چهارمین سمینار آمار فضایی و کاربردهای آن
4th Seminar on Spatial Statistics and Its Applications
۵–۶ آبان ماه ۱۴۰۰
27-28 Oct. 2021

تقریب لاپلاس برای برآورد مدل‌های آمیخته خطی تعمیم‌یافته فضایی، مطالعه موردی: داده‌های هزینه خانوار مناطق شهری

لیلا صالحی، محسن محمدزاده

گروه آمار، دانشگاه تربیت مدرس

چکیده: برآورد پارامترها در مدل‌های آمیخته خطی تعمیم‌یافته فضایی، با انتگرال‌های بعد بالا همراه است که استفاده از روش‌های تحلیلی برای حل آن‌ها غیر ممکن است. به‌علاوه استفاده از روش‌های عددی و مبتنی بر شبیه‌سازی، در بیشتر مواقع با محاسبات زمان‌بر و سنگین همراه است. یک پیشنهاد برای این موارد استفاده از روش‌های تقریبی است. در این مقاله با استفاده از تقریب لاپلاس، مدل آمیخته خطی تعمیم‌یافته فضایی دوجمله‌ای برای تحلیل داده‌های درآمد خانوار مناطق شهری به‌کار برده می‌شود. سپس کارایی این روش نتایج حاصل از روش ماکسیمم‌سازی امیدریاضی مونت کارلویی مورد ارزیابی و مقایسه قرار می‌گیرد.

واژه‌های کلیدی: تقریب لاپلاس، مدل‌های SGLM، ماکسیمم درستنمایی، MCEM.

کد موضوع‌بندی ریاضی (۲۰۱۰): 62H11، 62M30، 62F15.

۱ مقدمه

به دست آوردن تابع درستنمایی برای استنباط مدل‌های آمیخته خطی تعمیم‌یافته فضایی (SGLM)، نیازمند حل یک انتگرال بعد بالا نسبت به متغیرهای تصادفی فضایی است (دیگل و همکاران ، ۱۹۸۸). بعد این انتگرال‌ها برابر تعداد مشاهدات است. بنابراین حل آن‌ها یک مسئله بسیار مهم در برازش چنین مدل‌هایی است. مدل‌های SGLM در موارد زیادی با روش‌های زنجیر مارکفی مونت کارلو (MCMC) تحت استنباط بیزی برازش داده شده‌اند (مولر ، ۲۰۰۳؛ دیگل و ریبیرو، ۲۰۰۷). در همین راستا، ایدسویک و همکاران (۲۰۱۲) استنباط بیزی تقریبی را برای مدل‌های SGLM مطرح کردند. روش دیگری بر اساس الگوریتم همسان‌سازی داده‌ها نیز توسط باغیشنی و محمدزاده (۲۰۱۱) و ترابی (۲۰۱۵) ارائه شده

* نام وآدرس پست الکترونیکی ارائه‌دهنده مقاله: لیلا صالحی، l.salehi@modares.ac.ir

است. اخیرا محققان روش‌های تقریبی در قالب رهیافت بیزی و بسامدی را مورد توجه قرار داده‌اند که نسبت به روش‌های مبتنی بر الگوریتم‌های شبیه‌سازی، بار محاسباتی کمتر و سرعت بیشتری دارند. این الگوریتم ها، استنباط درستنمایی را برای مدل‌های SGLM فراهم می‌کنند اما همانند روش‌های MCMC به دلیل مبنای شبیه‌سازی آن‌ها، با محاسبات سنگین همراه هستند. به‌علاوه روش‌های مبتنی بر شبیه‌سازی، مقدار ماکسیمم درستنمایی را به دست نمی‌دهند. هگرتی و له له (۱۹۹۸) و وارین (۲۰۰۵) استفاده از تابع درستنمایی زوجی را برای برآورد مدل‌های SGLM پیشنهاد دادند. این روش از نظر محاسباتی آسان است زیرا تنها به حل یک انتگرال دوگانه نیاز دارد. مسئله اساسی در این روش، انتخاب زوج مرتب‌های مشاهدات است. بریسلو و لین (۱۹۹۵) روش شبه‌درستنمایی تاوانیده را برای مشاهدات دسته بندی شده دودویی انجام دادند. اوانگلو و همکاران (۲۰۱۱) نشان دادند تقریب لاپلاس اصلاح شده برای برآورد و پیش‌گویی در مدل‌های SGLM نتایج بهتری نسبت به روش‌های شبه‌درستنمایی و MCMC دارد. مولنبرگ و وربک (۲۰۰۵) تقریب لاپلاس را برای داده‌های طولی به کار بردند. بونات (۲۰۱۶) الگوریتمی‌برای برآورد ماکسیمم درستنمایی مدل‌های SGLM بر اساس تقریب لاپلاس را پیشنهاد دادند. در این مقاله بر اساس روش بونات (۲۰۱۶)، مدل SGLM را به داده‌های هزینه درآمد برازش داده و پارامترهای مدل را بر اساس الگوریتم پیشنهادی برآورد می‌کنیم. سپس برای این داده‌ها روش ماکسیمم سازی امید ریاضی مونت کارلویی[1] (MCEM) را نیز به کار برده و نتایج را باهم مقایسه می‌کنیم.

۲ تقریب لاپلاس برای SGLM

فرض کنید $Y(s)$ متغیر پاسخ فضایی گسسته و همچنین $Z_1(s),\ldots,Z_p(s)$ متغیرهای تبیینی فضایی قابل مشاهده و $\{X(s), s\in\mathbb{R}^2\}$ یک میدان تصادفی فضایی پنهان باشد، طوری که $X(s)$ نشان‌دهنده اثر تصادفی در موقعیت s است. دیگل و همکاران (۱۹۹۸) مدل‌های SGLM را به صورت زیر تعریف کردند.

الف- $\{S(x), x\in\mathbb{R}^2\}$ یک میدان تصادفی مانای گاوسی با میانگین صفر و با تابع کواریانس $\mathrm{cov}(S(x),S(x'))=\sigma^2\rho(x-x',\phi)$ است، که در آن $\rho(.,\phi)$ یک تابع معین مثبت و ϕ پارامتر همبستگی است.

ب- $\{Y(s), s\in\mathbb{R}^2\}$ به شرط $\{X(s), s\in\mathbb{R}^2\}$، مجموعه‌ای از متغیرهای تصادفی مستقل است و توزیع هر $Y(s)$ با میانگین شرطی $E[Y(s)|X(s)]$ مشخص می‌شود.

ج- برای هر تابع پیوند g داریم $g\{E[Y(s)|X(s)]\}=\sum_{j=1}^{p}Z_j(s)\beta_j+X(s)$، که در آن $\beta_1,\ldots,\beta_p$ ضرایب رگرسیونی هستند.

د- توزیع شرطی $[Y(s)|X(s)]$ عضو خانواده نمایی است.

برای برآورد ماکسیمم درستنمایی پارامترهای $\boldsymbol{\theta}=(\boldsymbol{\beta},\sigma^2,\phi)$ نیاز به انتگرال‌گیری از تابع توزیع توأم نسبت به متغیرهای تصادفی $\mathbf{S}(x)$ به صورت

$$L(\boldsymbol{\theta},\mathbf{y}(\mathbf{x}))=\int_{\Re^n} f(\mathbf{y}(\mathbf{x})|\mathbf{S}(\mathbf{x}))d\mathbf{S}(\mathbf{x}) \qquad (۱.۲)$$

چون بعد انتگرال (۱.۲) برابر تعداد مشاهدات است، حل آن از نظر محاسباتی مشکل است. روش‌هایی مانند مربع‌سازی عددی، گاوس-هرمیت[2] یا گاوس-هرمیت اصلاح شده مشکل هستند. انتگرال مونت کارلویی هم کند بوده و همگرایی آن به سختی صورت می‌پذیرد.

[1] Monte Carlo Expectation Maximaization
[2] Gauss-Hermite

یک روش کاربردی پیشنهادی، تقریب لاپلاس است (تایرنی و کادان ، ۱۹۸۶) که برای تقریب انتگرال‌هایی به‌صورت

$$\int_{\Re^n} \exp\mathbf{Q}(\mathbf{u})d\mathbf{u} \approx (۲\pi)^{n/۲}|-\mathbf{Q}''(\hat{\mathbf{u}})|^{-۱/۲}\exp\mathbf{Q}(\hat{\mathbf{u}}) \qquad (۲.۲)$$

طراحی شده است، که در آن $\mathbf{Q}(\mathbf{u})$ یک تابع تک‌ـمدی، کراندار و معین از متغیر n بعدی $\mathbf{u}$ و $\hat{\mathbf{u}}$ مقداری است که در آن $\mathbf{Q}(\mathbf{u})$ ماکسیمم شده است. اکنون برای تقریب لاپلاس مدل‌های SGLM، فرض کنید توزیع $f(\mathbf{y}(\mathbf{x})|\mathbf{S}(\mathbf{x})$ را بتوان به صورت یک خانواده تک پارامتری نمایی به‌صورت

$$f(\mathbf{y}(\mathbf{x})|\mathbf{S}(\mathbf{x});\boldsymbol{\beta}) = \exp\left\{\mathbf{y}(\mathbf{x})^T(\mathbf{D}\boldsymbol{\beta}+\mathbf{S}(\mathbf{x})) - \mathbf{1}^T b(\mathbf{D}\boldsymbol{\beta}+\mathbf{Q}(\mathbf{x})) + ۱^T c(\mathbf{y}(\mathbf{x}))\right\} \qquad (۳.۲)$$

نوشت، که در آن $b(\cdot)$ و $c(\cdot)$ توابعی معین هستند. تابع توزیع گاوسی چندمتغیره به‌صورت

$$f(\mathbf{S}(\mathbf{x});\Sigma) = (۲\pi)^{-n/۲}|\Sigma|^{۱/۲}\exp\left\{-\frac{۱}{۲}\mathbf{S}(\mathbf{x})^T\Sigma^{-۱}\mathbf{S}(\mathbf{x})\right\} \qquad (۴.۲)$$

است. تابع زیر انتگرال در (۱.۲) برابر حاصلضرب دو تابع (۳.۲) و (۴.۲) است. تابع درستنمایی را می‌توان به صورت

$$L(\boldsymbol{\theta};\mathbf{y}(\mathbf{x})) = \int_{\Re^n} \exp\mathbf{Q}(\mathbf{S}(\mathbf{x}))d\mathbf{S}(\mathbf{x})$$

نوشت که در آن

$$\begin{aligned}\mathbf{Q}(\mathbf{S}(\mathbf{x})) &= \mathbf{y}(\mathbf{x})^T(\mathbf{D}\boldsymbol{\beta}+\mathbf{S}(\mathbf{x})) - \mathbf{1}^T b(\mathbf{D}\boldsymbol{\beta}+\mathbf{Q}(\mathbf{x})) + \mathbf{1}^T c(\mathbf{y}(\mathbf{x})) \\ &- \frac{n}{۲}\log(۲\pi) - \frac{۱}{۲}\log|\boldsymbol{\Sigma}| - \frac{۱}{۲}\mathbf{Q}(\mathbf{x})^T\boldsymbol{\Sigma}^{-۱}\mathbf{S}(\mathbf{x}) \qquad (۵.۲)\end{aligned}$$

تقریب (۲.۲) نیازمند مقدار ماکسیمم $\hat{\mathbf{s}}$ برای تابع $\mathbf{Q}(\mathbf{S}(\mathbf{x}))$ است. یک روش برای به دست آوردن این مقدار استفاده از الگوریتم نیوتن‌ـرافسون[۳] است. ساختار این الگوریتم به صورت $\mathbf{s}_{i+۱} = \mathbf{s}_i - \mathbf{Q}''(\mathbf{s}_i)^{-۱}\mathbf{Q}'(\mathbf{s}_i)$ است، که پس از همگرایی مقدار $\hat{\mathbf{s}}$ به‌دست می‌آید و در این رابطه

$$\begin{aligned}\mathbf{Q}'(\mathbf{s}) &= \mathbf{y}(\mathbf{x}) - b'(\mathbf{D}\boldsymbol{\beta}+\mathbf{s})^T - \mathbf{s}^T\Sigma^{-۱} \\ \mathbf{Q}''(\mathbf{s}) &= -\mathrm{diag}\{b''(\mathbf{D}\boldsymbol{\beta}+\mathbf{s})\} - \Sigma^{-۱}\end{aligned}$$

و در نهایت تقریب لاپلاس برای SGLM به‌صورت

$$\begin{aligned}\ell(\boldsymbol{\theta};\mathbf{y}(\mathbf{x})) &= \frac{n}{۲}\log(۲\pi) - \frac{۱}{۲}\log\left|\mathrm{diag}\{b''(\mathbf{D}\boldsymbol{\beta}) + \hat{\mathbf{s}}(\boldsymbol{\theta})\} + \boldsymbol{\Sigma}^{-۱}\right| + \mathbf{y}(\mathbf{x})^T(\mathbf{D}\boldsymbol{\beta}+\hat{\mathbf{s}}(\boldsymbol{\theta})) \\ &- \mathbf{1}^T b(\mathbf{D}\boldsymbol{\beta}+\hat{\mathbf{s}}(\boldsymbol{\theta})) + \mathbf{1}^T c(\mathbf{y}(\mathbf{x})) - \frac{n}{۲}\log(۲\pi) - \frac{۱}{۲}\log|\boldsymbol{\Sigma}| - \frac{۱}{۲}\hat{\mathbf{s}}(\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-۱}\hat{\mathbf{s}}(\boldsymbol{\theta}) \qquad (۶.۲)\end{aligned}$$

است که می‌تواند به صورت عددی نسبت به پارامترها ماکسیمم شود. برای ماکسیمم کردن تابع از الگوریتم[۴] (BFGS) که یکی از روش‌های بهیه‌سازی در تابع optim() در نرم‌افزار R است، استفاده می‌شود. پارامترهای مدل به صورت $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^۲, \phi)$ در تابع وارد شده‌اند.

بونات (۲۰۱۶) برای همگرایی سریع تر الگوریتم، روشی را برای به دست آوردن مقادیر اولیه پیشنهاد دادند: ابتدا مدل خطی تعمیم‌یافته را به داده‌ها برازش داده و مقادیر اولیه برای $\boldsymbol{\beta}$ تعیین شود، بر اساس این مقادیر، مقدار $\hat{\boldsymbol{\mu}}$ محاسبه شود.

[3] Newthon-Raphson
[4] Broyden-Fletcher-Goldfarb-Shanno

سپس مقادیر باقیمانده بر اساس رابطه $\hat{\mathbf{r}} = (\mathbf{y} - \hat{\boldsymbol{\mu}})$ به دست آورده شوند. واریانس نمونه‌ای $\hat{\mathbf{r}}$ را می‌توان به عنوان مقدار اولیه σ^2 در نظر گرفت. در نهایت برای ϕ می‌توان ۱۰ درصد از بزرگترین فاصله بین دو نقطه مشاهدات را در نظر گرفت. الگوریتم نیوتن–رافسون نیز به مقادیر اولیه منطقی نیازمند است که برای آن می‌توان مقادیر امید ریاضی متغیرهای تصادفی را در نظر گرفت.

۱.۲ تحلیل داده‌ها

ابتدا برای بررسی درستی الگوریتم داده‌های شبیه سازی شده را تولید کرده و مقدار دقت برآورد پارامترها بر اساس معیار خطای استاندارد برآوردها به دست می‌آوریم. برای این کار ۱۰۰ داده از توزیع دوجمله‌ای با تابع پیوند لوژیت با مقادیر $\beta_0 = 2$، $\sigma^2 = 0/5$ و $\phi = 6$ را شبیه‌سازی کردیم. متغیر پنهان $\mathbf{X}$ را از توزیع گاوسی $N_{100}(\mathbf{100}; \boldsymbol{\Sigma}_\theta)$ در نظر گرفتیم، که در آن $\boldsymbol{\Sigma}_\theta = \sigma^2 \exp(-h/\phi)$ و h فاصله بین نقاط فضایی است. پارامترها را بر اساس تقرب لاپلاس برآورد کردیم که نتایج آن در جدول ۱ ارائه شده است. همانطور که ملاحظه می‌شود مقادیر خطای استاندارد برای هر سه پارامتر مقادیری قابل قبول هستند.

جدول ۱: نتایج برازش مدل به داده‌ها

پارامترها	مقدار واقعی	مقدار برآورد شده	خطای استاندارد
β_0	۲	۱/۹۸	۰/۰۶۰
σ^2	۰/۵	۰/۶۳	۰/۰۱۲
ϕ	۶	۵/۶۰	۰/۱۵۰

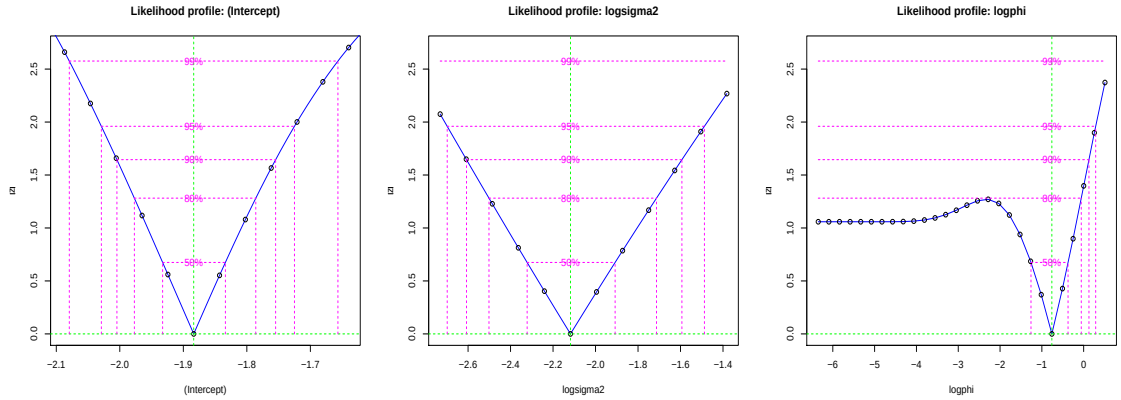
اکنون روش تقریب لاپلاس را برای داده‌های تعداد خانوارهای شهری بر حسب گروه‌های درآمد به تفکیک استان را مورد بررسی قرار می‌دهیم. این داده شامل ۴ ستون هستند که ستون اول و دوم طول و عرض جغرافیایی نقاط فضایی مربوط به مراکز استان ها، ستون سوم تعداد کل خانوارهای شهری در هر استان و ستون چهارم تعداد خانوارهای شهری با درآمد ۱۹۵۰۰۰۰ تا ۲۷۰۰۰۰۰ تومان است. می‌خواهیم همبستگی فضایی این داده‌های تعداد خانوارهای با درآمد مذکور بین تعداد کل خانوارهای یک استان را بررسی کرده و مدلی مناسب به آن‌ها برازش دهیم.

همان‌طور که در نمودار این داده‌ها در شکل ۱ ملاحظه می‌شود، استان‌های میانی کشور که در مجاورت هم هستند، از لحاظ درآمدی با هم شباهت دارند. به همین صورت استان‌های شرق و غرب کشور نیز وضعیتی مشابه هم دارند. در این داده‌ها تعداد خانوارهای با درآمد ۱۹۵۰۰۰۰ تا ۲۷۰۰۰۰۰ تومان را از بین کل خانوارهای هر استان در نظر گرفته و مدل دوجمله‌ای را برای آن‌ها در نظر گرفتیم. با توجه به همبستگی فضایی داده‌ها، مدل SGLM را با استفاده از تقریب لاپلاس به داده‌ها برازش داده و برآورد پارامترها را به دست آوردیم. همچنین براورد پارامترها را با استفاده از الگوریتم MCEM به‌دست آورده و با تقریب لاپلاس مقایسه کردیم. پارامترهای مدل شامل $\boldsymbol{\theta} = (\beta_0, \sigma^2, \phi)$ بودند که برای همگرایی بهتر الگوریتم‌ها، پارامترها را به صورت $\boldsymbol{\theta} = (\beta_0, \log\sigma^2, \log\phi)$ در مدل وارد کردیم. نتایج در جدول ۲ آمده است.

جدول ۲: نتایج برازش مدل به داده‌ها

پارامترها	Laplace	MCEM
β_0	−۱/۸۸	−۱/۷۱۸
σ^2	۰/۱۲	۰/۱۴
ϕ	۰/۴۶	۰/۴۷۰۱

شکل ۱: همبستگی فضایی داده‌های هزینه خانوار



شکل ۲: مقادیر تابع درستنمایی برای پارامترها در برازش مدل SGLM به داده‌های هزینه خانوار

در شکل ۲ مقادیر احتمال برای تابع درستنمایی مدل آمیخته خطی تعمیم‌یافته فضایی دوجمله‌ای نشان داده شده است. می‌توان از روی شکل مقادیر ماکسیمم شده تابع درستنمایی و همچنین بازه‌های اطمینان برای پارامترها را ملاحظه کرد. مقدار معیار Akaike برای مدل SGLM دوجمله‌ای بر اساس روش تقریب لاپلاس، ۸۰۰/۵۳ به دست آمده است.

بحث و نتیجه‌گیری

نتایج شبیه‌سازی در جدول ۱، نشان می‌دهد که از نظر معیار خطای استاندارد، روش تقریب لاپلاس نتایج خوبی را برای مدل‌های SGLM به دست می‌دهد و می‌توان این روش را برای داده‌های واقعی به کار برد. همچنین نتایج جدول ۲ نشان

می‌دهد که برای داده‌های درآمد خانوارهای شهری، دو روش تقریب لاپلاس و MCEM نتایج تقریبا مشابهی دارند.

مراجع

Baghishani, H. and Mohammadzadeh, M. (2011), Data Cloning Algorithm for Computing Maximum Likelihood Estimates in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **55**, 1748-1759.

Bonat, W. H., Ribeiro, Jr, P. J., (2016), Practical Likelihood Analysis for Spatial Generalized Linear Mixed Models, it Environmetrics, **77**(2): 83-89.

Breslow, N. E., Lin, X., (1995), Bias Correlation in Generalized Linear Mixed Models with a Single Component of Dispersion, *Biometrica*, **82**(1):81-91.

Cressie, N. (1993), *Statistics for Spatial Data*, John Wiley, New York.

Digle, P. J., Tawn, J. A. and Moyeed, RA., (1998), Model Based Geostatistics. *Journal of Royal Statistical Society*, Series B **70**(1):209226.

Digle, P. J. and Riberio, Jr P. J., (2007), *Model Based Geostatistics*, Springer: New York.

Eidsvik, J., Finley, A. O., Banerjee, S., and Rue, H. (2012), Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models, *Computational Statistics and Data Analysis*, **56**, 1362-1380.

Evangelou, E., Zhu, Z., Smith, R. L., (2011), Asymptotic Inference for Spatial GLMM Using High Order Laplace Approximation, *Jounal of Statistical Planning and Inference*, **141**(11):3564-3577.

Heagerti, P. J. and Lele, S. R., (1998), A Composite Likelihood Approch to Binary Spatial Data, *Journal of the American Statistical Association*, **93**(443):1099-1111.

Molenberghs, G. and Verbeke, G., (2005), *Models for Discrete Longitudinal Data*, Springer: New York.

Moller, J., (2003), *Spatial Statistics and Computational Methods*, Springer-Verlag: New York.

Tierney, L., Kadane, J., (1986), Accurate Approximation for Posterior Moments and Marginal Densities, *Journal of the American Statistical Association*, **81**(393):82-86.

Torabi, M. (2015), Likelihood Inference for Spatial Generalized Linear Mixed Models, *Communications in Statistics-Simulation and Computation*, **44**, 1692-1701.

Varin, C., Host, G., and Skare, O. (2005), Pairwise Likelihood Inference in Spatial Generalized Linear Mixed Models, *Computational Statistics and Data Analysis*, **49**, 1173-1191.