مجله علوم آماری، بهار و تابستان ۱۳۹۷
جلد ۱۲، شماره ۱، ص ۲۱ – ۳۷
DOI: 10.29252/jss.12.1.21

یک معیار جدید انتخاب مدل مبتنی بر داده تاگی

صدیقه اسحقی، حسین باغیشنی، نگار اقبال

گروه آمار، دانشگاه صنعتی شاهرود

تاریخ دریافت: ۱۳۹۶/۳/۲۸ تاریخ آخرین بازنگری: ۱۳۹۷/۱/۲۶

چکیده: یک چالش اساسی در استنباط مدل‌های آمیخته، معرفی معیارهای کارا برای انتخاب مدل است. منبع اصلی این چالش نیز برازش و محاسبه ماکسیمم تابع درستنمایی مدل می‌باشد. داده تاگی روش جدیدی است که برای برازش کارای مدل‌های آمیخته با روش ماکسیمم درستنمایی پیشنهاد شده است. این روش، اخیرا، طرفداران زیادی پیدا کرده است و مشکلات عمده سایر روش‌های استنباط مبتنی بر درستنمایی در مدل‌های آمیخته را ندارد. یکی از معایب این روش، عدم توانایی محاسبه مقدار ماکسیمم تابع درستنمایی است. این مقدار یک کمیت کلیدی در معرفی و محاسبه معیارهای انتخاب مدل محسوب می‌شود. بنابراین به‌نظر می‌رسد با روش داده تاگی نمی‌توان یک معیار اطلاع مناسب، به‌طور مستقیم، برای یافتن بهترین مدل در رده مدل‌های آمیخته، تعریف کرد. این پژوهش تلاشی است در جهت نقض این باور. در این مقاله، یک معیار مبتنی بر روش داده تاگی معرفی می‌شود و عملکرد آن در یک مطالعه شبیه‌سازی مورد ارزیابی قرار می‌گیرد.

واژه‌های کلیدی: الگوریتم MCMC، مدل آمیخته خطی تعمیم‌یافته، معیار انتخاب مدل، داده تاگی.

۱ مقدمه

در کاربردهای مختلف ماهیت متغیر پاسخ، ناگاوسی (گسسته) است و نوعی وابستگی بین مشاهدات آن وجود دارد. به عنوان نمونه می‌توان به داده‌های گسسته طولی، خوشه‌ای، و فضایی اشاره کرد. مدل‌های آمیخته

آدرس الکترونیکی نویسنده مسئول مقاله: حسین باغیشنی، hbaghishani@shahroodut.ac.ir
کد موضوع‌بندی ریاضی (۲۰۱۰): 62B10، 62B07

خطی تعمیم‌یافته [۱] (GLMMs) تعمیمی متداول از مدل‌های خطی تعمیم‌یافته [۲] (GLMs) هستند که برای تحلیل داده‌های وابسته گسسته استفاده می‌شوند. در این مدل‌ها، با افزودن اثرات تصادفی (متغیرهای پنهان) به پیش‌گوی خطی، وابستگی مشاهدات وارد می‌شود (بریسلو و کلیتون، ۱۹۹۳). برازش مدل و پیش‌گویی، دو هدف کلیدی در رده GLMM محسوب می‌شوند. دستیابی به این دو هدف با هر دو رهیافت بسامدی و بیزی دارای محدودیت‌ها و مشکلاتی است:

- استنباط‌های مبتنی بر درستنمایی، حل عددی انتگرال‌های با بعد بالا را شامل می‌شوند که می‌تواند بسیار پیچیده و هزینه‌بر باشد. به‌عبارتی، حضور انتگرال‌های رام‌نشدنی (با بعد بالا) عمده‌ترین مشکل بر سر راه استنباط‌های مبتنی بر درستنمایی در رده GLMM است که مساله انتخاب مدل را نیز شامل می‌شود.

- استنباط‌های بیزی، به دلیل پیدایش الگوریتم‌های نمونه‌گیری مونت کارلوی زنجیر مارکوفی [۳] (MCMC)، مشکلات حل انتگرال‌های رام‌نشدنی و محاسبه مشتق توابع پیچیده را ندارند. در مقابل، این استنباط‌ها همواره با مساله انتخاب توزیع پیشین و انتقاد وابستگی نتایج به این انتخاب همراه هستند. در اغلب موارد، در این مدل‌ها، از توزیع‌های پیشین ناآگاهی‌بخش عینی [۴] استفاده می‌شود که عدم وجود تعریفی واحد برای این نوع از توزیع‌های پیشین، به پیشنهادهای مختلف و در نتیجه استنباط‌های ممکن مختلف منتهی خواهد شد (ترابی و همکاران، ۲۰۱۵). البته، اخیرا، برگر و همکاران (۲۰۰۹) با ارایه یک تعریف یکتا و سرراست برای توزیع‌های پیشین مرجع [۵]، که دسته‌ای از پیشین‌های عینی محسوب می‌شوند، تا حدودی این مشکل را مرتفع کرده‌اند. با این وجود نویسندگان مایل هستند دو نکته را یادآوری کنند:

۱. آماردان‌های بیزی ذهنی‌گرا [۶] اعتقادی به استفاده از پیشین‌های عینی ندارند و استنباط بیزی صرفا به روش‌های بیزی عینی‌گرا محدود نمی‌شود. بنابراین آن‌ها پیشین‌های ذهنی خود را دارند که نحوه انتخاب آن‌ها می‌تواند از طرف سایرین مورد انتقاد قرار گیرد.

---

[1] Generalized linear mixed models
[2] Generalized linear models
[3] Markov chain Monte Carlo
[4] Objective
[5] Reference prior
[6] Subjective Bayesians

۲. سرعت کند و همگرایی ضعیف الگوریتم‌های MCMC در GLMM به انتخاب توزیع پیشین وابسته است.

تلاش‌های زیادی برای رفع این معایب در هر دو دیدگاه استنباطی مذکور صورت گرفته‌اند. به‌عنوان چند نمونه می‌توان به بریسلو و کلیتون (۱۹۹۳)، پینیرو و بیتس (۱۹۹۵)، و کریستنسن (۲۰۰۴) در رهیافت مبتنی بر درستنمایی، و کریستنسن و واگاپترسن (۲۰۰۲)، کریستنسن و همکاران (۲۰۰۶)، فونگ و همکاران (۲۰۱۰)، و حسینی و همکاران (۲۰۱۱) در دیدگاه بیزی اشاره کرد.

یک روش جانشین برای برازش مبتنی بر درستنمایی در رده GLMM، روش داده تاگی [۷] (DC) است که اولین بار توسط لله و همکاران (۲۰۰۷) معرفی و به‌کار گرفته شد. روش DC، روشی ساده برای محاسبه MLE پارامترهای مدل با بهره‌گیری از الگوریتم‌های MCMC است. این روش از یک رهیافت بیزی به‌عنوان ابزار محاسبه MLE پارامترهای مدل استفاده می‌کند و نتایج حاصل از آن نسبت به انتخاب توزیع‌های پیشین، ناوردا [۸] هستند. افزون بر این، مراحل برآورد پارامترهای مدل (و دقت آن‌ها) شامل محاسبه ساده مقادیر میانگین و واریانس نمونه‌ای است که با الگوریتم MCMC تولید می‌شود و نیازی به ماکسیمم‌سازی عددی و مشتق‌گیری از یک تابع درستنمایی پیچیده نیست.

روش DC محدودیت‌هایی را نیز به همراه دارد. یکی از آن‌ها به مقدار ماکسیمم تابع درستنمایی اختصاص دارد. این روش تنها برآورد ML پارامترهای مدل را فراهم می‌کند و قادر به محاسبه مقدار ماکسیمم تابع درستنمایی، در آن برآوردها، نیست. ماکسیمم تابع درستنمایی، کمیتی کلیدی در تعریف معیارهای انتخاب مدل به شمار می‌آید. بنابراین، با استفاده از روش DC نمی‌توان یک معیار انتخاب مدل (مستقیم) در رده‌ای از مدل‌های ممکن تعریف کرد. پیشنهادی برای رفع این مشکل توسط پونچیانو و همکاران (۲۰۰۹) ارایه شد. آن‌ها برای محاسبه نسبت درستنمایی، با استفاده از روش DC، یک الگوریتم ساده با نام الگوریتم نسبت درستنمایی DC [۹] معرفی کردند و از آن برای محاسبه اختلاف مقادیر معیارهای اطلاع مورد نیاز، به‌منظور مقایسه جفتی بین دو مدل، استفاده کردند. در روش پیشنهادی پونچیانو و همکارانش، برای مقایسه چند مدل نامزد و انتخاب بهترین مدل، از منظر یک معیار انتخاب مدل مثل معیار اطلاع آکاییک [۱۰] (AIC)، باید ابتدا مدل‌ها دو به دو با هم مقایسه و سپس مدل بهتر از ترکیب نتایج همه مقایسه‌های جفتی انتخاب شود. با این فرآیند، به‌راحتی قابل درک است که حتی اگر بعد رده

[7] Data cloning
[8] Invariant
[9] Data cloned likelihood ratio
[10] Akaike information criterion

مدل‌ها متوسط هم باشد، تعداد مقایسه‌های جفتی می‌تواند خیلی بزرگ و انتخاب بهترین، ناممکن شود.

در این مقاله، بر اساس روش داده تاگی، معیار اطلاعی معرفی می‌شود که به‌طور مستقل مدل‌های نامزد حاضر در یک رده را رتبه‌بندی می‌کند. این معیار تقریبا مشابه AIC تعریف می‌شود. در بخش ۲ مدل‌های آمیخته خطی تعمیم‌یافته و در بخش ۳ روش داده تاگی معرفی می‌شود. در بخش ۴ معیار اطلاع جدید مبتنی بر DC ارایه می‌شود و در بخش ۵ در یک مطالعه شبیه‌سازی، عملکرد معیار معرفی‌شده مورد ارزیابی قرار می‌گیرد. در پایان، بحث و نتیجه‌گیری ارایه خواهد شد.

۲ مدل‌های آمیخته خطی تعمیم‌یافته

در یک GLMM فرض می‌شود متغیرهای پاسخ با شرط معلوم بودن اثرات تصادفی، مستقل از هم بوده و دارای توزیعی از خانواده توزیع‌های نمایی هستند. برای تشریح مدل، داده‌های خوشه‌ای را در نظر بگیرید که اندازه‌های مکرر متغیر پاسخ در یک نمونه تصادفی از m خوشه رخ داده باشند. فرض کنید $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$، $i = 1, \ldots, m$، مشاهدات پاسخ خوشه iام، $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ بردار کل مشاهدات و $n = \sum_{i=1}^{m} n_i$ حجم نمونه کل باشند. همچنین فرض کنید $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iq})^T$ بردار q بعدی اثرات تصادفی در خوشه iام باشد. با شرط معلوم بودن اثرات تصادفی $\boldsymbol{u}_i$، توزیع هر یک از متغیرهای پاسخ عضوی از خانواده توزیع‌های نمایی با تابع چگالی

$$f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}) = \exp\left\{y_{ij}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{v}_{ij}^T\boldsymbol{u}_i) - a(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{v}_{ij}^T\boldsymbol{u}_i) + c(y_{ij})\right\}$$

است، که در آن $\boldsymbol{x}_{ij}$ و $\boldsymbol{v}_{ij}$، برای $i = 1, \ldots, m$ و $j = 1, \ldots, n_i$، بردارهای متغیرهای تبیینی به‌ترتیب p و q بعدی متناظر با اثرات ثابت $\boldsymbol{\beta}$ و تصادفی $\boldsymbol{u}$ و $a(\cdot)$ و $c(\cdot)$ توابعی معلوم هستند. در این مدل، پیش‌گوی خطی $\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{v}_{ij}^T\boldsymbol{u}_i$ پارامتر کانونی خانواده توزیع‌های نمایی است. اگر μ_{ij} میانگین شرطی مدل باشد، آن‌گاه $g(\mu_{ij}) = \eta_{ij}$ که در آن $g(\cdot)$ تابع پیوند مدل است. علاوه بر این، توزیع اثرات تصادفی $\boldsymbol{u}_i$ معمولا گاوسی $\boldsymbol{u}_i|\Sigma_{\boldsymbol{\theta}} \sim N(0, \Sigma_{\boldsymbol{\theta}})$ در نظر گرفته می‌شود که در آن $\boldsymbol{\theta}$ بردار r بعدی پارامترهای وابستگی مدل (مولفه‌های واریانس) است و ماتریس کوواریانس $\Sigma_{\boldsymbol{\theta}}$ معمولا به‌صورت پارامتری در نظر گرفته می‌شود. به عبارت دیگر با معلوم بودن پارامترهای θ ماتریس کوواریانس معلوم خواهد بود.

برای انجام استنباط مبتنی بر درستنمایی باید تابع درستنمایی کناری[11] مدل به صورت

$$\mathrm{L}(\boldsymbol{\beta},\boldsymbol{\theta}) \propto \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{u}_i,\boldsymbol{\beta})\phi_q(\boldsymbol{u}_i;0,\Sigma_{\boldsymbol{\theta}})d\boldsymbol{u}_i \qquad (1)$$

محاسبه شود، که در آن $\phi_q(\cdot;0,\Sigma_{\boldsymbol{\theta}})$ تابع چگالی نرمال q متغیره با بردار میانگین صفر و ماتریس کوواریانس $\Sigma_{\boldsymbol{\theta}}$ است. محاسبه تابع درستنمایی (۱) مستلزم حل انتگرالی با بعد q است و برای داده‌های حجیم، محاسبه آن کاری دشوار خواهد بود. ماکسیمم‌سازی انتگرال محاسبه‌شده، مرحله بعدی برای محاسبه MLE پارامترهای مدل است. داده تاگی راهی جذاب برای گریز از محاسبه چنین انتگرال‌هایی و ماکسیمم‌سازی نتیجه آن است.

۳ داده تاگی

داده تاگی، ترفندی محاسباتی است که به‌عنوان روشی کارا برای استنباط مبتنی بر درستنمایی در مدل‌های آمیخته به‌کار می‌رود. این روش از الگوریتم‌های نمونه‌گیری MCMC برای تولید نمونه از یک توزیع پسین به‌طور مصنوعی ساخته‌شده[12]، با نام توزیع پسین مبتنی بر داده[13] تاگی، برای محاسبه MLE پارامترها و برآورد واریانس آن‌ها استفاده می‌کند. مدل مبتنی بر داده تاگی، یک مدل بیزی کامل ساخته‌شده بر اساس k نسخه تکرارشده داده‌های اصلی است: ابتدا بردار nk بعدی $\boldsymbol{y}^{(k)} = (\boldsymbol{y},\ldots,\boldsymbol{y})$ از کپی کردن k بار بردار n بعدی مشاهدات $\boldsymbol{y}$ تشکیل می‌شود. به همین طریق مقادیر متغیرهای تبیینی نیز k بار تکرار می‌شوند. همچنین با استفاده از توزیع احتمالی اثرات تصادفی، k تحقق از آن تولید و بردار همسانه‌سازی اثرات تصادفی $\boldsymbol{u}^{(k)} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k)$ نتیجه می‌شود. برای پارامترهای نامعلوم مدل نیز توزیع پیشین $\pi(\boldsymbol{\beta},\boldsymbol{\theta})$ در نظر گرفته می‌شود. با فرض مستقل بودن نسخه‌های کپی‌شده داده‌ها از هم، تابع درستنمایی جدید که همان تابع درستنمایی اصلی به توان k است، به دست می‌آید و توزیع پسین مبتنی بر داده تاگی با یک رهیافت MCMC به‌طور تقریبی محاسبه می‌شود.

از آن‌جا که ایده پشتیبان روش DC بر ویژگی‌های مجانبی توزیع‌های پسین بنا نهاده شده است (باغیشنی و محمدزاده، ۲۰۱۱)، تعداد نسخه‌ها، یعنی k، باید به اندازه کافی بزرگ اختیار شود تا میانگین

[11] Marginal likelihood function
[12] Artificially constructed distribution
[13] DC-based posterior distibution

و ماتریس واریانس مقیاس‌بندی‌شده توزیع پسین نتیجه‌شده پارامترها به MLE و برآورد ماتریس واریانس متناظرشان همگرا شوند. برای مقیاس‌بندی ماتریس واریانس نیز از ضریب k^{-1} استفاده می‌شود. لازم به ذکر است که اگرچه در واقعیت k تکرار مستقل یک آزمایش مشابه، مجموعه‌های داده یکسانی را نتیجه نمی‌دهند و به‌طور نظری احتمال چنین رخدادی صفر است، اما این کار توسط برنامه‌های رایانه‌ای تولید مقادیر تصادفی امکان‌پذیر است.

توزیع پسین مبتنی بر DC به صورت

$$\pi^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y}) \propto \pi^{(k)}(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta})\pi^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$$

ساخته می‌شود. بر اساس ویژگی‌های مجانبی توزیع پسین مبتنی بر DC کناری پارامترها، $\pi^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$، باغیشنی و محمدزاده (۲۰۱۱) نشان دادند اگر $k \longrightarrow \infty$، آنگاه

$$\begin{aligned} \mathrm{E}^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}) &\longrightarrow (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \\ \mathrm{Var}^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}) &\longrightarrow k^{-1} \times \mathrm{Var}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \end{aligned}$$

که در آن $\mathrm{E}^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$ و $\mathrm{Var}^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$ به‌ترتیب امید ریاضی و واریانس توزیع پسین مبتنی بر DC کناری پارامترهای مدل و $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ برآوردهای ML پارامترهای $(\boldsymbol{\beta}, \boldsymbol{\theta})$ هستند. دقت شود که برای محاسبه توزیع کناری $\pi^{(k)}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y})$ نیازی به انتگرال‌گیری از توزیع پسین مبتنی بر DC نسبت به $\boldsymbol{u}$ نیست و به آسانی با کنار گذاشتن نمونه‌های تولیدشده $\boldsymbol{u}$ در الگوریتم MCMC، نتیجه نمونه‌ای از توزیع کناری مذکور خواهد بود. یکی از مزایای روش DC ناوردایی نتایج به‌دست آمده نسبت به انتخاب توزیع‌های پیشین است (لهله و همکاران، ۲۰۰۷).

۱.۳ تشخیص تعداد نسخه‌ها

همان‌طور که اشاره شد برای اطمینان از همگرایی برآوردهای نتیجه‌شده با روش DC به MLE، باید تعداد نسخه‌ها به اندازه کافی بزرگ اختیار شود. اما برای هر موقعیت عملی، سوال چند نسخه؟ مطرح خواهد بود. لهله و همکاران (۲۰۱۰) معیارهایی برای تشخیص همگرایی الگوریتم DC ارایه کردند که بر اساس توزیع پسین توام مبتنی بر DC تعریف می‌شوند. این معیارها عبارتند از ماکسیمم مقدار ویژه ماتریس کوواریانس توزیع پسین مبتنی بر DC ($\lambda_{\max}$)، میانگین توان دوم خطا (MSE) و یک آماره برازش

شبیه معیار همبستگی ($r^۲$). معیار ماکسیمم مقدار ویژه، تباهیدگی توزیع پسین مبتنی بر DC در نقاط MLE را نشان می‌دهد؛ در حالی‌که دو معیار دیگر بیانگر مناسب بودن تقریب نرمال هستند. هر سه معیار با افزایش تعداد نسخه‌ها، یعنی k، باید به صفر همگرا شوند. بنا به لله و همکاران (۲۰۱۰)، مقدار k که به ازای آن معیارهای معرفی‌شده کمتر از ۰٫۰۵ باشند، مناسب و کافی است.

۴ یک معیار اطلاع مبتنی بر روش DC

همان‌طور که اشاره کردیم، مقدار ماکسیمم تابع درستنمایی یک کمیت کلیدی در تعریف معیارهای انتخاب مدل است و روش DC قادر به محاسبه مقدار ماکسیمم تابع درستنمایی نیست. بنابراین، با استفاده از این روش نمی‌توان یک معیار انتخاب مدل (مستقیم) در رده‌ای از مدل‌های ممکن تعریف کرد. پونچیانو و همکاران (۲۰۰۹) توانستند الگوریتمی برای محاسبه اختلاف مقادیر معیارهای اطلاع معرفی کنند و مدل‌های نامزد را به‌صورت جفتی مقایسه کنند. این رهیافت، با توجه به بعد رده مدل‌ها، می‌تواند هزینه‌بر و طاقت‌فرسا باشد. در این بخش، معیاری معرفی می‌شود که بر مبنای روش داده تاگی به‌دست می‌آید و اعضای رده مدل‌ها را به‌طور مستقیم و تکی رتبه‌بندی می‌کند.

فرض کنید $g(y)$ تابع چگالی مدل واقعی باشد که داده‌ها از آن تولید شده‌اند و $f(y,\psi)$ تابع چگالی مدل پارامتری (نامزد) باشد که برای داده‌ها در نظر گرفته شده است. چندین روش برای اندازه‌گیری فاصله تقریب پارامتری $f(y,\psi)$ تا چگالی واقعی $g(y)$ وجود دارند. فاصله‌ای که مرتبط با روش ماکسیمم درستنمایی است، فاصله کولبک-لیبلر [۱۴] است که به صورت

$$\mathrm{KL}(g(\cdot), f(\cdot,\psi)) = \int g(y) \log \frac{g(y)}{f(y,\psi)} dy$$

تعریف می‌شود. لگاریتم تابع درستنمایی با $\ell_n(\psi) = \log \mathrm{L}_n(\psi)$ نشان داده می‌شود، که در آن اندیس n برای تاکید وابستگی به حجم نمونه است. تابع لگاریتم درستنمایی داده‌های کپی‌شده به صورت

$$\ell_n^k(\psi) = \log[\mathrm{L}_n(\psi)]^k = \log[\prod_{i=۱}^{n} f(y_i,\psi)]^k$$

[14] Kullback-Leibler distance

تعریف می‌شود. با توجه به استقلال نسخه‌های کپی‌شده، $g^k(y) = [g(y)]^k$ تابع چگالی مدل واقعی برای داده‌های کپی‌شده است. بنابراین باید به‌دنبال مدلی بود که فاصله

$$\mathrm{KL}(g^k(\cdot), f^k(\cdot,\psi)) = \int g^k(y) \log \frac{g^k(y)}{f^k(y,\psi)} dy \qquad (۲)$$

را می‌نیمم کند. برای هر مدل دلخواه $f(\cdot,\psi)$، مولفه $\int g^k(y) \log g^k(y) dy$ مقداری ثابت است. از طرفی، با استفاده از قانون قوی اعداد بزرگ

$$\frac{۱}{n}\ell_n^k(\psi) \xrightarrow{a.s} \int g^k(y) \log f^k(y,\psi) dy = \mathrm{E}_{g^k}\left(\log f^k(y,\psi)\right).$$

بنابراین چون برآوردگر $\hat{\psi}$ ماکسیمم‌کننده $\ell_n^k(\psi)$ است، تحت شرایط مناسب به $\psi_\circ$، یعنی می‌نیمم‌کننده فاصله (۲)، میل می‌کند، پس

$$\hat{\psi} \xrightarrow{a.s} \psi_\circ = \arg\min \mathrm{KL}(g^k(\cdot), f^k(\cdot,\psi)).$$

در ادامه، چند کمیت تعریف می‌شوند. کمیت‌های

$$u(y,\psi) = \frac{\partial}{\partial\psi} \log f^k(y,\psi), \quad I(y,\psi) = \frac{\partial^۲}{\partial\psi\partial\psi^T} \log f^k(y,\psi)$$

به‌ترتیب بردار امتیاز و ماتریس اطلاع متناظر با داده‌های کپی‌شده نامیده می‌شوند. چون $\psi_\circ$ فاصله (۲) را می‌نیمم می‌کند، پس

$$\mathrm{E}_{g^k}\left(u(y,\psi_\circ)\right) = \int g^k(y) u(y,\psi_\circ) dy = ۰.$$

با قرار دادن

$$J = -\mathrm{E}_{g^k} I(y,\psi_\circ), \quad K = \mathrm{Var}_{g^k} u(y,\psi_\circ).$$

وقتی $f^k(y,\psi_\circ)$ با $g^k(y)$ یکی باشد، J و K با بعد κ، ماتریس‌های همانی خواهند شد که در آن κ

بعد بردار پارامتر ψ است. با شرایط و تعاریف بالا ثابت می‌شود (هیورت و پلارد، ۱۹۹۳)

$$\hat{\psi} = \psi_{\circ} + J^{-1}\bar{U}_n + o_p(n^{-\frac{1}{2}}) \qquad (۳)$$

که در آن $\bar{U}_n = n^{-1}\sum_{i=1}^{n} u(y_i, \psi_{\circ})$. از قضیه حد مرکزی

$$\sqrt{n}\bar{U}_n \xrightarrow{d} U' \sim N_{\kappa}(\circ, K)$$

که در ترکیب با (۳) نتیجه می‌شود

$$V_n = \sqrt{n}(\hat{\psi} - \psi_{\circ}) \xrightarrow{d} J^{-1}U' = N_{\kappa}(\circ, J^{-1}KJ^{-1}). \qquad (۴)$$

با قرار دادن $\hat{\psi}$ در (۲) و ثابت بودن مولفه $\int g^k(y)\log g^k(y)dy$، کافی است

$$R_n = \int g^k(y)\log f^k(y, \hat{\psi})dy$$

به‌دست آورده شود، که یک متغیر تصادفی است و از طریق برآوردگر $\hat{\psi}$ به داده‌ها وابسته است. به همین دلیل از امید ریاضی آن، تحت تابع چگالی مدل واقعی، استفاده می‌شود. بنابراین

$$Q_n = \mathrm{E}_{g^k}(R_n) = \mathrm{E}_{g^k}(\int g^k(y)\log f^k(y, \hat{\psi})dy).$$

به‌طور کلی استراتژی معیارهای انتخاب مدل مبتنی بر ماکسیمم تابع درستنمایی، برآورد Q_n برای هر مدل نامزد و انتخاب مدل با بالاترین Q_n برآوردشده است. این فرآیند معادل با جستجو برای مدلی با کوچک‌ترین فاصله کولبک-لیبلر برآوردشده است. یک حالت ممکن برای برآورد Q_n، استفاده از توزیع تجربی داده‌های کپی‌شده است. در نتیجه

$$\hat{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\log f^k(y_i, \hat{\psi}) = \frac{1}{n}\ell_n^k(\hat{\psi}).$$

در داده‌های کپی‌شده، هر مشاهده k بار تکرار شده است. بنابراین مخرج $\hat{Q}_n$ برابر n است نه nk.

اکنون $Z_i = \log f^k(y_i, \psi_\circ) - Q_\circ$ قرار داده می‌شود، که در آن

$$Q_\circ = \int g^k(y) \log f^k(y, \psi_\circ) dy.$$

متغیرهای Z_i، متغیرهایی با میانگین صفر هستند. $\bar{Z}_n$ نیز میانگین Z_iها تعریف می‌شود. بنابراین می‌توان نشان داد

$$\hat{Q}_n - R_n = \bar{Z}_n + n^{-1} V_n^T J V_n + o_p(n^{-1}). \qquad (۵)$$

با توجه به (۴) می‌توان نوشت

$$V_n^T J V_n \xrightarrow{d} W = (U')^T J^{-1} J J^{-1} U' = (U')^T J^{-1} U'.$$

بنابراین

$$\begin{aligned} \mathrm{E}(\hat{Q}_n - R_n) = \mathrm{E}(\hat{Q}_n - Q_n) &= \mathrm{E}(\bar{Z}_n + n^{-1} V_n^T J V_n + o_p(n^{-1})) \qquad (۶) \\ &= \mathrm{E}(\bar{Z}_n) + n^{-1}\mathrm{E}(W) \approx \frac{\kappa^*}{n} \end{aligned}$$

که در آن $\kappa^* = \mathrm{E}(W)$ اثر[۱۵] ماتریس $J^{-1}K$ و جمله جریمه در تعریف معیارهای انتخاب مدل است.

تذکر ۱: اگر $f^k(y, \psi_\circ) = g^k(y)$، آن‌گاه $K = J$ و بنابراین $\kappa^* = \kappa$.

با نتیجه به‌دست آمده، یک نسخه با اریبی تصحیح‌شده از برآوردگر خام $\hat{Q}_n$ به صورت

$$\hat{Q}_n - \frac{\kappa^*}{n} = n^{-1}\{\ell_n^k(\hat{\psi}) - \kappa^*\}$$

حاصل می‌شود.

[15] Trace

۱.۴ معیار انتخاب مدل پیشنهادی

اگر داده‌ها مستقل باشند، آن‌گاه درجه آزادی مدل برابر حجم نمونه منهای تعداد پارامترهای برآوردشده مدل است. این تعریف روشن و عام هست و در معرفی جمله جریمه در معیارهای انتخاب مدل معروف مثل AIC و BIC[16] از آن استفاده می‌شود. اما زمانی که داده‌ها وابسته هستند، تعریف روشنی از درجه آزادی مدل وجود ندارد. یعنی مشخص نیست که چه تعداد از داده‌ها برای برآورد پارامترها کافی است. بنابراین، مساله اصلی در معرفی معیارهای انتخاب مدل برای مدل‌های آمیخته، تعریف یک جمله جریمه مناسب است.

کارهای متفاوتی به منظور معرفی یک جریمه مناسب برای AIC در مدل‌های آمیخته انجام شده‌اند. به‌عنوان نمونه می‌توان به وایدا و بلانچارد (۲۰۰۵)، رفتری و همکاران (۲۰۰۷)، گراون و نیب (۲۰۱۰)، و لیانگ و همکاران (۲۰۰۸) اشاره کرد. رفتری و همکاران (۲۰۰۷) با استفاده از میانگین هارمونیک مقادیر تابع درستنمایی به ازای نمونه‌های حاصل از توزیع پسین، برآوردی برای ماکسیمم تابع درستنمایی ارایه دادند. این برآوردگر در عین سادگی، به‌شدت ناپایدار است. آن‌ها برای رفع این مشکل، از توزیع (تقریبی) $\ell_{\max} - \ell_t$ استفاده کردند و نشان دادند

$$\ell_{\max} - \ell_t \sim Gamma(\alpha, ۱)$$

که در آن $\{\ell_t;\ t = ۱, \ldots, B\}$ مقادیر تابع درستنمایی به ازای B نمونه تولیدشده از توزیع پسین، $\ell_{\max}$ مقدار ماکسیمم تابع درستنمایی و $\alpha = \frac{\kappa}{۲}$ هستند. با توجه به آن‌که $\mathrm{E}(\ell_{\max} - \ell_t) = \alpha$ و $\mathrm{Var}(\ell_t) = \alpha$، برآوردهای گشتاوری به صورت

$$\hat{\ell}_{\max} = \bar{\ell} + s_\ell^۲, \quad \hat{\alpha} = s_\ell^۲$$

نتیجه می‌شوند، که در آن‌ها $\bar{\ell}$ و $s_\ell^۲$ میانگین نمونه و واریانس ℓ_t ها هستند. با توجه به معیار

$$\mathrm{AIC} = ۲\ell_{\max} - ۲\kappa$$

[16] Bayesian information criterion

رفتری و همکاران (۲۰۰۷) یک برآورد مونت کارلویی از آن را به صورت

$$\mathrm{AICM} = ۲\hat{\ell}_{\max} - ۲\hat{\kappa} = ۲(\bar{\ell} + s_{\ell}^{۲}) - ۴s_{\ell}^{۲} = ۲\bar{\ell} - ۲s_{\ell}^{۲} = ۲(\bar{\ell} - s_{\ell}^{۲})$$

ارایه دادند، که با الهام از آن و داشتن یک نمونه از توزیع پسین مبتنی بر DC، معیار

$$\mathrm{DCAIC} = ۲(\bar{\ell}^{k}(\hat{\psi}) - s_{\ell k}^{۲})$$

یا به‌طور معادل

$$\mathrm{DCAIC} = -۲\bar{\ell}^{k}(\hat{\psi}) + ۲s_{\ell k}^{۲} \qquad (۷)$$

پیشنهاد می‌شود. این معیار نسخه تعمیم‌یافته‌ای از AIC با استفاده از روش داده تاگی است. بنابراین کاملا منطقی است که تمام ویژگی‌های AIC، از جمله سازگاری را به ارث ببرد. یعنی با احتمال متمایل به یک قادر است مدل واقعی را در بین مدل‌های نامزد، در صورت وجود آن، انتخاب کند. همچنین AIC تمایل به انتخاب مدل‌های پیچیده‌تر دارد، یعنی مدل انتخابی توسط این معیار می‌تواند بیش‌برازش[۱۷] باشد. این موضوع برای DCAIC نیز برقرار است. مدل با کمترین مقدار DCAIC، برحسب رابطه (۷)، به‌عنوان مدل برتر انتخاب می‌شود.

۵ ارزیابی عملکرد معیار پیشنهادی

برای ارزیابی معیار ارایه‌شده، بردار پاسخ با حجم n از یک مدل آمیخته خطی تعمیم‌یافته پواسون به صورت

$$\begin{aligned} Y_i|\lambda_i &\sim Poisson(\lambda_i),\ i = ۱, \ldots, n \qquad (۸) \\ \lambda_i &= \exp(\boldsymbol{x}_i^T \beta + \alpha_i) \qquad (۹) \\ \alpha_i &\sim N(۰, \sigma^{۲}) \end{aligned}$$

[17] Overfitted

تولید شده است، که در آن $\boldsymbol{x}_i = (۱, x_{۱i}, x_{۲i}, x_{۳i})^T$ و مقادیر واقعی پارامترها

$$(\beta_۰, \beta_۱, \beta_۲, \beta_۳, \sigma) = (۱, -۰/۵, ۱/۵, ۰/۷, ۰/۵)$$

انتخاب شدند. متغیرهای تبیینی $x_۱$، $x_۲$ و $x_۳$ به‌ترتیب از توزیع‌های یکنواخت استاندارد، نرمال استاندارد، و برنولی با احتمال موفقیت ۰/۵ تولید شده‌اند. سه حجم نمونه مختلف $n = ۲۰, ۵۰, ۱۰۰$ نیز در نظر گرفته شده‌اند. علاوه بر این، تعداد نسخه‌ها برای داده‌های کپی‌شده برابر $k = ۱, ۵$ انتخاب شدند. دلیل انتخاب $k = ۵$ برای روش DC، مقادیر نتیجه‌شده معیارهای تشخیص همگرایی است که در جدول ۱ برای حجم‌های نمونه مختلف گزارش شده‌اند. تنها برای حجم نمونه $n = ۲۰$ این انتخاب می‌تواند کمی مورد تردید قرار گیرد. اما مقدار معیار $r^۲$ برای $k = ۵$ در این حالت نزدیک به ۰/۰۵ است. همه محاسبات در نرم‌افزار R

جدول ۱. مقادیر معیارهای تشخیص همگرایی الگوریتم DC

$r^۲$	MSE	$\lambda_{\max}$	k	n
۰/۳۰۵	۳۵/۲۶	۰/۸۱۳	۱	۲۰
۰/۰۷۹	۲/۶۴۲	۰/۱۱۳	۵	
۰/۰۵۶	۱/۱۲۲	۰/۰۰۷	۱۰	
۰/۰۰۷	۰/۲۲۸	۰/۰۶۶	۱	۵۰
۰/۰۰۲	۰/۰۲۷	۰/۰۱۳	۵	
۰/۰۰۱	۰/۰۴۵	۰/۰۰۷	۱۰	
۰/۰۰۲	۰/۰۱۹	۰/۱۲۷	۱	۱۰۰
۰/۰۰۰۴	۰/۰۰۴	۰/۰۲۱	۵	
۰/۰۰۱	۰/۰۲۹	۰/۰۱۱	۱۰	

(تیم هسته R، ۲۰۱۶) و به کمک بسته dclone (سلیموس، ۲۰۱۰) انجام شده‌اند.

برای بررسی توانایی DCAIC در انتخاب مدل واقعی و کنکاش ویژگی بیش‌برازشی آن، دو رده مدل (نادرست) کم‌برازش[۱۸] و بیش‌برازش در نظر گرفته شده‌اند. رده مدل‌های کم‌برازش، شامل سه مدل با پیش‌گوهای خطی به صورت

$$\mathrm{M}_۱ \quad : \quad \eta_i = \beta_۰ + \beta_۱ x_{۱i} + \alpha_i$$

[18] Underfitted

$$\mathrm{M}_{۲} \quad : \quad \eta_i = \beta_{۰} + \beta_{۲} x_{۲i} + \alpha_i$$
$$\mathrm{M}_{۳} \quad : \quad \eta_i = \beta_{۰} + \beta_{۲} x_{۲i} + \beta_{۳} x_{۳i} + \alpha_i$$

هستند. برای رده مدل‌های بیش‌برازش، تنها یک مدل با پیش‌گوی خطی

$$\mathrm{M}_{۴} : \eta_i = \beta_{۰} + \beta_{۱} x_{۱i} + \beta_{۲} x_{۲i} + \beta_{۳} x_{۳i} + \beta_{۴} x_{۴i} + \alpha_i$$

در نظر گرفته شد، که در آن $x_{۴}$ از توزیع یکنواخت استاندارد شبیه‌سازی شده است. در این مطالعه، ۱۰۰ مجموعه داده از مدل واقعی تولید شدند و برای هر مجموعه، هر ۵ مدل درست ($\mathrm{M}_{۰}$) و نادرست ($\mathrm{M}_{۱}$ تا $\mathrm{M}_{۴}$) برازش و معیار DCAIC محاسبه شد. نتایج برای تعداد نسخه‌های ۱ و ۵ در جدول ۲ گزارش شده‌اند که اعداد آن بیانگر تعداد دفعات انتخاب مدل واقعی توسط معیار DCAIC در مقابل مدل‌های رقیب در دو گروه مدل‌های کم‌برازش و بیش‌برازش، در ۱۰۰ مجموعه داده، هستند. برای $k = ۱$ که در واقع فقط از داده‌های اصلی استفاده می‌شود، معیار DCAIC همان معیار پیشنهادی رفتری و همکاران (۲۰۰۷) خواهد بود که تقریبی از AIC در مدل آمیخته شبیه‌سازی‌شده است. مشاهده می‌کنید که وقتی مدل رقیب یک مدل کم‌برازش است، با افزایش حجم نمونه، معیار پیشنهادی به انتخاب مدل واقعی تمایل دارد، در حالی‌که برای مدل رقیب بیش‌برازش شانس انتخاب مدل واقعی تقریبا برابر شانس عدم انتخاب آن است. از طرفی، در حالتی که حجم نمونه کوچک است، مدل واقعی معمولا در مقابل رقبای خود بازنده است. این نتیجه ویژگی بیش‌برازشی معیار DCAIC را که از معیار AIC به ارث می‌برد، به وضوح، نشان می‌دهد. در مقابل زمانی که تعداد نسخه‌های داده به $k = ۵$ افزایش می‌یابد، معیار DCAIC در اغلب موارد برای حالتی که مدل رقیب یک کم‌برازش است، مدل واقعی را انتخاب می‌کند. این ویژگی (سازگاری) با افزایش حجم نمونه نیز قوی‌تر می‌شود. در مقابل، زمانی که مدل رقیب یک بیش‌برازش است، در اغلب موارد، مدل واقعی شکست می‌خورد که نشان از نمود بارزتر ویژگی بیش‌برازشی در این حالت است.

با توجه به آن‌که شالوده یک توزیع پسین مبتنی بر DC بر پایه توزیع‌های پسین بیزی شکل می‌گیرد، بنا بر پیشنهاد یکی از داوران مقاله، مقایسه معیار پیشنهادی با یک معیار انتخاب مدل بیزی مثل معیار اطلاع کیبش[۱۹] (DIC) می‌تواند مفید باشد. برای این منظور، در مطالعه شبیه‌سازی تعداد دفعات انتخاب مدل واقعی توسط DIC نیز محاسبه شدند که در جدول ۲ گزارش شده‌اند. همانطور که ملاحظه می شود روند عملکرد دو معیار برای زمانی که از داده‌های واقعی استفاده می‌شود، تقریبا مشابه است، اما وقتی

[19] Deviance information criterion

جدول ۲. تعداد دفعات انتخاب مدل واقعی با دو معیار

k	مدل	DCAIC n = ۲۰	DCAIC n = ۵۰	DCAIC n = ۱۰۰	DIC n = ۲۰	DIC n = ۵۰	DIC n = ۱۰۰
۱	کم‌برازش	۳۹	۵۱	۷۲	۲۱	۴۲	۶۳
	بیش‌برازش	۴۶	۴۷	۵۹	۵۶	۴۶	۷۳
۵	کم‌برازش	۷۲	۷۲	۸۶	۲۶	۶۷	۶۵
	بیش‌برازش	۱۱	۱۹	۱۶	۶۲	۴۲	۶۳

از داده‌های کپی‌شده استفاده می‌شود، دو معیار نتایج متفاوتی را ارایه می‌دهند. در واقع معیار DCAIC در انتخاب مدل‌های کم‌برازش، به‌ویژه در حجم نمونه کوچک، توفیق خیلی بیشتری نسبت به DIC دارد و این نتیجه برای مدل‌های رقیب بیش‌برازش برعکس است. البته نویسندگان از معتبر بودن معیار DIC برای انتخاب مدل بر اساس داده‌های کپی‌شده، مطمئن نیستند. بنابراین مقایسه بین دو معیار برای حالت $k \neq ۱$، مورد سوال است.

بحث و نتیجه‌گیری

در این مقاله، یک معیار اطلاع جدید مبتنی بر روش داده تاگی ارایه شد که برای انتخاب مدل در رده مدل‌های آمیخته می‌تواند یک گزینه ممکن دلخواه باشد. عملکرد این معیار، بر اساس یک مطالعه شبیه‌سازی، نشان داد که هر دو ویژگی سازگاری و تمایل به انتخاب مدل‌های پیچیده معیار AIC در ذات این معیار جدید نیز وجود دارند. اما، این معیار نسخه قابل اعتمادتری از نسخه ارایه‌شده توسط رفتری و همکاران (۲۰۰۷) است.

با توجه به ویژگی بیش‌برازشی DCAIC، انتظار می‌رود بتوان نسخه سازواری از آن را که مشابه BIC عمل می‌کند، به‌دست آورد که مشکل بیش‌برازشی را با توجه به جریمه سنگین‌تر موجود در BIC نسبت به AIC، مرتفع کند. امکان‌سنجی بسط DCAIC به مواردی که ویژگی‌های سایر معیارهای اطلاع پرمصرف را دارا باشد، توسط نویسندگان در حال بررسی است و به‌عنوان یک موضوع پژوهشی آینده مطرح است.

تقدیر و تشکر

نویسندگان از داوران محترم برای بیان نظرات ارزشمندشان که در ارتقای کیفیت مقاله موثر بودند، قدردانی می‌کنند.

مراجع

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

Berger, J. O., Bernardo, J. M. and Sun, D. (2009), The Formal Definition of Reference Priors, *The Annals fo Statistics*, **37**, 905-938.

Breslow, N. and Clayton, D. G. (1993), Approximate Inference in Generalized Linear Mixed Models, *Jornal of the American Statistical Association*, **88**, 9-25.

Christensen, O. F. (2004), Monte Carlo Maximum Likelihood in Model-Based Geostatistics, *Journal of Computational and Graphical Statistics*, **13**, 702-718.

Christensen, O. F. and Waagepetersen, R. P. (2002), Bayesian Prediction of Spatial Count Data Using Generalized Linear Mixed Models, *Biometrics*, **58**, 280-286.

Christensen, O. F., Roberts, G. O. and Skold, M. (2006), Robust Markov Chain Monte Carlo Methods for Spatial Generalized Linear Mixed Models, *Journal of Computational and Graphical Statistics*, **15**, 1-17.

Fong, Y., Rue, H. and Wakefield, J. (2010), Bayesian Inference for Generalized Linear Mixed Models, *Biostatistics*, **11**, 397-412.

Greven, S. and Kneib, T. (2010), On the Behavior of Marginal and Conditional Akaike Information Criteria in Linear Mixed Models, *Biometrika*, **97**, 773-789.

Hjort, N. L. and Pollard, D. B. (1993), Asymptotics for Minimisers of Convex Processes, *Statistical Research Report*, Department of Mathematics, University of Oslo.

Hosseini, F., Eidsvik, J. and Mohammadzadeh, M. (2011), Approximate Bayesian Inference for Generalized Linear Mixed Models with Skew Normal Latent Variables, *Computational Statistics and Data Analysis*, **55**, 1791-1806.

Lele, S. R., Dennis, B. and Lutscher, F. (2007), Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models Using Bayesian Markov Chain Monte Carlo Methods, *Ecology Letters*, **10**, 551–563.

Lele, S. R., Nadeem, K. and Schmuland, B. (2010), Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning, *Journal of the American Statistical Association*, **105**, 1617-1625.

Liang, H., Wu, H. and Zou, G. (2008), A Note on Conditional AIC for Linear Mixed-effects Models, *Biometrika*, **95**, 773-778.

Pinheiro, J. C. and Bates, D. M. (1995), Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model, *Journal of Computational and Graphical Statistics*, **4**, 12-35.

Ponciano, J. M., Taper, M. L., Dennis, B. and Lele, S. R. (2009), Hierarchical Models in Ecology: Confidence Intervals, Hypothesis Testing, and Model Selection Using Data Cloning, *Ecology*, **90**, 356–362.

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. (2007), Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity, *In. Bernardo et al. (eds) Bayesian Statistics*, Oxford University Press.

R Core Team (2016), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Solymos P. (2010), dclone: Data Cloning in R. *The R Journal*, **2**, 29-37.

Torabi, M., Lele, S. R. and Parsad, N. G. N. (2015), Likelihood Inference for Small Area Estimation Using Data Cloning, *Computational Statistics and Data Analysis*, **89**, 158-171.

Vaida, F. and Blanchard, S. (2005), Conditional Akaike Information for Mixed-Effects Models, *Biometrika*, **92**, 351-370.