



تعیین طرح نمونه‌گیری بهینه برای تحلیل داده‌های فضایی

مجید جعفری خالدی* و فیروزه ریواز

دانشگاه تربیت مدرس

چکیده. یکی از موضوعات مهم در تحلیل داده‌های فضایی، نظیر پیشگویی فضایی یا برآورد پارامترهای مدل فضایی در نظر گرفته شده، انتخاب n موقعیت $\{t_1, \dots, t_n\}$ از فضای موقعیت‌های $D \subseteq \mathcal{R}^d$ است. برای این منظور هنگامی که D ناحیه‌ای پیوسته از \mathcal{R}^d است، با شبکه کردن آن به وسیله یک شبکه N عضوی، طرح نمونه‌گیری به‌گونه‌ای تعیین می‌شود که نتایج حاصل از تحلیل‌های فضایی مبتنی بر مشاهدات به دست آمده در این موقعیت‌ها، بهینه شود. این طرح، طرح نمونه‌گیری فضایی بهینه نامیده می‌شود. برای حالتی که تعداد انتخاب‌های ممکن n محل از N موقعیت موجود زیاد باشد، تعیین طرح نمونه‌گیری فضایی بهینه دشوار است. این مقاله الگوریتم‌های مناسبی برای تقریب آن ارائه می‌دهد.

واژگان کلیدی. داده‌های فضایی؛ طرح نمونه‌گیری فضایی؛ بهینگی.

۱ مقدمه

داده‌هایی که در فضای مورد مطالعه همبسته‌اند و این همبستگی تابعی از محل و موقعیت قرارگیری آن‌هاست، داده‌های فضایی نامیده می‌شوند. معمولاً برای مدل‌بندی داده‌های فضایی از یک میدان تصادفی

* نویسنده‌ی عهده‌دار مکاتبات.

$\{Z(t) : t \in D\}$ استفاده می‌شود، که در آن D مجموعه‌ای اندیس‌گذار در فضای اقلیدسی \mathbb{R}^d ، $d \geq 1$ است. پیشگویی مقدار نامعلوم میدان تصادفی در یک موقعیت دلخواه، برآورد پارامترهای تابع میانگین میدان که پارامترهای رگرسیونی نامیده می‌شوند و برآورد پارامترهای تابع کوواریانس میدان از موضوعات اساسی در تحلیل داده‌های فضایی به شمار می‌روند. برای این منظور لازم است n موقعیت $\{t_1, \dots, t_n\}$ در فضای موقعیت‌های D انتخاب و آن‌گاه کمیت مورد نظر اندازه‌گیری شود. یکی از مسائل اساسی در این ارتباط، انتخاب موقعیت‌ها به نحوی است که نتایج حاصل از تحلیل داده‌های فضایی نظیر پیشگویی فضایی یا برآورد پارامترها مبتنی بر مشاهدات به دست آمده در این موقعیت‌ها بهینه شوند. لذا با هدف تحلیل بهینه، معیار بهینگی طرح، انتخاب و طرح نمونه‌گیری فضایی بهینه به‌گونه‌ای تعیین می‌شود که این معیار را در بین تمامی طرح‌های ممکن مینیمم کند. هنگامی که فضای موقعیت‌های D ، ناحیه‌ای پیوسته از \mathbb{R}^d است، تعیین طرح نمونه‌گیری فضایی بهینه بسیار دشوار است و معمولاً در این حالت D به‌وسیله‌ی شبکه‌ای شامل N موقعیت $\{t_1, \dots, t_N\}$ تقریب زده و طرح نمونه‌گیری بهینه از موقعیت‌های این شبکه انتخاب می‌شود. در صورتی که تعداد حالت‌های ممکن انتخاب n محل از N موقعیت موجود زیاد باشد، مجدداً تعیین طرح نمونه‌گیری فضایی بهینه دشوار است و تقریبی از آن به‌عنوان طرح نمونه‌گیری بهینه انتخاب می‌شود. هنگامی که داده‌ها مستقل هستند، الگوریتم‌های مختلفی به‌منظور تعیین طرح نمونه‌گیری بهینه‌ی تقریبی توسط فدروف (۱۹۷۲)، بریمکولوف و همکاران (۱۹۸۶) و فدروف و مولر (۱۹۸۸) ارائه شده‌اند. اما در صورت همبسته بودن داده‌ها استفاده از این نوع طرح‌های نمونه‌گیری نامناسب است. در حالتی که داده‌ها از یک الگوی رگرسیونی با خطاهای همبسته پیروی کنند، با هدف برآورد پارامترهای رگرسیونی، آنجلیس و همکاران (۲۰۰۱) و مولر و پازمن (۲۰۰۳) الگوریتمی مناسب برای تقریب طرح نمونه‌گیری بهینه ارائه کرده‌اند.

در این مقاله با تصحیح این الگوریتم برای حالتی که همبستگی داده‌ها از نوع فضایی است، الگوریتمی به‌منظور تعیین طرح نمونه‌گیری فضایی بهینه‌ی تقریبی برای برآورد پارامترهای رگرسیونی ارائه می‌شود. سپس با هدف پیشگویی فضایی، الگوریتم دیگری ارائه می‌شود. در بخش ۲، برای یک میدان تصادفی که دارای ساختار میانگین و کوواریانس فضایی پارامتری است، برآورد پارامترها و پیشگویی فضایی تعیین شده است. در بخش ۳، با هدف برآورد بهینه‌ی پارامترهای رگرسیونی و پیشگویی فضایی بهینه، چگونگی تعیین یک طرح نمونه‌گیری فضایی بهینه و تقریب آن با استفاده از الگوریتم‌ها ارائه شده است. در بخش ۴ نحوه‌ی عملکرد الگوریتم‌ها بر اساس یک مثال عددی نشان داده شده است. بحث و موضوعات دیگر تحقیق در بخش ۵ ارائه شده است.

۲ تحلیل فضایی بر اساس یک طرح نمونه‌گیری معین

معمولاً میدان تصادفی $Z(\cdot) = \{Z(t) : t \in D\}$ را می‌توان به صورت

$$Z(t) = \mu(t) + \delta(t)$$

تجزیه کرد، که در آن $\mu(t) = f'(t)\beta$ تابع میانگین، $\beta = (\beta_1, \dots, \beta_p)$ بردار پارامترهای رگرسیونی، $f'(t) = (f_1(t), \dots, f_p(t))$ یک بردار p بعدی از مؤلفه‌های غیر تصادفی معلوم و $\delta(t)$ یک میدان تصادفی با میانگین صفر و کواریانس $\text{Cov}\{\delta(s), \delta(t)\} = c(s, t; \theta)$ است به طوری که این تابع کواریانس به پارامتر θ وابسته می‌باشد. با فرض آن که $Z = (Z(t_1), \dots, Z(t_n))'$ مقادیر میدان تصادفی در n موقعیت یک طرح نمونه‌گیری فضایی دلخواه $S_n = \{t_1, \dots, t_n\}$ را نشان دهد، خواهیم داشت:

$$E(Z) = X\beta, \quad \text{Var}(Z) = \sum_{\theta},$$

که در آن ماتریس $X_{n \times p}$ و ماتریس کواریانس فضایی \sum_{θ} به صورت

$$X = (f'(t_1), \dots, f'(t_n))', \quad \sum_{\theta} = (c(t_i, t_j; \theta))$$

تعریف می‌شوند. در این صورت برآورد کم‌ترین توان‌های دوم تعمیم‌یافته‌ی β و واریانس آن به ترتیب به صورت

$$\hat{\beta}_{gls} = \left(X' \sum_{\theta}^{-1} X \right)^{-1} X' \sum_{\theta}^{-1} Z$$

و

$$(۱) \quad \text{Var}(\hat{\beta}_{gls}) = \{M(S_n)\}^{-1}$$

هستند، که در آن $M(S_n) = X' \sum_{\theta}^{-1} X$. همچنین پیشگوی فضایی بهینه و واریانس پیشگوی در یک موقعیت دلخواه t_0 از میدان تصادفی $Z(\cdot)$ به ترتیب

$$\hat{Z}(t_0) = f'(t_0)\hat{\beta} + C'_{\theta} \sum_{\theta}^{-1} (Z - X\hat{\beta})$$

و

$$\sigma^2(t_0, S_n) = E\{Z(t_0) - \hat{Z}(t_0)\}^2$$

$$(۲) \quad = V^2(t_0, S_n) + \phi'(t_0, S_n)\{M(S_n)\}^{-1}\phi(t_0, S_n)$$

هستند (کرسی، ۱۹۹۳)، به طوری که

$$V^{\lambda}(t_0, S_n) = c(t_0, t_0; \theta) - C_{\theta}' \sum_{\theta}^{-1} C_{\theta}$$

$$\phi(t_0, S_n) = f(t_0) - X\{M(S_n)\}^{-1} C_{\theta}$$

و $C_{\theta}' = \text{Cov}\{Z(t_0), Z\}$ کواریانس بردار مشاهدات Z و $Z(t_0)$ است. همانگونه که از روابط (۱) و (۲) مشاهده می شود واریانس برآوردگر $\text{Var}(\hat{\beta}_{gls})$ و واریانس پیشگویی $\sigma^{\lambda}(t_0, S_n)$ تابعی از طرح نمونه گیری فضایی هستند و برای طرح های نمونه گیری مختلف، تغییر می کنند.

۳ طرح نمونه گیری فضایی بهینه

برای تعیین بهترین طرح نمونه گیری فضایی لازم است معیاری برای تشخیص بهینگی طرح انتخاب شود. معمولاً اندازه های از نزدیکی برآوردگر یا پیشگو به پارامتر واقعی یا داده، نظیر واریانس برآوردگر یا واریانس پیشگویی را می توان به عنوان معیاری برای سنجش بهینگی مورد استفاده قرار داد. ابتدا با هدف برآورد پارامتر β ، به دنبال تعیین طرح نمونه گیری هستیم که $\text{Var}(\hat{\beta}_{gls})$ را مینیمم کند. چون $\text{Var}(\hat{\beta}_{gls})$ یک ماتریس $p \times p$ است، مینیمم کردن تابعی از آن مانند $\Phi\{\text{Var}(\hat{\beta})\}$ به عنوان معیار بهینگی طرح انتخاب می شود. با فرض آن که تابع معیار $\Phi(\cdot)$ ، دترمینان ماتریس واریانس کواریانس $\hat{\beta}$ ؛ یعنی

$$\Phi\{\text{Var}(\hat{\beta})\} = |M(S_n)|^{-1}$$

اختیار شود، طرح نمونه گیری فضایی بهینه به صورت

$$S_n^* = \arg \min_{S_n \in D} |M(S_n)|^{-1}$$

خواهد شد. اگر D پیوسته باشد، تعیین S_n^* بسیار دشوار خواهد بود. لذا ناحیه D به وسیله ی مشبکه ای که شامل N گره است، بر اساس الگوی مشخصی مانند مشبکه های مربعی، مثلثی متساوی الاضلاع، شش وجهی و غیره تقریب زده می شود و مجموعه ی تمام گره های آن، فضای موقعیت جدید D_N را تشکیل می دهند. در این صورت، طرح نمونه گیری فضایی بهینه منوط به انتخاب یکی از $\binom{N}{n}$ طرح نمونه گیری ممکن است، به طوری که

$$S_n^* = \arg \min_{S_n \in D} |M(S_n)|^{-1}.$$

اما وقتی مقدار $\binom{N}{n}$ بزرگ باشد، به‌عنوان مثال اگر $N = 100$ و $n = 10$ اختیار شود، بررسی تمام طرح‌های نمونه‌گیری ممکن برای تعیین طرح نمونه‌گیری فضایی بهینه دشوار است و تقریبی از آن به‌عنوان طرح نمونه‌گیری بهینه انتخاب می‌شود. برای این منظور با لحاظ کردن ساختار همبستگی فضایی داده‌ها در الگوریتم‌های مکرر ارائه شده توسط آنجلیس و همکاران (۲۰۰۱) و مولر و پازمن (۲۰۰۳)، الگوریتمی برای تعیین طرح نمونه‌گیری فضایی بهینه تقریبی ارائه می‌شود.

الگوریتم ۱: طرح نمونه‌گیری فضایی بهینه با هدف برآورد β

۱. $z = 0$ قرار داده شود.

۲. یک طرح نمونه‌گیری اولیه $S_n^{(0)}$ دلخواه انتخاب شود.

۳. با فرض آن‌که برای $x \in D_N - S_n^{(j)}$ تابع $g(x) = \frac{|M(S_{n+1}^{(j)})|}{|M(S_n^{(j)})|}$ باشد، نقطه‌ی $x^+ = \arg \min_{x \in D_N - S_n^{(j)}} \{g(x)\}$ به طرح نمونه‌گیری $S_n^{(j)}$ اضافه شود و مجموعه‌ی $S_{n+1}^{(j)}$ تشکیل شود.

۴. نقطه‌ی $x^- = \arg \min_{x \in S_{n+1}^{(j)}} \{g(x)\}$ از مجموعه‌ی $S_{n+1}^{(j)}$ حذف و طرح نمونه‌گیری $S_n^{(j+1)}$ تشکیل شود.

۵. $z = z + 1$ قرار داده شود و به مرحله‌ی ۳ بازگشت شود.

الگوریتم تا همگرایی؛ یعنی انطباق طرح $S_n^{(j+1)}$ بر $S_n^{(j)}$ ، ادامه می‌یابد. برای سهولت محاسبات این الگوریتم، مولر و پازمن (۲۰۰۳) شکلی ساده برای تابع $g(x)$ تعیین کردند که به‌طور متناظر در تحلیل فضایی آن را می‌توان به‌صورت زیر نوشت.

$$g(x) = \frac{V^\dagger(x, S_n) + \phi'(x, S_n)(X' \Sigma_\theta^{-1} X)^{-1} \phi(x, S_n)}{V^\dagger(x, S_n)}$$

هنگامی که هدف تحلیل پیشگویی فضایی است، طرح نمونه‌گیری که واریانس پیشگویی $\sigma^2(x, S_n)$ را برای $x \in D$ ، مینیمم کند به‌عنوان طرح نمونه‌گیری بهینه انتخاب می‌شود. در این حالت تابعی از واریانس‌های پیشگویی مانند $\{\sigma^2(x, S_n) : x \in D\}$ به‌عنوان تابع معیار انتخاب می‌شود. اگر این تابع به یکی از دو صورت میانگین یا ماکسیمم واریانس پیشگویی روی ناحیه‌ی D ؛ یعنی

$$V(S_n) = \frac{1}{|D|} \int_D \sigma^2(x, S_n) dx \quad \text{یا} \quad H(S_n) = \min_{x \in D} \sigma^2(x, S_n)$$

در نظر گرفته شود، طرح نمونه‌گیری فضایی بهینه به‌صورت

$$S_n^* = \arg \min_{S_n \in D} V(S_n) \quad \text{یا} \quad S_n^* = \arg \min_{S_n \in D} H(S_n)$$

خواهد شد. لازم به ذکر است که هر یک از این طرح‌ها طبق معیار متناظرشان بهینه بوده و لزوماً بر اساس معیارهای دیگر بهینه نیستند. به دلیل پیوسته بودن D ، در این حالت نیز تعیین طرح نمونه‌گیری بهینه بسیار دشوار است. لذا آن‌را با یکی از الگوهای مثالی، مربعی یا شش وجهی، مشبکه‌ای کرده و فضای D_N تشکیل می‌شود. در این صورت معیارهای بهینگی طرح نمونه‌گیری میانگین و ماکسیمم برابر

$$(۳) \quad V(S_n) = \frac{1}{N} \sum_{x \in D_N} \sigma^2(x, S_n)$$

$$(۴) \quad H(S_n) = \min_{x \in D_N} \sigma^2(x, S_n)$$

خواهند شد. اگر مقدار $\binom{N}{n}$ بزرگ باشد، برای تعیین تقریبی S_n^* روی D_N با فرض آن‌که معیار بهینگی طرح میانگین واریانس‌های پیشگویی باشد، الگوریتم زیر را می‌توان مورد استفاده قرار داد.

الگوریتم ۲: طرح نمونه‌گیری فضایی بهینه با هدف پیشگویی

۱. $z = 0$ قرار داده می‌شود.
۲. طرح نمونه‌گیری اولیه $S_n^{(0)}$ به طور تصادفی انتخاب و $V(S_n^{(0)})$ از رابطه (۳) محاسبه شود.
۳. طرح نمونه‌گیری $S_n^{(j+1)}$ به صورت تصادفی یا سیستماتیک انتخاب شود.
۴. اگر $V(S_n^{(j+1)}) \leq V(S_n^{(j)})$ ، طرح نمونه‌گیری $S_n^{(j+1)}$ با $S_n^{(j)}$ جایگزین شود.
۵. $z = z + 1$ قرار داده شود و به مرحله سوم بازگشت شود.

الگوریتم تا رسیدن به همگرایی در $V(S_n)$ ادامه می‌یابد. در هر مرحله از این الگوریتم طرح نمونه‌گیری بهبود می‌یابد تا در نهایت طرحی تقریباً بهینه؛ یعنی، طرحی که $V(S_n)$ را به طور تقریبی حد اقل می‌کند، تعیین شود. هرگاه معیار بهینگی طرح ماکسیمم واریانس‌های پیشگویی اختیار شود، با جایگزینی $H(S_n)$ با $V(S_n)$ در الگوریتم بالا تقریب طرح نمونه‌گیری بهینه به روش مشابه انجام‌پذیر است. لازم به ذکر است طرح نمونه‌گیری به دست آمده به وسیله هر یک از دو الگوریتم ۱ و ۲، لزوماً طرح نمونه‌گیری بهینه‌ی مطلق نیست.

۴ مثال عددی

اکنون نحوه تعیین طرح نمونه‌گیری فضایی بهینه بر اساس یک مثال عددی نشان داده می‌شود. برای این منظور فرض کنید پیشگویی میدان تصادفی مانای مرتبه‌ی دوم $Z(\cdot)$ روی ناحیه‌ی $D = [0, 3] \times [0, 3]$

با میانگین ثابت $E\{Z(t)\} = \beta$ ؛ یعنی $p = 1$ و $f_1(t) = 1$ تابع کوواریانس همسانگرد نمایی

$$c(t, s; \theta) = \theta^{\|t-s\|}, \quad s, t \in D,$$

با $\theta = \frac{1}{4}$ مورد نظر باشد. برای تعیین طرح نمونه‌گیری فضایی بهینه با اندازه‌ی نمونه‌ی $n = 4$ فضای موقعیت‌های D توسط مشبکه‌ی مربعی با $N = 16$ گره به صورت

$$D_N = \{(0, 0), (0, 1), (0, 2), (0, 3), (1, 0), (1, 1), (1, 2), (1, 3), \\ (2, 0), (2, 1), (2, 2), (2, 3), (3, 0), (3, 1), (3, 2), (3, 3)\}$$

تقریب زده می‌شود. چون برای محاسبه‌ی معیارهای میانگین و ماکسیمم همه‌ی $(\frac{1}{4})$ طرح نمونه‌گیری، زمان زیادی مورد نیاز است، تعیین طرح نمونه‌گیری فضایی بهینه دشوار است و برای این منظور از الگوریتم ۲ استفاده می‌شود. در این صورت طرح‌های نمونه‌گیری بهینه بر اساس معیارهای میانگین و ماکسیمم به ترتیب برابر $S^*[\text{mean}] = \{(0, 1), (1, 2), (2, 2), (3, 3)\}$ و $S^*[\text{max}] = \{(1, 0), (0, 2), (2, 3), (3, 2)\}$ می‌شوند. همان طور که مشاهده می‌شود این دو طرح نمونه‌گیری از یکدیگر متفاوتند. موقعیت‌های $S^*[\text{max}]$ همگی روی مرز قرار دارند. این موضوع می‌تواند به این دلیل باشد که بیش‌ترین مقادیر واریانس پیشگویی در نقاط مرزی است و طرح نمونه‌گیری بر اساس معیار ماکسیمم به شکلی است که مقادیر بالای واریانس‌های کریکینگ کاهش یابند.

۵ بحث و موضوعات دیگر تحقیق

از آن‌جا که برای تحلیل فضایی بهینه نظیر پیشگویی فضایی و برآورد پارامترها نیاز به انتخاب یک طرح نمونه‌گیری فضایی بهینه است، در این مقاله نحوه‌ی تعیین این طرح با فرض معلوم بودن پارامتر تابع کوواریانس فضایی، مورد بررسی قرار گرفت. هنگامی که پارامتر کوواریانس نامعلوم است، می‌توان با استفاده از یک نمونه‌گیری مقدماتی آن را برآورد کرد و مقدار برآورد را به عنوان مقدار واقعی پارامتر در تعیین طرح نمونه‌گیری بهینه به کار برد. اما در بعضی از مسائل آمار فضایی امکان نمونه‌گیری مقدماتی وجود ندارد. به عنوان مثال برای تعیین مکان‌های مناسب نصب و راه‌اندازی ایستگاه‌های سنجش آلودگی هوای یک شهر، نمونه‌گیری مقدماتی امکان‌پذیر نیست. در این حالت تعیین طرح نمونه‌گیری بهینه از مسائل مهم در این خصوص به شمار می‌رود که توسط نویسندگان این مقاله در حال بررسی است.

مرجعها

- Angelis, L.; Bora-Senta, E.; Moyssiadis, C. (2001). Optimal exact experimental designs with autocorrelated errors through a simulated annealing algorithm, *Comput. Statist. Data. Anal.*, **37**, 275-296.
- Brimkulov, U.; Krug, G.; Savanov, V. (1986). *Design of Experiments for Random Fields and Processes*, Nauka, Moscow.
- Cressie, N. (1993). *Statistics for Spatial Data*, Wiley, New York.
- Fedorov, V.V. (1972). *Theory of Optimal Experiments*, Academic Press, New York.
- Fedorov, V.V.; Muller, W. (1988). *Two Approches in Optimization of Observing Networks*. In optimal design and analysis of experiments, Amsterdam/New York.
- Muller, W.G.; Pazman, A. (2003). Measures for designs in experiments with correlated errors, *Biometrika*, **90**, 423-434.

دریافت: ۲۵ دی ۱۳۸۴
 آخرین اصلاح: ۲۳ اردیبهشت ۱۳۸۵

<p>فیروزه ریواز گروه آمار، دانشکدهی علوم پایه، دانشگاه تربیت مدرس، پل نصر، بزرگراه جلال آل احمد، تهران، ایران. پیام‌نگار: rivaz@modares.ac.ir</p>	<p>مجید جعفری خالدي گروه آمار، دانشکدهی علوم پایه، دانشگاه تربیت مدرس، پل نصر، بزرگراه جلال آل احمد، تهران، ایران. پیام‌نگار: jafari-m@modares.ac.ir</p>
---	--