مجله‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار
دوره‌ی ۶، شماره‌ی ۱، بهار و تابستان ۱۳۸۸، صص ۱–۲۳

J. Statist. Res. Iran **6** (2009): 1–23

ارزیابی و کاربرد مدل فضایی گاوسی-لگ گاوسی برای پیشگویی بیزی استوار داده‌های آلودگی هوای تهران

حمیدرضا زارعی فرد و مجید جعفری خالدی*

دانشگاه تربیت مدرس

چکیده. تحلیل داده‌های فضایی از جمله پیشگویی معمولاً با فرض نرمال بودن داده‌ها انجام می‌شود. این در حالی است که چنین فرضی اغلب در عمل برقرار نمی‌باشد. گاهی اوقات نرمال نبودن داده‌ها از وجود داده‌های دورافتاده ناشی می‌شود. در این حالت پالاسیوس و استیل (۲۰۰۶) تعمیمی از مدل گاوسی تحت عنوان مدل گاوسی-لگ گاوسی پیشنهاد نموده و تحلیل بیزی آن را با استفاده از روش‌های مونت کارلوی زنجیر مارکوفی ارایه کردند. از جمله بر مبنای عامل بیزی به شناسایی داده‌های دورافتاده پرداختند. از آن‌جا که محاسبه‌ی عامل بیزی بسیار دشوار می‌باشد، در این مقاله چگال‌ترین ناحیه‌ی پسینی برای شناسایی داده‌های دورافتاده پیشنهاد می‌شود. چون توزیع پسین دارای فرم بسته‌ای نمی‌باشد، برای تعیین این ناحیه از الگوریتم چن و شو استفاده می‌شود. همچنین بر اساس یک مثال شبیه‌سازی و بکارگیری معیار میانگین مجذور خطای پیشگویی، قابلیت مدل گاوسی-لگ گاوسی برای پیشگویی بیزی استوار نشان داده می‌شود. سپس با استفاده از این مدل، پیشگویی بیزی داده‌های آلودگی هوای شهر تهران ارایه شده و عملکرد آن مورد ارزیابی قرار می‌گیرد.

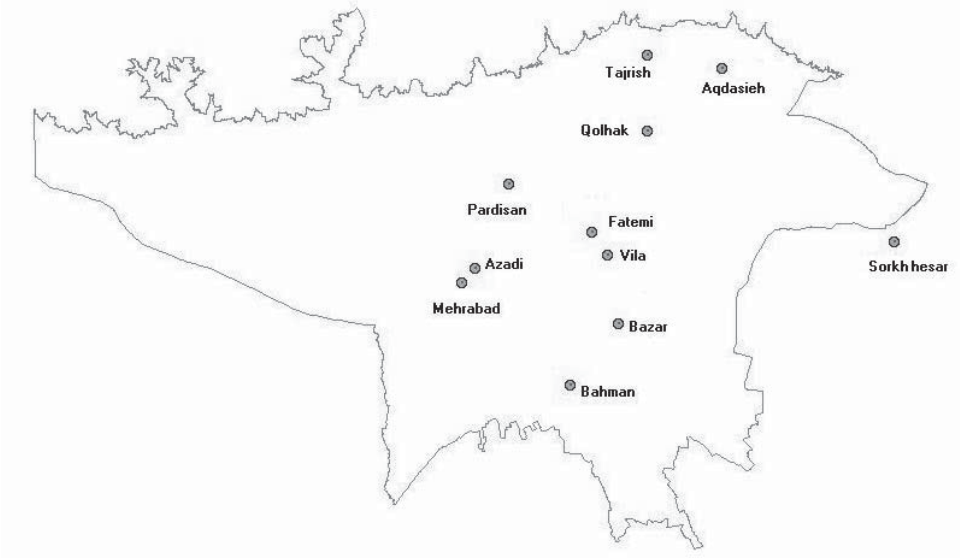
واژگان کلیدی. مدل فضایی گاوسی- لگ گاوسی؛ پیشگویی فضایی استوار؛ رهیافت بیزی؛ چگال‌ترین ناحیه‌ی پسینی؛ روش‌های مونت کارلوی زنجیر مارکوفی؛ میانگین مجذور خطای پیشگویی.

*نویسنده‌ی عهده‌دار مکاتبات

۱ مقدمه

مسئله‌ی آلودگی هوا یکی از معضلات عمده‌ی کلان‌شهر تهران بشمار می‌رود. با توجه به این‌که شهر تهران از سه سمت در محاصره‌ی رشته کوه‌های البرز قرار دارد، آلودگی ناشی از تردد خودروها و دیگر وسایل آلوده‌ساز باعث می‌شود آلاینده‌ها در سطح شهر محبوس شوند و بدون وزش باد مساعد راه خروجی نداشته باشند.

یکی از مهم‌ترین منابع آلودگی هوای شهر تهران گاز منوکسید کربن (CO) است. غلظت منوکسید کربن در نواحی شهری با ترافیک سنگین، به میزان قابل توجهی افزایش می‌یابد. به‌دلیل تأثیر منفی این گاز در متابولیسم تنفسی و فعالیت‌های مغزی افراد، مدل‌بندی و پهنه‌بندی مقادیر (CO) به‌منظور کنترل و کاهش آن، بسیار مورد توجه است. در این خصوص ریواز و همکاران (۲۰۰۷) با استفاده از یک مدل گاوسی، تحلیل فضایی-زمانی آلودگی هوای شهر تهران را بر اساس مشاهدات بدست آمده از ۱۱ ایستگاه سنجش آلودگی هوا (شکل ۱) ارایه نمودند. گرچه فرض گاوسی بودن مشاهدات موجب سادگی استنباط‌ها از قبیل پیشگویی می‌شود، اما اغلب چنین فرضی در عمل برقرار نمی‌باشد. یکی از عوامل تخطی از فرض نرمال بودن مشاهدات، داده‌های دورافتاده می‌باشد.



شکل ۱. محل قرارگیری ایستگاه‌های سنجش آلاینده‌های هوا در شهر تهران

به‌عنوان مثال در مسئله‌ی آلودگی هوای تهران، ایستگاه سرخ حصار به‌علت قرار گرفتن در یک منطقه جنگلی میزان آلودگی بسیار کمی در مقایسه با دیگر ایستگاه‌ها دارد. لذا چنین مشاهده‌ای می‌تواند یک داده‌ی دورافتاده محسوب شود. از آن‌جا که حضور این گونه داده‌ها منجر به ضخامت دم‌های توزیع و بزرگی ضریب کشیدگی می‌شود، در این حالت نمی‌توان از توزیع نرمال که دم‌های باریک‌تری دارد، استفاده کرد. به‌طور کلی شناسایی و مدل‌بندی داده‌های دورافتاده یکی از مشکلاتی است که آماردانان از دیرباز با آن روبه‌رو بوده‌اند و تاکنون رویکردهای مختلفی برای غلبه بر مشکلات ناشی از حضور این مشاهدات ارایه شده است. از آن جمله می‌توان به روش‌های استوار اشاره کرد (میلیتینو و همکاران، ۲۰۰۶ و سریولی و ریانی، ۱۹۹۹). در این روش‌ها تحت فرض نرمال بودن مشاهدات، ارایه‌ی یک تحلیل استوار مورد نظر می‌باشد. اما ممکن است یک مشاهده‌ی دورافتاده به همان الگویی که برای بقیه‌ی داده‌ها برقرار است، متعلق باشد. در این حالت بکارگیری توزیع‌هایی که دارای دم‌های ضخیم‌تر از توزیع نرمال می‌باشند، می‌تواند رهگشا باشد. این موضوع اولین بار توسط جفریز (۱۹۶۱) مورد بررسی قرار گرفت. مارونا (۱۹۷۶) و لانگ و همکاران (۱۹۸۹) برای مدلی که در آن خطاها از توزیع t-استیودنت پیروی می‌کنند، براورد ماکسیمم درستنمایی را مورد بررسی قرار دادند. وست (۱۹۸۴) نیز از خانواده‌ی توزیع‌های مقیاس آمیخته از نرمال برای مدل‌بندی داده‌های دورافتاده استفاده کرد. فرناندز و استیل (۲۰۰۰) نیز با معرفی توزیع‌های پیشین ناسره برای مدل وست، وجود توزیع پسین و گشتاورهایش را بررسی کردند. در زمینه‌ی داده‌های زمین آماری نیز پالاسیوس و استیل (۲۰۰۶) مدل گاوسی-لگ گاوسی را مبتنی بر توزیع‌های مقیاس آمیخته از نرمال معرفی کرده و رهیافت بیزی را برای تحلیل‌ها بکار بردند. آن‌ها همچنین از عامل بیزی به‌منظور شناسایی داده‌های دورافتاده استفاده کردند.

اما از آن‌جا که محاسبه‌ی عامل بیزی بسیار پیچیده و وقت‌گیر است، در این مقاله استفاده از چگال‌ترین ناحیه‌ی پسین (HPD)[1] برای تعیین داده‌های دورافتاده پیشنهاد می‌شود. چون در این حالت توزیع پسین دارای فرم بسته‌ای نمی‌باشد، الگوریتم چن و شو (۱۹۹۹) را برای تعیین این ناحیه به‌کار می‌بریم. به‌علاوه بر اساس یک مثال شبیه‌سازی و به‌کارگیری معیار میانگین مجذور خطای پیشگویی قابلیت مدل گاوسی-لگ گاوسی برای پیشگویی بیزی استوار نشان داده می‌شود. سپس با استفاده از این مدل، پیشگویی بیزی داده‌های آلودگی هوای شهر تهران ارایه شده و عملکرد آن ارزیابی می‌شود.

برای این اهداف، ابتدا در بخش ۲ یک مدل فضایی گاوسی- لگ گاوسی معرفی می‌شود. در بخش ۳ به‌منظور تحلیل بیزی مدل با ترکیب تابع درستنمایی و توزیع پیشین پارامترها، توزیع پسین تعیین می‌شود و بر اساس آن پیشگویی فضایی بیزی انجام می‌شود. لازم به ذکر است که برای انجام محاسبات بیزی،

روش‌های مونت کارلوی زنجیر مارکوفی مورد استفاده قرار می‌گیرد. به‌علاوه نحوه‌ی تعیین چگال‌ترین ناحیه‌ی پسینی ارایه شده و با عامل بیزی مقایسه می‌شود. سپس در بخش ۴ با استفاده از داده‌ی شبیه‌سازی‌شده عملکرد مدل در پیشگویی بیزی استوار و مناسب بودن چگال‌ترین ناحیه‌ی پسینی در شناسایی داده‌های دورافتاده ارزیابی می‌شود. سرانجام در بخش ۵ نحوه‌ی کاربست مدل در خصوص داده‌های آلودگی هوای تهران ارایه شده و عملکرد آن ارزیابی می‌شود.

۲ مدل فضایی گاوسی-لگ گاوسی

فرض کنید میدان تصادفی گاوسی $Z(\cdot) = \{Z(x); x \in D \subseteq R^d\}$، $d \geqslant ۱$ به‌صورت

$$Z(x) = f'(x)\beta + \sigma\varepsilon(x) + \tau\rho(x) \qquad (۱)$$

تجزیه شده باشد، که در آن رویه‌ی میانگین یک تابع خطی از بردار توابع معلوم

$$f'(x) = \{f_۱(x), \ldots, f_k(x)\}$$

و بردار ضرایب رگرسیونی β بوده، مانده‌ی $\varepsilon(\cdot)$ یک میدان مانای مرتبه‌ی دوم با میانگین صفر، واریانس ۱ و تابع همبستگی همسانگرد وابسته به بردار θ به‌صورت

$$\mathrm{corr}\{\varepsilon(x_i), \varepsilon(x_j)\} = C_\theta(||x_i - x_j||) = C_\theta(||h||)$$

در نظر گرفته شده است. لازم به ذکر است هنگامی که $x_i = x_j$ باشد، $\mathrm{corr}\{\varepsilon(x_i), \varepsilon(x_j)\} = ۱$ بوده و بنا بر این $C_\theta(۰) = ۱$ می‌باشد. در ادامه تابع همبستگی $C_\theta(\cdot)$ از رده‌ی انعطاف‌پذیر ماترن

$$C_\theta(||h||) = \frac{۱}{۲^{\theta_۲ - ۱}\Gamma(\theta_۲)} \left(\frac{||h||}{\theta_۱}\right)^{\theta_۲} \kappa_{\theta_۲}\left(\frac{||h||}{\theta_۱}\right)$$

در نظر گرفته می‌شود، که در آن $\theta_۱ > ۰$ پارامتر دامنه، $\theta_۲$ پارامتر همواری و $\kappa_{\theta_۲}$ تابع‌بسل اصلاح‌شده[۲] از نوع سوم و از مرتبه‌ی $\theta_۲$ می‌باشد (اشتین، ۱۹۹۹). همچنین $\rho(\cdot)$ یک میدان تصادفی گاوسی ناهمبسته با میانگین صفر و واریانس ۱ می‌باشد که برای مدل‌بندی خطاهای اندازه‌گیری و اغتشاش یا اصطلاحاً اثر قطعه‌ای به‌کار می‌رود. لازم به ذکر است که در این مدل میدان‌های تصادفی $\rho(\cdot)$ و $\varepsilon(\cdot)$ مستقل در نظر گرفته شده است. پارامترهای σ و τ مثبت بوده و نسبت $\omega^۲ = \frac{\tau^۲}{\sigma^۲}$ بیانگر اهمیت نسبی اثر قطعه‌ای $\tau^۲$ در قیاس با واریانس $\sigma^۲$ است.

اگر در مدل گاوسی (۱) توزیع مانده‌ها از خانواده‌ی توزیع‌های مقیاس آمیخته از نرمال در نظر گرفته شود، مدل گاوسی تعمیم‌یافته به‌صورت

$$Z(x) = f'(x)\beta + \sigma \frac{\varepsilon(x)}{\sqrt{\lambda(x)}} + \tau\rho(x) \qquad (۲)$$

است، که در این مدل $\varepsilon(\cdot)$ و $\rho(\cdot)$ دارای توزیع‌های مشابه با مدل (۱) بوده و میدان تصادفی $\ln\lambda(\cdot)$ از میدان‌های $\rho(\cdot)$ و $\varepsilon(\cdot)$ مستقل در نظر گرفته می‌شود. به‌علاوه فرض می‌شود میدان تصادفی $\ln\lambda(\cdot)$ گاوسی با توزیع‌های متناهی بعد

$$\ln(\lambda) = (\ln\lambda_1, \ldots, \ln\lambda_n)' \sim N_n\left(-\frac{\nu}{2}\underline{1}, \nu C_\theta\right) \qquad \nu > 0 \qquad (۳)$$

باشد، که در آن $\underline{1}$ بردار یک‌ها و $C_\theta = (C_\theta(||x_i - x_j||))$ است. با توجه به توزیع توأم قبل و این موضوع که $C_\theta(0) = 1$ است، به‌وضوح می‌توان مشاهده کرد که هر یک از λ_iها دارای توزیع $\ln(-\frac{\nu}{2}, \nu)$ می‌باشند. در نتیجه برای مقادیر کوچک ν توزیع λ_iها به شدت در حوالی ۱ خواهد بود و هنگامی که ν افزایش پیدا می‌کند توزیع پراکنده‌تر و چوله‌تر می‌شود و مقدار مد توزیع، به سمت صفر انتقال می‌یابد. مدل (۲) با توزیع معرفی شده در (۳)، مدل گاوسی-لگ گاوسی (GLG)[۳] نامیده می‌شود.

لازم به ذکر است برای مدل‌بندی داده‌های دورافتاده می‌توان یک فرایند ناهمبسته‌ی مقیاس آمیخته از نرمال برای مولفه‌ی اثر قطعه‌ای مدل (۱) در نظر گرفت. در این حالت مقدار کوچک متغیرهای آمیزنده بیانگر یک اثر قطعه‌ای بزرگ می‌باشد، که این بیانگر تفاوت قابل ملاحظه‌ی مشاهده‌ی متناظر نسبت به مقادیر مجاورش است. لذا این حالت منطبق با مفهوم سنتی دورافتاده می‌باشد. اما مدل GLG امکان تحلیل داده‌ها هنگامی که یک ناحیه با واریانس بزرگ وجود دارد را فراهم می‌کند. لذا تحت این مدل تفسیر میدان تصادفی آمیزنده تا حدی متفاوت است. در واقع در این حالت مشاهدات متناظر مقادیر کوچک متغیر آمیزنده به یک ناحیه با واریانس مشاهداتی بزرگ متعلق هستند. با این وجود این مشاهدات دورافتاده نامیده می شوند.

حال اگر $Z = (Z_1, \ldots, Z_n)$ داده‌های مربوط به موقعیت‌های نمونه‌ای $(x_1, \ldots, x_n)$ را نمایش دهند، پالاسیوس و استیل (۲۰۰۶) نشان دادند مدل GLG از ویژگی‌های مناسبی برخوردار است. از جمله این‌که ضریب کشیدگی به‌صورت

$$\mathrm{Kurt}(Z_i) = \frac{E(Z_i - \mu_i)^4}{E^2(Z_i - \mu_i)^2} = \frac{3e^{3\nu}}{e^{2\nu}} = 3e^{\nu}$$

است، که به‌وضوح مقداری بزرگ‌تر از ۳ می‌باشد. با توجه به این ضریب کشیدگی هر چه مقدار پارامتر ν بزرگتر باشد، ضریب کشیدگی نیز بزرگ‌تر می‌شود. لذا از این پارامتر می‌توان به‌عنوان یک معیار برای تعیین داده‌های دورافتاده استفاده کرد. از آن‌جا که Z دارای توزیع نرمال شرطی

$$p(z|\beta,\sigma^2,\tau^2,\theta,\Lambda)=N_n\left(X\beta,\sigma^2\left(\Lambda^{-\frac{1}{2}}C_\theta\Lambda^{-\frac{1}{2}}\right)+\tau^2 I_n\right)$$

است، که در آن $\lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ و $X=(f(x_1),\ldots,f(x_n))'$ می‌باشند، تابع درستنمایی به‌صورت

$$L(\beta,\sigma^2,\tau^2,\theta;z)=p(z|\beta,\sigma^2,\tau^2,\theta)=\int_{R^+}\cdots\int_{R^+}p(z|\beta,\sigma^2,\tau^2,\theta,\lambda)\,dP_{\lambda_1}\ldots dP_{\lambda_n}$$

خواهد بود. به‌دلیل مشکلات مبتلا به‌روش بسامدی برای تحلیل مدل از جمله ماکسیمم نمودن تابع درستنمایی قبل، در بخش بعد رهیافت بیزی ارایه می‌شود.

۳ تحلیل بیزی مدل GLG

برای تحلیل بیزی، نخست لازم است توزیع‌های پیشین پارامترهای مدل تعیین شود. از آن‌جا که هر یک از پارامترهای مدل ویژگی خاصی از میدان را کنترل می‌کنند، لذا فرض می‌شود تمام پارامترها از یکدیگر مستقل هستند. بنا بر این توزیع پیشین را می‌توان به‌صورت

$$\pi(\beta,\sigma^2,\omega^2,\nu,\theta)=\pi(\beta)\pi(\sigma^2)\pi(\omega^2)\pi(\nu)\pi(\theta)$$

نوشت. برگر و همکاران (۲۰۰۱) نشان دادند توزیع‌های پسین متناظر با پیشین‌های ناسره‌ی معمول، نظیر پیشین جفریز[۴] ممکن است ناسره گردد. لذا به‌منظور اطمینان از سره بودن پسین، پیشین‌های سره برای هر یک از پارامترهای مدل در نظر گرفته می‌شود. لازم به ذکر است که مقادیر ابرپارامترها را می‌توان با استفاده از روش بیز تجربی تعیین نمود (خالدی و ریواز، ۲۰۰۹). ولی به‌دلیل این‌که در این مدل تعداد آن‌ها زیاد بوده است، بنا بر این براورد آن‌ها دشوار می‌باشد، لذا مقادیر پایه برای آن‌ها اختیار می‌شود. حال برای هر یک از پارامترها توزیع پیشین متناظرشان معرفی می‌شود.

- برای پارامتر β که مؤلفه‌های آن حقیقی مقدار هستند، یک پیشین مبهم[۵] نرمال به‌صورت $\beta\sim N_k(0,c_1 I_k)$ اختیار می‌شود، که در آن c_1 مقدار دلخواه بزرگی است. اگر پیشینی مناسب اختیار شود، انتظار آن است که پسین نیز از همان خانواده از توزیع‌ها باشد، یا به‌عبارت دیگر مزدوج باشد. لذا با توجه به مزدوج بودن پیشین نرمال، از این توزیع برای پارامتر β استفاده شده است.

- اغلب برای پارامتر مثبت σ^{-2} یک پیشین مزدوج مبهم گاما به فرم $\sigma^{-2} \sim Ga(c_2, c_3)$ در نظر گرفته می‌شود، که در آن c_2 و c_3 مقادیر کوچک دلخواهی هستند.

- برای ω^2 پیشین گاوسی معکوس تعمیم‌یافته[۶] به‌صورت $\omega^2 \sim \mathrm{GIG}(\gamma, c_4, c_5)$ اختیار می‌شود، که در آن c، (γ, c_4, c_5) مقادیر دلخواهی هستند. لازم به ذکر است که این توزیع توسط براندوف-نیلسن و همکاران (۱۹۸۲) به‌صورت

$$\mathrm{GIG}(\gamma, c_4, c_5) = \frac{\left(\frac{c_5}{c_4}\right)^{\gamma}}{2\kappa_\gamma(c_4 c_5)} (\omega^2)^{\gamma-1} e^{-\frac{1}{2}\{c_4^2(\omega^2)^{-1} + c_5^2\omega^2\}} \qquad c_4, c_5 \in R^+,\ \gamma \in R$$

معرفی گردید، که در آن κ_γ دلالت به تابع بسل اصلاح‌شده‌ی نوع سوم از مرتبه‌ی γ می‌کند. با توجه به آن‌که پارامتر نامنفی مقدار ω^2 نمایانگر خطای اندازه گیری است و معمولاً شانس بالای مقادیر کوچک این پارامتر مورد انتظار است، توزیع‌های چوله به راست مانند گاما و گامای معکوس را برای این پارامتر می‌توان در نظر گرفت. چون خانواده‌ی توزیع‌های گاوسی معکوس تعمیم‌یافته بسیار کلی بوده و به‌علاوه در حالت خاص $c_4 = 0$ و $\gamma > 0$ توزیع گاما و در حالت $c_5 = 0$ و $\gamma < 0$ توزیع گامای معکوس را در بردارد، به‌کارگیری آن می‌تواند منطقی باشد. این توزیع همچنین در حالت $\gamma = 0$ نیز از انعطاف‌پذیری بالایی برخوردار بوده و با تغییر پارامترهای c_4 و c_5 فرم‌های چگالی متنوعی ارایه می‌کند (بیبی و سورنسن، ۲۰۰۳). لذا استفاده از آن به‌عنوان پیشین پارامترهایی که مثبت بوده و توزیع آن‌ها ممکن است چوله به راست باشد، توجیه‌پذیر به نظر می‌رسد.

- برای پارامتر مثبت ν نیز با استدلال مشابه با ω^2 از توزیع $\nu \sim \mathrm{GIG}(0, c_6, c_7)$ استفاده می‌کنیم، که در آن (c_6, c_7) مقادیر دلخواهی هستند.

- اشتین (۱۹۹۹) نشان داد پارامترهای θ_1 و θ_2 به یکدیگر وابسته می‌باشند، به گونه‌ای که با تغییر در درجه‌ی همواری میدان، دامنه یا شعاع همبستگی فضایی نیز تغییر می‌کند. برای لحاظ نمودن این همبستگی در پیشین توأم $\boldsymbol{\theta} = (\theta_1, \theta_2)$، او پاراماتربندی دیگر از θ_1 به‌صورت $\rho = 2\theta_1\sqrt{\theta_2}$ پیشنهاد کرد و نشان داد این پارامتر جدید همبستگی کم‌تری با پارامتر θ_2 دارد. همچنین پارامتر دامنه‌ی θ_1 ارتباط معکوسی با فاصله‌ی اقلیدسی $||h||$ دارد به گونه‌ای که با افزایش $||h||$ شاهد کاهش وابستگی فضایی هستیم. بر این اساس توزیع پیشین توأم (θ_1, θ_2) به‌صورت

$$\pi(\theta_1, \theta_2) = \pi(\theta_1|\theta_2)\pi(\theta_2) = \exp\left\{\frac{c_8\sqrt{2\theta_2}}{\mathrm{med}(||h||)}\right\}\exp(c_9)$$

اختیار می‌شود، که در آن $\text{med}(||h||)$ میانه‌ی تمام فواصل بین داده‌ها است. توزیع نمایی اختیار شده برای $\pi(\theta_1|\theta_2)$ و $\pi(\theta_2)$ به‌دلیل ساده بودن فرم این توزیع بوده و به‌علاوه بررسی‌های صورت گرفته نشان می‌دهند معمولاً مقادیر کوچک همبستگی و همواری نسبت به مقادیر بزرگ از شانس وقوع بالاتری برخوردار هستند.

۳٫۱ پیشگویی فضایی بیزی

برای پیشگویی به‌روش بیزی لازم است توزیع پیشگوی بیزی تعیین شود. با تعیین این توزیع می‌توان در ارتباط با جنبه‌های مورد علاقه‌ی دیگر از قبیل احتمال تجاوز از یک مقدار آستانه استنباط کرد.

اگر $Z_0 = Z(x_0)$ را به‌عنوان مقدار متغیر پاسخ در مکان x_0 در نظر بگیریم، در این صورت توزیع پیشگو برابر است با

$$f(z_0|z) = \int f(z_0|z,\eta,\lambda,\lambda_0)p(\lambda_0|z,\lambda,\eta)\pi(\lambda,\eta|z)d\eta\, d\lambda\, d\lambda_0, \qquad (۴)$$

که در آن $\boldsymbol{\eta} = (\beta, \sigma^2, \omega^2, \theta, \nu)'$ بردار پارامترهای $p+5$ بعدی مدل و λ_0 متغیر آمیزنده‌ی متناظر با موقعیت x_0 است. اما با توجه به این‌که میدان به‌طور شرطی گاوسی است، بنا بر این

$$f(z_0, z|\eta,\lambda,\lambda_0) = N_{n+1}\left(\mu^*, \sigma^2\lambda^{*\frac{1}{2}}C_\theta^*\lambda^{*\frac{1}{2}} + \tau^2 I_{n+1}\right)$$

است، که در آن $\lambda^* = \text{diag}(\lambda_0, \boldsymbol{\lambda})$ بوده و به‌علاوه

$$\mu^* = \begin{pmatrix} f'(t_0)\beta \\ X\beta \end{pmatrix} \qquad C_\theta^* = \begin{pmatrix} 1 & r_\theta' \\ r_\theta & C_\theta \end{pmatrix}$$

و بردار $r_\theta' = (C_\theta(||x_0 - x_i||))$ است. در نتیجه $Z_0|z,\boldsymbol{\eta},\boldsymbol{\lambda},\lambda_0$ دارای توزیع نرمال با میانگین و واریانس زیر است:

$$E(Z_0|z,\eta,\lambda,\lambda_0) = f'(t_0)\beta + \lambda_0^{-\frac{1}{2}}r_\theta'\lambda^{-\frac{1}{2}}\left(\lambda^{-\frac{1}{2}}C_\theta\lambda^{-\frac{1}{2}} + \omega^2 I_n\right)^{-1}(z - X\beta)$$

$$\text{var}(Z_0|z,\eta,\lambda,\lambda_0) = \sigma^2\left\{\lambda_0^{-1} + \omega^2 - \lambda_0^{-1}r_\theta'\lambda^{-\frac{1}{2}}\left(\lambda^{-\frac{1}{2}}C_\theta\lambda^{-\frac{1}{2}} + \omega^2 I_n\right)^{-1}\lambda^{-\frac{1}{2}}r_\theta\right\}.$$

همچنین $p(\ln\lambda_0|\lambda,z,\eta) = p(\ln\lambda_0|\ln\lambda,\nu)$ است که با توجه به گاوسی بودن میدان تصادفی $\ln\lambda(\cdot)$، توزیع شرطی $\ln\lambda_0|\lambda,\nu$ نرمال با میانگین و واریانس

$$E(\ln\lambda_0|\lambda,\nu) = -\frac{\nu}{2} + r_\theta' C_\theta^{-1}\left(\ln\lambda + \frac{\nu}{2}\underline{1}\right)$$

$$\mathrm{var}(\ln\lambda_\circ|\lambda,\nu)=\nu(1-r_\theta' C_\theta^{-1} r_\theta)$$

می‌باشد. بنا بر این پیشگوی فضایی بیزی به‌صورت

$$\hat{Z}_\circ = E(Z_\circ|z) = \int E(Z_\circ|z,\eta,\lambda,\lambda_\circ)p(\lambda_\circ|z,\lambda,\eta)\pi(\lambda,\eta|z)d\eta\, d\lambda\, d\lambda_\circ \qquad (۵)$$

خواهد بود. از آن‌جا که محاسبه‌ی تحلیلی توزیع پیشگو بیزی (۴) و به دنبال آن پیشگوی فضایی بیزی (۵) بسیار دشوار و بعضاً نشدنی است، با استفاده از روش‌های مونت کارلوی زنجیر مارکوفی از توزیع پسین $p(\lambda,\eta|z)$ نمونه‌گیری کرده و سپس با قرار دادن نمونه‌های به دست آمده در توزیع $p(\lambda_\circ|z,\lambda,\eta)$ و نمونه‌گیری از این توزیع و مجدداً جایگزینی نمونه‌های به دست آمده در $f(z_\circ|z,\eta,\lambda,\lambda_\circ)$ و نمونه‌گیری از این توزیع، می‌توان نمونه‌هایی از توزیع پیشگوی بیزی $f(z_\circ|z)$ به‌صورت $\{z_\circ^{(k)}\}_{k=1}^{l}$ تولید کرد. بنا بر این تقریبی از پیشگویی فضایی بیزی و واریانس پیشگویی به‌صورت

$$\hat{Z}_\circ = \frac{\sum_{k=1}^{l} z_\circ^{(k)}}{l}$$

$$\mathrm{var}(Z_\circ|z) \approx \frac{\sum_{k=1}^{l}(z_\circ^{(k)})^2}{l} - \left\{\frac{\sum_{k=1}^{l} z_\circ^{(k)}}{l}\right\}^2$$

خواهد شد. اما نمونه‌گیری از توزیع پسین $\pi(\lambda,\eta|z)$ و به دنبال آن انجام این فرایند بسیار دشوار است. یک راه حل برای این مسئله، نمونه‌گیری از توزیع پسین $\pi(\lambda,\varepsilon,\eta|z)$ می‌باشد که در آن $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)'$ مانده‌های حاصل در موقعیت‌های $x_1,\ldots,x_n$ می‌باشد. بر این اساس اگر $\{\lambda^{(i)},\varepsilon^{(i)},\eta^{(i)}\}_{i=1}^{m}$ نمونه‌های تولید شده از توزیع پسین $\pi(\lambda,\varepsilon,\eta|z)$ باشند، $\{\lambda^{(i)},\eta^{(i)}\}_{i=1}^{m}$ نمونه‌هایی تولید شده از توزیع پسین $\pi(\lambda,\eta|z)$ می‌باشند.

حال یک الگوریتم MCMC برای نمونه‌گیری از $\pi(\lambda,\varepsilon,\eta|z)$ اجرا می‌شود. برای این منظور یک چارچوب مبتنی بر روش نمونه‌گیری گیبز با تعیین توزیع‌های شرطی کامل اتخاذ می‌شود.

ابتدا برای محاسبه‌ی توزیع شرطی کامل ε، با فرض آن‌که $d=(d_1,\ldots,d_n)$ با $d_i=\frac{\sigma\{Z_i-f'(x_i)\beta\}}{\tau^2\sqrt{\lambda_i}}$ باشد، داریم

$$\begin{aligned}\pi(\varepsilon|z,\lambda,\eta) &\propto f(z|\lambda,\varepsilon,\eta)f(\varepsilon|\theta)\\ &\propto e^{-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{1}{\tau^2}\left\{z_i-f'(x_i)\beta-\sigma\frac{\varepsilon_i}{\sqrt{\lambda_i}}\right\}^2+\varepsilon C_\theta^{-1}\varepsilon\right]}\\ &\propto e^{-\frac{1}{2}(\varepsilon'-W_1^{-1}d)W_1(\varepsilon-W_1^{-1}d)}\end{aligned}$$

که در آن $W_1 = C_\theta^{-1} + \omega^2 \Lambda^{-1}$ است. بنا بر این

$$p(\varepsilon|z, \varepsilon, \lambda, \beta, \sigma^2, \omega^2, \theta, \nu) = N_n(W_1^{-1} d, W_1^{-1}) \quad (۶)$$

در نتیجه نمونه‌گیری از آن به‌سادگی امکان‌پذیر است.

برای محاسبه‌ی توزیع شرطی کامل β با فرض آن‌که $a = (a_1, \ldots, a_n)$ با $a_i = \frac{1}{\tau}(z_i - \sigma \frac{\varepsilon_i}{\sqrt{\lambda_i}})$ باشد، داریم

$$\begin{aligned}
\pi(\beta|z, \varepsilon, \lambda, \beta, \sigma^2, \omega^2, \theta, \nu) &\propto f(z|\lambda, \varepsilon, \eta)\pi(\beta) \\
&\propto e^{-\frac{1}{2}\left[\sum_{i=1}^n \left\{a_i - \frac{f'(x_i)\beta}{\tau}\right\}^2 + \frac{1}{c_1}\beta'\beta\right]} \\
&\propto e^{-\frac{1}{2}\left[\sum_{i=1}^n \left\{\beta' \frac{f'(x_i)f(x_i)}{\tau^2}\beta\right\} - \frac{2}{\tau}\beta' \sum_{i=1}^n \{f(x_i)a_i\} + \frac{1}{c_1}\beta'\beta\right]} \\
&= e^{-\frac{1}{2}\left\{\beta'\left(\frac{X'X}{\tau^2} + \frac{1}{c_1} I_k\right)\beta - \frac{2}{\tau}\beta' X' a\right\}} \\
&\propto e^{-\frac{1}{2}(\beta - W_2^{-1} X' a)' W_2 (\beta - W_2^{-1} X' a)}
\end{aligned}$$

که در آن $W_2 = \frac{1}{c_1} I_k + \frac{1}{\tau^2} X'X$ است. با توجه به رابطه‌ی قبل

$$p(\beta|z, \varepsilon, \lambda, \sigma^2, \omega^2, \theta, \nu) = N_n\left(\frac{1}{\tau} W_2^{-1} X' a, W_2^{-1}\right) \quad (۷)$$

می‌باشد و در نتیجه نمونه‌گیری از آن نیز به‌سادگی امکان‌پذیر است.

توزیع شرطی کامل $\gamma = \sigma^{-2}$ متناسب است با

$$\begin{aligned}
\pi(\gamma|z, \varepsilon, \lambda, \beta, \omega^2, \theta_1, \theta_2, \nu) &\propto f(z|\lambda, \varepsilon, \eta)\pi(\gamma) \\
&\propto \gamma^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^n \frac{\gamma}{\omega}\left\{z_i - f'(x_i)\beta - \frac{\varepsilon_i}{\sqrt{\gamma}\sqrt{\lambda_i}}\right\}^2} \pi(\gamma).
\end{aligned}$$

در نتیجه با توجه به این‌که توزیع شرطی کامل قبل دارای صورت تحلیلی مشخصی نمی‌باشد، برای نمونه‌گیری از آن می‌توان از الگوریتم متروپلیس–هستینگز استفاده کرد. در روش متروپلیس–هستینگز اگر $\gamma^{(t)}$ مقدار شبیه‌سازی‌شده در مرحله‌ی tام باشد، نمونه‌ی جدید را از تابع نامزد $q(\gamma^*|\gamma^{(t)})$ بدست آورده و سپس آن

را با احتمال

$$r=\frac{\gamma^{*\frac{n}{2}}e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{\gamma^*}{\omega^2}\left(z_i-f'(x_i)\beta-\frac{\varepsilon_i}{\sqrt{\gamma^*}\sqrt{\lambda_i}}\right)^2}\pi(\gamma^*)\frac{1}{\gamma^{(t)}}}{\gamma^{(t)\frac{n}{2}}e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{\gamma^{(t)}}{\omega^2}\left(z_i-f'(x_i)\beta-\frac{\varepsilon_i}{\sqrt{\gamma^{(t)}}\sqrt{\lambda_i}}\right)^2}\pi(\gamma^{(t)})\frac{1}{\gamma^*}}$$

می‌پذیریم. برای نمونه‌گیری از توزیع شرطی کامل ω^2 نیز الگوریتم مشابهی به کار می‌رود.

توزیع شرطی کامل θ_1 متناسب با

$$\begin{aligned}\pi(\theta_1|z,\varepsilon,\lambda,\beta,\omega^2,\sigma^2,\theta_2,\nu) &\propto f(z|\varepsilon,\lambda,\eta)f(\varepsilon|\lambda,\theta_1,\theta_2)p(\lambda|\theta_1,\theta_2,\nu)\pi(\theta_1|\theta_2)\\ &\propto f(\varepsilon|\theta_1,\theta_2)p(\lambda|\theta_1,\theta_2,\nu)\pi(\theta_1|\theta_2)\\ &\propto |C_\theta|^{-\frac{1}{2}}e^{-\frac{1}{2}\varepsilon' C_\theta\varepsilon}|C_\theta|^{-\frac{1}{2}}e^{-\frac{1}{2}(\ln\lambda+\frac{\nu}{2}I)C_\theta^{-1}(\ln\lambda+\frac{\nu}{2}I)}\\ &\quad\times e^{-\frac{c_\lambda\sqrt{2\theta_2}}{\mathrm{med}(d)}\theta_1}\end{aligned}$$

می‌باشد. برای نمونه‌گیری از توزیع شرطی کامل θ_1 به‌طور مشابه با توزیع شرطی کامل σ^{-2} می‌توان الگوریتم متروپلیس-هستینگز را بکار برد.

توزیع شرطی کامل θ_2 به‌صورت

$$\pi(\theta_2|z,\varepsilon,\lambda,\beta,\omega^2,\sigma^2,\theta_1,\nu)\propto p(\varepsilon|\theta_1,\theta_2)p(\lambda|\theta_1,\theta_2,\nu)\pi(\theta_1|\theta_2)\pi(\theta_2)$$

می‌باشد. برای نمونه‌گیری از این توزیع شرطی، مشابه با توزیع شرطی θ_1 عمل می‌کنیم.

توزیع شرطی کامل ν متناسب است با

$$\begin{aligned}\pi(\nu|z,\varepsilon,\lambda,\beta,\omega^2,\sigma^2,\theta_1,\theta_2) &\propto f(z|\varepsilon,\lambda,\eta)f(\varepsilon|\lambda,\theta)p(\lambda|\theta,\nu)\pi(\nu)\\ &\propto p(\lambda|\nu,\theta)\pi(\nu)\\ &\propto \frac{1}{\nu^{\frac{n}{2}}}e^{-\frac{1}{2}(\ln\lambda+\frac{\nu}{2}I)C_\theta^{-1}(\ln\lambda+\frac{\nu}{2}I)}\pi(\nu).\end{aligned}$$

برای نمونه‌گیری از آن می‌توان از الگوریتم متروپلیس-هستینگز استفاده کرد.

به‌دلیل وجود همبستگی بین متغیرهای $\lambda_1,\ldots,\lambda_n$ و به‌علاوه بزرگی بعد λ برای اندازه‌ی نمونه‌ی بزرگ، شبیه‌سازی از $p(\lambda|z,\varepsilon,\eta)$ دشوار است. برای رفع این مشکل عناصر λ را به بلوک‌های مختلف به گونه‌ای افراز می‌کنیم که عناصر نسبتاً همگن در یک بلوک قرار گیرند. با این عمل بیش‌تر وابستگی بین λ_iها به عناصر واقع در خوشه‌ها محدود می‌شود. حال نمونه‌گیری از توزیع شرطی کامل λ، مبتنی

بر یک چارچوب نمونه‌گیری گیبز صورت می‌پذیرد. برای توضیح این موضوع فرض کنید $\lambda_{(i)}$ دلالت به n_i عنصر همگن واقع در خوشه‌ی iام کند و $\lambda_{-(i)}$ دلالت به بقیه‌ی عناصر کند. بردار λ به‌صورت $\lambda' = (\lambda_{-(i)}, \lambda_{(i)})'$ در نظر گرفته شده و به‌طور متناظر ماتریس همبستگی به‌صورت

$$C_\theta = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

افراز می‌شود. در این صورت توزیع پسین شرطی کامل $\lambda_{(i)}$ به‌صورت

$$p(\lambda_{(i)}|z, \lambda_{(-i)}, \varepsilon, \eta) \propto e^{\frac{-1}{2\tau^2}\sum_{i\in(i)}\left\{z_i - f'(x_i)\beta - \sigma\frac{\varepsilon_i}{\sqrt{\lambda_i}}\right\}^2} p(\lambda_{(i)}|\lambda_{(-i)}, \theta, \nu) \qquad (۸)$$

خواهد بود، زیرا

$$\begin{aligned} p(\lambda_{(i)}|z, \lambda_{(-i)}, \varepsilon, \eta) &= \frac{p(\lambda_{(i)}, \lambda_{(-i)}, \varepsilon, z, \beta, \sigma^2, \omega^2, \theta, \nu)}{p(\lambda_{(-i)}, \varepsilon, z, \beta, \sigma^{-2}, \omega^2, \theta, \nu)} \\ &\propto p(\lambda_{(i)}|\lambda_{(-i)}, \theta, \nu)\frac{p(z, \varepsilon, \beta, \sigma^{-2}, \omega^2|\lambda, \theta, \nu)}{p(\lambda_{(-i)}, z, \varepsilon, \beta, \sigma^{-2}, \omega^2, \nu)} \\ &\propto p(z|\varepsilon, \lambda, \beta, \sigma^{-2}, \omega^2, \theta, \nu).p(\lambda_{(i)}|\lambda_{(-i)}, \theta, \nu) \\ &\propto e^{\frac{-1}{2\tau^2}\sum_{i\in(i)}\left\{z_i - f'(x_i)\beta - \sigma\frac{\varepsilon_i}{\sqrt{\lambda_i}}\right\}^2}.p(\lambda_{(i)}|\lambda_{(-i)}, \theta, \nu) \end{aligned}$$

که در آن $i \in (i)$ به همه‌ی مشاهدات واقع در خوشه‌ی iام دلالت می‌کند. اولین عامل در رابطه‌ی (۸) در واقع متناسب با حاصل‌ضرب توزیع‌های نرمال بریده‌شده[۷] در صفر به‌صورت

$$\lambda_i^{-\frac{1}{2}}|z_i, \varepsilon_i, \beta, \sigma, \omega \sim NT\left\{\frac{z_i - f'(x_i)\beta}{\sigma\varepsilon_i}, \frac{\omega^2}{\varepsilon_i^2}\right\}$$

است. $p(\lambda_{(i)}|\lambda_{(-i)}, \theta, \nu)$ پیشین شرطی است که با توجه به توزیع پیشین λ در رابطه‌ی (۳) تعیین می‌شود. بنا بر این

$$\ln\lambda_{(i)}|\lambda_{(-i)}, \theta, \nu \sim N(w_{(i)}, R_{(i)})$$

می‌باشد، که در آن

$$\begin{aligned} w_{(i)} &= \frac{\nu}{2}(C_{21}C_{11}^{-1}j - j) + C_{21}C_{11}^{-1}\ln\lambda_{-i} \\ R_{(i)} &= \nu(C_{22} - C_{21}C_{11}^{-1}C_{12}). \end{aligned}$$

با توجه به اینکه توزیع شرطی کامل $\lambda_{(i)}$ دارای فرم تحلیلی مشخصی نمی‌باشد، برای نمونه‌گیری از آن می‌توان الگوریتم متروپلیس-هستینگز استفاده کرد.

۳٫۲ شناسایی داده‌های دورافتاده

در مدل GLG مشاهدات با λ_iهای کوچک تمایل دارند دور از رویه‌ی میانگین قرار گیرند و به نوعی دورافتاده محسوب می‌شوند. به‌عبارت دیگر این مشاهدات به ناحیه‌ای با واریانس نمونه‌ای بزرگ‌تر نسبت به سایر مشاهدات متعلق هستند. بنا بر این برای شناسایی داده‌هایی که مشکوک به دورافتاده بودن هستند می‌توان از آزمون فرضیه‌های

$$\begin{cases} H_۰ : \lambda_i = ۱ \\ H_۱ : \lambda_i \neq ۱ \end{cases}$$

استفاده نمود. در این حالت پالاسیوس و استیل (۲۰۰۶) عامل بیزی

$$B_i = \alpha \frac{p(\lambda_i|z)}{p(\lambda_i)}|_{\lambda_i=۱} \qquad (۹)$$

را برای آزمودن این فرضیه‌ها پیشنهاد نمودند، که در آن $\frac{p(\lambda_i|z)}{p(\lambda_i)}$ را نسبت چگالی ساواج و دیکی می‌نامند و عامل تصحیح‌کننده $\alpha = E\left\{\frac{p(\ln \lambda_{-i})}{p(\ln \lambda_{-i}|\lambda_i=۱)}\right\}$ است، که در آن λ_{-i} از حذف iامین عنصر λ بدست آمده است. این امید ریاضی بر اساس توزیع $p(\lambda_{-i}|z, \lambda_i = ۱)$ محاسبه می‌شود. برای تعیین α با توجه به رابطه‌ی (۳) به‌سادگی می‌توان نشان داد

$$\ln \lambda_{-i}|\lambda_i = ۱ \sim N_{n-۱}\left(-\frac{\nu}{۲}(۱ - c_i), \nu(C_\theta^{-i} - c_i c_i')\right)$$

که در آن C_θ^{-i} از حذف iامین سطر و ستون C_θ بدست می‌آید و c_i یک بردار از عناصر $C_\theta(||x_i - x_j||)$، $j \neq i$، می‌باشد. بنا بر این اگر به‌صورت مشابه با بخش قبل نمونه‌های $\{\lambda_{-i}^{(j)}\}_{j=۱}^m$ از توزیع پسین $p(\lambda_{-i}|z, \lambda_i = ۱)$ تولید شوند، براورد عامل تصحیح‌کننده به‌صورت

$$\hat{\alpha} = \hat{E}\left\{\frac{p(\ln \lambda_{-i})}{p(\ln \lambda_{-i}|\lambda_i = ۱)}\right\} = \sum_{j=۱}^{m} \frac{p(\ln \lambda_{-i}^{(j)})}{p(\ln \lambda_{-i}^{(j)}|\lambda_i = ۱)} \qquad (۱۰)$$

می‌باشد. برای محاسبه‌ی نسبت چگالی ساواج و دیکی با توجه به رابطه‌ی $\frac{p(\lambda_i|z)}{p(\lambda_i)}|_{\lambda_i=۱} = \frac{f(z|\lambda_i=۱)}{f(z)}$، روش نیوتن و رفتری (۱۹۹۴) را می‌توان بکار برد. برای این منظور $f(z|\lambda_i)$ و $f(z)$ به‌ترتیب بر اساس نمونه‌های حاصل از توزیع پسین

$$\pi(\lambda_{-i}, \varepsilon, \beta, \sigma^۲, \theta, \omega^۲, \nu|\lambda_i = ۱, z) \quad \text{و} \quad \pi(\lambda, \varepsilon, \beta, \sigma^۲, \theta, \omega^۲, \nu|z)$$

براورد می‌شود. به این ترتیب محاسبه‌ی عامل بیزی برای هر مشاهده مستلزم شبیه‌سازی‌های جداگانه از توزیع پسین $\pi(\lambda_{-i}, \varepsilon, \beta, \sigma^۲, \theta, \omega^۲, \nu|\lambda_i = ۱, z)$ می‌باشد. لذا هر چه اندازه‌ی نمونه بزرگ‌تر باشد، حجم محاسبات نیز افزایش بیش‌تری می‌یابد. در این حالت محاسبه‌ی عامل بسیار دشوار و زمان‌بر خواهد بود. در این مقاله برای آزمون فرض مورد نظر، استفاده از ناحیه‌ی HPD پیشنهاد می‌شود. یک ناحیه‌ی HPD $۱۰۰(۱-\alpha)$٪ برای λ_i به‌صورت

$$R^i(\pi_\alpha) = \{\lambda_i : \pi(\lambda_i|z) \geqslant \pi_\alpha\} \qquad (۱۱)$$

تعریف می‌شود، که در آن π_α بزرگ‌ترین مقداری است که به ازای آن $p(\lambda_i \in R^i(\pi_\alpha)) \geqslant ۱-\alpha$ باشد. این ناحیه به‌دلیل داشتن دو ویژگی

- نقاط درون ناحیه‌ی HPD دارای بیش‌ترین چگالی پسینی نسبت به نقاط بیرون فاصله هستند،
- ناحیه‌ی HPD کوتاه‌ترین فاصله برای یک احتمال مفروض است،

بسیار مورد توجه است. اما از آن‌جا که معمولاً تعیین ناحیه‌ی HPD با روش‌های تحلیلی به‌دلیل مشخص نبودن فرم بسته‌ی توزیع پسین غیر ممکن است، در این حالت چن و شو (۱۹۹۹) روشی بر پایه‌ی نمونه‌های MCMC حاصل از توزیع پسین ارایه کردند. در این روش اگر $\{\lambda_{i_j}\}_{j=۱}^n$ نمونه‌های MCMC بدست آمده از توزیع پسین $\pi(\lambda_i|z)$ باشد و $\lambda_{i_{(j)}}$ بیانگر jامین آماره‌ی ترتیبی باشد، یک ناحیه‌ی HPD برای λ_i به‌صورت

$$R^i_{k^*}(n) = (\lambda_{i_{(k^*)}}, \lambda_{i_{(k^*+[(۱-\alpha)n])}})$$

بدست می‌آید، که در آن $[(۱-\alpha)n]$ بیانگر جزء صحیح $(۱-\alpha)n$ می‌باشد و k^* به گونه‌ای انتخاب می‌شود که

$$\lambda_{i_{(k^*+[(۱-\alpha)n])}} - \lambda_{i_{(k^*)}} = \min_{۱\leqslant k\leqslant n-[(۱-\alpha)n]} (\lambda_{i_{(k+[(۱-\alpha)n])}} - \lambda_{i_{(k)}}).$$

چن و شو (۱۹۹۹) همچنین نشان دادند که اگر این رابطه تنها یک جواب داشته باشد، آن‌گاه

$$R^i_{j^*}(n) \longrightarrow R^i(\pi_\alpha) \quad a.s. \quad as \quad n \longrightarrow \infty$$

اکنون برای تعیین ناحیه HPD بر اساس الگوریتم زیر عمل می‌کنیم:

۱. یک نمونه‌ی MCMC از $\pi(\lambda_i|z)$ به‌صورت $\{\lambda_{i_j}, j = ۱, ۲, \ldots, m\}$ بدست آورده می‌شود.

۲. برای تعیین آماره‌های ترتیبی، مقادیر $\{\lambda_{i_j}, j = ۱, ۲, \ldots, m\}$ به‌صورت

$$\lambda_{i_{(۱)}} \leqslant \lambda_{i_{(۲)}} \leqslant \cdots \leqslant \lambda_{i_{(m)}}$$

منظم می‌شوند.

۳. بازه‌ی باورمندی $۱۰۰(۱-\alpha)٪$ برای هر یک از λ_iها به‌صورت

$$R_k^i(n) = (\lambda_{i_{(k)}}, \lambda_{i_{(k+[(۱-\alpha)n])}}) \quad k = ۱, ۲, \ldots, n - [(۱-\alpha)n]$$

محاسبه می‌شوند.

۴. فاصله‌ای که دارای کم‌ترین طول در میان تمامی فواصل فوق باشد، به‌عنوان ناحیه‌ی HPD $۱۰۰(۱-\alpha)٪$ برای λ_i بکار برده می‌شود.

با توجه به این الگوریتم به‌وضوح مشخص است که با یک مرتبه تولید نمونه از توزیع پسین $\pi(\lambda|z)$، قادر به محاسبه‌ی ناحیه‌ی HPD برای تمام مشاهدات هستیم. لذا استفاده از این ناحیه برای شناسایی داده‌های دورافتاده، ساده بوده و به‌علاوه منجر به صرفه‌جویی چشمگیر در زمان می‌شود. لازم به ذکر است که در این مقاله $\alpha = ۰/۰۵$ اختیار می‌شود.

۴ بررسی شبیه‌سازی

در این بررسی، نحوه‌ی تعیین ناحیه‌ی HPD و عملکرد مدل GLG در پیشگویی استوار ارزیابی می‌شود. برای این منظور ناحیه‌ی مورد مطالعه‌ی D به‌صورت مربع $[۰, ۱] \times [۰, ۱]$ اختیار و ۵۰ موقعیت نمونه‌ای به‌طور تصادفی در آن انتخاب شده است. سپس یک میدان تصادفی گاوسی با پارامتر روند $\beta_۱ = ۵$، واریانس $\sigma^۲ = ۱$، همبستگی فضایی و همواری $\theta_۱ = \theta_۲ = ۰/۵$ در موقعیت‌های نمونه‌ای با استفاده از تابع grf محیط Fields Random از نرم‌افزار R شبیه‌سازی می‌شود. در مرحله‌ی بعد، سه مشاهده‌ی ۱، ۵ و ۴۴ را به‌طور تصادفی انتخاب کرده و از مشاهده‌ی ۱ دو واحد کم و به مشاهده‌ی ۵ دو واحد اضافه و به مشاهده‌ی ۴۴ یک واحد اضافه کردیم. این حالت را داده‌های دورافتاده ملایم نامیدیم. به‌علاوه، حالت داده‌های دورافتاده‌ی قوی با کم کردن چهار واحد از مشاهده‌ی ۱ و اضافه کردن چهار واحد به مشاهده‌ی ۵ و اضافه کردن پنج واحد به مشاهده‌ی ۴۴ ایجاد گردید. برای پارامترهای روند و واریانس از توزیع‌های توصیف شده در بخش سوم متناظر با حالت مبهم $c_۱ = ۱۰^۴$ و $c_۲ = c_۳ = ۱۰^{-۶}$ استفاده گردید.

جدول ۱. عامل بیزی برای مشاهدات انتخاب شده

شماره‌ی مشاهده	moderate outliers		strong outliers	
	BF برای $\lambda_i = ۱$	HPD	BF برای $\lambda_i = ۱$	HPD
۱	۰/۲	(۰/۰۰۷, ۱/۱)	۰	(۰/۰۰۰۱, ۰/۲)
۵	۰/۰۷۹	(۰/۰۰۸, ۰/۹۶)	۰	(۰/۰۰۰۴, ۰/۱۷)
۴۴	۰/۳	(۰/۰۳, ۱/۵)	۰	(۰/۰۰۱, ۰/۱۵)

جدول ۲. ارزیابی مدل گاوسی و GLG برای پیشگویی بیزی استوار بر اساس معیار MSPE

سطح حضور داده‌های پرت	مدل گاوسی	مدل GLG
Outliers Modarete	۰/۲۲	۰/۱۶
Outliers Strong	۰/۸۸	۰/۱۶

ابرپارامترهای توزیع‌های پیشین سهم اثر قطعه‌ای، پارامتر متغیر آمیزنده و پارامترهای تابع همبستگی نیز به گونه‌ای در نظر گرفته شد که میانگین و واریانس توزیع پیشین مقادیر متوسطی باشند. بنا بر این

$$(c_۴, c_۵) = (۰/۶, ۱), \qquad (c_۶, c_۷) = (۰/۵, ۲)$$
$$c_۸ = ۰/۹, \qquad c_۹ = ۰/۵$$

اختیار می‌کنیم. با توجه به نمودارهای زمان داغیدن دوره‌ی داغیدن ۳۰۰۰۰۰ اختیار شد. ضمن آن‌که تعداد تکرارهای بعد از زمان داغیدن ۲۰۰۰۰۰ در نظر گرفته شد. برای کاهش وابستگی موجود بین نمونه‌های حاصل از الگوریتم MCMC و انجام تحلیل مناسب‌تر، هر پنجمین نمونه از تکرارهای بعد از زمان داغیدن استخراج می‌شود.

با استفاده از ناحیه‌ی HPD ۹۵٪ و عامل بیزی می‌توان حضور داده‌های دورافتاده را بررسی کرد. در جدول ۱ این معیارها برای مشاهده ۱، ۵ و ۴۴ در دو حالت داده‌های دورافتاده‌ی ملایم و قوی آورده شده است. با توجه به نتایج این جدول، در حالت داده‌های دورافتاده‌ی قوی هر دو معیار دلالت به دورافتاده بودن مشاهدات مد نظر می‌کنند. در حالت داده‌های دورافتاده‌ی ملایم نیز هر دو معیار نتایج مشابهی را ارایه کرده‌اند. در واقع برای داده‌ی ۱ و ۴۴ عامل بیزی به ترتیب ۰/۲ و ۰/۳ است که نشان از چندان دورافتاده نبودن این دو دارد. ناحیه‌ی HPD نیز برای این دو داده مقدار $\lambda_i = ۱$ را در بر می‌گیرد، که به نوعی دلالت

بر این موضوع دارد. این درحالی است که فاصله‌ی HPD با صرف زمان بسیار کم‌تری این نتایج را ارایه کرده است.

برای ارزیابی عملکرد پیشگو، ۱۶ نقطه‌ی دیگر بر روی یک شبکه‌ی منظم به مختصات‌های

$$\{۰/۲, ۰/۴, ۰/۶, ۰/۸\} \times \{۰/۲, ۰/۴, ۰/۶, ۰/۸\}$$

انتخاب و داده‌ها در این موقعیت‌ها شبیه‌سازی گردید. سپس معیار میانگین مجذور خطای پیشگویی $\text{MSPE} = \frac{۱}{۱۶}\sum_{i=۱}^{۱۶}(Z_i - \hat{Z}_i)^۲$ محاسبه شد، که در آن Z_i مقدار واقعی موقعیت iام از شبکه‌ی منظم و $\hat{Z}_i$ مقدار پیشگویی‌شده در این موقعیت را نمایش می‌دهد. با توجه به این معیار مشاهده می‌شود که مدل GLG عملکرد بسیار بهتری داشته است که این موضوع به ویژه در حالت داده‌های دورافتاده‌ی قوی مشخص‌تر است. ضمن آن‌که عملکرد پیشگویی مدل گاوسی به شدت تغییر یافته است. اما پیشگویی در مدل GLG تحت تأثیر داده‌های دورافتاده قرار نگرفته و استوار بوده است.

۵ تحلیل داده‌های آلودگی هوای تهران

در این بخش، متوسط‌های هفتگی میزان غلظت CO به ppm مربوط به سال ۱۳۸۵ که توسط سازمان محیط زیست و شرکت کنترل کیفیت هوای تهران در ۱۱ ایستگاه سنجش آلاینده‌های هوا اندازه‌گیری و ثبت شده‌اند، مورد تحلیل قرار می‌گیرد. چون در جمع‌آوری داده‌ها خطای اندازه‌گیری و خطای ثبت وجود دارد، لذا منطقی است که مشاهدات نوفه‌دار فرض شوند. جدول ۳، متوسط‌های هفتگی CO را برای هر یک از ایستگاه‌ها به همراه طول و عرض جغرافیایی آن‌ها بر حسب درجه-دقیقه و متر نشان می‌دهد. لازم به توضیح است که مختصات ایستگاه‌ها بر حسب متر، نسبت به مبدأ مختصات تقاطع نصف‌النهار مبدأ (گرینویچ) و خط استوا ارایه شده‌اند. همان‌طور که از این جدول مشاهده می‌شود میزان آلودگی در ایستگاه سرخ حصار به‌دلیل قرارگیری آن در یک مکان جنگلی، به‌طور قابل ملاحظه‌ای از دیگر ایستگاه‌ها کم‌تر است. بنا بر این بکارگیری مدل GLG منطقی به نظر می‌رسد. مدل تابعی روند به‌صورت

$$x = (x_۱, x_۲)$$

$$\mu(x) = f'(x)\beta = \beta_۱ + \beta_۲x_۱ + \beta_۳x_۱^۲ + \beta_۴x_۲ + \beta_۵x_۱x_۲ + \beta_۶x_۲^۲ \qquad (۱۲)$$

در نظر گرفته می‌شود، که در آن $x_۱$ و $x_۲$ به‌ترتیب بیانگر طول و عرض جغرافیایی هستند. لازم به ذکر است به‌دلیل نسبتاً کلی بودن فرم تابعی (۱۲)، معمولاً در عمل از آن برای مدل‌بندی روند استفاده می‌شود. ضمن آن‌که در مسئله‌ی آلودگی هوای تهران، مرکز شهر آلوده‌تر نسبت به سایر مکان‌ها بوده و هر اندازه از مرکز دورتر

جدول ۳. موقعیت جغرافیایی ۱۱ ایستگاه سنجش آلودگی هوا در تهران

ایستگاه	بر حسب درجه		بر حسب متر		میانگین CO
	طول جغرافیایی	عرض جغرافیایی	طول جغرافیایی	عرض جغرافیایی	
آزادی	۵۱°۲۰′	۳۵°۴۲′	۵۳۰۶۰۶	۳۹۵۱۱۷۱	۵/۷۷
اقدسیه	۵۱°۲۹′	۳۵°۴۸′	۵۴۳۸۵۲	۳۹۶۱۸۸۱	۵/۷۱
بازار	۵۱°۲۵′	۳۵°۴۰′	۵۳۸۲۹۸	۳۹۴۸۲۳۳	۶/۷۲
بهمن	۵۱°۲۳′	۳۵°۳۵′	۵۳۵۷۴۳	۳۹۴۴۹۰۵	۵/۶۴
پردیسان	۵۱°۲۲′	۳۵°۴۴′	۵۳۳۶۰۸	۳۹۵۵۳۱۷	۶/۰۱
تجریش	۵۱°۲۶′	۳۵°۴۸′	۵۳۹۸۴۳	۳۹۶۲۶۱۳	۵/۹۱
سرخ حصار	۵۱°۳۵′	۳۵°۴۲′	۵۵۳۰۷۹	۳۹۵۲۵۸۹	۱/۳۱
فاطمی	۵۱°۲۴′	۳۵°۴۳′	۵۳۶۸۹۳	۳۹۵۳۱۰۵	۷/۱۰
قلهک	۵۱°۲۶′	۳۵°۴۶′	۵۳۹۸۶۱	۳۹۵۸۵۲۸	۴/۹۷
مهرآباد	۵۱°۱۹′	۳۵°۴۱′	۵۲۹۹۰۰	۳۹۵۰۳۸۰	۴/۵۰
ویلا	۵۱°۲۵′	۳۵°۴۲′	۵۳۷۶۹۰	۳۹۵۱۸۸۲	۷/۱۰

می‌شویم، آلودگی کاهش می‌یابد (ریواز و همکاران، ۲۰۰۷). بنا بر این فرم درجه‌ی دوم (۱۲) برای تابع روند منطقی به نظر می‌رسد. اگر چه می‌توان منابع انتشار آلودگی مانند کارخانه‌ها، ترافیک شدید و غیره را به‌عنوان متغیر کمکی در تابع روند لحاظ نمود و نتایج تحلیل را بهبود بخشید، اما در این مقاله به‌دلیل در اختیار نداشتن این اطلاعات، تنها از طول و عرض جغرافیایی محل ایستگاه‌ها استفاده شده است. شکل ۲ زمان داغیدن تعدادی از پارامترهای مدل را نشان می‌دهد. با توجه به این نمودار دوره‌ی داغیدن ۴۰۰۰۰۰ اختیار شد، تعداد تکرارهای بعد از زمان داغیدن نیز ۴۰۰۰۰۰ در نظر گرفته شد. ضمن آن‌که برای کاهش وابستگی موجود بین نمونه‌های حاصل از الگوریتم MCMC، مجدداً نتایج بر اساس هر پنجمین استخراج از تکرارهای بعد از زمان داغیدن ارایه می‌شود.

سپس با استفاده از عامل بیزی و چگال‌ترین ناحیه‌ی پسینی، معنی داری هر یک از ضرایب رگرسیونی در رابطه‌ی (۱۲) تعیین شد. بر اساس این معیارها یک روند از درجه‌ی اول برحسب طول و عرض جغرافیایی مناسب تشخیص داده شد. سپس، مقدار عامل بیزی در حمایت از مدل GLG در حدود ۵ بدست آمده، که بیانگر حمایت داده‌ها از مدل GLG است. برای بررسی حضور داده‌ی دورافتاده در بین مشاهدات، میانگین پسین متغیرهای آمیزنده محاسبه شد، که همان‌طور که قابل پیش‌بینی بود کوچک‌ترین مقدار مربوط به ایستگاه سرخ حصار بود. در جدول ۴ با استفاده از معیارهای معرفی‌شده برای شناسایی

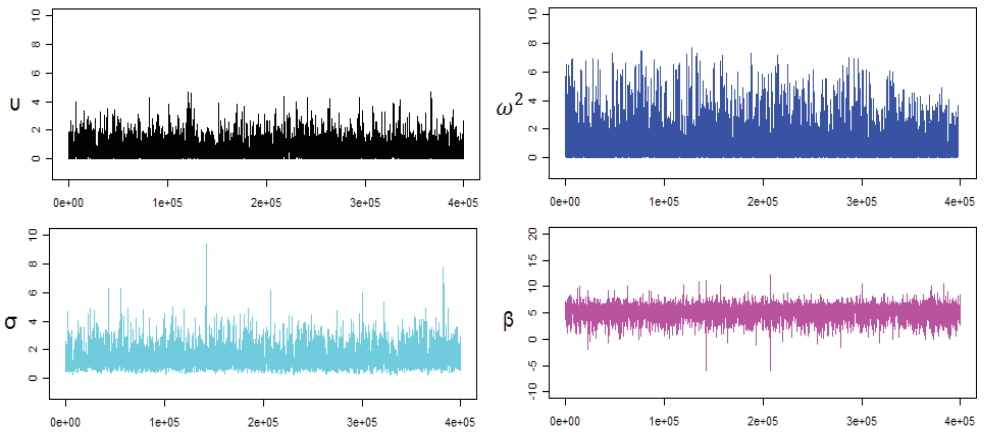
جدول ۴. معیارهای بررسی داده‌ی دورافتاده برای ایستگاه

ناحیه‌ی HPD	فاکتور بیزی	ایستگاه
$(۰/۰۳, ۰/۸)$	۰/۲۷	سرخ حصار

داده‌های دورافتاده، این ایستگاه به‌طور دقیق‌تر مورد ارزیابی قرار گرفت. با توجه به این جدول، معیارها دلالت به دورافتاده بودن مشاهده‌ی حاصل از این ایستگاه می‌کند.

در ادامه معیار اعتبارسنجی متقابل برای ارزیابی عملکرد دو مدل در پیشگویی مورد بررسی قرار گرفت. این معیار به‌صورت $\frac{۱}{n}\sum_{i=۱}^{n}\varepsilon_{(i)}^{۲}$ تعریف می‌شود که در آن $\varepsilon_{(i)}$ باقی‌مانده‌ی حاصل از اختلاف مقادیر مشاهده‌شده و پیش‌بینی‌شده بر اساس سایر مشاهدات می‌باشد. مقدار این معیار برای دو مدل گاوسی و GLG به ترتیب برابر ۵/۰۱ و ۲/۶ بدست آمد که بیانگر عملکرد بسیار بهتر مدل GLG در پیشگویی می‌باشد. بررسی نتایج مربوط به باقی‌مانده‌های اعتبارسنجی، نشان می‌دهد که این مقدار برای ایستگاه سرخ حصار تحت مدل گاوسی بسیار بیش‌تر از مدل GLG بوده است.

اکنون تحت مدل GLG می‌توان برای هر نقطه‌ی دلخواه پیشگویی بیزی و واریانس پیشگویی را بدست آورد. به‌عنوان مثال در موقعیت میدان انقلاب با طول جغرافیایی ۵۳۸۶۱۴ و عرض جغرافیایی



شکل ۲. نمودار دوره‌ی داغیدن تعدادی از پارامترها

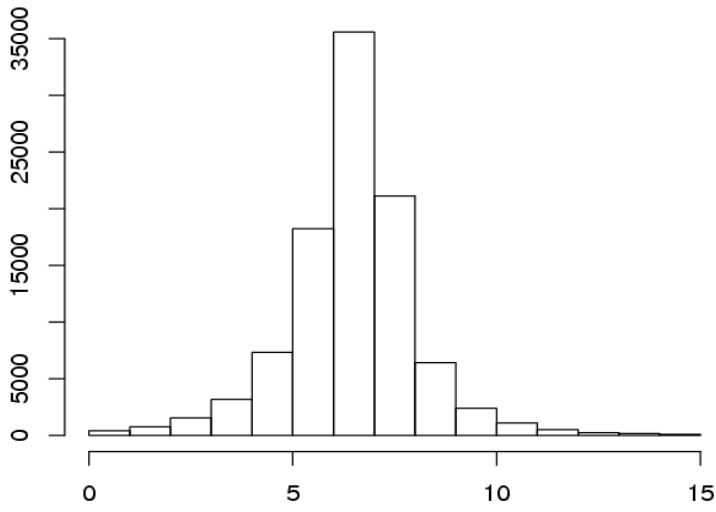
۳۹۵۱۲۱۹، پیشگویی بیزی میزان آلودگی هوا و انحراف استاندارد این پیشگویی به ترتیب ۶/۴۳ و ۱/۵۸ خواهد بود. در شکل ۳ نمودار هیستوگرام برای توزیع پیشگویی این موقعیت به نمایش گذاشته شده است.

۶ بحث و نتیجه‌گیری

در این مقاله، مدل فضایی گاوسی–لگ گاوسی که به‌عنوان تعمیمی از مدل گاوسی به‌شمار رفته و بر مبنای توزیع‌های مقیاس آمیخته از نرمال تعریف می‌شود، در مثال‌های کاربردی و شبیه‌سازی مورد بررسی قرار گرفت و نتایج ذیل مشاهده گردید:

۱. در یک مدل گاوسی–لگ گاوسی شناسایی داده‌های دورافتاده با استفاده از ناحیه‌ی HPD به‌نحوی مطلوب امکان‌پذیر است. ضمن آن‌که زمان محاسبه‌ی ناحیه‌ی HPD به‌نحوی چشمگیر و قابل توجه کوتاه‌تر بوده و به‌سادگی قابل تعیین است.

۲. بر اساس معیارهای MSPE ارایه‌شده در جدول ۴ و اعتبار سنجی متقابل در مثال کاربردی (برای مدل GLG و گاوسی به‌ترتیب برابر ۲/۶ و ۵/۰۱ می‌باشد)، می‌توان عملکرد مناسب پیشگویی



شکل ۳. نمودار هیستوگرام توزیع پیشگوی بیزی در موقعیت میدان انقلاب

مدل GLG در مقایسه با مدل گاوسی را به ویژه برای پیشگویی داده‌های آلودگی هوای شهر تهران مشاهده کرد. این موضوع می‌تواند ناشی از اثرات نامطلوب داده‌های دورافتاده بر نتایج مدل گاوسی و استواری مدل GLG نسبت به حضور این گونه داده‌ها باشد.

موضوعات زیر نیز به‌عنوان مسائل جدید قابل بررسی هستند:

۱. فرض اساسی در سراسر مقاله همسانگردی مدل تابع کوواریانس بود. در صورتی‌که این فرض برقرار نباشد لازم است مدل گاوسی-لگ گاوسی به‌نحوی مناسب‌تر تعمیم یابد.

۲. الگوریتم متروپلیس-هستینگز به‌دلیل همگرایی در زمان‌های طولانی با مشکل محاسباتی روبه‌رو است. لذا ارایه‌ی الگوریتم مناسبی در این خصوص از جمله به کارگیری الگوریتم بازنمونه‌گیری از نقاط نمونه‌گیری‌شده لازم به بررسی می‌باشد.

۳. بررسی حساسیت نتایج تحلیل به مقدار ابرپارامترهای اختیار شده و در صورت حساس بودن ارایه‌ی راه حلی مناسب در این خصوص مورد توجه قرار دارد.

توضیحات

۱. Highest posterior density

۲. Modified bessel function

۳. Gaussian-Log Gaussian

۴. Jefferi's Prior

۵. Vague Prior

۶. Generalized Inverse Gaussian

۷. Truncated Normal Distribution

مرجع‌ها

Barndorff-Nielsen, O., Kent, J. and Srensen, M. (1982). Normal variance-mean mixtures and Z-distributions, *International Statistical Review*, **50**, 145-159.

Berger, J.O., De Oliveira, V. and Sanso, B. (2001). Objective Bayesian analysis of spatially correlated data, *Journal of the American Statistical Association*, **93**, 1361-1374.

Bibby, B.M. and Sorensen, M. (2003). *Hyperbolic Processes in Finance*, in Handbook of Heavy-Tailed Distributions in Finance, ed. S. T. Rachev, New York: Elsevier, 211-248.

Cerioli, A. and Riani, M. (1999). The ordering of spatial data and the detection of multiple outliers, *Journal of Computational and Graphical Statistics*, **8**, 239-258.

Chen, M. and Shao, Q. (1998). Monte Carlo estimation of Bayesian credible and HPD intervals, *Journal of Computational and Graphical Statistics*, **7**, 69-92.

Fernandez, C. and Steel, M.F.J. (2000). Bayesian regression analysis with scale mixtures of normals, *Econometric Theory*, **16**, 80-101.

Jeffreys, H. (1961). *Theory of Probability*, Oxford University Press, London.

Khaledi, M.J. and Rivaz, F. (2009). Empirical Bayes spatial prediction using a Monte Carlo EM algorithm, *Statistical Methods and Applications*, **18**, 35-47

Lange, K.L., Little, R.J.A. and Taylor J.M.G. (1989). Robust statistical modeling using the T-distribution, *Journal of the American Statistical Association*, **84**, 881-896.

Militino, A.F., Palacios, M.B. and Ugarte, M.D. (2006). Outliers detection in multivariate spatial linear models, *Journal of Statistical Planning and Inference*, **136**, 125-146.

Maronna, R. (1976). Robust M-estimators of multivariate location and scatter, *Annals of Statistics*, **4**, 51- 67.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap, *Journal of the Royal Statistical Society*, ser. B, **56**, 3-48.

Palacios, M.B. and Steel M.F.J. (2006). Non-Gaussian Bayesian geostatistical modeling, *Journal of the American Statistical Association*, **101**, 604-618.

Rivaz, F., Mohammadzadeh, M. and Khaledi, M.J. (2007). Empirical Bayes prediction for space-Time data under separable models, *Journal of statistical Research*, **1**, 45-61

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging*, Springer-Verlag, New York.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression, *Journal of the Royal Statistical Society*, Series B, **46**, 431-439.

حمیدرضا زارعی فرد
گروه آمار، دانشکده‌ی علوم ریاضی،
دانشگاه تربیت مدرس،
تهران، ایران.
رایانشانی: *zareifard@modares.ac.ir*

مجید جعفری خالدی
گروه آمار، دانشکده‌ی علوم ریاضی،
دانشگاه تربیت مدرس،
تهران، ایران.
رایانشانی: *jafari-m@modares.ac.ir*