)(n-q)} \sum_{i=۱}^{m-p}\sum_{j=۱}^{n-q}\{z(i,j)-z(i+p,j+q)\}^۲ \\ \hat{\gamma}(p,-q) &= \frac{۱}{۲(m-p)(n-q)} \sum_{i=۱}^{m-p}\sum_{j=q+۱}^{n}\{z(i,j)-z(i+p,j-q)\}^۲.\end{aligned} \tag{۲.۲}$$

در این‌جا p و q به ترتیب تعداد گام‌های سطر و ستون در طول توری هستند. به‌طور کلی شبکه برابر با فاصله‌ی گام است. محاسبات تغییرنگار از q تا $-q$ و از ۰ تا p است. نقشه‌ی دو بعدی تغییرنگار در شکل ۴.۲ نشان داده شده است. سوم، تغییرنگار می‌تواند در تمام جهت‌ها برای شبکه‌های منظم یا نامنظم محاسبه شود.



شکل ۴.۲: اجزاء تشکیل دهنده‌ی تغییرنگار

برآورد تغییرنگارسه بعدی

محاسبه تغییرنگار سه بعدی در میدان با توری‌های حجمی منظم یا نامنظم به این شکل است که می‌توان توری را به‌صورت یک سری در نظر گرفت و در سه بعد طول، عرض و عمق مورد بررسی قرار داد. معادله‌ی (۳.۲) مربوط به تغییرنگار سه بعدی است که برای بررسی همسانگردی باید در چند جهت مختلف از طول، عرض و عمق یک توری محاسبه شود.

$$\hat{\gamma}(p,q,r) = \frac{۱}{۲(m-p)(n-q)(l-r)} \sum_{i=۱}^{m-p} \sum_{j=۱}^{n-q} \sum_{k=۱}^{l-r} \{z(i,j,k) - z(i+p, j+q, k+r)\}^{۲} \quad (۳.۲)$$

که در آن p، q و r به ترتیب اندازه گام‌های طول، عرض و عمق و m، n و l تعداد گام‌های طول، عرض و عمق هستند.

۲.۳.۲ برآوردگر استوار

برآوردگر گشتاوری که در معادله‌ی (۵.۱) آورده شده است وقتی داده‌ها فاقد روند باشند، ناریب است، یعنی $E[۲\hat{\gamma}] = ۲\gamma$. اما این برآوردگر استوار[۸] نیست و دارای خواص پایداری ضعیفی است. چرا که برای فاصله‌های بزرگ از داده‌های کمتری استفاده می‌کند و نیز نسبت به نقاط پرت و دورافتاده بسیار تاثیرپذیر است. کرسی و هاوکینز (۱۹۸۰) دو برآورگر استوار را معرفی کردند که از معادلات

$$\bar{\gamma}(h) = \frac{\{\frac{۱}{N}\sum_{N(h)} |Z(s_i) - Z(s_j)|^{\frac{۱}{۲}}\}^{۴}}{۰/۹۱۴ + \frac{۰/۹۸۸}{N(h)}} \quad (۴.۲)$$

و

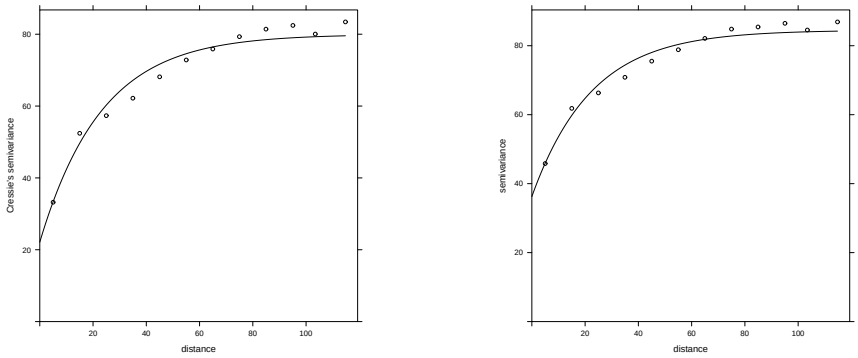
$$\tilde{\gamma}(h) = \frac{(Med\{|Z(s_i) - Z(s_j)|^{\frac{۱}{۲}} \ (s_i, s_j) \in N(h)\})^{۴}}{۰/۹۱۴} \quad (۵.۲)$$

[۸] Robust

پیروی می‌کند، که در آن $Med(\cdot)$ مقدار میانه در مجموعه $N(h)$ است.

نتایج برآورد تغییرنگار سه بعدی برای میدان نفتی

برازش مدل مناسب به تغییرنگار امر مهمی است که در پیشگویی تاثیر بسزایی دارد. در این‌جا از دو روش رسم نمودار (روش چشمی) و مقایسه مجموع توان‌های دوم خطا[۹] (SSErr) برای انتخاب مدل مناسب تغییرنگار تجربی استفاده کردیم. بعد از رسم نمودار نیم‌تغییرنگار تجربی کلاسیک منحنی مدل‌هایی که به نظر مناسب بودند به نمودار نیم‌تغییرنگار برازش دادیم. سپس مدلی که دارای کمترین مقدار SSErr بود را انتخاب کردیم. در این‌جا مدل پارامتری نمایی، SSErr کمتری نسبت به سایر مدل‌ها داشت. برای اطمینان بیشتر از برآوردگر استوار هم استفاده کردیم، که مقدار SSErr بیشتری نسبت به برآوردگر کلاسیک داشت. نمودار ۵.۲ منحنی برآوردگر کلاسیک و استوار مدل پارامتری نمایی برازش داده شده به نیم‌تغییرنگار را نشان داده‌است. مقدار ازاره، دامنه و اثر قطعه‌ای به ترتیب برابر با ۸۴/۵، ۶۷/۲۳ و ۳۶/۲۷ است.



(ب) (آ)

شکل ۵.۲: نمودار سه بعدی نیم‌تغییرنگار تجربی منحنی مدل پارامتری نمایی برازش داده شده (آ) برآوردگر کلاسیک، (ب) برآوردگر استوار

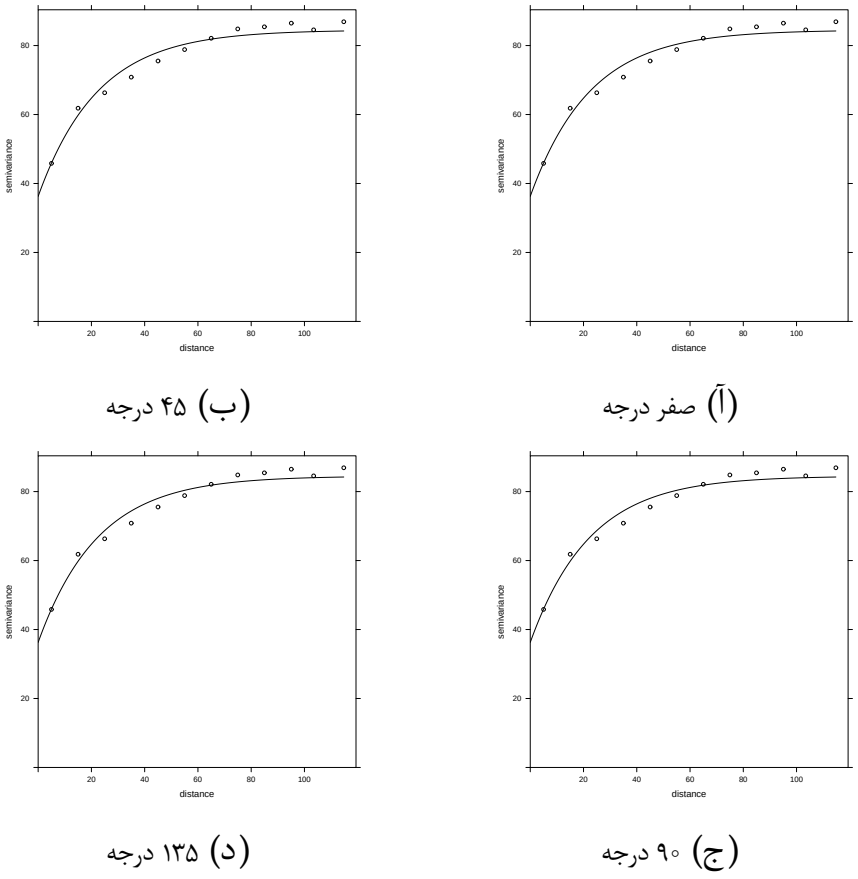
۳.۳.۲ بررسی ناهمسانگردی میدان نفتی

با توجه به اهمیت همسانگرد بودن میدان و تأثیر آن بر برآورد درست ساختار وابستگی، وجود ناهمسانگردی در میدان نفتی را بررسی می‌کنیم.

نمودار سه بعدی نیم‌تغییرنگار تجربی داده‌های مقاومت فشاری در چهار جهت ۰، ۴۵، ۹۰ و ۱۳۵ درجه برای بررسی ناهمسانگردی همراه با منحنی مدل پارامتری نمایی برازش داده‌شده

[۹] Sum of squers error

در شکل ۶.۲ آورده شده است. با توجه به شکل می‌توان دریافت که نیم‌تغییرنگار در این چهار جهت مشابه و در نتیجه میدان همسانگرد است.



(آ) صفر درجه (ب) ۴۵ درجه

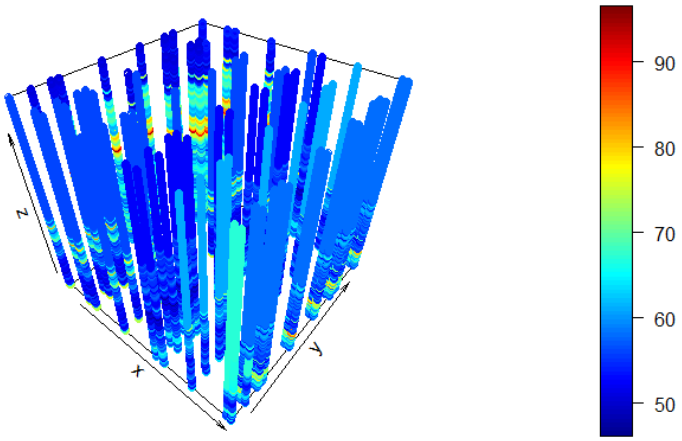
(ج) ۹۰ درجه (د) ۱۳۵ درجه

شکل ۶.۲: نمودار سه بعدی نیم‌تغییرنگار و منحنی مدل پارامتری نمایی برازش داده شده در چهار جهت

۴.۲ پیشگویی سه بعدی مقاومت فشاری تک محوری سازند

برای پیش‌گویی داده‌های مقاومت فشاری تک محوری سازند، با توجه به پذیره گوسی بودن میدان تصادفی (۱۴.۱)، ابتدا ناحیه تحت مطالعه را مشبکه‌بندی کردیم. به این صورت که طول و عرض میدان نفتی را به ۶ قسمت مساوی تقسیم کردیم. سپس عمق‌ها را از کمترین مقدار تا بیشترین مقدارش به برش‌های ۱۵ سانتی‌متری افراز کردیم. از ضرب طول، عرض و عمق یک مشبکه که به مکعب‌های مستطیلی کوچکی به طول ۱۷۴، عرض ۸۳ و ارتفاع ۰/۱۵ متر افراز شده است به‌دست می‌آید. مشبکه ۹۵۰۰۰ گره را شامل می‌شود. سپس مختصات جغرافیایی ۵ چاهی که به‌عنوان داده‌های آزمون در نظرگرفتیم را به این مشبکه برای پیشگویی

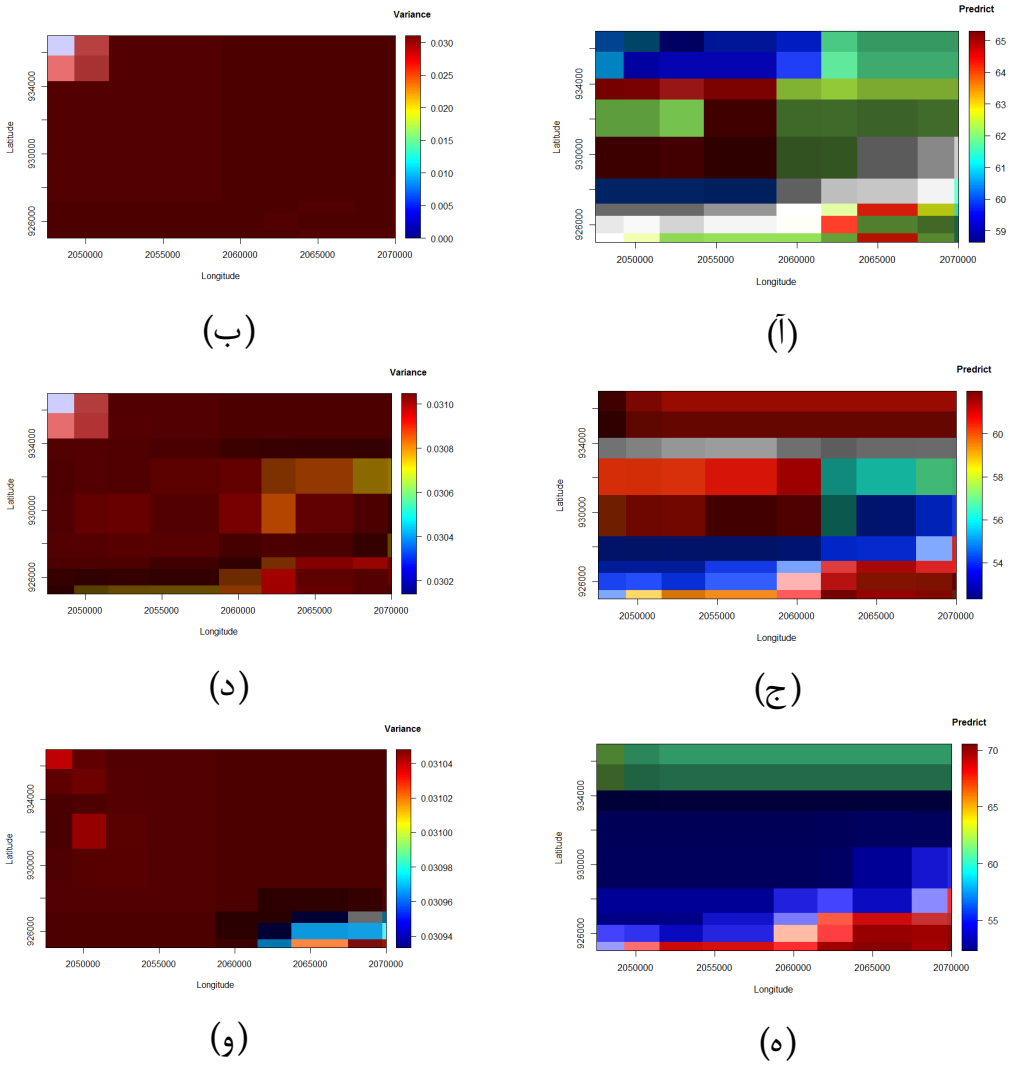
اضافه کردیم. در شکل ۷.۲ با استفاده از کریگینگ معمولی و مشبکه‌بندی ناحیه تحت مطالعه،



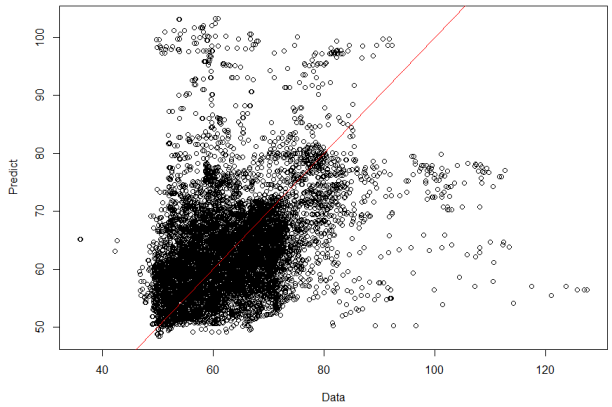
شکل ۷.۲: پیش‌گویی سه بعدی میدان نفتی تحت مطالعه

نقشه پهنه‌بندی مقادیر پیش‌گویی مقدار مقاومت فشاری را به‌دست آوردیم. شکل ۸.۲ نمودار دو بعدی مقادیر پیش‌گویی و واریانس پیش‌گویی را در سه عمق مختلف نشان می‌دهد.

محور افقی شکل ۹.۲ نشان دهنده‌ی مقادیر واقعی مجموعه داده‌های آزمون Z_i و محور عمودی مقادیر پیشگویی به دست‌آمده از کریگینگ معمولی $\hat{Z}_i$ برای مجموعه‌ی آزمون است. برای UCS کمتر از ۸۰ توزیع پراکنش داده‌ها تقریباً حول نیم‌ساز ربع اول و سوم $(Z_i = \hat{Z}_i)$ است و بیانگر مناسب بودن مدل و مقدار ناچیز باقی‌مانده‌ها است. با افزایش مقادیر داده‌ها، پیشگو و مقادیر اصلی از نیم‌ساز فاصله گرفته و اندازه‌ی خطا بیشتر شده است. به‌عبارتی واریانس خطا ثابت نیست ولی همچنان نیم‌ساز بین داده‌ها قرار دارد و عملکرد روش کریگینگ سه بعدی مناسب است.



شکل ۸.۲: نقشه پهنه بندی پیشگویی مقادیر مقاومت فشاری در عمق‌های ۱۷۰۳، ۲۱۵۵ و ۲۴۸۰ متر به ترتیب در (آ)، (ج) و (ه). همچنین نقشه پهنه‌بندی واریانس پیشگویی در همین عمق‌ها به ترتیب در (ب)، (د) و (و).



شکل ۹.۲: مقادیر پیشگویی در مقابل مقادیر واقعی داده‌های آزمون

فصل ۳

روش‌های ناپارامتری فضایی در میدان‌های تصادفی سه بعدی

در این فصل کریگینگ میانه پالایش سه بعدی را توضیح می‌دهیم. سپس از دو مثال شبیه‌سازی شده برای پیش‌گویی با استفاده از کریگینگ عام که در آن از مدل‌بندی اسپلاین استفاده شده و کریگینگ میانه‌پالایش بهره‌می‌بریم. در نهایت به مقایسه‌ی روش‌های مذکور می‌پردازیم.

۱.۳ میانه پالایش

توکی (۱۹۷۰) و ولمن و هوگلین (۱۹۸۱) میانه پالایش دو طرفه [۱] را روشی برای استخراج اثرات ستونی و سطری با به‌کارگیری میانه به‌جای میانگین معرفی کردند. میانه پالایش کاربردهای زیادی دارد. برای نمونه در آمار فضایی، کرسی در سال ۱۹۹۳ با استفاده از شبکه‌بندی داده‌ها از این روش به‌عنوان یک روش تنومند برای حذف هرگونه روند در دامنه فضایی استفاده کرد. میانه پالایش معمولاً برازشی را برای یک مدل جمعی ارائه می‌دهد که از دید کمترین قدرمطلق باقی‌مانده‌ها تقریباً بهینه است (موستلر و توکی، ۱۹۷۷). در این پایان‌نامه، برای کاربردهای سه بعدی در زمین‌شناسی و نفت، مشابه الگوریتم ارائه شده توسط کرسی (۱۹۹۳)، برای استخراج اثرات طول، عرض و بعد سوم (عمق یا ارتفاع) در تابع روند میدان تصادفی برای اولین بار یک

[۱] Two-way

الگوریتم میانه‌پالایش سه بعدی را پیشنهاد و معرفی می‌کنیم و کارایی آن را با دو روش دیگر در کیفیت پیشگویی فضایی مقایسه می‌کنیم.

۱.۱.۳ الگوریتم میانه پالایش سه بعدی

در این بخش، روش کار را به‌صورت جبری توضیح می‌دهیم. برای درک بیشتر شماره‌ی هر مرحله از تکرار میانه پالایش را بالای نماد و داخل پرانتز قرار می‌دهیم. پس $R_a^{(n)}$ ، $C_b^{(n)}$ و $L_c^{(n)}$ نشان‌دهنده‌ی برآورد اثرات میانه در سطوح مختلف به ترتیب از عوامل سطر، ستون و لایه هستند که در مرحله‌ی n تأثیرشان بر روی داده‌ها حذف شده است. مقدار اولیه میانه را برای هر سه عامل برابر صفر قرار می‌دهیم. یعنی برای تمام a، b و cها

$$R_a^{(۰)} = C_b^{(۰)} = L_c^{(۰)} = ۰$$

اگر $r_{abc}^{(n)}$ باقی‌مانده‌ها پس از حذف اثرات سطری، ستونی و لایه‌ای در مرحله‌ی n باشند، پس می‌توانیم از $r_{abc}^{(۰)} = Z(\mathbf{s}_{abc})$ شروع کنیم که $Z(\mathbf{s}_{abc})$ همان داده‌های اصلی در موقعیت‌های $\mathbf{s}_{abc}$ هستند. اکنون جزئیات مراحل میانه پالایش را مورد بررسی قرار می‌دهیم. میانه‌ها را باید برای هر عامل به ترتیب حذف کنیم. گاهی نتایج به ترتیب انتخاب عامل‌ها بستگی دارد.

وقتی i $(i = ۱, \ldots, n)$ را شمارنده‌ی هر تکرار در نظر می‌گیریم، هر مرحله از تکرار میانه‌پالایش از دستورات زیر تبعیت می‌کند: ابتدا میانه سطری را در هر سطح محاسبه می‌کنیم، یعنی

$$R_a^{(i)} = \underset{bc}{\text{med}}\{r_{abc}^{(i-۱)}\} \qquad a = ۱, \ldots, U$$

که در آن U تعداد کل سطرهاست. میانه را روی b و c حساب می‌کنیم و a را ثابت در نظر می‌گیریم. سپس میانه‌ها را از $r_{abc}^{(i-۱)}$ کم می‌کنیم و میانه ستونی را حساب می‌کنیم.

$$C_b^{(i)} = \underset{ac}{\text{med}}\{r_{abc}^{(i-۱)} - R_a^{(i)}\} \qquad b = ۱, \ldots, V$$

که در آن V تعداد کل ستون‌هاست. در نهایت مقدار $C_b^{(i)}$ را کم می‌کنیم و میانه را برای هر مرحله از عامل لایه به تعداد کل W، محاسبه می‌کنیم. داریم

$$L_c^{(i)} = \underset{ab}{\text{med}}\{r_{abc}^{(i-۱)} - R_a^{(i)} - C_b^{(i)}\} \qquad c = ۱, \ldots, W$$

با کم کردن مقدار $L_c^{(i)}$ باقی‌مانده‌ها در مرحله اول به‌صورت زیر به‌دست می‌آیند:

$$r_{abc}^{(i)} = r_{abc}^{(i-۱)} - R_a^{(i)} - C_b^{(i)} - L_c^{(i)}.$$

اگر بخواهیم برآورد اثرات را در این مرحله مرکزی کنیم، میانه را برای هر مجموعه از اثر حساب

می‌کنیم. سپس برآورد اثرات به‌صورت زیر محاسبه می‌شوند:

$$\hat{R}_a = \sum_{i=1}^{n} R_n^{(i)} - \underset{a}{\text{med}}\{\sum_{i=1}^{n} R_a^{(i)}\}, \quad a = 1, \ldots, U$$
$$\hat{C}_b = \sum_{i=1}^{n} R_n^{(i)} - \underset{b}{\text{med}}\{\sum_{i=1}^{n} C_b^{(i)}\}, \quad b = 1, \ldots, V$$
$$\hat{L}_c = \sum_{i=1}^{n} R_n^{(i)} - \underset{c}{\text{med}}\{\sum_{i=1}^{n} L_c^{(i)}\}, \quad c = 1, \ldots, W$$
$$\hat{\mu} = \underset{a}{\text{med}}\{\sum_{i=1}^{n} R_a^{(i)}\} + \underset{b}{\text{med}}\{\sum_{i=1}^{n} C_b^{(i)}\} + \underset{c}{\text{med}}\{\sum_{i=1}^{n} L_c^{(i)}\}.$$

در این‌جا $\hat{\mu}$ برآورد میانه کل است. برآورد نهایی به ترتیب انتخاب عامل‌ها برای حذف اثر آن‌ها بستگی دارد. اما در کاربرد معمولاً این موضوع از اهمیت زیادی برخوردار نیست. در عمل برای روش میانه پالایش تکرارهای ثابتی در نظر گرفته می‌شود (حداقل ۲ یا ۳ تکرار) و از آن به‌عنوان تحلیل استاندارد استفاده می‌شود. در روش میانه پالایش سه طرفه[۲] محاسبات برای رسیدن به هدف شرط جانبی یا حذف کامل اثرات هدایت می‌شوند. یعنی برای رسیدن به وضعیتی که

$$\begin{aligned} &\underset{bc}{\text{med}}\{r_{abc}\} = 0, \ \text{برای هر } a && \underset{a}{\text{med}}\{\hat{R}_a\} = 0 \\ &\underset{ac}{\text{med}}\{r_{abc}\} = 0, \ \text{برای هر } b && \underset{b}{\text{med}}\{\hat{C}_b\} = 0 \\ &\underset{ab}{\text{med}}\{r_{abc}\} = 0, \ \text{برای هر } c && \underset{c}{\text{med}}\{\hat{L}_c\} = 0 \end{aligned} \qquad (۱.۳)$$

روابط (۱.۳) نشان می‌دهند که اثرات سطری، ستونی و لایه‌ای در داده‌ها به‌طور کامل حذف شده‌اند.

مثال ۱.۱.۳. برای درک بیشتر الگوریتم میانه‌پالایش سه طرفه، از یک مثال ساده با داده‌های مصنوعی، استفاده می‌کنیم. ماتریس سه بعدی ($۵ \times ۲ \times ۳$) در (۲.۳) دارای ۳ سطر، ۲ ستون

[۲] Three-way

و ۵ لایه است.

$$
\begin{array}{c} \\ b_۱ \\ \\ \\ \\ \\ b_۲ \\ \\ \\ \\ \end{array}
\begin{array}{c}
\begin{array}{ccc} a_۱ & a_۲ & a_۳ \end{array} \\
\left(\begin{array}{ccc}
\begin{array}{cc} c_۱ & \begin{bmatrix} ۳ \end{bmatrix} \\ c_۲ & \begin{bmatrix} ۱ \end{bmatrix} \\ c_۳ & \begin{bmatrix} ۶ \end{bmatrix} \\ c_۴ & \begin{bmatrix} ۱۲ \end{bmatrix} \\ c_۵ & \begin{bmatrix} ۹ \end{bmatrix} \end{array} &
\begin{bmatrix} -۷ \\ -۵ \\ ۲ \\ ۰ \\ ۱ \end{bmatrix} &
\begin{bmatrix} ۵ \\ ۸ \\ ۲ \\ -۱ \\ ۰ \end{bmatrix} \\
\begin{array}{cc} c_۱ & \begin{bmatrix} ۰ \end{bmatrix} \\ c_۲ & \begin{bmatrix} -۵ \end{bmatrix} \\ c_۳ & \begin{bmatrix} ۴ \end{bmatrix} \\ c_۴ & \begin{bmatrix} ۱۲ \end{bmatrix} \\ c_۵ & \begin{bmatrix} ۱ \end{bmatrix} \end{array} &
\begin{bmatrix} ۳ \\ -۱ \\ -۱۰ \\ ۶ \\ -۸ \end{bmatrix} &
\begin{bmatrix} -۴ \\ -۶ \\ ۳ \\ ۱۱ \\ ۷ \end{bmatrix}
\end{array}\right)
\end{array}
\tag{۲.۳}
$$

با انتخاب سطر به‌عنوان اولین عامل تاثیرگذار بر روی داده‌ها، ابتدا میانه‌ی سطری را با ثابت در نظر گرفتن سطوح آن، روی سطوح ستون و لایه محاسبه می‌کنیم. میانه سطری متغیر[۳] را با نماد CRE و اثر سطری را با RE نشان می‌دهیم. مقدار اولیه اثرات سطری برای تمام سطوح را برابر با صفر قرار می‌دهیم.

$$
\begin{aligned}
\mathrm{CRE}_۱^{(۱)} &= \mathrm{median}(\{۳,۱,۶,۱۲,۹\},\{-۷,-۵,۲,۰,۱\},\{۵,۸,۲,-۱,۰\}) = ۲ \\
\mathrm{CRE}_۲^{(۱)} &= \mathrm{median}(\{۰,-۵,۴,۱۲,۱\},\{۳,-۱,-۱۰,۶,-۸\},\{-۴,-۶,۳,۱۱,۷\}) = ۱ \\
\mathrm{RE}_۱^{(۰)} &= \mathrm{RE}_۲^{(۰)} = ۰ \\
\mathrm{RE}_۱^{(۱)} &= \mathrm{CRE}_۱^{(۱)} - \mathrm{RE}_۱^{(۰)} = ۲ - ۰ = ۲ \\
\mathrm{RE}_۲^{(۱)} &= \mathrm{CRE}_۲^{(۱)} - \mathrm{RE}_۲^{(۰)} = ۱ - ۰ = ۱
\end{aligned}
$$

میانه اثر سطری مرحله اول که با $\mathrm{RE}^{(۱)}$ نمایش می‌دهیم، به‌صورت زیر حساب می‌شود:

$$
\mathrm{RE}^{(۱)} = \mathrm{median}(\mathrm{RE}_۱^{(۱)}, \mathrm{RE}_۱^{(۱)}) = \mathrm{median}(۲,۱) = ۱/۵
$$

سپس اثرات سطری را از داده‌های مربوط به همان سطر کم می‌کنیم که ماتریس (۳.۳) نتیجه می‌شود.

[۳] Change row effect

$$
\begin{array}{c} \\ b_1 \\ \\ \\ \\ \\ \\ b_2 \\ \\ \\ \\ \\ \end{array}
\begin{array}{c}
\begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\
\left(\begin{array}{ccc}
\begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array} &
\begin{bmatrix} ۱ \\ -۱ \\ ۴ \\ ۱۰ \\ ۷ \end{bmatrix} &
\begin{bmatrix} -۹ \\ -۷ \\ ۰ \\ -۲ \\ -۱ \end{bmatrix} &
\begin{bmatrix} ۳ \\ ۶ \\ ۰ \\ -۳ \\ -۲ \end{bmatrix} \\
\begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array} &
\begin{bmatrix} -۱ \\ -۶ \\ ۳ \\ ۱۱ \\ ۰ \end{bmatrix} &
\begin{bmatrix} ۲ \\ -۲ \\ -۱۱ \\ ۵ \\ -۹ \end{bmatrix} &
\begin{bmatrix} -۵ \\ -۷ \\ ۲ \\ ۱۰ \\ ۶ \end{bmatrix}
\end{array}\right)
\end{array}
\tag{۳.۳}
$$

برای حذف اثر ستون مشابه سطر رفتار می‌کنیم، مقدار اولیه اثرات ستون[۴] (CE)، برای تمام سطوح برابر صفر است. اثر متغیر ستون[۵] را با CCE نمایش می‌دهیم.

$$
\begin{aligned}
\mathrm{CCE}_1^{(1)} &= \mathrm{median}(\{۱,-۱,۴,۱۰,۷\},\{-۱,-۶,۳,۱۱,۰\}) = ۲ \\
\mathrm{CCE}_2^{(1)} &= \mathrm{median}(\{-۹,-۷,۰,-۲,-۱\},\{۲,-۲,-۱۱,۵,-۹\}) = -۲ \\
\mathrm{CCE}_3^{(1)} &= \mathrm{median}(\{۳,۶,۰,-۳,-۲\},\{-۵,-۷,۲,۱۰,۶\}) = ۱ \\
\mathrm{CE}_1^{(0)} &= \mathrm{CE}_2^{(0)} = \mathrm{CE}_3^{(0)} = ۰ \\
\mathrm{CE}_1^{(1)} &= \mathrm{CCE}_1^{(1)} - \mathrm{CE}_1^{(0)} = ۲ - ۰ = ۲ \\
\mathrm{CE}_2^{(1)} &= \mathrm{CCE}_2^{(1)} - \mathrm{CE}_2^{(0)} = -۲ - ۰ = -۲ \\
\mathrm{CE}_3^{(1)} &= \mathrm{CCE}_3^{(1)} - \mathrm{CE}_3^{(0)} = ۱ - ۰ = ۱
\end{aligned}
$$

اکنون میانه اثر ستونی مرحله اول که با $\mathrm{CE}^{(1)}$ نشان می‌دهیم، به‌صورت زیر حساب می‌شود:

$$
\mathrm{CE}^{(1)} = \mathrm{median}(\mathrm{CE}_1^{(1)}, \mathrm{CE}_2^{(1)}, \mathrm{CE}_3^{(1)}) = \mathrm{median}(۲,-۲,۱) = ۱.
$$

سپس اثرات ستون را از داده‌های مربوط به همان ستون کم می‌کنیم. نتایج در ماتریس (۴.۳) آورده شده‌اند .

---

[۴] Column effect

[۵] Change column effect

$$
\begin{array}{c} \\ b_۱ \\ \\ \\ \\ \\ b_۲ \\ \\ \\ \\ \\ \end{array}
\begin{array}{c}
\begin{array}{ccc} a_۱ & a_۲ & a_۳ \end{array} \\
\left(\begin{array}{ccc}
\begin{array}{c} c_۱ \\ c_۲ \\ c_۳ \\ c_۴ \\ c_۵ \end{array}
\begin{bmatrix} -۱ \\ -۳ \\ ۲ \\ ۸ \\ ۵ \end{bmatrix} &
\begin{bmatrix} -۷ \\ -۵ \\ ۲ \\ ۰ \\ ۱ \end{bmatrix} &
\begin{bmatrix} ۲ \\ ۵ \\ -۱ \\ -۴ \\ -۳ \end{bmatrix} \\
\begin{array}{c} c_۱ \\ c_۲ \\ c_۳ \\ c_۴ \\ c_۵ \end{array}
\begin{bmatrix} -۳ \\ -۸ \\ ۱ \\ ۹ \\ -۲ \end{bmatrix} &
\begin{bmatrix} ۴ \\ ۰ \\ -۹ \\ ۷ \\ -۷ \end{bmatrix} &
\begin{bmatrix} -۶ \\ -۸ \\ ۱ \\ ۹ \\ ۵ \end{bmatrix}
\end{array}\right)
\end{array}
\tag{۴.۳}
$$

پس از حذف اثرات سطر و ستون، اثرات لایه را حذف می‌کنیم. در ماتریس (۴.۳) با ثابت در نظر گرفتن سطوح لایه، میانه را روی سطوح سطر و ستون محاسبه می‌کنیم. همانند قبل مقادیر اولیه اثرات لایه‌ای[۶] (LE) را برای تمام سطوح، برابر صفر در نظر می‌گیریم. اثر متغیر لایه‌ای را با CLE[۷] نشان می‌دهیم.

$$
\begin{aligned}
\mathrm{CLE}_۱^{(۱)} &= \mathrm{median}(-۱,-۷,۲,-۳,۴,-۶) = -۲ \\
\mathrm{CLE}_۲^{(۱)} &= \mathrm{median}(-۳,-۵,۵,-۸,۰,-۸) = -۴ \\
\mathrm{CLE}_۳^{(۱)} &= \mathrm{median}(۲,۲,-۱,۱,-۹,۱) = ۱ \\
\mathrm{CLE}_۴^{(۱)} &= \mathrm{median}(۸,۰,-۴,۹,۷,۹) = ۷/۵ \\
\mathrm{CLE}_۵^{(۱)} &= \mathrm{median}(۵,۱,-۳,-۲,-۷,۵) = -۰/۵ \\
\mathrm{LE}_۱^{(۰)} &= \mathrm{LE}_۲^{(۰)} = \mathrm{LE}_۳^{(۰)} = \mathrm{LE}_۴^{(۰)} = \mathrm{LE}_۵^{(۰)} = ۰ \\
\mathrm{LE}_۱^{(۱)} &= \mathrm{CLE}_۱^{(۱)} - \mathrm{LE}_۱^{(۰)} = -۲ - ۰ = -۲ \\
\mathrm{LE}_۲^{(۱)} &= \mathrm{CLE}_۲^{(۱)} - \mathrm{LE}_۲^{(۰)} = -۴ - ۰ = -۴ \\
\mathrm{LE}_۳^{(۱)} &= \mathrm{CLE}_۳^{(۱)} - \mathrm{LE}_۳^{(۰)} = ۱ - ۰ = ۱ \\
\mathrm{LE}_۴^{(۱)} &= \mathrm{CLE}_۴^{(۱)} - \mathrm{LE}_۴^{(۰)} = ۷/۵ - ۰ = ۷/۵ \\
\mathrm{LE}_۵^{(۱)} &= \mathrm{CLE}_۵^{(۱)} - \mathrm{LE}_۵^{(۰)} = -۰/۵ - ۰ = -۰/۵
\end{aligned}
$$

میانه اثر لایه‌ای نیز از میانه‌گیری روی اثرات در تمام سطوح محاسبه می‌شود. میانه‌ی لایه‌ای

[۶] Liyer effect

[۷] Change liyer effect

مرحله اول با نماد $\mathrm{LE}^{(۱)}$ نشان داده می‌شود و به‌صورت زیر به‌دست می‌آید:

$$\begin{aligned}\mathrm{LE}^{(۱)} &= \mathrm{median}(\mathrm{LE}^{(۱)}_{۱}, \mathrm{LE}^{(۱)}_{۲}, \mathrm{LE}^{(۱)}_{۳}, \mathrm{LE}^{(۱)}_{۴}, \mathrm{LE}^{(۱)}_{۵}) \\ &= \mathrm{median}(-۲, -۴, ۱, ۷/۵, -۰/۵) = -۰/۵\end{aligned}$$

حال با کم کردن اثر لایه‌ای از داده‌ها ماتریس (۵.۳) به‌دست می‌آید. این ماتریس مقدار باقی‌مانده‌های مرحله اول را ارائه می‌دهد.

$$\begin{array}{c} \\ b_۱ \\ \\ \\ \\ \\ \\ b_۲ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} a_۱ \\ \left(\begin{array}{c} c_۱ \\ c_۲ \\ c_۳ \\ c_۴ \\ c_۵ \end{array}\right. \\ \begin{bmatrix} ۱ \\ ۱ \\ ۱ \\ -۰/۵ \\ ۵/۵ \end{bmatrix} \\ \\ \begin{array}{c} c_۱ \\ c_۲ \\ c_۳ \\ c_۴ \\ c_۵ \end{array} \\ \begin{bmatrix} -۱ \\ -۴ \\ ۰ \\ ۱/۵ \\ -۱/۵ \end{bmatrix} \end{array} \begin{array}{c} a_۲ \\ \\ \begin{bmatrix} -۵ \\ -۱ \\ ۱ \\ -۷/۵ \\ ۱/۵ \end{bmatrix} \\ \\ \\ \begin{bmatrix} ۶ \\ ۴ \\ -۱۰ \\ -۰/۵ \\ -۶/۵ \end{bmatrix} \end{array} \begin{array}{c} a_۳ \\ \\ \begin{bmatrix} ۴ \\ ۹ \\ -۲ \\ -۱۱/۵ \\ -۲/۵ \end{bmatrix} \\ \\ \\ \begin{bmatrix} -۴ \\ -۴ \\ ۰ \\ ۱/۵ \\ ۵/۵ \end{bmatrix} \left.\begin{array}{c} \\ \\ \\ \\ \\ \end{array}\right) \end{array} \tag{۵.۳}$$

در نهایت مقدار میانه کل مرحله اول $\mu^{(۱)}$ از مجموع سه میانه سطر، ستون و لایه حاصل می‌شود.

$$\mu^{(۱)} = \mathrm{RE}^{(۱)} + \mathrm{RE}^{(۱)} + \mathrm{LE}^{(۱)} = ۱/۵ + ۱ - ۰/۵ = ۲$$

اگر اثرات عوامل همچنان در باقی‌مانده‌های مرحله اول مشهود بود، می‌توان میانه‌پالایش‌های مراحل دیگر را نیز ادامه داد. با این تفاوت که باقی‌مانده‌های هر مرحله، داده‌های ورودی مرحله بعد محسوب می‌شوند. تکرار مراحل تا زمانی ادامه پیدا می‌کند که محقق به هدف خود برای پالایش داده‌ها برسد و اثر عوامل به‌طور کامل از داده‌ها خارج شده باشد.

۲.۳ کریگینگ سه بعدی میانه پالایش

کرسی (۱۹۹۳) کریگینگ میانه‌پالایش[۸] (MPK) را ارائه داد. او از میانه پالایش دو طرفه (توکی، ۱۹۷۷؛ امرسون و هوگلین، ۲۰۰۰) برای استخراج اثرات سطح متوسطی از داده‌های فضایی و

[۸] Median polish kriging

استفاده از کریگینگ معمولی (OK) برای پیشگویی فضایی فرآیند باقی‌مانده‌ها استفاده کرد. MPK با جزئیات بیشتر و کاربرد آن بر روی داده‌های زمین‌آماری که به‌صورت منظم یا منظم توزیع شده‌اند و همینطور بر روی داده‌های شمارشی در کرسی (۱۹۹۳) توضیح داده شده است. این روش نسبت به کریگینگ عام (UK) دارای دو مزیت است. در وهله اول مؤلفه‌ی میانگین در (۱۴.۱) در روش میانه پالایش نسبت به داده‌های دورافتاده مقاوم هستند و باقی‌مانده‌هایی با کمترین اریبی را برای تابع ساختار پیشگویی می‌کند (کرسی و گلونک، ۱۹۸۴). دوم اینکه پیشین معلومی برای تابع ساختار، تغییرنگار، فرض نمی‌شود.

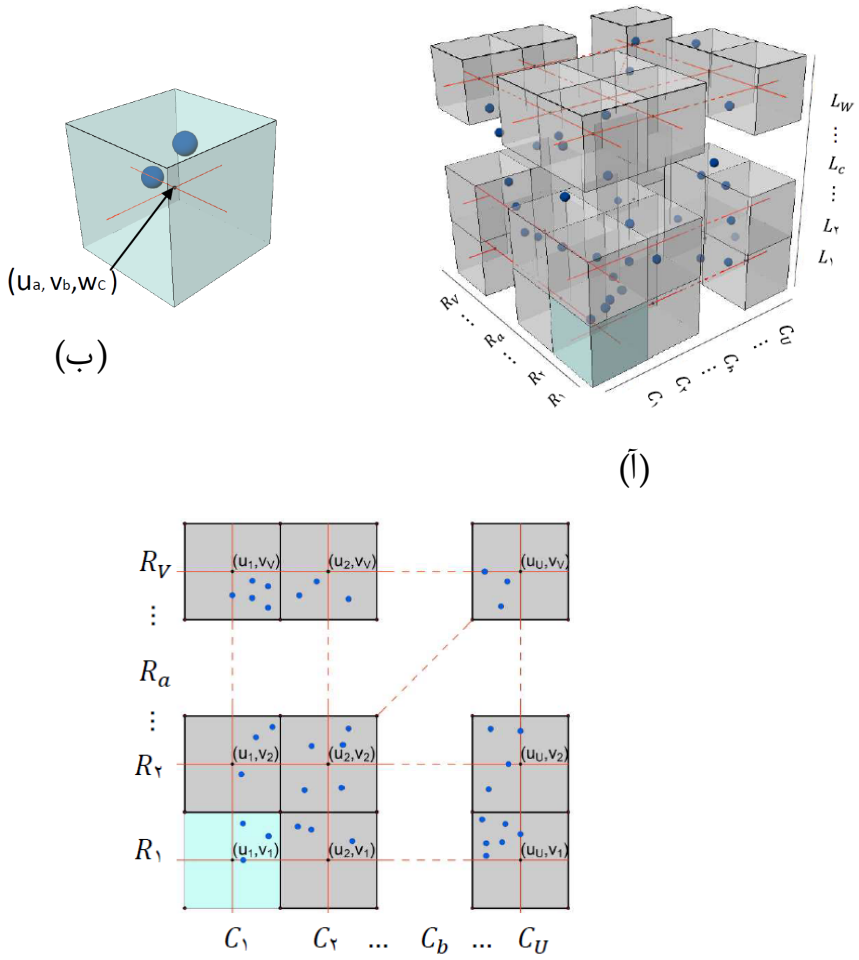
۱.۲.۳ تجزیه فرآیند فضایی

در این بخش فرض می‌کنیم، هدف تحلیل و پیشگویی فضایی در میدان $\{Z(\mathbf{s}), \mathbf{s} \in D\}$ است که در آن شاخص فضایی $\mathbf{s}$ در $D \subseteq \mathbb{R}^3$ تغییر می‌کند. این فرآیند را می‌توان به‌صورت زیر تجزیه کرد:

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}) \qquad (۶.۳)$$

که در آن $\mu(\cdot)$ مؤلفه‌ی روند فضایی و $\delta(\cdot)$ مؤلفه‌ی باقی‌مانده‌های تصادفی است که نشان دهنده‌ی تغییرات تصادفی فرآیند است. ما در این پایان‌نامه از رویکرد بریک (۲۰۰۱) و مارتینز و همکاران (۲۰۱۷) برای داده‌هایی که توزیع فضایی نامنظم دارند استفاده می‌کنیم ولی تحلیل ساختار را از دو بعد به سه بعد $U \times V \times W$ گسترش می‌دهیم. فرض می‌کنیم سه بعد به‌صورت جداگانه در یک شکل جمعی روی متغیر پاسخ تأثیر می‌گذارند. داده‌ها را در یک آرایه سه بعدی مرتب می‌کنیم. انتظار می‌رود که مشارکت طول، عرض و لایه بر متغیر پاسخ مشخص باشد.

در اینجا R_a، C_b و L_c به ترتیب عوامل سطر، ستون و لایه، با سطح‌های U، V و W هستند. اگر مختصات فضایی داده‌ها نامنظم باشند، با توجه به فاصله‌ی بین مختصات در هر بعد، یک شبکه مکعبی $p \times q \times r$ تعریف می‌کنیم تا در منطقه‌ی تحقیق قرار گیرد. این شبکه‌ی مکعبی داده‌ها را در سلول‌های مکعبی خود جای می‌دهد. توجه کنید که بعضی از سلول‌ها ممکن است خالی باشند. سپس هر داده، مختصات نزدیک‌ترین گره را دریافت می‌کند. در داده‌های چندبعدی، اگر چند داده مختصات یک گره را اختیار کنند، باید مقادیر داده‌ها با یک مقدار مناسب مانند میانه جایگزین شود. شکل ۱.۳ مشبکه‌بندی به روش میانه‌پالایش برای سه‌بعد و برشی از لایه، دو بعد را نشان داده است.



شکل ۱.۳: شماتیک مشبکه‌بندی به روش میانه‌پالایش (آ) و (ج) سه‌بعد، (ب) دو بعد

روند داده‌ها، $\mu(\mathbf{s})$، ممکن است به جهت مولفه‌ها و اثرات متقابل آن‌ها وابسته باشد، اما در اینجا اثرات را به‌صورت ساده در نظر گرفته‌ایم. پیرو روش بریک (۲۰۰۱) با استفاده از روش میانه پالایش در یک ساختار سه بعدی به استخراج مکرر اثرات سطر، ستون و لایه‌ها بر اساس الگوریتم بخش قبلی می‌پردازیم. شرط توقف ممکن است همگرایی باشد یا ممکن است توسط خود محقق تعریف شود. فرمول ریاضی الگوریتم میانه‌پالایش در بخش ۱.۱.۳ شرح داده شد.

برای داده‌ای در موقعیت $\mathbf{s}_{abc} \in D$، $Z(\mathbf{s}_{abc})$ $a = ۱, \ldots, U, b = ۱, \ldots, V, c = ۱, \ldots, W$ فرآیند تصادفی به‌صورت

$$Z(\mathbf{s}_{abc}) = \mu_z(\mathbf{s}_{abc}) + \delta_{abc} \qquad (۷.۳)$$

است که در آن

$$\mu_z(\mathbf{s}_{abc}) = \mu + R_a + C_b + L_c. \qquad (۸.۳)$$

و δ_{abc} تغییرات ناشی از نوسانات طبیعی و فرآیندهای اندازه‌گیری است. علاوه بر این μ میانگین کل و R_a، C_b و L_c به ترتیب اثر سطر aام، اثر ستون bام و اثر لایه‌ی cام است. سپس به حذف اثرات سه عامل می‌پردازیم تا جایی که به شرط (۱.۳) برسیم.

۲.۲.۳ درون‌یابی و برون‌یابی اثرات میانه پالایش

مقدار $\mu_z(\mathbf{s}_{abc})$ داده‌شده در فرمول (۸.۳) فقط مقدار میانگین هر سلول را برآورد می‌کند. پس به نوعی از درون‌یابی یا برون‌یابی برای تعریف مؤلفه‌ی روند در مکان جدید پیشگویی نیاز داریم. بنابراین هر مکان فضایی $\mathbf{s} = (u, v, w)$ را با ناحیه کراندار و ترکیبی از خطوط شبکه یعنی $u_a \le u < u_{a+1}$، $v_b \le v < v_{b+1}$ و $w_c \le w < w_{c+1}$ در نظر می‌گیریم. به‌طوری که $a \in \{1, \ldots, U\}$، $b \in \{1, \ldots, V\}$ و $c \in \{1, \ldots, W\}$. در این صورت برآورد روند فضایی به‌صورت زیر تعریف می‌شود:

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_a + \frac{u - u_a}{u_{a+1} - u_a}(\hat{R}_{a+1} - \hat{R}_a) \\ &+ \hat{C}_b + \frac{v - v_b}{v_{b+1} - v_b}(\hat{C}_{b+1} - \hat{C}_b) \\ &+ \hat{L}_c + \frac{w - w_c}{w_{c+1} - w_c}(\hat{L}_{c+1} - \hat{L}_c).\end{aligned}$$

در این‌جا برآورد روند میانه پالایش $\hat{\mu}_{mpk}$ حاصل از درون‌یابی خطی است و $\hat{R}_a$، $\hat{C}_b$ و $\hat{L}_c$ به ترتیب برآورد اثرهای سطری aام، ستونی bام و لایه‌ای cام است.

برای برآورد روند خارج از شبکه، بریک (۲۰۰۱) برون‌یابی را پیشنهاد کرده است. ما نیز برون‌یابی پیشنهادی او را به حالت سه بعدی گسترش می‌دهیم. حالت‌های مختلف برون‌یابی برآورد روند به‌صورت زیر هستند:

- وقتی $\mathbf{s} = (u, v, w)$ با $u < u_1$، $v_1 \le v_b < v < v_{b+1} < v_V$ و $w_1 \le w_c < w < w_{c+1} < w_W$، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_1 + \frac{u - u_1}{u_2 - u_1}(\hat{R}_2 - \hat{R}_1) \\ &+ \hat{C}_b + \frac{v - v_b}{v_{b+1} - v_b}(\hat{C}_{b+1} - \hat{C}_b) \\ &+ \hat{L}_c + \frac{w - w_c}{w_{c+1} - w_c}(\hat{L}_{c+1} - \hat{L}_c).\end{aligned}$$

- وقتی $\mathbf{s} = (u, v, w)$ با $u_1 \le u_a < u < u_{a+1} < u_U$، $v < v_1$ و $w_1 \le w_c < w < w_{c+1} < w_W$، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_a + \frac{u - u_a}{u_{a+1} - u_a}(\hat{R}_{a+1} - \hat{R}_a) \\ &+ \hat{C}_1 + \frac{v - v_1}{v_2 - v_1}(\hat{C}_2 - \hat{C}_1) \\ &+ \hat{L}_c + \frac{w - w_c}{w_{c+1} - w_c}(\hat{L}_{c+1} - \hat{L}_c).\end{aligned}$$

- وقتی $\mathbf{s}=(u,v,w)$ با $u_1 \le u_a < u < u_{a+1} < u_U$، $v_1 \le v_b < v < v_{b+1} < v_V$ و $w_1 < w$، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_a + \frac{u-u_a}{u_{a+1}-u_a}(\hat{R}_{a+1}-\hat{R}_a)\\ &+ \hat{C}_b + \frac{v-v_b}{v_{b+1}-v_b}(\hat{C}_{b+1}-\hat{C}_b)\\ &+ \hat{L}_1 + \frac{w-w_1}{w_2-w_1}(\hat{L}_2-\hat{L}_1).\end{aligned}$$

- وقتی $\mathbf{s}=(u,v,w)$ با $u < u_1$، $v < v_1$ و $w_1 \le w_c < w < w_{c+1} < w_W$، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_1 + \frac{u-u_1}{u_2-u_1}(\hat{R}_2-\hat{R}_1)\\ &+ \hat{C}_1 + \frac{v-v_1}{v_2-v_1}(\hat{C}_2-\hat{C}_1)\\ &+ \hat{L}_c + \frac{w-w_c}{w_{c+1}-w_c}(\hat{L}_{c+1}-\hat{L}_c).\end{aligned}$$

- وقتی $\mathbf{s}=(u,v,w)$ با $u < u_1$، $v_1 \le v_b < v < v_{b+1} < v_V$ و $w < w_1$، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_1 + \frac{u-u_1}{u_2-u_1}(\hat{R}_2-\hat{R}_1)\\ &+ \hat{C}_b + \frac{v-v_b}{v_{b+1}-v_b}(\hat{C}_{b+1}-\hat{C}_b)\\ &+ \hat{L}_1 + \frac{w-w_1}{w_2-w_1}(\hat{L}_2-\hat{L}_1).\end{aligned}$$

- وقتی $\mathbf{s}=(u,v,w)$ با $u_1 \le u_a < u < u_{a+1} < u_U$، $v < v_1$ و $w < w_1$، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_a + \frac{u-u_a}{u_{a+1}-u_a}(\hat{R}_{a+1}-\hat{R}_a)\\ &+ \hat{C}_1 + \frac{v-v_1}{v_2-v_1}(\hat{C}_2-\hat{C}_1)\\ &+ \hat{L}_1 + \frac{w-w_1}{w_2-w_1}(\hat{L}_2-\hat{L}_1).\end{aligned}$$

- وقتی $\mathbf{s}=(u,v,w)$ با $u < u_1$، $v < v_1$ و $w < w_1$، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} &+ \hat{R}_1 + \frac{u-u_1}{u_2-u_1}(\hat{R}_2-\hat{R}_1)\\ &+ \hat{C}_1 + \frac{v-v_1}{v_2-v_1}(\hat{C}_2-\hat{C}_1)\\ &+ \hat{L}_1 + \frac{w-w_1}{w_2-w_1}(\hat{L}_2-\hat{L}_1).\end{aligned}$$

- وقتی $\mathbf{s}=(u,v,w)$ با $u_U < u$، $v_1 \le v_b < v < v_{b+1} < v_V$ و $w_1 \le w_c < w < w_{c+1} <$

w_W، داریم

$$\begin{aligned}\hat{\mu}_{mpk}(\mathbf{s}_{abc}) = \hat{\mu} + \hat{R}_{U-1} + \frac{u-u_{U-1}}{u_U-u_{U-1}}(\hat{R}_U - \hat{R}_{U-1}) \\ + \hat{C}_b + \frac{v-v_b}{v_{b+1}-v_b}(\hat{C}_{b+1} - \hat{C}_b) \\ + \hat{L}_c + \frac{w-w_c}{w_{c+1}-w_c}(\hat{L}_{c+1} - \hat{L}_c).\end{aligned}$$

به‌طور مشابه برای حالت‌های $v_V < v$ و $w_W < w$ عمل می‌کنیم.

۳.۲.۳ کریگینگ معمولی روی باقی‌مانده‌ها

در نهایت پیشگویی توسط کریگینگ میانه پالایش از مجموع برآورد روند میانه پالایش با مقدار پیشگویی حاصل از کریگینگ معمولی روی باقی‌مانده‌های به‌دست آمده از میانه پالایش در همان سطح به‌دست می‌آید. یعنی مقدار پیشگویی میدان در نقطه مکانی جدید $\mathbf{s}_0$ عبارت است از

$$\hat{Z}_{mpk}(\mathbf{s}_0) = \hat{\mu}_{mpk}(\mathbf{s}_0) + \hat{\delta}_{mpk}(\mathbf{s}_0)$$

به‌طوری که

$$\hat{\delta}_{mpk}(\mathbf{s}_0) = \hat{\theta}_\delta + \mathbf{C^T}_\delta \mathbf{\Sigma}_\delta^{-1}(\boldsymbol{\delta} - \mathbf{1}\hat{\theta}_\delta) \qquad \mathbf{s}_0 \in \mathbb{R}^3$$

که در آن $\mathbf{Z}_{mpk} = (Z_{mpk}(\mathbf{s}_1), \ldots, Z_{mpk}(\mathbf{s}_n))^T$ با $\boldsymbol{\delta} = \mathbf{Z}_{mpk} - \mu_{mpk}$، $\mathbf{\Sigma}_\delta = \mathrm{Cov}(\boldsymbol{\delta})$، $\mathbf{1} = (1, \ldots, 1)^T$ و $\boldsymbol{\delta} = (\delta(\mathbf{s}_1), \ldots, \delta(\mathbf{s}_n))^T$، $\delta_0 = \delta(\mathbf{s}_0)$ با $\mathbf{C}_\delta = \mathrm{Cov}(\boldsymbol{\delta}, \delta_\mathbf{0})$، $\hat{\mu}_{mpk} = (\hat{\mu}_{mpk}(\mathbf{s}_1), \ldots, \hat{\mu}_{mpk}(\mathbf{s}_n))^T$ و $\hat{\theta}_\delta = (\mathbf{1}^T \sum_\delta^{-1} \mathbf{1})^{-1} \mathbf{1}^T \sum_\delta^{-1} \boldsymbol{\delta}$ نشان‌گر برآوردگر کمترین توان‌های دوم تعمیم‌یافته با میانگین ثابت، برای فرآیند باقی‌مانده‌های $\delta(\cdot)$ و بر اساس ساختار فضایی ماتریس کوواریانس $\mathbf{\Sigma}_\delta$ است.

کمترین توان‌های دوم خطای پیشگو[۹] (MSPE) برای کریگینگ میانه‌پالایش نیز برابر با مقدار واریانس کریگینگ معمولی باقی‌مانده‌ها است. بنابراین

$$\begin{aligned}&\mathrm{MSPE}(Z(\mathbf{s}_0), \hat{Z}_{mpk}(\mathbf{s}_0)) \approx \mathrm{MSPE}(\delta(\mathbf{s}_0), \hat{\delta}_{mpk}(\mathbf{s}_0)) \\ &= \hat{\sigma}^2_{mpk}(\mathbf{s}_0) = \sigma^2 - \mathbf{C}^T{}_\delta \Sigma_\delta^{-1} \mathbf{C}_\delta + (1 - \mathbf{1}^T \mathbf{\Sigma}_\delta^1 \mathbf{c}_\delta)^2 / (\mathbf{1}^T \mathbf{\Sigma}_\delta^1 \mathbf{1})\end{aligned}$$

که در آن $\sigma^2 = \mathrm{Cov}(0)$ همان ازاره مدل برازش داده‌شده به تغییرنگار است.

به‌منظور ارزیابی عملکرد رهیافت پیشنهادی ما بر اساس روش میانه‌پالایش، نتایج حاصل از پیشگویی را در قالب یک مطالعه شبیه‌سازی با روش کریگینگ عام که در آن برآورد تابع روند با روش اسپلاین محاسبه می‌شود، مقایسه کردیم. برای این منظور ابتدا روش برآورد ناپارامتری یک تابع روند با روش اسپلاین را معرفی می‌کنیم.

[۹] Mean squared prediction error

۳.۳ روش اسپلاین

در تحلیل رگرسیونی داده‌های فضایی، روش مرسوم استفاده از مدل‌هایی با اثرات آمیخته خطی پارامتری است، به‌طوری‌که وابستگی فضایی به‌عنوان یک اثر تصادفی فضایی در نظر گرفته می‌شود و معمولاً یک فرآیند گوسی با میانگین صفر و تابع کوواریانس پارامتری فرض می‌شود (استروپ، ۲۰۱۲). بسیاری از نویسندگان از جمله ریپلی (۱۹۸۱) و کرسی (۱۹۹۳) از روش‌های پارامتری برای استنباط آماری در مدل‌های فضایی استفاده کرده‌اند. با این حال اغلب اوقات، با شرایطی رو به رو می‌شویم که یک مدل پارامتری خاص را با اطمینان نمی‌توان به‌کار گرفت. بنابراین باید از یک روش ناپارامتری جایگزین استفاده کرد.

روش اسپلاین می‌تواند به‌عنوان یک روش جایگزین برای مدل‌بندی تابع روند فضایی مورد استفاده قرار گیرد. این روش، یک روش نیمه‌پارامتری به‌وجود می‌آورد که برای محاسبه‌ی اثرات روند فضایی نیازی به مشخصات ساختاری پارامتری ندارد (لودویگ و همکاران، ۲۰۲۰).

یکی از رایج‌ترین روش‌های هموارسازی رگرسیون ناپارامتری، اسپلاین‌ها هستند که برای اولین بار توسط شوئنبرگ (۱۹۴۶) معرفی شد. وی برای نخستین بار قضیه‌ای برای اثبات وجود اسپلاین درون‌یاب مطرح کرد و بعد مطالعه‌ی چندجمله‌ای‌های تکه‌ای برای توابع اسپلاین آغاز شد. با فرض $s \in \mathbb{R}$، فاصله‌ی $[a,b]$ را می‌توان به‌صورت اجتماع دو مجموعه $[a,t] \cup [t,b]$ در نظر گرفت. اسپلاین مرتبه یک، تابعی است که در هر یک از بازه‌های فوق یک تابع خطی و در نقطه‌ی t پیوسته باشد. اسپلاین خطی مرتبه اول را می‌توان به‌صورت

$$\mu(s) = \begin{cases} \beta_0 + \beta_1 s & s \le t \\ \beta_0' + \beta_1' s & s > t \end{cases} \tag{۹.۳}$$

نمایش داد. با در نظر گرفتن شرط پیوستگی تابع μ در نقطه‌ی $s = t$ داریم

$$\begin{aligned} \beta_0 + \beta_1 t &= \beta_0' + \beta_1' t \\ \beta_0' &= \beta_0 + \beta_1 t - \beta_1' t \end{aligned} \tag{۱۰.۳}$$

و با جایگذاری β_0' در ضابطه‌ی دوم μ داریم

$$\beta_0' + \beta_1' s = \beta_0 + \beta_1 t - \beta_1' t + \beta_1' s = \beta_0 + \beta_1 t + \beta_1'(s-t)$$

پس

$$\mu(s) = \begin{cases} \beta_0 + \beta_1 s & s \le t \\ \beta_0 + \beta_1 t + \beta_1'(s-t) & s > t \end{cases} \tag{۱۱.۳}$$

که می‌توان تابع دو ضابطه‌ای را به‌صورت

$$\mu(s) = \beta_0 + \beta_1 s + \beta_2 (s-t)_+$$

نوشت که در آن $(s-t)_+ = \mu(s)(s-t,0) = (s-t)I(s>t)$، $I(\cdot)$ تابع نشانگر و $\beta_2 = \beta_1' - \beta_1$ اختلاف بین شیب خطوط اول و دوم تابع را نشان می‌دهد. به‌عبارتی می‌توان به این نتیجه رسید که تابع μ ترکیب خطی از توابع پایه ۱، s و $(s-t)_+$ است. اسپلاین مرتبه اول را اسپلاین خطی نیز می‌نامند.

اگر $[a,b]$ به $k+1$ زیربازه افراز شده باشد، یعنی

$$[a,t_1), [t_1,t_2), \ldots, [t_K,b]$$

به‌طوری که $a < t_1 < \ldots < t_K < b$، نقاط $\{t_i\}_{i=1}^K$ را گره می‌نامند. با این شرایط اسپلاین مرتبه اول را می‌توان به‌صورت

$$\mu(s) = \beta_0 + \beta_1 s + \sum_{k=1}^{K} \beta_{k+1}(s-t_k)_+$$

نشان داد که ترکیب خطی از توابع پایه ۱، s، $(s-t_1)_+$، ... و $(s-t_K)_+$ است. به‌طور مشابه، توابع اسپلاین مرتبه‌های بالاتر تعریف می‌شوند. اگر هر یک از بازه‌های مذکور یک چندجمله‌ای از مرتبه‌ی d باشد که $\mu(\cdot)$ و مشتق‌های آن از مرتبه‌ی ۱ تا $d-1$ در گره‌ها پیوسته باشند، آن‌گاه تابع $\mu(\cdot)$ را یک اسپلاین مرتبه d می‌نامند. تابع اسپلاین مرتبه d را می‌توان به‌صورت

$$\mu(s) = \beta_0 + \beta_1 s + \beta_2 s^2 + \ldots + \beta_d s^d + \sum_{k=1}^{K} \beta_{k+d}(s-t_k)_+^d$$

بیان کرد، که در آن $(s-t_k)_+^d = (s-t_k)^d I(s>t)$ تابع بریده‌شده توانی از مرتبه‌ي d است. به بیان ساده‌تر، این تابع ترکیب خطی از یک چندجمله‌ای مرتبه d و توابع بریده‌شده است. مجموعه‌ی

$$\{1, s, s^2, \ldots, s^d, (s-t_1)_+^d, \ldots, (s-t_K)_+^d\}$$

یک پایه برای فضای اسپلاین مرتبه d با گره‌های t_1، t_2، ... و t_k است. اگر $d=3$ باشد، تابع مورد نظر را اسپلاین مکعبی[۱۰] می‌نامند. اسپلاین مکعبی به دلیل انعطاف‌پذیری بالایی که دارد یکی از رایج‌ترین و پرکاربردترین نوع اسپلاین‌هاست که در بسیاری از علوم مورد استفاده قرار می‌گیرد.

مدل رگرسیون ناپارامتری $\mathbf{Z} = \boldsymbol{\mu}(s) + \boldsymbol{\delta}$ را در نظر بگیرید، که در آن

$$\begin{aligned} \mathbf{Z} &= (Z_1, \ldots, Z_n)^T \\ \boldsymbol{\mu}(s) &= (\mu(s_1), \ldots, \mu(s_n))^T \\ \boldsymbol{\delta} &= (\delta_1, \ldots, \delta_n) \end{aligned} \qquad (۱۲.۳)$$

[۱۰] Cubic spline

اگر به‌جای تابع $\mu(\cdot)$ از تابع اسپلاین استفاده شود، رگرسیون به‌دست آمده را رگرسیون اسپلاین می‌نامند. بنابراین می‌توان مدل زیر را ارائه کرد:

$$Z_i = \beta_0 + \beta_1 s_i + \beta_2 s_i^2 + \ldots + \beta_d s_i^d + \sum_{k=1}^{K} \beta_{d+k}(s_i - t_k)_+^d + \delta_i \qquad i = 1, \ldots, n$$

که در آن δ_i در پذیره‌های زیربنایی رگرسیون صدق می‌کند. با استفاده از نمادگذاری ماتریسی، رگرسیون اسپلاین به‌صورت

$$\mathbf{Z} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\delta}$$

نمایش داده می‌شود که در آن $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{d+K})$ و ماتریس طرح $\mathbf{B}$ به‌صورت

$$\mathbf{B} = \begin{pmatrix} 1 & s_1 & s_1^2 & \ldots & s_1^d & (s_1 - t_1)_+^d & (s_1 - t_2)_+^d & \ldots & (s_1 - t_K)_+^d \\ 1 & s_2 & s_2^2 & \ldots & s_2^d & (s_2 - t_1)_+^d & (s_2 - t_2)_+^d & \ldots & (s_2 - t_K)_+^d \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_n & s_n^2 & \ldots & s_n^d & (s_n - t_1)_+^d & (s_n - t_2)_+^d & \ldots & (s_n - t_K)_+^d \end{pmatrix}_{n \times (d+k+1)}$$

(۱۳.۳)

است. در $\mathbf{B}$ درایه‌های ستون اول همگی ۱ هستند و درایه‌های d ستون بعدی مربوط به مشاهدات متغیرهای s، s^2، ... و s^d و درایه‌های ستون‌های $d+2$ تا $d+K+1$ مربوط به مشاهدات حاصل از متغیرهای $(s-t_1)_+^d$، ... و $(s-t_K)_+^d$ به ازای گره‌های از پیش تعیین‌شده t_1، t_2، ... و t_K است. رگرسیون اسپلاین تعریف شده متعلق به رده رگرسیون خطی است. به‌عبارتی تابع $\mu(\cdot)$ نسبت به پارامترها خطی است و هزینه‌های محاسباتی برآورد و منحنی رگرسیون اندک است. پس می‌توان نتیجه گرفت که برازش یک مدل رگرسیون اسپلاین بر روی داده‌ها را به راحتی می‌توان با استفاده از روش‌های مرسوم پیاده‌سازی کرد.

با استفاده از روش کمترین توان‌های دوم خطا، برآوردگر بردار $\boldsymbol{\beta}$ به‌صورت $\hat{\boldsymbol{\beta}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Z}$ به‌دست می‌آید. عیبی که به این برآوردگر وارد است، هم‌خطی بین پارامترهای توان بریده‌شده

$$(s - t_1)_+^d, (s - t_2)_+^d, \ldots, (s - t_K)_+^d$$

است که موجب عدم وارون‌پذیری یا بدشرطی ماتریس $\mathbf{B}^T\mathbf{B}$ می‌شود و این موضوع ناپایداری ضرایب رگرسیونی را در پی دارد.

یک روش مناسب پیشنهادشده برای حل این مشکل، استفاده از B-اسپلاین به‌جای پایه‌های توان بریده‌شده $(s - t_i)_+^d$، $i = 1, \ldots, K$، است. بازه‌ی $[a, b]$ و تعداد K گره به‌صورت

$$a = t_0 < t_1 < \ldots < t_K < t_{K+1} = b$$

را در نظر بگیرید. متداول‌ترین روش محاسبه‌ی پایه‌های B-اسپلاین مرتبه‌ی d استفاده از پایه‌های B-اسپلاین مرتبه‌ی $d-1$ است که بر اساس رابطه‌ی بازگشتی زیر است:

$$\mathbf{B}_{i,d}(s) = (\frac{s - t_i}{t_{i+d} - t_i})\mathbf{B}_{i,d-۱}(s) + (\frac{t_{i+d+۱} - s}{t_{i+d+۱} - t_{i+۱}})\mathbf{B}_{i+۱,d-۱}(s), \qquad i = -d, -d+۱, \ldots, K$$

که در آن $t_{-d} = t_{-d+۱} = \ldots = t_{-۲} = t_{-۱} = T_۰ = a$ و $t_{K+۱} = b$. تعداد پایه‌های B-اسپلاین مرتبه‌ی d برابر $d + K + ۱$ است. با استفاده از این پایه‌ها می‌توان چندجمله‌ای‌های متعامد را به‌دست آورد.

در این پایان‌نامه از روش رگرسیون اسپلاین و بر اساس توابع B-اسپلاین در کنار رهیافت میانه‌پالایش، تابع روند را برآورد و کریگینگی مشابه کریگینگ میانه‌پالایش اجرا می‌کنیم. نتایج این دو روش را با هم مقایسه می‌کنیم.

۴.۳ مطالعه شبیه‌سازی

در این بخش عملکرد کریگینگ میانه پالایش را با داده‌های شبیه‌سازی‌شده سه بعدی منظم بررسی می‌کنیم. سپس این روش را با کریگینگ عامی که روند آن با استفاده از روش اسپلاین مدل شده است، مقایسه می‌کنیم.

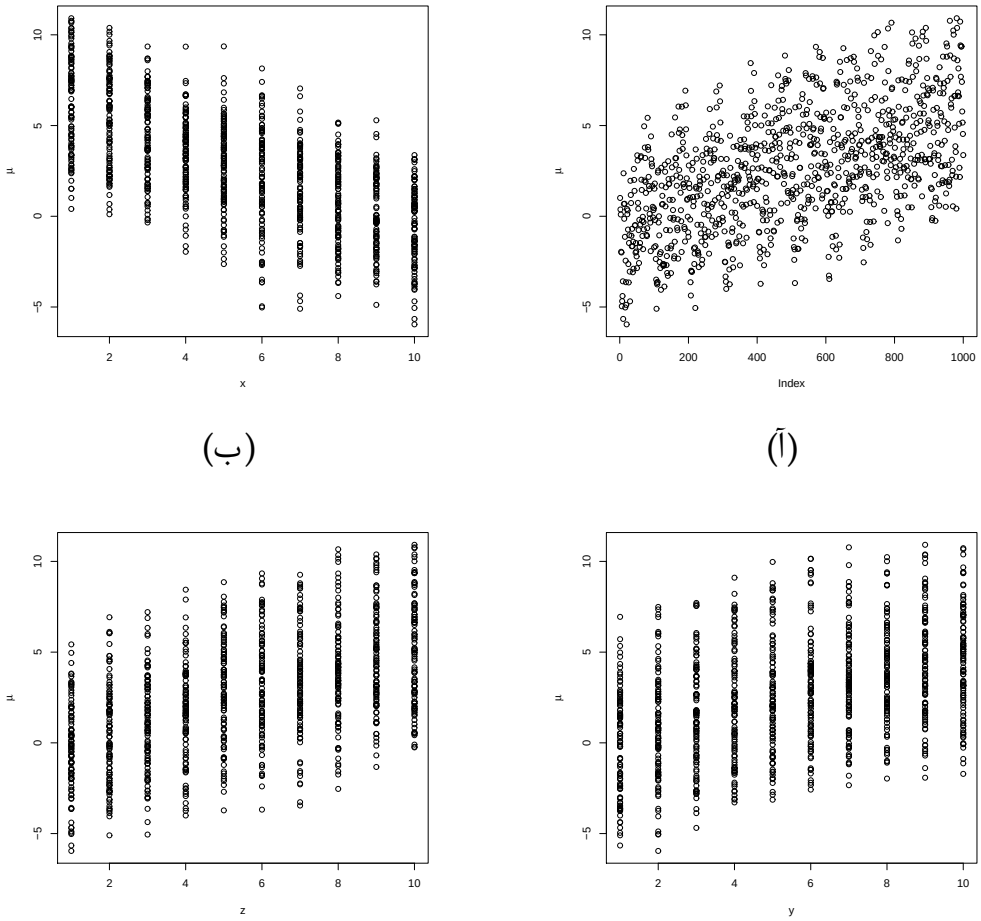
در مطالعه شبیه‌سازی دو حالت خطی و غیرخطی را برای روند واقعی داده‌ها در نظر گرفتیم که به تفکیک آن‌ها را تشریح می‌کنیم.

۱.۴.۳ روند خطی

برای تولید ۱۰۰۰ داده‌ی فضایی با مختصات سه بعدی (طول، عرض و لایه) از بسته geoR در محیط نرم‌افزار R استفاده کردیم. در مدل (۶.۳) داده‌ها را از رابطه‌ی زیر بر اساس یک تابع روند خطی تولید کردیم:

$$\mu = ۱ - (۰/۸ \times \mathbf{x}) + (۰/۵ \times \mathbf{y}) + (۰/۶ \times \mathbf{z})$$

اثر تصادفی فضایی $\delta(\mathbf{s})$ را از یک میدان تصادفی گاوسی با میانگین صفر و ماتریس کواریانسی که از مدل نمایی (۷.۱) به دست می‌آید تولید کردیم، که در آن پارامتر مدل $\phi = ۰/۵$ انتخاب شد. اثرقطعه‌ای، دامنه و ازاره را نیز به ترتیب برابر با ۰، ۱ و ۱ در نظرگرفتیم. نمودارهای پراکنش داده‌های خام و داده‌ها بر حسب سه محور مختصات در شکل ۲.۳ نمایش داده شده‌اند.



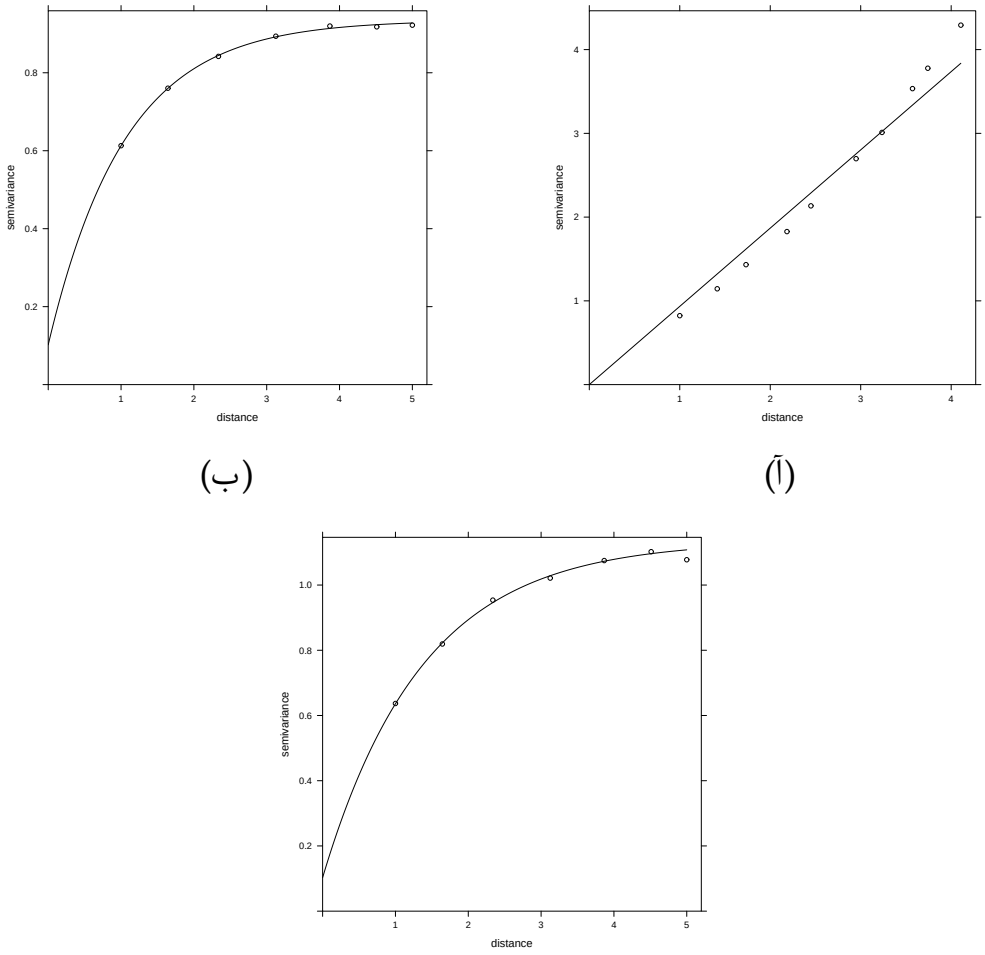
شکل ۲.۳: نمودار پراکنش داده‌های خام و پراکنش نسبت به مختصات x ، y و z به ترتیب (آ)، (ب)، (ج) و (د).

نمودار نیم‌تغییرنگار تجربی و مدل برازش داده شده بر اساس سه رهیافت مختلف در شکل ۳.۳ آورده شده است. تغییرنگار برآورد شده بر اساس داده‌های اصلی (مدل خطی) در شکل ۲.۳(آ) نشان می‌دهد که میدان مانا نیست و استفاده از داده‌ها قبل از حذف روند بیانگر خطای زیادی در برآورد ساختار وابستگی فضایی داده‌ها است. بنابراین وجود روند در داده‌ها کاملاً تایید می‌شود.

پس داده‌ها باید ابتدا روندزدایی شوند. قبل از این کار برای ارزیابی عملکرد روش‌های پیشنهادی داده‌ها را به سه روش ۵ بسته[۱۱] ، ۱۰ بسته[۱۲] و LOOCV[۱۳] (فریدمن و همکاران، ۲۰۰۱) به دو مجموعه آموزشی و آزمون تقسیم کردیم و با داده‌های آموزشی دو الگوریتم میانه‌پالایش و

[۱۱] 5 Fold
[۱۲] 10 Fold
[۱۳] Leave one out cross validation

رگرسیون خطی را روی سه مختصات x، y و z برای برآورد روند و محاسبه باقی‌مانده‌ها اجرا کردیم. نتیجه برآورد تغییرنگار بر اساس باقی‌مانده‌ها برای دو روش، مدل مشابه نمایی را پیشنهاد می‌دهد.



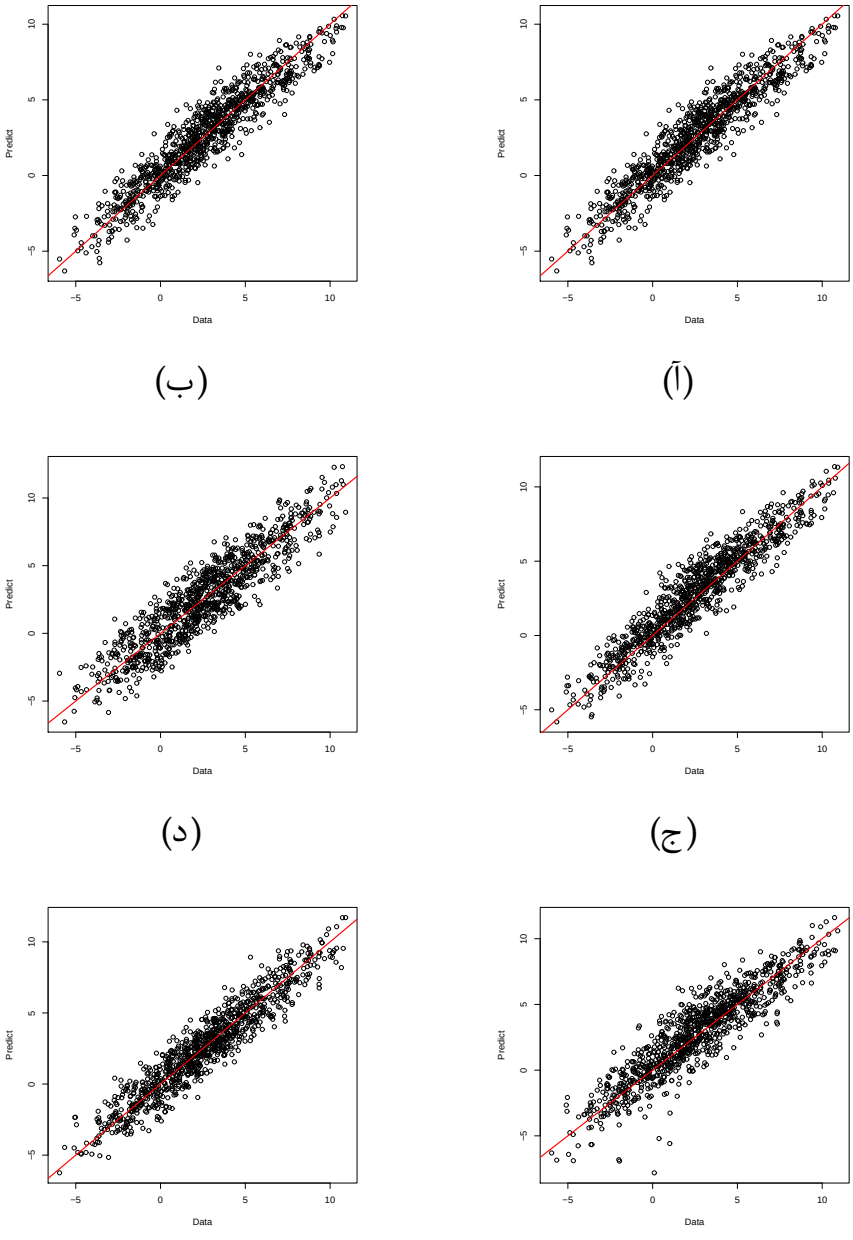
(ب) (آ)

(ج)

شکل ۳.۳: نمودار سه بعدی نیم‌تغییرنگار و (آ): منحنی مدل پارامتری خطی برازش داده شده روی داده‌های اصلی. (ب) منحنی مدل پارامتری نمایی برازش داده شده روی باقی‌مانده‌های حاصل از میانه‌پالایش. (ج) منحنی مدل پارامتری نمایی برازش داده شده روی باقی‌مانده‌های مدل برازش داده شده با استفاده از رگرسیون خطی

محور افقی نمودارهای پراکنش در شکل ۴.۳ نشان‌دهنده‌ی مقادیر واقعی داده‌های آزمون و محور عمودی آن‌ها بیانگر مقادیر پیشگویی به‌دست آمده از کریگینگ میانه‌پالایش و عام است. نمودارهای سمت راست نتایج حاصل از روش میانه‌پالایش و سمت چپ روش کریگینگ عام را نشان می‌دهند. در این حالت (روند خطی) در هر دو روش توزیع پراکنش داده‌ها حول نیم‌ساز ربع اول و سوم است و بیانگر مناسب بودن مدل و مقدار ناچیز باقی‌مانده‌ها است. همچنین نیم‌ساز بین داده‌ها قرار دارد و می‌توان نتیجه گرفت میانگین خطا برابر با صفر است.

با استفاده از روش‌های اعتبارسنجی متقابل ۵ بسته، ۱۰ بسته و و LOOCV مقادیر MSPE را برای دو روش محاسبه کردیم که در جدول ۱.۳ گزارش شده است.



شکل ۴.۳: مقادیر پیشگویی در مقابل مقادیر واقعی داده‌های آزمون با استفاده از کریگینگ میانه‌پالایش (آ) ۵ بسته، (ج) ۱۰ بسته و (ه) LOOCV و با استفاده از کریگینگ عام (ب) ۵ بسته، (د) ۱۰ بسته و (و) LOOCV.

جدول ۱.۳: مقادیر MSPE

کریگینگ عام			کریگینگ میانه‌پالایش		
۵ بسته	۱۰ بسته	LOOCV	۵ بسته	۱۰ بسته	LOOCV
۱.۹۶	۱.۹۱	۱.۱۱	۱.۲۰	۱.۱۸	۱.۶

بر اساس مقادیر جدول، مشخص است که روش میانه‌پالایش با استفاده از روش‌های ۵ بسته و ۱۰ بسته برتر از روش کریگینگ عام است، اما با استفاده از اعتبارسنجی متقابل LOOCV روش کریگینگ‌عام بهتر عمل می‌کند. نمودارهای پراکنش شکل ۴.۳ در دو رویکرد اعتبارسنجی‌متقابل نیز این موضوع را تأیید می‌کنند.

۲.۴.۳ روند غیر خطی

در شبیه‌سازی دوم روند داده‌ها را غیرخطی در نظر گرفتیم و از رابطه‌ی زیر آن‌ها را تولید کردیم:

$$\mu = ۱ + ۰/۸\log(x) + ۰/۶\sin(y) - ۰/۸\cos(z) + ۰/۰۵z \times (-۰/۰۵)y \qquad (۱۴.۳)$$

اثر تصادفی فضایی را مشابه روند خطی تولید کردیم. نمودارهای پراکنش داده‌های خام و به تفکیک سه بعد در شکل ۵.۳ آورده شده‌اند.

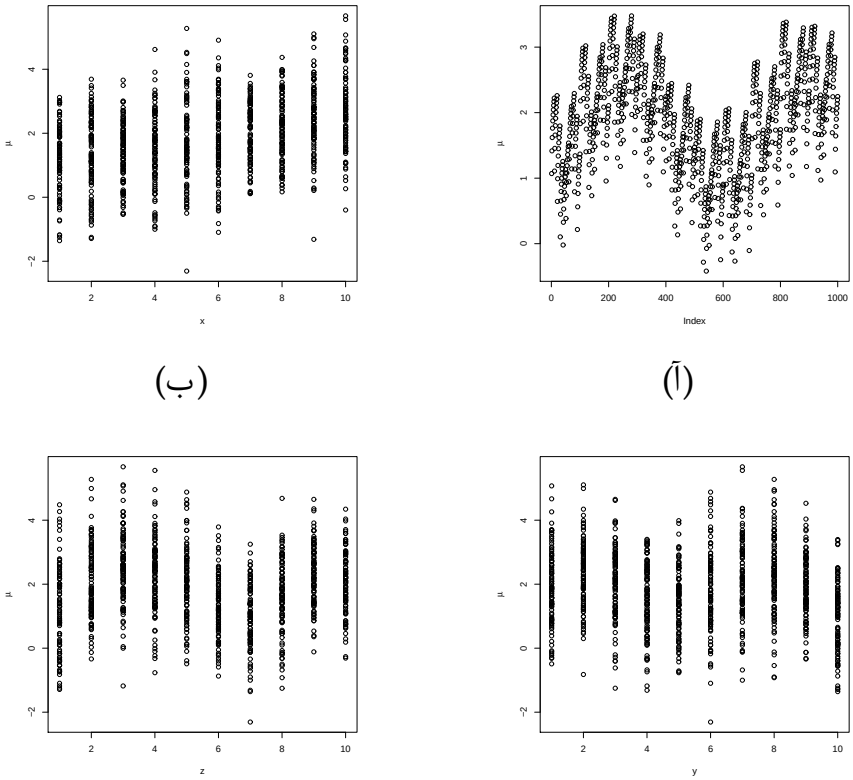
با اجرای کریگینگ سه بعدی عادی روی باقی‌مانده‌های حاصل از سه روش رگرسیون خطی، میانه‌پالایش و اسپلاین، نتایج اعتبارسنجی متقابل پیشگویی داده‌های آزمون را محاسبه و نمودارهای پراکنش آن‌ها در شکل ۶.۳ گزارش کردیم.

مشابه شبیه‌سازی اول، ابتدا پس از حذف روند تابع تغییرنگار بر اساس هر سه روش میانه‌پالایش کریگینگ عام با روند خطی و کریگینگ عام با روند ناپارامتری حاصل از اسپلاین برآورد شدند و کریگینگ معمولی روی باقی‌مانده‌های حاصل پیاده‌سازی شد. نتایج پیشگویی برای داده‌های مجموعه‌های آزمون حاصل از سه روش ۵ بسته، ۱۰ بسته و LOOCV در شکل ۶.۳ گزارش شده‌اند. در این حالت نیز توزیع پراکنش داده‌ها نسبتاً حول نیم‌ساز ربع اول و سوم است و بیانگر مناسب بودن روش‌ها است.

مقدار MSPE را برای سه روش مذکور محاسبه کردیم که در جدول ۲.۳ گزارش شده‌اند.

جدول ۲.۳: مقادیر MSPE

کریگینگ عام			کریگینگ میانه‌پالایش			کریگینگ عام با روند اسپلاین		
۵ بسته	۱۰ بسته	LOOCV	۵ بسته	۱۰ بسته	LOOCV	۵ بسته	۱۰ بسته	LOOCV
۱.۹۸	۱.۹۵	۱.۱۳	۱.۶۵	۱.۳۴	۱.۲۱	۰.۶۳۵	۰.۶۲۴	۰.۶۱۸

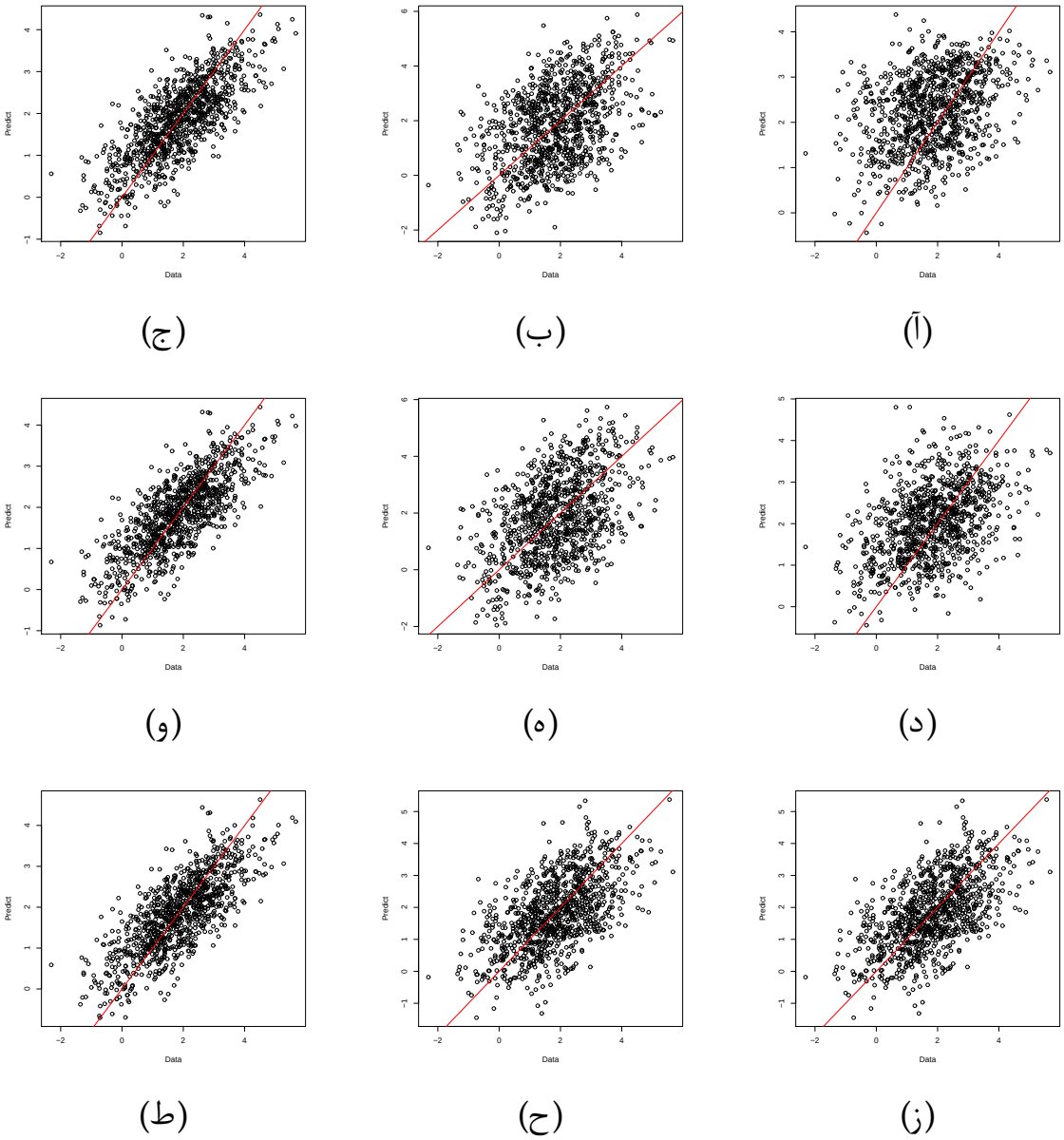


(آ) (ب)

(ج) (د)

شکل ۵.۳: نمودار پراکنش داده‌ها و پراکنش نسبت به مختصات x ، y و z به ترتیب (آ)، (ب)، (ج) و (د).

بر اساس مقادیر جدول، مشابه حالت روند خطی، روش میانه‌پالایش در پیشگویی فضایی در دو روش ۵ بسته و ۱۰ بسته بهتر از کریگینگ عام عمل می‌کند ولی در روش LOOCV، کریگینگ عام عملکرد بهتری نسبت به کریگینگ میانه‌پالایش دارد. البته با توجه به مقادیر جدول می‌توان درک کرد که روش اسپلاین بهتر از دو رقیب دیگر عمل کرده است، که نشان از توانایی بالای این روش در برآورد توابع غیرخطی دارد. نمودارهای پراکنش شکل ۶.۳ در سه رویکرد اعتبارسنجی‌متقابل نیز این موضوع را تأیید می‌کنند.



(آ) (ب) (ج)

(د) (ه) (و)

(ز) (ح) (ط)

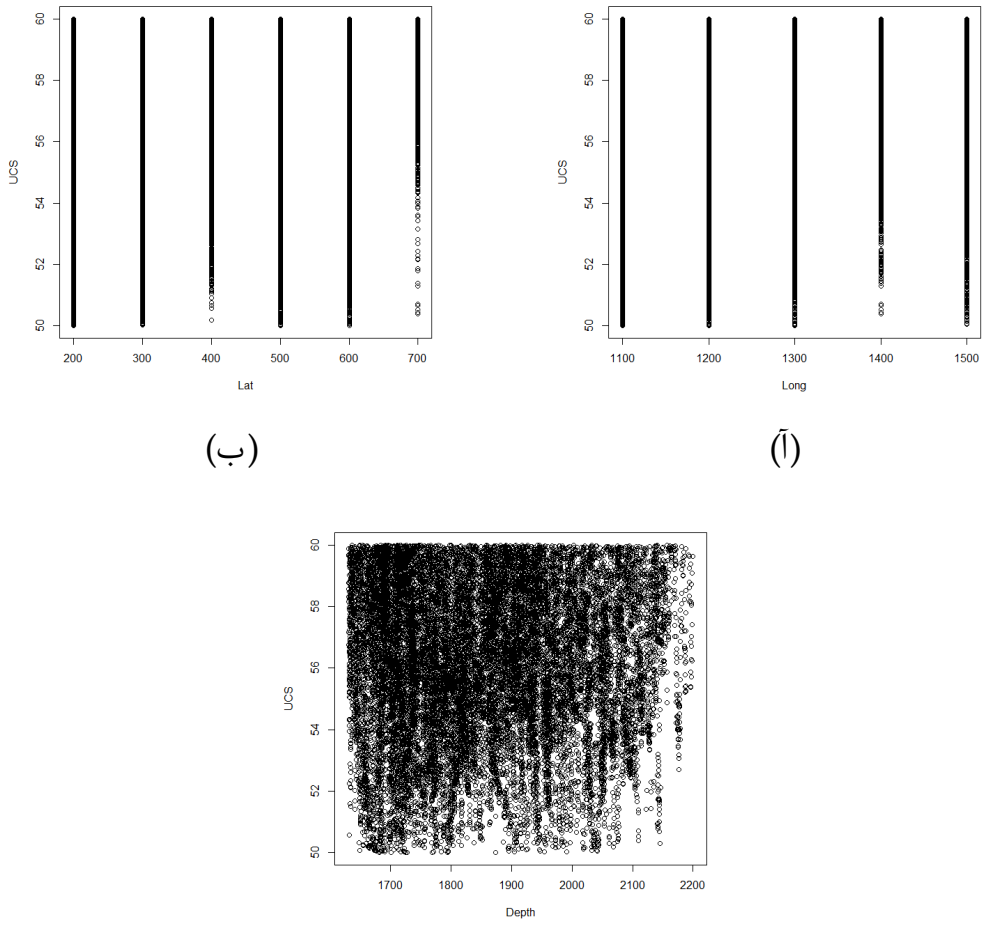
شکل ۶.۳: مقادیر پیشگویی در مقابل مقادیر واقعی داده‌های آزمون با استفاده از کریگینگ میانه‌پالایش (آ) ۵ بسته، (د) ۱۰ بسته و (ز) LOOCV، کریگینگ عام با روند خطی (ب) ۵ بسته، (ه) ۱۰ بسته و (ح) LOOCV و با استفاده از کریگینگ عام روی باقی‌مانده‌های حاصل از اسپلاین (ج) ۵ بسته، (و) ۱۰ بسته و (ط) LOOCV.

۵.۳ مطالعه پیشگویی سه بعدی ناپارامتری مقاومت فشاری تک محوری سنگ

در اینجا به پیشگویی سه بعدی مقاومت فشاری تک محوری سازند با استفاده از روش‌های ناپارامتری می‌پردازیم. با به‌کارگیری از رویکرد ارزیابی متقابل، داده‌های UCS را به‌طور تصادفی

به دو مجموعه‌ی آزمایشی و آزمون تقسیم کردیم. پیش‌تر گفته‌شد که میدان تصادفی را می‌توان به‌صورت $Z(\mathrm{s}) = \mu(\mathrm{s}) + \delta(\mathrm{s}), \quad \mathrm{s} \in D$ تجزیه کرد، پس ابتدا روند، $\mu(\mathrm{s})$، را با استفاده از میانه‌پالایش و اسپلاین برآورد کردیم، سپس مقادیر برآورد شده‌ی روند را با $\delta(\mathrm{s})$، مقادیر پیشگویی شده با استفاده از کریگینگ معمولی روی باقی‌مانده‌های حاصل از دو روش فوق جمع شدند. در نهایت مقادیر به‌دست آمده با مقادیر واقعی داده‌های مجموعه‌ی آزمون UCS مقایسه شدند. برای بررسی کارکرد پیشگویی سه بعدی هر سه روش میانه‌پالایش، اسپلاین و کریگینگ عام از مقدار عددی و نمودار جعبه‌ای MSPE بهره بردیم.

نمودارهای پراکنش داده‌ها بر حسب سه محور مختصات در شکل ۷.۳ نشان داده شده‌اند.



شکل ۷.۳: نمودار پراکنش داده‌ها نسبت به مختصات طول ، عرض و عمق به ترتیب (آ)، (ب) و (ج).

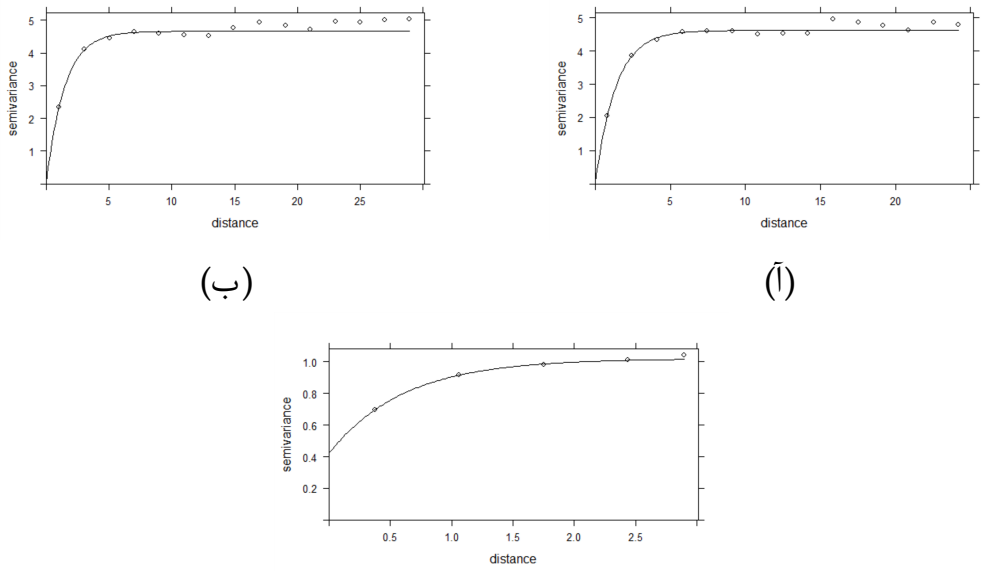
شکل ۸.۳ نمودار نیم‌تغییرنگار روی باقی‌مانده‌های حاصل از سه‌روش را نمایش می‌دهد، شایان ذکر است که نمودار نیم‌تغییرنگار تجربی و مدل منتخب برازش داده‌شده روی داده‌های اصلی در شکل ۵.۲ (آ)، نمایش داده شده، همچنین مقادیر اثرقطعه‌ای، ازاره و دامنه‌ی

نیم‌تغییرنگار تجربی روی باقی‌مانده‌های به‌دست آمده از روش‌های مذکور در جدول ۳.۳ ارائه شده است.

جدول ۳.۳: پارامترهای تغییرنگار

	کریگینگ عام	کریگینگ میانه‌پالایش	کریگینگ اسپلاین
ازاره	۴.۶۰۵	۰.۵۹۷	۴.۵۸۱
دامنه	۱.۴۴۵	۰.۶۱۱	۱.۳۶۴
اثر قطعه‌ای	۰.۰۵۴	۰.۴۲۳	۰.۰۳۵

بر اساس مقادیر MSPE که در جدول ۴.۳ ارائه شده، می‌توان نتیجه گرفت که پیشگویی فضایی با روش اسپلاین از MSPE کوچک‌تری نسبت به دو رقیب خود برخوردار است، پس این روش نسبت به دو روش دیگر در پیش‌گویی فضایی برتری دارد. در نهایت می‌توان نتیجه گرفت که روش کریگینگ عام نسبت به میانه‌پالایش برتری دارد. نمودارهای جعبه‌ای شکل ۹.۳ نیز این موضوع را تایید می‌کنند. دلیل ضعیف عمل کردن میانه پالایش با توجه به شکل ۷.۳ را می‌توان خطی بودن روند داده‌ها نسبت به سه مختصات طول، عرض و عمق دانست، که در این‌صورت برآورد روند با رگرسیون که بر حسب میانگین عمل می‌کند کاراتر از میانه‌پالایش است که بر حسب میانه عمل می‌کند.



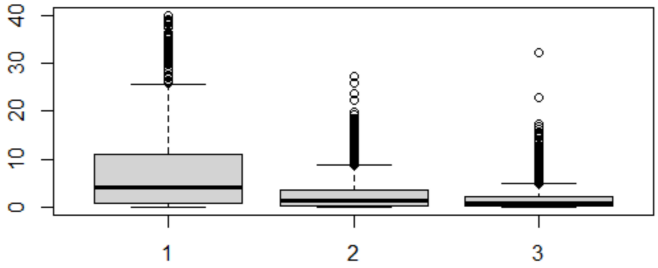
(ب) (آ)

(ج)

شکل ۸.۳: نمودار سه بعدی نیم‌تغییرنگار و منحنی مدل پارامتری نمایی برازش داده شده به باقی‌مانده‌های حاصل از روش‌های (آ): اسپلاین، (ب): ماکسیمم درستنمایی و (ج): میانه‌پالایش.

جدول ۴.۳: پارامترهای تغییرنگار

	کریگینگ عام	کریگینگ میانه‌پالایش	کریگینگ اسپلاین
مقادیر MSPE	۲.۷۹	۹.۸۹	۱.۷۸



شکل ۹.۳: نمودار جعبه‌ای مقادیر MSPE حاصل از روش‌های ۱:کریگینگ میانه‌پالایش، ۲: کریگینگ عام و ۳: کریگینگ معمولی با استفاده از اسپلاین.

فصل ۴

رهیافت بیزی تقریبی برای مدل‌بندی فضایی سه بعدی

۱.۴ مقدمه

آمار بیزی یک روش عقلانی در به‌کارگیری عدم قطعیت حالت‌های طبیعت به بیان احتمال است. در چارچوب فکری آمار بیزی، میزان باورها به حالت طبیعت مبنای کار است. این میزان از باورها با انتخاب باورهای پیشین قبل از مشاهده‌ی داده آغاز و با به‌کار گرفتن داده‌های به‌دست آمده از مشاهده یا آزمایش به روز می‌شوند و از آن با عنوان باورهای پسین یاد می‌کنیم. در روش‌های بیزی تحلیل بر باورهای پسین استوار است (گلمن و همکاران، ۲۰۱۳).

در استنباط مبتنی بر درست‌نمایی، پارامترها[1] ثابت در نظر گرفته می‌شوند که مقادیر آن‌ها ناشناخته است و به‌طور مستقیم نمی‌توان مقدار آن را اندازه‌گیری کرد. در استنباط بیزی، پارامترها متغیرهای تصادفی بدون مشاهده (و غیرقابل مشاهده) در نظر گرفته می‌شوند (دیگل و جورجی، ۲۰۱۹).

روش‌های MCMC به‌طور گسترده برای استنباط بیزی استفاده می‌شوند، اما محدودیت آن‌ها در بار محاسباتی‌اشان است. پیشرفت‌های موجود در جمع‌آوری داده‌ها، منجر به در

[1] ترجمه تحت اللفظی یونان باستان از پارامتر اصطلاح "Beyond measurement" به معنای ''فراتر از اندازه‌گیری'' است.

دسترس بودن مجموعه داده‌های بزرگ[۲] شده است که با موقعیت‌های تعیین شده‌ی فضایی، فضایی–زمانی و همچنین با منابع مختلف مشخص می‌شوند. پیچیدگی مدل برای در نظر گرفتن ساختارهای مکانی داده‌های بزرگ می‌تواند منجر به چندین روز زمان محاسبات برای انجام استنباط بیزی از طریق الگوریتم MCMC شود. به‌ویژه زمانی که بعد میدان پنهان[۳] بزرگ باشد، همچنین وجود همبستگی بین عناصر میدان پنهان و ویژگی‌های همگرایی و درهم‌تنیدگی، این الگوریتم‌ها را با مشکل روبرو می‌کنند. برای غلبه بر این چالش، رو و همکاران در سال ۲۰۰۹ در مقاله‌ی خود یک روش با نام تقریب لاپلاس آشیانه‌ای جمع‌بسته ارائه دادند. آن‌ها با استفاده از این الگوریتم ثابت کردند که می‌توانند به جواب‌های سریع و دقیقی دست یابند. الگوریتم روش INLA ابتدا به‌صورت یک برنامه مستقل شروع به‌کار کرد، سپس در R (به عنوان بسته‌ای با نام R-INLA) مورد دسترسی قرار گرفت و از آن پس، بین آماردان‌ها و محققان به‌ویژه در زمینه آمار فضایی از محبوبیت زیادی برخوردار شد. سایت http://www.r-inla.org یک سایت عالی و فعال در زمینه‌ی روش INLA است که مقالات و آموزش‌ها را در اختیار کاربران قرار می‌دهد و در تالار گفتگو محققان می‌توانند سوالات و مشکلات خود را برای بررسی و پاسخگویی، مطرح کنند (بلانگیادو و کملتی، ۲۰۱۵).

طی چندین سال، محققانی همچون گیلز و همکاران (۱۹۹۶) و بروکس و همکاران (۲۰۱۱) از رویکرد MCMC برای محاسبه‌ی توزیع پسین توأم پارامترهای مدل استفاده می‌کردند. اما این روش بار محاسباتی بسیار بالایی دارد چرا که توزیع در ابعاد بالاتر محاسبه می‌شود.

اولین بار رو و همکاران (۲۰۰۹) INLA را پیشنهاد دادند که در استنباط بیزی سریع‌تر عمل می‌کند. آن‌ها متذکر شدند که به‌جای برآورد توزیع پسین توأم پارامترهای مدل، می‌توان روی توزیع‌های پسین کناری پارامترهای مدل متمرکز شد. در بسیاری از موارد، استنباط کناری برای دانش از پارامترهای مدل و اثرات پنهان کافی است و نیازی به توزیع چندمتغیره پسین توأم نیست. در ادامه آن‌ها توجه خود را به مدل‌هایی معطوف کردند که می‌توانند به‌صورت میدان‌های گوسی تصادفی مارکوفی[۴] (GMRF) بیان شوند (گومز، ۲۰۲۰). این موضوع زمان محاسبات مدل را کاهش می‌دهد (رو و هلد، ۲۰۰۵). INLA در رده گسترده‌ای از مدل‌ها به نام مدل‌های گوسی پنهان[۵] (LGMs) کاربرد دارد. عضویت در این رده مستلزم این است که برخی از پارامترهای موجود در مدل دارای توزیع گوسی باشند (وانگ و همکاران، ۲۰۱۸). به دنبال پیشنهاد این رویکرد جدید، مارتینز و همکاران (۲۰۱۳) و اخیراً رو و همکاران (۲۰۱۷) مقالاتی را در این رابطه ارائه داده‌اند.

---

[۲] Big datasets
[۳] Latent field
[۴] Gaussian Markov random field
[۵] Latent Gaussian Models

۲.۴ مدل‌های گوسی پنهان

مدل‌های گوسی پنهان زیرمجموعه‌ای از مدل‌های بیزی سلسله‌مراتبی هستند. در این نوع از مدل‌ها، پارامتر در سطح نهفته و دارای توزیع پیشین توأم به شرط ابرپارامترها است. رده LGMها یک زیررده از مدل‌های رگرسیونی با ساختار جمعی هستند. رو و همکاران در پژوهش خود فرض کردند توزیع مشاهده‌ها y_i، برای $i = ۱, \ldots, n$، از یک خانواده و اغلب از خانواده توزیع‌های نمایی تبعیت می‌کنند. میانگین یا یک چندک مشخص مثل μ_i، و پیشگوی جمعی مدل رگرسیونی η_i از طریق تابع پیوند $g(\cdot)$ به‌شکل $g(\mu_i) = \eta_i$ مدل‌بندی می‌شود. در واقع

$$\eta_i = \alpha + \sum_{j=۱}^{J} f_j(\mathbf{u}_{i,j}) + \sum_{k=۱}^{K} \beta_k z_{i,k} + \varepsilon_i \qquad (۱.۴)$$

که در آن α عرض از مبدا، $\{f_j\}_{j=۱}^{J}$ اثرات ناپارامتری نامعلوم با برخی وابستگی‌های ساختاری خاص با وزن‌های مشخص $\mathbf{u}$ و ضرایب $\{\beta_k\}_{k=۱}^{K}$ اثرات ثابت خطی متغیرهای تبیینی $z_{i,۱}, \ldots, z_{i,K}$ روی متغیر پاسخ است و $\varepsilon_۱, \ldots, \varepsilon_n$ خطاهای غیرساختاری مدل هستند. گومز در کتاب خود در سال ۲۰۲۰، اشاره کرده است که تابع درستنمایی مشاهدات y_iها، باید باهم مرتبط باشند و الزامی در اینکه از خانواده‌ی توزیع‌های نمایی باشند وجود ندارد. این رده از مدل‌ها، چون تابع $\{f_j\}$ ساختار نامعلومی دارد می‌تواند کاربردهای زیادی داشته باشد. مدل‌های گوسی پنهان، یک زیرمجوعه از تمام مدل‌های جمعی بیزی با ساختار پیشگوی جمعی است، یعنی پیشین گوسی به α، $\{\beta_k\}$، $\{f_j\}$ و $\{\epsilon_i\}$ اختصاص داده می‌شود. در چارچوب INLA بردار اثرات پنهان $\mathbf{x}$ به‌صورت زیر نمایش داده می‌شود که دارای ابرپارامترهای پنهان است:

$$\mathbf{x} = \{\eta_i, \alpha, \beta_k, f_j\}.$$

پارامترها در $\mathbf{x}$ دارای توزیع توأم گوسی هستند که روی ابرپارامترهای $\boldsymbol{\theta} = (\theta_۱, \ldots, \theta_p)^T$ شرطی شده است. توزیع پسین توأم اثرات پنهان $\mathbf{x}$ و ابرپارامترهای $\boldsymbol{\theta}$ را می‌توان به‌صورت زیر نوشت:

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

در معادله‌ی قبلی، $\pi(\mathbf{y})$ درستنمایی کناری مدل را نشان می‌دهد که یک ثابت نرمال‌ساز است و اغلب نادیده گرفته می‌شود، چرا که محاسبه‌ی آن معمولاً دشوار است. با این حال INLA، یک تقریب دقیق از آن را ارائه می‌دهد (هوبین و استورویک، ۲۰۱۶). ابرپارامترها معمولاً دارای چگالی پسین گوسی نیستند و اغلب واریانس کناری، دامنه همبستگی و همواری یا سایر پارامترهای همبستگی اثرات تصادفی را مشخص می‌کنند. به‌صورت شماتیک LGMها با تابع پیوند یک متغیره ممکن است از نظر سطح داده، سطح پنهان و سطح ابرپارامترها به گونه‌ی سلسله‌مراتبی نشان داده شوند (رانکسن و همکاران، ۲۰۲۰).

سطوح داده: فرض کنید مشاهدات $\mathbf{y} = (y_۱, \ldots, y_n)^T$ به پارامترهای پنهان $\mathbf{x}$ وابسته و دارای چگالی

$\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})$ است. برای سادگی فرض می‌شود با استقلال شرطی، توزیع n مشاهده به وسیله‌ی درستنمایی زیر محاسبه می‌شود:

$$\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i|\mathbf{x},\boldsymbol{\theta}) \qquad (۲.۴)$$

به‌طوری که $\pi(y_i|\mathbf{x},\boldsymbol{\theta}) = \pi(y_i|\eta_i,\boldsymbol{\theta})$ با $\mu_i = \mathbf{Q}_\pi(y_i|\eta_i) = g^{-1}(\eta_i)$ که در آن μ_i میانگین یا یک چندک مناسب از تابع چندک $\mathbf{Q}_p$ است. در رابطه‌ی (۲.۴) هر عضو از y_i تنها به یک عضو از x_i از میدان پنهان $\mathbf{x}$ مرتبط است.

سطوح پنهان: پارامتر پنهان $\mathbf{x}$، دارای چگالی پیشین گوسی است و به‌طور بالقوه به برخی از ابرپارامترها وابسته است. تابع چگالی $\mathbf{x}$ ممکن است به‌صورت

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = N(\mathbf{x}|\mu(\boldsymbol{\theta}),\boldsymbol{\Sigma}(\boldsymbol{\theta}))$$

نوشته شود. در رابطه‌ی فوق، عبارات سمت راست چگالی گوسی با بعد برابر بعد $\mathbf{x}$ با بردار میانگین $\mu(\boldsymbol{\theta})$ و ماتریس کوواریانس $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ را نشان می‌دهد (گلمن و همکاران، ۲۰۱۳).

سطوح ابرپارامترها: ابرپارامترهای $\boldsymbol{\theta}$، دارای چگالی پیشین $\pi(\boldsymbol{\theta})$ هستند.

فرض می‌کنیم $\mathbf{x}$ دارای توزیع چندمتغیره گوسی با بردار میانگین ۰ و ماتریس دقت $Q(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$ با ابرپارامتر $\boldsymbol{\theta}$ باشد، یعنی

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = (۲\pi)^{-\frac{n}{۲}} |Q(\boldsymbol{\theta})|^{\frac{۱}{۲}} \exp(-\frac{۱}{۲}\mathbf{x}^T Q(\boldsymbol{\theta})\mathbf{x}) \qquad (۳.۴)$$

به‌طوری که مولفه‌های میدان گوسی $\mathbf{x}$ به‌طور شرطی مستقل هستند. از این موضوع می‌توان نتیجه گرفت که ماتریس دقت $Q(\boldsymbol{\theta})$، تُنُک[۶] است. رو و هلد (۲۰۰۵) این خصوصیت را GMRF تعریف کردند. زمانی که استنباط با استفاده از GMRF صورت گیرد، تُنُک بودن ماتریس دقت به کم شدن زمان محاسبات می‌انجامد.

توزیع پسین $\mathbf{x}$ و $\boldsymbol{\theta}$ از ضرب درستنمایی (۲.۴) و چگالی GMRF (۳.۴) و توزیع پیشین ابرپارامترها به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned}
\pi(\mathbf{x},\boldsymbol{\theta}|y) &\propto \pi(\boldsymbol{\theta}) \times \pi(\mathbf{x}|\boldsymbol{\theta}) \times \pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) \\
&\propto \pi(\boldsymbol{\theta}) \times \pi(\mathbf{x}|\boldsymbol{\theta}) \times \prod_{i=1}^{n} \pi(y_i|x_i,\boldsymbol{\theta}) \\
&\propto \pi(\boldsymbol{\theta}) \times (۲\pi)^{-\frac{n}{۲}} |Q(\boldsymbol{\theta})|^{\frac{۱}{۲}} \exp\left(-\frac{۱}{۲}\mathbf{x}^T Q(\boldsymbol{\theta})\mathbf{x}\right) \times \prod_{i=1}^{n} \exp(\log(\pi(y_i|x_i,\boldsymbol{\theta}))) \\
&\propto \pi(\boldsymbol{\theta}) \times |Q(\boldsymbol{\theta})|^{\frac{۱}{۲}} \exp\left(-\frac{۱}{۲}\mathbf{x}^T Q(\boldsymbol{\theta})\mathbf{x} + \sum_{i=1}^{n} \log(\pi(y_i|x_i,\boldsymbol{\theta}))\right).
\end{aligned}$$

[۶] Sparse

۳.۴ تقریب لاپلاس آشیانه‌ای جمع بسته

هدف استنباط بیزی، تقریب پسین کناری اعضای میدان گوسی پنهان $\mathbf{x}$

$$\pi(x_i|\mathbf{y}) = \int \pi(x_i, \boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \int \pi(x_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \quad (۴.۴)$$

و اعضای بردار ابرپارامتر

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j} \quad (۵.۴)$$

است، که در آن $\boldsymbol{\theta}_{-j}$ بردار ابرپارامترها در غیاب عضو jام θ_j است. این تقریب در سه گام انجام می‌شود.

گام اول: ابتدا تقریب پسین کناری $\boldsymbol{\theta}$ با استفاده از تقریب لاپلاس محاسبه می‌شود. رو و همکاران (۲۰۰۹) معتقدند اساس رویکرد INLA از تقریب پسین کناری $\boldsymbol{\theta}$ یعنی $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ پیروی می‌کند.

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})}\Bigg|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \quad (۶.۴)$$

که در آن، $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ تقریب گوسی برای چگالی شرطی کامل $\mathbf{x}$ و $\mathbf{x}^*(\boldsymbol{\theta})$ مقدار مُد شرطی $\mathbf{x}$ برای مقادیر داده‌شده $\boldsymbol{\theta}$ است. تقریب چگالی(شرطی) θ_j که $j = ۱, \ldots, J$ به‌صورت

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}$$

است.

گام دوم: تقریب لاپلاس پسین کناری اعضای میدان $\mathbf{x}$ را محاسبه می‌کنیم. به‌طور ساده‌تر، تقریب لاپلاس $\tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})$ برای $\boldsymbol{\theta}$های تعیین‌شده به‌صورت زیر محاسبه می‌شود:

$$\tilde{\pi}(x_i|\mathbf{y}) = \int \tilde{\pi}(x_i, \boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} = \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}.$$

گام سوم: در آخرین گام، دو مرحله‌ی قبل را با استفاده از انتگرال‌گیری عددی با هم ترکیب کرده و انتگرال توزیع پسین میدان پنهان را با استفاده از جمع متناهی به‌دست می‌آوریم. در واقع داریم

$$\tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y}) = \mathrm{N}\{x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^{۲}(\boldsymbol{\theta})\}$$

که در آن $\mu(\boldsymbol{\theta})$ (بردار) میانگین تقریب گوسی و $\sigma^{۲}(\boldsymbol{\theta})$ (بردار) واریانس کناری متناظر است. در نهایت

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_{k=۱}^{n} \tilde{\pi}(x_i|\theta_k, \mathbf{y})\tilde{\pi}(\theta_k|\mathbf{y})\Delta_k \quad (۷.۴)$$

که در آن Δ_k وزن‌هایی هستند که به θ_k نسبت داده می‌شود. رو و مارتینو (۲۰۰۷) نشان دادند که تقریب کناری پسین $\boldsymbol{\theta}$ دقیق است.

۱.۳.۴ تقریب چگالی پسین توأم بردار ابرپارامترها

چگالی پسین توأم بردار ابرپارامترها به‌صورت

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}) &= \frac{\pi(\boldsymbol{\theta},\mathbf{y})}{\pi(\mathbf{y})}\\ &= \frac{\pi(\boldsymbol{\theta},\mathbf{y})}{\pi(\mathbf{y})}\frac{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\\ &= \frac{\pi(\mathbf{x},\boldsymbol{\theta},\mathbf{y})}{\pi(\mathbf{y})\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\\ &= \frac{\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\\ &\propto \frac{\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\end{aligned} \tag{۸.۴}$$

محاسبه می‌شود. تقریب لاپلاس آشیانه‌ای جمع بسته‌ی معادله‌ی (۸.۴) برای هر مقدار $\boldsymbol{\theta}$ که پیش‌تر در معادله‌ی (۶.۴) به آن اشاره شد به شکل

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \left.\frac{\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\pi(\mathbf{x}|,\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \tag{۹.۴}$$

به‌دست می‌آید. در معادله‌ی (۹.۴) مخرج کسر از توزیع گوسی حاصل می‌شود. با استفاده از بسط تیلور در تقریب گوسی، توزیع شرطی کامل

$$\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y}) \propto \left(-\frac{۱}{۲}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \sum_{i\in\mathcal{I}}\log\pi(y_i|x_i,\boldsymbol{\theta})\right)$$

حول مُد توزیع شرطی کامل $\mathbf{x}$، یعنی $\mathbf{x}^*(\boldsymbol{\theta}) = (x_۱^*(\boldsymbol{\theta}),\ldots,x_m^*(\boldsymbol{\theta}))$ به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned}\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta},\mathbf{y}) &\propto \left(-\frac{۱}{۲}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \sum_{i\in\mathcal{I}} g_i(x_i)\right)\\ &\propto \left(-\frac{۱}{۲}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \sum_{i\in\mathcal{I}}(a_i + b_i x_i - \frac{۱}{۲}c_i x_i^۲)\right)\\ &\propto \left(-\frac{۱}{۲}\mathbf{x}^T(\mathbf{Q}+\mathrm{diag}(\mathbf{c}))\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{x}\right)\end{aligned} \tag{۱۰.۴}$$

که در آن $c_i = -g''(x_i^*(\boldsymbol{\theta}))$، $b_i = g'(x_i^*(\boldsymbol{\theta})) + x_i^*(\boldsymbol{\theta})c_i$ و a_i از تناسب حذف می‌شود. در رابطه‌ی (۱۰.۴) مُد با استفاده از روش نیتون–رافسون[۷] که با نام الگوریتم امتیازدهی فیشر[۸] نیز شناخته می‌شود، محاسبه می‌شود (فارمر و تاتز، ۲۰۰۱). اگر $\boldsymbol{\mu}^{(۰)}$ حدس اولیه باشد، $g_i(x_i)$ حول $\mu_i^{(۰)}$

[۷] Newton–Raphson method

[۸] Fisher scoring algorithm

به مرحله دوم بسط داده می‌شود و می‌توان نوشت

$$g_i(x_i) \approx g_i(\mu_i^{(0)}) + b_i x_i - \frac{1}{2} c_i x_i^2$$

که $\{b_i\}$ و $\{c_i\}$ به $\boldsymbol{\mu}^{(0)}$ وابسته هستند. این یک تقریب گوسی بهینه با ماتریس دقت $\mathbf{Q}+\text{diag}(\mathbf{c})$ است و مُد از حل رابطه‌ی $\mathbf{Q}+\text{diag}(\mathbf{c})\boldsymbol{\mu}^{(1)} = \mathbf{b}$ به‌دست می‌آید. این فرآیند تا زمانی که به یک توزیع پنهان با میانگین $\mathbf{x}^*$ و ماتریس دقت $\mathbf{Q}^* = \mathbf{Q}+\text{diag}(\mathbf{c}^*)$ همگرا شود، تکرار می‌شود (رو و همکاران، ۲۰۰۹).

تقریب لاپلاس چگالی پسین $\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{x},\boldsymbol{\theta},\mathbf{y})}{\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}$ با بسط مخرج کسر حول $\mathbf{x} = \mathbf{x}^*(\boldsymbol{\theta})$ به‌صورت

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

به‌دست می‌آید.

در سه مرحله از فرآیند اجرای روش INLA، از تقریب $\tilde{\pi}_{LA}(\boldsymbol{\theta}|\mathbf{y})$ استفاده می‌شود. مهمترین و اولین کاربرد آن در به‌دست آوردن تقریب بیزی پسین میدان پنهان است. در مرحله‌ی بعد از رابطه‌ی (۹.۴) برای محاسبه‌ی تقریب کناری درستنمایی $\pi(\mathbf{y})$ استفاده می‌شود. در نهایت از این تقریب در جهت محاسبه‌ی چگالی کناری پسین بعضی از اعضای ابرپارامترهای $\boldsymbol{\theta}$، یعنی $\pi(\theta_i|\mathbf{y})$ استفاده می‌شود. باید توجه داشت که با انتگرال‌گیری روی محدوده‌ی چند بعدی $\boldsymbol{\theta}$ می‌توان از (۹.۴) برای محاسبه‌ی سه مرحله‌ی مذکور اقدام کرد. پس ضرورت انتخاب نقاط مناسب انتگرال‌گیری کاملاً مشهود است. رو و همکاران (۲۰۰۹) دو رویکرد توری[۹] و طرح مرکب مرکزی[۱۰] (CCD) را برای انتخاب نقاط انتگرال‌گیری مناسب پیشنهاد دادند. رویکرد توری، مشتمل از نقاط یک توری است که $\tilde{\pi}_{LA}(\boldsymbol{\theta}|\mathbf{y})$ چگالی بیشتری در آن نقاط دارد. زمانی که بعد $\boldsymbol{\theta}$ بزرگ نباشد (مثلاً کمتر از ۶)، این روش بسیار دقیق و مناسب است. وقتی بعد $\boldsymbol{\theta}$ بین ۶ تا ۱۲ باشد، استفاده از رویکرد CCD مفید است. در هر دو رویکرد، $\tilde{\pi}_{LA}(\boldsymbol{\theta}|\mathbf{y})$ تک مُدی فرض می‌شود. در فرآیند انتگرال‌گیری به مُد $\tilde{\pi}_{LA}(\boldsymbol{\theta}|\mathbf{y})$ نیاز است و با نماد $\boldsymbol{\theta}^*$ نشان داده می‌شود. با توجه به پیچیدگی تابع مورد نظر می‌توان از انتگرال‌گیری بهینه‌سازی نیوتن–رافسون یا شبه‌نیوتن[۱۱] برای بهینه کردن $\boldsymbol{\theta}^*$ استفاده کرد.

۲.۳.۴ روش توری

روش توری طی سه گام انجام می‌شود که در ادامه به بررسی آن‌ها می‌پردازیم.

گام اول: مقدار $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ را نسبت به $\boldsymbol{\theta}$ بهینه می‌کنیم تا مقدار مُد $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ محاسبه شود. این کار را می‌توان با روش نیوتن–رافسون انجام داد.

[۹] Grid

[۱۰] Central composite design

[۱۱] Quasi-Newton

گام دوم: ماترس منفی هسین با استفاده از تفاضل‌های متناهی در پیکره‌ی $\boldsymbol{\theta}^*$ محاسبه می‌شود. با فرض $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$، اگر چگالی گوسی را داشته باشیم، آنگاه $\boldsymbol{\Sigma}$ ماتریس کوواریانس $\boldsymbol{\theta}$ خواهد بود. اگر ماتریس $\boldsymbol{\Sigma}$ به‌صورت $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}$ تجزیه شود، برای ساده‌سازی متغیر استاندارد $\mathbf{z}$ به‌جای $\boldsymbol{\theta}$ به‌صورت

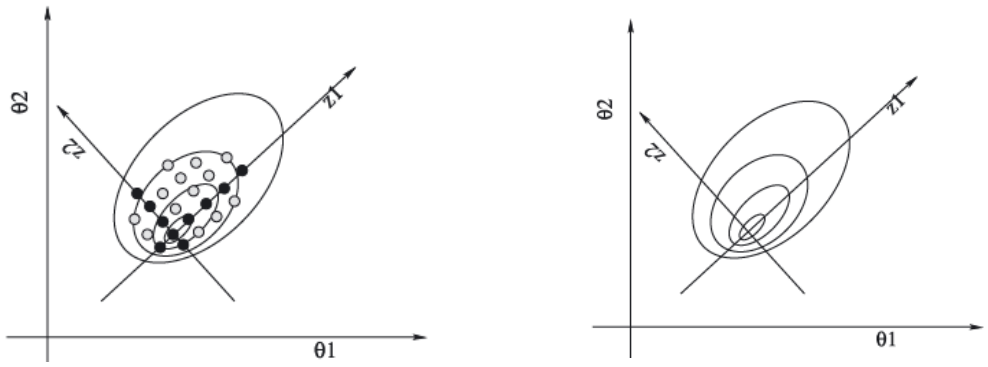
$$\boldsymbol{\theta}(\mathbf{z}) = \boldsymbol{\theta}^* + \mathbf{V}\boldsymbol{\Lambda}^{1/2}\mathbf{z}$$

استفاده می‌شود.

گام سوم: بررسی $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ با استفاده از متغیر $\mathbf{z}$. برای توضیح ساده و قابل درک، شکل ۱.۴ نمودار تراز $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ را وقتی $m = 2$ است برای مُد و مختصات جدید نشان می‌دهد. در واقع هدف کاوش $\log\{\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\}$ برای یافتن بیشترین چگالی احتمال است. برای این‌کار از نقطه‌ی مُد $\mathbf{z} = 0$ شروع می‌کنیم. سپس یک گام در راستای مثبت z_1 به طول δ_z برمی‌داریم تا جایی که شرط

$$\log[\{\tilde{\pi}(\boldsymbol{\theta}(0)|\mathbf{y})\}] - \log[\{\tilde{\pi}(\boldsymbol{\theta}(\mathbf{z})|\mathbf{y})\}] < \delta_\pi \qquad (۱۱.۴)$$

برقرار شود. برای z_2 نیز به همین ترتیب رفتار می‌کنیم. با این کار نقاط سیاه به‌وجود می‌آیند. اکنون از همین روش می‌توان تمام نقاط میانی (نقاط خاکستری) را با ترکیبات مختلف نقاط سیاه ایجاد کرد و مختصات جدید تولید می‌شوند. رو و همکاران در پژوهش خود، $\delta_z = 1$ و $\delta\pi = 2/5$ را در نظر گرفتند (رو و همکاران، ۲۰۰۹).



(ب) (آ)

شکل ۱.۴: (آ) مکان مُد، و سیستم مختصات $\mathbf{z}$، (ب) نمایش نقاط منتخب انتگرال‌گیری با راهبرد توری

۳.۳.۴ روش طرح مرکب مرکزی

انتخاب نقاط انتگرال‌گیری در روش توری، وقتی تعداد ابرپارامترها، m، زیاد می‌شود خیلی پرهزینه است، حتی اگر تعداد ابرپارامترها متوسط (۶ تا ۱۲) باشد. مثلاً اگر چگالی $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{z})$

گوسی باشد و $\delta_z = ۱$ و $\delta_\pi = ۲/۵$ آن‌گاه به $\mathcal{O}(۵^m)$ نقطه برای به‌دست آوردن (۷.۴) نیاز است و اگر این روش به سه نقطه در هر بعد محدود شود به $\mathcal{O}(۳^m)$ نقطه نیاز است که از مرتبه‌ی نمایی است (رو و همکاران، ۲۰۰۹).

در این بخش رویکرد CCD شرح داده می‌شود که هزینه‌ی محاسبات را برای mهای بزرگ کاهش می‌دهد. روش CCD به عنوان مساله‌ی طرح رویه‌ی پاسخ در نظر گرفته می‌شود و نقاط روی فضای m بعدی پخش می‌شوند و بر اساس پاسخ‌های به‌دست آمده در آن نقاط خمیدگی پسینی حول مُد برآورد می‌شود. در این بخش، تنها سطوحِ پاسخ از مرتبه دو را در نظر گرفته و از طرح درجه دو CCD استفاده می‌شود (باکس و ویلسون، ۱۹۵۱). در اغلب زمینه‌های پژوهشی مختلف، پژوهشگران برای کشف اطلاعاتی درباره‌ی فرآیند سیستم‌های خاص، آزمایش‌هایی را انجام می‌دهند. آزمایش‌گر برای تشخیص عوامل تغییرات در پاسخ، باید عوامل متغیر در آزمایش، دامنه‌ی تغییرات هر عامل و سطوح خاصی که در هر بار اجرا باید مورد استفاده قرار گیرند را تعیین کند. طرح یک آزمایش، ابزار مهمی برای تعیین عملکرد درست فرآیند تولید است. در علم آمار منظور از طرح آزمایش آماری این است که فرآیندی را طراحی کنند که بتوان با آن داده‌های مناسبی به روش‌های مختلف آماری جمع‌آوری و تحلیل کرد.

طرح ۲^k عاملی، طرح‌های k عاملی هستند که هر عامل فقط از دو سطح کمی یا کیفی برخوردار است. طرح عاملی V-تجزیه[۱۲]، به طرح‌های عاملی می‌گویند که اثرهای اصلی و اثرهای متقابل دوعاملی با اثرهای اصلی و اثرهای متقابل در عامل دیگری با یکدیگر، هم‌اثر نباشند. اگر تعداد عامل‌ها در این نوع طرح‌ها زیاد باشد، اجرا از حیث تعداد و تعیین نقاط طرح کار دشواری است. بزرگترین طرح عاملی V-تجزیه، $۲^{۱۰-۳}$ برابر با $۲^{۱۱-۴}$ است (مونتگومری، ۲۰۰۰؛ باکس و همکاران، ۱۹۷۸). سانچز و سانچز (۲۰۰۵)، با استفاده از ماتریس مرتب‌سازی شده‌ی هادامارد[۱۳] (بیوچمپ، ۱۹۸۴) و اندیس والش[۱۴]، راهکار یادگیری را برای تعیین طرح‌های عاملی کسری[۱۵]، حتی تا ۱۲۰ عامل، ارائه دادند که به‌صورت مختصر در ادامه بیان می‌شود.

ماتریس هادامارد H_k، یک ماترس $۲^k \times ۲^k$ با درایه‌های $+۱$ یا -۱ است که می‌توان آن را با روابط بازگشتی به‌صورت

$$H_۱ = [۱], \qquad H_۲ = \begin{bmatrix} ۱ & ۱ \\ ۱ & -۱ \end{bmatrix}, \qquad H_{k+۱} = \begin{bmatrix} H_k & H_k \\ H_k & -H_k \end{bmatrix}, \tag{۱۲.۴}$$

محاسبه کرد. اگر ستون‌های H_k با $h_۰, h_۱, \ldots, h_{N-۱}$ تعیین شوند که در آن $N = ۲$، هر ستون h_i به اثراصلی یا اثر متقابل دو عاملی با استفاده از الگوریتم بازگشتی برای اندیس‌گذاری والش

[۱۲] V-Resoulation
[۱۳] Hadamard ordered matrix
[۱۴] Walsh index
[۱۵] Fractional factorial design

طرح عاملی V-تجزیه کسری تجزیه می‌شود. در k مرحله این اندیس‌گذاری انجام می‌شود که در مرحله‌ی iام آن، اندیس عامل x_i تعیین می‌شود. اگر اندیس والش عامل‌های $x_{j۱}$ و $x_{j۲}$ به ترتیب $a_۱$ و $a_۲$ باشند، اندیس اثر متقابل بین این دو عامل، یعنی اثر $x_{j۱}x_{j۲}$، از رابطه‌ی $a_۱ \oplus a_۲$ محاسبه می‌شود، که در آن بُعد از تغییر مبنای اعداد به دو، پای انحصاری $\oplus$، با استفاده از جدول ۱.۴ محاسبه می‌شود.

جدول ۱.۴: نقاط طرح هادامارد والش برای $k = ۳$

A	B	A $\oplus$ B
۰	۰	۰
۱	۰	۱
۰	۱	۱
۱	۱	۰

به عنوان مثال $۷ \oplus ۴ = (۱۱۱)_۲ \oplus (۱۰۰)_۲ = (۰۱۱)_۲ = (۱ \times ۲^۰) + (۱ \times ۲^۱) + (۰ \times ۲^۳) = ۳$. در انتهای مرحله‌ی iام، $i = ۱, \ldots, k$ دو مجموعه‌ی A_i و T_i به‌صورت‌های

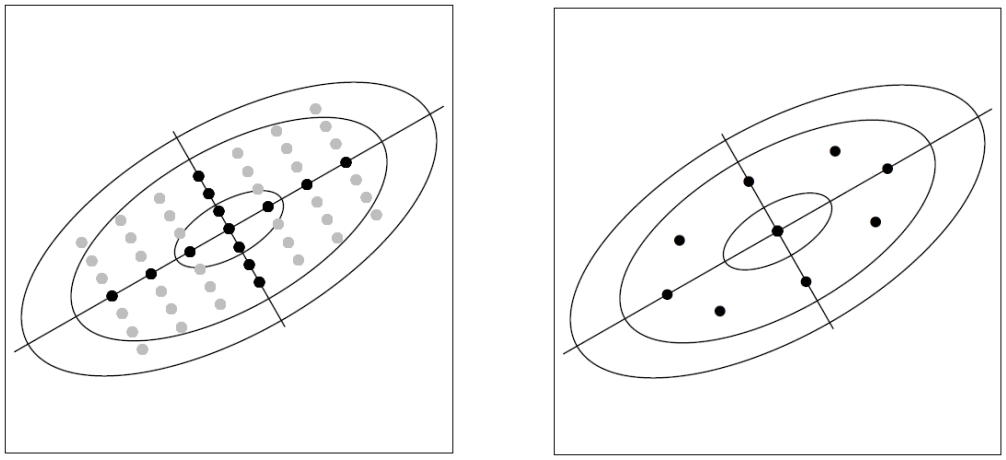
$$A_i = \{\text{اندیس والش اثرهای اصلی عامل‌های } X_۱, \ldots, X_k\}$$
$$T_i = \{\text{اثرهای متقابل دو عاملی بین مولفه‌های اندیس والش متناظر } A_i\}$$

به‌دست می‌آید که $A_۰$ و $T_۰$ تهی فرض می‌شوند. برای اعتبار اندیس والش، اندیس پیشنهاد شده‌ی عامل X_j که با b نشان داده می‌شود، باید دارای دو شرط زیر باشد:

۱. نباید متعلق به $A_{j-۱}$ و $T_{j-۱}$ باشد.

۲. اندیس اثرهای متقابل دو عاملی بین x_j و $x_۱, \ldots, x_{j-۱}$ نباید متعلق به A_j و T_j باشند.

در شکل ۲.۴ نقاط منتخب انتگرال‌گیری به روش‌های CCD و روش توری نشان داده شده‌اند.



(آ) (ب)

شکل ۲.۴: نقاط منتخب انتگرال‌گیری بااستفاده از (آ) روش CCD و (ب) روش توری

الگوریتم تعیین اندیس والش در ادامه ذکر شده‌است.

الگوریتم ۱.۳.۴. تعیین اندیس والش

۱. قرار می‌دهیم $m = ۱$.

۲. $b = \max\{i : i \in A_{m-۱}\}$.

۳. تا وقتی $b \in T_{m-۱}$ یا $\{\exists j \in A_{m-۱} : b \oplus i \in A_{m-۱} \cup T_{m-۱}\}$ قرار می‌دهیم $b = b + ۱$.

۴. دو مجموعه‌ی $A_m = \{b\} \cup A_{m-۱}$ و $T_m = T_{m-۱} \cup \{i \oplus b : i \in A_m\}$ را محاسبه می‌کنیم.

۵. قرار می‌دهیم $m = m + ۱$ و تا وقتی $m < k$ به گام ۲ برمی‌گردیم.

بعد از اندیس‌گذاری به روش والش، هر ستون ماتریس هادامارد متعلق به یک اثر اصلی یا اثر متقابل دو عاملی خواهد بود و هر سطر ماتریس مشخص‌کننده‌ی سطوح عامل در یک اجرا خواهد بود که به آن نقاط طرح می‌گویند.

جدول ۳.۴ نقاط طرح را برای $k = ۳$ نشان می‌دهد. در این طرح‌ها به تعداد نقاط غیرتکراری، طرح نیاز به اجرا دارد.

جدول ۲.۴: نقاط طرح هادامارد والش برای $k = ۳$

نماد طرح	۶ : $X_۲X_۳$	۵ : $X_۱X_۳$	۴ : $X_۳$	۳ : $X_۱X_۲$	۲ : $X_۲$	۱ : $X_۱$
۱	+۱	+۱	+۱	+۱	+۱	+۱
۲	+۱	−۱	+۱	−۱	+۱	−۱
۳	−۱	+۱	+۱	−۱	−۱	+۱
۴	−۱	−۱	+۱	+۱	−۱	−۱
۵	−۱	−۱	−۱	+۱	+۱	+۱
۶	−۱	+۱	−۱	−۱	+۱	−۱
۷	+۱	−۱	−۱	−۱	−۱	+۱
۸	+۱	+۱	−۱	+۱	−۱	−۱

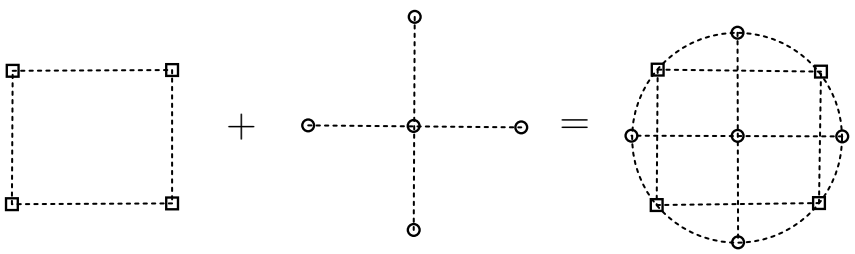
جدول ۳.۴: تعداد اجرا در طرح عاملی کسری

بعد θ	۱-۲	۳	۴-۵	۶	۷-۸	۹-۱۱	۱۲-۱۷
تعداد نقاط	۴	۸	۱۶	۳۲	۶۴	۱۲۸	۲۵۶

تعیین نقاط و تعداد نقاط در طرح مکعب مرکزی

طرح مرکب مرکزی، طرح عامل کسری را در خود دارد که با یک نقطه‌ی مرکزی و یک گروه متشکل از $۲m$ نقطه با عنوان ستاره‌ای ترکیب شده است. بنابراین، تعداد k نقطه از مجموع تعداد طرح و $۲m + ۱$ نقطه جدید به‌دست می‌آیند. مثلاً در طرح CCD، تعداد نقاط مورد نیاز برای $m = ۵$ برابر $۲۷ = ۱ + (۲ \times ۵) + ۱۶$ است که در مقایسه با تعداد نقاط مورد نیاز در روش توری $۳۱۲۵ = ۵^۵$ یا $۲۴۳ = ۳^۵$ بسیار اندک است.

در این روش، تمام نقاط روی سطح کروی m بعدی به شعاع $\sqrt{m}$ قرار می‌گیرند. نقاط ستاره‌ای در فاصله‌ی $\pm\sqrt{m}$ روی محورها و نقطه‌ی مرکزی در مبدا قرار می‌گیرند. شکل ۳.۴ موقعیت نقاط را برای CCD وقتی $m = ۲$ است را نشان می‌دهد.



شکل ۳.۴: موقعیت نقاط در طرح مرکب مرکزی

برای مشخص شدن وزن‌های انتگرال در (۷.۴) فرض‌شده است که $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ گوسی استاندارد است و انتگرال $\boldsymbol{\theta}^T\boldsymbol{\theta}$ برابر m باشد. با این شرایط فرض‌شده، وزن انتگرال‌گیری روی نقاط کره به شعاع $f_\circ\sqrt{m}$ به‌صورت زیر محاسبه می‌شود:

$$\Delta_k = \left[(k-1)(f_\circ^2-1)\{1+\exp(-\frac{mf_\circ^2}{2})\}\right]^{-1}$$

که در آن $f_\circ > 1$. وزن انتگرال‌گیری نقطه‌ی مرکزی برابر $1-(k-1)\Delta_k$ است (باکس و ویلسون، ۱۹۵۱).

درون‌یابی گوسی نامتقارن

بعضی از چگالی‌های کناری $\pi(\theta_k|\mathbf{y})$ را می‌توان با تقریب توزیع توأم $\pi(\boldsymbol{\theta}|\mathbf{y})$ با یک توزیع نرمال چند متغیره $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ محاسبه کرد. تقریب گوسی $\pi(\theta_k|\mathbf{y})$ محاسبات اضافی ندارد، زیرا مُد $\boldsymbol{\theta}^*$ و ماتریس منفی هسین H با استفاده از (۷.۴) تقریب شده است.

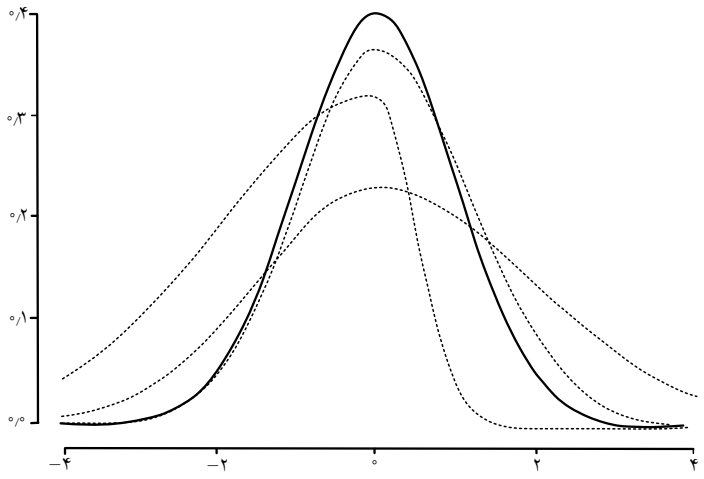
اگر $\pi(\theta_k|\mathbf{y})$ چوله باشد، در این صورت تقریب گوسی مناسب نیست. در این حالت، برای تصحیح تقریب گوسی با هزینه‌ی کمتر به شکل زیر عمل می‌شود. تابع $f(\boldsymbol{\theta})$ را وقتی با نقاط $\mathbf{z}(\boldsymbol{\theta}) = (z_1(\boldsymbol{\theta}), \ldots, z_k(\boldsymbol{\theta}))$ فرض‌شده در $\mathbf{z}$ باشند، می‌توان به‌صورت زیر تعریف کرد:

$$f(\boldsymbol{\theta}) = \prod_{k=1}^{m} f_k(z_k(\boldsymbol{\theta})) \qquad (۱۳.۴)$$

که در آن

$$f_k(z) \propto \begin{cases} \exp\left(-\frac{1}{2(\sigma^{k+})^2}z^2\right) & z \geq 0 \\ \exp\left(-\frac{1}{2(\sigma^{k-})^2}z^2\right) & z < 0 \end{cases} \qquad (۱۴.۴)$$

برای رفع نامتقارنی $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$، پارامترهای $(\sigma^{k+}, \sigma^{k-})$ که $k = 1, \ldots, m$، را به گونه‌ای تعریف می‌کنیم که بر پایه m محور مختلف تغییری نداشته باشد، اما بر اساس مسیر، جهت مثبت و منفی هر محور تغییر کند. برای محاسبه این توابع، ابتدا باید توجه داشته باشیم که در یک توزیع گوسی، در لگاریتم چگالی وقتی مد به ± 2 می‌رسد، انحراف معیار برابر -2 می‌شود. این پارامترها را در روشی که به‌طور دقیق برای تمام مسیرها تقریب زده می‌شوند، حساب می‌کنیم. سپس تقریب‌های $\pi(\theta_k|\mathbf{y})$ با انتگرال‌گیری عددی (۱۳.۴) محاسبه می‌شوند، به‌طوری که محاسبه‌ی آن برای پارامترهای معلوم است. شکل ۴.۴ انعطاف پذیری $f_k(z)$ را در معادله‌ی (۱۴.۴) برای نقاط مختلف σ^{k+} و σ^{k-} نشان می‌دهد.



شکل ۴.۴: توزیع نرمال استاندارد (خط ممتد) و چگالی معادله‌ی (۱۴.۴) برای مقادیر مختلف پارامتر مقیاس (خطوط خط‌چین)

الگوریتم انتگرال‌گیری عددی

تقریب چگالی‌های پسین کناری $\tilde{\pi}(\theta_k|\mathbf{y})$ با استفاده از الگوریتم انتگرال‌گیری عددی به‌صورت زیر است:

$$\tilde{\pi}(\theta_k|\mathbf{y}) = \begin{cases} N(۰, \sigma^۲_{k+}) & \sigma > ۰ \\ N(۰, \sigma^۲_{k-}) & \sigma \leq ۰. \end{cases} \qquad (۱۵.۴)$$

سپس این معادله برای محاسبه‌ی $\sigma^۲_{k+}$ و $\sigma^۲_{k-}$ بدون محاسبه با انتگرال‌گیری عددی به‌صورت الگوریتم قبلی، به‌کار می‌رود. لم زیر برای این عمل مفید است.

لم ۱.۳.۴. با فرض $\mathbf{x} = (x_۱, \ldots, x_n)^T \sim N(۰, \boldsymbol{\Sigma})$

$$-\frac{۱}{۲}(x_۱, E(\mathbf{x}_{-۱}|x_۱)^T)\boldsymbol{\Sigma}^{-۱}\begin{pmatrix} x_۱ \\ E(\mathbf{x}_{-۱}|x_۱)^T \end{pmatrix} = -\frac{۱}{۲}\frac{x_۱^۲}{\sum_{۱۱}}.$$

این لم بیان می‌کند که توزیع توأم $\boldsymbol{\theta}$ با عنوان تابعی از θ_i یا $\boldsymbol{\theta}_{-i}$، ارزیابی‌شده در میانگین شرطی $E(\boldsymbol{\theta}_{-i}|\theta_i)$، رفتار می‌کند. چون $\mathbf{x}$ گوسی نیست، این محاسبه یک تقریب خواهد بود.

برای محور $k = ۱, \ldots, m$ این الگوریتم میانگین شرطی $E(\boldsymbol{\theta}_{-i}|\theta_i)$ را با گوسی بودن $\boldsymbol{\theta}$، طوری محاسبه می‌کند که در θ_j خطی و تنها به مُد $\boldsymbol{\theta}^*$ و کوواریانس $\boldsymbol{\Sigma}$ در راهبرد توری وابسته است. سپس از لم ۱.۳.۴ برای دستیابی به چگالی پسین کناری تقریب‌شده θ_j در هر مسیر از این محور استفاده می‌کنیم. برای هر مسیر از این محور باید فقط سه نقطه از این چگالی را که به وسیله‌ی لم ۱.۳.۴ تقریب زده شده‌اند، ارزیابی شوند به‌طوری که برای محاسبه مشتق دوم کافی است و انحراف معیارهای σ_k^+ و σ_k^- معادله به‌دست می‌دهند.

۴.۳.۴ تقریب چگالی پسین کناری عناصر میدان پنهان

پس از انتخاب نقاط انتگرال‌گیری، برای آن‌که بتوان با استفاده از دو تقریب $\tilde{\pi}(\theta_k|\mathbf{y})$ و $\tilde{\pi}(x_i|\theta_k,\mathbf{y})$، چگالی (۴.۴) را تقریب زد، باید چگالی پسین کناری x_iها را محاسبه کرد. از این‌رو، در ادامه، با استفاده از سه روش تقریب گوسی، لاپلاس و لاپلاس ساده‌شده[۱۶]، تقریب پسین $\pi(x_i|\theta_k,\mathbf{y})$ را محاسبه می‌کنیم و محاسن و معایب هر یک را بیان خواهیم کرد.

الف: تقریب گوسی

استفاده از $\tilde{\pi}_G(\mathbf{x}|\theta_k,\mathbf{y})$ و استخراج کناری آن ساده‌ترین راه محاسبه تقریب $\pi(x_i|\theta_k,\mathbf{y})$ است. با توجه به محاسبه‌ی $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ در مرحله انتخاب نقاط انتگرال‌گیری θ_k و همچنین محاسبه‌ی تقریب گوسی $\tilde{\pi}_G(\mathbf{x}|\theta_k,\mathbf{y})$، مقدار بردار میانگین نیز به‌دست آمده و می‌توان واریانس کناری را با استفاده از وارون ماتریس دقت Q به‌دست آورد. از آن‌جایی‌که معمولا بعد میدان پنهان بزرگ است، وارون کردن ماتریسی با ابعاد بزرگ دشوار به‌نظر می‌آید. رو و مارتینو (۲۰۰۷) رابطه‌ی بازگشتی را بر اساس معادلات تاکاهاشی[۱۷] (تاکاهاشی و همکاران، ۱۹۷۳) معرفی کردند. این رابطه قادر است تا واریانس کناری را برای GMRFهایی با ابعاد بزرگ به سرعت محاسبه کند. بنابراین تقریب

$$\tilde{\pi}_G(x_i|\theta_k,\mathbf{y}) \sim N(\mu_i(\theta_k),\sigma_i^{2}(\theta_k)) \qquad (۱۶.۴)$$

که در آن $\sigma_i^{2}(\theta_k)$ واریانس کناری مولفه x_i است، که برای $\pi(x_i|\theta_k,\mathbf{y})$ عموماً نتایج قابل قبولی را ارائه می‌دهد.

اگر $\mathbf{Q}=\mathbf{L}\mathbf{L}^T$ و $\mathbf{z}\sim N(0,\mathbf{I})$ باشد، آن‌گاه $\mathbf{x}$ در معادله $\mathbf{L}^T\mathbf{x}=\mathbf{z}$ دارای توزیع $N(0,\boldsymbol{\Sigma})$ است. لازم به ذکر است که می‌توان معادله‌ی $\mathbf{L}^T\mathbf{x}=\mathbf{z}$ را به‌صورت

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

هم نوشت. با ضرب طرفین این معادله در x_j به ازای $j\geq i$ و امید گرفتن از آن، درایه‌های ماتریس $\boldsymbol{\Sigma}$ توسط رابطه‌ی

$$\Sigma_{ij} = \frac{\delta_{ij}^{2}}{L_{ii}} - \frac{1}{L_{ii}}\sum_{k=i+1}^{n} L_{k_i}\Sigma_{k_j}, \qquad j\geq i, i=n,\ldots,1 \qquad (۱۷.۴)$$

محاسبه می‌شوند که در آن $\delta_{ii}=1$ و اگر $i=j$ آن‌گاه $\delta_{ii}=0$ خواهد بود. اما اگر $i\neq j$، معادله‌ی (۱۷.۴) دقیقاً معادل با معادله‌ی تاکاهاشی است که قادر به محاسبه‌ی سریع عناصر معینی از وارون یک ماتریس با استفاده از تجزیه چولسکی آن می‌باشد.

[۱۶] Simplified Laplace approximation
[۱۷] Takahashi equation

بدین ترتیب اگر $\mathbf{Q} = \mathbf{L}\mathbf{L}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T$ باشد، به گونه‌ای که $\mathbf{L} = \mathbf{V}\mathbf{D}^{\frac{1}{2}}$ در نتیجه $\mathbf{D}$ یک ماتریس قطری و $\mathbf{V}$ یک ماتریس پایین مثلثی با درایه‌های قطری یک است، بنابراین خواهیم داشت

$$\mathbf{\Sigma} = \mathbf{D}^{-1}\mathbf{V}^{-1} + (\mathbf{I} - \mathbf{V}^T)\mathbf{\Sigma} \qquad (۱۸.۴)$$

چرا که $\mathbf{Q}\mathbf{\Sigma} = \mathbf{I}$ بنابراین $\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{\Sigma} = \mathbf{I}$. پس

$$(\mathbf{V}\mathbf{D})^{-1}\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{\Sigma} = (\mathbf{V}\mathbf{D})^{-1}.$$

بنابراین

$$\mathbf{I}\mathbf{V}^T\mathbf{\Sigma} = \mathbf{D}^{-1}\mathbf{V}^{-1}.$$

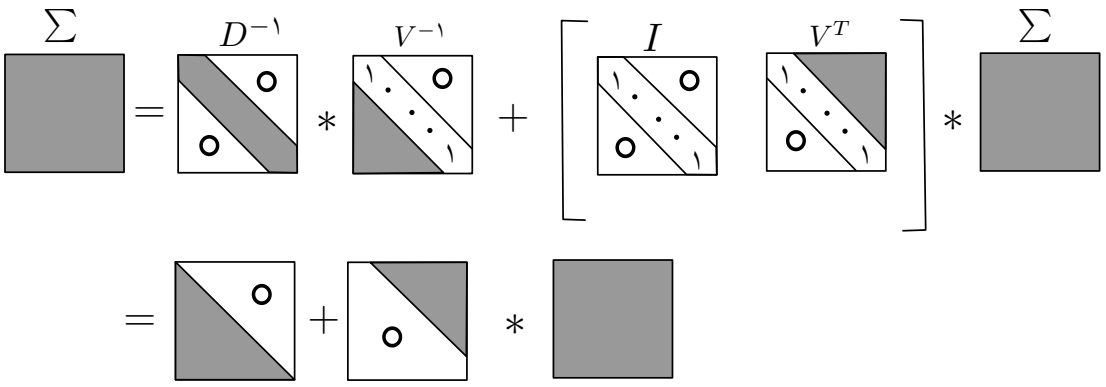
پس

$$\mathbf{D}^{-1}\mathbf{V}^{-1} - \mathbf{V}^T\mathbf{\Sigma} = 0.$$

در نتیجه

$$\mathbf{D}^{-1}\mathbf{V}^{-1} - \mathbf{V}^T\mathbf{\Sigma} + \mathbf{\Sigma} = \mathbf{\Sigma}.$$

شکل ۵.۴ نواحی غیر صفر را برای رابطه (۱۸.۴) نشان می‌دهد. همان‌طور که ملاحظه می‌شود $\mathbf{D}^{-1}\mathbf{V}^{-1}$ یک ماتریس پایین مثلثی است، به‌طوری‌که $(\mathbf{D}^{-1}V_{ii}^{-1}) = (\mathbf{D}^{-1})_{ii}$.



شکل ۵.۴: شماتیک ماتریسی معادله‌ی تاکاهاشی

از این‌رو خوشبختانه نیازی به محاسبه‌ی V^{-1} نخواهیم داشت. بنابراین درایه‌های ماتریس $\mathbf{\Sigma}$ به‌صورت

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

محاسبه شده که پایه‌ای را برای محاسبه‌ی واریانس‌های کناری x_1 تا x_n فراهم می‌کند.

ب: تقریب لاپلاس

یکی از مشکلات به‌کارگیری تقریب گوسی، آن است که خطای ناشی از تشخیص نادرست مُد یا درنظر نگرفتن چولگی توزیع، ممکن است مقادیر آن را تحت تاثیر قرار دهد. برای رفع این مشکل، تقریب لاپلاس به‌صورت

$$\tilde{\pi}_{LA}(x_i|\boldsymbol{\theta},\mathbf{y}) \propto \frac{\pi(\mathbf{x},\boldsymbol{\theta},\mathbf{y})}{\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i,\mathbf{x},\boldsymbol{\theta},\mathbf{y})}\bigg|_{\mathbf{x}_{-i}=\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})} \qquad (۱۹.۴)$$

معرفی شد که در آن $\tilde{\pi}_{GG}$ تقریب گوسی و $\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})$ مد چگالی شرطی $\pi(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})$ است. همچنین لازم به ذکر است که $\tilde{\pi}_{GG}$ متفاوت از چگالی شرطی متناظر با $\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ است. در $\tilde{\pi}$ ماتریس دقت نسبت به x_i ثابت و میانگین آن تابعی خطی از x_i است. در حالی‌که در $\tilde{\pi}_{GG}$ ابتدا مُد $\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})$ و سپس لگارتیم درستنمایی حول آن محاسبه می‌شود که ماتریس دقت آن با تغییر x_i تغییر می‌کند. از آن‌جا که چگالی $\tilde{\pi}_{GG}$ بر اساس شرطی کردن روی x_i و سپس به‌کار بردن تقریب لاپلاس بر روی سایر متغیرها قابل محاسبه است، از تقریب (۱۶.۴) که تنها بر اساس تقریب گوسی است، دقیق‌تر می‌باشد.

برای محاسبه‌ی عبارت (۱۹.۴) بایستی $\tilde{\pi}_{GG}$ را برای x_i و $\boldsymbol{\theta}$ محاسبه کنیم که امری زمان‌بر و طاقت‌فرسا است. برای حل این مشکل، رو و همکاران (۲۰۰۹) دو تعدیل را برای تسهیل در محاسبات ارائه نمودند. در تعدیل اول به دلیل آن‌که میانگین شرطی $E[\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y}]$ و مُد شرطی $\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})$ به شرط گوسی بودن توزیع $(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})$ بر یکدیگر منطبق خواهند بود و به دلیل درنظر گرفتن توزیع پیشین گوسی برای $\pi(\mathbf{x},\boldsymbol{\theta})$، گوسی بودن توزیع $(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})$ نیز چندان دور از انتظار نخواهد بود. بنابراین با استفاده از

$$x^*_{-i}(x_i,\boldsymbol{\theta}) \approx E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i) \qquad (۲۰.۴)$$

از میانگین به‌جای مُد استفاده و از گام بهینه‌سازی $\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})$ صرف‌نظر می‌کنیم. هاسیو و همکاران (۲۰۰۴) نشان دادند که استفاده از $E_{\tilde{\pi}}(\mathbf{x}_{-i}|x_i)$ به‌جای $\mathbf{x}^*_{-i}(x_i,\boldsymbol{\theta})$ نتایج تقریبا مشابهی را در پی خواهد داشت.

در تعدیل دوم از این ایده استفاده می‌شود که تنها x_iهایی که در همسایگی x_i هستند، بر متغیر کناری x_i اثر می‌گذارند. اگر نزدیکی بین x_i و x_j به وابستگی بین دو رأس i و j تعبیر شود، تنها آن x_jهایی که در شعاع مشخصی از رأس i هستند در محاسبه‌ی چگالی کناری x_i لحاظ خواهند شد که این شعاع همسایگی با $R(\boldsymbol{\theta})$ نشان داده می‌شود. با استفاده از (۲۰.۴) رابطه

$$\frac{E_{\tilde{\pi}_G}(x_j|x_i)-\mu_j(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} = a_{ij}(\boldsymbol{\theta})\frac{x_i-\mu_i(\boldsymbol{\theta})}{\sigma_j(\boldsymbol{\theta})} \qquad (۲۱.۴)$$

برای بعضی a_{ij} که $i \neq j$ برقرار است. از این‌رو یک روش برای تعیین شعاع همسایگی، ساخت مجموعه‌ی $R(\boldsymbol{\theta})$ به‌صورت $R(\boldsymbol{\theta}) = \{j : |a_{ij}| > ۰/۰۰۱\}$ است.

ج: تقریب لاپلاس ساده شده

در تقریب لاپلاس، حتی اگر از میانگین به‌جای مُد استفاده کنیم، همچنان فرآیند محاسباتی سنگین و زمان‌بر خواهد بود. با توجه به (۱۰.۴) و $\pi(\mathbf{x}_{-i}|x_i) = \frac{\pi(\mathbf{x})}{\pi(x_i)} \propto \pi(\mathbf{x})$ مخرج کسر (۱۹.۴) به‌صورت

$$\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})\Big|_{x_i=E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)} \propto (2\pi)^{-\frac{n}{2}}\left|\mathbf{Q}+\operatorname{diag}(\mathbf{c})\right|^{\frac{1}{2}} \times \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}+b^T\mathbf{x}-\frac{1}{2}\mathbf{x}^T(\operatorname{diag}(\mathbf{c})\mathbf{x})\right)\Big|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)} \tag{۲۲.۴}$$

است، که در آن $c_i = -\dfrac{\partial^2\pi(y_j|x_j,\boldsymbol{\theta})}{\partial x_j^2}$. بنابراین برای هر $i=1,\ldots,n$ ماتریس دقت $\mathbf{x}_{-i}|x_i$ به‌صورت زیرماتریس اصلی $Q_{[-i,-j]}$ است. از طرفی رابطه

$$\begin{aligned}\pi(\mathbf{x}) &= \pi(x_i)\pi(\mathbf{x}_{-i}|x_i)\\ &\propto \operatorname{Var}(x_i)^{-\frac{1}{2}}|Q_{[-i,-j]}|^{\frac{1}{2}}\exp-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}\end{aligned}$$

برقرار است. بنابراین رابطه‌ی (۲۲.۴) را می‌توان به‌صورت

$$\begin{aligned}&\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)}\\ &\propto (2\pi)^{-\frac{n}{2}}\operatorname{Var}(x_i)^{-\frac{1}{2}}|Q_{[-i,-i]}+\operatorname{diag}(\mathbf{c})|^{\frac{1}{2}}\exp\left(-\frac{1}{2}(x_i,E(\mathbf{x}_{-i}|x_i)^T)Q\begin{pmatrix}x_i\\E(\mathbf{x}_{-i}|x_i)\end{pmatrix}\right)\\ &\exp\left(\mathbf{b}^T\begin{pmatrix}x_i\\E(\mathbf{x}_{-i}|x_i)\end{pmatrix}-\frac{1}{2}(x_i,E(\mathbf{x}_{-i}|x_i)^T)(\operatorname{diag}(\mathbf{c}))\begin{pmatrix}x_i\\E(\mathbf{x}_{-i}|x_i)\end{pmatrix}\right)\\ &\propto |Q_{[-i,-i]}+\operatorname{diag}(\mathbf{c})|^{\frac{1}{2}}\exp\left(-\frac{1}{2}(x_i,E(\mathbf{x}_{-i}|x_i)^T)(\operatorname{diag}(\mathbf{c}))\begin{pmatrix}x_i\\E(\mathbf{x}_{-i}|x_i)\end{pmatrix}\right)\end{aligned}$$

نوشت. رو و همکاران (۲۰۰۹)، نشان دادند که اگر $\mathbf{x}\sim N(0,\boldsymbol{\Sigma})$، آن‌گاه

$$-\frac{1}{2}(x_i,E(\mathbf{x}_{-i}|x_i)^T)\boldsymbol{\Sigma}^{-1}\begin{pmatrix}x_i\\E(\mathbf{x}_{-i}|x_i)\end{pmatrix} = -\frac{1}{2}\frac{x_i^2}{\boldsymbol{\Sigma}_{ii}},\quad i=1,\ldots,n.$$

بنابراین

$$\log\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i,\boldsymbol{\theta},\mathbf{y})|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_{GG}}(\mathbf{x}_{-i}|x_i)} \propto c_0+\frac{1}{2}\log|Q^*+\operatorname{diag}(\mathbf{c})|. \tag{۲۳.۴}$$

در محاسبه‌ی دترمینان موجود در (۲۳.۴) باید برای هر x_i یک ماتریس $(N-1)\times(N-1)$ محاسبه شود که امری بسیار زمان‌بر است. رو و همکاران (۲۰۰۹) برای حل این مشکل تقریب لاپلاس ساده‌شده را معرفی کردند. در این روش با تعریف $x_i^{(s)} = \frac{x_i-\mu_i(\boldsymbol{\theta})}{\sigma_i(\boldsymbol{\theta})}$ صورت کسر (۱۹.۴) حول $x_i=\mu_i(\boldsymbol{\theta})$ به‌صورت

$$\begin{aligned}&\log\{\pi(\mathbf{x},\boldsymbol{\theta},\mathbf{y})\}|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)} =\\ &-\frac{1}{2}(x_i^{(s)})^2+\frac{1}{6}(x_i^{(s)})^3\sum_{j\in\frac{\mathcal{I}}{i}}d_j^{(3)}\{\mu_i(\boldsymbol{\theta}),\boldsymbol{\theta}\}\{\sigma_i(\boldsymbol{\theta})a_{ij}(\boldsymbol{\theta})\}^3+\ldots\end{aligned} \tag{۲۴.۴}$$

بسط داده می‌شود. جملات مرتبه‌ی اول و دوم تقریب گوسی و جمله‌ی مرتبه‌ی سوم تصحیحی برای چولگی فراهم می‌کند. در بسط لگارتیم مخرج کسر (۱۹.۴) با توجه به این‌که برای هر ماتریس M رابطه‌ی $\frac{\partial \log|M|}{\partial m} = \text{Trace}(M^{-1}\frac{\partial M}{\partial m})$ برقرار است (ماردیا و همکاران، ۱۹۸۹)، رابطه‌ی

$$\begin{aligned}\frac{d\log|Q^* + \text{diag}(\mathbf{c})|}{\text{d}x_i} &= \text{Trace}[Q^* + \text{diag}(\mathbf{c})]^{-1}\frac{\text{d}[Q^* + \text{diag}(\mathbf{c})]}{\text{d}x_i} \\ &= \text{Trace}[Q^* + \text{diag}(\mathbf{c})]^{-1}\text{diag}[d^{(3)}(x_i, \boldsymbol{\theta})] \\ &= \sum_j \text{Var}(x_j|x_i)d_j^{(3)}(x_i, \boldsymbol{\theta}) \\ &= \sum_j \sigma_j(\boldsymbol{\theta})\{1 - \text{corr}_{\tilde{\pi}_G}(x_i, x_j|\boldsymbol{\theta})^2\}d_j^{(3)}(x_i, \boldsymbol{\theta}) \qquad (۲۵.۴)\end{aligned}$$

به‌دست می‌آید، که در آن $\text{Var}(x_j|x_i) = \sigma_i(\boldsymbol{\theta})\{1 - \text{corr}_{\tilde{\pi}_G}(x_i, x_j|\boldsymbol{\theta})^2\}$. با استفاده از رابطه‌ی (۲۵.۴)، بسط تیلور مخرج کسر (۱۹.۴) حول $x_i = \mu_i(\boldsymbol{\theta})$ به‌صورت

$$\begin{aligned}&\log\{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})\}|_{\mathbf{x}_{-i}=E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)} = \\ &c_0 - \frac{1}{2}(X_i^{(s)})\sum_{j\in\frac{\mathcal{I}}{i}}\sigma_j(\boldsymbol{\theta})\{1 - \text{corr}_{\tilde{\pi}_G}(x_i, x_j|\boldsymbol{\theta})^2\}\text{d}_j^{(3)}(\mu_i(\theta)_i, \boldsymbol{\theta})\sigma_j(\boldsymbol{\theta})a_{ij}(\boldsymbol{\theta}) + \ldots \qquad (۲۶.۴)\end{aligned}$$

حاصل می‌شود که در آن c_0 مقداری ثابت است. با توجه به (۲۴.۴) و (۲۶.۴) خواهیم داشت

$$\gamma^{(1)}(\boldsymbol{\theta}) = \frac{1}{2}(x_i^{(s)})\sum_{j\in\frac{\mathcal{I}}{i}}\sigma_i(\theta)\{1 - \text{corr}_{\tilde{\pi}_G}(x_i, x_j|\boldsymbol{\theta})^2\}\text{d}_j^{(3)}(\mu_i(\boldsymbol{\theta})_i, \boldsymbol{\theta})a_{ij}(\boldsymbol{\theta})$$

و

$$\gamma^{(3)}(\boldsymbol{\theta}) = \sum_{j\in\frac{\mathcal{I}}{i}}\text{d}_j^{(3)}\{\mu_i(\boldsymbol{\theta}), \theta\}\{\sigma_j(\boldsymbol{\theta})a_{ij}(\boldsymbol{\theta})\}^3.$$

بنابراین تقریب لاپلاس ساده‌شده به‌صورت زیر به‌دست می‌آید:

$$\tilde{\pi}_{SLA} = c_0 - \frac{1}{2}(x_i^{(s)}) + \gamma_i^{(1)}(\boldsymbol{\theta})(x_i^{(s)}) + \frac{1}{6} + (x_i^{(s)})\gamma_i^{(3)}(\boldsymbol{\theta}) + \ldots$$

۴.۴ ملاک‌های ارزیابی

یکی از راه‌های متداول برای مقایسه مدل‌ها در تحلیل بیزی، استفاده از عامل بیزی[۱۸] است. فرض کنید بردار مشاهدات y را داریم و هدف ما مقایسه دو مدل

$$M_1 : \pi_1(\mathbf{y}|\theta_1), \qquad M_2 : \pi_2(\mathbf{y}|\theta_2)$$

[۱۸] Bayes factor

باشد. توزیع‌های پیشین $\pi_1(\theta_1)$ و $\pi_2(\theta_2)$ را برای پارامترها و احتمالات پیشین $\pi(M_1)$ و $\pi(M_2)$ را برای مدل‌ها درنظر می‌گیریم. بر اساس قانون بیز، بخت پسین‌ها برای دو مدل به‌صورت

$$\frac{\pi(M_1|\mathbf{y})}{\pi(M_2|\mathbf{y})} = \frac{\pi(M_1)}{\pi(M_2)} \cdot \frac{\int_{\theta_1} \pi_1(\mathbf{y}|\theta_1)\pi_1(\theta_1)d\theta_1}{\int_{\theta_2} \pi_2(\mathbf{y}|\theta_2)\pi_2(\theta_2)d\theta_2}$$

$$= \text{عامل بیزی} \times \text{بخت پیشین مدل‌ها}$$

است. بعد از مرتب‌سازی، عامل بیزی به‌صورت

$$\begin{aligned}\beta(\mathbf{y}) &= \frac{\pi(M_1|\mathbf{y})}{\pi(M_2|\mathbf{y})} \cdot \frac{\pi(M_2)}{\pi(M_1)} \\ &= \frac{\pi(M_1|\mathbf{y})/\pi(M_2|\mathbf{y})}{\pi(M_1)/\pi(M_2)} \\ &= \frac{\text{بخت پسین مدل‌ها}}{\text{بخت پیشین مدل‌ها}}\end{aligned}$$

به‌دست می‌آید. اگر $\pi(M_1) = \pi(M_2)$ و فضای پارامتر θ_1 و θ_2 یکسان باشند، آن‌گاه

$$\begin{aligned}\beta(\mathbf{y}) &= \frac{\pi(M_1|\mathbf{y})}{\pi(M_2|\mathbf{y})} \cdot \frac{\pi(M_2)}{\pi(M_1)} \\ &= \frac{\frac{\pi(M_1,\mathbf{y})}{\pi(\mathbf{y})\pi(M_1)}}{\frac{\pi(M_2,\mathbf{y})}{\pi(\mathbf{y})\pi(M_2)}} \\ &= \frac{\pi(\mathbf{y}|M_1)}{\pi(\mathbf{y}|M_2)}.\end{aligned}$$

یعنی عامل بیزی با نسبت درستنمایی‌ها برابر می‌شود. بنابراین مدل‌ها را می‌توان با استفاده از درستنمایی‌های دو مدل، یعنی $\pi(\mathbf{y}|M_1)$ و $\pi(\mathbf{y}|M_k)$ مقایسه کرد. برای این منظور تقریب درستنمایی‌های $\pi(\mathbf{y}|M_k)$ با استفاده از (۹.۴)، به روش INLA به‌صورت

$$\tilde{\pi}(\mathbf{y}|M_k) = \int \frac{\tilde{\pi}_k(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\theta, \mathbf{y})} d\boldsymbol{\theta}, \quad k = 1, 2 \qquad (۲۷.۴)$$

محاسبه می‌شود.

۱.۴.۴ ملاک اطلاع انحراف

برای مقایسه مدل‌های سلسله‌مراتبی پیچیده از ملاک اطلاع انحراف[۱۹] (DIC) که اندازه‌ای از پیچیدگی[۲۰] (تعداد موثر پارامترها) و برازش[۲۱] مدل است، استفاده می‌شود. با توجه به این‌که افزایش تعداد پارامترها تأثیر مستقیمی بر پیچیدگی و قدرت پیشگویی مدل دارد، بهتر است به هنگام مقایسه مدل‌ها تعداد پارامترهای موثر نیز مورد توجه قرار گیرد. دمپستر (۱۹۷۴)،

[۱۹] Deviance information criterion
[۲۰] Complexity
[۲۱] Fit

ملاکی تحت عنوان انحراف را بر اساس لگاریتم درستنمایی توزیع پسین معرفی کرد که از آن برای استخراج کمیت‌های برازش و پیچیدگی استفاده می‌شود و نهایتا با ترکیبی از این دو ملاک DIC تعریف می‌شود (اشپیگل هالترو و همکاران، ۲۰۰۲).

فرض کنید بردار مشاهدات $\mathbf{y}$ در اختیار باشد و هدف انتخاب بهترین مدل از بین مدل‌های برازش‌یافته به این مجموعه باشد. با استفاده از رهیافت بیزی، توزیع توأم $\pi(\mathbf{y}, \theta)$ به‌صورت

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

است که در آن $\phi = (\mathbf{x}, \boldsymbol{\theta})$ و با شرط معلوم بودن $\mathbf{x}$ و $\boldsymbol{\theta}$، y_iها مستقل شرطی هستند. دمپستر (۱۹۷۴)، ملاک انحراف را به‌صورت

$$D(\mathbf{x}) = -۲ \log \pi(\mathbf{y}|\mathbf{x}) + ۲ \log \pi(\mathbf{y}) \qquad (۲۸.۴)$$

تعریف کرد، که در آن $\pi(\mathbf{y})$، عبارت استانداردی است که فقط تابعی از داده‌ها است و بنابراین تأثیری در مقایسه مدل‌ها ندارد. لازم به ذکر است که عبارت (۲۸.۴) را انحراف بیزی[۲۲] نیز می‌نامند. اشپیگل هالترو و همکاران (۲۰۰۲)، کمیت برازش را بر اساس میانگین پسین انحرافات به‌صورت

$$\bar{D} = E_{\mathbf{x}|\mathbf{y}}[D]$$

و کمیت پیچیدگی را اختلاف بین میانگین پسین انحرافات و انحراف بر اساس میانگین پسین پارامترها به‌صورت

$$\begin{aligned} pD &= E_{\mathbf{x}|\mathbf{y}}[D] - D[E_{\mathbf{x}|\mathbf{y}}(\mathbf{x})] \\ &= \bar{D} - D(\bar{\mathbf{x}}) \end{aligned}$$

تعریف کردند. ملاک DIC با درنظر گرفتن دو مفهوم برازش و پیچیدگی به‌صورت

$$\begin{aligned} \text{DIC} &= \bar{D} + p_D \\ &= D(\bar{\mathbf{x}}) + ۲p_D \end{aligned}$$

تعریف می‌شود. مدل‌های با DIC کوچکتر، داده‌ها را بهتر پشتیبانی می‌کنند.

۲.۴.۴ ملاک پیش‌گویی شرطی مولفه‌ها

برای مقایسه مدل‌ها، از نظر اعتبار پیش‌گویی می‌توان از ملاک پیش‌گویی شرطی مولفه‌ها[۲۳] (CPO) برای مشاهدات به‌صورت

$$\text{CPO}_\text{i} = \pi(y_i|\mathbf{y}_{-i}), \quad i = ۱, \ldots, n$$

[۲۲] Bayesian deviance

[۲۳] Conditional precdictive ordinates

استفاده کرد، که در آن $\mathbf{y}_{-i}$ همان بردار $\mathbf{y}$ به‌جز درایه iام است. مقادیر صفر و خیلی کوچک CPO بیان‌گر پرت بودن y_i متناظر است. از منهای میانگین لگارتیم CPO_iها برای مقایسه مدل‌ها استفاده می‌شود، که ملاک نمره لگاریتمی اعتبارسنجی متقابل[۲۴] (LogScore) نامیده می‌شود (گنتینگ و رفتری، ۲۰۰۷). با توجه به این‌که مقادیر بزرگ CPO_iها گویای پیش‌گویی برتر مدل است، بنابراین هرچه مقدارLogScore کوچک باشد، مدل برای پیش‌گویی مناسب‌تر خواهد بود.

۵.۴ داده‌های فضایی سه بعدی

میدان‌های تصادفی گوسی (معروف به GRF) از مولفه‌های مهم مدل‌بندی در آمار فضایی هستند و نقش برجسته‌ای به‌ویژه در زمین‌آمار دارند (کرسی، ۱۹۹۳؛ استین، ۱۹۹۹؛ دیگل و ریبریو، ۲۰۰۷؛ چلیس و دلفینر، ۲۰۱۲). همچنین GRFها در مدل‌های فضایی سلسله‌مراتبی مدرن، یک مولفه ساختاری مهم را تشکیل می‌دهند (بارتلت، ۱۹۳۸) که ویژگی‌های اصلی میدان فضایی توسط تابع کوواریانس آن‌ها بیان می‌شود. میدان‌های تصادفی گوسی از معدود مدل‌های چند متغیره‌ی مناسب با قابلیت محاسبه‌ی ثابت نرمال‌ساز هستند که از ویژگی‌های تحلیلی خوبی نیز برخوردارند. اما از منظر محاسباتی، GRFها برای حجم‌های نمونه بزرگ با مشکل مواجه می‌شوند، زیرا هزینه‌ی تجزیه ماتریس‌های چگال (غیرتُنُک) با بعد بالا می‌تواند بسیار گزاف باشد. امروزه اگرچه قدرت محاسباتی در بالاترین سطح ممکن خود قرار دارد، اما در واقعیت هنوز کندی محاسبات سد راه بسیاری از پژوهش‌های کاربردی است.

قدرت محاسباتی و کارایی روش INLA زمانی است که میدان تصادفی فضایی، مارکوفی باشد و بتوان ماتریس دقت (که عکس ماتریس کوواریانس میدان است) را تُنُک در نظر گرفت. بنابراین، این روش برای داده‌های زمین‌آماری به‌ویژه زمانی که حجم داده‌ها بزرگ است، نمی‌تواند کارا باشد. برای رفع این مشکل و استفاده از روش تقریبی INLA برای تحلیل داده‌های زمین‌آماری، لیندگرین و همکاران (۲۰۱۱) روشی را با عنوان معادلات دیفرانسیل جزیی تصادفی[۲۵] (SPDE) معرفی کردند که با مکانیسمی یک GRF را با یک تقریب GMRF به‌صورت تقریبی بازنویسی می‌کند. با روش پیشنهادی آن‌ها می‌توان یک مدل GRF را برای داده‌های زمین‌آماری توسط یک مدل مارکوفی GMRF نمایش داد و سپس از روش INLA برای برازش آن به داده‌ها بهره برد. به این شکل، کارایی محاسباتی روش تقریبی بیزی با ترکیب رهیافت SPDE حفظ می‌شود.

پس، هدف اصلی رهیافت SPDE یافتن یک فرآیند تصادفی مارکوفی GMRF، با همسایگی فضایی و ماتریس دقت تُنُک $\mathbf{Q}$ توسط گسسته‌سازی ناحیه فضایی پیوسته است. با توجه به تمرکز پایان‌نامه بر تحلیل داده‌های زمین‌آماری میدان نفتی، از رهیافت ترکیبی INLA+SPDE

[۲۴] Cross Validated Logarithmic Score

[۲۵] Stochastic partial differential equation

برای مدل‌بندی آن‌ها استفاده می‌کنیم. استفاده از این رهیافت برای داده‌های فضایی سه بعدی به ندرت و در چارچوب‌های غیرمنسجم با ابزارهای غیرمتحد در نرم‌افزارهای مختلف انجام شده است. به‌عنوان نمونه می‌توان به ژانگ و همکاران (۲۰۱۶) اشاره کرد.

ما در این پایان‌نامه، برای اولین بار از رهیافت ترکیبی مذکور و مختص داده‌های سه بعدی برای تحلیل داده‌های فضایی استفاده می‌کنیم. برای این کار، ابتدا رهیافت SPDE را معرفی می‌کنیم.

۶.۴ رهیافت معادلات دیفرانسیل جزئی تصادفی

فرض کنید $\{\omega(\mathbf{s}),\ \mathbf{s}\in\mathbf{D}\subseteq\mathbb{R}^d\}$ یک میدان تصادفی گوسی مانای مرتبه دوم با تابع کوواریانس مترن (معرفی‌شده در رابطه (۸.۱) در فصل اول) با پارامتر همواری k و مقیاس ϕ است. افزون بر این، فرض کنید فرآیند $\omega(\mathbf{s})$ در موقعیت‌های $\boldsymbol{s}_i$، $i=1,\cdots,n$، مشاهده شده است. روش SPDE یک GMRF با همسایگی فضایی و ماتریس تُنُک Q می‌یابد که در بهترین حالت نشان‌دهنده یک GRF مترن است. در این رهیافت، میدان تصادفی مترن به‌صورت یک ترکیب خطی از توابع تکه‌ای خطی که بر یک مثلث‌بندی از دامنه‌ی D تعریف شده است، بیان می‌شود. به‌عبارت ساده‌تر دامنه فضای D به مجموعه‌ای از مثلث‌ها تقسیم‌بندی می‌شود که حداکثر در یک یال یا زاویه‌ی مشترک متقاطع اشتراک دارند. موقعیت‌های $s_1,\cdots,s_n$ رئوس مثلث‌بندی اولیه در نظر گرفته می‌شوند و سپس دیگر رئوس به‌منظور دست یافتن به یک مثلث‌بندی مناسب برای پیش‌گویی فضایی و سایر اهداف اضافه می‌شوند.

لیندگرین و همکاران (۲۰۱۱) نشان دادند تنها جواب‌های پایدار برای یک SPDE میدان‌های تصادفی مترن هستند، به‌طوری که یک میدان تصادفی گوسی $\omega(\mathbf{s})$ با تابع کوواریانس مترن و پارامترهای هموارسازی k و مقیاس ϕ تعریف می‌شود، که در آن $\sigma^2=\frac{\Gamma(k)}{\Gamma(k+\frac{d}{2})(4\pi)^{\frac{d}{2}}\pi^{2k}}$ و $K_k(\cdot)$ تابع بسل اصلاح شده نوع دوم است (ویتل ۱۹۵۴ و ۱۹۵۴). یک جواب پایدار برای معادله‌ی دیفرانسیل جزئی تصادفی به‌صورت

$$(\phi^2-\Delta^2)^{\frac{\alpha}{2}}\tau(\mathbf{s})\omega(\mathbf{s})=W(\mathbf{s}),\quad \mathbf{s}\in\mathbb{R}^d,\alpha=k+\frac{d}{2},\phi>0,k>0 \tag{۲۹.۴}$$

است، که در آن W میدان تصادفی نوفه سفید[۲۶]، Δ عمل‌گر دیفرانسیلی مرتبه دوم یا همان عمل‌گر لاپلاس[۲۷]، $\Delta=\sum_{i=1}^d\sigma^2/\omega_i^2$، و τ پارامتر کنترل‌گر واریانس است. عبارت $(\phi^2-\Delta^2)^{\frac{\alpha}{2}}$ جمله‌ی عمل‌گر دیفرانسیل[۲۸] است. با به کار بردن تعریف فوریه، یک لاپلاس کسری[۲۹] به شکل

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

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

در نظر گرفته می‌شود، که در آن ψ تابعی در $\mathbb{R}^d$ است. لیندگرین و همکاران (۲۰۱۱) ابتدا یک صورت ضعیف تصادفی[۳۰] معادله‌ی دیفرانسیل جزئی (۲۹.۴) را برای یک میدان تصادفی مارکوفی گوسی که بیان‌گر میدان تصادفی مترن روی شبکه مثلث‌بندی باشد، در نظر گرفتند. شبکه‌بندی مثلثی نسبت به روش‌های دیگر مانند روش‌های مبتنی بر شبکه‌های معمول و سنتی، منعطف‌تر است.

یک روش عددی برای حل تقریبی معادلات دیفرانسیل جزئی، روش عناصر متناهی[۳۱] است و مثلث‌بندی روشی معمول برای حل معادلات با استفاده از روش عناصر متناهی است (برنر و اسکات، ۲۰۰۷). در این روش دقت جواب‌ها به نحوه‌ی مثلث‌بندی ناحیه فضایی بستگی دارد. لیندگرین و همکاران (۲۰۱۱) برای نمایش عناصر متناهی SPDE، یک ترکیب خطی به صورت

$$\omega(\mathbf{s}) = \sum_{g=1}^{G} \psi_g(\mathbf{s})\hat{\omega}_g \qquad (۳۰.۴)$$

تعریف کردند، که در آن $\{\psi_g\}$ مجموعه‌ای از توابع پایه، $\hat{\omega}$ها وزن‌های ترکیب خطی دارای توزیع گوسی با میانگین صفر و G تعداد کل رئوس مثلث‌بندی است. توابع پایه طوری انتخاب می‌شوند که در رأس g برابر یک و در سایر رئوس صفر باشند.

در بسته‌ی نرم‌افزاری R-INLA برای $d = 2$ معادله‌ی دیفرانسیل جزئی تصادفی (۲۹.۴) به‌صورت عمومی

$$(\phi^2(\mathbf{s}) - \Delta^2)^{\frac{\alpha}{2}}\tau(\mathbf{s})\omega(\mathbf{s}) = W(\mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^2, \alpha = k + \frac{d}{2}, \phi > 0, k > 0$$

در نظر گرفته می‌شود که میدان‌های تصادفی نامانا را نیز شامل می‌شود، یعنی میدان‌هایی که واریانس $W(\mathbf{s})$ و پارامتر ϕ در آن‌ها ثابت نیست و به موقعیت فضایی وابسته است. البته برای سادگی، واریانس $W(\mathbf{s})$ را ثابت و پارامتر مقیاس را برای $\omega(\mathbf{s})$ به شکل $\tau(\mathbf{s})$ در نظر می‌گیرند. در این صورت، اگر یک یا هر دو پارامتر τ و ψ ثابت نباشند، میدان نامانا خواهد شد و بر اساس رابطه‌ی (۸.۱) واریانس کناری میدان به شکل $\sigma^2 = \frac{1}{4\pi\tau^2\phi^2}$ تغییر می‌یابد. لیندگرین و همکاران (۲۰۱۱) نشان دادند وزن‌های گوسی در نظر گرفته‌شده در (۳۰.۴) یک ساختار مارکوفی دارند و میدان تصادفی مترن را به یک GMRF تبدیل می‌کند. ماتریس دقت GMRF تولیدشده، وقتی $\alpha = 2$ باشد، برای وزن‌های $\omega = \{\hat{\omega}_1, \ldots, \hat{\omega}_G\}$ با استفاده از شرایط مرزی نیومن (چنگ و چنگ، ۲۰۰۵) به‌صورت

$$\mathbf{Q} = \tau(\phi^2 C + 2\phi^2 \mathbf{G} + \mathbf{G}C^{-1}\mathbf{G}) \qquad (۳۱.۴)$$

است، که در آن درایه‌های ماتریس C به‌صورت

$$\mathbf{C}_{ii} = \int \psi_i(\mathbf{s})d\mathbf{s}$$

[۳۰] Stochastic weak formulation

[۳۱] Finite element method

و درایه‌های ماتریس تُنُک $\mathbf{G}$ به‌صورت

$$\mathbf{G}_{ij} = \int \nabla\psi_i(\mathbf{s})\nabla\psi_g(\mathbf{s})d\mathbf{s}$$

است، که در آن منظور از ∇ عمل‌گر گرادیان است. ماتریس دقت $\mathbf{Q}$ تُنُک است و درایه‌های آن به پارامترهای ϕ و τ وابسته است. به این شکل، ω یک فرآیند تصادفی مارکوفی با توزیع نرمال با میانگین صفر و واریانس $\mathbf{Q}^{-1}$ است که برای اهداف محاسباتی کارای INLA مناسب خواهد بود.

۷.۴ رهیافت INLA+SPDE برای مدل‌بندی فضایی سه بعدی

با بازنویسی پیشگوی خطی (۱.۴) برای هر کدام از اثرات ثابت، غیرخطی ناپارامتری و اثرات تصادفی (شامل اثرات فضایی)، می‌توان پیشگوی خطی را به‌عنوان یک ترکیب خطی از مولفه‌های اثرات پنهان $\mathbf{x}$ نوشت. این بازنویسی کمک می‌کند تا بتوان در تقریب مبتنی بر SPDE برازش مدل با روش INLA را به‌صورتی واحد و کارا اجرا کرد.

به‌طور دقیق‌تر می‌توان پیشگوی (۱.۴) را به‌صورت زیر بازنویسی کرد:

$$\eta_i = \sum_k h_k(z_i^k)$$

که در آن k بیان‌گر kامین مولفه از پیشگو است و z_i^k مقدار متغیر تبیینی با اثر ثابت یا اثر (ناپارامتری) غیرخطی است یا متغیری است که اندیس مربوط به اثرات تصادفی را در خود جای داده است. تابع $h_k(\cdot)$ نیز تابعی است که z_i^k را به kامین مولفه این معادله پیوند می‌زند. در برخی موارد هر مشاهده y_i به ترکیب خطی مولفه‌های $\mathbf{x}$ وابسته است و توزیع مشاهدات پاسخ را می‌توان به‌صورت زیر بازنویسی کرد:

$$y_i|\mathbf{x},\boldsymbol{\theta} \sim \pi\left(y_i|\sum_j \mathbf{A}_{ij}\mathbf{x},\boldsymbol{\theta}\right).$$

مولفه‌های $\mathbf{A}_{ij}$ عناصر ماتریس $\mathbf{A}$ هستند که ماتریس مشاهده نامیده می‌شود. ماتریس مشاهده نگاشتی بین میدان پنهان فضایی (تعریف شده روی مشبکه مثلث‌بندی شده) و مشاهدات (تعریف شده در مجموعه‌ای از مکان‌ها) ایجاد می‌کند. روش R-INLA، پیش‌گوی خطی جدید $\boldsymbol{\eta}^*$ را به‌صورت ترکیب خطی $\boldsymbol{\eta}$ تولید می‌کند، یعنی

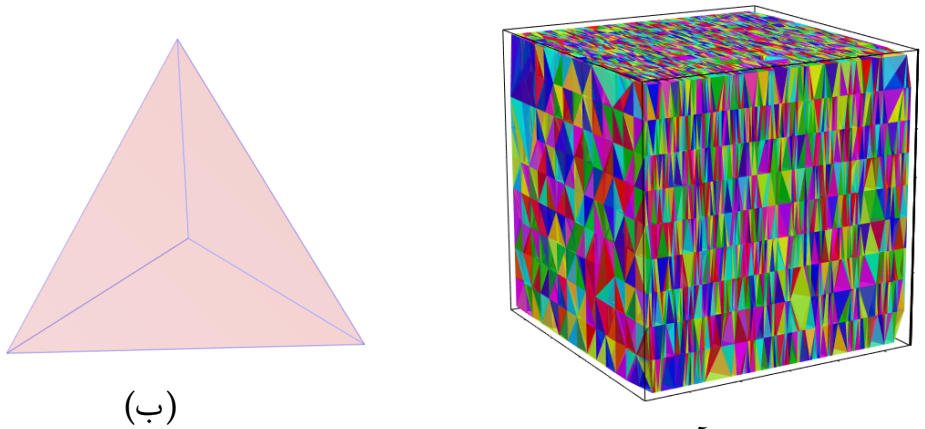
$$\boldsymbol{\eta}^* = A\boldsymbol{\eta}$$

و درستنمایی به‌وسیله‌ي η_i^* به‌جای η_i به میدان پنهان مرتبط می‌شود. بنابراین

$$\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi(y_i|\eta_i^*,\boldsymbol{\theta}).$$

۱.۷.۴ مثلث‌بندی ناحیه سه بعدی

برای اجرای روش SPDE ابتدا باید ناحیه تحت مطالعه را مثلث‌بندی کرد. در بسته نرم‌افزاری R-INLA برای حالت یک و دو بعدی، توابع تعریف‌شده‌ای برای این کار وجود دارند. اما در حالت عمومی سه بعدی تا کنون چنین ابزاری وجود نداشته است. با طرح مساله مورد نظر این پایان‌نامه با توسعه‌دهنده اصلی رهیافت SPDE، یعنی فین لیندگرین، یک بسته نرم‌افزاری جدید که به بسته INLA افزوده می‌شود، نوشته شد. در این بسته با نام inlamesh3d می‌توان مثلث‌بندی سه بعدی را که در آن از چهاروجهی[۳۲] به‌جای مثلث استفاده می‌شود، اجرا کرد. برای این کار از روش مثلث‌بندی دلاونی[۳۳] که در بسته نرم‌افزاری geometry موجود است، استفاده می‌شود. تصویر یک چهاروجهی منتخب از یک ناحیه مثلث‌بندی شده سه بعدی در ۶.۴ (آ و ب) نمایش داده شده است.



(ب) (آ)

شکل ۶.۴: گسسته‌سازی ناحیه مثال شبیه‌سازی در فضای سه بعدی با چهاروجهی‌ها (آ) و یک چهاروجهی منتخب (ب)

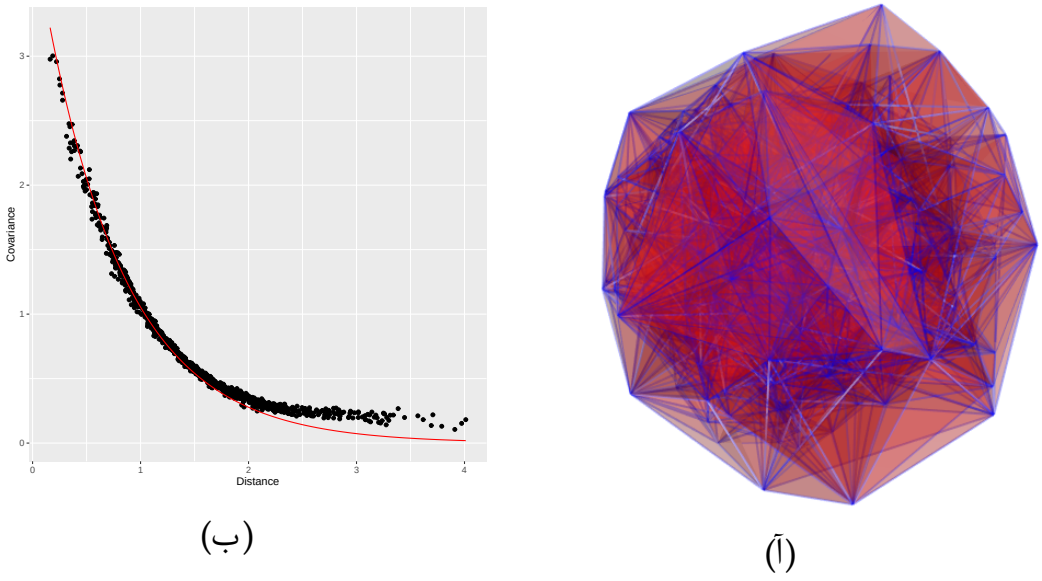
ماتریس‌های ساختاری لازم برای روش عناصر متناهی، در این حالت، مشابه حالت دوبعدی قابل محاسبه هستند. این ماتریس‌های ساختاری بر اساس رابطه (۳۱.۴) برای ساخت ماتریس دقت میدان تقریبی لازم هستند. این کار را می‌توان به صورت سرراست انجام داد و محاسبات پیچیده‌ای را نیاز ندارد.

[۳۲] Tetrahedron
[۳۳] Delaunay triangulation

تعیین پارامترهای میدان مترن در اجرای روش SPDE نیز شبیه حالت دوبعدی و با استفاده از توابع استاندارد موجود در روش مذکور در بسته R-INLA قابل انجام است. برای ارزیابی روش پیشنهادی جدید از یک مثال شبیه‌سازی استفاده می‌کنیم.

مثال شبیه‌سازی

در این مثال نقاط مکانی سه بعدی به حجم ۱۰۰۰ را از توزیع نرمال سه متغیره استاندارد تولید کردیم. بنابراین در یک ناحیه سه بعدی، هزار نقطه مکانی نامنظم قرار می‌گیرند. مقادیر واقعی پارامترهای دامنه و واریانس را به ترتیب ۱/۵ و ۲ انتخاب کردیم. همچنین مقادیر ۱ و ۱ را به‌عنوان میانه توزیع‌های پیشین پارامترهای دامنه و واریانس میدان مترن در نظر گرفتیم. بر اساس این نقاط و ناحیه مثلث‌بندی شده (شکل ۷.۴ (آ) را ببینید)، ماتریس دقت برآورد شد و تابع کوواریانس میدان برآوردشده بر حسب فاصله نقاط محاسبه و در قسمت ب شکل ۷.۴ رسم شد. در این شکل منحنی تابع کوواریانس واقعی که یک تابع نمایی است نیز رسم شده است. همانطور که مشاهده می‌شود، به‌جز در مرزها (با کمی انحراف از واقعیت به دلیل محدودیت ناحیه) کوواریانس برآوردشده با واقعیت تطابق دارد. این نتیجه نشانی از درستی تقریب GMRF حاصل از گسسته‌سازی سه بعدی در ناحیه تحت مطالعه است.



(ب) (آ)

شکل ۷.۴: (آ) گسسته‌سازی سه بعدی ناحیه شبیه‌سازی شده، (ب) منحنی تابع کوواریانس واقعی در مثال شبیه‌سازی به همراه تقریب حاصل از روش SPDE

برای ارزیابی توانایی بازیابی مقادیر واقعی پارامترهای میدان تصادفی تولیدشده، از میدان مترن با ماتریس دقت مذکور و میانگین $A\mathbf{x}$ در ۱۰۰ نقطه مکانی شبیه‌سازی شده، مقادیر پاسخ را تولید کردیم. به داده‌های تولید شده یک نوفه نرمال با انحراف معیار ۰/۱ نیز اضافه کردیم. نتایج برآوردهای میانه پسینی به همراه فواصل باورمند بیزی ۹۵ درصد در جدول ۴.۴ گزارش

شده‌اند. دو پارامتر تابع کوواریانس مترن به خوبی بازیابی شده‌اند ولی پارامتر انحراف معیار جمله نوفه کم برآورد شده است.

جدول ۴.۴: برآوردهای میانه پسینی به همراه فواصل باورمند بیزی ۹۵ درصد

پارامتر	مقدار واقعی	برآورد	فاصله باورمندپذیری	
			L	U
دامنه	۱/۵	۱/۷۸۴	۱/۲۶۶	۲/۶۹۹
واریانس	۲	۱/۹۷۴	۱/۷۰۲	۲/۳۲۵
انحراف معیار نوفه	۰/۱	۰/۰۰۹	۰/۰۰۴	۰/۰۲۸

۸.۴ نتیجه‌گیری

هدف اصلی این پایان‌نامه، ارایه و تشریح روش‌های نوین مدل‌بندی سه بعدی داده‌های زمین‌آماری به‌ویژه در زمینه داده‌های ژئومکانیکی مخازن نفتی است. برای این‌کار رهیافت‌های مختلفی را مطرح کردیم که می‌توان به صورت زیر خلاصه کرد:

۱. روش کریگینگ معمولی در حالت کلاسیک که در فصل دوم به آن پرداخته شد و بر روی یک مجموعه داده از یک میدان نفتی پیاده‌سازی شد.

۲. روش‌های کریگینگ میانه پالایش و عام با مدل‌بندی اسپلاین که در فصل سوم مطرح شدند و با دو مثال شبیه‌سازی عملکرد آن‌ها تشریح و با هم مقایسه شدند. روش میانه پالایش در مسایل سه بعدی برای اولین بار در این پایان‌نامه ارایه شده است.

۳. روش کریگینگ بیزی تقریبی با استفاده از رهیافت ترکیبی INLA+SPDE که در این زمینه نیز روش تازه‌ای است. عملکرد روش نیز در یک مثال شبیه‌سازی بررسی شد.

این مجموعه نشان می‌دهد که در این زمینه کارهای پژوهشی زیادی صورت نگرفته است و با توجه به کاربردهای جذاب هم از جنبه کاربردی و هم مالی، تمرکز بر روی آن‌ها می‌تواند بسیار مفید باشد.

مراجع

[۱] حسنی‌پاک ع آ. (۱۳۷۷)، *زمین‌آمار (ژئواستاتیک)*، چاپ اول، موسسه انتشارات و چاپ دانشگاه تهران، تهران.

[۲] محمدزاده م. (۱۳۹۸)، *آمار فضایی و کاربردهای آن*، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران.

[3] Abdel-Aal, H. K. and Alsahlawi, M. A. (2014), *Petroleum Economics and Engineering*, CRC Press, Boca Raton.

[4] Bartlett, M. S. (1938), The approximate recovery of information from replicated field experiments with large blocks, *Journal of Agricultural Science (Cambridge)*, **28**, 418-427.

[5] Bartlett, M. S. (1978), Nearest neighbour models in the analysis of field experiments, *Journal of the Royal Statistical Society: Series B (Methodological)*, **40**(2), 147-158.

[6] Beauchamp, K. G. (1984), *Applications of Walsh and Related Functions: with an Introduction to Sequency Theory*, Academic press, London.

[7] Berke, O. (2001), Modified median polish kriging and its application to the Wolfcamp–Aquifer data, *Environmetrics: The Official Journal of the International Environmetrics Society*, **12**(8), 731-748.

[8] Besag, J. (1986), On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society: Series B (Methodological)*, **48**(3), 259-279.

[9] Blangiardo, M. and Cameletti, M. (2015), *Spatial and Spatio-Temporal Bayesian Models with R-INLA*, John Wiley and Sons, UK, West Sussex.

[10] Box, G. E. and Wilson, K. B. (1951), On the experimental attainment of optimum conditions, *Journal of the Royal Statistical Society: Series B (Methodological)*, **13**(1), 1-38.

[11] Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978), *Statistics for Eexperimenters an Introduction to Design, Data Analysis and Model Building*, John Wiley and Sons, New York.

[12] Brenner, S. and Scott, R. (2007), *The Mathematical Theory of Finite Element Methods*, Springer Science and Business Media, New York.

[13] Brooks, S. Gelman, A., Jones, G. and Meng, X. L. (2011), *Handbook of Markov Chain Monte Carlo*, CRC Press, Taylor and Francis, Boca Raton.

[14] Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Brooks/COle Cengage Learning, CA, Belmont.

[15] Cheng, A. H. D. and Cheng, D. T. (2005), Heritage and early history of the boundary element method, *Engineering Analysis with Boundary Elements*, **29**(3), 268-302.

[16] Chilés, J. P. and Delfiner, P. (2009), *Geostatistics: Modeling Spatial Uncertainty*, Second Edition, John Wiley and Sons, New Jersey.

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

[18] Cressie, N. (2015), *Statistics for Spatial Data*, John Wiley and Sons, New York.

[19] Cressie, N. and Hawkins, D. M. (1980), Robust estimation of the variogram: I, *Journal of the International Association for Mathematical Geology*, **12**(2), 115-125.

[20] Cressie, N. and Wikle, C. K. (2015), *Statistics for Spatio-Temporal Data*, John Wiley and Sons, New York.

[21] Cressie, N. and Glonek, G. (1984), Median based covariogram estimators reduce bias, *Statistics and Probability Letters*, **2**(5), 299-304.

[22] Dempster, A. P. (1974), The direct use of likelihood for significance testing , *Proceedings of Conference on Foundational Questions in Statistical Inference*, Department of Theoretical Statistics, University of Aarhus, 335-352.

[23] Diggle, P. J. and Ribeiro, P. J.(2007), *Model based Geostatistics*, Springer series in statistics, New York.

[24] Diggle, P. J. and Giorgi, E. (2019), *Model-based Geostatistics for Global Public Health: Methods and Applications*, CRC Press, Boca Raton.

[25] Eriksson, M. and Siska, P. P. (2000), Understanding anisotropy computations, *Mathematical Geology*, **32**(6), 683-700.

[26] Emerson, J. D. and Hoaglin, D. C. (2000), Analysis of two-way tables by medians, *Understanding Robust and Exploratory Data Analysis*, Edited by David C. Hoaglin, Frederick Mosteller and John W. Tukey.

[27] Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd edn, Springer, Berlin.

[28] Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd, Edinburgh.

[29] Fischer, M. M. and Wang, J. (2011), *Spatial Data Analysis: Models, Methods and Techniques*, Springer Science , Berlin.

[30] Friedman, J., Hastie, T. and Tibshirani, R. (2001), *The Elements of Statistical Learning*, Springer series in statistics, New York.

[31] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013), *Bayesian Data Analysis*, Third Edition, CRC press, Hoboken .

[32] Gneiting, T. and Raftery, A. E. (2007), Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, **102**(477), 359-378.

[33] Gómez-Rubio, V. (2020), *Bayesian Inference with INLA*, CRC Press, Boca Raton.

[34] Gilks, W. R., Richardson, S. S. and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, Chapman k Hall/CRC, London.

[35] Gomez, M. and Hazen, K. (1970), *Evaluating Sulfur and Ash distribution in Coal Seems by Statistical Response Surface Regression Analysis*, US Department of Interior, Bureau of Mines.

[36] Halley, E. (1686), An historical account of the trade winds, and monsoons, observable in the seas between and near the tropicks; with an attempt to assign the phisical cause of said winds, *Philosophical Transactions*, **183**, 153-168.

[37] Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**(1), 97-109.

[38] Hrafnkelsson, B., Siegert, S. Huser, R., Bakka, H. and Jóhannesson, Á. V. (2020), Max-and-Smooth: a two-step approach for approximate Bayesian inference in latent Gaussian models, *Bayesian Analysis*.

[39] Hsiao, C. K., Huang, S. Y. and Chang, C. W. (2004), Bayesian marginal inference via candidate's formula, *Statistics and Computing*, **14**(1), 59-66.

[40] Hubin, A. and Storvik, G. (2016), Estimating the marginal likelihood with Integrated nested Laplace approximation (INLA), *ArXiv Preprint ArXiv*, 1611.01450.

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

[42] Lindgren, F. and Rue, H. (2015), Bayesian spatial modelling with R-INLA, *Journal of Statistical Software*, **63**(19), 1-25.

[43] Ludwig, G., Zhu, J., Reyes, P., Chen, C. S. and Conley, S. P. (2020), On spline-based approaches to spatial linear regression for geostatistical data, *Environmental and Ecological Statistics*, 1-28.

[44] Mardia, K. V., Kent, J.T. and Bibby, J. M. (1989), *Multivariate Analysis Academic Press*, London.

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

[46] Martínez, W. A., Melo, C. E. and Melo, O. O. (2017). Median Polish Kriging for space–time analysis of precipitation. *Spatial Statistics*, **19**, 1-20.

[47] Matheron, G. (1971), *The Theory of Regionalized Variables and its Applications*, École National Supérieure des Mines, Paris.

[48] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), Equation of state calculations by fast computing machines, *The journal of Chemical Physics*, **21**(6), 1087-1092.

[49] Montero, J. M., Fernández-Avilés, G. and Mateu, J. (2015),*S*patial and Spatio-Temporal Geostatistical Modeling and Kriging, John Wiley and Sons, West Sussex.

[50] Montgomery, D. C. (2000), *Design and Analysis of Experiments*, 5th Ed, Wiley, New York.

[51] Mood, A. M., Graybill, F. A. and Boes, D. C. (2017), *Introduction to the Theory of Statistics*, Chennai: McGraw Hill Education (India), Pvt. Ltd.

[52] Mosteller, F. and Tukey, J. W. (1977), *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading.

[53] Nazari Ostad, M., Niri, M. E. and Darjani, M. (2018), 3D modeling of geomechanical elastic properties in a carbonate-sandstone reservoir: a comparative study of geostatistical co-simulation methods, *Journal of Geophysics and Engineering*, **15**(4), 1419-1431.

[54] Oliver, M. A. and Webster, R. (2015), *Basic Steps in Geostatistics: the Variogram and Kriging*, Springer International Publishing, New York.

[55] Papadakis, J. S. (1937), Méthode statistique pour des expériences sur champ.*B*ull. Inst. Amel. Plantes a Salonique, **23**, 13-29.

[56] Pawitan, Y. (2001), *In all Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press.

[57] Ripley, B. (1981), *Spatial Statistics*, Wiley, NewYork.

[58] Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, CRC press, Boca Raton.

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

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

[61] Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. K. (2017), Bayesian computing with INLA: a review, *Annual Review of Statistics and Its Application*, **4**, 395-421.

[62] Sanchez, S. M. and Sanchez, P. J. (2005), Very large fractional factorial and central composite designs, *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **15**(4), 362-377.

[63] Sarma, D. D. (2010), *Geostatistics with Applications in Earth Sciences*, Second Edition, Springer Science and Business Media, New York.

[64] Schoenberg, I. J. (1946), Contributions to the problem of approximation of equidistant data by analytic functions. Part B. On the problem of osculatory interpolation. A second class of analytic approximation formulae, *Quarterly of Applied Mathematics*, **4**(2), 112-141.

[65] Smith, H. (1938), An empirical law describing heterogeneity in the yields of agricultural crops, *Journal of Agricultural Science (Cambridge)*, **28**, 1-23.

[66] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002), Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583-639.

[67] Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, New York.

[68] Stroup, W. W. (2012), *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*, CRC press, Hoboken.

[69] Student, (1907), On the error of counting with a haemacytometer, *Biometrika*, **5**, 351-360.

[70] Takahashi, K., Fagan, J. and Chin, M. S. (1973), Formation of a sparse bus impedance matrix and its application to short circuit study, *In Proc. 8th PICA Conference*, 4-6.

[71] Tukey, J. W. (1970), *Exploratory Data Analysis*, Addison-Wesley.

[72] Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley Company.

[73] Velleman, P. F. and Hoaglin, D. C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Scituate: Duxbury Press, New York.

[74] Wang, X., Yue, Y. R. and Faraway, J. J. (2018),*Bayesian Regression Modeling with INLA*, CRC Press.

[75] Whittle, P. (1954), On stationary processes in the plane, *Biometrika*, **41**(3-4) 434-449.

[76] Whittle, P. (1963), Stochastic-processes in several dimensions, *Bulletin of the International Statistical Institute*, **40**(2), 974-994.

[77] Wilkinson, G. N., Eckert, S. R., Hancock, T. W. and Mayo, O. (1983), Nearest neighbor (NN) analysis with field experiments, *Journal of the Royal Statistical Society B*, **45**, 151-178.

[78] Yates, F. (1938), The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments, *Biometrika*, **30**, 444-466.

[79] Zhang, R. Czado, C. and Sigloch, K. (2016), Bayesian spatial modelling for high dimensional seismic inverse problems, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**(2), 187-213.

[80] Zimmerman, D. L. (1993), Another look at anisotropy in geostatistics, *Mathematical Geology*, **25**(4), 453-470.

[81] Zoback, M. D. (2010), *Reservoir Geomechanics*, Cambridge University Press.

Abstract

The use of mechanical rock properties to study underground spaces is called geo-mechanic. Geo-mechanic is one of the most important branches of science in exploring and drilling underground reservoirs, especially oil reservoirs. Most of the data obtained in geo-mechanical studies are geostatistical data. Modeling and analysis of such data require spatial statistics methods. However, much of the geostatistical data in geo-mechanical studies are three-dimensional, making it challenging to develop spatial statistical models and efficient computational methods for their analysis. In this thesis, we propose two new methods for analyzing three-dimensional spatial random fields, for the first time, to address the mentioned challenges. First, we introduce a nonparametric method called three-dimensional median polish to estimate the trend function of a spatial random field. Employing this method, we also develop a median polish kriging method. We also compare the proposed median polish kriging's performance with universal kriging, in which the trend function is estimated using the spline method. The performance is assessed using both simulated and real examples of uniaxial compressive unconfined compressive strength in one of Iran's oil reservoirs. Second, using a Bayesian approach, called Integrated Nested Laplace Approximation (INLA), and its combination with Stochastic Partial Differential Equations (SPDE) method, we will develop the analysis of three-dimensional continuous indexed spatial random fields.

Keywords: Geostatistics, Bayesian inference, three-dimensional median polish Kriging, Integrated Nested Laplace Approximation, INLA+SPDE approach.

Faculty Of Mathematical Sciences

MSc Thesis in Mathematical Statistics

Bayesian 3d geo-mechanical modeling by using geostatistical models: a case study of an Iranian oil field

By: Fatane Fakhri

Supervisor:

Dr. Hossein Baghishani

Advisor:

Dr. Behzad Tokhmechi

November 2020