مجله علوم آماری، بهار و تابستان ۱۳۸۹
جلد ۴، شماره ۱، ص ۳۵-۵۸

بهبود الگوریتم ساختاری مونت کارلوی زنجیر مارکوف در مدل‌های چندسطحی با متغیر پاسخ نرمال

عاطفه فرخی، موسی گل‌علی‌زاده
گروه آمار، دانشگاه تربیت مدرس

تاریخ دریافت: ۱۳۸۸/۷/۱۶ تاریخ آخرین بازنگری: ۱۳۸۹/۲/۲۳

چکیده : مدل‌های چندسطحی در علوم کاربردی شامل علوم اجتماعی، جامعه‌شناسی، پزشکی و اقتصاد برای تحلیل داده‌های همبسته مورد استفاده قرار می‌گیرند. روش‌های متفاوتی برای برآورد این مدل‌ها با متغیر پاسخ نرمال وجود دارند. در این مقاله برای به‌کارگیری روش بیزی از تعمیم الگوریتم مونت کارلوی زنجیر مارکوف استفاده می‌شود که قالبی ساده داشته و باعث حذف همبستگی بین نمونه‌های شبیه‌سازی شده برای پارامترهای ثابت و خطاهای منتسب به گروه‌ها می‌شود. چون بعد ماتریس کواریانس بردار خطای جدید افزایش می‌یابد، برای تسریع همگرایی این روش دو راهکار بر مبنای تجزیه چولسکی ماتریس‌های کواریانس پیشنهاد می‌شود. سپس عملکرد این روش‌ها در مطالعه شبیه‌سازی و مثالی کاربردی مورد ارزیابی قرار می‌گیرد.

واژه‌های کلیدی : د اده‌های چندسطحی، مدل‌های عرض از مبداء تصادفی، الگوریتم مونت کارلوی زنجیر مارکوف، تجزیه چولسکی.

آدرس الکترونیک مسئول مقاله: موسی گل‌علی‌زاده، golalizadeh@modares.ac.ir
کد موضوع‌بندی ریاضی (۲۰۰۰): ۶۲F۱۵ و ۶۲J۹۹

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

۱ مقدمه

معمولاً بررسی‌های آماری روابط بین متغیرهای تبیینی و وابسته که به‌صورت ترکیب‌های مختلف از مشاهدات کمی و کیفی باشند در قالب رگرسیون، تحلیل واریانس و تحلیل‌های کواریانس صورت می‌گیرد. یکی از فرض‌های اساسی در کاربرد این گونه مدل‌ها استقلال آماری بین مشاهدات است. گاهی اوقات این فرض برای موضوع مورد مطالعه صادق نبوده، در نتیجه به‌کارگیری مدل‌های رگرسیونی متداول دارای اشکال می‌باشد. به‌عنوان نمونه پینهریو و بیتس (۲۰۰۰) نشان دادند، نادیده گرفتن فرض همبستگی بین مشاهدات منجر به کم‌برآوردی خطای برآورد ضرایب رگرسیونی می‌شود. مثال‌های زیادی در حوزه علوم کاربردی شامل علوم اجتماعی، علوم پزشکی، جامعه‌شناسی و غیره وجود دارند که در آن‌ها همبستگی بین مشاهدات کاملا مشهود است. به‌عنوان مثال، در بررسی میزان کلسترول خون تعدادی بیمار، می‌توان انتظار داشت شباهتی نسبی بین وضعیت بیمارانی که تحت درمان یک پزشک هستند وجود داشته باشد (توایسک، ۲۰۰۶). مدل مناسب تحلیل داده‌هایی مانند مثال فوق، مدل چندسطحی است. ویژگی اصلی داده‌های چندسطحی خصوصیت گروه‌بندی آن‌ها است که در مدل‌بندی آماری لحاظ می‌شود.

روش‌های استنباط آماری زیادی راجع به پارامترهای مدل‌های چندسطحی معرفی شده است (گلداستاین، ۱۹۹۹). طبق معمول چنین استنباط‌هایی به دو شیوه بسامدی و بیزی صورت می‌گیرد. در این بین روش‌های بیزی شامل موضوعات نظری و روش‌های مبتنی بر شبیه‌سازی کامپیوتری و به‌ویژه روش‌های مونت کارلوی زنجیر مارکوف[۱] (MCMC) هستند. مقایسه این روش‌ها توسط براون و درایپر (۲۰۰۰) صورت گرفت. گلفند و همکاران (۱۹۹۵) نیز کاربرد روش‌های مختلف بیزی را در مدل‌های چندسطحی مورد مطالعه قرار دادند.

برای برآورد کردن پارامترهای مدل چندسطحی با روش‌های MCMC و برطرف کردن مشکل همگرایی نمونه‌های تولید شده سرجنت و همکاران (۲۰۰۰) روشی بنام روش ساختاری مونت کارلوی زنجیر مارکوف[۲] (SMCMC) را پیشنهاد کردند.

[۱] Markov Chain Monte Carlo

[۲] Structured Markov Chain Monte Carlo

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

این روش کارایی بسیار خوبی در مقایسه با روش‌های بازپارامتری داشته و در عین حال از ساختار ساده‌ای برخوردار است، به‌طوری که اکثر مدل‌های چندسطحی را می‌توان در آن گنجاند. اما نکته قابل تأمل در روش SMCMC، حجم بالای محاسبات ناشی از بزرگ بودن ماتریس کواریانس مدل تغییر یافته است.

در بخش دوم این مقاله روش SMCMC به اختصار معرفی می‌شود. سپس در بخش سوم ابتدا راهکارهایی برای بهبود روش SMCMC پیشنهاد شده و به‌منظور تشریح بهتر مطالب دو مدل عرض از مبدا تصادفی در نظر گرفته می‌شود. در بخش چهارم با شبیه‌سازی و یک مثال کاربردی روش‌های موجود و پیشنهادی مورد مقایسه قرار می‌گیرند. بحث و نتیجه‌گیری در بخش پنجم ارائه خواهد شد.

۲ روش ساختاری مونت کارلوی زنجیر مارکوف

در روش‌های MCMC تعیین زمان همگرایی، اخذ نمونه‌های تقریباً مستقل، کاهش برآورد خطای برآوردگرها و خصوصیات کارای دیگر زنجیر از موضوعات بسیار اساسی به‌شمار می‌رود. برای آنکه بتوان نمونه‌هایی تقریبا مستقل از توزیع هدف تولید کرد، لازم است نحوه تولید زنجیر مارکوف، زمان همگرایی آن، کاهش همبستگی نمونه‌های حاصل از زنجیر و معیارهای دیگر مورد توجه قرار گیرد. مقاله‌های متعددی در بررسی این مطالب چاپ شده‌اند. به‌عنوان مثال می‌توان به هیل و اسمیت (۱۹۹۲)، گیلکس و رابرت (۱۹۹۶) اشاره کرد. در بین روش‌های MCMC موضوع بهبود زمان دستیابی به همگرایی از طریق کاهش همبستگی نمونه‌های حاصل از زنجیر و در عین حال رسیدن به برآورد دقیق انگیزه‌ای شد تا محققین در این زمینه تحقیقات بیشتری به‌عمل آورند. روش‌های بازپارامتری[۳] یکی از نتایج این فعالیت‌ها بوده است. روش‌های بازپارامتری شامل روش‌های متعامدسازی[۴] (هیل و اسمیت، ۱۹۹۲)، روش‌های مرکزی کردن سلسله مراتبی[۵]

[۳] Reparameterization
[۴] Orthogonalization
[۵] Hierarchical centering

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

(گلفند و همکاران، ۱۹۹۵) و روش‌های بسط پارامتر[۶] (گلمن و همکاران، ۲۰۰۸) است. پاپاس پیلیوپوس و همکاران (۲۰۰۷) نیز یک صورت کلی برای پارامتربندی مدل‌های سلسله مراتبی ارائه کردند.

سرجنت و همکاران (۲۰۰۰) به نیت کاهش همبستگی نمونه‌های حاصل از زنجیر مربوط به پارامترهای مدل‌های چندسطحی روش جدیدی تحت عنوان SMCMC ارائه کردند که شباهت زیادی با مدل‌بندی رگرسیون چندگانه دارد. آن‌ها مطالعه خود را روی مدل‌هایی با پارامترهای زیاد متمرکز کردند. مدل‌هایی با پارامتر زیاد شامل مدل‌های سلسله مراتبی، مدل‌های مولفه واریانس و بعضی از مدل‌های فضایی است (بسیگ و همکاران، ۱۹۹۲). قابل ذکر است که روش SMCMC در تعدادی از مدل‌های اقتصادی و همچنین مدل‌های رشد نیز به‌کار گرفته شد (میرا و سرجنت، ۲۰۰۳). همچنین استفاده عملی آن در مدلی خاص از آمار فضایی و نمایش کارایی آن نسبت به مدل‌گزینی‌های دیگر توسط بانرجی و همکاران (۲۰۰۴) صورت گرفت.

ساده‌ترین مدل سلسله مراتبی، مدلی با اثرات تصادفی یک طرفه متعادل به‌صورت

$$y_{ij} = \beta_0 + u_j + \epsilon_{ij}, \qquad i = 1, \ldots, n_j,\ j = 1, \ldots, J \qquad (1)$$

است، که در آن اندیس‌های i و j به ترتیب بیانگر سطح اول و سطح دوم، $\sum_{j=1}^{J} n_j = N$ تعداد کل مشاهدات، β_0 پارامتر ثابت بیانگر عرض از مبدا، و u_jها و ϵ_{ij}ها مستقل و دارای توزیع نرمال با میانگین صفر و واریانس‌های به ترتیب σ_u^2 و σ_e^2 هستند. مدل (۱) مدل عرض از مبدا تصادفی با متغیر پاسخ پیوسته y_{ij} نامیده می‌شود. نمادگذاری مدل‌های چندسطحی در این مقاله بر اساس نمادگذاری‌های گلداستاین (۱۹۹۹) است که علاوه بر مطالعه آن‌ها برآورد بسامدی پارامترهای مدل را نیز ارائه کرده است. برآورد بیزی پارامترهای این مدل و مدل‌های چندسطحی دیگر در براون (۲۰۰۹) آمده است.

[۶] Parameter expantion

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

با پیروی از سرجنت و همکاران (۲۰۰۰) و به‌منظور به‌کارگیری روش SMCMC قرار داده می‌شود $\beta_{\circ j} = \beta_\circ + u_j$. این رابطه را می‌توان به صورت $-\beta_{\circ j} + \beta_\circ + u_j = \circ$ نیز نوشت. با توجه به این تساوی و رابطه (۱) داریم

$$\begin{aligned} y_{ij} &= \beta_{\circ j} + \epsilon_{ij} \\ \circ &= -\beta_{\circ j} + \beta_\circ + u_j. \end{aligned} \qquad (۲)$$

حال به ازای $i = ۱,\ldots,n_j$ و $j = ۱,\ldots,J$ رابطه (۲) را می‌توان به‌صورت برداری

$$\begin{pmatrix} y_{۱۱} \\ y_{۲۱} \\ \vdots \\ y_{n_J J} \\ 0_J \end{pmatrix} = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \ldots & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \ldots & 0_{n_2} \\ \vdots & \ddots & \vdots & \vdots \\ 0_{n_J} & \ldots & 1_{n_J} & 0_{n_J} \\ & -I_J & & 1_J \end{pmatrix} \begin{pmatrix} \beta_{\circ ۱} \\ \vdots \\ \beta_{\circ J} \\ \beta_\circ \end{pmatrix} + \begin{pmatrix} \epsilon_{۱۱} \\ \vdots \\ \epsilon_{n_J J} \\ u_۱ \\ \vdots \\ u_J \end{pmatrix} \qquad (۳)$$

نوشت، که در آن 0_q بردار ستونی q بعدی از صفرها، 1_q بردار ستونی q بعدی از یک‌ها و I_J ماتریس همانی با بعد J است. واضح است برابری (۳) را می‌توان به‌صورت ماتریسی

$$Y = X\Theta + E \qquad (۴)$$

نوشت، که در آن $Y = (y_{۱۱},\ldots,y_{n_J J},0_J)^T$ بردار متغیرهای وابسته،

$$X = \begin{pmatrix} diag\{1_{n_1},\ldots,1_{n_J}\} & 0_N \\ -I_J & 1_J \end{pmatrix}$$

ماتریس طرح، $\Theta = (\beta_{\circ ۱},\ldots,\beta_{\circ J},\beta_\circ)^T$ بردار پارامترهای مدل و $E = (\epsilon_{۱۱},\ldots,\epsilon_{n_J J},u_۱,\ldots,u_J)^T = (\epsilon,u)^T$ بردار خطاها در ساختار جدید هستند. در ادامه رابطه رگرسیونی جدید (۴) مدل جدید یا مدل ساختاری نامیده می‌شود. واضح است بردار خطاهای E در رابطه (۴) دارای توزیع نرمال چندمتغیری با بردار میانگین 0_{N+J} و ماتریس کواریانس $\Sigma_{N+J} = \Sigma$

$$\Sigma = \begin{pmatrix} Cov(\epsilon) & 0_{N\times J} \\ 0_{J\times N} & Cov(u) \end{pmatrix} = \begin{pmatrix} \sigma_e^۲ I_N & 0_{N\times J} \\ 0_{J\times N} & \sigma_u^۲ I_N \end{pmatrix}.$$

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

است (رنچر، ۱۹۹۵). به‌کمک اطلاعات موجود و در نظر گرفتن پیشین ناآگاهی بخش تخت برای بردار $J+1$ بعدی پارامتر Θ، یعنی $\pi(\Theta) \propto 1$، توزیع شرطی کامل آن عبارت است از

$$\Theta\Big|(Y,X,\Sigma) \sim MVN\Big((X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}Y),(X^T\Sigma^{-1}X)^{-1}\Big). \qquad (۵)$$

به کمک الگوریتم نمونه‌گیری گیبس می‌توان نمونه‌های تصادفی از توزیع شرطی کامل Θ تولید کرد و از طریق آن به استنباط بیزی راجع به پارامترهای β_0 و $\beta_{0J},\ldots,\beta_{01}$ یا به طور معادل $u_J,\ldots,u_1$ پرداخت.

روش SMCMC تنها تغییر اساسی را در به‌روز رسانی Θ اعمال می‌کند. در حالی‌که به‌روز نمودن σ_e^2 و σ_u^2، بدون تغییر و با استفاده از روش‌های معمول نمونه‌گیری گیبز و از توزیع شرطی کامل آن‌ها که گاما یا وارون گاما است انجام می‌شود (براون، ۲۰۰۹). به عنوان مثال برای مدل (۱) با انتخاب پیشین $\Gamma(a_u,b_u)$ و $\Gamma(a_e,b_e)$ به ترتیب برای $1/\sigma_u^2$ و $1/\sigma_e^2$ توزیع‌های شرطی کامل به ترتیب $\Gamma(J/2+a_u, b_u+\sum_{j=1}^{J}u_j^2/2)$ و $\Gamma(N/2+a_e, b_e+\sum_{ij}e_{ij}^2/2)$ خواهند بود، که در آن‌ها $e_{ij}=y_{ij}-\beta_0-u_j$. این موضوع برای مدل‌های دیگری که در ادامه مطالعه می‌شوند نیز صادق است جز این که تغییرات جزئی در پارامترهای توزیع‌های گاما رخ خواهد داد. به منظور جلوگیری از افزایش حجم مطالب مقاله از بیان جزئیات بیشتر راجع به انتخاب توزیع‌های پیشین و تاثیر آن‌ها بر توزیع‌های پسین و شرطی کامل پارامترهای درگیر مدل خودداری می‌شود. خواننده علاقه‌مند به مطالعه بیشتر در این زمینه به براون و درایپر (۲۰۰۰) و براون (۲۰۰۹) ارجاع داده می‌شود.

در روش‌های معمول MCMC پارامترهای β_0، $u_J,\ldots,u_1$ مجزا به‌روز می‌شوند و چون در طی تولید زنجیر بین آن‌ها همبستگی به‌وجود می‌آید روش معمول باعث تاخیر در زمان همگرایی می‌شود (براون، ۲۰۰۹). لذا اخذ نمونه‌های تقریبا مستقل از توزیع‌های پسین آن‌ها مستلزم اجرای الگوریتم MCMC برای مدت طولانی است تا بتوان از تاثیر همبستگی بین نمونه‌ها کاست. اما نکته حائز اهمیت در به‌کارگیری روش SMCMC این است که به‌دلیل به‌روز شدن توام پارامترهای β_0 و $\beta_{01},\ldots,\beta_{0J}$ در یک بلوک زمان همگرایی الگوریتم MCMC کاهش می‌یابد

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

(سرجنت و همکاران، ۲۰۰۰). به عبارتی دیگر با به‌روز نمودن توام این پارامترها همبستگی درونی بین آن‌ها در کل زنجیر کاهش می‌یابد و لذا زنجیر بهتر ترکیب می‌شود. واضح است که این امر خود موجب افزایش اندازه نمونه موثر[۷] (ESS) می‌شود. این نکته در ادامه در مثال‌های شبیه‌سازی و کاربردی منعکس خواهد شد.

تعمیم مدل (۱)، مدل عرض از مبدا تصادفی همراه با حضور متغیرهای مستقل است. مدل چند سطحی عرض از مبدا تصادفی با حضور تنها یک متغیر مستقل x_{ij} و با متغیر پاسخ پیوسته y_{ij} عبارت است از

$$y_{ij} = \beta_0 + u_j + x_{ij}\beta_1 + \epsilon_{ij} \qquad (۶)$$

فرض‌های مدل (۶) همان فرض‌های مدل (۱) می‌باشد جز این‌که در مدل اخیر پارامتر ثابت اضافی β_1 بیانگر شیب ثابت مدل است (گلداستاین، ۱۹۹۹).

برای بکارگیری روش SMCMC ابتدا رابطه (۶) را به‌صورت دو معادله

$$\begin{aligned} y_{ij} &= \beta_{0j} + x_{ij}\beta_1 + \epsilon_{ij} \\ 0 &= -\beta_{0j} + \beta_0 + u_j. \end{aligned}$$

نوشته و سپس به ازای کلیه مشاهدات آن‌را به‌صورت برداری

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{n_J J} \\ 0_J \end{pmatrix} = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \dots & 0_{n_1} & x_1 & 0_{n_1} \\ 0_{n_2} & 1_{n_2} & \dots & 0_{n_2} & x_2 & 0_{n_2} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0_{n_J} & \dots & 1_{n_J} & 0_{n_J} & x_J & 0_{n_J} \\ & -I_J & & & 0_J & 1_J \end{pmatrix} \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0J} \\ \beta_1 \\ \beta_0 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_J J} \\ u_1 \\ \vdots \\ u_J \end{pmatrix}$$

می‌نویسیم، که در آن $x_q = (x_{1q}, x_{2q}, \dots, x_{n_q q})^T$. واضح است این تساوی نیز می‌تواند به صورت ماتریسی $Y = X\Theta + E$ با ماتریس طرح جدید نوشته شود. لذا توزیع شرطی کامل Θ همان توزیع (۵) خواهد بود، جز اینکه ماتریس طرح X تغییر خواهد کرد. به عنوان مثال، بنا به مدل جدید و رابطه (۵) معکوس ماتریس

[۷] Effective Sample Size

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

کواریانس توزیع شرطی Θ به شرط سه‌تایی (X, Y, Σ) به‌صورت زیر خواهد بود

$$X^T\Sigma^{-1}X = \begin{pmatrix} \frac{1}{\sigma_u^2}+\frac{n_1}{\sigma_e^2} & \cdots & 0 & \frac{\sum x_{i1}}{\sigma_e^2} & \frac{-1}{\sigma_u^2} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \frac{1}{\sigma_u^2}+\frac{n_J}{\sigma_e^2} & \frac{\sum x_{iJ}}{\sigma_e^2} & \frac{-1}{\sigma_u^2} \\ \frac{\sum x_{i1}}{\sigma_e^2} & \cdots & \frac{\sum x_{iJ}}{\sigma_e^2} & \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} x_{ij}^2}{\sigma_e^2} & 0 \\ \frac{-1}{\sigma_u^2} & \cdots & \frac{-1}{\sigma_u^2} & 0 & \frac{J}{\sigma_u^2} \end{pmatrix}. \quad (۷)$$

مجددا برآورد بردار پارامترهای $\Theta = (\beta_{01}, \ldots, \beta_{0J}, \beta_0)^T$ به‌کمک نمونه‌گیری گیبز از توزیع (۵) با ماتریس طرح جدید X به‌دست می‌آید. همچنین σ_e^2 و σ_u^2 با روش معمول نمونه‌گیری گیبس و به کمک نمونه‌های تولید شده از توزیع وارون گاما برآورد می‌شوند.

به‌منظور به‌کارگیری الگوریتم $SMCMC$ برای مدل‌های عرض از مبدا تصادفی و همچنین مدل‌های چندسطحی دیگر ضروری است معادله مدل به صورت معادله رگرسیونی $Y = X\Theta + E$ نوشته شود. سپس به‌کمک رابطه (۵) برآورد تعدادی از پارامترهای مدل به‌کمک روش‌های شبیه‌سازی MCMC انجام می‌گیرد. در روش SMCMC به دلیل تعریف مدل ساختاری همبستگی بین پارامترهای ثابت و خطاهای مربوط به سطوح بالا کم خواهد شد. اما نکته مهم روش SMCMC این است که معمولا بعد ماتریس کواریانس مدل جدید خیلی زیاد است و به دلیل این‌که به‌روز نمودن این ماتریس در هر گام از الگوریتم MCMC ضروری است، به‌کارگیری این روش زمان همگرایی زنجیر را افزایش خواهد داد. نکته دیگر این‌که با افزایش تعداد گروه‌ها و همچنین ورود متغیرهای مستقل و لحاظ نمودن مدل‌های مختلف چندسطحی بعد ماتریس $(X^T\Sigma^{-1}X)^{-1}$ بزرگ‌تر شده و محاسبه آن حجم وسیعی از محاسبات مربوط به تولید نمونه برای پارامترها را به خود اختصاص خواهد داد.

در مقاله حاضر، برای رفع مشکلاتی شبیه آن‌چه در بالا به آن‌ها اشاره شد دو روش پیشنهاد و کارایی آن‌ها نسبت به روش SMCMC مورد ارزیابی قرار می‌گیرد. روش‌های پیشنهادی را بهبودهای روش SMCMC می‌نامیم. بیان این نکته ضروری است که توجه این مقاله تنها معطوف به مدل عرض از مبدا تصادفی بوده و مدل‌های دیگر چندسطحی مثل مدل شیب تصادفی مورد مطالعه قرار نگرفته است. یکی از

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

دلایل این امر عدم بهبود مناسب روش‌های پیشنهادی برای مدل‌های دیگر شامل مدل شیب تصادفی بوده است. اما امید می‌رود در تحقیقات آتی این موضوع نیز با راهکارهای دیگر مورد مطالعه قرار گیرد.

۳ بهبود روش ساختاری مونت کارلوی زنجیر مارکوف

به دلیل افزایش بیش از حد بعد ماتریس کواریانس به روش SMCMC انتظار می‌رود دقت برآورد پارامترها در مسائل کاربردی کاهش یابد. این امر از تقریب وارون ماتریس کواریانس و ترکیب آن با ماتریس طرح ناشی می‌شود. یک راه حل مناسب برای بهبود روش SMCMC استفاده از تجزیه ماتریس است.

روش‌های مختلفی برای تجزیه یک ماتریس وجود دارد که هر کدام دارای محاسن و معایب خاص خود هستند. اما تجزیه چولسکی به دلیل سرعت بالای محاسبات نسبت به تجزیه‌های دیگر ارجحیت دارد (بیکر، ۲۰۰۳). اگر A ماتریسی متقارن معین مثبت باشد، تجزیه چولسکی آن به صورت $A = LL^T$ است که در آن L ماتریسی بالا مثلثی است. توجه کنید برای مدل‌های این مقاله شاید تجزیه‌های دیگر ساختار مناسب‌تری برای معکوس ماتریس کواریانس ارائه دهند ولی بنا به بیکر (۲۰۰۳) استفاده از آن‌ها بهبود قابل ملاحظه‌ای در دقت و زمان همگرایی روش SMCMC ایجاد نخواهند کرد.

بیان این نکته ضروری است که برای پیاده‌سازی روش‌های بهبود یافته SMCMC، ابتدا با نرم‌افزار Maple نسخه ۱۱ تجزیه چولسکی معکوس ماتریس‌های مورد نیاز محاسبه و سپس مدل‌بندی و دستیابی به برآوردهای مورد نیاز با نرم‌افزار R نسخه ۲.۹.۰، انجام شده‌اند. همچنین برای برازش مدل‌های چندسطحی تعدادی از کتابخانه‌های ضروری شامل lme4، MASS و nlme در نرم‌افزار آماری R به خدمت گرفته شده است.

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

۱.۳ تجزیه چولسکی وارون ماتریس کواریانس پارامترهای بلوک بندی‌شده

در این بخش برای بهبود روش SMCMC، از تجزیه چولسکی ماتریس کواریانس استفاده می‌شود. از آنجا که محاسبه ماتریس کواریانس پارامترهای بلوک‌بندی شده زمان‌بر است، با استفاده از تجزیه چولسکی مشکل افزایش بعد ماتریس کواریانس مدل ساختاری و تاثیر آن در زمان همگرایی روش SMCMC مرتفع می‌شود. روش پیشنهادی برای دو مدل مطرح شده در بخش قبل به کار گرفته خواهد شد.

با توجه به مدل عرض از مبدا تصادفی (۱) و بنا به رابطه (۵) تجزیه چولسکی وارون ماتریس کواریانس مدل ساختاری یعنی $X^T\Sigma^{-1}X$ عبارت است از

$$\Sigma^*\Sigma^{*T} = \begin{pmatrix} aI_{J+1} & \frac{-1}{a\sigma_u^2} \\ 0_J & \sqrt{\frac{n_J}{n\sigma_u^2+\sigma_e^2}} \end{pmatrix}^T \begin{pmatrix} aI_{J+1} & \frac{-1}{a\sigma_u^2} \\ 0_J & \sqrt{\frac{n_J}{n\sigma_u^2+\sigma_e^2}} \end{pmatrix} \qquad (۸)$$

که در آن

$$a = \sqrt{\frac{n\sigma_u^2+\sigma_e^2}{\sigma_u^2\sigma_e^2}}.$$

با توجه به این تساوی و رابطه (۵) به نظر می‌رسد محاسبات مربوط به اجرای الگوریتم MCMC خیلی ساده‌تر از حالت معمول آن (عدم استفاده از تجزیه چولسکی) باشد. به ویژه با جایگذاری تجزیه چولسکی $\Sigma^*\Sigma^{*T}$ به جای $X^T\Sigma^{-1}X$ در رابطه (۵)، ماتریس کواریانس توزیع شرطی کامل Θ به‌صورت $(\Sigma^{*T})^{-1}(\Sigma^*)^{-1}$ خواهد بود. بنابراین توزیع شرطی کامل Θ به‌صورت

$$\Theta\Big|(Y,X,\Sigma) \sim MVN\Big((\Sigma^*\Sigma^{*T})^{-1}(X^T\Sigma^{-1}Y),(\Sigma^*\Sigma^{*T})^{-1}\Big)$$

تغییر خواهد کرد، که با توجه به رابطه (۸) محاسبه واریانس ساده‌تر خواهد شد. از آن‌جا که عبارت $X^T\Sigma^{-1}Y$ برابر حاصل‌ضرب تعدادی مقادیر معلوم در $1/\sigma_e^2$ است، با محاسبه مقادیر ثابت آن خارج از تکرار الگوریتم MCMC می‌توان زمان همگرایی را بیش از پیش کاهش داد. به‌عنوان مثال برای مدل عرض از مبدا تصادفی و بدون حضور متغیر مستقل (رابطه (۱) را ببینید) کمیت $X^T\Sigma^{-1}Y$ عبارت است از

$$\frac{1}{\sigma_e^2}\Big(\sum_{i=1}^{n_1} y_{ij}, \sum_{i=1}^{n_2} y_{ij}, \ldots, \sum_{i=1}^{n_J} y_{ij}\Big)^T.$$

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

حال نحوه به کارگیری روش SMCMC در مدل عرض از مبدا تصادفی و با حضور یک متغیر تبیینی تشریح می‌شود. واضح است روش‌های پیشنهادی برای وقتی که تعداد متغیرهای تبیینی بیشتر باشند نیز قابل تعمیم است. توجه کنید تجزیه چولسکی ماتریس $X^T\Sigma^{-1}X$ در مدل چندسطحی عرض از مبدا تصادفی و با حضور متغیر تبیینی در مدل‌هایی که تعداد واحدهای درون سطوح نابرابر باشند دارای فرم استانداردی نیست. به عبارتی دیگر تجزیه چولسکی بسته به تعداد واحد هر گروه دارای ساختار متفاوتی است. لذا نمی‌توان یک دستورالعمل سرراستی را برای چنین مدل‌هایی پیشنهاد نمود. از اینرو برای آشنایی بیشتر با نحوه اجرای روش SMCMC و در عین حال سادگی محاسبات آتی واحدهای سطوح متفاوت مدل چندسطحی عرض از مبدا تصادفی و با حضور متغیر تبیینی به‌صورت برابر در نظر گرفته می‌شود. چنین طرحی به طرح متعادل معروف است.

وارون ماتریس کواریانس مدل چندسطحی عرض از مبدا تصادفی و با حضور متغیر مستقل در (۷) آمده است. با لحاظ نمودن فرض طرح متعادل، تجزیه چولسکی ماتریس (۷) عبارت است از

$$\Sigma^*\Sigma^{*T} = \begin{pmatrix} f_\circ I_J & & & 0_{J\times 2} & \\ \frac{\sum x_{i1}}{\sqrt{f_\circ}\sigma_e^2} & \cdots & \frac{\sum x_{iJ}}{\sqrt{f_\circ}\sigma_e^2} & f_1 & \circ \\ \frac{-1}{\sqrt{f_\circ}\sigma_u^2} & \cdots & \frac{-1}{\sqrt{f_\circ}\sigma_u^2} & f_3 & f_2 \end{pmatrix} \times \begin{pmatrix} f_\circ I_J & & & 0_{J\times 2} & \\ \frac{\sum x_{i1}}{\sqrt{f_\circ}\sigma_e^2} & \cdots & \frac{\sum x_{iJ}}{\sqrt{f_\circ}\sigma_e^2} & f_1 & \circ \\ \frac{-1}{\sqrt{f_\circ}\sigma_u^2} & \cdots & \frac{-1}{\sqrt{f_\circ}\sigma_u^2} & f_3 & f_2 \end{pmatrix}^T$$

که در آن $f_\circ$، f_1، f_2 و f_3 به‌صورت زیر تعریف می‌شوند

$$f_\circ^2 = (n\sigma_u^2 + \sigma_e^2)/(\sigma_u^2\sigma_e^2),$$

$$f_1^2 = \left(\sigma_e^2 \sum x_{ij}^2 + (n-1)\sum x_{ij}^2\sigma_u^2 - 2\sigma_u^2 \sum x_{ij}x_{mj}\right)/(f_\circ\sigma_e^2),$$

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

$$f_{۲}^{۲} = \frac{\left((nm-۱)\sigma_e^{۲} + nm(n-۱)\sigma_u^{۲}\right)\sum x_{ij}^{۲} - ۲(nm\sigma_u^{۲}+\sigma_e^{۲})\sum x_{ij}x_{mj}}{f_{۰}\sigma_e^{۲}\sigma_u^{۲}\sigma_e^{۲}\sum x_{ij}^{۲} + (n-۱)\sum x_{ij}^{۲}\sigma_u^{۲} - ۲\sigma_u^{۲}\sum x_{ij}x_{mj}},$$

$$f_{۳}^{۲} = \frac{\sum x_{ij}\sqrt{\sigma_e^{۲}}}{\left(f_{۰}\sigma_u^{۲}\sigma_e^{۲}\sigma_e^{۲}\sum x_{ij}^{۲} + (n-۱)\sum x_{ij}^{۲}\sigma_u^{۲} - ۲\sigma_u^{۲}\sum x_{ij}x_{mj}\right)}.$$

فرآیند به‌کارگیری این تجزیه در به‌روز کردن پارامتر Θ مشابه وضعیتی است که برای مدل عرض از مبدا تصادفی بدون حضور متغیر تبیینی توضیح داده شد. تنها تغییر جزئی ورود متغیر تبیینی x_{ij} در مدل و در نتیجه در تجزیه چولسکی است. با این حال همان‌گونه که روابط مربوط به ساختار تجزیه چولسکی برای مدل عرض از مبدا تصادفی و با حضور یک متغیر تبیینی نشان می‌دهد با اضافه نمودن متغیرهای مستقل دیگر تجزیه چولسکی وضعیتی به مراتب پیچیده‌تر به خود خواهد گرفت. لذا اعمال آن در فرآیند به‌روز کردن پارامترها تأثیر بسزایی در کاهش زمان همگرایی الگوریتم MCMC نخواهد داشت.

روش پیشنهادی به کارگرفته شده در این بخش را روش SMCMC1 نامیده و با یک مثال کاربردی و مطالعه شبیه‌سازی دقت برآوردهای حاصل و همچنین زمان همگرایی سیستم با استفاده از این روش با روش‌های دیگر مقایسه می‌شوند. اما قبل از آن روش پیشنهادی دوم این مقاله که در مورد نحوه به کارگیری روش تجزیه چولسکی ماتریس کواریانس مدل اولیه برای دو مدل عرض از مبدا و عرض از مبدا در حضور متغیر مستقل است مورد مطالعه قرار می‌گیرد.

۲.۳ تجزیه چولسکی ماتریس کواریانس مدل ساختاری

پیشنهاد دیگر برای بهبود روش SMCMC استفاده از تجزیه چولسکی برای ماتریس کواریانس مدل ساختاری است. به عبارتی دیگر در این روش ابتدا تجزیه چولسکی معکوس ماتریس کواریانس مدلی که به صورت رابطه (۴) نوشته شده باشد را محاسبه و سپس با دخالت آن در رابطه (۵) مدلی ساده‌تر به‌دست آورده می‌شود. این روش را SMCMC2 نامگذاری می‌نماییم.

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

برای مدل عرض از مبدا تصادفی (۱) ماتریس کواریانس مدل ساختاری (۴) ماتریسی قطری متشکل از N تا σ_e^2 و J تا σ_u^2 است. لذا ماتریس بالا مثلثی حاصل از تجزیه چولسکی، که با Σ_1 نمایش داده می‌شود، ماتریسی قطری بترتیب شامل N مقدار σ_e^{-1} و J مقدار σ_u^{-1} است. حال با جای‌گذاری این تجزیه در عبارت $X^T\Sigma^{-1}X$ نمونه‌گیری از توزیع شرطی کامل Θ به‌صورت

$$\Theta\Big|(Y,X,\Sigma) \sim MVN\Big((X^T\Sigma_1\Sigma_1^TX)^{-1}X^T\Sigma_1^T\Sigma_1Y,(X^T\Sigma_1\Sigma_1^TX)^{-1}\Big). \quad (۹)$$

خواهد بود. اکنون پیش‌بینی می‌شود با اختیار متغیر جدید $Z = X\Sigma_1^{-1}$ روش SMCMC، بهبود بیشتری نسبت به روش SMCMC1 داشته باشد. به‌کمک ماتریس جدید، میانگین و واریانس توزیع شرطی کامل Θ در (۹) به‌ترتیب برابر $(Z^TZ)^{-1}Z^T\Sigma_1Y$ و $(Z^TZ)^{-1}$ خواهند بود. به عبارتی دیگر به‌منظور به‌روز کردن Θ ضروری است از توزیع نرمال چندمتغیره با میانگین $(Z^TZ)^{-1}Z^T\Sigma_1Y$ و واریانس $(Z^TZ)^{-1}$ نمونه‌گیری کرد. لذا محاسبه Z و ذخیره‌سازی آن به جای محاسبه مستقیم $(X^T\Sigma_1\Sigma_1^TX)^{-1}$ بر سرعت همگرایی خواهد افزود.

شکل ماتریسی تجزیه چولسکی ماتریس کواریانس به روش SMCMC2 در مدل عرض از مبدا تصادفی و در حضور تنها یک متغیر تبیینی مانند مدل عرض از مبدا تصادفی و بدون حضور متغیر مستقل است. تنها تفاوت این دو مدل در ماتریس طرح آن‌ها خواهد بود. در این حالت اگر تجزیه چولسکی ماتریس کواریانس مدل ساختاری مربوطه با Σ_* نشان داده شود رابطه (۵) به‌صورت

$$\Theta\Big|(Y,X,\Sigma) \sim MVN\Big((X^T\Sigma_*\Sigma_*^TX)^{-1}X^T\Sigma_*\Sigma_*^TY,(X^T\Sigma_*\Sigma_*^TX)^{-1}\Big).$$

تغییر خواهد کرد. همان‌طور که ملاحظه می‌شود تعدادی از کمیت‌های موجود در ماتریس کواریانس در بردار میانگین نیز ظاهر شده‌اند. لذا با ذخیره اولیه تعدادی از کمیت‌های ثابت که در مراحل شبیه‌سازی SMCMC نیازی به به‌روز سازی آن‌ها نیست بر سرعت الگوریتم SMCMC افزوده خواهد شد.

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

۴ تحلیل یک مثال کاربردی و مطالعه شبیه‌سازی

در این بخش ابتدا یک مثال واقعی مورد مطالعه قرار خواهد گرفت. سپس با شبیه‌سازی داده‌هایی که تا حدودی مشابه مثال کاربردی بخش ابتدایی هستند تحلیل آماری مربوطه صورت می‌گیرد. به ویژه اجرای الگوریتم‌های پیشنهادی با روش‌های موجود قبلی مورد مقایسه قرار می‌گیرد.

۱.۴ مثال کاربردی

پایگاه http://www.emgo.nl/quality-of-our-research/research-tools/multilevel شامل داده‌هایی مربوط به ۴۴۱ بیمار در محدوده سنی ۴۴ تا ۸۶ سال است که تحت درمان ۱۲ پزشک بوده و میزان کلسترول خون آن‌ها بر حسب میلی مول در لیتر ثبت شده است. بنا به نمادگذاری مدل چندسطحی گروه‌بندی بیماران بر اساس پزشکان است. از آن جا که از نقطه نظر کلسترول خون بیماران تحت پزشک خاص تا حدودی با هم شباهت دارند، برای تحلیل مشاهدات نمی‌توان از روش‌های متداول رگرسیونی که همبستگی بین بیماران را نادیده می‌گیرد استفاده کرد. از این رو از مدل دو سطحی، که سطح اول بیماران تحت درمان و سطح دوم پزشکان هستند، استفاده می‌شود. با توجه به داده‌ها تعداد بیماران هر گروه، یعنی n_j، به ازای $j = ۱, \ldots, ۱۲$ به ترتیب عبارتند از ۳۶، ۳۶، ۳۹، ۳۶، ۳۶، ۳۹، ۳۶، ۳۶، ۳۹، ۳۶، ۳۶ و ۳۶. این داده‌ها با ترکیب‌های مختلفی از مدل‌های چندسطحی با روش کمترین توان‌های دوم تعمیم یافته بازگشتی مورد تحلیل قرار گرفت (توایسک، ۲۰۰۶). اما برآورد بیزی پارامترهای مدل و به‌ویژه با روش MCMC تعیین نشده است.

در ادامه ابتدا پارامترهای مدل چندسطحی با استفاده از روش‌های MCMC، SMCMC و بهبودهای روش SMCMC برآورد خواهند شد. سپس زمان همگرایی الگوریتم در روش‌های متفاوت مورد بررسی قرار خواهد گرفت. هنگام اجرای روش‌های MCMC، SMCMC و بهبودهای آن زنجیر ۱۰۰۰۰۰ بار تکرار کرده و برای محاسبه میانگین‌های توزیع‌های شرطی کامل، به‌عنوان برآورد پارامترها، گام داغیدن برابر ۱۰۰۰ در نظر گرفته شده است. همچنین توزیع‌های پیشین پارامترهای

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

مدل به‌صورت

$$\pi(\beta_1) \propto 1, \quad \pi(\beta_0) \propto 1, \quad \pi(1/\sigma_u^2) \sim \Gamma(a_u, b_u), \quad \pi(1/\sigma_e^2) \sim \Gamma(a_e, b_e)$$

اختیار شده است به‌طوری که مقادیر ابر پارامترها همان مقادیر پیشنهاد شده توسط براون و درایپر (۲۰۰۰) و براون (۲۰۰۹) یعنی $a_u = a_e = b_u = b_e = 0.001$ هستند. توجه شود که انتخاب پیشین‌های یکنواخت برای پارامترهای اثر ثابت منجر به تساوی برآوردگرهای بیزی و ماکسیمم درستنمایی خواهد شد. برای برآورد پارامترها در مدل‌های چندسطحی ابتدا پارامترهای مدل ساده عرض از مبدا تصادفی و بدون حضور متغیر مستقل برآورد می‌شوند. سپس مدل عرض از مبدا تصادفی و با حضور متغیر مستقل سن مورد بررسی قرار می‌گیرد.

برآورد پارامترها همراه با خطای برآورد عرض از مبدا در یک مدل دوسطحی عرض از مبدا تصادفی در جدول ۱ آمده است. برآورد پارامترهای مدل عرض از مبدا تصادفی با روش‌های MCMC، SMCMC، SMCMC1 و SMCMC2 محاسبه شده است. تغییرات کم برآورد β_0 حاکی از دقت روش‌های برآورد است. نتایج نشان می‌دهند که با استفاده از روش‌های متفاوت، تفاوت زیادی بین برآورد پارامترها وجود ندارد. اما در ادامه نشان داده خواهد شد که روش پیشنهادی از دو مزیت عمده برخوردار است.

جدول ۱: برآورد و خطای استاندارد برآورد پارامترهای مدل عرض از مبداء تصادفی با روش‌های مختلف MCMC.

$\widehat{\sigma}_e^2$	$\widehat{\sigma}_{u_0}^2$	$\widehat{\beta}_0(SE(\widehat{\beta}_0))$	روش
۰/۵۴۱	۰/۵۲۲	۵/۹۰۰ (۰/۰۷۲)	$MCMC$
۰/۵۳۶	۰/۵۲۳	۵/۹۸۱ (۰/۰۸۱)	$SMCMC$
۰/۵۳۶	۰/۵۲۸	۶/۰۲۹ (۰/۰۸۲)	$SMCMC1$
۰/۵۳۶	۰/۵۲۳	۵/۹۸۱ (۰/۰۸۱)	$SMCMC2$

تجزیه چولسکی ماتریس کواریانس ساختاری برای مدل عرض از مبدا تصادفی و در حضور متغیر تبیینی در حالتی که تعداد واحدها در گروه‌های متفاوت یکسان نباشد روند منظمی را دنبال نمی‌کند. اما برای انجام یک مقایسه مناسب

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

تعداد مشاهدات هر گروه ۳۶ در نظر گرفته شد و برآورد پارامترهای مدل به روش SMCMC1 انجام شده است. قابل ذکر است متغیر مستقل مدل سن بیماران تحت درمان است. نتایج آن همراه با روش‌های MCMC، SMCMC و SMCMC2 در جدول ۲ ارائه شده است. نتایج نشان می‌دهد بین برآورد پارامترها با روش‌های متفاوت تنها تغییرات جزئی وجود دارد. اما برآورد پارامترها در این مدل با برآورد پارامترها در مدل عرض از مبدا تصادفی و بدون حضور متغیر مستقل متفاوت است. به علاوه مشاهده می‌شود مقدار عرض از مبدا (β_0) نسبت به قبل کاهش چشمگیری داشته و در واریانس‌های مدل نیز تغییراتی ایجاد شده است. همان‌گونه که در

جدول ۲: برآورد و خطای استاندارد برآورد پارامترهای مدل عرض از مبداء تصادفی با حضور متغیر تبیینی سن با روش‌های مختلف MCMC.

روش	$\widehat{\beta}_0(SE(\widehat{\beta}_0)\times 10^5)$	$\widehat{\beta}_1(SE(\widehat{\beta}_1))$	$\widehat{\sigma}^2_{u_0}$	$\widehat{\sigma}^2_e$
$MCMC$	۲/۹۲۹(۹/۸۱۲)	۰/۰۴۹(۰/۰۷۷)	۰/۴۹۱	۰/۳۲۵
$SMCMC$	۲/۹۴۸(۳/۲۴۱)	۰/۰۴۹(۰/۰۹۲)	۰/۴۹۳	۰/۳۳۴
$SMCMC1$	۲/۹۳۰(۹/۴۳۹)	۰/۰۴۹(۰/۰۷۸)	۰/۴۹۴	۰/۳۳۰
$SMCMC2$	۲/۹۰۱(۱/۱۵۱)	۰/۰۵۰(۰/۰۸۸)	۰/۴۹۸	۰/۳۳۲

دو مدل مورد بررسی ملاحظه شد، روش SMCMC و بهبودهای آن برآوردهایی حدودا مشابه روش‌های موجود قبلی ارائه می‌نمایند. اما نکته قابل تامل در مورد روش‌های SMCMC و بهبودهای آن نتایج حاصل از مقدار ESS و میزان زمان مصرفی سیستم (به ثانیه) در ارائه برآورد پارامترهای مدل‌های چندسطحی متفاوت شامل مدل عرض از مبدا تصادفی و عرض از مبدا تصادفی در حضور متغیر تبیینی سن است. جدول ۳ مقدار ESS و میزان زمان مصرفی سیستم در دستیابی به برآورد پارامترهای مدل عرض از مبدا تصادفی و بدون حضور متغیر مستقل به روش‌های MCMC، SMCMC، SMCMC1 و SMCMC2 را نشان می‌دهد. همان‌گونه که ملاحظه می‌شود هم زمان محاسبه برآورد پارامترهای مدل مورد نظر با روش تجزیه چولسکی ماتریس کواریانس به روش SMCMC2 خیلی کمتر است و هم ESS آن از همه بیشتر است. گرچه زمان کاربر در روش SMCMC2 تنها

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

بیشتر از روش SMCMC است، اما این خسارت ناشی از تعدادی دستورات اضافی ضروری در این روش بوده و زیاد چشمگیر نمی‌باشد.

جدول ۳: مقدار ESS و زمان دستیابی برآورد پارامترها در مدل عرض از مبداء تصادفی با روش‌های مختلف MCMC.

روش	ESS	زمان کاربر (به ثانیه)	زمان سیستم (به ثانیه)
$MCMC$	۳۷۴۲۱	۱۴۴/۰۷	۰/۱۱
$SMCMC$	۹۵۷۴۱	۱۰۸/۷۸	۰/۱۷
$SMCMC1$	۹۵۸۸۲	۱۹۹/۹۲	۰/۱۳
$SMCMC2$	۹۶۳۹۶	۱۱۰/۱۵	۰/۰۶

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

جدول ۴: مقدار ESS و زمان (برحسب ثانیه) دستیابی به برآورد پارامترها در مدل عرض از مبداء تصادفی با حضور متغیر تبیینی سن با روش‌های مختلف MCMC.

روش	ESS	زمان کاربر	زمان سیستم
$MCMC$	۲۵۸۷۶	۱۸۰/۰۷	۲۳/۴
$SMCMC$	۸۸۷۳۵	۱۲۴/۱۶	۰/۳۶
$SMCMC$۱	۸۸۴۳۲	۱۶۷/۶۱	۰/۱۶
$SMCMC$۲	۸۸۹۳۲	۱۵۲/۴۴	۰/۱۱

۲.۴ مطالعه شبیه‌سازی

در این بخش بر اساس داده‌های شبیه‌سازی از مدل‌های تحت بررسی، روش‌های پیشنهادی مقایسه می‌شوند. بنا به خلاصه‌های آماری مربوط به متغیرهای کلسترول خون و سن بیماران در بخش قبل، سعی شد داده‌های شبیه‌سازی تا حدودی مشابه آن‌ها باشند. با استفاده از ۴۴۱ بار شبیه‌سازی از توزیع‌های U(۳ , ۹) و $N(۶, ۱)$ برای متغیر وابسته و از توزیع‌های U(۴۴ , ۸۶) و $N(۶۵, ۸۱)$ برای متغیر تبیینی به تحلیل آماری روش‌های پیشنهادی پرداخته شد. برای اجرای روش‌های SMCMC و بهبودهای متناظر با آن زنجیر ۱۰۰۰۰۰ بار تکرار و در محاسبه میانگین‌های توزیع‌های شرطی کامل، به عنوان برآورد پارامترها، ۱۰۰۰ مرحله داغیدن در نظر گرفته شده‌اند. فرض‌های زیر بنایی مدل همانند بخش‌های قبل در نظر گرفته شده است. برآورد پارامترهای مدل دوسطحی عرض از مبدا تصادفی به روش‌های مختلف SMCMC و تحت شبیه‌سازی از دو توزیع مشروحه فوق در جدول ۵ آمده است.

همان طور که ملاحظه می‌شود، برآورد پارامترها تحت توزیع نرمال برای متغیر وابسته تفاوت چندانی با برآوردهای متناظر آن هنگام کار با داده‌های واقعی ندارد. با این حال در هنگام استفاده از توزیع یکنواخت برآورد واریانس خطای سطح اول خیلی کم شده است. این موضوع ناشی از تغییرپذیری زیاد مشاهدات هنگام استفاده از توزیع یکنواخت است. در نتیجه استواری برآوردها نسبت به تغییرپذیری توزیع‌ها کم است. نکته مهم شباهت نسبی برآوردها تحت روش‌های مختلف اما با فرض

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

جدول ۵: برآورد و خطای استاندارد برآورد پارامترهای مدل عرض از مبداء تصادفی با روش‌های مختلف MCMC.

توزیع	روش	$\widehat{\beta}_{\circ}(SE(\widehat{\beta}_{\circ}))$	$\widehat{\sigma}_u^2$	$\widehat{\sigma}_e^2$
یکنواخت	$SMCMC$	۶/۰۲۹ (۰/۰۴۷)	۰/۰۷۷	۲/۹۵۵
	$SMCMC1$	۶/۱۲۹ (۰/۰۴۷)	۰/۰۸۴	۲/۹۵۹
	$SMCMC2$	۶/۰۲۹ (۰/۰۴۶)	۰/۰۷۶	۲/۹۵۵
نرمال	$SMCMC$	۶/۰۱۰ (۰/۰۵۳)	۰/۰۵۸	۵/۰۶۴
	$SMCMC1$	۶/۰۷۸ (۰/۰۵۳)	۰/۰۱۹	۶/۰۷۸
	$SMCMC2$	۵/۹۸۲ (۰/۰۵۲)	۰/۰۱۶	۶/۰۶۴

یک توزیع خاص برای متغیرهای مورد مطالعه است. این موضوع مقایسه روش‌های برآورد براساس مقدار ESS و زمان مصرفی سیستم را میسر می‌سازد.

با فرض ثابت بودن شیب خط رگرسیون برای گروه‌های متفاوت متغیر تبیینی را وارد مدل نموده و با روش SMCMC و بهبود آن پارامترهای مدل برآورد شد. نتایج در جدول ۶ ارائه شده‌اند. همان‌طور که ملاحظه می‌شود بین برآورد پارامترها هنگام استفاده از توزیع‌های یکنواخت و نرمال و برآوردهای متناظر آن‌ها هنگام استفاده از داده‌های واقعی تفاوت عمده‌ای وجود دارد. بویژه برآورد واریانس خطای سطح اول در هر دو حالت کاهش یافته ولی مقدار آن در سطح دوم افزایش یافته است. به علاوه، مثل حالت مدل عرض از مبداء تصادفی، مقایسه برآوردهای حاصل از یک توزیع خاص، مثلا نرمال، نشانگر شباهت نسبی بین برآوردها و خطای برآورد است. در نتیجه می‌توان آن‌ها را بر اساس مقدار ESS و زمان مصرفی سیستم مورد مقایسه قرار داد.

علاوه بر بررسی دقت و نیرومندی برآوردها با استفاده از شبیه‌سازی مدل، لازم است میزان بهبود زمان اجرای الگوریتم‌های برآورد بر اساس روش‌های موجود مورد ارزیابی قرار گیرد. برای این منظور مقدار ESS، زمان‌های کاربر و سیستم برای هر کدام از وضعیت‌ها و مدل‌ها در جدول‌های ۷ و ۸ ارائه شده‌اند. نتایج حاکی از آن است که روش تجزیه چولسکی وارون ماتریس کواریانس مدل اولیه کمترین زمان سیستم (به ثانیه) در دستیابی به برآورد پارامترها را به خود اختصاص داده

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

است. به‌علاوه مقدار ESS حاصل از این روش در هر دو وضعیت شبیه‌سازی داده‌ها بیشترین مقدار ممکن را دارد. در نتیجه روش تجزیه چولسکی معکوس ماتریس کواریانس مدل اولیه روش برتر در بین روش‌های مورد بررسی است.

برای مدل عرض از مبدا تصادفی همراه با حضور یک متغیر تبیینی نیز مقدار ESS و مقایسه زمان سیستم در تولید نمونه‌های کارا مورد محاسبه قرار گرفت. همان‌گونه که در جدول ۸ ملاحظه می‌شود تجزیه چولسکی وارون ماتریس کواریانس مدل اولیه و سپس اجرای الگوریتم SMCMC از زمان مصرفی کمتری برخوردار است. به علاوه مقدار ESS حاصل از این روش در هر دو وضعیت شبیه‌سازی داده‌ها بیشترین مقدار ممکن را دارد. بنابراین اجرای الگوریتم SMCMC همراه با تجزیه چولسکی ماتریس کواریانس مدل ساختاری بهبود قابل ملاحظه‌ای در کاهش زمان مصرفی سیستم برای تولید نمونه‌های کارا اعمال می‌نماید. جالب اینکه کاهش زمان مصرفی متاثر از توزیع‌های آماری مربوط به متغیرهای مورد مطالعه نیست. اجرای شبیه‌سازی یک نتیجه فرعی دیگری را نیز ارائه نموده است. پایداری برآوردها شدیدا تحت تاثیر توزیع‌ها است. در حالی که با فرض نرمال بودن متغیرهای تبیینی و وابسته برآوردها مورد اعتمادند، اما این وضعیت با فرض توزیع یکنواخت از اعتبار لازم برخوردار نیست.

بحث و نتیجه‌گیری

با گسترش توانایی کامپیوتر برای برآورد بیزی پارامترها در مدل‌های چندسطحی محققین تعمیم‌های متفاوتی از الگوریتم MCMC را معرفی نمودند. یکی از آن روش‌ها روش SMCMC است. در این مقاله نشان داده شد که اگرچه روش SMCMC به‌کمک تعریف ساختار جدید باعث حذف همبستگی بین پارامترهای ثابت و خطای تصادفی مربوط به سطوح بالا می‌شود اما به‌دلیل افزایش بعد ماتریس کواریانس مدل ساختاری و به‌روز نمودن آن در هر گام MCMC زمان همگرایی روش SMCMC کاهش می‌یابد. برای رفع این مشکل دو روش بر اساس تجزیه چولسکی ماتریس کواریانس پیشنهاد شد. همچنین کارکرد روش‌های پیشنهادی بر

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

جدول ۶: برآورد و خطای استاندارد برآورد پارامترهای مدل عرض از مبداء تصادفی در حضور متغیر تبیینی با روش‌های مختلف MCMC.

توزیع	روش	$\widehat{\beta}_{\circ}(SE(\widehat{\beta}_{\circ}) \times ۱۰^{۵})$	$\widehat{\beta}_{۱}(SE(\widehat{\beta}_{۱}))$	$\widehat{\sigma}_{u}^{۲}$	$\widehat{\sigma}_{e}^{۲}$
	$SMCMC$	۶/۰۰۴ (۳/۲۴)	۰/۰۱۰ (۰/۱۵۵)	۰/۰۱۳	۲/۹۵۰
یکنواخت	$SMCMC۱$	۶/۰۰۶ (۵/۵۹)	۰/۰۱۰ (۰/۲۰۰)	۰/۰۱۲	۲/۹۴۹
	$SMCMC۲$	۶/۰۰۶ (۵/۵۹)	۰/۰۱۰ (۰/۲۴۲)	۰/۰۱۳	۲/۹۴۹
	$SMCMC$	۵/۵۰۸ (۲/۹۰)	۰/۰۰۷ (۰/۱۷۴)	۰/۰۱۸	۶/۰۷۶
نرمال	$SMCMC۱$	۵/۵۰۹ (۳/۵۰)	۰/۰۰۷ (۰/۱۸۴)	۰/۰۱۸	۶/۰۵۰
	$SMCMC۲$	۵/۵۰۷ (۳/۳۰)	۰/۰۰۷ (۰/۱۸۳)	۰/۰۱۸	۶/۰۴۹

جدول ۷: مقدار ESS و زمان (بر حسب ثانیه) محاسبه برآورد پارامترها در مدل عرض از مبداء تصادفی با روش‌های مختلف $MCMC$.

توزیع	روش	ESS	زمان کاربر	زمان سیستم
	$SMCMC$	۶۴۳۲۱	۱۱۱/۱۷	۰/۰۴
یکنواخت	$SMCMC۱$	۸۴۳۲۹	۱۶۶/۹۷	۰/۰۷
	$SMCMC۲$	۸۶۵۶۷	۱۱۱/۰۰	۰/۰۳
	$SMCMC$	۵۷۳۰۰	۱۰۸/۵۸	۰/۱
نرمال	$SMCMC۱$	۷۴۲۸۷	۱۶۲/۴۸	۰/۱۲
	$SMCMC۲$	۷۶۹۳۱	۱۰۹/۷۰	۰/۰

جدول ۸: مقدار ESS و زمان (برحسب ثانیه) محاسبه برآورد پارامترها در مدل عرض از مبداء تصادفی در حضور متغیر تبیینی با روش‌های مختلف $MCMC$.

توزیع	روش	ESS	زمان کاربر	زمان سیستم
	$SMCMC$	۵۴۶۵۴	۱۲۵/۴۳	۰/۰۲
یکنواخت	$SMCMC۱$	۷۴۵۰۰	۱۳۰/۸۸	۰/۰۴
	$SMCMC۲$	۷۶۵۰۰	۱۲۰/۷۸	۰/۰۱
	$SMCMC$	۴۹۱۰۰	۱۵۵/۵۸	۰/۸۴
نرمال	$SMCMC۱$	۷۰۵۰۱	۱۶۳/۷۸	۰/۸۶
	$SMCMC۲$	۷۳۸۹۳	۱۵۲/۰۲	۰/۰۷

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

اساس یک مثال کاربردی و یک مطالعه شبیه‌سازی مورد ارزیابی قرار گرفت و تجزیه ماتریس کواریانس مدل ساختاری و سپس اجرای روش شبیه‌سازی MCMC به عنوان یک راه‌کار مفید و موثر به منظور تولید نمونه‌های موثر مستقل بیشتر و همچنین کاهش زمان اجرای الگوریتم SMCMC پیشنهاد گردید. به علاوه نشان داده شد که کارائی روش SMCMC با فرض نرمال بودن متغیرهای درگیر بیشتر از فرض توزیع یکنواخت است و پیش‌بینی می‌شود این مطلب با فرض توزیع‌های دیگر نیز صادق باشد.

تقدیر و تشکر

نویسندگان از داوران محترم که پیشنهادهای سازنده‌ای برای بهبود این مقاله ارائه کردند تقدیر و تشکر می‌نمایند.

مراجع

Baker, A. J. (2003), *Matrix Groups: An Introduction to Lie Group Theory*, New York, Springer-Verlag.

Banerjee, S, Carlin, B. P. and Gelfand, A. E. (2004), *Hierarchical Modelling and Analysis for Spatial Data*, London, Chapman and Hall.

Besag, J., York, J. and Mollie, A. (1992), Bayesian Image Restoration with Two Applications in Spatial Statistics, *Annals of Institute of Statistical Mathematics*, **43**, 1-59.

Browne, W. J. (2009), *MCMC Estimation in MLwiN (Version 2.10)*, Center for Multilevel Modelling, University of Bristol.

Browne, W. J. and Draper, D. (2000), Implementation Issues in Bayesian Fitting of Multilevel Models, *Computational Statistics*, **15**, 391-420.

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995), Efficient Parameterisations for Linear Mixed Models, *Biometerika*, **83**, 479-488.

Gelman, A., van Dyk, D. A., Huang, Z.Y. and Boscardin, W. J. (2008), Using Redundant Parameterization to Fit Hierarchical Models, *Journal of Computational and Graphical Statistics*, **17**, 95-122.

Gilks, W.R. and Robert, G. O. (1996), Strategies for Improving MCMC, In: Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds), *Markov Chain Monte Carlo in Practice*, 89-140, London, Chapman and Hall.

Goldstein, H. (1999), *Multilevel Statistical Models*, London, Institute of Education, Multilevel Project.

Hills, S. E. and Smith, A. F. M. (1992), Parameterization Issues in Bayesian Inference, In Bernardo, J., Berger, J., Dawid, A. and Smith, A. *Bayesian Statistics,* **4**, 227-246.

Mira. A. and Sargent. D. J. (2003), A New Strategy for Speeding Markov Chain Monte Carlo Algorithms. *Statistical Methods and Applications*, **12**, 49-60.

Papaspiliopoulos, O., Roberts, G. O. and Skold, M. (2007), A General Framework for Parametrisation of Hierarchical Models, *Statistical Sciences*, **22**, 59-73.

Pinherio, J. C. and Bates, D. M. (2000), *Mixed-Effects Models in R and S-plus*, New York, Springer-Verlag.

Rencher, A. C. (1995), *Methods of Multivariate Analysis*, New York, Wiley.

Sargent, D. J., Hodges, J. S. and Carlin, B. P. (2000), Structured Markov Chain Monte Carlo, *Journal of Computational and Graphical Statistics*, **9**, 217-234.

Twisk, J. W. R. (2006), *Applied Multilevel Analysis: Practical Guides to Biostatisticals and Epidemiology*, Cambridge, Cambridge University Press.

[DOR: 20.1001.1.17358183.1389.4.1.16.4]

Journal of Statistical Sciences, Spring & Summer 2010 **145**

Vol. 4, No. 1, pp 35-58.

Improving of Structured Markov Chain Monte Carlo Algorithm in Multilevel Models

Farokhi, A. and Golalizadeh, M.

Deptartment of Statistics, Tarbiat Modares University, Tehran, Iran.

Abstract: The multilevel models are used in applied sciences including social sciences, sociology, medicine, economic for analysing correlated data. There are various approaches to estimate the model parameters when the responses are normally distributed. To implement the Bayesian approach, a generalized version of the Markov Chain Monte Carlo algorithm, which has a simple structure and removes the correlations among the simulated samples for the fixed parameters and the errors in higher levels, is used in this article. Because the dimension of the covariance matrix for the new error vector is increased, based upon the Cholesky decomposition of the covariance matrix, two methods are proposed to speed the convergence of this approach. Then, the performances of these methods are evaluated in a simulation study and real life data.

Keywords: Multilevel data, Random intercept models, MCMC algorithm, Cholesky decomposition.

Mathematics Subject Classification (2000): 62F15, 62J99