Statistical Research and Training Center
©(2005) J. Stat. Res. Iran **2** 47-54

مجلّه‌ی پژوهش‌های آماری ایران
دوره‌ی ۲، شماره‌ی ۱، بهار و تابستان ۱۳۸۴، صص ۵۳-۶۰

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 نظرگرفته شده، انتخاب n موقعیت $\{t_1, \ldots, t_n\}$ از فضای موقعیت‌های $D \subseteq \Re^d$ است. برای این منظور هنگامی که D ناحیه‌ای پیوسته از $\Re^d$ است، با مشبکه کردن آن به‌وسیله‌ی 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.

[DOI: 10.18869/acadpub.jsri.2.1.53]

$\{Z(t) : t \in D\}$ استفاده می‌شود، که در آن D مجموعه‌ای اندیس‌گذار در فضای اقلیدسی $\Re^d$، $d \geqslant ۱$، 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 $\{t_۱, \ldots, t_n\}$ در فضای موقعیت‌های D انتخاب و آنگاه کمیت مورد نظر اندازه‌گیری شود. یکی از مسائل اساسی در این ارتباط، انتخاب موقعیت‌ها به‌نحوی 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 $\Re^d$ است، تعیین طرح نمونه‌گیری فضایی بهینه بسیار دشوار است و معمولاً در این حالت D به‌وسیله‌ی مشبکه‌ای شامل N موقعیت $\{t_۱, \ldots, t_N\}$ تقریب زده و طرح نمونه‌گیری بهینه از موقعیت‌های این مشبکه 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.

[DOI: 10.18869/acadpub.jsri.2.1.53]

۲ تحلیل فضایی بر اساس یک طرح نمونه‌گیری معین

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

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

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

$$E(Z) = X\beta, \quad \mathrm{Var}(Z) = \sum\nolimits_{\theta},$$

که در آن ماتریس $X_{n\times p}$ و ماتریس کوواریانس فضایی $\sum_\theta$ به‌صورت

$$X = (f'(t_1), \ldots, f'(t_n))', \qquad \sum\nolimits_{\theta} = (c(t_i, t_j : \theta))$$

تعریف می‌شوند. در این صورت برآورد کم‌ترین توان‌های دوم تعمیم‌یافته‌ی β و واریانس آن به‌ترتیب به‌صورت

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

و

$$\mathrm{Var}(\hat{\beta}_{gls}) = \{M(S_n)\}^{-1} \qquad (1)$$

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

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

و

$$\begin{aligned} \sigma^2(t_0, S_n) &= E\{Z(t_0) - \hat{Z}(t_0)\}^2 \\ &= V^2(t_0, S_n) + \phi'(t_0, S_n)\{M(S_n)\}^{-1}\phi(t_0, S_n) \end{aligned} \qquad (2)$$

[DOI: 10.18869/acadpub.jsri.2.1.53]

هستند (کرِسی، ۱۹۹۳)، به‌طوری که

$$V^{۲}(t_{۰}, S_n) = c(t_{۰}, t_{۰}; \theta) - C'_{\theta} \sum\nolimits_{\theta}^{-۱} C_{\theta}$$
$$\phi(t_{۰}, S_n) = f(t_{۰}) - X\{M(S_n)\}^{-۱} C_{\theta}$$

و $C'_{\theta} = \mathrm{Cov}\{Z(t_{۰}), Z\}$ کوواریانس بردار مشاهدات Z و $Z(t_{۰})$ است. همان‌گونه که از روابط (۱) و (۲) مشاهده می‌شود واریانس برآوردگر $\mathrm{Var}(\hat{\beta}_{gls})$ و واریانس پیشگویی $\sigma^{۲}(t_{۰}, S_n)$ توابعی از طرح نمونه‌گیری فضایی هستند و برای طرح‌های نمونه‌گیری مختلف، تغییر می‌کنند.

۳ طرح نمونه‌گیری فضایی بهینه

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

$$\Phi\{\mathrm{Var}(\hat{\beta})\} = |M(S_n)|^{-۱}$$

اختیار شود، طرح نمونه‌گیری فضایی بهینه به‌صورت

$$S_n^{*} = \arg \min_{S_n \in D} |M(S_n)|^{-۱}$$

خواهد شد. اگر D پیوسته باشد، تعیین S_n^{*} بسیار دشوار خواهد بود. لذا ناحیه‌ی D به‌وسیله‌ی مشبکه‌ای که شامل N گره است، بر اساس الگوی مشخصی مانند مشبکه‌های مربعی، مثلثی متساوی الاضلاع، شش وجهی و غیره تقریب زده می‌شود و مجموعه‌ی تمام گره‌های آن، فضای موقعیت جدید D_N را تشکیل می‌دهند. در این صورت، طرح نمونه‌گیری فضایی بهینه منوط به انتخاب یکی از $\binom{N}{n}$ طرح نمونه‌گیری ممکن است، به‌طوری که

$$S_n^{*} = \arg \min_{S_n \in D} |M(S_n)|^{-۱}.$$

[DOI: 10.18869/acadpub.jsri.2.1.53]

اما وقتی مقدار $\binom{N}{n}$ بزرگ باشد، به‌عنوان مثال اگر $N = ۱۰۰$ و $n = ۱۰$ 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. $j = ۰$ قرار داده شود.

۲. یک طرح نمونه‌گیری اولیه‌ی دلخواه $S_n^{(۰)}$ انتخاب شود.

۳. با فرض آنکه برای $x \in D_N - S_n^{(j)}$، تابع $g(x) = \frac{|M(S_{n+۱}^{(j)}|}{|M(S_n^{(j)})|}$ باشد، نقطه‌ی $x^+ = \arg \min_{x \in D_N - S_n^{(j)}} \{g(x)\}$ به طرح نمونه‌گیری $S_n^{(j)}$ اضافه شود و مجموعه‌ی $S_{n+۱}^{(j)}$ تشکیل شود.

۴. نقطه‌ی $x^- = \arg \min_{x \in S_{n+۱}^{(j)}} \{g(x)\}$ از مجموعه‌ی $S_{n+۱}^{(j)}$ حذف و طرح نمونه‌گیری $S_n^{(j+۱)}$ تشکیل شود.

۵. $j = j + ۱$ قرار داده شود و به مرحله‌ی ۳ بازگشت شود.

الگوریتم تا همگرایی؛ یعنی انطباق طرح $S_n^{(j+۱)}$ 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.

$$g(x) = \frac{V^۲(x, S_n) + \phi'(x, S_n)(X' {\sum_\theta}^{-۱} X)^{-۱} \phi(x, S_n)}{V^۲(x, S_n)}.$$

هنگامی که هدف تحلیل پیشگویی فضایی است، طرح نمونه‌گیری که واریانس پیشگویی $\sigma^۲(x, S_n)$ را برای $x \in D$، مینیمم کند به‌عنوان طرح نمونه‌گیری بهینه انتخاب می‌شود. در این حالت تابعی از واریانس‌های پیشگویی مانند $\phi\{\sigma^۲(x, S_n) : x \in D\}$ به‌عنوان تابع معیار انتخاب می‌شود. اگر این تابع به یکی از دو صورت میانگین یا ماکسیمم واریانس پیشگویی روی ناحیه‌ی D؛ یعنی

$$V(S_n) = \frac{۱}{|D|} \int_D \sigma^۲(x, S_n)\, dx \qquad \text{یا} \qquad H(S_n) = \min_{x \in D} \sigma^۲(x, S_n)$$

در نظر گرفته شود، طرح نمونه‌گیری فضایی بهینه به‌صورت

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

[DOI: 10.18869/acadpub.jsri.2.1.53]

خواهد شد. لازم به‌ذکر 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 مثلثی، مربعی یا شش وجهی، مشبکه‌ای کرده و فضای D_N تشکیل می‌شود. در این صورت معیارهای بهینگی طرح نمونه‌گیری میانگین و ماکسیمم برابر

$$V(S_n) = \frac{1}{N} \sum_{x \in D_N} \sigma^2(x, S_n) \qquad (3)$$

$$H(S_n) = \min_{x \in D_N} \sigma^2(x, S_n) \qquad (4)$$

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. $j = 0$ قرار داده می‌شود.
۲. طرح نمونه‌گیری اولیه $S_n^{(0)}$ به‌طور تصادفی انتخاب و $V(S_n^{(0)})$ از رابطه‌ی (۳) محاسبه شود.
۳. طرح نمونه‌گیری $S_n^{(j+1)}$ به‌صورت تصادفی یا سیستماتیک انتخاب شود.
۴. اگر $V(S_n^{(j+1)}) \leqslant V(S_n^{(j)})$، طرح نمونه‌گیری $S_n^{(j+1)}$ با $S_n^{(j)}$ جایگزین شود.
۵. $j = j + 1$ قرار داده شود و به مرحله‌ی 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 به‌طور 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 به‌وسیله‌ی هر یک از دو الگوریتم ۱ و ۲، لزوماً طرح نمونه‌گیری بهینه‌ی مطلق نیست.

۴ مثال عددی

اکنون نحوه‌ی 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 $D = [0, 3] \times [0, 3]$

[DOI: 10.18869/acadpub.jsri.2.1.53]

با میانگین ثابت $E\{Z(t)\} = \beta$؛ یعنی $p = ۱$ و $f_۱(t) = ۱$ و تابع کوواریانس همسانگرد نمایی

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

با $\theta = \frac{۱}{۲}$ مورد نظر باشد. برای تعیین طرح نمونه‌گیری فضایی بهینه با اندازه‌ی نمونه‌ی $n = ۴$، فضای موقعیت‌های D توسط مشبکه‌ی مربعی با $N = ۱۶$ گره به‌صورت

$$D_N = \{(۰,۰),(۰,۱),(۰,۲),(۰,۳),(۱,۰),(۱,۱),(۱,۲),(۱,۳),$$
$$(۲,۰),(۲,۱),(۲,۲),(۲,۳),(۳,۰),(۳,۱),(۳,۲),(۳,۳)\}$$

تقریب زده می‌شود. چون برای محاسبه‌ی معیارهای میانگین و ماکسیمم همه‌ی $\binom{۱۶}{۴} = ۱۸۲۰$ 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 به‌ترتیب برابر $S_۴^*[\text{mean}] = \{(۰,۱),(۱,۲),(۲,۲),(۳,۳)\}$ و $S_۴^*[\text{max}] = \{(۱,۰),(۰,۲),(۲,۳),(۳,۲)\}$ می‌شوند. همان طور که مشاهده می‌شود این دو طرح نمونه‌گیری از یکدیگر متفاوتند. موقعیت‌های $S_۴^*[\text{max}]$ 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 به‌عنوان مقدار واقعی پارامتر در تعیین طرح نمونه‌گیری بهینه به کار برد. اما در بعضی از مسائل آمار فضایی امکان نمونه‌گیری مقدماتی وجود ندارد. به‌عنوان 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.

[DOI: 10.18869/acadpub.jsri.2.1.53]

مرجع‌ها

Angelis, L.; Bora-Senta, E.; Moyssiadis, C. (2001). Optimal exact experimental designs with autocorrelated errors throught 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

مجید جعفری خالدی
گروه آمار،
دانشکده‌ی علوم پایه،
دانشگاه تربیت مدرس،
پل نصر، بزرگراه جلال آل احمد،
تهران، ایران.
پیام‌نگار: *jafari-m@modares.ac.ir*

فیروزه ریواز
گروه آمار،
دانشکده‌ی علوم پایه،
دانشگاه تربیت مدرس،
پل نصر، بزرگراه جلال آل احمد،
تهران، ایران.
پیام‌نگار: *rivaz@modares.ac.ir*

[DOI: 10.18869/acadpub.jsri.2.1.53]