



یک مدل فضایی چوله رتبه کاسته برای داده‌های با اندازه بزرگ

حسین بوجاری* و مجید جعفری خالدي†

دانشگاه تربیت مدرس، hossein.boojari@modares.ac.ir،
دانشگاه تربیت مدرس، jafari-m@modares.ac.ir

چکیده

در بسیاری از داده‌های آمار فضایی، چولگی مشاهده می‌شود و برای تحلیل چنین داده‌هایی، مدل‌های چوله گاوسی متعددی ارائه شده است. اما امروزه با توجه به پیشرفتهای تکنولوژی، اغلب، داده‌های بسیار حجیم فضایی چوله تولید می‌شوند. در نتیجه، تحت این شرایط، امکان استفاده مستقیم از مدل‌های چوله گاوسی موجود در آمار فضایی بدلیل پیچیدگی و مشکلات محاسباتی وجود ندارد. از این رو در این مقاله یک مدل چوله گاوسی رتبه کاسته معرفی می‌شوند که با استفاده از ترکیب خطی تعداد مشخصی از توابع پایه و متغیر تصادفی پنهان چوله نرمال ساخته می‌شود بگونه‌ای که کاملاً با داده‌های فضایی چوله با حجم زیاد سازگار است.

واژه‌های کلیدی: چوله گاوسی، رتبه کاسته، متغیر پنهان، توابع پایه.

رده‌بندی موضوعی ریاضی (2010): 62H11, 62H12.

۱ مقدمه

مدل‌های فضایی گاوسی با توجه به خواصی که توزیع نرمال دارد، بعنوان مدل پایه در تحلیل داده‌های فضایی بکار می‌روند. اما در عمل با موارد متعددی روبرو می‌شویم که در توزیع داده‌ها شواهدی از تعارض با فرض گاوسی مانند چولگی بروز می‌کند و بنابر این استفاده از مدل گاوسی را به چالش می‌کشاند. از این رو در آمار فضایی معرفی الگوهای ناگوسی مورد توجه آماردانان قرار گرفته است. با توجه به ضعف‌های استنباطی که تبدیلات باکس-کاکس به دنبال دارد، رویکردهای جدیدی برای تحلیل داده‌های فضایی چوله براساس توزیع‌های چوله نرمال ارائه و توسعه داده شده است. کیم و مارلیک (۲۰۰۴) میدان تصادفی چوله گاوسی را براساس توزیع چوله نرمال چند متغیره معرفی کردند. آلارد و ناوا (۲۰۰۷) میدان تصادفی چوله‌ای را بر اساس توزیع چوله نرمال بسته ارائه داده‌اند. زارعی‌فرد و خالدي (۲۰۱۳) بر مبنای توزیع چوله نرمال چند متغیره یکپارچه مدلی را معرفی کردند.

اما روش‌ها و مدل‌های فضایی که تاکنون در این زمینه مطرح شده‌اند و به بعضی از آنها اشاره شد، فقط قادر به مدل‌بندی داده‌های چوله با حجم کم هستند و به دلیل پیچیدگی‌های ساختاری این مدل‌ها امکان بکارگیری آنها برای داده‌های حجیم وجود ندارد. یک دلیل اصلی برای بروز چنین مشکلی ماتریس کوواریانس متغیر پاسخ است. چون ساختار همبستگی فضایی به صورت ماتریس وابستگی $\Sigma_{n \times n}$ برای n مشاهده فضایی در تحلیل این گونه مدل‌ها به کار می‌رود و بخصوص، معکوس این ماتریس یعنی Σ^{-1} مورد نیاز است در نتیجه برای داده‌های فضایی با حجم زیاد، انجام محاسبات برای یافتن Σ^{-1} حتی با وجود کامپیوترهای بسیار قوی نیز با مشکل روبرو است. می‌توان نشان داد که تعداد محاسبات لازم برای محاسبه Σ^{-1} تابعی از n^2 است. به عنوان

*سخنران
†مسئول مکاتبات

مثال اگر فرض کنیم تعداد مشاهدات ۱۰۰۰۰ باشد، آنگاه ماتریس کوواریانس دارای ۵۰۰۰۵۰۰۰ مقدار مختلف خواهد بود که حتی در ذخیره‌سازی این ماتریس نیز مشکل وجود دارد و معکوس‌گیری از این ماتریس بخصوص برای روش‌هایی مانند تحلیل بیزی که با شبیه‌سازی مونت‌کارلو به صورت بازگشتی عمل می‌کنند، غیر ممکن خواهد شد. به خصوص که در اکثر ساختارهای همبستگی یک یا چند پارامتر مجهول نیز وجود دارند و این پارامترها در درایه‌های ماتریس Σ نیز وجود دارند. از طرفی امروزه با پیشرفت تکنولوژی و توسعه سیستم‌های اطلاعاتی و وجود ماهواره‌ها بخصوص در زمینه جغرافیا و محیط زیست، در عمل با داده‌های حجیم علمی روبرو هستیم که ممکن است بعنوان مثال داده‌هایی را از کل کره زمین با حجم نمونه بسیار بزرگ در دسترس قرار دهند. نکته قابل ذکر این است که این مشکل برای مدل‌های گاوسی و روش‌های متداول آمار فضایی مانند روش کریگینگ نیز وجود دارد و این روش‌ها نیز در برخورد با داده‌های حجیم با چالش مواجه می‌شوند. بعنوان مثال، برآورد ضرایب رگرسیون فضایی برای پارامتر q بعدی β ، بصورت $\hat{\beta} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}$ است، که در آن \mathbf{X} ماتریس مشاهده‌های متغیرهای تبیینی و \mathbf{Y} بردار مشاهده‌های متغیر پاسخ است. بنابراین مشهود است که اگر حجم نمونه بزرگ باشد، محاسبات برآورد پارامترهای مدل و نیز پیشگویی فضایی در مدل‌های موجود، عملاً امکان‌پذیر نخواهد بود. برای رفع این مشکل، رویکردهایی جهت حل مشکل محاسبه Σ^{-1} مورد توجه قرار گرفته است. به عنوان مثال، رو و تیلمند (۲۰۰۲) در حالتی که ماتریس Σ تنگ باشد تقریبی را برای Σ^{-1} ارائه دادند. رو و هلد (۲۰۰۵) در مدل‌های گاوسی بر اساس روش‌های تبدیل فوریه، محاسبات را انجام دادند. استفاده از این تکنیک‌ها در مدل‌های چوله گاوسی شاید به نظر راهگشا برسد، اما همیشه شرایط بکارگیری این روش‌ها مهیا نیست و یا اینکه وجود تقریب در این گونه محاسبات ممکن است باعث بروز خطاهایی همچون اریبی برآوردها و پیشگویی‌ها شود. از این رو بعنوان یک رویکرد متفاوت و با نگاه مدل محور، به نظر می‌رسد که به کمک یک مدل رتبه کاسته می‌توان کل فرایند فضایی را تقریب زد به گونه‌ای که هم چولگی را در نظر بگیرد و هم با داده‌های حجیم سازگار باشد. در این روش میدان خطای مدل، بصورت یک میدان تصادفی که ترکیبی خطی از توابع پایه و متغیرهای پنهان با توزیع چوله نرمال است در نظر گرفته می‌شود. با این دیدگاه، در این مقاله مدل رتبه کاسته چوله‌ای معرفی می‌شود که با داده‌های بزرگ سازگاری کامل دارد. واضح است که مدل ارائه شده در حالتی که توزیع متغیر پنهان نرمال باشد برای داده‌های نرمال با حجم زیاد نیز کارایی خود را حفظ می‌نماید.

۲ مدل فضایی چوله سازگار با حجم نمونه بزرگ

فرض کنید $\{Y(\mathbf{s}) : \mathbf{s} \in \mathbf{D} \subseteq R^d\}$ یک میدان تصادفی فضایی با مقادیر حقیقی باشد که معمولاً d برابر ۲ یا ۳ است و $\mathbf{Y} = (y(\mathbf{s}_1), y(\mathbf{s}_2), \dots, y(\mathbf{s}_n))'$ تحقق‌های این میدان تصادفی در n موقعیت فضایی $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$ باشند. در این میدان، مدل رگرسیون فضایی به صورت زیر مفروض است:

$$y(\mathbf{s}) = x(\mathbf{s})'\beta + w(\mathbf{s}) + \varepsilon(\mathbf{s}). \quad (1.2)$$

در این مدل، فرایند فضایی پنهان $w(\mathbf{s})$ بیانگر ساختار فضایی باقی مانده‌ها و فرایند مستقل $\varepsilon(\mathbf{s})$ اغلب به عنوان اثر قطعه‌ای مطرح می‌شود و یک نوفه‌ی ناهمبسته تصادفی است که دارای توزیع $N(0, \sigma^2)$ است. همچنین $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), x_2(\mathbf{s}), \dots, x_q(\mathbf{s}))'$ بردارمشاهدات متغیرهای تبیینی است که به موقعیت وابسته‌اند و از طریق پارامتر q بعدی β با متغیر وابسته $y(\mathbf{s})$ وابستگی دارد. بر اساس نمونه‌ی n تایی، مدل (۱،۲) بصورت:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{W} + \epsilon, \quad (2.2)$$

خواهد بود که در آن \mathbf{W} و ϵ به ترتیب بردارهای شامل $w(s_i)$ و $\epsilon(s_i)$ هستند. همچنین ماتریس $n \times q$ بعدی \mathbf{X} ، شامل مشاهدات متغیرهای تبیینی $x_j(s_i)$ است. حال به منظور ایجاد یک مدل فضایی چوله رتبه کاسته، بردار تصادفی $\gamma = (\gamma_1, \dots, \gamma_m)'$ به عنوان یک فرایند پنهان در نظر گرفته شده است و در این صورت رابطه

$$w(\mathbf{s}) = \sum_{j=1}^m B_j(\mathbf{s})\gamma_j = \mathbf{B}(\mathbf{s})'\gamma, \quad (3.2)$$

برقرار است که در آن $\mathbf{B}(\mathbf{s}) = (B_1(\mathbf{s}), \dots, B_m(\mathbf{s}))'$ و m مقداری مشخص و بسیار کوچکتر از n است. با در نظر گرفتن توزیع چوله نرمال چند متغیره معرفی شده توسط آرلانو-واله و همکاران (۲۰۰۷)، برای γ ، تابع چگالی γ به صورت

$$\tau^m \phi_m(\gamma; \circ, \mathbf{K} + \Delta\Delta'(\cdot\Phi_m)\Delta'(\mathbf{K} + \Delta\Delta')^{-1}\gamma; \circ, (I_m + \Delta'\mathbf{K}\Delta)^{-1}), \quad (4.2)$$

به دست می‌آید، که در آن \mathbf{K} یک ماتریس پراکندگی معین مثبت $m \times m$ بعدی، $\phi_m(\cdot; \mu, \mathbf{A})$ و $\Phi_m(\cdot; \mu, \mathbf{A})$ به ترتیب توابع چگالی و توزیع نرمال m متغیره با میانگین μ و ماتریس کوواریانس \mathbf{A} هستند. همچنین $\Delta = \alpha I_m$ یک ماتریس قطری است که در آن α پارامتر چولگی است. در این صورت مدل (۱،۲) یک مدل چوله گاوسی رتبه کاسته است. براین اساس، می‌توان نشان داد که تابع میانگین و کوواریانس $Y(\cdot)$ به ترتیب به صورت

$$E(Y(\mathbf{s})) = \mathbf{x}'(\mathbf{s})\beta + \sqrt{\frac{\tau}{\pi}} \alpha \sum_{j=1}^m B_j(\mathbf{s}) \quad (5.2)$$

$$c(\mathbf{s}, \mathbf{s}') = cov(Y(\mathbf{s}), Y(\mathbf{s}')) = \mathbf{B}'(\mathbf{s}) \left[\mathbf{K} + \left(1 - \frac{\tau}{\pi}\right) \alpha^2 \mathbf{I}_m \right] \mathbf{B}(\mathbf{s}') + \tau^2 \mathbf{I}_{\{\mathbf{s}=\mathbf{s}'\}}. \quad (6.2)$$

بدست می‌آیند. در نتیجه روابط

$$E(\mathbf{Y}) = \mathbf{X}\beta + \alpha \sqrt{\frac{\tau}{\pi}} \mathbf{B}\mathbf{1}_n, \quad (7.2)$$

$$\Sigma = \mathbf{B} \left[\mathbf{K} + \left(1 - \frac{\tau}{\pi}\right) \alpha^2 \mathbf{I}_m \right] \mathbf{B}' + \tau^2 \mathbf{I}_n, \quad (8.2)$$

برقرارند که در آن‌ها، $\mathbf{1}_n = (1, 1, \dots, 1)'$ و $\mathbf{B} = (\mathbf{B}'(\mathbf{s}_1), \dots, \mathbf{B}'(\mathbf{s}_n))$ نتیجه جالبی که به کمک رابطه شرمن-موریسون-وودبری بدست می‌آید آن است که

$$\Sigma^{-1} = \tau^{-2} \mathbf{I}_n - \tau^{-2} \mathbf{B} \left[(\mathbf{K} + \left(1 - \frac{\tau}{\pi}\right) \alpha^2 \mathbf{I}_m)^{-1} + \tau^{-2} \mathbf{B}'\mathbf{B} \right]^{-1} \mathbf{B}'. \quad (9.2)$$

بنابر این برای محاسبه ماتریس $n \times n$ بعدی Σ^{-1} فقط نیاز به معکوس کردن ماتریس $m \times m$ بعدی زیر است:

$$\left[(\mathbf{K} + \left(1 - \frac{\tau}{\pi}\right) \alpha^2 \mathbf{I}_m)^{-1} + \tau^{-2} \mathbf{B}'\mathbf{B} \right]$$

بنابراین مدل ارائه شده علاوه بر در نظر گرفتن چولگی، سازگاری کامل با داده‌های حجیم دارد. توجه شود که در این مدل مقدار m در مقایسه با مقدار n بسیار کوچکتر در نظر گرفته می‌شود. با تعیین تابع پایه در این مدل می‌توان روابط فوق را قابل محاسبه نمود.

اگر برای نقاط ثابت گره s_1^*, \dots, s_m^* فرض کنیم $B_j(s) = b(s - s_j^*)$ ، آنگاه فقط لازم است تابع پایه b تعیین شود. انتخاب توابع هسته به عنوان تابع پایه توسط بسیاری از محققان صورت گرفته است و در این میان تابع هسته‌ی بیزر که به صورت

$$b(\mathbf{s}, \phi) = \left[1 - \left(\frac{\|\mathbf{s}\|}{r} \right)^2 \right]^{l_p + \phi(u_p - l_p)} I_{(\|\mathbf{s}\| < r)} \quad , \phi \in (0, 1) \quad (10.2)$$

است، با توجه به خصوصیات مناسب فضایی که دارد می‌تواند انتخاب معقولی در این خصوص باشد. l_p و u_p به عنوان حدود بالا و پایین توان تابع هستند که از قبل تعیین می‌شوند و تابع تا مرتبه‌ی $2(l_p + \phi(u_p - l_p))$ مشتق پذیر است. پارامتر ϕ میزان همواری میدان تصادفی را بیان می‌کند.

انتخاب نقاط گره نیز یک مساله‌ی مهم است که می‌توان از طرح شبکه‌ای منظم برای مشخص کردن موقعیت گره‌ها استفاده نمود. با استفاده از این طرح می‌توان یک پوشش منظم از ناحیه‌ی تحت مطالعه را بدست آورد.

حال، با توجه به آنکه بیان تصادفی بردار γ به صورت $\gamma \stackrel{d}{=} \alpha |\mathbf{Z}_0| + \mathbf{Z}_1$ است که در آن $\mathbf{Z}_0 \sim N_m(0, \mathbf{I}_m)$ و $\mathbf{Z}_1 \sim N_m(0, \mathbf{K})$ بردارهای تصادفی نرمال m متغیره مستقل هستند، از این رو انجام تحلیل بیزی بر روی پارامترهای مدل معرفی شده و پیشگویی فضایی با در نظر گرفتن توزیع‌های پیشین برای پارامترهای مدل به سادگی از طریق بیان تصادفی سلسله مراتبی داده افزایشی شده به صورت زیر امکان پذیر است (بوجاری و خالدی، ۲۰۱۶).

$$\begin{aligned} \mathbf{Y} | \gamma, \beta, \tau^2 &\sim N_n(\mathbf{X}\beta + \mathbf{B}\gamma, \tau^2 \mathbf{I}_n), \\ \gamma | \mathbf{K}, \alpha, |\mathbf{Z}_0| = \mathbf{t} &\sim N_m(\alpha \mathbf{t}, \mathbf{K}), \\ |\mathbf{Z}_0| &\sim HN_m(0, \mathbf{I}_m). \end{aligned}$$

۳ بررسی شبیه‌سازی

به منظور بررسی شناسایی پذیری پارامتر چولگی در مدل معرفی شده، با در نظر گرفتن مقادیر $\mu(\mathbf{s}) = 0$ ، $\tau = 1$ ، $\mathbf{K} = \mathbf{I}_m$ و مقادیر $lp = 2$ و $\theta = 0$ و $r = \sqrt{0.5}$ برای تابع هسته بیزر و انتخاب تصادفی $n = 200$ نقطه در مربع $[0, 1] \times [0, 1]$ ، هفت مجموعه داده بر اساس مدل چوله گاوسی برای مقادیر مختلف $\alpha \in \{0, 0.1, 0.5, 1, 2, 3, 5\}$ تولید شدند. انجام استنباط بیزی بر اساس الگوریتم MCMC با دروه داغیدن ۲۵۰۰۰ و نمونه ۵۰۰۰ تایی بصورت ۱۰ درمیان و با در نظر گرفتن توزیع‌های پیشین تخت

$$\beta \sim N(0, 100 \mathbf{I}_q), \tau^2 \sim IG(0.01, 0.01), \alpha \sim N(0, 100), \mathbf{K} \sim IW(\mathbf{I}_m, m),$$

انجام گرفت. نتایج بر اساس مقادیر مختلف m که بصورت منظم انتخاب شده‌اند، در جدول ۱ آورده شده است که نشان از شناسایی‌پذیری و استنباط معتبر بر روی پارامتر چولگی است.

۴ مثال کاربردی

کرسی و کاتفوس (۲۰۱۱)، داده‌های مربوط به مقادیر co_2 را که از ۲۶۶۳۳ نقطه کره زمین جمع آوری شده بود را با یک مدل گاوسی رتبه کاسته (GM) مورد بررسی قرار دادند. اما، از آنجا که توزیع این داده‌ها چوله است (نمودار ۱)، در نظر گرفتن مدل چوله گاوسی رتبه کاسته (SGM) می‌تواند توجه بسیار مناسبی برای بهبود تحلیل این داده‌ها باشد. با حفظ نقاط گره و پیشین‌های بکار رفته توسط آنها، مدل SGM همراه مدل GM به داده‌ها برازش شد که نتایج آن در جدول ۲ آمده است. معیار فاکتور بیزی مدل SGM در مقابل مدل GM برابر ۷/۴ بدست آمد که نشان از برتری مدل چوله گاوسی دارد.

m=۳۶	m=۲۵	m=۹	مقدار واقعی
۰٫۱۵۱(۰٫۲۴۵)	۰٫۰۰۴(۰٫۲۱۳)	۰٫۰۳۱(۰٫۵۲۶)	۰
۰٫۲۴۴(۰٫۲۴۳)	۰٫۱۱۴(۰٫۱۹۹)	۰٫۱۴۰(۰٫۵۳۶)	۰٫۱
۰٫۶۲۱(۰٫۲۶۶)	۰٫۵۳۵(۰٫۱۸۳)	۰٫۵۶۰(۰٫۵۹۱)	۰٫۵
۱٫۱۱۴(۰٫۲۹۵)	۱٫۰۳۹(۰٫۲۳۰)	۱٫۰۵۷(۰٫۶۶۵)	۱
۲٫۰۹۴(۰٫۴۳۶)	۲٫۰۶۱(۰٫۳۳۴)	۲٫۰۴۶(۰٫۹۳۹)	۲
۳٫۰۱۷(۰٫۳۸۷)	۳٫۱۶۲(۰٫۴۹۸)	۲٫۹۷۲(۱٫۰۱۸)	۳
۴٫۲۴۳(۰٫۴۳۶)	۴٫۷۴۰(۰٫۵۶۰)	۴٫۸۲۴(۱٫۶۴۶)	۵

جدول ۱: میانگین پسین (انحراف معیار) برای پارامتر چولگی تحت مقادیر مختلف m .

SGM				GM				پارامتر
۹۷٫۵%	۲٫۵%	انحراف معیار	میانگین	۹۷٫۵%	۲٫۵%	انحراف معیار	میانگین	
۳۷۶٫۵	۳۷۵٫۴	۰٫۲۹۵	۳۷۶	۳۷۶٫۴	۳۷۴٫۶	۰٫۳۹۵	۳۷۵٫۸	β_0
۰٫۵۲۴	-۰٫۷۲۰	۰٫۳۲۶	-۰٫۰۰۸	۰٫۶۳۶	-۰٫۶۱۱	۰٫۲۹۸	۰٫۰۴۳	β_1
-۰٫۰۳۵	-۰٫۱۰۵	۰٫۰۱۸	-۰٫۰۰۷۱	-۰٫۰۳۶	-۰٫۱۰۶	۰٫۰۱۸	-۰٫۰۷۱	β_2
۰٫۷۵۸	۰٫۷۱۸	۰٫۱۰۷	۰٫۷۳۸	۰٫۷۳۸	۰٫۷۱۷	۰٫۰۱۱	۰٫۷۳۸	τ
-۰٫۰۰۳	-۰٫۴	۰٫۱۰۷	-۰٫۱۲۶	-	-	-	-	α

جدول ۲: میانگین پسین، انحراف معیار و فاصله اطمینان ۹۵٪ پارامترهای دو مدل بر اساس داده‌های CO_2 .

۵ نتیجه گیری

بدلیل پیچیدگی مدل‌های چوله گاوسی موجود، امکان استفاده از آنها برای داده‌های حجیم فضایی وجود ندارد. بنابراین، با یک رویکرد متفاوت، مدل چوله گاوسی رتبه کاسته‌ای ارائه گردید که قابلیت تحلیل داده‌های فضایی حجیم و چوله را دارا است. نتایج شبیه سازی و تحلیل داده‌های واقعی نیز توانایی این مدل را تایید می‌نمایند.

مراجع

- [1] D. Allard, P. Naveau, A New Spatial Skew-Normal Random Field Model, *Communications in Statistics* **36** (2007), 1821-1834.
- [2] H. Boojari, M. Jafari Khaledi, A non-homogeneous skew-Gaussian Bayesian spatial model, *Statistical Methods and Application*, **25(1)** (2016), 55-73.
- [3] Katzfuss, M., Cressie, N., (2011). Tutorial on Fixed Rank Kriging (FRK) of CO2 data. *Technical Report*. No. **858**.
- [4] H. M. Kim, B. K. Mallik, A Bayesian Prediction Using the Skew- Gaussian Distribution, *Journal of Statistical Planning and Inference*, **120** (2004), 85-101.
- [5] H. Rue, H. Tjelmeland, Fitting Gaussian Markov random fields to Gaussian fields, *Scandinavian Journal of Statistics*, **29 (1)** (2002), 31-49.
- [6] H. Rue, L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall, London, UK, 2005.
- [7] H. Zareifard, M. Jafari Khaledi, Non-Gaussian Modeling of Spatial Data Using Scale Mixing of a Unified Skew Gaussian Process, *Journal of Multivariate Analysis*, **114** (2013), 16-28.