



A Bayesian Approach and Vecchia Grouping for Nonparametric Estimation of Nonstationary Spatial Covariance in Big Data

Ghasemi, F.¹ , Mohammadian Mosammam, A.² , Mateu, J.³ 

¹Department of Mathematics, University of Zanjan, Zanjan, Iran.

²Department of Statistics, University of Zanjan, Zanjan, Iran.

³Department of Mathematics, Jaume I University, Castellon, Spain.

Corresponding author: A. Mohammadian Mosammam, a.m.mosammam@znu.ac.ir

Received: 16/9/2025 **Revised:** 2/2/2026 **Accepted and Published Online:** 4/2/2026.

Introduction

Modeling spatial dependence in large datasets is challenging, as nearby observations are correlated and classical independence assumptions fail. Gaussian processes are flexible but limited by cubic computational costs. Among scalable alternatives, the Vecchia approximation is attractive due to its conditional likelihood formulation. We propose a Bayesian grouped Vecchia method for complex, nonstationary spatial dependence with replicated data. By partitioning locations into approximately independent blocks, we induce a sparse, block-structured precision matrix, enabling fast, accurate inference via conjugate regression. This reduces computational burden and improves uncertainty quantification for large spatial problems.

Material and Methods

We consider multiple independent replications of a spatial process observed at a set of ordered spatial locations. Each replication is modeled as a realization from a zero-mean multivariate Gaussian distribution with a spatial covariance matrix that captures the dependence structure among locations. To enable scalable inference for large spatial datasets, the joint likelihood is approximated using the Vecchia framework, which represents the joint distribution as a product of conditional distributions based on a fixed number of nearest previously ordered neighbors. This approximation induces

sparsity in the associated precision matrix and yields a modified Cholesky representation, enabling the model to be reformulated as a sequence of independent Bayesian linear regression problems. A fully Bayesian approach is adopted, assigning conjugate normal-inverse-gamma priors to the regression coefficients and the conditional variances. This formulation enables efficient posterior inference through closed-form updates, while the remaining hyperparameters are estimated using an adaptive Metropolis-Hastings algorithm. To further enhance computational efficiency and approximation accuracy, a grouping strategy is employed that partitions spatial locations into disjoint blocks and defines neighbor sets at the group level. By jointly computing likelihood contributions within each group, the proposed grouped Vecchia approximation reduces computational and memory costs while preserving the conditional dependence structure, making it well-suited for large-scale spatial covariance estimation.

Results and Discussion

The performance of the proposed approach was assessed through simulation studies and real data analysis. Simulation results showed that the grouping strategy substantially reduced computational time while achieving lower Frobenius norm errors in covariance and precision matrix estimation. In addition, posterior credible intervals obtained from the grouped approach were consistently narrower, indicating more concentrated posterior distributions and improved uncertainty quantification. These findings were further confirmed by analyses of real data using climate model outputs.

Keywords: Bayesian linear regression, Spatial covariance estimation, Conditional independence, Grouping strategy, Vecchia approximation

Mathematics Subject Classification (2020): 62M30, 62F15, 65C05.



©The Author(s). The Publisher is Iranian Statistical Society.

This is an open access article distributed under the terms and conditions of [\(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)

رویکرد بیزی و گروه‌بندی وکیا در برآورد ناپارامتری کوواریانس فضایی نامانا برای مه‌داده‌ها

فاطمه قاسمی^۱، علی محمدیان مصمم^۲ و خورخه متیو^۳

^۱گروه ریاضی، دانشگاه زنجان

^۲گروه آمار، دانشگاه زنجان

^۳گروه ریاضی، دانشگاه جومه، کاستلون، اسپانیا

نویسنده مسئول: علی محمدیان مصمم، a.m.mosammam@znu.ac.ir

تاریخ دریافت: ۱۴۰۴/۰۰/۰۰ تاریخ بازنگری: ۱۴۰۴/۰۰/۰۰ تاریخ پذیرش و انتشار: ۱۴۰۴/۰۰/۰۰

چکیده: در این مقاله، به برآورد ناپارامتری ساختار همبستگی نامانا در داده‌های فضایی با اندازه بزرگ پرداخته می‌شود. روش پیشنهادی، رویکردی بیزی مبتنی بر گروه‌بندی است که توسعه‌ای از تقریب وکیا محسوب می‌شود و بر فرض استقلال شرطی داده‌های مرتب‌شده بنا شده است. این فرض منجر به تنک‌سازی ماتریس دقت و تجزیه چولسکی تنک می‌گردد و امکان مدل‌سازی فرایند گاوسی n -متغیره را به‌صورت دنباله‌ای از رگرسیون‌های خطی بیزی فراهم می‌آورد. مرتب‌سازی داده‌ها با روش ماکسیم کردن کمترین فاصله، عملکرد مدل را بهبود می‌بخشد. افزون‌براین، اعمال الگوریتم گروه‌بندی روی داده‌های مرتب‌شده با حذف وابستگی‌های ضعیف بین موقعیت‌ها، ساختار کوواریانس را به‌صورت بلوک‌بندی‌شده و فوق‌العاده تنک تعریف می‌کند که منجر به کاهش چشمگیر بار محاسباتی و افزایش دقت مدل می‌شود. نتایج شبیه‌سازی و تحلیل داده‌های واقعی نشان می‌دهد که نمونه‌های به‌دست‌آمده از توزیع پسین، بازه‌های عدم قطعیت کوچکتری نسبت به روش‌های بدون گروه‌بندی دارند.

واژه‌های کلیدی: رگرسیون خطی بیزی، عامل چولسکی، استقلال شرطی، گروه‌بندی، تقریب وکیا

کد موضوع‌بندی ریاضی (۲۰۲۰): 62H11، 60G60 .



©نویسندگان). ناشر انجمن آمار ایران است.

این مقاله با دسترسی آزاد تحت شرایط و ضوابط (CC BY-NC 4.0) توزیع شده است.

۱ مقدمه

در تحلیل‌های آماری کلاسیک، فرض استقلال مشاهدات نقش مهمی در ساده‌سازی مدل‌ها دارد؛ اما در داده‌های فضایی، این فرض اغلب نقض می‌شود. مشاهدات فضایی به‌طور طبیعی وابسته به موقعیت مکانی خود هستند و نزدیکی جغرافیایی منجر به همبستگی بیشتر میان داده‌ها می‌شود. این وابستگی به موقعیت، نیازمند مدل‌سازی دقیق ساختار کوواریانس است تا اطلاعات ارزشمند موجود در داده‌ها از دست نرود (محمدزاده، ۱۳۹۴).

فرآیند گاوسی، ابزار اصلی در زمین‌آمار برای مدل‌سازی وابستگی فضایی بوده است. با این حال، در عصر داده‌های بزرگ، پیچیدگی محاسباتی این فرآیند به‌ویژه در وارون‌سازی ماتریس کوواریانس با مرتبه $O(n^3)$ مانعی جدی برای تحلیل داده‌های با حجم بالا ایجاد کرده است (شابنبرگر و گاتوی، ۲۰۰۴؛ کرسی، ۲۰۱۵؛ بنرجی و همکاران، ۲۰۰۳؛ کرسی و ویکل، ۲۰۱۱). برای مدل‌سازی داده‌های فضایی در مقیاس بزرگ، رویکردهایی نظیر توزیع‌های شرطی مرتبه پایین (استاین و همکاران، ۲۰۰۴؛ وکیا، ۱۹۸۸)، توابع چگالی شبه‌درست‌نمایی (وارین و همکاران، ۲۰۱۱؛ ایدسویک و همکاران، ۲۰۱۴) و مدل‌سازی در دامنه طیفی (فوننتس، ۲۰۰۷) و نیمه‌طیفی (مصمم، ۲۰۲۵) پیشنهاد شده‌اند. افزون‌براین، رویکردهای تقریب مرتبه پایین، مانند کریگینگ رتبه ثابت کرسی و یوهانسون (۲۰۰۸)؛ کانگ و کرسی (۲۰۱۱)؛ کتفز و کرسی (۲۰۱۱) و فرآیندهای پیش‌گوی بنرجی و همکاران (۲۰۰۸) و فیئلی و همکاران (۲۰۰۹) نیز توسعه یافته‌اند. با این حال، حتی در این روش‌ها نیز، در شرایطی که داده‌ها به شدت همبسته هستند یا نوفه زیادی دارند، ممکن است کارایی لازم را نداشته و منجر به از دست رفتن اطلاعات مهم شوند. رویکردهای تنک‌سازی، از جمله کوواریانس نازک‌شده کافمن و همکاران (۲۰۰۸)؛ دو و همکاران (۲۰۰۹)؛ شیبی و روپرت (۲۰۱۲) و همچنین معرفی تنکی در ماتریس دقت رو و هلد (۲۰۰۵)، نیز مورد بررسی قرار گرفته‌اند.

اگرچه این روش‌ها در برآورد پارامترها و درونیایی مؤثرند، اما لزوماً برای استنباط روی فرآیندهای پنهان یا برآورد دقیق عدم قطعیت توسعه نیافته‌اند. در این راستا، رویکرد فرآیند گاوسی نزدیکترین همسایه^۱ (NNGP)، که توسط داتا و همکاران (۲۰۱۶) معرفی شده است، مدلی خوش‌تعریف برای داده‌های فضایی بزرگ فراهم می‌کند که ماتریس‌های دقت تنک ایجاد کرده و به استنباط کارآمد محاسباتی کمک می‌کند. رویکرد NNGP ساختار همسایگی را به‌صورت ثابت و مبتنی بر موقعیت‌های نزدیک تعریف می‌کند که در داده‌هایی با وابستگی فضایی پیچیده یا ناهمگن، منجر به برازش ضعیف و از دست رفتن اطلاعات ساختاری می‌شود. از این‌رو، در این پژوهش روشی معرفی می‌گردد که به توسعه رویکردی انعطاف‌پذیرتر می‌پردازد که می‌تواند ساختار همبستگی را به‌صورت داده محور و قابل کنترل تعریف کند. این رویکرد که بر پایه روش وکیا (۱۹۸۸) بنا نهاده شده و مشاهدات تکراری را نیز در نظر می‌گیرد، امکان استنباط بیزی دقیق و سریع را فراهم می‌آورد (کید و کتفز، ۲۰۲۲؛ مصمم و همکاران، ۱۴۰۱). نوآوری اصلی این پژوهش، معرفی الگوریتم گروه‌بندی موقعیت‌های فضایی به بلوک‌های مستقل است (مصمم و محمدی، ۱۳۹۷)، که با هدف حذف وابستگی‌های غیرضروری و افزایش تنک‌سازی ماتریس کوواریانس طراحی شده است. برخلاف روش‌های قبلی که وابستگی بین تمام موقعیت‌ها را الزاماً حفظ می‌کنند، روش پیشنهادی با تعریف ساختار کوواریانس در سطح

¹Nearest neighbor Gaussian process

بلوک‌ها، امکان مدل‌سازی دقیق‌تر و کنترل‌پذیرتر را فراهم می‌سازد. این طراحی، به‌ویژه در داده‌هایی با وابستگی‌های موضعی یا ساختار فضایی پیچیده، منجر به کاهش چشمگیر تعداد موقعیت‌های فضایی و در نتیجه کاهش قابل‌توجه بار محاسباتی می‌شود. الگوریتم گروه‌بندی با ایجاد ساختار بلوکی-تنگ از ماتریس دقت، در چارچوب تقریب‌های وکیا و ساختارهای تنگ قرار می‌گیرد. نتایج نظری موجود نشان می‌دهند که خطای تقریب کوواریانس و فاصله کولبک-لایبلر میان پسین دقیق و تقریبی با افزایش اندازه همسایگی کاهش می‌یابد و تقریب رفتار پسین سازگاری از خود نشان می‌دهد؛ که در مطالعات وکیا (۱۹۸۸)، گینس (۲۰۱۸) و کترفز و گینس (۲۰۲۱) قابل مشاهده است. علاوه بر این، تقریب‌های بلوکی و کوواریانس‌های نازک‌شده نیز دارای نتایج نظری مشخصی در زمینه خطای تقریب و رفتار پسین هستند. مطالعات سنگ و هوانگ (۲۰۱۲) و نیز بنرجی و همکاران (۲۰۲۳) نشان می‌دهند که خطای تقریب در این روش‌ها با کوچک‌تر شدن اندازه بلوک‌ها و کاهش طول برد وابستگی فضایی کاهش می‌یابد. از آنجا که روش حاضر یک تنگ‌سازی داده‌محور بر ساختار چولسکی اعمال می‌کند، مستقیماً تحت پوشش این تضمین‌های نظری قرار می‌گیرد. نتایج شبیه‌سازی نشان می‌دهد که نمونه‌های به‌دست‌آمده از توزیع پسین، بازه‌های عدم قطعیت کوچک‌تری نسبت به روش‌های بدون گروه‌بندی دارند، که دلیلی بر بهبود دقت آماری و برتری روش پیشنهادی است. از این‌رو، این روش برای مدل‌سازی فرآیندهایی که با چالش ابعاد بالا و ساختارهای وابستگی پیچیده مواجه هستند، بسیار کارآمد، منعطف و قابل‌توسعه است.

در ادامه، مطالب به شرح زیر سازماندهی شده‌اند: بخش ۲ به توصیف روش‌شناسی می‌پردازد. در بخش ۳، مقایسه‌های عددی با استفاده از داده‌های شبیه‌سازی شده ارائه می‌شوند. بخش ۴، به تحلیل حساسیت روش‌های مختلف مرتب‌سازی و پارامتر m پرداخته است. در بخش ۵ روش پیشنهادی روی داده‌های واقعی اجرا شده است و بخش ۶ نیز نتیجه‌گیری را شامل می‌شود.

۲ برآورد ماتریس کوواریانس فضایی

فرض کنید $\mathbf{Y} = \{y_i^{(\ell)} : i = 1, \dots, n, \ell = 1, \dots, N\}$ نشان‌دهنده N تکرار از یک میدان تصادفی مارکوف گاوسی فضایی نامانا باشد که در موقعیت‌های $S = \{s_1, \dots, s_n\}$ مشاهده شده است. موقعیت‌های S بر اساس مرتب‌سازی ماکسیمم کردن کمترین فاصله (شفر و همکاران، ۲۰۲۱ب؛ گینس، ۲۰۱۸) مرتب شده‌اند. در این رویکرد، ابتدا یک نقطه به‌صورت دلخواه یا بر اساس معیاری خاص، انتخاب می‌شود که با $\bar{s}$ نمایش داده می‌شود. سپس نقاط دیگر به طور متوالی به مجموعه اضافه می‌شوند، به‌گونه‌ای که حداقل فاصله بین نزدیک‌ترین نقاط در مجموعه انتخابی افزایش یابد. برای وضوح بیشتر، فرض کنید $\tau : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ یک جایگشت از اعداد ۱ تا n و S^T بردار جایگشت یافته S باشد، به‌طوری‌که $s_i^T = s_{\tau(i)}$. در این صورت شکل ساختاریافته

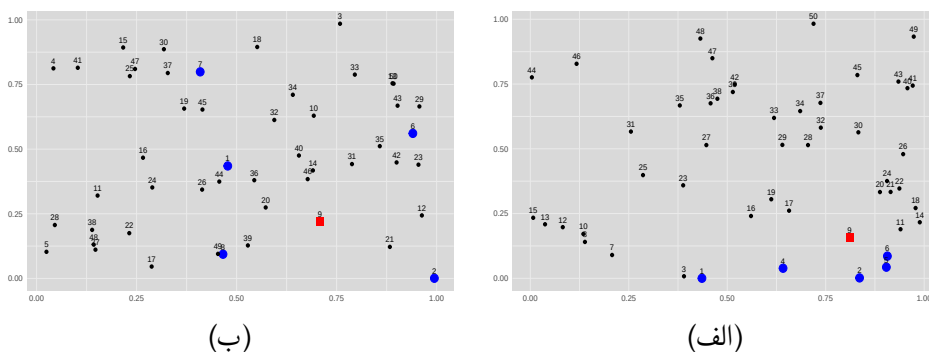
۲۰۸ برآورد ناپارامتری کوواریانس فضایی نامانا برای مه داده‌ها

مرتب‌سازی مذکور به صورت

$$\tau(1) = \arg \max_{i \in \{1, \dots, n\}} \|s_i - \bar{s}\|, \quad \bar{s} = \frac{1}{n} \sum_{i=1}^n s_i$$

$$\tau(j) = \arg \max_{i \in \{1, \dots, n\} \setminus \{\tau(1), \dots, \tau(j-1)\}} \min_{k \in \{1, \dots, j-1\}} \|s_i - s_{\tau(k)}\|, \quad j > 1$$

خواهد بود. با توجه به این مرتب‌سازی، $J_m(i) = \{j_{i1}, \dots, j_{im}\}$ زیرمجموعه‌ای از اعداد صحیح بین ۱ و $i-1$ شامل m همسایه نزدیک به موقعیت s_i است که پیش‌تر مرتب شده‌اند. در شکل ۱ که به مقایسه دو نوع مرتب‌سازی می‌پردازد؛ 50° موقعیت به‌طور تصادفی در مربع واحد انتخاب شده‌اند. نقاط قرمز نشان‌دهنده نقطه هدف (i) هستند، در حالی که نقاط آبی نشان‌دهنده ۵ همسایه نزدیک آن با اندیس‌های کوچکتر هستند. سایر نقاط با رنگ مشکی مشخص شده‌اند. به عنوان مثال، برای $m = 5$ ، مقدار $J_5(9)$ برای مرتب‌سازی بر اساس مختصات، برابر با $(1, 2, 4, 5, 6)$ و برای مرتب‌سازی بر اساس ماکسیم کردن کمترین فاصله، برابر با $(1, 2, 6, 7, 8)$ است.



شکل ۱. مرتب‌سازی بر اساس (الف) مختصات؛ (ب) ماکسیم کردن کمترین فاصله.

پس از مرتب‌سازی موقعیت‌ها، نماد y_i برای نمایش بردار مشاهدات در موقعیت s_i ، در نظر گرفته می‌شود، که به صورت $y_i = (y_i^{(1)}, \dots, y_i^{(N)})'$ مشخص می‌گردد. همچنین، برای هر $\ell = 1, \dots, N$ ، بردار $y^{(\ell)} = (y_1^{(\ell)}, \dots, y_n^{(\ell)})'$ به عنوان مشاهده فضایی در تکرار ℓ ام تعریف می‌شود. $y^{(\ell)}$ به صورت فرآیند گاوسی n - متغیره به صورت

$$y^{(\ell)} \Big| \Sigma \stackrel{iid}{\sim} \mathcal{N}_n(o, \Sigma), \quad \ell = 1, \dots, N, \quad (1)$$

در نظر گرفته شده است. با بهره‌گیری از تقریب وکیا، تابع چگالی توأم مشاهدات که به صورت حاصل ضرب توزیع‌های شرطی قابل بیان است، با زیرمجموعه‌ای از بردارهای شرطی کامل جایگزین می‌شود. تقریب وکیا با فرض استقلال

شرطی مرتب‌شده به صورت

$$p(y_1^{(\ell)}, \dots, y_n^{(\ell)} | \Sigma) = p(y_1^{(\ell)}) \prod_{i=2}^n p(y_i^{(\ell)} | y_{j_{i1}}^{(\ell)}, \dots, y_{j_{im}}^{(\ell)}, \Sigma), \quad \ell = 1, \dots, N, \quad (2)$$

بیان می‌گردد. این ساختار استقلال شرطی مرتب‌شده، نه تنها موجب کاهش پیچیدگی محاسباتی در مدل‌سازی داده‌های بزرگ می‌شود، بلکه به صورت ضمنی ساختار خاصی را برای ماتریس دقت القا می‌کند. این ساختار در قالب تجزیه چولسکی اصلاح‌شده^۱ (MCD) قابل نمایش است که به صورت

$$\Sigma^{-1} = \mathbf{U} \mathbf{D}^{-1} \mathbf{U}' \quad (3)$$

تعریف می‌شود. در این رابطه $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$ یک ماتریس قطری با درایه‌های مثبت d_i ، برای $i = 1, \dots, n$ و $\mathbf{U}$ ماتریس بالا مثلثی با قطر یک است. استقلال شرطی مرتب شده مفروض در (۲) موجب تنک شدن ماتریس $\mathbf{U}$ می‌شود، به طوری که هر ستون $\mathbf{U}$ حداکثر m درایه غیرقطری غیرصفر دارد (کترفز و گینس، ۲۰۲۱). درایه‌های غیرصفر هر ستون $\mathbf{U}$ به صورت $\mathbf{u}_i = \mathbf{U}_{J_m(i), i}$ تعریف می‌شود، که نشان‌دهنده $J_m(i)$ درایه غیرصفر، در ستون i ام ماتریس $\mathbf{U}$ است. بر اساس رابطه (۳)، می‌توان $O(n^2)$ درایه نامعلوم از Σ را با استنتاج متغیرهای $d_1, \dots, d_n$ و $\mathbf{u}_1, \dots, \mathbf{u}_n$ با هزینه محاسباتی $O(nm)$ برآورد کرد. برای دستیابی به این مهم، مدل (۱) را می‌توان به صورت یک سری از n مدل رگرسیون خطی به صورت

$$p(\mathbf{Y} | \Sigma) = p(\mathbf{y}_1) \prod_{i=2}^n p(\mathbf{y}_i | \mathbf{y}_{1:i-1}, \Sigma) = \prod_{i=1}^n \mathcal{N}(\mathbf{y}_i | \mathbf{X}_i \mathbf{u}_i, d_i \mathbf{I}_N) \quad (4)$$

نوشت (هوانگ و همکاران، ۲۰۰۶)، که در آن $\mathbf{I}_N$ ماتریس همانی و $\mathbf{X}_i$ ماتریس طرح، شامل مشاهدات در m موقعیت همسایه $\mathbf{s}_i$ است، که همسایه‌ها توسط تابع $J_m(i)$ تعیین می‌شوند. به طور دقیق‌تر، سطر ℓ ام ماتریس $\mathbf{X}_i$ برابر است با $-y_{J_m(i)}^{(\ell)}$ ، که $y_{J_m(i)}^{(\ell)}$ مقدار مشاهده ℓ ام از همسایگان مکان i ام است. $\mathbf{u}_i$ بردار ضرایب رگرسیون است که نشان‌دهنده تأثیر همسایه‌ها بر $\mathbf{y}_i$ است. d_i یک مقدار اسکالر است که نشان‌دهنده واریانس نوفه است. برای ضرایب رگرسیون $\mathbf{u}_i$ و واریانس d_i ، از توزیع پیشین نرمال-گامای معکوس (NIG) به صورت

$$\mathbf{u}_i | d_i, \boldsymbol{\theta} \stackrel{\text{ind}}{\sim} \mathcal{N}(\circ, d_i \mathbf{V}_i), \quad d_i | \boldsymbol{\theta} \stackrel{\text{ind}}{\sim} \mathcal{FG}(\alpha_i, \beta_i), \quad i = 1, \dots, n, \quad (5)$$

استفاده شده است. در این رابطه، $\boldsymbol{\theta}$ بردار ابرپارامترهاست که پارامترهای $\mathbf{V}_i$ ، α_i و β_i و همچنین طول بردار $\mathbf{u}_i$ یعنی m را مشخص می‌کند. این ابرپارامترها به طور غیرمستقیم پارامترهای توزیع‌های پیشین را کنترل می‌کنند و

¹Modified Cholesky decomposition

۲۱۰ برآورد ناپارامتری کوواریانس فضایی نامانا برای مه داده ها

طراحی آن ها با الهام از رفتار توابع کوواریانس ماترن صورت گرفته است. توزیع های پسین، برآوردهای به روز شده $\mathbf{u}_i$ و d_i را بر اساس داده های مشاهده شده ارائه می دهند، که به صورت

$$p(\mathbf{u}_1, \dots, \mathbf{u}_n, d_1, \dots, d_n \mid \mathbf{Y}, \boldsymbol{\theta}) \propto \prod_{i=1}^n \mathcal{N}(\mathbf{u}_i \mid \hat{\mathbf{u}}_i, d_i \mathbf{G}_i) \mathcal{IG}(d_i \mid \tilde{\alpha}_i, \tilde{\beta}_i) \quad (۶)$$

بیان می شود و برآورد پارامترهای آن نیز عبارت خواهند بود از

$$\begin{aligned} \hat{\mathbf{u}}_i &= \mathbf{G}_i \mathbf{X}_i' \mathbf{y}_i, \quad \mathbf{G}_i = (\mathbf{X}_i' \mathbf{X}_i + \mathbf{V}_i^{-1})^{-1}, \quad \tilde{\alpha}_i = \alpha_i + \frac{N}{\gamma}, \\ \tilde{\beta}_i &= \beta_i + \frac{1}{\gamma} (\mathbf{y}_i' \mathbf{y}_i - \hat{\mathbf{u}}_i' \mathbf{G}_i^{-1} \hat{\mathbf{u}}_i) = \beta_i + \frac{1}{\gamma} \mathbf{y}_i' (\mathbf{I}_N + \mathbf{X}_i \mathbf{V}_i \mathbf{X}_i')^{-1} \mathbf{y}_i. \end{aligned}$$

۲.۱ پارامتر بندی توزیع های پیشین

در این بخش، پارامتر بندی توزیع های پیشین NIG، با توجه به رابطه (۵) برای پارامترهای $\mathbf{u}_i$ و d_i مورد بررسی قرار می گیرد. به جای تعیین مستقیم پارامترهای توزیع NIG، از برداری از ابرپارامترها به صورت $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$ استفاده می شود. در این چارچوب، θ_1 با واریانس حاشیه ای، θ_2 با دامنه، و θ_3 با میزان همواری تابع کوواریانس مرتبط هستند. این ابرپارامترها به گونه ای انتخاب می شوند که ساختار همبستگی بین پارامترهای مدل، مشابه ساختار همبستگی توصیف شده توسط توابع کوواریانس ماترن در مدل های فضایی باشد. برای پارامتر بندی d_i از تابع کوواریانس نمایی با واریانس θ_1 و دامنه $\frac{\gamma}{\theta_2}$ ، استفاده شده است؛ که رابطه $- \theta_3 \|\mathbf{s}_i - \mathbf{s}_j\|$ را $K_{\boldsymbol{\theta}}(\mathbf{s}_i, \mathbf{s}_j) = \theta_1 \exp(- \theta_3 \|\mathbf{s}_i - \mathbf{s}_j\|)$ به صورت d_i شرطی $m = 1$ ، واریانس شرطی d_i به صورت

$$d_i = \text{Var}(y_i^{(\ell)} \mid \mathbf{y}_{J_{m(i)}}^{(\ell)}) = \theta_1 (1 - e^{-\theta_3 \|\mathbf{s}_i - \mathbf{s}_g\|}),$$

محاسبه می شود؛ که در آن $g = J_1(i)$ نزدیک ترین همسایه قبلی به $\mathbf{s}_i$ است. برای شبکه ای منظم روی مکعب واحد $[0, 1]^p$ ، فاصله $\|\mathbf{s}_i - \mathbf{s}_g\|$ به صورت منظم به اندازه $(i)^{-1/p}$ کاهش می یابد، بنابراین مقدار پیشین d_i به سمت رابطه $d_i \approx \theta_1 (1 - e^{-\theta_3 (i)^{-1/p}})$ میل می کند. مطالعات تجربی نشان داده اند که این فرم تابعی برای توابع ماترن در ابعاد بالاتر نیز تقریباً برقرار است. لذا میانگین پیشین به صورت

$$E(d_i \mid \boldsymbol{\theta}) = \beta_i / (\alpha_i - 1) = \theta_1 (1 - e^{-\theta_3 (i)^{-1/p}}) = \theta_1 f_{\theta_3}(i) \quad (۷)$$

تعریف می شود؛ که در آن $f_{\theta_3}(i) = (1 - e^{-\theta_3 (i)^{-1/p}})$ در نظر گرفته می شود. از رابطه d_i با i و با توجه به واریانس مشاهدات تجربی برای عناصر d_i ، انحراف معیار پیشین d_i برابر با نصف میانگین در نظر گرفته شده است، که با جایگذاری در رابطه بین میانگین و انحراف معیار، $\alpha_i = 6$ و $\beta_i = 5\theta_1 (1 - e^{-\theta_3 (i)^{-1/p}})$

حاصل خواهد شد. در ادامه، برای پارامتر بندی $\mathbf{u}_i$ از ماتریس قطری $\mathbf{V}_i = \text{diag}(v_{i1}, \dots, v_{im})$ استفاده شده است، که در آن $v_{ij} = \exp(-\theta_3 j) / (\theta_1 f_{\theta_1}(i))$. این فرم تابعی با توجه به نتایج حاصل از مسائل مقدار مرزی بیضوی (شفر و همکاران، ۲۰۲۱b) و رفتار کاهش همبستگی در توابع ماترن انتخاب شده است. به عبارت دیگر، تأثیر همسایگان دورتر به صورت نمایی کاهش می‌یابد و همسایگان نزدیک‌تر نقش بیشتری در پیش‌بینی دارند. مقدار پارامتر m به صورت داده‌محور تعیین می‌شود و برابر با بزرگ‌ترین مقدار j است که در رابطه $\exp(-\theta_3 j) > 0.001$ صدق می‌کند. در این رابطه، j نشان‌دهنده تعداد همسایه‌هاست و مقدار θ_3 از داده‌ها قابل یادگیری است. این مکانیسم انتخاب، تضمین می‌کند که تنها موقعیت‌هایی با وابستگی معنادار به عنوان همسایه در نظر گرفته شوند، که منجر به کاهش مؤثر بار محاسباتی و حفظ دقت مدل می‌شود. برای تعیین مقادیر مناسب θ ، از روش ماکسیم درستنمایی حاشیه‌ای استفاده می‌شود، که با ترکیب روابط (۴) و (۵) به دست می‌آید. با خارج کردن ضرایب از انتگرال و انتگرال‌گیری نسبت به $\mathbf{u}_i$ و d_i ، و با استفاده از نتایج استاندارد مدل‌های گاوسی مزدوج، عبارت

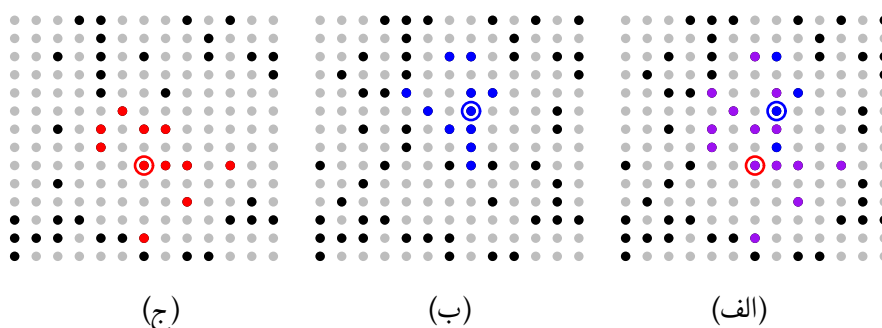
$$p(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^n \int_{d_i} \int_{\mathbf{u}_i} \mathcal{N}(\mathbf{y}_i | \mathbf{X}_i \mathbf{u}_i, d_i \mathbf{I}_N) \mathcal{N}(\cdot, d_i \mathbf{V}_i) \mathcal{FG}(\alpha_i, \beta_i) d\mathbf{u}_i dd_i \\ \propto \prod_{i=1}^n \left(\frac{|\mathbf{G}_i|^{1/2}}{|\mathbf{V}_i|^{1/2}} \cdot \frac{\beta_i^{\alpha_i}}{\tilde{\beta}_i^{\alpha_i}} \cdot \frac{\Gamma(\tilde{\alpha}_i)}{\Gamma(\alpha_i)} \right)$$

حاصل می‌شود (Γ تابع گاما است). اگر برای $\boldsymbol{\theta}$ توزیع پیشین یکنواخت در نظر گرفته شود، تابع درستنمایی حاشیه‌ای و تابع چگالی پسین یکسان خواهند بود، یعنی $P(\mathbf{Y}|\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}|\mathbf{Y})$. در نتیجه، استنباط بیزی کامل و تجربی هر دو رویکردهای مستقیمی برای حل مسائل بیزی ارائه می‌دهند. از آن‌جا که توزیع پسین $P(\boldsymbol{\theta}|\mathbf{Y})$ به صورت تحلیلی قابل محاسبه نیست، از نسخه تطبیقی الگوریتم متروپلیس-هستینگز (MH) برای نمونه‌برداری از این توزیع استفاده می‌شود. در این مطالعه، از پیاده‌سازی موجود در نرم‌افزار R توسط **شیدگر (۲۰۱۲)** استفاده شده است.

پس از تولید نمونه‌هایی از بردار ابرپارامترهای $\boldsymbol{\theta}$ توسط الگوریتم MH، پارامترهای توزیع پسین قابل برآورد خواهند بود. این عمل در داده‌های بزرگ هزینه محاسباتی بالایی دارد. با اعمال گروه‌بندی بر روی موقعیت‌ها، اطلاعات موجود از بین نمی‌رود، اما تعداد موقعیت‌ها و در نتیجه تعداد ماتریس‌های کوواریانس مورد نیاز برای برآورد کاهش می‌یابد. در فرآیند گروه‌بندی، تعداد موقعیت‌ها از n نقطه به K بلوک کاهش می‌یابد. در این حالت، تعداد همسایگان هر بلوک معمولاً بیشتر از تعداد همسایگان هر موقعیت در حالت بدون گروه‌بندی است. به جای بررسی جداگانه هر موقعیت، موقعیت‌ها در قالب بلوک‌ها تحلیل می‌شوند. در واقع، گروه‌بندی امکان دستیابی به تقریب‌های دقیق‌تر را بدون افزایش هزینه، و حتی با کاهش آن، فراهم می‌سازد.

۲.۲ گروه بندی و مجموعه همسایگان جدید

فرض کنید $B = \{B_1, \dots, B_K\}$ یک افراز از مجموعه $\{1, \dots, n\}$ باشد. این افراز، نشان دهنده گروه بندی مشاهدات به بلوک ها است. برای هر بلوک $B_k, k = 1, \dots, K$ ، به صورت $U_k = \bigcup_{i \in B_k} J_i$ تعریف می شود. در این رابطه، U_k ، اجتماع همسایگان تمام مشاهدات موجود در بلوک B_k و J_i همسایگان مرتب شده موقعیت s_i است. سپس، فرض کنید اندیس i عضوی از بلوک B_k باشد؛ مجموعه همسایگان جدید $\bar{J}_i$ به صورت $\bar{J}_i = \{j \in U_k : j \leq i\}$ تعریف می شود. این تعریف بدان معناست که $\bar{J}_i$ شامل تمام همسایگان مشاهدات در بلوک B_k است، با این شرط اضافی که تمام عناصر $\bar{J}_i$ باید کوچک تر یا مساوی i باشند. بنابراین، $\bar{J}_i$ مجموعه ای از اعداد صحیح بین ۱ تا i است که i را نیز شامل می شود. با این تعریف، $\bar{J}_i$ همچنان با قاعده اصلی وکیا (که همسایگان باید از نقاط قبلی در ترتیب بندی باشند و خود نقطه نیز جزء همسایگان خودش باشد) سازگار است و مجموعه همسایگان جدید $\bar{J}_i$ به انتخاب های همسایگان اصلی $J_1, \dots, J_n$ و همچنین به افراز B بستگی دارد. نماد $p_{\theta, J}$ یک تقریب از رابطه (۲) ارائه می دهد که از این پس، به جای استفاده مستقیم از فرمول (۲)، از همین نماد استفاده خواهد شد. همچنین $J = \{J_1, \dots, J_n\}$ به عنوان یک مجموعه از زیر مجموعه های $J_i := \{j_{i1}, \dots, j_{im}, i\}$ که به عنوان همسایگان مشاهده i در نظر گرفته می شوند، تعریف می شود. تقریب جدید $p_{\theta, \bar{J}}$ نسخه گروه بندی شده تقریب بدون گروه بندی $p_{\theta, J}$ نامیده می شود. شکل ۲ مثالی از نحوه ساخت $\bar{J}_i$ را برای گروهی که شامل دو نقطه است، نشان می دهد. دایره قرمز موقعیت با اندیس ۴۵ و دایره آبی موقعیت با اندیس ۶۰ را مشخص می کند و نقاط هم رنگ، همسایگان متناظر هر نقطه را نشان می دهند. نقاط مشکی مجموعه ای از نقاط با اندیس های ۱ تا ۶۰ را تشکیل می دهند. در نسخه گروه بندی شده، همسایگان نقطه ۴۵ با رنگ بنفش و همسایگان نقطه ۶۰ با ترکیبی از رنگ های آبی و بنفش نمایش داده شده اند که منجر به افزایش تعداد همسایگان نقطه ۴۵ به ۳ و نقطه ۶۰ به ۷ همسایه می شود.



شکل ۲. مراحل گروه بندی: (الف) مجموعه نقاط ۱ تا ۶۰ و گروه بندی نقاط (ب) همسایگان نقطه ۶۰، (ج) همسایگان نقطه ۴۵

در ادامه این بخش، با استفاده از تجزیه چولسکی ماتریس دقت ارائه شده در رابطه (۳)، تأثیر گروه بندی مشاهدات

بر دقت مدل مورد بررسی قرار می‌گیرد. ماتریس غیرهمبسته‌کننده Ψ ، به صورت $\Psi = \mathbf{D}^{-1/2} \mathbf{U}'$ تعریف می‌شود، که در آن $\Psi_{ii} = d_i^{-1/2}$ و $\Psi_{ij} = d_i^{-1/2} \mathbf{U}'_{ij}$ است. این ماتریس امکان تبدیل داده‌ها به بردارهای غیرهمبسته‌شده $\mathbf{z}^{(\ell)} = \Psi \mathbf{y}^{(\ell)}$ را فراهم می‌آورد. در این تبدیل، هر مؤلفه $z_i^{(\ell)}$ به‌عنوان باقی‌مانده استاندارد شده از بهترین پیش‌بینی نااریب خطی^۱ (BLUP) $y_i^{(\ell)}$ با استفاده از مشاهدات $(y_1^{(\ell)}, \dots, y_{i-1}^{(\ell)})$ تفسیر می‌شود. برای هر $i = 2, \dots, n$ $z_i^{(\ell)}$ ها به صورت

$$z_i^{(\ell)} = \sum_{j=1}^i \Psi_{ij} y_j^{(\ell)} = \Psi_{ii} \left[y_i^{(\ell)} - \sum_{j=1}^{i-1} \left(\frac{-\Psi_{ij}}{\Psi_{ii}} \right) y_j^{(\ell)} \right], \quad \ell = 1, \dots, N \quad (8)$$

بیان می‌شوند. ضرایب رگرسیون در این پیش‌بینی نقش وزن‌های بهینه در BLUP را ایفا می‌کنند. تعبیری که می‌تواند به یک تابع میانگین خطی تعمیم داده شود (استاین و همکاران، ۲۰۰۴). از آنجایی که $z_i^{(\ell)}$ ها غیرهمبسته هستند، می‌توان چگالی توأم آن‌ها را به صورت حاصل ضرب چگالی حاشیه‌ای $z_i^{(\ell)}$ ها نوشت. با این تفاسیر، چگالی شرطی $y_i^{(\ell)}$ به شرط $y_1^{(\ell)}, \dots, y_{i-1}^{(\ell)}$ به صورت

$$\log p_{\theta}(y_i^{(\ell)} | y_1^{(\ell)}, \dots, y_{i-1}^{(\ell)}) = -\frac{1}{2} \log 2\pi + \log \Psi_{ii} - \frac{1}{2} (z_i^{(\ell)})^2,$$

نوشته می‌شود و بنابراین سطر i ام Ψ خلاصه‌ای از اطلاعات مربوط به چگالی احتمال شرطی $y_i^{(\ell)}$ با توجه به $y_1^{(\ell)}$ تا $y_{i-1}^{(\ell)}$ را نتیجه می‌دهد. تقریب وکیا ماتریس Ψ را با یک نسخه تنک $\tilde{\Psi}$ جایگزین می‌کند، که در آن بیشتر عناصر صفر هستند. به عبارت دیگر، $y_i^{(\ell)}$ بر روی یک زیرمجموعه محدود از مشاهدات قبلی بازگشت داده می‌شود. اگر $y_j^{(\ell)}$ در مجموعه شرطی‌سازی برای $y_i^{(\ell)}$ نباشد، عنصر $\tilde{\Psi}_{ij}$ صفر خواهد بود. به دلیل تفسیر BLUP، واریانس برآورد $\tilde{\Psi}_{ii}$ همیشه کمتر یا مساوی واریانس واقعی Ψ_{ii} است.

فرض کنید $\tilde{\Psi}^1$ و $\tilde{\Psi}^2$ دو تقریب وکیا هستند که از یک مرتب‌سازی برای مشاهدات استفاده می‌کنند، اما مجموعه‌های شرطی‌سازی متفاوتی به ترتیب J^1 و J^2 دارند. اگر مجموعه شرطی‌سازی J_i^1 زیرمجموعه‌ای از J_i^2 ، برای هر i باشد، آنگاه $\tilde{\Psi}_{ii}^1$ کمتر یا مساوی $\tilde{\Psi}_{ii}^2$ خواهد بود. این ویژگی به این دلیل است که با اضافه کردن متغیرهای بیشتر به مجموعه شرطی‌سازی، واریانس BLUP نمی‌تواند افزایش یابد، در نتیجه وارون آن افزایش می‌یابد.

واگرایی کولبک-لایبلر^۲ (KL) معیاری برای سنجش میزان تفاوت بین دو توزیع احتمال p_1 و p_2 است. این معیار با فرمول $KL(p_2 || p_1) = E(\log(\frac{p_2}{p_1}))$ تعریف می‌شود، که در آن E امید ریاضی نسبت به توزیع p_2 است. قضیه ۱ بیان می‌کند که این واگرایی با اضافه شدن داده‌ها به مجموعه‌های شرطی‌سازی کاهش پیدا می‌کند. این نتیجه برای اثبات این موضوع که تقریب گروهی می‌تواند از تقریب غیرگروهی بهتر باشد، مورد استفاده قرار می‌گیرد.

قضیه ۱. اگر برای هر $i = 1, \dots, n$ ، $J_i^1 \subset J_i^2$ باشد، آنگاه $KL(p_{\theta} || p_{\theta, J^2}) \leq KL(p_{\theta} || p_{\theta, J^1})$.

¹Best linear unbiased predictor

²Kullback Leibler

۲۱۴ برآورد ناپارامتری کوواریانس فضایی نامانا برای مه داده ها

فرع ۱. تقریب گروه بندی شده، KL کوچک تری نسبت به همتای گروه بندی نشده خود دارد. یعنی

$$KL(p_{\theta}||p_{\theta,\bar{J}}) \leq KL(p_{\theta}||p_{\theta,J}).$$

۲.۳ الگوریتم گروه بندی

هدف اصلی، دسته بندی مشاهدات به گروه های مجزا به گونه ای است که هم دقت مدل افزایش یابد و هم بار محاسباتی کاهش پیدا کند. بدین منظور، یک الگوریتم حریصانه سریع پیشنهاد می شود که تنها بر ترتیب داده ها و ساختار همسایگی هر مشاهده تکیه دارد. در این روش، هر مشاهده در ابتدا در یک گروه مستقل قرار می گیرد و سپس الگوریتم به صورت مرحله ای، جفت بلوک هایی را که شرایط مشخصی را برآورده می کنند، با یکدیگر ادغام می کند.

الگوریتم ۱. الگوریتم گروه بندی موقعیت های فضایی:

ورودی: آرایه همسایگی $NNarray \in \mathbb{N}^{n \times (m+1)}$

خروجی: افراز مجموعه مشاهدات به بلوک های $\{B_k\}$

گام ۱- مقدار اولیه بلوک ها به صورت $B = \{\{1\}, \dots, \{n\}\}$ تنظیم شود.

گام ۲- برای هر $\ell = 1$ تا m ، مراحل زیر انجام شود:

گام ۱.۲- برای هر $i = 1$ تا n ، مراحل زیر انجام شود:

- بلوک B_k شامل i شناسایی شود، به طوری که $i \in B_k$.

- مقدار $j = j_{i\ell}$ از $NNarray$ استخراج شود.

- بلوک $B_{k'}$ شامل j شناسایی شود، به طوری که $j \in B_{k'}$.

- اگر $k \neq k'$ ، آنگاه:

- مجموعه همسایگی ها به صورت $U = U_k \cup U_{k'}$ تعریف شود.

- اگر شرط $(\#U)^2 \leq (\#U_k)^2 + (\#U_{k'})^2$ برقرار بود، آنگاه:

- بلوک B_k به صورت $B_k \leftarrow B_k \cup B_{k'}$ به روزرسانی شود.

- بلوک $B_{k'}$ تهی شود، یعنی $B_{k'} \leftarrow \emptyset$.

۲.۴ پیچیدگی محاسباتی

روش پیشنهادی با بهره گیری از ساختار ماتریس کوواریانس و عامل چولسکی معکوس آن، امکان محاسبه هم زمان سهم مشاهدات یک گروه را فراهم می کند و منجر به کاهش قابل توجهی در بار محاسباتی و حافظه می شود. در روش گروه بندی، مجموعه مشاهدات به گروه های $B_1, \dots, B_K$ تقسیم می شوند و برای هر گروه B_k مجموعه همسایگان مشترک U_k استفاده می شود. با این کار، عملیات مربوط به ساخت و فاکتورگیری ماتریس کوواریانس تنها یک بار برای هر گروه انجام می شود، نه برای تک تک مشاهدات.

پیچیدگی حافظه در نسخه گروه‌بندی شده برابر است با $\sum_{k=1}^K (\#U_k)^2$ و تعداد عملیات ممیز شناور برابر با $\sum_{k=1}^K (\#U_k)^3$ است. در مقابل، نسخه بدون گروه‌بندی که برای هر مشاهده از مجموعه همسایگی با اندازه $m+1$ استفاده می‌کند، هزینه حافظه به صورت $n(m+1)^2$ و تعداد عملیات ممیز شناور به صورت $n(m+1)^3$ است. بنابراین، هرگاه چند مشاهده دارای همسایگان مشترک باشند، مقدار $(\#U_k)$ به طور قابل توجهی کوچک‌تر از $m+1$ شده و در نتیجه دو مجموع $\sum_k (\#U_k)^2$ و $\sum_k (\#U_k)^3$ به میزان چشم‌گیری کاهش می‌یابد. این موضوع منجر به صرفه‌جویی قابل توجه در حافظه و سرعت اجرا می‌شود.

جدول ۱. مقایسه پیچیدگی محاسباتی نسخه گروه‌بندی شده و نسخه بدون گروه‌بندی

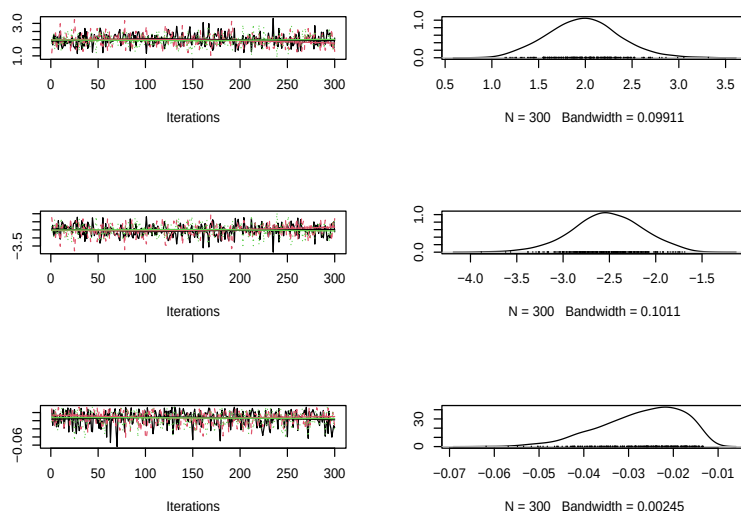
روش	هزینه حافظه	تعداد عملیات ممیز شناور (FLOPs)
بدون گروه‌بندی	$n(m+1)^2$	$n(m+1)^3$
گروه‌بندی شده	$\sum_{k=1}^K (\#U_k)^2$	$\sum_{k=1}^K (\#U_k)^3$

۳ مطالعه شبیه‌سازی

برای ارزیابی عملکرد روش پیشنهادی برآورد ماتریس کوواریانس، ۲۰ تکرار مستقل از یک فرآیند گاوسی با تابع کوواریانس ماترن شبیه‌سازی شد. پارامترهای مدل شامل واریانس حاشیه‌ای برابر با ۳، همواری برابر با ۱، و دامنه برابر با ۲۵٪ بودند. موقعیت‌های فضایی به صورت تصادفی از مربع واحد $[0, 1] \times [0, 1]$ انتخاب شدند و شبیه‌سازی‌ها در ۹۰۰ نقطه انجام شد. برای نمونه‌گیری از توزیع پسین ابرپارامترهای θ ، از الگوریتم تطبیقی MH مطابق با رویکرد ارائه شده در **کید و کتفز (۲۰۲۲)** استفاده شد. سه زنجیره مارکوف مونت کارلو اجرا گردید و از هر زنجیره ۳۰۰۰۰ نمونه پسین استخراج شد. پس از حذف ۱۵۰۰۰ نمونه ابتدایی به عنوان دوره سوختن، نمونه‌های باقی‌مانده با فاکتور ۵۰ نازک‌سازی شدند و در نهایت ۹۰۰ نمونه مستقل حاصل گردید. سپس نمونه‌ها به اشیای MCMC تبدیل شدند تا بتوان همگرایی چندزنجیره‌ای را ارزیابی کرد. نمودارهای اثر برای هر زنجیره و هر پارامتر در شکل ۳ نشان می‌دهد که زنجیره‌ها به خوبی ترکیب شده‌اند و هیچ روند قابل توجهی در طول نمونه‌گیری مشاهده نمی‌شود. ردیف اول، دوم و سوم نمودار به ترتیب مربوط به θ_1 ، θ_2 و θ_3 است. نمودارهای اثر نشان‌دهنده همگرایی پایدار و رفتار ترکیب مناسب هستند، در حالی که نمودارهای چگالی، توزیع‌های پسین مشخصی را با پهنای باند انتخاب شده به طور دقیق نمایش می‌دهند و اعتبار فرآیند نمونه‌گیری و پایداری برآوردهای پارامتر را تأیید می‌کنند.

برای ارزیابی همگرایی زنجیره‌های مارکوف مونت‌کارلو، از معیار کاهش مقیاس بالقوه^۱ (PSRF)، استفاده شد که توسط **گلن و روبین (۱۹۹۲)** معرفی شده است. این معیار با مقایسه واریانس درون‌زنجیره‌ای و بین‌زنجیره‌ای، میزان همگرایی به توزیع هدف را ارزیابی می‌کند. مقدار گزارش شده برای زنجیره‌ها ۱/۰۱ بود. مقدار بسیار نزدیک

¹Potential scale reduction factor



شکل ۳. نمودارهای اثر و چگالی پسین برای سه ابرپارامتر مدل در طول ۳۰۰ تکرار الگوریتم MCMC

به ۱، یک شاخص قابل قبول برای تأیید همگرایی زنجیره‌های مارکوف مونت کارلو و نشان‌دهنده ترکیب مناسب آن‌ها است. مقدار برآورد نقطه‌ای PSRF برای سه پارامتر مدل به ترتیب برابر با $1/003$ ، $1/002$ و $0/998$ بوده که نشان‌دهنده همگرایی مناسب زنجیره‌ها است. اندازه نمونه موثر^۱ (ESS) برای پارامترهای اول و دوم برابر با ۹۰۰ و برای پارامتر سوم برابر با ۸۱۵/۲۹ است. این مقادیر نشان‌دهنده کیفیت مناسب نمونه‌گیری و کاهش وابستگی بین نمونه‌ها در زنجیره‌های مارکوف مونت کارلو است.

بعد از اطمینان از به‌دست آوردن نمونه‌های پسین مناسب از ابرپارامترهای θ ، از این نمونه‌های تولید شده برای تولید نمونه‌های جدید از توزیع پسین، یعنی رابطه (۶) استفاده گردید. برای برآورد درایه‌های ماتریس دقت از دو رویکرد مختلف استفاده شد. رویکرد اول بر اساس روش کید و کتفز (۲۰۲۲) بود که در آن موقعیت‌ها با روش ماکسیم‌کردن کمترین فاصله، مرتب شدند و ماتریس همسایگی بر اساس موقعیت‌های مرتب‌شده تشکیل گردید. در رویکرد دوم، از روش گروه‌بندی بعد از مرتب‌سازی داده‌ها استفاده شد. نتایج در جدول ۲ نشان‌دهنده عملکرد بهتر روش گروه‌بندی، از نظر سرعت در برآورد درایه‌های ماتریس دقت است. افزون‌براین، خطای نسبی نرم فروبنیوس^۲ (دریناس و همکاران، ۲۰۰۶) برای روش گروه‌بندی کمتر از روش کید و کتفز است. این بهبود عملکرد در کاربردهایی که نیاز به محاسبات سریع و دقیق دارند، اهمیت پیدا می‌کند.

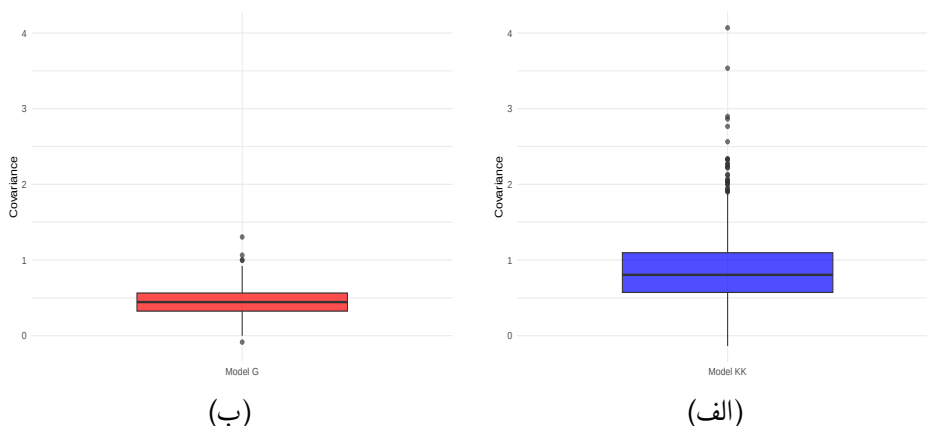
برای مقایسه کوواریانس‌های برآوردشده به روش کید و کتفز و رویکرد گروه‌بندی، از نمودارهای جعبه‌ای به‌عنوان یک معیار گرافیکی استفاده شد. با توجه به تنگ بودن ماتریس کوواریانس حاصل از روش گروه‌بندی، یک درایه

^۱Effective sample size

^۲Frobenius norm error

جدول ۲. مقایسه زمان و خطای نرم فروبنیوس روش‌های برآورد ماتریس کوواریانس		
روش	زمان برآورد (دقیقه)	خطای نسبی نرم فروبنیوس
کید و کتفز	۵/۵	۱/۰۶۶۸
گروه‌بندی	۳	۰/۹۹۲۵

غیرصفر به صورت تصادفی انتخاب و مقدار متناظر آن در هر یک از ۹۰۰ ماتریس کوواریانس برآورد شده استخراج گردید. این فرآیند به طور مشابه برای روش کید و کتفز نیز انجام شد. نمودارهای جعبه‌ای مقادیر حاصل در شکل ۴ ارائه شده‌اند. نتایج نشان می‌دهد که رویکرد گروه‌بندی توزیعی متمرکز با میانه ۰/۵ و پراکندگی بسیار کم دارد که بیانگر واریانس پایین و قطعیت بالای مدل است. در مقابل، روش کید و کتفز پراکندگی بیشتری با میانه ۰/۸ نشان می‌دهد که حاکی از واریانس بالاتر و عدم قطعیت بیشتر در برآوردهاست. این مقایسه عملکرد پایدارتر و اطمینان بالاتر رویکرد گروه‌بندی را در تخمین کوواریانس تأیید می‌کند.

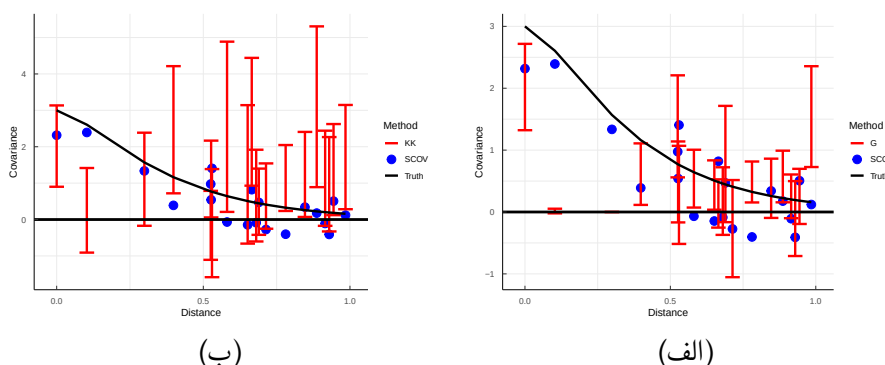


شکل ۴. نمودارهای جعبه‌ای برآورد یک درایه تصادفی از ماتریس کوواریانس برای (الف) روش کید و کتفز و (ب) رویکرد گروه‌بندی

برای تحلیل دقیق‌تر عدم قطعیت موجود در برآوردهای ماتریس کوواریانس، از بازه‌های باورمندی^۱ پسین استفاده شده است. در این راستا، دو رویکرد مدل‌سازی شامل روش گروه‌بندی و روش کید و کتفز مورد بررسی قرار گرفته‌اند. به منظور ارزیابی عملکرد این دو روش، ۲۰ تکرار از فرآیند گاوسی با تابع کوواریانس ماترن در ۹۰۰ موقعیت تصادفی تولید شده‌اند. سپس برای هر تکرار، ماتریس کوواریانس نمونه‌ای محاسبه گردیده و بازه‌های باورمندی ۹۵ درصد برای درایه‌های منتخب از ماتریس کوواریانس استخراج شده‌اند. برآوردهای نمونه‌ای، مقادیر کوواریانس را نشان می‌دهند که مستقیماً از داده‌های حاصل از نمونه‌برداری فرآیند گاوسی محاسبه شده‌اند. در مقابل، بازه‌های اعتبار ۹۵ درصد

¹Credible interval

پسین، دامنه‌ای را مشخص می‌کنند که با احتمال ۹۵ درصد، مقدار واقعی کوواریانس بین دو موقعیت تصادفی در آن قرار دارد. این بازه‌ها، میزان عدم قطعیت موجود در برآوردهای مدل بیزی را منعکس می‌سازند و امکان مقایسه دقت و تمرکز هر روش را فراهم می‌کنند (شکل ۵). روش گروه‌بندی با ارائه بازه‌های اعتبار باریک‌تر، برآوردهایی با تمرکز



شکل ۵. برآورد کوواریانس با روش (الف) گروه‌بندی، (ب) کید و کتفز

بالتر و قابلیت اعتماد بیشتر فراهم می‌آورد. این ویژگی به‌ویژه در کاربردهایی که دقت در برآورد روابط بین متغیرها اهمیت دارد، از جمله در مدل‌سازی فضایی و تحلیل داده‌های مکانی، بسیار حائز اهمیت است. در مقابل، روش کید و کتفز با بازه‌های اعتبار گسترده‌تر، نشان‌دهنده حساسیت بیشتر به تغییرات داده و نوسانات فرآیند تصادفی است. این تفاوت‌ها می‌توانند راهنمایی مؤثری در انتخاب روش مناسب برای تحلیل‌های بیزی در مسائل مختلف فراهم کنند.

۴ تحلیل حساسیت

۴.۱ پایداری مدل نسبت به روش‌های مختلف مرتب‌سازی

به‌منظور بررسی حساسیت روش گروه‌بندی نسبت به ترتیب نقاط، سه روش مرتب‌سازی شامل مرتب‌سازی مبتنی بر ماکسیمم‌سازی کمترین فاصله، مرتب‌سازی بر اساس مؤلفه طول جغرافیایی، و مرتب‌سازی مبتنی بر همبستگی (کید و کتفز، ۲۰۲۲) مورد ارزیابی قرار گرفتند. در این راستا، ابتدا ۷۵۰ نمونه از توزیع پسین ابرپارامترهای مدل با استفاده از الگوریتم MCMC استخراج شد. برای هر مجموعه نمونه، ماتریس دقت $Q = UU'$ و سپس ماتریس کوواریانس متناظر $\Sigma = Q^{-1}$ محاسبه گردید. برای ارزیابی اثر نوع مرتب‌سازی بر استنباط‌های پسین، فاصله اقلیدسی بین تمام جفت‌های غیرقطری نقاط محاسبه شد و ۵۰۰ جفت با کمترین فاصله انتخاب گردیدند. سپس برای هر یک از این جفت‌ها، بازه‌های باورمندی ۹۵٪ مبتنی بر توزیع پسین عناصر کوواریانس محاسبه شد. این بازه‌ها بیانگر عدم قطعیت پسین در برآورد کوواریانس بین نقاط بوده و به پارامترهای ماترن یا سایر پارامترهای ساختاری

مدل وابسته نیستند. نتایج نشان داد که نرخ پوشش تجربی بازه‌های باورمندی ۹۵٪، که بر اساس ۵۰۰ جفت منتخب محاسبه شده‌اند، در حدود ۷۵٪ قرار دارد. نتایج حاصل از سایر روش‌های مرتب‌سازی نیز رفتار مشابهی را نشان می‌دهند. این یافته‌ها با نتایج نظری گزارش شده در مطالعات پیشین (فورر و همکاران، ۲۰۰۶؛ داتا و همکاران، ۲۰۱۶) سازگار است؛ به‌طوری‌که در چارچوب تقریب‌ها و روش‌های نازک‌سازی، انتظار می‌رود بازه‌های عدم قطعیت از سطح اسمی فاصله بگیرند، در حالی‌که همچنان اطلاعات معناداری درباره عدم قطعیت پسین فراهم می‌کنند. علاوه بر این، مقایسه میان روش‌های مرتب‌سازی نشان داد که دو روش ماکسیم‌سازی کمترین فاصله و مرتب‌سازی مبتنی بر همبستگی به‌طور میانگین منجر به بازه‌های باورمندی باریک‌تری شده و عملکرد بهتری نسبت به روش مبتنی بر طول جغرافیایی دارند (جدول ۳).

جدول ۳. مقایسه روش‌های مرتب‌سازی			
روش مرتب‌سازی	خطای نرم فروبنیوس	میانگین عرض بازه‌ها	زمان اجرا (دقیقه)
ماکسیم کمترین فاصله	۰/۹۹۳	۰/۲۰	۱۲:۳۱
طول جغرافیایی	۰/۸۹۹	۱/۲۸	۰۹:۲۴
همبستگی	۰/۹۹۴	۰/۲۹	۱۱:۳۹

۴.۲ تحلیل حساسیت نسبت به پارامتر m

در روش‌های مبتنی بر تقریب وکیا و نیز در ساختارهای چولسکی معکوس، پارامتر m تعداد همسایگان مورد استفاده در شرطی‌سازی را تعیین می‌کند. انتخاب این پارامتر اهمیت زیادی دارد؛ زیرا اگر m بسیار کوچک باشد، بخشی از ساختار وابستگی فضایی بازسازی نمی‌شود و دقت پیش‌بینی کاهش می‌یابد. در مقابل، اگر m بسیار بزرگ باشد، تعداد پارامترها و هزینه محاسباتی به‌طور چشمگیری افزایش یافته و سرعت کاهش می‌یابد، بدون آنکه دقت بهبود قابل‌توجهی پیدا کند. بنابراین ثابت در نظر گرفتن m همواره مناسب نیست و لازم است داده‌ها در تعیین مقدار بهینه آن نقش داشته باشند. در این مقاله همانند مقاله کید و کترفز این نکته مورد توجه قرار گرفته و مقدار m بر اساس نرخ کاهش ضرایب چولسکی تعیین می‌شود؛ به‌گونه‌ای که همسایه‌ای با واریانس پیشین کمتر از آستانه، اطلاعات جدیدی اضافه نمی‌کند. برای ارزیابی پایداری مدل، پارامتر m در روش ماکسیم کردن کمترین فاصله در چند مقدار مختلف تنظیم شد و خطای فروبنیوس، میانگین عرض بازه‌های باورمندی ۹۵٪ و زمان اجرا محاسبه گردید (جدول ۴). نتایج نشان می‌دهد که افزایش m از ۵ به ۱۰ دقت مدل را به‌طور محسوسی بهبود می‌دهد، اما مقادیر بزرگ‌تر تغییرات اندک ایجاد می‌کنند و زمان اجرا افزایش قابل توجهی دارد. بنابراین پس از حد مشخصی، افزودن همسایگان بیشتر اطلاعات جدیدی به مدل اضافه نمی‌کند و تنها هزینه محاسباتی را بالا می‌برد.

جدول ۴. تحلیل حساسیت پارامتر m از نظر دقت و هزینه محاسباتی برای ۵۰۰ جفت اندازه همسایگی m خطای نرم فروبنیوس میانگین عرض بازه‌ها زمان اجرا (دقیقه)

۵	۱/۱۸۷	۰/۴۳	۰۷:۱۸
۱۰	۰/۹۹۶	۰/۲۵	۱۰:۵۴
۱۵	۰/۹۹۴	۰/۲۲	۱۳:۲۱
۲۰	۰/۹۹۳	۰/۲۲	۱۷:۴۸

۵ تحلیل داده‌های واقعی

برای ارزیابی عملکرد عملی روش پیشنهادی، از داده‌های مدل اقلیمی^۱ (CESM) استفاده شد. این داده‌ها بخشی از پروژه مجموعه بزرگ هستند که توسط کی و همکاران (۲۰۱۵) معرفی شده‌اند و از طریق لینک <https://rda.ucar.edu/datasets/d651027> قابل دریافت می‌باشند. داده‌های مورد استفاده شامل اطلاعات آب‌وهوایی از سال ۴۰۲ میلادی به مدت ۹۸ سال متوالی هستند. بررسی ویژگی‌های توصیفی این مجموعه نشان می‌دهد که تقویم مورد استفاده در آن از نوع تقویم بدون سال کبیسه است. همچنین متغیری به نام `time_var` در داخل مجموعه، حاوی تعداد روزهای گذشته از تاریخ مبدأ است که در استخراج روزهای مورد نظر کمک کننده است. با توجه به این توضیحات و با در نظر گرفتن این موضوع که داده‌های استفاده شده در مقاله کید و کترفز مربوط به میانگین دمای روز اول جولای از سال ۴۰۲ هستند، در این پژوهش ما داده‌های مربوط به اول سپتامبر را انتخاب کرده‌ایم تا رفتار دمایی اواخر تابستان نیز مورد بررسی قرار گیرد.

داده‌های مورد استفاده شامل میانگین دمای روزانه سطح زمین (بر حسب کلونین) در تاریخ اول سپتامبر، طی ۵۰ سال متوالی از سال ۴۱۱ میلادی به بعد هستند. این داده‌ها بر روی شبکه‌ای با تفکیک مکانی ۱/۲۵ درجه در طول و ۰/۹ درجه در عرض جغرافیایی تعریف شده‌اند که شامل $۹۰ \times ۵۵ = ۴۹۵۰$ نقطه شبکه‌ای می‌شود. ناحیه انتخاب شده بخش نسبتاً وسیعی از قاره آمریکا را در بر می‌گیرد و شامل اقیانوس، خشکی، جزایر و رشته‌کوه‌هاست. این تنوع جغرافیایی منجر به ساختاری پیچیده و نامانا در وابستگی‌های فضایی داده‌ها می‌شود. برای آماده‌سازی داده‌ها، دمای ثبت شده در تاریخ اول سپتامبر هر سال استخراج شد. در هر نقطه شبکه‌ای، میانگین و واریانس دما در طول ۵۰ سال محاسبه و داده‌ها به صورت ناهنجاری دمایی با میانگین صفر و واریانس یک استاندارد شدند. با تحلیل ساختار زمانی داده‌ها، شواهدی از همبستگی قوی زمانی مشاهده نشد. بنابراین، فرض استقلال نمونه‌ها در طول ۵۰ سال به‌طور تقریبی برقرار است. برای نمونه‌گیری از توزیع پسین ابرپارامترهای θ ، از الگوریتم تطبیقی MH مطابق با رویکرد ارائه شده در کید و کترفز (۲۰۲۲) استفاده شد. زنجیره مارکوف مونت کارلو با ۳۰۰۰۰ تکرار اجرا گردید. پس از حذف ۱۵۰۰۰ نمونه ابتدایی به‌عنوان دوره سوختن، نمونه‌های باقی‌مانده با فاکتور ۵۰ نازک‌سازی شدند و در نهایت ۳۰۰ نمونه مستقل حاصل گردید. نمودارهای اثر نوسان‌های پایدار و بدون روند را نشان دادند که تأییدی بر کیفیت نمونه‌گیری و همگرایی موفقیت‌آمیز الگوریتم MCMC در برآورد ابرپارامترهای مدل فضایی می‌باشند.

¹Community Earth System Model

برای برآورد درایه‌های چولسکی ماتریس دقت، دو روش کید و کتفز و روش گروه‌بندی به کار گرفته شد. روش کید و کتفز با وجود دقت بالا، از نظر محاسباتی نسبتاً سنگین است و زمان اجرای قابل توجهی دارد. در روش پیشنهادی، نقاط شبکه‌ای به گروه‌هایی با ساختار آماری مشابه تقسیم می‌شوند. سپس در هر گروه، تخمین‌های محلی برای ماتریس دقت محاسبه شده و در نهایت ماتریس‌های چولسکی به دست می‌آیند که برای اجرای آن زمان کمتری نسبت به روش کید و کتفز صرف می‌شود. در این مطالعه برای سنجش دقت بازسازی ماتریس‌های کوواریانس، از معیار خطای نسبی نرم فروبنیوس استفاده شده است. این معیار نه تنها کل ساختار ماتریس را در نظر می‌گیرد، بلکه با نرمال‌سازی نسبت به بزرگی ماتریس مرجع، امکان مقایسه منصفانه بین روش‌های مختلف را فراهم می‌سازد. در مقابل، معیارهایی مانند میانگین مربعات خطا (MSE) بیشتر برای بردارهای پیش‌بینی‌شده مناسب هستند و در تحلیل ساختارهای ماتریسی ممکن است اطلاعات فضایی را از دست بدهند.

از نظر دقت نیز، روش گروه‌بندی توانست نتایجی قابل مقایسه با روش مرجع ارائه دهد. ماتریس کوواریانس داده‌های استخراج شده محاسبه گردید و سپس با ماتریس کوواریانس برآورد شده به روش کید و کتفز و روش گروه‌بندی مقایسه گردید. از آنجایی که در مقاله کید و کتفز، مقایسه روش با انواع روش‌های برآورد ماتریس‌های کوواریانس انجام شده و برتری روش اثبات گردیده است، در ادامه مطالعات ایشان، به مقایسه روش گروه‌بندی با روش کید و کتفز بسنده شد. بر اساس جدول ۵، روش گروه‌بندی علاوه بر زمان محاسباتی کمتر، خطای بازسازی

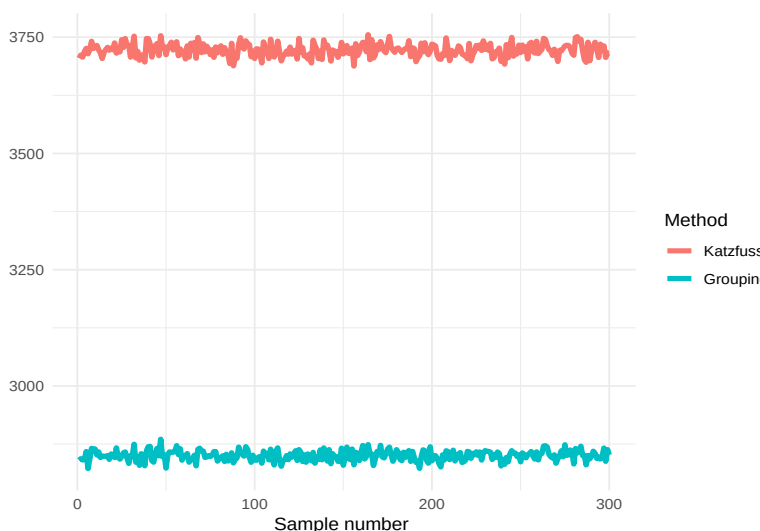
جدول ۵. زمان اجرا و خطای نسبی بازسازی ماتریس کوواریانس میانگین در دو روش مدل‌سازی فضایی

روش	زمان اجرا (دقیقه)	خطای نسبی نرم فروبنیوس
گروه‌بندی	۴۳	۱/۰۰۵
کید و کتفز	۶۷	۱/۳۳۵

پایین‌تری نسبت به روش کید و کتفز دارد. مقایسه خطای نسبی نرم فروبنیوس بین ماتریس‌های کوواریانس برآورد شده و ماتریس مرجع نشان می‌دهد که روش گروه‌بندی با خطای نسبی ۱/۰۰۵ عملکرد دقیق‌تری نسبت به روش کید و کتفز با خطای نسبی ۱/۳۳۵ داشته است. برای بررسی بیشتر، با توجه به ماتریس‌های دقت به دست آمده از هر دو روش که لیستی شامل ۳۰۰ ماتریس است، نرم فروبنیوس برای هر ماتریس دقت در هر دو روش به دست آمد و در شکل ۶ رسم شد که با توجه به آن، نرم به دست آمده در روش گروه‌بندی کمتر از روش کید و کتفز است که گویای کوچک بودن بازه‌های عدم قطعیت در برآورد درایه‌های ماتریس چولسکی نسبت به روش کید و کتفز است. ماتریس‌هایی با نرم بالاتر ممکن است منجر به تخمین‌های با واریانس بیشتر شوند، در حالی که ماتریس‌های با نرم پایین‌تر معمولاً ساختار هموارتر و محافظه‌کارانه‌تری را مدل می‌کنند. این موضوع در انتخاب روش مناسب برای کاربردهای خاص فضایی باید مدنظر قرار گیرد. نتایج حاصل از تحلیل آماری نرم فروبنیوس ماتریس‌های چولسکی نشان می‌دهد که روش گروه‌بندی نسبت به روش کید و کتفز از واریانس کمتری برخوردار است. به طور مشخص، انحراف معیار در روش گروه‌بندی برابر با ۱۲/۱۵ و در روش کید و کتفز برابر با ۱۵/۴۶ گزارش شده است. این اختلاف بیانگر همگنی عددی و پایداری بیشتر نمونه‌های تولیدشده در روش گروه‌بندی است. واریانس کمتر در این

۲۲۲ برآورد ناپارامتری کوواریانس فضایی نامانا برای مه داده‌ها

روش نشان‌دهنده نوسان محدودتر در ساختار ماتریس‌های کوواریانس بازسازی شده بوده و می‌تواند به تخمین‌های قابل اعتمادتر و کاهش حساسیت مدل نسبت به تغییرات تصادفی در داده‌ها منجر شود؛ موضوعی که برای کاربردهای نیازمند پایداری عددی بالا اهمیت ویژه‌ای دارد. لازم به ذکر است؛ کدهای استفاده شده در این مطالعه از سوی نویسندگان، در صورت درخواست خوانندگان علاقمند، در دسترس آنها قرار خواهد گرفت.



شکل ۶. مقایسه نوسان عددی ماتریس‌های چولسکی به دست آمده از روش گروه‌بندی و روش کید و کتفز

۶ بحث و نتیجه‌گیری

در این مقاله، یک چارچوب بیزی مقیاس‌پذیر برای برآورد ماتریس کوواریانس در داده‌های فضایی بزرگ مقیاس ارائه شد. این روش با بهره‌گیری از فرض استقلال شرطی در داده‌های مرتب‌شده، ساختار تنک ماتریس چولسکی دقت را به صورت داده‌محور تعیین کرده و هزینه محاسباتی را از مرتبه درجه دوم به مرتبه‌ای نزدیک به خطی کاهش می‌دهد. همچنین یک الگوریتم گروه‌بندی معرفی شد که با تجمیع موقعیت‌ها و حذف وابستگی‌های ضعیف، ساختار کوواریانس را در قالب بلوک‌های منسجم و کم‌هزینه بازنمایی می‌کند. نتایج شبیه‌سازی و تحلیل داده‌های واقعی نشان دادند که رویکرد پیشنهادی، ضمن حفظ دقت و برازش مناسب، عدم قطعیت پسین را به طور قابل قبولی کنترل کرده و در سناریوهای مختلف از نظر پایداری و مقیاس‌پذیری عملکرد مطلوبی دارد. انتخاب داده‌محور تنظیمات مدل، از جمله تعداد همسایه‌ها، نقش مهمی در افزایش انعطاف‌پذیری و کارایی روش ایفا می‌کند. به منظور افزایش تفسیرپذیری، از نمونه‌های پسین ماتریس کوواریانس، پروفایل‌های فاصله-کوواریانس محلی استخراج و مدل ماترن محلی در چند

نقطه نماینده برازش شد. نتایج نشان داد که روش گروه‌بندی، با وجود ساختار تنک و کاهش ابعاد، الگوهای وابستگی محلی را به‌طور پایدار بازتولید می‌کند؛ امری که با مشاهدات ارائه‌شده در شکل ۵ نیز سازگار است. از این‌رو، تفسیر پارامترهای ماترن باید به‌صورت کیفی و مبتنی بر روند کلی توزیع پسین انجام شود، نه به‌عنوان برآوردهای نقطه‌ای. در ادامه این پژوهش، بررسی روش‌های استنتاج سریع‌تر مانند VI و INLA، تحلیل دقیق‌تر حساسیت ساختار گروه‌بندی، و مقایسه با سایر روش‌های تقریبی متداول می‌تواند به توسعه و تثبیت جایگاه رویکرد پیشنهادی در ادبیات مدل‌سازی فضایی کمک کند.

تقدیر و تشکر

از سردبیر، داوران و ویراستار محترم مجله برای بازبینی دقیق، نظرات سازنده و پیشنهادهای ارزشمندشان که نقش مهمی در بهبود کیفیت این مقاله ایفا کرده‌اند، صمیمانه سپاسگزاریم.

مراجع

- محمدزاده، م.، (۱۳۹۸)، آمار فضایی و کاربردهای آن، چاپ سوم، مرکز نشر آثار علمی دانشگاه تربیت مدرس، تهران،
- محمدیان مصمم، ع. و محمدی، س. (۱۳۹۷)، برآورد تابع کواریانس فضایی با استفاده از روش درستمایی مرکب بلوکی تفاضلی، مجله علوم آماری، ۱۲(۲)، ۵۱۳-۵۲۵.
- محمدیان مصمم، ع.، عباسی، ا. و متیو، خ. (۱۴۰۱)، کاربرست رهیافت بیزی در مدل‌سازی فضایی-زمانی داده‌های جرم، مجله علوم آماری، ۱۶(۲)، ۴۳۵-۴۴۸.
- Banerjee, S., Carlin, B. P., and Gelfand, A. E., (2003), Hierarchical Modeling and Analysis for Spatial Data, Chapman and Hall/CRC.
- Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H., (2008), Gaussian Predictive Process Models for Large Spatial Data Sets, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(4), 825-848.
- Banerjee, S., Guhaniyogi, R., and Finley, A.O. (2023), Localized Gaussian Processes for Large Spatial Data, *Biometrika*, **110**(2), 327-345.
- Cressie, N. (2015), *Statistics for Spatial Data*. John Wiley & Sons.

Cressie, N., and Johannesson, G, (2008), Fixed Rank Kriging for Very Large Spatial Data Sets, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 209–226.

Cressie, N., and Wikle, C. K, (2011), Statistics for Spatio-Temporal Data, John Wiley & Sons.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E, (2016), Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets, *Journal of the American Statistical Association*, **111**(514), 800–812.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S, (2006, August), Subspace Sampling and Relative-Error Matrix Approximation: Column-Based Methods, *In International Workshop on Approximation Algorithms for Combinatorial Optimization*, 316–326, Berlin, Heidelberg: Springer Berlin Heidelberg.

Du, J., Zhang, H., and Mandrekar, V, (2009), Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators, *Journal of the American Statistical Association*, **37**, 3330–3361.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J, (2014), Estimation and Prediction in Spatial Models with Block Composite Likelihoods, *Journal of Computational and Graphical Statistics*, **23**(2), 295–315.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E, (2009), Improving the Performance of Predictive Process Modeling for Large Datasets, *Computational Statistics & Data Analysis*, **53**(8), 2873–2884.

Fuentes, M, (2007), Approximate Likelihood for Large Irregularly Spaced Spatial Data, *Journal of the American Statistical Association*, **102**(477), 321–331.

Furrer, R., Genton, M. G., and Nychka, D, (2006), Covariance Tapering for Interpolation of Large Spatial Datasets, *Journal of Computational and Graphical Statistics*, **15**(3), 502–523.

- Gelman, A., and Rubin, D. B, (1992), Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, **7**(4), 457–472.
- Guinness, J, (2018), Permutation and Grouping Methods for Sharpening Gaussian Process Approximations, *Technometrics*, **60**(4), 415–429.
- Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L, (2006), Covariance Matrix Selection and Estimation via Penalised Normal Likelihood, *Biometrika*, **93**(1), 85–98.
- Kang, E. L., and Cressie, N, (2011), Bayesian Inference for the Spatial Random Effects Model, *Journal of the American Statistical Association*, **106**(495), 972–983.
- Katzfuss, M., and Cressie, N, (2011), Spatio-Temporal Smoothing and EM Estimation for Massive Remote-Sensing Data Sets, *Journal of Time Series Analysis*, **32**(4), 430–446.
- Katzfuss, M., and Guinness, J, (2021), A General Framework for Vecchia Approximations of Gaussian Processes, *Statistical Science*, **36**(1), 124–141.
- Kaufman, C. G., Schervish, M. J., and Nychka, D. W, (2008), Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets, *Journal of the American Statistical Association*, **103**(484), 1545–1555.
- Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., ... and Vertenstein, M. (2015), The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, *Bulletin of the American Meteorological Society*, **96**(8), 1333–1349.
- Kidd, B., and Katzfuss, M, (2022), Bayesian Nonstationary and Nonparametric Covariance Estimation for Large Spatial Data (with discussion), *Bayesian Analysis*, **17**(1), 291–351.

- Mosammam, A. M, (2025), On the Frequency Domain Composite Likelihood Methods for Estimating Space-Time Covariance Functions for Large Datasets, *Communications in Statistics—Simulation and Computation*,**54**(4), 1146–1160.
- Rue, H., and Held, L, (2005), Gaussian Markov Random Fields: Theory and Applications, Chapman and Hall/CRC.
- Sang, H., and Huang, J. Z. (2012), A Full Scale Approximation of Covariance Functions for Large Spatial Data Sets, *Journal of the Royal Statistical Society: Series B*, **74**(1), 111–132.
- Schabenberger, O., and Gotway, C. A, (2004), Statistical Methods for Spatial Data Analysis, Chapman and Hall/CRC Texts in Statistical Science, Chapman and Hall/CRC.
- Schafer, F., Sullivan, T. J., and Owhadi, H, (2021b), Compression, Inversion, and Approximate PCA of Dense Kernel Matrices at Near-Linear Computational Complexity, *Multiscale Modeling & Simulation*, **19**(2), 688–730.
- Scheidegger, A, (2012), Implementation of A Generic Adaptive Monte Carlo Markov Chain sampler. R package `adaptMCMC`, version 1.4. Available at <https://cran.r-project.org/package=adaptMCMC>.
- Shaby, B., and Ruppert, D, (2012), Tapered covariance: Bayesian Estimation and Asymptotics, *Journal of Computational and Graphical Statistics*, **21**(2), 433–452.
- Stein, M. L., Chi, Z., and Welty, L. J, (2004), Approximating Likelihoods for Large Spatial Data Sets, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(2), 275–296.
- Varin, C., Reid, N., and Firth, D, (2011), An Overview of Composite Likelihood Methods, *Statistica Sinica*, **21**, 5–42.
- Vecchia, A. V, (1988), Estimation and Model Identification for Continuous Spatial Processes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **50**(2), 297–312.