

ارزیابی برآورد 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. لسی و همکاران [۱۰] الگوریتمی از نوع امیدگیری-ماکسیمم‌سازی (EM) را معرفی کردند که برای برآورد 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 برآورد 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

برآوردهای ماکسیمم درست‌نمایی به‌وسیله‌ی ماکسیمم‌سازی مستقیم این تابع بسیار مشکل است. لی و همکاران [۱۰] برای به دست آوردن برآورد 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 y_i بردار تصادفی آشکار $p \times 1$ باشد که در مدل اندازه‌گیری زیر صدق می‌کند.

$$(۱) \quad y_i = \mu + \Lambda \zeta_i + \varepsilon_i$$

که در آن Λ یک ماتریس $(q < p) \times p$ از محموله‌های عاملی، μ برداری از ثابت‌ها، ζ_i بردار تصادفی $q \times 1$ از عامل‌های تصادفی و ε_i بردار $p \times 1$ از مقادیر خطا با توزیع $N_p(\mathbf{0}, \Psi)$ است که Ψ یک ماتریس کواریانس قطری و ζ_i مستقل از ε_i است. در موقعیت‌های پیچیده‌تر، ζ_i را به‌صورت (η_i^T, ξ_i^T) افراز می‌کنیم. فرض می‌شود که بردار متغیرهای پنهان افراز‌شده در مدل معادله‌ی ساختاری غیر خطی زیر صدق می‌کند.

$$(۲) \quad \eta_i = B\eta_i + \Gamma H(\xi_i) + \delta_i$$

که در آن $\boldsymbol{\eta}_i$ و $\boldsymbol{\zeta}_i$ به ترتیب زیربردارهای پنهان $q_1 \times 1$ و $q_r \times 1$ از $\boldsymbol{\zeta}_i$ هستند $(q_1 + q_r = q)$ ، همچنین تابع $H(\boldsymbol{\zeta}_i) = (h_1(\boldsymbol{\zeta}_i), \dots, h_t(\boldsymbol{\zeta}_i))^T$ تابع بردار-مقدار $t \times 1$ با توابع مشتق پذیر است که به طور خطی مستقل از h_1, \dots, h_t است $(t > q_r)$ و \mathbf{B} و $\mathbf{\Gamma}$ به ترتیب ماتریس های $q_1 \times q_1$ و $q_1 \times t$ از ضرایب رگرسیون $\boldsymbol{\eta}_i$ روی $\boldsymbol{\eta}_i$ و $H(\boldsymbol{\zeta}_i)$ هستند. همانند لی و زو [۱۱] فرض می شود که $\mathbf{B}_0 = \mathbf{I} - \mathbf{B}$ ناکین است. علاوه بر این $\boldsymbol{\zeta}_i$ و $\boldsymbol{\delta}_i$ به ترتیب به طور مستقل هم توزیع با $N_p[\boldsymbol{\omega}, \boldsymbol{\Phi}]$ و $N_p[\boldsymbol{\omega}, \boldsymbol{\Psi}_\delta]$ هستند که در آن ها ماتریس های کواریانس $\boldsymbol{\Phi}$ و $\boldsymbol{\Psi}_\delta$ قطری و متقارن می باشند. با فرض $\boldsymbol{\Pi} = (\mathbf{B}, \mathbf{\Gamma})$ و $G(\boldsymbol{\zeta}_i) = (\boldsymbol{\eta}_i^T, H(\boldsymbol{\zeta}_i)^T)^T$ ، معادله ی ساختاری غیر خطی (۲) را می توان به صورت زیر بازنویسی کرد.

$$(3) \quad \boldsymbol{\eta}_i = \boldsymbol{\Pi} G(\boldsymbol{\zeta}_i) + \boldsymbol{\delta}_i .$$

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 مشاهده شده و $\mathbf{y}_{i,mis}$ مقادیر گم شده هستند. در این جا فرض می کنیم که داده ها به طور تصادفی و با یک مکانیسم قابل چشم پوشی گم شده اند. این نوع داده های گم شده در مطالعات رفتاری و اجتماعی معمول هستند.

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

فرض کنید $\mathbf{Z} = (\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n)$ ماتریس متغیرهای پنهان در NSEM باشد. بر اساس ایده ی اصلی الگوریتم EM، داده های مشاهده شده ی \mathbf{Y}_{obs} با داده افزایی داده های پنهان \mathbf{Z} و داده های گم شده ی واقعی \mathbf{Y}_{mis} کامل می شود. فرض کنید $\mathbf{X} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis}, \mathbf{Z})$ مجموعه ی داده های کامل و $L_c(\mathbf{X}; \boldsymbol{\theta}) = \log \Pr(\mathbf{X} | \boldsymbol{\theta})$ تابع لگاریتم درستنمایی داده های کامل باشد. ابتدا توجه کنید که

$$(۴) \quad \Pr(\mathbf{y}_i, \zeta_i | \boldsymbol{\theta}) = \Pr(\mathbf{y}_i | \zeta_i, \boldsymbol{\theta}_\lambda) \Pr(\boldsymbol{\eta}_i | \zeta_i, \boldsymbol{\theta}_\nu) \Pr(\zeta_i | \boldsymbol{\theta}_\nu)$$

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

$$(۵) \quad L_c(\mathbf{X}; \boldsymbol{\theta}) = L_\lambda(\mathbf{X}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) + L_\nu(\mathbf{Z}; \boldsymbol{\Pi}, \boldsymbol{\Psi}_\delta) + L_\gamma(\mathbf{Z}; \boldsymbol{\Phi}),$$

که در آن

$$(۶) \quad L_\lambda(\mathbf{X}, \boldsymbol{\theta}_\lambda) = C_\lambda - \frac{n}{\nu} \log |\boldsymbol{\Psi}| - \frac{1}{\nu} \sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda} \zeta_i)^T \boldsymbol{\Psi}^{-1} (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda} \zeta_i),$$

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

$$(۸) \quad L_\gamma(\mathbf{Z}, \boldsymbol{\theta}_\gamma) = C_\gamma - \frac{n}{\nu} \log |\boldsymbol{\Phi}| - \frac{1}{\nu} \sum_{i=1}^n \zeta_i^T \boldsymbol{\Phi}^{-1} \zeta_i.$$

مقادیر $C_\nu = -\left(\frac{q \cdot n}{\nu}\right) \log(\nu \pi)$ و $C_\gamma = -\left(\frac{p \cdot n}{\nu}\right) \log(\nu \pi)$ ، $C_\lambda = -\left(\frac{p \cdot n}{\nu}\right) \log(\nu \pi)$ ثابت هستند.

۴- الگوریتم MCEM

برای NSEM معرفی شده با داده‌های گم‌شده، به واسطه‌ی 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 r امین تکرار با مقدار جاری $\boldsymbol{\theta}^{(r)}$ دو گام زیر اجرا شود.

گام E: برآورد کردن

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

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

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

محاسبه‌ی مستقیم امیدهای شرطی که در گام E لازم است به‌خاطر پیچیدگی $G(\xi_i)$ و وجود داده‌های گم‌شده مشکل است. با الهام از ایده‌ی داده‌شده در وی و تنر [۲۱]، گام E به‌وسیله‌ی تعداد به‌اندازه‌ی کافی بزرگ از مشاهدات که از توزیع شرطی $(\mathbf{Y}_{mis}, \mathbf{Z})$ به شرط \mathbf{Y}_{obs} و $\boldsymbol{\theta}^{(r)}$ شبیه‌سازی شده‌اند، تقریب زده می‌شود. الگوریتم دورگه‌ای که نمونه‌گیر گیبس و الگوریتم متروپلیس-هستینگس (MH) را ترکیب می‌کند 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]) به دست خواهد آمد. بنا بر این الگوریتم EM معرفی شده، 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 عمومی از \mathbf{Y}_{mis} و \mathbf{Z} باشد که در $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ وارد شده است. امید شرطی آن به‌وسیله‌ی

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

تقریب زده می‌شود که $\{\mathbf{Y}_{mis}^{(m)}, \mathbf{Z}^{(m)}; m = 1, \dots, M\}$ نمونه‌ی به‌اندازه‌ی کافی بزرگی است که از توزیع توأم شرطی $\Pr(\mathbf{Y}_{mis}, \mathbf{Z} | \mathbf{Y}_{obs}, \boldsymbol{\theta})$ شبیه‌سازی شده است.

نمونه‌گیر گیبس [۶] زیر برای گرفتن این نمونه‌ها مورد استفاده قرار می‌گیرد.

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

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

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

برای $i = 1, \dots, n$ چون y_i ها متقابلاً مستقل هستند، $y_{i,mis}$ ها نیز متقابلاً مستقل هستند. چون Ψ_ε قطری است، $y_{i,mis}$ به‌طور شرطی از $y_{i,obs}$ به شرط ζ_i مستقل است. با استفاده از رابطه‌ی (۱) به دست می‌آوریم

$$\Pr(\mathbf{Y}_{mis} | \mathbf{Y}_{obs}, \boldsymbol{\theta}, \mathbf{Z}) = \prod_{i=1}^n \Pr(y_{i,mis} | \boldsymbol{\theta}, \zeta_i),$$

و

$$(۱۰) \quad \left[y_{i,mis} | \boldsymbol{\theta}, \zeta_i \right]^D = N \left[\boldsymbol{\mu}_{i,mis} + \boldsymbol{\Lambda}_{i,mis} \zeta_i, \boldsymbol{\Psi}_{\varepsilon_i,mis} \right]$$

که بردار $\boldsymbol{\mu}_{i,mis}$ زیربردار $1 \times p_i$ 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 حاصل ضرب 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 $\Pr(\zeta_i | y_i, \boldsymbol{\theta}) = \Pr(\zeta_i | \mathbf{Y}, \boldsymbol{\theta}) = \Pr(\zeta_i | \mathbf{Y}_{obs}, \boldsymbol{\theta}, \mathbf{Y}_{mis})$ که با استفاده از [۱۰] متناسب است با

$$(۱۱) \quad \exp \left\{ \frac{1}{\varphi} \zeta_i \boldsymbol{\Phi}^{-1} \zeta_i - \frac{1}{\varphi} (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda} \zeta_i)^T \boldsymbol{\Psi}_\varepsilon^{-1} (\mathbf{y}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda} \zeta_i) - \frac{1}{\varphi} [\boldsymbol{\eta}_i - \boldsymbol{\Pi} G(\zeta_i)]^T \boldsymbol{\Psi}_\delta^{-1} [\boldsymbol{\eta}_i - \boldsymbol{\Pi} G(\zeta_i)] \right\}.$$

توزیع غیر استاندارد فوق نسبتاً پیچیده است. بنا بر این، مجبوریم که از الگوریتم MH [۱۶] و [۱۷] برای شبیه‌سازی کارآمدتر مشاهدات از آن استفاده کنیم. الگوریتم MH روش شناخته‌شده‌ای است که به‌طور گسترده برای شبیه‌سازی مشاهدات از چگالی هدف استفاده می‌شود. این کار با کمک توزیع پیشنهادی، توزیعی که نمونه‌گیری از

آن راحت تر است، انجام می گیرد. این جا $\Pr(\zeta_i | y_i, \theta)$ به عنوان چگالی هدف در نظر گرفته می شود. براساس پیشنهاد داده شده در [۱۹]، طبیعی است که $N(\xi, \sigma^2 \Omega)$ را به عنوان توزیع پیشنهادی انتخاب کنیم، که σ^2 مقدار انتخابی و

همچنین $\Omega^{-1} = \Sigma_{\zeta}^{-1} + \Lambda^T \Psi_{\varepsilon}^{-1} \Lambda$

$$(۱۲) \quad \Sigma_{\zeta}^{-1} = \begin{bmatrix} \mathbf{B}_0^T \Psi_{\delta}^{-1} \mathbf{B}_0 & -\mathbf{B}_0^T \Psi_{\delta}^{-1} \Gamma \Lambda \\ -\Lambda^T \Gamma^T \Psi_{\delta}^{-1} \mathbf{B}_0 & \Phi^{-1} + \Lambda^T \Gamma^T \Psi_{\delta}^{-1} \Gamma \Lambda \end{bmatrix}$$

که $\Delta = \partial H(\xi) / \partial \xi^T \Big|_{\xi=\xi_0}$

۲-۴- گام ماکسیم سازی

در گام M نیاز داریم تا $Q[\theta | \theta^{(r)}]$ را نسبت به θ ماکسیم کنیم. این عمل معادل با حل مجموعه ای از معادلات به صورت زیر است.

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

برای $i = 1, \dots, p$ و $j = 1, \dots, q$ فرض کنید $y_{i(k)}$ ، k امین عنصر y_i ، $\eta_{i(j)}$ ، k امین و j امین سطر Λ و Π باشند. در این صورت

$$\begin{aligned} \frac{\partial L_c(\mathbf{X} | \theta)}{\partial \boldsymbol{\mu}} &= \Psi_{\varepsilon}^{-1} \sum_{i=1}^n (y_i - \boldsymbol{\mu} - \Lambda \zeta_i), \\ \frac{\partial L_c(\mathbf{X} | \theta)}{\partial \Phi} &= \frac{1}{2} \Phi^{-1} \sum_{i=1}^n (\xi_i \xi_i^T - \Phi) \Phi^{-1}, \\ \frac{\partial L_c(\mathbf{X} | \theta)}{\partial \Lambda_k} &= \Psi_{\varepsilon_k}^{-1} \sum_{i=1}^n [y_{i(k)} - \boldsymbol{\mu}_k - \Lambda_k \zeta_i] \zeta_i^k, \\ \frac{\partial L_c(\mathbf{X} | \theta)}{\partial \Pi_j} &= \Psi_{\delta_j}^{-1} \sum_{i=1}^n [\eta_{i(j)} - \Pi_j G(\zeta_i)] G(\zeta_i)^T, \end{aligned}$$

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

9

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

(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 $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 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 $\mathbf{y}_i = \{y_1, \dots, y_\varepsilon\}$ ، $\boldsymbol{\zeta}_i = (\eta_{i1}, \xi_{i1}, \xi_{i2})^T$ و $\eta_{i1} = \gamma_{11}\xi_{i1} + \gamma_{12}\xi_{i1} + \gamma_{13}\xi_{i1}\xi_{i2} + \gamma_{14}\xi_{i1}\xi_{i2} + \gamma_{15}\xi_{i2}\xi_{i2}$ و ساختار ماتریس محموله‌ها در معادله‌ی (۱) را به‌وسیله‌ی

$$\boldsymbol{\Lambda}^T = \begin{bmatrix} 1 & \lambda_{\eta_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \lambda_{\eta_2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{\eta_3} \end{bmatrix}$$

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.

$$\lambda_{\eta_1} = \lambda_{\eta_2} = \lambda_{\eta_3} = 0.6, \quad \boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_\varepsilon)^T = (0, 0, \dots, 0)^T,$$

$$\boldsymbol{\Gamma} = (\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{14}, \gamma_{15}) = (0.3, 0.3, 0.5, -0.5, 0.5),$$

$$\Psi = \text{diag}(\psi_{ii}) = \text{diag}(0/3, \dots, 0/3), \quad i = (1, \dots, 6) \text{ برای}$$

$$\Psi_{\delta} = 0/25, \quad \mathbf{B} = \mathbf{0}, \quad \Phi = (\phi_{11}, \phi_{12}, \phi_{22}) = (1, 0/2, 1).$$

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

همان‌طور که جدول ۱ نشان می‌دهد برآورد 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
λ_{γ_1}	۰/۶	۰/۵۳۹	۰/۵۴	۰/۵۳۴	۰/۵۳۶	۰/۵۴۴	۰/۵۴۱	۰/۵۴۲	۰/۵۳۹
λ_{γ_2}	۰/۶	۰/۶۱۴	۰/۶۱۴	۰/۶۲۱	۰/۶۲	۰/۶۱۹	۰/۶۲۱	۰/۶۲۱	۰/۶۲
λ_{γ_3}	۰/۶	۰/۵۷۱	۰/۵۷۲	۰/۵۸	۰/۵۷۱	۰/۵۴	۰/۵۲۳	۰/۶۱۷	۰/۵۰۸
γ_{11}	۰/۳	۰/۲۸	۰/۲۸۷	۰/۲۹	۰/۲۹۱	۰/۲۷۹	۰/۲۹۵	۰/۲۹۱	۰/۲۹۲
γ_{12}	۰/۳	۰/۲۲۲	۰/۲۱۹	۰/۲۱۵	۰/۲۱۸	۰/۲۲۱	۰/۲۱۳	۰/۱۹۹	۰/۲۱۱
γ_{13}	۰/۵	۰/۵۲۱	۰/۵۲۳	۰/۵۳۴	۰/۵۳۱	۰/۵۲۴	۰/۵۲۳	۰/۵۲۵	۰/۵۳۵
γ_{14}	-۰/۵	-۰/۴۵۵	-۰/۴۵۲	-۰/۴۷۲	-۰/۴۵۱	-۰/۴۳۵	-۰/۴۳۷	-۰/۴۳۷	-۰/۴۵۳
γ_{15}	۰/۵	۰/۴۲	۰/۴۱۵	۰/۴۳۱	۰/۴۲۴	۰/۴۰۲	۰/۳۹۳	۰/۳۸۷	۰/۴۰۴
μ_1	.	-۰/۰۶۷	-۰/۰۶۳	-۰/۰۶۶	-۰/۰۶۷	-۰/۰۵۸	-۰/۰۶۲	-۰/۰۶۱	-۰/۰۵۶
μ_2	.	-۰/۰۰۸	-۰/۰۰۶	-۰/۰۶۳	-۰/۰۰۷	-۰/۰۰۴	-۰/۰۰۶	-۰/۰۰۶	-۰/۰۰۲
μ_3	.	-۰/۰۳۶	-۰/۰۳۳	-۰/۰۳۲	-۰/۰۳۱	-۰/۰۳	-۰/۰۳	-۰/۰۳۳	-۰/۰۳۱
μ_4	.	-۰/۰۶۲	-۰/۰۶۳	-۰/۰۶۳	-۰/۰۶۲	-۰/۰۶	-۰/۰۶۱	-۰/۰۶۳	۰/۰۶۲
μ_5	.	-۰/۱۱	-۰/۱۱۳	-۰/۱۱۴	-۰/۱۱۳	-۰/۱۰۹	-۰/۱۰۷	-۰/۱۱۲	-۰/۱۱
μ_6	.	-۰/۰۵۶	-۰/۰۶۱	-۰/۰۵	-۰/۰۵۶	-۰/۰۴۳	-۰/۰۹۲	-۰/۰۰۶	-۰/۰۰۷
ψ_δ	۰/۲۵	۰/۲۳۳	۰/۲۳۹	۰/۲۲۱	۰/۲۲۷	۰/۲۵۲	۰/۲۷۷	۰/۲۸۶	۰/۲۵۶
ψ_{11}	۰/۳	۰/۲۱۹	۰/۲۲۵	۰/۲۲۹	۰/۲۲۹	۰/۲۳۹	۰/۲۲۷	۰/۲۳۱	۰/۲۲۱
ψ_{22}	۰/۳	۰/۳۵۲	۰/۳۵۱	۰/۳۵۹	۰/۳۵۶	۰/۳۴۶	۰/۳۵	۰/۳۴۹	۰/۳۵۲
ψ_{33}	۰/۳	۰/۳۱۸	۰/۳۱۹	۰/۳۳	۰/۳۲۹	۰/۳۲۶	۰/۳۲۹	۰/۳۲۹	۰/۳۲۷
ψ_{44}	۰/۳	۰/۳۳۱	۰/۳۳	۰/۳۲۸	۰/۳۲۷	۰/۳۲۷	۰/۳۲۶	۰/۳۲۶	۰/۳۲۸
ψ_{55}	۰/۳	۰/۲۶۱	۰/۲۶۱	۰/۲۷۵	۰/۲۷۷	۰/۲۵۱	۰/۲۱۳	۰/۲۰۳	۰/۲۵۳
ψ_{66}	۰/۳	۰/۳۳۱	۰/۳۲۴	۰/۳۳۴	۰/۳۲۶	۰/۳۷	۰/۳۶۴	۰/۳۸۵	۰/۳۳۹
ϕ_{11}	۱	۱/۰۹۹	۱/۰۹۹	۱/۰۸۵	۱/۰۸۸	۱/۰۹	۱/۰۸۹	۱/۰۸۸	۱/۰۹۱
ϕ_{12}	۰/۲	۰/۳۰۴	۰/۳۰۶	۰/۳۰۶	۰/۲۸۷	۰/۲۷۹	۰/۲۸۸	۰/۲۸۹	۰/۳۰۱
ϕ_{22}	۱	۱/۲۸۱	۱/۲۸	۱/۲۶۳	۱/۲۶۳	۱/۲۶۳	۱/۲۹۳	۱/۳۲۱	۱/۲۸۶

۶- نتیجه گیری

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 برآورد پارامترهای مدل معادلات ساختاری غیر خطی مورد بررسی قرار گرفت. نتایج مطالعه‌ی شبیه‌سازی، کارایی الگوریتم معرفی‌شده توسط لی و همکاران [۱۰] را در برآورد 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. Assoc.*, **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 Markov 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). Structural 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 algorithm. *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

