گزیده‌مطالب آماری
سال ۱۹، شماره‌ی ۲، پاییز وزمستان ۱۳۸۷، صص ۱۹۸-۱۸۵

ارزیابی براورد ماکسیمم درستنمایی مدل‌های معادلات ساختاری غیر خطی با داده‌های به‌طور تصادفی گم‌شده تحت نرخ‌های گم‌شدگی مختلف

زینب صنم‌نو،[۱] مجتبی گنجعلی[۲،*]

[۱] دانشجوی فوق لیسانس آمار، دانشگاه شهید بهشتی
۲ دانشیار آمار، دانشگاه شهید بهشتی

چکیده. در علوم رفتاری و اجتماعی برخورد با متغیرهای پنهان بسیار متداول است. یکی از بهترین روش‌ها برای مدل‌بندی این‌گونه متغیرها، مدل معادلات ساختاری است که از دو معادله‌ی اندازه‌گیری و ساختاری تشکیل یافته است و روابط بین متغیرهای پنهان با معادله‌ی ساختاری نشان داده می‌شوند. با وجود این، نظریه‌ی ماکسیمم درستنمایی و نرم‌افزارهای کامپیوتری موجود نظیر لیزرل [۸] و EQS [۱] که در مطالعات روان‌شناسی و اجتماعی برای ارزیابی ارتباطات بین متغیرهای پنهان به کار می‌روند، بر اساس روابط خطی بین متغیرها و وجود داده‌های کامل بنا نهاده شده‌اند. وجود داده‌های گم‌شده از یک طرف و از طرف دیگر وجود ارتباطات غیر خطی بین متغیرهای پنهان برای به دست آوردن مدل‌های معنی‌دار از اهمیت بسیاری برخوردار است. لی و همکاران [۱۰] الگوریتمی از نوع امیدگیری-ماکسیمم‌سازی (EM) را معرفی کردند که برای براورد ماکسیمم درستنمایی پارامترهای معادلات ساختاری غیر خطی با داده‌های به‌طور تصادفی گم‌شده (MAR) به کار می‌رود. در این الگوریتم برای به دست آوردن انتگرال‌های پیچیده در امید شرطی، گام E به‌وسیله‌ی الگوریتم دورگه‌ای کامل می‌شود که نمونه‌گیر گیبس [۶] و الگوریتم متروپلیس-هستینگس را ترکیب می‌کند درحالی‌که گام M به‌طور کارایی به‌وسیله‌ی ماکسیمم‌سازی شرطی [۱۵] کامل می‌گردد. در این مقاله قصد داریم تا با استفاده از یک مطالعه‌ی شبیه‌سازی، کارایی

واژگان کلیدی: الگوریتم متروپلیس-هستینگس؛ داده‌های گم‌شده؛ معادلات ساختاری غیر خطی؛ نمونه‌گیر گیبس.

* نویسنده‌ی عهده‌دار مکاتبات

این روش را زمانی که نرخ گم‌شدگی افزایش می‌یابد مورد بررسی قرار دهیم.

۱- مقدمه

مدل تحلیل عاملی و مدل‌هایی از نوع لیزرل مدل‌هایی هستند که در آن‌ها متغیرهای پنهان با تابع‌های خطی در ارتباط‌اند. اخیراً مشخص شده است که روابط غیر خطی بین متغیرهای پنهان برای تشکیل مدل‌های درست‌تر و معنی‌دارتر گاهی اوقات مفید است. به‌عنوان مثال برای اثرهای متقابل و درجه دوی متغیرهای پنهان در مطالعات کاربردی به [۲]، [۲۰] و مرجع‌های موجود در آن‌ها رجوع کنید.

به‌دلیل پیچیدگی توزیع‌های مرتبط با متغیرهای پنهان غیر خطی، تحلیل مدل‌های معادلات ساختاری غیر خطی (NSEM) دشوار است. مدل‌های تحلیل عاملی غیر خطی ابتدا توسط مک‌دونالد [۱۴] مورد مطالعه قرار گرفت و سپس توسط اعتضادی آملی و مک‌دونالد [۵] و مویجارت و بنتلر [۱۷] توسعه داده شد. اخیراً برای تحلیل برخی مدل‌های معادلات ساختاری غیر خطی با اثرات متقابل متغیرهای پنهان، روش‌هایی پیشنهاد شده‌اند که برنامه‌ی لیزرل [۸] را به‌کار می‌برند. این روش‌ها را می‌توان در منابع کنی و جود [۹]، پینگ [۱۸] و مارش و همکاران [۱۳] ملاحظه کرد.

لی و همکاران [۱۰] الگوریتمی از نوع EM را معرفی کردند که برای براورد ماکسیمم درستنمایی پارامترهای معادلات ساختاری غیر خطی با داده‌های به‌طور تصادفی گم‌شده به‌کار می‌رود. رویکرد ماکسیمم درستنمایی، یک روش مهم آماری است که بسیاری از خصوصیات بهینه‌ی آماری نظیر سازگاری، کارایی و ... را دارد. به‌دلیل این‌که برخورد با داده‌های گم‌شده در عمل بسیار معمول است، لی و همکاران [۱۰] روش ماکسیمم درستنمایی را در حضور داده‌های به‌طور تصادفی گم‌شده بسط دادند. تحلیل داده‌های گم‌شده در علم آمار مورد توجه زیادی قرار گرفته است و هنوز هم حوزه‌ی فعالی در پژوهش‌هاست (برای مثال لیتل و روبین [۱۲] و افرون [۴]). برای مدل معادلات ساختاری غیر خطی با داده‌های گم‌شده، معرفی شده به‌واسطه‌ی روابط غیر خطی بین متغیرهای پنهان و حضور داده‌های گم‌شده، توزیع توأم داده‌های مشاهده‌شده پیچیده است. در نتیجه به دست آوردن تابع لگاریتم درستنمایی داده‌های مشاهده‌شده دشوار بوده و محاسبه‌ی

براوردهای ماکسیمم درستنمایی به‌وسیله‌ی ماکسیمم‌سازی مستقیم این تابع بسیار مشکل است. لی و همکاران [۱۰] برای به دست آوردن براورد ماکسیمم درستنمایی از الگوریتم شناخته‌شده‌ی EM [۳] استفاده کردند. در روش آن‌ها علاوه بر داده‌های گم‌شده‌ی واقعی، با متغیرهای پنهان نیز همانند داده‌های گم‌شده‌ی فرضی رفتار می‌شود. به هر حال هنوز به‌علت وجود داده‌های گم‌شده و غیر خطی بودن مدل، پیچیدگی وجود دارد. برای حل این مشکل در گام E، امیدهای شرطی با استفاده از توزیع شرطی مقادیر گم‌شده به‌شرط مقادیر مشاهده‌شده، با میانگین‌های نمونه‌ای تقریب زده می‌شوند. گام M نیز برای حل، فرم بسته ندارد و از ماکسیمم‌سازی شرطی برای تکمیل این گام استفاده می‌شود. در این مقاله ابتدا مدل‌های معادلات ساختاری غیر خطی معرفی می‌شوند، سپس الگوریتم معرفی‌شده توسط لی و همکاران [۱۰] ارائه می‌شود و در انتها کارایی این روش زمانی که تعداد داده‌های گم‌شده افزایش می‌یابد با استفاده از یک مطالعه‌ی شبیه‌سازی مورد بررسی قرار می‌گیرد.

۲- مدل‌های معادلات ساختاری غیر خطی

۲-۱- توصیف مدل

فرض کنید $\mathbf{y}_i$ بردار تصادفی آشکار $p\times 1$ باشد که در مدل اندازه‌گیری زیر صدق می‌کند.

$$\mathbf{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\zeta}_i + \boldsymbol{\varepsilon}_i \qquad (۱)$$

که در آن $\boldsymbol{\Lambda}$ یک ماتریس $p\times q$ $(q<p)$ از محموله‌های عاملی، $\boldsymbol{\mu}$ برداری از ثابت‌ها، $\boldsymbol{\zeta}_i$ بردار تصادفی $q\times 1$ ازعامل‌های تصادفی و $\boldsymbol{\varepsilon}_i$ بردار $p\times 1$ از مقادیر خطا با توزیع $N_p(\mathbf{0},\boldsymbol{\Psi})$ است که $\boldsymbol{\Psi}$ یک ماتریس کواریانس قطری و $\boldsymbol{\zeta}_i$ مستقل از $\boldsymbol{\varepsilon}_i$ است. در موقعیت‌های پیچیده‌تر، $\boldsymbol{\zeta}_i$ را به‌صورت $\left(\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T\right)$ افراز می‌کنیم. فرض می‌شود که بردار متغیرهای پنهان افرازشده در مدل معادله‌ی ساختاری غیر خطی زیر صدق می‌کند.

$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}H(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i \qquad (۲)$$

که در آن $\boldsymbol{\eta}_i$ و $\boldsymbol{\xi}_i$ به‌ترتیب زیربردارهای پنهان $q_1 \times 1$ و $q_2 \times 1$ از $\boldsymbol{\zeta}_i$ هستند $(q_1 + q_2 = q)$، همچنین تابع $H(\boldsymbol{\xi}_i) = (h_1(\boldsymbol{\xi}_i), ..., h_t(\boldsymbol{\xi}_i))^T$ تابع بردار-مقدار $t \times 1$ با توابع مشتق‌پذیر است که به‌طور خطی مستقل از $h_1, ..., h_t$ است $(t > q_2)$ و $\mathbf{B}$ و $\boldsymbol{\Gamma}$ به‌ترتیب ماتریس‌های $q_1 \times q_1$ و $q_1 \times t$ از ضرایب رگرسیون $\boldsymbol{\eta}_i$ روی $\boldsymbol{\eta}_i$ و $H(\boldsymbol{\xi}_i)$ هستند. همانند لی و زو [۱۱] فرض می‌شود که $\mathbf{B}_0 = \mathbf{I} - \mathbf{B}$ ناتَکین است. علاوه بر این $\boldsymbol{\xi}_i$ و $\boldsymbol{\delta}_i$ به‌ترتیب به‌طور مستقل هم‌توزیع با $N_p[0, \boldsymbol{\Phi}]$ و $N_p[0, \boldsymbol{\Psi}_\delta]$ هستند که در آن‌ها ماتریس‌های کواریانس $\boldsymbol{\Phi}$ و $\boldsymbol{\Psi}_\delta$ قطری و متقارن می‌باشند. با فرض $\boldsymbol{\Pi} = (\mathbf{B}, \boldsymbol{\Gamma})$ و $G(\boldsymbol{\zeta}_i) = (\boldsymbol{\eta}_i^T, H(\boldsymbol{\xi}_i)^T)^T$، معادله‌ی ساختاری غیر خطی (۲) را می‌توان به‌صورت زیر بازنویسی کرد.

$$\boldsymbol{\eta}_i = \boldsymbol{\Pi} G(\boldsymbol{\zeta}_i) + \boldsymbol{\delta}_i \ . \qquad (۳)$$

۳- تابع درستنمایی در تحلیل مدل

فرض کنید $\mathbf{y}_i = \{\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis}\}$ که $\mathbf{y}_{i,obs}$ مقادیر مشاهده‌شده و $\mathbf{y}_{i,mis}$ مقادیر گم‌شده هستند. در این‌جا فرض می‌کنیم که داده‌ها به‌طور تصادفی و با یک مکانیسم قابل چشم‌پوشی گم شده‌اند. این نوع داده‌های گم‌شده در مطالعات رفتاری و اجتماعی معمول هستند.

فرض کنید $\mathbf{Y}_{obs} = \{\mathbf{y}_{i,obs}; i = 1, ..., n\}$، $\mathbf{Y}_{mis} = \{\mathbf{y}_{i,mis}; i = 1, ..., n\}$ و $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. لی و همکاران [۱۰] الگوریتم EM مونت کارلو (MCEM) را برای به دست آوردن براوردهای درستنمایی ماکسیمم NSEM با داده‌های MAR بر اساس Y_{obs} گسترش دادند.

فرض کنید $\mathbf{Z} = (\zeta_1, ..., \zeta_n)$ ماتریس متغیرهای پنهان در NSEM باشد. بر اساس ایده‌ی اصلی الگوریتم EM، داده‌های مشاهده‌شده‌ی $\mathbf{Y}_{obs}$ با داده‌افزایی داده‌های پنهان $\mathbf{Z}$ و داده‌های گم‌شده‌ی واقعی $\mathbf{Y}_{mis}$ کامل می‌شود. فرض کنید $\mathbf{X} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{Z})$ مجموعه‌ی داده‌های کامل و $L_c(\mathbf{X}; \boldsymbol{\theta}) = \log \Pr(\mathbf{X}|\boldsymbol{\theta})$ تابع لگاریتم درستنمایی داده‌های کامل باشد. ابتدا توجه کنید که

$$\Pr(\mathbf{y}_i, \zeta_i \,|\boldsymbol{\theta}) = \Pr(\mathbf{y}_i \,|\zeta_i, \boldsymbol{\theta}_1) \Pr(\boldsymbol{\eta}_i \,|\boldsymbol{\xi}_i, \boldsymbol{\theta}_2) \Pr(\boldsymbol{\xi}_i \,|\boldsymbol{\theta}_3) \quad (۴)$$

که در آن $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ ، $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi})$ ، $\boldsymbol{\theta}_2 = (\boldsymbol{\Pi}, \boldsymbol{\Psi}_\delta)$ و $\boldsymbol{\theta}_3 = \boldsymbol{\Phi}$. از روابط (۱) و (۲) ، تابع لگاریتم درستنمایی داده‌های کامل شامل سه قسمت است که در زیر آورده شده‌اند.

$$L_c(\mathbf{X};\boldsymbol{\theta}) = L_1(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\Psi}) + L_2(\mathbf{Z};\boldsymbol{\Pi},\boldsymbol{\Psi}_\delta) + L_3(\mathbf{Z};\boldsymbol{\Phi}), \quad (۵)$$

که در آن

$$L_1(\mathbf{X},\boldsymbol{\theta}_1) = C_1 - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\zeta_i)^T \boldsymbol{\Psi}^{-1}(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\zeta_i), \quad (۶)$$

$$L_2(\mathbf{Z},\boldsymbol{\theta}_2) = C_2 - \frac{n}{2}\log|\boldsymbol{\Psi}_\delta| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\eta}_i - \boldsymbol{\Pi}G(\zeta_i))^T \boldsymbol{\Psi}_\delta^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\Pi}G(\zeta_i)), \quad (۷)$$

$$L_3(\mathbf{Z},\boldsymbol{\theta}_3) = C_3 - \frac{n}{2}\log|\boldsymbol{\Phi}| - \frac{1}{2}\sum_{i=1}^{n}\xi_i^T \boldsymbol{\Phi}^{-1}\xi_i . \quad (۸)$$

$C_1 = -\left(\frac{pn}{2}\right)\log(2\pi)$ ، $C_2 = -\left(\frac{q_1 n}{2}\right)\log(2\pi)$ و $C_3 = -\left(\frac{q_2 n}{2}\right)\log(2\pi)$ مقادیر ثابت هستند.

۴- الگوریتم MCEM

برای NSEM معرفی‌شده با داده‌های گم‌شده، به‌واسطه‌ی روابط غیر خطی بین متغیرهای پنهان و حضور داده‌های گم‌شده، توزیع توأم داده‌های کامل، پیچیده است. همان‌گونه که قبلاً نیز ذکر شده است الگوریتم EM به‌عنوان مهم‌ترین روش برای حل مسائل ML با داده‌های گم‌شده به‌خوبی شناخته شده است. این روش به‌صورت زیر به کار گرفته می‌شود. در rامین تکرار با مقدار جاری $\boldsymbol{\theta}^{(r)}$ دو گام زیر اجرا شود.

گام E: براورد کردن

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \mathrm{E}\left\{L_c(\mathbf{X},\boldsymbol{\theta} \,|\, \mathbf{Y}_{obs}, \boldsymbol{\theta}^{(r)})\right\} = \mathrm{E}\left\{L_c(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{Z}, \boldsymbol{\theta} \,|\, \mathbf{Y}_{obs}, \boldsymbol{\theta}^{(r)})\right\}$$

که امید ریاضی نسبت به توزیع شرطی توأم $(\mathbf{Y}_{mis},\mathbf{Z})$ بـه شـرط $\mathbf{Y}_{obs}$ و $\boldsymbol{\theta}^{(r)}$ گرفتـه می‌شود.

گام M: ماکسیمم کردن $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(r)}\right)$ تا $\boldsymbol{\theta}^{(r)}$ به $\boldsymbol{\theta}^{(r+1)}$ به‌روز شود.

محاسبه‌ی مستقیم امیـدهای شـرطی کـه در گـام E لازم اسـت بـه‌خـاطر پیچیـدگی $G(\zeta_i)$ و وجود داده‌های گم‌شده مشکل است. با الهام از ایـده‌ی داده‌شـده در وی و تنـر [۲۱]، گام E به‌وسیله‌ی تعداد به‌اندازه‌ی کافی بـزرگ از مـشاهدات کـه از توزیـع شـرطی $(\mathbf{Y}_{mis},\mathbf{Z})$ به‌شرط $\mathbf{Y}_{obs}$ و $\boldsymbol{\theta}^{(r)}$ شبیه‌سازی شـده‌انـد، تقریـب زده مـی‌شـود. الگـوریتم دورگه‌ای که نمونه‌گیر گیبس و الگوریتم متروپلیس–هستینگس (MH) را ترکیب می‌کنـد برای این هدف به وجود آمده است. چون گام M نیز حل فرم بسته ندارد، $\boldsymbol{\theta}^{(r+1)}$ از طریق دنباله‌ای از گام‌های ماکسیمم‌سازی شرطی (رجوع کنید به [۱۵]) بـه دسـت خواهـد آمـد. بنا بر این الگوریتم EM معرفی‌شده، به‌عنوان یک الگوریتم مونت کارلویی تلقی می‌گردد.

۱-۴- اجرای گام E از طریق الگوریتم دورگه

فرض کنید $h(\mathbf{Y}_{mis},\mathbf{Z})$ تابعی عمـومی از $\mathbf{Y}_{mis}$ و $\mathbf{Z}$ باشـد کـه در $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(r)}\right)$ وارد شده است. امید شرطی آن به‌وسیله‌ی

$$\hat{E}\left\{h(\mathbf{Y}_{mis},\mathbf{Z})\middle|\mathbf{Y}_{obs},\boldsymbol{\theta}\right\}=\frac{1}{M}\sum_{m=1}^{M}h\left[\mathbf{Y}_{mis}^{(m)},\mathbf{Z}^{(m)}\right] \qquad (۹)$$

تقریب زده می‌شـود کـه $\left\{\left[\mathbf{Y}_{mis}^{(m)},\mathbf{Z}^{(m)}\right];m=1,\ldots,M\right\}$ نمونـه‌ی بـه‌انـدازه‌ی کـافی بزرگی است که از توزیع توأم شرطی $\Pr(\mathbf{Y}_{mis},\mathbf{Z}|\,\mathbf{Y}_{obs},\boldsymbol{\theta})$ شـبیه‌سـازی شـده اسـت. نمونه‌گیر گیبس [۶] زیر برای گرفتن این نمونه‌ها مورد استفاده قرار می‌گیرد.

در kامین تکرار با مقادیر جاری $\mathbf{Y}_{mis}^{(k)}$ و $\mathbf{Z}^{(k)}$،

گام ۱: $\mathbf{Y}_{mis}^{(k+1)}$ را از توزیع $\Pr\left(\mathbf{Y}_{mis}\,|\,\mathbf{Y}_{obs},\boldsymbol{\theta},\mathbf{Z}^{(k)}\right)$ استخراج می‌کنیم و

گام ۲: $\mathbf{Z}^{(k+1)}$ را از توزیع $\Pr\left(\mathbf{Z}|\,\mathbf{Y}_{obs},\boldsymbol{\theta},\mathbf{Y}_{mis}^{(k+1)}\right)$ استخراج می‌کنیم.

برای $i = ۱,\ldots,n$ چون $\mathbf{y}_i$ها متقابلاً مستقل هستند، $\mathbf{y}_{i,mis}$ها نیز متقابلاً مستقل هستند. چون $\boldsymbol{\Psi}_\varepsilon$ قطری است، $\mathbf{y}_{i,mis}$ به‌طور شرطی از $\mathbf{y}_{i,obs}$ به‌شرط ζ_i مستقل است. با استفاده از رابطه‌ی (۱) به دست می‌آوریم

$$\Pr\left(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \boldsymbol{\theta}, \mathbf{Z}\right) = \prod_{i=۱}^{n} \Pr\left(\mathbf{y}_{i,mis} \mid \boldsymbol{\theta}, \zeta_i\right),$$

و

$$\left[\mathbf{y}_{i,mis} \mid \boldsymbol{\theta}, \zeta_i\right] \overset{D}{=} N\left[\boldsymbol{\mu}_{i,mis} + \boldsymbol{\Lambda}_{i,mis}\zeta_i, \boldsymbol{\Psi}_{\varepsilon_i,mis}\right] \quad (۱۰)$$

که بردار $\boldsymbol{\mu}_{i,mis}$ زیربردار $p_i \times ۱$ از $\boldsymbol{\mu}$ است، $\boldsymbol{\Lambda}_{i,mis}$ زیرماتریس $p_i \times q$ از $\boldsymbol{\Lambda}$ با سطرهای مطابق با اجزای مشاهده‌نشده است که حذف شده‌اند و $\boldsymbol{\Psi}_{\varepsilon_i,mis}$ زیرماتریس $p_i \times p_i$ از $\boldsymbol{\Psi}_\varepsilon$ با سطرها و ستون‌های حذف‌شده‌ی متناظر است. در حالت کلی، مکانیسم Y_{mis} ممکن است خیلی پیچیده با موقعیت‌های متفاوتی از اقلام گم‌شده باشد. با وجود این، توزیع شرطی متناظر، تنها حاصل‌ضرب توزیع‌های نرمال ساده است.

NSEM پیچیده که به‌وسیله‌ی معادلات (۱) و (۲) تعریف شده است، به‌شرط بردارهایی از متغیرهای پنهان $\mathbf{Z}$ و داده‌های گم‌شده‌ی $\mathbf{Y}_{mis}$، به مدل معادلات هم‌زمان آشناتر کاهش پیدا می‌کند. در نتیجه، پیچیدگی ناشی از غیر خطی بودن متغیرهای پنهان تصادفی و اقلام گم‌شده، به‌شدت کاسته شده است. برای ζ_i در $\mathbf{Z}$، براساس تعریف مدل و فرض‌های آن داریم $\Pr\left(\zeta_i \mid \mathbf{Y}_{obs}, \boldsymbol{\theta}, \mathbf{Y}_{mis}\right) = \Pr\left(\zeta_i \mid \mathbf{Y}, \boldsymbol{\theta}\right) = \Pr\left(\zeta_i \mid \mathbf{y}_i, \boldsymbol{\theta}\right)$ که با استفاده از [۱۰] متناسب است با

$$\exp\{-\frac{۱}{۲}\boldsymbol{\xi}_i \boldsymbol{\Phi}^{-۱}\boldsymbol{\xi}_i - \frac{۱}{۲}\left(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\zeta_i\right)^T \boldsymbol{\Psi}_\varepsilon^{-۱}\left(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\zeta_i\right)$$
$$-\frac{۱}{۲}\left[\boldsymbol{\eta}_i - \boldsymbol{\Pi}G\left(\zeta_i\right)\right]^T \boldsymbol{\Psi}_\delta^{-۱}\left[\boldsymbol{\eta}_i - \boldsymbol{\Pi}G\left(\zeta_i\right)\right]\}. \quad (۱۱)$$

توزیع غیر استاندارد فوق نسبتاً پیچیده است. بنا بر این، مجبوریم که از الگوریتم MH ([۱۶] و [۷]) برای شبیه‌سازی کارامدتر مشاهدات از آن استفاده کنیم. الگوریتم MH روش MCMC شناخته‌شده‌ای است که به‌طور گسترده برای شبیه‌سازی مشاهدات از چگالی هدف استفاده می‌شود. این کار با کمک توزیع پیشنهادی، توزیعی که نمونه‌گیری از

آن راحت‌تر است، انجام می‌گیرد. این‌جا $\Pr(\zeta_i \,|\mathbf{y}_i, \boldsymbol{\theta})$ به‌عنوان چگالی هدف در نظر گرفته می‌شود. براساس پیشنهاد داده شده در [۱۹]، طبیعی است که $N\left(0, \sigma^2 \boldsymbol{\Omega}\right)$ را به‌عنوان توزیع پیشنهادی انتخاب کنیم، که σ^2 مقدار انتخابی و $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}_\zeta^{-1} + \boldsymbol{\Lambda}^T \boldsymbol{\Psi}_\varepsilon^{-1} \boldsymbol{\Lambda}$. همچنین

$$\boldsymbol{\Sigma}_\zeta^{-1} = \begin{bmatrix} \mathbf{B}_0^T \boldsymbol{\Psi}_\delta^{-1} \mathbf{B}_0 & -\mathbf{B}_0^T \boldsymbol{\Psi}_\delta^{-1} \boldsymbol{\Gamma} \boldsymbol{\Delta} \\ -\boldsymbol{\Delta}^T \boldsymbol{\Gamma}^T \boldsymbol{\Psi}_\delta^{-1} \mathbf{B}_0 & \boldsymbol{\Phi}^{-1} + \boldsymbol{\Delta}^T \boldsymbol{\Gamma}^T \boldsymbol{\Psi}_\delta^{-1} \boldsymbol{\Gamma} \boldsymbol{\Delta} \end{bmatrix} \quad (۱۲)$$

که $\boldsymbol{\Delta} = \partial H(\boldsymbol{\xi}) / \partial \boldsymbol{\xi}^T \Big|_{\boldsymbol{\xi}=0}$.

۲-۴- گام ماکسیمم‌سازی

در گام M نیاز داریم تا $Q\left[\boldsymbol{\theta} \middle| \boldsymbol{\theta}^{(r)}\right]$ را نسبت به $\boldsymbol{\theta}$ ماکسیمم کنیم. این عمل معادل با حل مجموعه‌ای از معادلات به‌صورت زیر است.

$$\frac{\partial Q\left[\boldsymbol{\theta} \middle| \boldsymbol{\theta}^{(r)}\right]}{\partial \boldsymbol{\theta}} = \mathrm{E}\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} L_c\left(\mathbf{X} \middle| \boldsymbol{\theta}\right) \middle| \mathbf{U}_{obs}, \boldsymbol{\theta}^{(r)} \right\} = 0 \quad (۱۳)$$

برای $i = 1,\ldots,p$ و $j = 1,\ldots,q_1$ فرض کنید $y_{i(k)}$، kامین عنصر $\mathbf{y}_i$، $\eta_{i(j)}$، jامین عنصر $\boldsymbol{\eta}_i$ و $\boldsymbol{\Lambda}_k$ و $\boldsymbol{\Pi}_j$ به‌ترتیب، kامین و jامین سطر $\boldsymbol{\Lambda}$ و $\boldsymbol{\Pi}$ باشند. در این صورت

$$\frac{\partial L_c\left(\mathbf{X} \middle| \boldsymbol{\theta}\right)}{\partial \boldsymbol{\mu}} = \boldsymbol{\Psi}_\varepsilon^{-1} \sum_{i=1}^{n} \left(\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda} \boldsymbol{\zeta}_i\right),$$

$$\frac{\partial L_c\left(\mathbf{X} \middle| \boldsymbol{\theta}\right)}{\partial \boldsymbol{\Phi}} = \frac{1}{2} \boldsymbol{\Phi}^{-1} \sum_{i=1}^{n} \left(\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T - \boldsymbol{\Phi}\right) \boldsymbol{\Phi}^{-1},$$

$$\frac{\partial L_c\left(\mathbf{X} \middle| \boldsymbol{\theta}\right)}{\partial \boldsymbol{\Lambda}_k} = \boldsymbol{\Psi}_{\varepsilon_k}^{-1} \sum_{i=1}^{n} \left[y_{i(k)} - \boldsymbol{\mu}_k - \boldsymbol{\Lambda}_k \boldsymbol{\zeta}_i \right] \boldsymbol{\zeta}_i^k,$$

$$\frac{\partial L_c\left(\mathbf{X} \middle| \boldsymbol{\theta}\right)}{\partial \boldsymbol{\Pi}_j} = \boldsymbol{\Psi}_{\delta_j}^{-1} \sum_{i=1}^{n} \left[\eta_{i(j)} - \boldsymbol{\Pi}_j G\left(\boldsymbol{\zeta}_i\right) \right] G\left(\boldsymbol{\zeta}_i\right)^T,$$

$$\frac{\partial L_c\left(\mathbf{X}|\boldsymbol{\theta}\right)}{\partial diog\left(\boldsymbol{\Psi}_\varepsilon\right)}=\frac{1}{2}diog\{\boldsymbol{\Psi}_\varepsilon^{-1}\sum_{i=1}^{n}[\left(\mathbf{y}_i-\boldsymbol{\mu}-\boldsymbol{\Lambda}\boldsymbol{\zeta}_i\right)\left(\mathbf{y}_i-\boldsymbol{\mu}-\boldsymbol{\Lambda}\boldsymbol{\zeta}_i\right)^T-\boldsymbol{\Psi}_\varepsilon]\boldsymbol{\Psi}_\varepsilon^{-1}\},$$

و

$$\frac{\partial L_c\left(\mathbf{X}|\boldsymbol{\theta}\right)}{\partial diog\left(\boldsymbol{\Psi}_\delta\right)}=\frac{1}{2}diog\left\{\boldsymbol{\Psi}_\delta^{-1}\sum_{i=1}^{n}\left(\left[\boldsymbol{\eta}_i-G\left(\boldsymbol{\zeta}_i\right)\right]\left[\boldsymbol{\eta}_i-G\left(\boldsymbol{\zeta}_i\right)\right]^T-\boldsymbol{\Psi}_\delta\right)\boldsymbol{\Psi}_\delta^{-1}\right\}.$$

(۱۴)

این معادلات هم‌زمان نمی‌توانند به فرم بسته حل شوند. بر اساس ایده‌ی داده‌شده در [۱۵] حل مورد نیاز در گام M می‌تواند از طریق چندین ماکسیمم‌سازی شرطی محاسباتی به دست آید.

۵- شبیه‌سازی داده‌ها

مجموعه داده‌ی $\mathbf{Y}=\{\mathbf{y}_1,\ldots,\mathbf{y}_n\}$ را از مدل معادلات ساختاری غیر خطی معرفی‌شده در معادلات (۱) و (۲) تولید می‌کنیم. داده‌ها را با مشخصات زیر تولید می‌کنیم.

فرض می‌کنیم که $\mathbf{y}_i=\{y_1,\ldots,y_6\}$، $\zeta_i=\left(\eta_{i1},\xi_{i1},\xi_{i2}\right)^T$ و $\eta_{i1}=\gamma_{11}\xi_{i1}+\gamma_{12}\xi_{i1}+\gamma_{13}\xi_{i1}\xi_{i2}+\gamma_{14}\xi_{i1}\xi_{i1}+\gamma_{14}\xi_{i2}\xi_{i2}$، و ساختار ماتریس محموله‌ها در معادله‌ی (۱) را به‌وسیله‌ی

$$\boldsymbol{\Lambda}^T=\begin{bmatrix}1 & \lambda_{21} & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & \lambda_{42} & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & \lambda_{63}\end{bmatrix}$$

مشخص می‌کنیم. عناصری از ماتریس $\boldsymbol{\Lambda}$ که با ۰ یا ۱ مشخص شده‌اند پارامترهای ثابت در نظر گرفته شده‌اند. در نتیجه در کل، ۲۴ پارامتر نامعلوم در مدل وجود دارد. مقادیر واقعی پارامترهای نامعلوم جامعه به‌وسیله‌ی مقادیر زیر داده می‌شوند.

$$\lambda_{21}=\lambda_{42}=\lambda_{63}=0/6\,,\quad \boldsymbol{\mu}=\left(\mu_1,\mu_2,\ldots,\mu_6\right)^T=\left(0,0,\ldots,0\right)^T,$$

$$\boldsymbol{\Gamma}=(\gamma_{11},\gamma_{12},\gamma_{13},\gamma_{14},\gamma_{15})=(0/3,0/3,0/5,-0/5,0/5)\,,$$

$$\boldsymbol{\Psi} = \text{diog}(\psi_{ii}) = \text{diog}(۰/۳,\ldots,۰/۳), \qquad i = (۱,\ldots,۶) \text{ برای}$$

$$\boldsymbol{\Psi}_{\delta} = ۰/۲۵, \ \mathbf{B} = ۲, \ \boldsymbol{\Phi} = (\phi_{۱۱}, \phi_{۱۲}, \phi_{۲۲}) = (۱, ۰/۲, ۱).$$

حجم نمونه، n، را ۴۰۰ در نظر می‌گیریم. ابتدا ۱۰ درصد از داده‌های متغیر $\mathbf{y}_{۶}$ را به‌طور تصادفی حذف می‌کنیم و پارامترهای مدل را با استفاده از الگوریتم MCEM معرفی‌شده براورد می‌کنیم. در گام بعدی ۲۰ درصد داده‌های متغیر $\mathbf{y}_{۶}$ را حذف می‌کنیم و پارامترهای مدل را براورد می‌کنیم. و گام‌های بعدی را نیز به همین ترتیب تا حذف ۸۰ درصد داده‌های متغیر $\mathbf{y}_{۶}$ انجام می‌دهیم. نتایج مربوط به براورد پارامترها در جدول ۱ آورده شده‌اند. ستون اول نام پارامترها، ستون دوم مقادیر اولیه، ستون سوم براورد پارامترها وقتی که ۱۰ درصد داده‌های متغیر $\mathbf{y}_{۶}$ گم شده‌اند، و ستون آخر براورد پارامترها وقتی که ۸۰ درصد داده‌های متغیر $\mathbf{y}_{۶}$ گم شده‌اند را نشان می‌دهد.

همان‌طور که جدول ۱ نشان می‌دهد براورد پارامترها زمانی که نرخ گم‌شدگی افزایش می‌یابد تغییر چندانی نمی‌کنند و به مقادیر واقعی پارامترها نزدیک هستند. با توجه به روابط غیر خطی میان متغیرهای پنهان در مدل معادلات ساختاری غیر خطی و وجود داده‌های گم‌شده، تحلیل مدل معادلات ساختاری غیر خطی با داده‌های گم‌شده امری متعارف نیست.

جدول ۱- MLEهای پارامترها در مطالعه‌ی شبیه‌سازی

پارامتر	مقدار	٪۱۰	٪۲۰	٪۳۰	٪۴۰	٪۵۰	٪۶۰	٪۷۰	٪۸۰
$\lambda_{۲۱}$	۰/۶	۰/۵۳۹	۰/۵۴	۰/۵۳۴	۰/۵۳۶	۰/۵۴۴	۰/۵۴۱	۰/۵۴۲	۰/۵۳۹
$\lambda_{۴۲}$	۰/۶	۰/۶۱۴	۰/۶۱۴	۰/۶۲۱	۰/۶۲	۰/۶۱۹	۰/۶۲۱	۰/۶۲۱	۰/۶۲
$\lambda_{۶۳}$	۰/۶	۰/۵۷۱	۰/۵۷۲	۰/۵۸	۰/۵۷۱	۰/۵۴	۰/۵۲۳	۰/۶۱۷	۰/۵۰۸
$\gamma_{۱۱}$	۰/۳	۰/۲۸	۰/۲۸۷	۰/۲۹	۰/۲۹۱	۰/۲۷۹	۰/۲۹۵	۰/۲۹۱	۰/۲۹۲
$\gamma_{۱۲}$	۰/۳	۰/۲۲۲	۰/۲۱۹	۰/۲۱۵	۰/۲۱۸	۰/۲۲۱	۰/۲۱۳	۰/۱۹۹	۰/۲۱۱
$\gamma_{۱۳}$	۰/۵	۰/۵۲۱	۰/۵۲۳	۰/۵۳۴	۰/۵۳۱	۰/۵۲۴	۰/۵۲۳	۰/۵۲۵	۰/۵۳۵
$\gamma_{۱۴}$	-۰/۵	-۰/۴۵۵	-۰/۴۵۲	-۰/۴۷۲	-۰/۴۵۱	-۰/۴۳۵	-۰/۴۳۷	-۰/۴۳۷	-۰/۴۵۳
$\gamma_{۱۵}$	۰/۵	۰/۴۲	۰/۴۱۵	۰/۴۳۱	۰/۴۲۴	۰/۴۰۲	۰/۳۹۳	۰/۳۸۷	۰/۴۰۴
$\mu_{۱}$	۰	-۰/۰۶۷	-۰/۰۶۳	-۰/۰۶۶	-۰/۰۶۷	-۰/۰۵۸	-۰/۰۶۲	-۰/۰۶۱	-۰/۰۵۶
$\mu_{۲}$	۰	-۰/۰۰۸	-۰/۰۰۶	-۰/۰۶۳	-۰/۰۰۷	-۰/۰۰۴	-۰/۰۰۶	-۰/۰۰۶	-۰/۰۰۲
$\mu_{۳}$	۰	-۰/۰۳۶	-۰/۰۳۳	-۰/۰۳۲	-۰/۰۳۱	-۰/۰۳	-۰/۰۳	-۰/۰۳۳	-۰/۰۳۱
$\mu_{۴}$	۰	-۰/۰۶۲	-۰/۰۶۳	-۰/۰۶۳	-۰/۰۶۲	-۰/۰۶	-۰/۰۶۱	-۰/۰۶۳	۰/۰۶۲
$\mu_{۵}$	۰	-۰/۱۱	-۰/۱۱۳	-۰/۱۱۴	-۰/۱۱۳	-۰/۱۰۹	-۰/۱۰۷	-۰/۱۱۲	-۰/۱۱
$\mu_{۶}$	۰	-۰/۰۵۶	-۰/۰۶۱	-۰/۰۵	-۰/۰۵۶	-۰/۰۴۳	-۰/۰۹۲	-۰/۰۰۶	-۰/۰۷
ψ_{δ}	۰/۲۵	۰/۲۳۳	۰/۲۳۹	۰/۲۲۱	۰/۲۲۷	۰/۲۵۲	۰/۲۷۷	۰/۲۸۶	۰/۲۵۶
$\psi_{۱۱}$	۰/۳	۰/۲۱۹	۰/۲۲۵	۰/۲۲۹	۰/۲۲۹	۰/۲۳۹	۰/۲۲۷	۰/۲۳۱	۰/۲۲۱
$\psi_{۲۲}$	۰/۳	۰/۳۵۲	۰/۳۵۱	۰/۳۵۹	۰/۳۵۶	۰/۳۴۶	۰/۳۵	۰/۳۴۹	۰/۳۵۲
$\psi_{۳۳}$	۰/۳	۰/۳۱۸	۰/۳۱۹	۰/۳۳	۰/۳۲۹	۰/۳۲۶	۰/۳۲۹	۰/۳۲۹	۰/۳۲۷
$\psi_{۴۴}$	۰/۳	۰/۳۳۱	۰/۳۳	۰/۳۲۸	۰/۳۲۷	۰/۳۲۷	۰/۳۲۶	۰/۳۲۶	۰/۳۲۸
$\psi_{۵۵}$	۰/۳	۰/۲۶۱	۰/۲۶۱	۰/۲۷۵	۰/۲۷۷	۰/۲۵۱	۰/۲۱۳	۰/۲۰۳	۰/۲۵۳
$\psi_{۶۶}$	۰/۳	۰/۳۳۱	۰/۳۲۴	۰/۳۳۴	۰/۳۲۶	۰/۳۷	۰/۳۶۴	۰/۳۸۵	۰/۳۳۹
$\phi_{۱۱}$	۱	۱/۰۹۹	۱/۰۹۹	۱/۰۸۵	۱/۰۸۸	۱/۰۹	۱/۰۸۹	۱/۰۸۸	۱/۰۹۱
$\phi_{۲۱}$	۰/۲	۰/۳۰۴	۰/۳۰۶	۰/۳۰۶	۰/۲۸۷	۰/۲۷۹	۰/۲۸۸	۰/۲۸۹	۰/۳۰۱
$\phi_{۲۲}$	۱	۱/۲۸۱	۱/۲۸	۱/۲۶۳	۱/۲۶۳	۱/۲۹۳	۱/۳۲۱	۱/۳۲۹	۱/۲۸۶

۶- نتیجه‌گیری

در این مقاله روش لی و همکاران [۱۰] معرفی شد که با استفاده از ابزارهای محاسباتی مفید در آمار نظیر الگوریتم EM، نمونه‌گیر گیبس و الگوریتم متروپلیس-هستینگس همراه با ایده‌ی داده‌افزایی، این مدل پیچیده را برازش می‌دهد. در این مطالعه تأثیر نرخ گم‌شدگی بر براورد پارامترهای مدل معادلات ساختاری غیر خطی مورد بررسی قرار گرفت. نتایج مطالعه‌ی شبیه‌سازی، کارایی الگوریتم معرفی‌شده توسط لی و همکاران [۱۰] را در براورد پارامترهای مدل معادلات ساختاری غیر خطی با افزایش نرخ گم‌شدگی نشان داد.

مرجع‌ها

[1] Bentler, P.M. (1992). EQS: Structural equation program manual. Los Angeles: BMDP Statistical Software.

[2] Busmeyer, J.R.; Jones, L.E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, **93**, 549-562.

[3] Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, **39**, 1-38.

[4] Efron, B. (1994). Missing data, imputation, and the Bootstrap (with discussion). *J. Amer. Statist. Asoc.*, **89**, 463-479.

[5] Etezadi-Amoli, J.; McDonald, R.P. (1983). A second generation nonlinear factor analysis. *Psychometrika*, **48**, 315-342.

[6] Geman, S.; Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machie Intell.*, **6**, 721-741.

[7] Hastings, W.K. (1970). Monte Carlo sampling methods using Merkov chains and their applications. *Biometrika*, **57**, 97-109.

[8] Joreskog, K.G.; Sorbom, D. (1996). LISREL 8: Structural equation modeling with the SIMPLIs command language. Scientific Software International: Hove and London.

[9] Kenny, D.A.; Judd, C.M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychometrika Bulletin*, **96**, 201-210.

[10] Lee, S.Y.; Song, X.Y.; Lee, C.K. (2003). Maximum likelihood estimation of

nonlinear structural equation models with ignorable missing data. *J. Educat. Behav. Statist*, **28**, 111-134.

[11] Lee, S.Y.; Zhu, H.T. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British J. Math. Statist. Psych.*, **53**, 209-232.

[12] Little, R.J.A.; Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[13] Marsh, H.W.; Wen, Z.; Hau, K.T. (2004). Stractural equation models of latent interaction: evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, **9**, 275-300.

[14] McDonald, R.P. (1962). A general approach to nonlinear factor analysis. *Psychometrika*, **27**, 123-157.

[15] Meng, X.L.; Rubin, D.B. (1993). Maximum likelihood estimation via ECM algorithm: a general framework. *Biometrika*, **80**, 267-278.

[16] Metropolis, N.; Rosenbluth, M.N.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. (1953). Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1092.

[17] Moorijat, A.; Bentler, P. (1986). Random polynomial factor analysis. In Diday, E. at al. (eds),Data Analysis and Informatics, IV, 241-250.

[18] Ping, R.A. (1996). Estimating latent variable interaction and quadratics: the state of this art. *Journal of management*, **22**, 163-183.

[19] Roberts, C.P.; Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

[20] Schumacker, R.E.; Marcoulides, G.A. (1998). Interaction and nonlinear effects in structural equation models. Mahwah, NJ: Lawrence Erlbaum Associate.

[21] Wei, G.C.G.; Tanner, M.A. (1990). A Monte Carlo implementation of EM algorithm and poor man's data augmentation algorihm. *J. Amer. Statist. Assoc.*, **85**, 699-704.

زینب صنم‌نو
دانشجوی فوق لیسانس آمار
تهران، اوین، دانشگاه شهید بهشتی، دانشکده‌ی علوم ریاضی، گروه آمار.
پیام‌نگار:

مجتبی گنجعلی
استاد آمار
تهران، اوین، دانشگاه شهید بهشتی، دانشکده‌ی علوم ریاضی، گروه آمار.
پیام‌نگار: m-ganjali@sbu.ac.ir