مجله علوم آماری، پاییز و زمستان ۱۳۹۹
جلد ۱۴، شماره ۲، ص ۳۶۷ – ۳۸۸
DOI: 10.29252/jss.14.2.367

انتخاب متغیر در مدل‌های خطی-جزئی با اثرات آمیخته برای داده‌های طولی با بعد بالا

مژگان تعاونی و محمد آرشی

گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود

تاریخ دریافت: ۱۳۹۸/۰۶/۲۹ تاریخ آخرین بازنگری: ۱۳۹۹/۰۱/۲۳

چکیده: در این مقاله، مسئله برآورد و انتخاب متغیر همزمان در مدل‌های خطی-جزئی با اثرات آمیخته برای داده‌های طولی با بعد بالا در نظر گرفته شده است. مولفه ناپارامتری موجود در مدل با اسپلاین‌های رگرسیونی تقریب زده شده و سپس از طریق بهینه‌سازی تابع هدف مبتنی بر تابع تاوان ، برآورد و انتخاب متغیر به طور همزمان انجام می‌شود. در ادامه، رفتار حدی برآوردگرهای حاصل در چارچوب داده‌های طولی با بعد بالا که در آن تعداد پارامترها متناسب با افزایش حجم نمونه افزایش می‌یابد، مورد مطالعه قرار می‌گیرد. به منظور پیاده‌سازی روش برآورد پیشنهادی، یک الگوریتم تکراری مناسب برای انتخاب متغیرهای مهم و برآورد ضرایب غیر صفر ارائه گردیده است. در نهایت، عملکرد روش پیشنهادی با مطالعه شبیه‌سازی و تحلیل یک مجموعه داده واقعی مورد ارزیابی قرار گرفته است.

واژه‌های کلیدی: داده طولی، انتخاب متغیر، برآوردگر تاوانیده، مدل خطی-جزئی، اسپلاین هموارساز، بعد بالا، بیماری ایدز.

آدرس الکترونیکی نویسنده مسئول مقاله: محمد آرشی، m_arashi_stat@yahoo.com
کد موضوع‌بندی ریاضی (۲۰۱۰): 62G05, 62G20, 62H20.

۱ مقدمه

برای درنظر گرفتن همبستگی بین پاسخ‌ها در مطالعات طولی روش‌های متفاوتی گسترش یافته است. یکی از متداول‌ترین روش‌ها مدل اثرات آمیخته (لرد و ویر ، ۱۹۸۲) است که ناهمگنی بین پاسخ‌ها توسط مدل‌بندی مولفه تصادفی بررسی می‌گردد. این مدل با وجود سادگی در محاسبات و تفسیر، به دلیل فرضیات پارامتری محدود کننده از جمله اعمال رابطه خطی بین متغیرهای پاسخ و تبیینی، در مدل‌سازی بسیاری از روابط پیچیده با شکست مواجه می‌شود و مثال‌های بسیاری را می‌توان یافت که شناسایی و تشخیص اثرات رگرسیون بر اساس فرضیات بیان‌شده امکان‌پذیر نیست. در چنین شرایطی، آماردانان در راستای تحلیل هر چه بهتر داده‌های طولی با خلق ترکیبی از مدل‌های موجود، کلاس دیگری از مدل‌ها را ایجاد نموده‌اند که از جمله متداول‌ترین آن‌ها می‌توان به مدل خطی-جزئی[1] با اثرات آمیخته اشاره نمود. این مدل تعمیم مدل زیگر و دیگل (۱۹۹۴) است که برای تحلیل داده‌های خود از یک مدل نیمه پارامتری با عرض از مبدأ تصادفی استفاده کرده‌اند.

امروزه توسعه و گسترش سریع علوم در زمینه‌های مختلف و افزایش سرعت ثبت و تحلیل اطلاعات، آماردانان را با حجم عظیمی از اطلاعات پیچیده مواجه نموده است که بکارگیری روش‌های موجود را ناممکن ساخته است. رخداد مساله بعد بالا[2] در داده‌های طولی نیز اجتناب ناپذیر است. به منظور سهولت در امر تجزیه و تحلیل، اغلب منطقی و مفید است که فرض شود تنها تعداد کمی از متغیرها برای مدل‌سازی متغیر پاسخ مناسب هستند. این فرض، رویکرد نوینی تحت عنوان رگرسیون تاوانیده [3] را گسترش داد. پس از آن که فو (۲۰۰۳) و فن و لی (۲۰۰۴) به تعمیم رگرسیون تاوانیده به ترتیب در مدل‌های خطی تعمیم یافته و مدل‌های خطی-جزئی با داده‌های طولی پرداختند، مساله انتخاب متغیر در زمینه داده‌های طولی با پیشرفت چمشگیری روبرو شد. در این میان می‌توان به تحقیقات کانتونی و همکاران (۲۰۰۵)، وانگ و کوای (۲۰۰۹) و سوای و همکاران (۲۰۱۰) در مدل‌های خطی تعمیم یافته، باندل و همکاران (۲۰۱۰) در مدل اثرات آمیخته، نی و همکاران (۲۰۱۰) و ما و همکاران (۲۰۱۳) در مدل‌های خطی-جزئی اشاره نمود. با این وجود، مساله انتخاب متغیر برای داده‌های طولی با بعد بالا به واسطه چالش‌های تحمیل شده از جمله وجود همبستگی درون گروهی پیشینه چندانی نداشته و تنها می‌توان فعالیت‌های سو و همکاران (۲۰۱۳) و وانگ و همکاران (۲۰۱۲) را نام برد. اخیرا کاظمی و همکاران (۱۳۹۷)، در مدل‌های خطی-جزئی با بعد بالا مساله انتخاب متغیر را بررسی کرده است. در عین حال، تنها تحقیق موجود درباره انتخاب

[1] Partially linear model
[2] High dimension
[3] Penalized regression

متغیر در داده‌های طولی با بعد بالا، تعاونی و آرشی (۲۰۱۹) است، که تفاوت عمده مقاله حاضر با آن در فرضیات پایه‌ای و در نتیجه روش متفاوت برآورد پارامترها در توزیع نرمال بوده که بسیار ساده‌تر از مرجع اخیر است. در این مقاله برآورد و انتخاب متغیر در مدل خطی-جزئی با اثرات آمیخته مبتنی بر رهیافت توابع تاوان، در نظر گرفته شد. روش پیشنهادی به علت وجود توام اثرات تصادفی و مولفه ناپارامتری در مدل، نسبتاً با تحقیقات انجام شده متفاوت است. بخش ۲ مقاله به معرفی نمادگذاری متداول مدل می‌پردازد. در مرحله اول ابتدا مولفه ناپارامتری موجود در مدل با اسپلاین‌های رگرسیونی تقریب زده شده و سپس در بخش ۳ از طریق بهینه‌سازی تابع هدف مبتنی بر تابع تاوان مانند ماکسیمم درستنمایی تاوانیده[1] برآورد و انتخاب متغیر انجام می‌گیرد. از آنجایی که برآوردگر معرفی شده در تعامل با ماتریس کوواریانس پاسخ‌ها و اثرات تصادفی هستند، به منظور بهبود کارآیی برآوردگر ماکسیمم درستنمایی تاوانیده، از الگوریتم تکراری امیدگیری ماکسیمم‌سازی [2] (EM) استفاده می‌شود. علاوه بر این، نحوه انتخاب پارامترهای تابع تاوان در مورد بررسی قرار می‌گیرد. بخش ۴ به مطالعه خواص مجانبی برآوردگرها در چارچوب داده‌های بعد بالا که در آن تعداد پارامترها متناسب با افزایش حجم نمونه افزایش می‌یابد، می‌پردازد. برای بررسی عملکرد روش معرفی شده، مطالعه‌ای شبیه‌سازی و همچنین تحلیل داده‌های واقعی به ترتیب در بخش‌های ۵ و ۶ فراهم آمده است. بحث و نتیجه گیری بخش پایانی مقاله است.

۲ مدل خطی-جزئی با اثرات آمیخته

فرض کنید نمونه‌ای تصادفی از n آزمودنی وجود داشته باشد، به طوری که نمونه iام دارای n_i مشاهده مکرر در طول زمان است. لذا تعداد کل مشاهدات برابر با $N = \sum_{i=1}^{n} n_i$ است. همچنین، فرض کنید $y_i(t_{ij})$، $(i = 1, \ldots, n; j = 1, \ldots, n_i)$، پاسخ آزمودنی iام در زمان t_{ij} باشد. مدل خطی-جزئی با اثرات آمیخته به‌صورت

$$y_i(t_{ij}) = \boldsymbol{x}_i^{\top}(t_{ij})\boldsymbol{\beta} + g(t_{ij}) + \boldsymbol{z}_i^{\top}(t_{ij})\boldsymbol{b}_i + \varepsilon_i(t_{ij}),$$

است، که در آن $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_n})^{\top}$ برداری $p_n \times 1$ بعدی از ضرایب ثابت رگرسیونی وابسته به متغیرهای تبیینی $\boldsymbol{x}_i(t_{ij})$؛ $g(\cdot)$ تابعی نامعلوم، هموار و دارای مشتق دوم؛ $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})^{\top}$

[1] Penalized maximum likelihood
[2] Expectation maximization (EM)

برداری $q \times ۱$ بعدی از ضرایب تصادفی، مستقل از هم، دارای توزیع نرمال با میانگین ۰ و ماتریس کوواریانس $\boldsymbol{D}_i$ و وابسته به متغیرهای تبیینی $\boldsymbol{z}_i(t_{ij})$؛ $\varepsilon_i(\cdot)$ خطای تصادفی نرمال و مستقل از $\boldsymbol{b}_i$ها با $E\{\varepsilon_i(\cdot)\} = ۰$ هستند. بدون کاستن از کلیت مساله می‌توان فرض کرد که t_{ij}ها در بازه $[۰, ۱]$ هستند.

مشابه فن و همکاران (۲۰۰۷)، فرض می‌شود $Var\{\varepsilon_i(t_{ij})\} = \sigma^۲(t_{ij})$، یعنی واریانس خطاهای تصادفی تابعی ناپارامتری و هموار نسبت به زمان هستند. از طرف دیگر فرض می‌شود تابع همبستگی بین $\varepsilon_i(t_{ij})$ و $\varepsilon_i(t_{ik})$ دارای شکل تابعی $Corr\{\varepsilon_i(t_{ij}), \varepsilon_i(t_{ik})\} = \rho(t_{ij}, t_{ik}, \boldsymbol{\theta})$ است، به طوری که $\rho(t_{ij}, t_{ik}, \boldsymbol{\theta})$ تابعی معین مثبت و $\boldsymbol{\theta}$ بردار پارامترهای همبستگی است.

فرض کنید $\boldsymbol{y}_i = \big(y_i(t_{i۱}), \ldots, y_i(t_{in_i})\big)^\top$، $\boldsymbol{X}_i = \big(\boldsymbol{x}_i^\top(t_{i۱}), \ldots, \boldsymbol{x}_i^\top(t_{in_i})\big)^\top$، $\boldsymbol{g}(t_i) = (g(t_{i۱}), \ldots, g(t_{in_i}))^\top$، $\boldsymbol{Z}_i = \big(\boldsymbol{z}_i^\top(t_{i۱}), \ldots, \boldsymbol{z}_i^\top(t_{in_i})\big)^\top$، $\boldsymbol{\varepsilon}_i = \big(\varepsilon_i(t_{i۱}), \ldots, \varepsilon_i(t_{in_i})\big)^\top$، که در آن $\boldsymbol{y}_i$ بردار پاسخ n_i بعدی، $\boldsymbol{X}_i$ ماتریس $n_i \times p$ بعدی از اثرات ثابت، $\boldsymbol{Z}_i$ ماتریس $n_i \times q$ بعدی از اثرات تصادفی، $\boldsymbol{g}(t_i)$ بردار n_i بعدی از توابع نامعلوم و $\boldsymbol{\varepsilon}_i$ بردار n_i بعدی از خطاهای مدل است. مدل ماتریسی که عناصر آن براساس آزمودنی iام مشخص شده‌اند به صورت

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{g}(t_i) + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \qquad (۱)$$

نمایش داده می‌شود. فرض کنید $\boldsymbol{V}_i = \boldsymbol{Z}_i^\top\boldsymbol{D}_i\boldsymbol{Z}_i + \boldsymbol{\Sigma}_i$ ماتریس کوواریانس $n_i \times n_i$ بعدی مربوط به پاسخ $\boldsymbol{y}_i$ باشد که در آن $\boldsymbol{D}_i$ ماتریس کوواریانس $q \times q$ بعدی مربوط به $\boldsymbol{b}_i$ها و $\boldsymbol{\Sigma}_i$ ماتریس کوواریانس $n_i \times n_i$ بعدی مربوط به $\boldsymbol{\varepsilon}_i$ها است. به تبعیت از لیانگ و زیگر (۱۹۸۶)، $\boldsymbol{\Sigma}_i$ به صورت $\boldsymbol{\Sigma}_i = \boldsymbol{A}_i\boldsymbol{R}_i(\boldsymbol{\theta})\boldsymbol{A}_i$ در نظر گرفته می‌شود، به طوری که $\boldsymbol{A}_i = \mathrm{diag}\big(\sigma(t_{i۱}), \ldots, \sigma(t_{in_i})\big)$ یک ماتریس قطری $n_i \times n_i$ بعدی با عناصر قطری $\sigma(t_{ij})$ و $\boldsymbol{R}_i(\boldsymbol{\theta})$ ماتریس همبستگی بین $\varepsilon_i(t_{ij})$ و $\varepsilon_i(t_{ik})$ است که عضو (j,k)ام آن به صورت $\rho(t_{ij}, t_{ik}, \boldsymbol{\theta})$ تعریف می شود.

مدل (۱) را می‌توان به صورت مدل ماتریسی کلی $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{g}(t) + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}$ بازنویسی کرد، به طوری که $\boldsymbol{y} = \big(\boldsymbol{y}_۱^\top, \ldots, \boldsymbol{y}_n^\top\big)^\top$، $\boldsymbol{X} = \big(\boldsymbol{X}_۱^\top, \ldots, \boldsymbol{X}_n^\top\big)^\top$، $\boldsymbol{Z} = \mathrm{diag}\big(\boldsymbol{Z}_۱, \ldots, \boldsymbol{Z}_n\big)^\top$، $\boldsymbol{\varepsilon} = \big(\boldsymbol{\varepsilon}_۱^\top, \ldots, \boldsymbol{\varepsilon}_n^\top\big)^\top$ و $\boldsymbol{b} = \big(\boldsymbol{b}_۱^\top, \ldots, \boldsymbol{b}_n^\top\big)^\top$. همچنین $\boldsymbol{V} = \mathrm{diag}\big(\boldsymbol{V}_۱, \ldots, \boldsymbol{V}_n\big)$ ماتریس کوواریانس $N \times N$ بعدی مربوط به $\boldsymbol{y}$ بوده که در آن $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{D}_۱, \ldots, \boldsymbol{D}_n)$ $\boldsymbol{Z}^\top\boldsymbol{D}\boldsymbol{Z} + \boldsymbol{\Sigma}$ ماتریس کوواریانس $nq \times nq$ بعدی مربوط به بردار اثر تصادفی $\boldsymbol{b}$ و $\boldsymbol{\Sigma} = \mathrm{diag}\big(\boldsymbol{\Sigma}_۱, \ldots, \boldsymbol{\Sigma}_n\big)$ ماتریس کوواریانس $N \times N$ بعدی مربوط به $\boldsymbol{\varepsilon}$ است. سپس $g(t)$ با اسپلاین‌های رگرسیونی به صورت

$\boldsymbol{g}(t) = \boldsymbol{B}(t)^\top \boldsymbol{\alpha}$ تقریب می‌شود، که در آن $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{h_n})^\top$ بردار ضرایب رگرسیونی،

$$\boldsymbol{B}(t) = \big(1, t, \ldots, t^d, (t-t_1)_+^d, \ldots, (t-t_{L_n})_+^d\big)^\top,$$

$a_+ = \max(0, a)$ و $h_n = L_n + d + 1$ تعداد کل توابع پایه در $\boldsymbol{B}(t)$ است. این روش با تعداد کمی از گره‌ها تقریب خوبی را از یک تابع هموار فراهم می‌کند، همچنین با درنظر گرفتن توابع پایه به عنوان شبه‌متغیر تابعی خطی از مولفه‌های ناپارامتری در اختیار قرار می‌دهد. به این ترتیب مدل (۱) را می‌توان به صورت مدل رگرسیون خطی

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{B}_i(\boldsymbol{t}_i)\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \\ &= \tilde{\boldsymbol{X}}_i\boldsymbol{\theta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \end{aligned} \qquad (۲)$$

بیان نمود، که در آن $\boldsymbol{B}_i(\boldsymbol{t}_i) = \big(\boldsymbol{B}^\top(t_{i1}), \ldots, \boldsymbol{B}^\top(t_{in_i})\big)^\top$ یک ماتریس $n_i \times h_n$ بعدی است و $\tilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i, \boldsymbol{B}_i(\boldsymbol{t}_i))^\top$ ماتریس طرح ادغام شده $n_i \times (p_n + h_n)$ بعدی از ماتریس اثرات ثابت و اثرات اسپلاین و $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$ بردار پارامتر ادغام شده $(p_n + h_n) \times 1$ بعدی هستند.

۳ الگوریتم EM برای ماکسیمم درستنمایی تاوانیده

فرض کنید $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{D}, \boldsymbol{\Sigma},)$ مجموعه تمام پارامترهای مدل است. اگر اثرات تصادفی $\boldsymbol{b}_i$ را به عنوان مقادیر گمشده و مجموعه $\{(\boldsymbol{y}_i, \boldsymbol{b}_i), i = 1, \ldots, n\}$ داده‌های کامل در نظر گرفته شود، آنگاه لگاریتم تابع درستنمایی بدون در نظر گرفتن جمله ثابت به صورت

$$\ell_c(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \frac{1}{2}\big\{ \log|\boldsymbol{\Sigma}_i^{-1}| + \log|\boldsymbol{D}_i^{-1}| - [\boldsymbol{\varepsilon}_i^\top \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\varepsilon}_i + \boldsymbol{b}_i^\top \boldsymbol{D}_i^{-1}\boldsymbol{b}_i]\big\}, \qquad (۳)$$

است. برای ارزیابی امید شرطی (۳) به شرط داده‌های مشاهده شده $\boldsymbol{y} = \{\boldsymbol{y}_i, i = 1, \ldots, n\}$ و برآوردهای جاری $\widehat{\boldsymbol{\Theta}}^{(r)} = (\widehat{\boldsymbol{\theta}}^{(r)}, \widehat{\boldsymbol{D}}^{(r)}, \widehat{\boldsymbol{\Sigma}}^{(r)})$، از لم زیر استفاده می‌شود که نتیجه‌ای از قضیه ۳ در وانگ و فن (۲۰۱۱) است.

لم ۱. تحت مفروضات این بخش و مدل (۱)، داریم

$$\begin{aligned}\begin{bmatrix}\boldsymbol{y}_i\\ \boldsymbol{b}_i\end{bmatrix} &\sim \mathrm{N}_{n_i+q}\left(\begin{bmatrix}\tilde{\boldsymbol{X}}_i\boldsymbol{\theta}\\ \boldsymbol{0}\end{bmatrix},\ \begin{bmatrix}\boldsymbol{V}_i & \boldsymbol{Z}_i\boldsymbol{D}_i\\ \boldsymbol{D}_i\boldsymbol{Z}_i^\top & \boldsymbol{D}_i\end{bmatrix}\right),\\ \boldsymbol{b}_i|(\boldsymbol{y}_i) &\sim \mathrm{N}_q\big(\boldsymbol{D}_i\boldsymbol{Z}_i^\top\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\theta}),(\boldsymbol{D}_i^{-1}+\boldsymbol{Z}_i^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{Z}_i)^{-1}\big).\end{aligned}$$

مراحل اصلی الگوریتم EM به شرح زیر انجام می‌شود:

گام E: محاسبه تابع $Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})=\mathrm{E}\big(\ell_c(\boldsymbol{\Theta})|\boldsymbol{y},\widehat{\boldsymbol{\Theta}}^{(r)}\big)$، که برای $i=1,\ldots,n$ مجموع عبارت $Q_i(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})=\mathrm{E}\big(\ell_c^{[i]}(\boldsymbol{\Theta})|\boldsymbol{y}_i,\widehat{\boldsymbol{\Theta}}^{(r)}\big)$ است و

$$Q_i(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})=\tfrac{1}{2}\big\{\log|\boldsymbol{\Sigma}_i^{-1}|+\log|\boldsymbol{D}_i^{-1}|-\mathrm{tr}(\boldsymbol{D}_i^{-1}\widehat{\boldsymbol{B}}_i^{(r)})-\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}\widehat{\boldsymbol{\Psi}}_i^{(r)}(\boldsymbol{\theta}))\big\},$$

که در آن

$$\begin{aligned}\widehat{\boldsymbol{B}}_i^{(r)} &= \mathrm{E}(\boldsymbol{b}_i\boldsymbol{b}_i^\top|\boldsymbol{y}_i,\widehat{\boldsymbol{\Theta}}^{(r)})=\widehat{\boldsymbol{b}}_i^{(r)}\widehat{\boldsymbol{b}}_i^{(r)^\top}+\widehat{\boldsymbol{V}}_{\boldsymbol{b}_i}^{(r)},\\ \widehat{\boldsymbol{\Psi}}_i^{(r)}(\boldsymbol{\beta}) &= \mathrm{E}(\boldsymbol{e}_i\boldsymbol{e}_i^\top|\boldsymbol{y}_i,\widehat{\boldsymbol{\Theta}}^{(r)})\\ &= (\boldsymbol{y}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\theta}-\boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i^{(r)})(\boldsymbol{y}_i-\tilde{\boldsymbol{X}}_i\boldsymbol{\theta}-\boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i^{(r)})^\top+\boldsymbol{Z}_i\widehat{\boldsymbol{V}}_{\boldsymbol{b}_i}^{(r)}\boldsymbol{Z}_i^\top,\\ \widehat{\boldsymbol{b}}_i^{(r)} &= E(\boldsymbol{b}_i|\boldsymbol{y}_i,\widehat{\boldsymbol{\Theta}}^{(r)})=\widehat{\boldsymbol{D}}_i^{(r)}\boldsymbol{Z}_i^\top\widehat{\boldsymbol{V}}_i^{(r)^{-1}}(\boldsymbol{y}_i-\tilde{\boldsymbol{X}}_i\widehat{\boldsymbol{\theta}}^{(r)}),\\ \widehat{\boldsymbol{V}}_{\boldsymbol{b}_i}^{(r)} &= \mathrm{Cov}(\boldsymbol{b}_i|\boldsymbol{y}_i,\widehat{\boldsymbol{\Theta}}^{(r)})=\big(\widehat{\boldsymbol{D}}_i^{(r)}+\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\Sigma}}_i^{(r)^{-1}}\boldsymbol{Z}_i\big)^{-1}.\end{aligned}$$

گام M: به هنگام کردن $\widehat{\boldsymbol{\Theta}}^{(r)}$ توسط $\widehat{\boldsymbol{\Theta}}^{(r+1)}=\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})$.

برای انتخاب متغیرهای مهم و برآورد ضرایب آن‌ها، به لگاریتم تابع درستنمایی (۳) به صورت

$$\ell_c^{\mathrm{Pen}}(\boldsymbol{\Theta})=\ell_c(\boldsymbol{\Theta})-n\sum_{k=1}^{p}p_{\lambda_n}(|\beta_k|). \qquad (۴)$$

تاوانیده می‌شود، که در آن p_{λ_n} را تابع تاوان و λ_n را پارامتر تاوان می‌نامند. به طور مشابه تابع Q تاوانیده در گام E به صورت $Q_i^{\mathrm{Pen}}(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})=Q_i(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})-n\sum_{k=1}^{p_n}p_{\lambda_n}(|\beta_k|)$ در نظر

گرفته می‌شود. فن و لی (۲۰۰۱) یک تابع تاوان مقعر به نام اسکد[1] را معرفی کردند که دارای هر سه ویژگی مطلوب تنکی، نااریبی و پیوستگی است. این تابع تاوان دارای مشتق مرتبه اول به صورت

$$p'_{\lambda_n}(|\beta|) = \lambda_n\{I(|\beta| \le \lambda_n) + \frac{(a\lambda_n - |\beta|)_+}{(a-1)\lambda_n} I(|\beta| > \lambda_n)\}; \quad a > 2.$$

است. برای حل مسئله بهینه‌سازی درگام M، از روش تکراری تقریب موضعی توان دوم[2] (LQA) استفاده می‌شود (فن و لی ، ۲۰۰۱) . در این روش با استفاده از بسط تیلور و داشتن یک مقدار اولیه $\beta_{0,k}$، $|\beta_{0,k}| > 0$، مشتق تابع تاوان به صورت

$$[p_{\lambda_n}(|\beta_k|)]' = q_{\lambda_n}(|\beta_k|)\text{sign}(\beta_k) \approx \frac{q_{\lambda_n}(|\beta_{0,k}|)}{|\beta_{0,k}|}\beta_k,$$

تقریب زده می‌شود، که در آن $q_{\lambda_n}(|\beta_k|) = p'_{\lambda_n}(|\beta_k|)$ و $\text{sign}(a) = I(a > 0) - I(a < 0)$ مشتق مرتبه اول $p_{\lambda_n}(\cdot)$ است. هانتر و لی (۲۰۰۵) خاصیت همگرایی الگوریتم LQA را بررسی و برای محاسبه برآوردگر تاوانیده، الگوریتم MM را پیشنهاد دادند. در این الگوریتم برای کنترل عدم شناساپذیری برآوردگر در نقطه ۰، با قبول اندکی انحراف به میزان ϵ، تقریب LQA را بهبود دادند. سپس برآوردگر $\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}}\{Q_i(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)}) - n\sum_{k=1}^{p}\frac{q_{\lambda_n}(|\beta_{0,k}|)}{2(\epsilon+|\beta_{0,k}|)}\beta_k^2\}$ را براساس الگوریتم LQA با حل مسئله بهینه‌سازی به صورت بدست می‌آوردند. آنها خاطر نشان کردند که انتخاب مناسب ϵ برای شرط تنکی برآوردگر و همچنین سرعت همگرایی تقریب، ضروری است. فن و لی (۲۰۰۱) نشان دادند که الگوریتم LQA پس از تعداد کمی تکرار به همگرایی می‌رسد. در هر تکرار به محض اینکه برخی از مؤلفه‌های $|\beta_k|$ کوچکتر از $\epsilon = 10^{-6}$ شود، متغیرهای مربوط به آن‌ها به عنوان متغیر بی‌تاثیر شناخته و مقدار صفر برای آن‌ها در نظر گرفته می‌شود.

[1] SCAD
[2] Local quadratic approximation (LQA)

گام M: به هنگام کردن $\widehat{\boldsymbol{\beta}}^{(r)}$، $\widehat{\boldsymbol{D}}^{(r)}$ و $\widehat{\boldsymbol{\Sigma}}^{(r)}$ از طریق ماکسیمم کردن $Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(r)})$ که به صورت

$$\widehat{\boldsymbol{\theta}}^{(r+1)} = \Big(\sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i^{\top} \widehat{\boldsymbol{\Sigma}}_i^{(r)^{-1}} \tilde{\boldsymbol{X}}_i + n\boldsymbol{E}_n(\widehat{\boldsymbol{\theta}}^{(r)})^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i^{\top} \widehat{\boldsymbol{\Sigma}}_i^{(r)^{-1}} (\boldsymbol{y}_i - \boldsymbol{Z}_i \widehat{\boldsymbol{b}}_i^{(r)}),$$

$$\widehat{\boldsymbol{D}}^{(r+1)} = n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{B}}_i^{(r)}, \quad \widehat{\boldsymbol{\Sigma}}^{(r+1)} = n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{\Psi}}_i^{(r)}(\boldsymbol{\beta}),$$

حاصل می‌شوند، که در آن $\boldsymbol{E}_n(\widehat{\boldsymbol{\theta}}^{(r)}) = \text{diag}\{\frac{q_{\lambda_n}(|\beta_1|)}{\epsilon+|\beta_1|}, \ldots, \frac{q_{\lambda_n}(|\beta_p|)}{\epsilon+|\beta_p|}, \mathbf{0}_{h_n}\}$، $\epsilon = 10^{-6}$ و $\mathbf{0}_{h_n}$ برداری h_n بعدی با عناصر صفر است.

گام‌های E و M تا زمان همگرایی به مقداری ثابت تکرار می‌شود تا برآوردگر ماکسیمم درستنمایی تاوانیده برای پارامترهای $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{D}}, \widehat{\boldsymbol{\Sigma}})$ ، بدست آید.

۳.۱ انتخاب پارامترهای تاوان

در مطالعات عددی مقاله، از اسپلاین مکعبی ($d = 3$) استفاده می‌شود. این نکته که تعداد گره‌ها با افزایش حجم نمونه افزایش یابند اهمیت داشته و از سوی دیگر تعداد زیاد گره‌ها ممکن است واریانس برآوردگرها را افزایش دهد. بنابراین، تعداد گره‌ها باید به درستی انتخاب شود تا بین اریبی و واریانس برآوردگرها تعادل ایجاد گردد. برای سهولت در محاسبات معمولا گره‌هایی با فواصل برابر و تعداد گره‌ها به صورت $L_n \approx n^{1/(2r+1)}$ انتخاب می‌شود. این استراتژی در هی و همکاران (۲۰۰۲) ، چن و چو (۲۰۰۷) و سینها و ستار (۲۰۱۵) اجرا شده است. واضح است که $r > 0$ زیرا $r = 0$ تعداد n گره را در نظر می‌گیرد که انتخاب درستی نیست. معمولاً r را برابر ۱ در نظر می‌گیرند؛ به طور مثال برای نمونه‌هایی به حجم ۵۰، ۱۰۰ و ۲۰۰، تعداد گره‌ها به ترتیب برابر ۴، ۵ و ۶ خواهد بود. بنابراین اگر اسپلاین مکعبی و تعداد گره‌ها به صورت $L_n \approx n^{1/(2r+1)}$ در نظر گرفته شود، تعداد توابع پایه به صورت $h_n \approx n^{1/(2r+1)} + 4$ است؛ که برای حجم نمونه‌های ۵۰، ۱۰۰ و ۲۰۰ به ترتیب ۸، ۹ و ۱۰ تابع پایه وجود خواهد داشت. فن و لی (۲۰۰۱) از دیدگاه بیزی مقدار $a = 3.7$ را به عنوان یک مقدار مناسب برای مسائل مختلف پیشنهاد دادند و برای انتخاب مقدار مناسب پارامتر تنظیم λ_n، از معیار اعتبار سنجی متقابل تعمیم یافته به صورت

$$GCV(\lambda_n) = \frac{RSS(\lambda_n)/n}{(1 - tr\big(d(\lambda_n)\big)/n)^2}$$

استفاده کردند، که در آن $RSS(\lambda_n) = (\boldsymbol{y} - \tilde{\boldsymbol{X}}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{Z}^\top \widehat{\boldsymbol{b}})^\top \widehat{\boldsymbol{V}}^{-1} (\boldsymbol{y} - \tilde{\boldsymbol{X}}^\top \widehat{\boldsymbol{\theta}} - \boldsymbol{Z}^\top \widehat{\boldsymbol{b}})$ و

$$\begin{aligned} d(\lambda_n) &= \boldsymbol{X}(\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X} + \lambda_n \boldsymbol{I})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \\ &+ \boldsymbol{Z}\widehat{\boldsymbol{D}}\boldsymbol{Z}^\top (\boldsymbol{I} - \boldsymbol{X}[(\boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \widehat{\boldsymbol{V}}^{-1}]). \end{aligned}$$

مقدار λ_n از حل مساله بهینه‌سازی $\widehat{\lambda}_n = \underset{\lambda_n \in \Re^+}{\arg\min}\, GCV(\lambda_n)$ برآورد می‌شود.

۴ بررسی رفتار مجانبی برآوردگرها

این بخش به بیان توزیع مجانبی برآوردهای $\widehat{\boldsymbol{\beta}}_n$ و $\widehat{g}$ پرداخته است. در اکثر متون آماری در حیطه انتخاب متغیر به منظور سادگی در محاسبات، پارامترهای درست $\boldsymbol{\beta}_{n^\circ}$ به صورت $\boldsymbol{\beta}_{n^\circ} = (\boldsymbol{\beta}_{n^\circ 1}^\top, \boldsymbol{\beta}_{n^\circ 2}^\top)^\top$ و ماتریس طرح مدل به صورت $\boldsymbol{X}_i = \big(\boldsymbol{X}_{i(1)}, \boldsymbol{X}_{i(2)}\big)$ تفکیک می‌شوند. بدون کاستن از کلیت مساله فرض می شود $\boldsymbol{\beta}_{n^\circ 2} = \boldsymbol{0}$ و تمام اعضای $\boldsymbol{\beta}_{n^\circ 1}$ که برداری s_n^* بعدی است غیر صفر باشد. در مقاله حاضر، ضرایب رگرسیونی درست به صورت $\boldsymbol{\theta}_{n^\circ} = (\boldsymbol{\beta}_{n^\circ 1}^\top, \boldsymbol{\beta}_{n^\circ 2}^\top, \boldsymbol{\alpha}_{n^\circ}^\top)^\top$ است، که در آن $\boldsymbol{\alpha}_{n^\circ}$ برداری h_n بعدی و وابسته به تابع ناپارامتری $g_\circ$ است. به منظور جداسازی فضای پارامتر به فضای صفر و غیر صفر، $\boldsymbol{\theta}_{n^\circ}$ را به صورت $(\boldsymbol{\theta}_{n^\circ 1}^\top, \boldsymbol{\theta}_{n^\circ 2}^\top)^\top$ بازتفکیک کرده به طوری که $\boldsymbol{\theta}_{n^\circ 1} = (\boldsymbol{\beta}_{n^\circ 1}^\top, \boldsymbol{\alpha}_{n^\circ}^\top)^\top$ یک بردار $(s_n = s_n^* + h_n)$ بعدی است که تمام اعضای آن غیر صفراند و $\boldsymbol{\theta}_{n^\circ 2} = \boldsymbol{\beta}_{n^\circ 2} = \boldsymbol{0}$. متعاقبا، مقادیر برآورد و ماتریس طرح به ترتیب به صورت $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\theta}}_{n1}^\top, \widehat{\boldsymbol{\theta}}_{n2}^\top)^\top$ و $\tilde{\boldsymbol{X}}_i = \big(\tilde{\boldsymbol{X}}_{i(1)}^\top, \tilde{\boldsymbol{X}}_{i(2)}^\top\big)^\top$ بازتفکیک می‌شوند، که در آن $\widehat{\boldsymbol{\theta}}_{n1} = (\widehat{\boldsymbol{\beta}}_{n1}^\top, \widehat{\boldsymbol{\alpha}}_n^\top)^\top$ ، $\tilde{\boldsymbol{X}}_{i(1)} = \big(\boldsymbol{X}_{i(1)}^\top, \boldsymbol{B}_i(\boldsymbol{t}_i)^\top\big)^\top$، $\widehat{\boldsymbol{\theta}}_{n2} = \widehat{\boldsymbol{\beta}}_{n2}$ و $\tilde{\boldsymbol{X}}_{i(2)} = \boldsymbol{X}_{i(2)}$. فرض کنید $a_n = \max_{1 \le k \le p_n}\{p'_{\lambda_n}|\beta_{n^\circ,k}|,\ \beta_{n^\circ,k} \neq 0\}$ و $b_n = \max_{1 \le k \le p_n}\{p''_{\lambda_n}|\beta_{n^\circ,k}|,\ \beta_{n^\circ,k} \neq 0\}$ به طوری که، $\beta_{n^\circ,k}$ مولفه kام $\boldsymbol{\beta}_{n^\circ}$ است.

اگر گام M در الگوریتم EM چند جواب داشته باشد، تنها دنباله‌ای از برآوردگر سازگار $\widehat{\boldsymbol{\theta}}_n$ در نظر گرفته می‌شود.، یعنی وقتی $n \to \infty$ آنگاه $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n^\circ} \to \boldsymbol{0}$ و $\sup_t |\boldsymbol{B}^\top(t)\widehat{\boldsymbol{\alpha}}_n - g_\circ(t)| \xrightarrow{p} 0$. علاوه بر این، فرض کنید به ازای هر i، $n_i = m < \infty$ باشد؛ مجموعه سه تایی $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{t}_i)$، $i = 1, \ldots, n$ متغیرهایی مستقل و هم توزیع باشند و مشتق kام تابعی مانند $d(\cdot)$ به صورت $d^{(k)}(\cdot)$ تعریف شود. برای مطالعه رفتار حدی برآوردگرها، شرایط نظم زیر را در نظر بگیرید.

(۱.A) به ازای هر $k \ge 2$، مشتق kام تابع $\boldsymbol{g}_\circ(\boldsymbol{t}_i)$ کراندار است.

(۲.A) ماتریس کوواریانس برآورد شده $\widehat{\boldsymbol{\Sigma}}_i$ در شرط $\|\widehat{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}_i\| = O_p(n^{-1/2})$ صدق می‌کند.

(۳.A) $b_1 \leq \lambda_{\min}(n^{-1}\sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i^\top \boldsymbol{\Sigma}_i^{-1}\tilde{\boldsymbol{X}}_i) \leq \lambda_{\max}(n^{-1}\sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i^\top \boldsymbol{\Sigma}_i^{-1}\tilde{\boldsymbol{X}}_i) \leq b_2$، که در آن b_1 و b_2 مقادیر ثابت و مثبت و $\lambda_{\min}$ و $\lambda_{\max}$ کوچکترین و بزرگترین مقدار ویژه هستند.

(۴.A) وقتی $n \longrightarrow \infty$، آنگاه $\min_{1\leq k\leq s_n} |\theta_{n0k}|/\lambda_n \to \infty$، $\lambda_n \to 0$، $s_n^3 n^{-1} = o(1)$، $s_n^2(\log n)^4 = o(n\lambda_n^2)^2$، $\log(p_n) = o(n\lambda_n^2/(\log n)$، $p_n s_n^4(\log n)^6 = o(n^2\lambda_n^2)$ و $p_n s_n^3(\log n)^8 = o(n^2\lambda_n^4)$.

(۵.A) تابع تاوان در شرایط $\liminf_{n\to\infty}\liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$، $a_n = O(n^{-\frac{1}{2}})$، $b_n \to \infty$، صدق می‌کند و همچنین مقادیر ثابت C و D وجود دارند به طوری که به ازای هر $\theta_1, \theta_2 > \lambda_n C$ خواهیم داشت $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \leq D|\theta_1 - \theta_2|$.

ملاحظه ۱. *در مدل اولیه فرض می‌شود که $g(\cdot)$ تابعی هموار است، یعنی کلیه مشتقات آن وجود داشته باشد، لذا شرط (۱.A) برقرار است. شرط (۲.A) مشابه شرط استفاده شده در لیانگ و زیگر (۱۹۸۶) است که فرض می‌کند پارامتر همبستگی $\widehat{\boldsymbol{\theta}}$ در شرط $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = O_p(1)$ صدق می‌کند. با توجه به این که دنباله برآوردگرهای سازگار در الگوریتم EM در نظر گرفته می‌شود، این شرط نیز برقرار است. شرط (۳.A) یکی از شرایط معمول در رگرسیون برای داده‌های مستقل است. شرایط (۴.A) و (۵.A) از شرایط معمول در متون انتخاب متغیر است که توسط فن و لی (۲۰۰۱) اثبات شده است.*

قضیه ۱. *با درنظرگرفتن شرایط (۱.A)-(۵.A)، وقتی $n \longrightarrow \infty$ و $n^{-1}p_n^2 = o(1)$، آنگاه $\widehat{\boldsymbol{\theta}}_n$ یک جواب الگوریتم EM است، به طوری که*

$$i.\ \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{n0}\| = O_p\big(\sqrt{p_n}(n^{-1/2} + a_n)\big),$$
$$ii.\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big(\widehat{g}(t_{ij}) - g_0(t_{ij})\big)^2 = O_p(n^{-2m/(2m+1)}).$$

برهان: برای اثبات قسمت (i)، فرض کنید $\alpha_n = \sqrt{p_n}(n^{-1/2} + a_n)$ و $\|\boldsymbol{U}\| = C$، به طوری که C یک مقدار ثابت و به اندازه کافی بزرگ است. نشان داده می‌شود که به ازای هر $\xi > 0$ یک مقدار ثابت C وجود دارد که $\mathcal{P}\big\{\inf_{\|\boldsymbol{U}\|=C} \ell_c^{Pen}(\boldsymbol{\theta}_{n0} + \alpha_n\boldsymbol{U}) \geq \ell_c^{Pen}(\boldsymbol{\theta}_{n0})\big\} \geq 1 - \xi$. به عبارت دیگر با احتمال نزدیک به ۱، در مجموعه $\{\boldsymbol{\theta}_{n0} + \alpha_n\boldsymbol{U} : \|\boldsymbol{U}\| \leq C\}$ یک جواب محلی برای $\ell_c^{Pen}(\boldsymbol{\theta}_{n0})$

وجود دارد. با استفاده از $p_{\lambda_n}(۰) = ۰$، داریم:

$$\begin{aligned} D_n(\boldsymbol{U}) &= \ell_c^{Pen}(\boldsymbol{\theta}_{n^\circ} + \alpha_n \boldsymbol{U}) - \ell_c^{Pen}(\boldsymbol{\theta}_{n^\circ}) \\ &\geq \ell_c(\boldsymbol{\theta}_{n^\circ} + \alpha_n \boldsymbol{U}) - \ell_c(\boldsymbol{\theta}_{n^\circ}) + n \sum_{i=۱}^{n} \{p_{\lambda_n}(|\theta_{n^\circ,k} + \alpha_n \boldsymbol{U}_k|) - p_{\lambda_n}(|\theta_{n^\circ,k}|)\} \\ &\equiv I + II \end{aligned}$$

با انجام محاسبات جبری داریم

$$I = \frac{n\alpha_n^۲}{۲} \boldsymbol{U}^\top \widehat{\boldsymbol{\Gamma}}_n \boldsymbol{U} - \alpha_n \boldsymbol{U}^\top \sum_{i=۱}^{n} \tilde{\boldsymbol{X}}_i \boldsymbol{\Sigma}_i^{-۱} (\boldsymbol{Y}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\theta} - \boldsymbol{Z}_i \boldsymbol{b}_i) \equiv I_۱ - I_۲,$$

که در آن $\boldsymbol{\Gamma}_n = n^{-۱} \sum_{i=۱}^{n} \tilde{\boldsymbol{X}}_i^\top \boldsymbol{\Sigma}_i^{-۱} \tilde{\boldsymbol{X}}_i$ و $\widehat{\boldsymbol{\Gamma}}_n = n^{-۱} \sum_{i=۱}^{n} \tilde{\boldsymbol{X}}_i^\top \widehat{\boldsymbol{\Sigma}}_i^{-۱} \tilde{\boldsymbol{X}}_i$. عبارت‌های $I_۱$ و $I_۲$ را به صورت

$$\begin{aligned} I_۱ &= \frac{n\alpha_n^۲}{۲} \boldsymbol{U}^\top (\boldsymbol{\Gamma}_n - \widehat{\boldsymbol{\Gamma}}_n + \widehat{\boldsymbol{\Gamma}}_n) \boldsymbol{U} = \frac{n\alpha_n^۲}{۲} \boldsymbol{U}^\top \widehat{\boldsymbol{\Gamma}}_n \boldsymbol{U} + O_p(n^{-۱/۲}) n\alpha_n^۲ \|\boldsymbol{U}\|^۲ \\ I_۲ &= \alpha_n \boldsymbol{U}^\top \sum_{i=۱}^{n} \tilde{\boldsymbol{X}}_i \boldsymbol{\Sigma}_i^{-۱} \boldsymbol{\epsilon}_i \end{aligned}$$

خواهیم داشت. با استفاده از شرط (۳.A) می‌توان نشان داد

$$\begin{aligned} |\alpha_n \boldsymbol{U}^\top \sum_{i=۱}^{n} \tilde{\boldsymbol{X}}_i \boldsymbol{\Sigma}_i^{-۱} \boldsymbol{\epsilon}_i| &\leq \alpha_n \| \sum_{i=۱}^{n} \tilde{\boldsymbol{X}}_i \boldsymbol{\Sigma}_i^{-۱} \varepsilon_i \| \|\boldsymbol{U}\| \\ &= O_p(\alpha_n \sqrt{n p_n}) \|\boldsymbol{U}\| \\ &= O_p(\alpha_n^۲ n) \|\boldsymbol{U}\|. \end{aligned}$$

بنابراین عبارت $I_۲$ به صورت $I_۲ = O_p(\alpha_n^۲ n)\|\boldsymbol{U}\|$ بدست می‌آید. مشابه اثبات فن و لی (۲۰۰۱) می‌توان نشان داد

$$II = O_p(\sqrt{s_n} n \alpha_n a_n) \|\boldsymbol{U}\| + O_p(n\alpha_n^۲ b_n) \|\boldsymbol{U}\|.$$

بنابراین از شرط (۳.A) و اینکه $\|\boldsymbol{U}\|$ به اندازه کافی بزرگ باشد، می‌توان دریافت که عبارت‌های $I_۲$ و II تحت الشعاع عبارت $I_۱$ هستند که این عبارت مقداری مثبت است. بنابراین قسمت (i) اثبات می‌شود.

برای اثبات (ii)، بنابر قضیه ۱۲.۷ در اسچوماکر (۱۹۸۱)، تحت شرط (۱.A) مقدار ثابت C وجود دارد که $\sup_{t\in[۰,۱]}|g_۰(\boldsymbol{t})-\boldsymbol{B}(\boldsymbol{t})\boldsymbol{\alpha}_۰|\leq CL_n^{-m}$. بنابراین با استفاده از $L_n\approx n^{۱/(۲m+۱)}$ داریم

$$\begin{aligned}\big(\widehat{g}(t)-g_۰(t)\big)^۲ &= \big(\boldsymbol{B}(t)\widehat{\boldsymbol{\alpha}}_n-g_۰(t)\big)^۲\simeq\big(\boldsymbol{B}(t)\boldsymbol{\alpha}_{n^۰}-g_۰(\boldsymbol{t})\big)^۲\\ &\leq |\boldsymbol{B}(t)\boldsymbol{\alpha}_{n^۰}-g_۰(t)||\boldsymbol{B}(t)\boldsymbol{\alpha}_{n^۰}-g_۰(t)|\\ &\leq \sup_{t\in[۰,۱]}|\boldsymbol{B}(t)\boldsymbol{\alpha}_{n^۰}-g_۰(t)|\sup_{t\in[۰,۱]}|\boldsymbol{B}(t)\boldsymbol{\alpha}_{n^۰}-g_۰(t)|\\ &\leq CL_n^{-m}CL_n^{-m}=C^۲n^{-۲m/(۲m+۱)}=O_p(n^{-۲m/(۲m+۱)}).\end{aligned}$$

قضیه ۲. *تحت شرایط (۱.A) تا (۵.A)، وقتی $n\to\infty$، اگر $L_n=O_p(n^{۱/(۲m+۱)})$ و $n^{-۱}p_n^۳=o(۱)$، آنگاه*

$$\begin{aligned}&i.\ \mathbb{P}(\widehat{\boldsymbol{\beta}}_{n^۲}=\mathbf{0})\to ۱,\\ &ii.\ \widehat{\boldsymbol{\Theta}}\xrightarrow{p}\boldsymbol{\Theta},\ \ \sqrt{n}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta})\xrightarrow{d}N(\mathbf{0},\boldsymbol{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-۱}),\ \ (\widehat{\boldsymbol{\beta}}_{n^۱}-\boldsymbol{\beta}_{n^۰۱})\xrightarrow{d}\boldsymbol{N}_{s_n}(\mathbf{0},\boldsymbol{I}^{-۱}_{\boldsymbol{\beta}_{n^۰۱}\boldsymbol{\beta}_{n^۰۱}}),\end{aligned}$$

که در آن $\boldsymbol{\eta}=(vec(D)^\top,vec(\boldsymbol{\Sigma})^\top)^\top$، $n^{-۱}\boldsymbol{J}_{\boldsymbol{\beta}_{n^۰۱}\boldsymbol{\beta}_{n^۰۱}}\to\boldsymbol{I}_{\boldsymbol{\beta}_{n^۰۱}\boldsymbol{\beta}_{n^۰۱}}$، $n^{-۱}\boldsymbol{J}_{\boldsymbol{\eta}\boldsymbol{\eta}}\to\boldsymbol{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}$، $\boldsymbol{J}_{\boldsymbol{\beta}_{n^۰۱}\boldsymbol{\beta}_{n^۰۱}}$ و $\boldsymbol{J}_{\boldsymbol{\eta}\boldsymbol{\eta}}$ ماتریس‌های اطلاع فیشر هستند و $\xrightarrow{p}$ ، $\xrightarrow{d}$ به ترتیب بیانگر همگرایی در احتمال و توزیع هستند.

برهان: برای اثبات قسمت (i) باید نشان داد که به ازای هر $\boldsymbol{\theta}_{n۱}$ ای که $\|\boldsymbol{\theta}_{n۱}-\boldsymbol{\theta}_{n^۰۱}\|=O_p(\sqrt{p_n/n})$ و هر مقدار ثابت C، $(\boldsymbol{\theta}_{n۱}^\top,\mathbf{0})^\top$ جواب بهینه تابع ماکسیمم درستنمایی تاوانیده است. به عبارت دیگر

$$\ell_c^{Pen}(\boldsymbol{\theta}_{n۱}^\top,\mathbf{0})^\top\}=\max_{\|\boldsymbol{\theta}_{n^۲}\|\leq C\sqrt{p_n/n}}\ell_c^{Pen}(\boldsymbol{\theta}_{n۱}^\top,\boldsymbol{\theta}_{n^۲}^\top)^\top\}.$$

برای اثبات کافی است نشان داده شود عبارت

$$\frac{\partial \ell_c^{Pen}(\boldsymbol{\theta}_n)}{\partial \theta_{n,k}} = \begin{cases} < 0 & -\xi_n < \theta_{n,k} < 0 \\ > 0 & 0 < \theta_{n,k} < \xi_n \end{cases}$$

به ازای هر $\theta_{n,k} \neq 0$ و $k = s_n + 1, \ldots, p_n$ برقرار است، که در آن $\xi_n = C\sqrt{p_n/n}$. واضح است که $\frac{\partial \ell_c^{Pen}(\boldsymbol{\theta}_n)}{\partial \theta_{n,k}} = \frac{\partial \ell_c(\boldsymbol{\theta}_n)}{\partial \theta_{n,k}} + np'_{\lambda_n}(|\theta_{n,k}|)\text{sgn}(\theta_{n,k})$ و مشابه فن و لی (۲۰۰۱) می‌توان نشان داد $\frac{\partial \ell_c(\boldsymbol{\theta}_n)}{\partial \theta_{n,k}} = O_p(\sqrt{np_n})$. بنابراین

$$\frac{\partial \ell_c^{Pen}(\boldsymbol{\theta}_n)}{\partial \theta_{n,k}} = n\lambda_n \Big\{ \frac{p'_{\lambda_n}(|\theta_{n,k}|)}{\lambda_n} \text{sgn}(\theta_{n,k}) + O_p\Big(\frac{\sqrt{p_n/n}}{\lambda_n}\Big) \Big\}.$$

از آنجایی که $\frac{\sqrt{p_n/n}}{\lambda_n} \to 0$ و $\liminf_{n\to\infty} \liminf_{\theta \to 0_+} p'_{\lambda_n}(\theta)/\lambda_n > 0$، علامت $\theta_{n,k}$ تعیین کننده علامت $\frac{\partial \ell_c^{Pen}\{(\boldsymbol{\theta}_n)}{\partial \theta_{n,k}}$ است. بنابراین اثبات کامل می‌شود.

بخش همگرایی در احتمال و همگرایی در توزیع مستقیما از طریق قانون ضعیف اعداد بزرگ، قضیه حد مرکزی و قضیه اسلاتسکی اثبات می‌شود.

۵ مطالعه شبیه‌سازی

در این بخش، فرایند برازش مدل، برآورد پارامترهای مدل و انتخاب متغیر به صورت عددی ارزیابی می‌شود. ابتدا مجموعه‌ای از داده‌های شبیه‌سازی شده توسط روش پیشنهادی مقاله یعنی برآوردگر ماکسیمم درستنمایی تاوانیده در مدل خطی-جزئی با اثرات آمیخته (P-PLMM) مورد تحلیل قرار می‌گیرد. سپس به منظور بررسی عملکرد روش P-PLMM در مسئله برآورد و انتخاب متغیر همزمان، نتایج حاصل از تحلیل داده‌ها بدون رهیافت تابع تاوان یعنی روش برآوردگر ماکسیمم درستنمایی در مدل خطی-جزئی با اثرات آمیخته (PLMM) نیز گزارش می‌شود. در نظر داشته باشید که اگر داده‌ها ذاتاً دارای جزء ناپارامتری باشند، مدل خطی-جزئی بهتر از مدل خطی عمل می‌کند. برای نشان دادن این موضوع روش P-PLMM با روش برآوردگر ماکسیمم درستنمایی تاوانیده در مدل خطی با اثرات آمیخته (P-LMM) مقایسه می‌شود.

متغیرهای پاسخ از مدل زیر شبیه‌سازی می‌شوند

$$y_i(t_{ij}) = x_{۱,i}(t_{ij})\beta_۱ + x_{۲,i}(t_{ij})\beta_۲ + g(t_{ij}) + b_i + \varepsilon_i(t_{ij}),$$

که در آن مولفه‌ها به صورت $i = ۱, \ldots, n$، $n = ۵۰, ۱۰۰, ۲۰۰$، $j = ۱, \ldots, n_i$، $n_i = ۵$، $b_i \sim N(۰, \sigma_b^۲)$، $\sigma_b^۲ = ۰/۲۵$، $\boldsymbol{\beta}^\top = (۲, ۳, ۱/۵, ۲, ۰, \ldots, ۰)$، $\boldsymbol{X}_{ij}^\top = (x_{ij}^{(۱)}, \ldots, x_{ij}^{(p_n)})$، $x_{ij}^{(k)} \sim U(-۱, ۱)$ و $t_{ij} \sim U(۰, ۱)$ تعیین شده‌اند. به منظور انتخاب تعداد مولفه‌های پارامتری مدل، نویسندگان پیشنهادهای متفاوتی از جمله $p_n = [\frac{n}{۲}]$، $p_n = [۴/۵n^{۱/۴}]$ و $p_n = [\frac{n}{b\log(n)}]$ داشته‌اند که در آن‌ها $b > ۱$ و $[a]$ جزء صحیح مقدار a است. موارد بیان شده برای حالت $p_n < n$ است. برای زمانی که $p_n > n$ نیز نرخ‌هایی مانند $\log(p_n) = o_p(n^b)$ پیشنهاد شده است که در آن $۰ < b < ۱$. انتخاب مقادیر p_n بسیار بزرگتر از n چالش‌هایی را به همراه خواهد داشت، به طوری که در این حالت اگر چه احتمال انتخاب متغیرهای موثر در مدل افزایش می‌یابد در عین حال متغیرهای بی‌تاثیر بیشتری در مدل وارد خواهند شد که این موضوع بر عملکرد روش‌های برآورد و انتخاب متغیر تاثیر منفی خواهد گذاشت. به منظور همپوشانی تمامی نرخ‌های پیشنهادی، موارد $(۵۰, ۱۱), (۱۰۰, ۱۴), (۲۰۰, ۱۶)$ برای حالت $p_n < n$ و موارد $(۳۰, ۱۰۰), (۱۰۰, ۵۰۰), (۲۰۰, ۲۰۰۰)$ برای حالت $p_n > n$ در نظر گرفته شده است.

دقت برآورد هر یک از روش‌ها توسط میانگین توان دوم خطاها (MSE) براساس ۱۰۰ تکرار از شبیه‌سازی، مورد سنجش قرار گرفت که به صورت $MSE = \sum_{s=۱}^{۱۰۰} \|\widehat{\boldsymbol{\beta}}_n^s - \boldsymbol{\beta}_{n۰}\|/۱۰۰$ تعریف می‌شود که در آن $\widehat{\boldsymbol{\beta}}_n^s$ برآوردگر $\boldsymbol{\beta}_{n۰}$ در s-امین مجموعه داده تولید شده است. ارزیابی عملکرد روش‌ها در مساله انتخاب متغیر توسط معیارهای (C, I) سنجیده می‌شود به طوری که C میانگین ضرایب صفر که به درستی صفر تخمین زده شده‌اند و I میانگین ضرایب غیر صفر است که به اشتباه صفر شده‌اند. برای ارائه توصیف جامع‌تر از عملکرد مساله انتخاب متغیر روش‌ها از معیارهای دیگری نیز استفاده شده است. UF بیانگر نسبت تکرارهایی است که ضرایب تاثیر گذار را به اشتباه صفر درنظر گرفته است، CF بیانگر نسبت تکرارهایی است که مدل را به درستی تعیین کرده‌اند و نسبت تکرارهایی که علاوه بر متغیرهای مهم، متغیرهای بی‌تاثیر را نیز در مدل وارد کرده‌اند با OF نشان داده شده است.

جدول ۱، نتایج مربوط به عملکرد روش‌های P-LMM، P-PLMM و PLMM در برآورد و انتخاب مدل برای مقادیر مختلف (n, p_n) است. از نظر دقت برآورد روش PLMM، MSE بیشتری نسبت به دو روش دیگر دارد، در حالیکه دو روش P-LMM و P-PLMM بسیار نزدیک به هم عمل می‌کنند. با

جدول ۱. مقایسه روش‌های ۱: PLMM، ۲: P-LMM و ۳: P-PLMM

	حالت $p_n < n$						حالت $p_n > n$					
	$(n,p) = (۵۰, ۱۱)$						$(n,p) = (۵۰, ۱۰۰)$					
روش	MSE	$C(۷)$	$I(۰)$	UF	CF	OF	MSE	$C(۷)$	$I(۰)$	UF	CF	OF
۱	۰٫۴۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰	۸٫۰۵	۰٫۱۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰
۲	۰٫۱۸	۶٫۲۸	۰٫۰۰	۰٫۰۰	۰٫۵۰	۰٫۵۰	۰٫۴۶	۹۴٫۱۹	۰٫۰۰	۰٫۰۰	۰٫۲۳	۰٫۷۷
۳	۰٫۱۹	۶٫۴۸	۰٫۰۰	۰٫۰۰	۰٫۵۴	۰٫۴۶	۰٫۳۹	۹۴٫۶۵	۰٫۰۰	۰٫۰۰	۰٫۲۷	۰٫۷۳
	$(n,p) = (۱۰۰, ۱۴)$						$(n,p) = (۱۰۰, ۵۰۰)$					
۱	۰٫۲۷	۰٫۰۲	۰٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰	۱۴۹۹٫۱۴	۴۷٫۰۲	۰٫۰۳	۰٫۰۳	۰٫۰۰	۱٫۰۰
۲	۰٫۰۹	۹٫۴۷	۰٫۰۰	۰٫۰۰	۰٫۶۲	۰٫۳۸	۰٫۰۶	۴۹۵٫۷۲	۰٫۰۰	۰٫۰۰	۰٫۳۰	۰٫۷۰
۳	۰٫۰۹	۹٫۵۶	۰٫۰۰	۰٫۰۰	۰٫۶۹	۰٫۳۱	۰٫۰۴	۴۹۶٫۲۵	۰٫۰۰	۰٫۰۰	۰٫۵۴	۰٫۴۶
	$(n,p) = (۲۰۰, ۱۶)$						$(n,p) = (۲۰۰, ۲۰۰۰)$					
۱	۰٫۱۴	۰٫۰۹	۰٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰	۱۲۵٫۴۱	۱۱۳۷٫۶۲	۰٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰
۲	۰٫۰۴	۱۱٫۹۳	۰٫۰۰	۰٫۰۰	۰٫۹۳	۰٫۰۷	۰٫۰۲	۱۹۹۶٫۸۹	۰٫۰۰	۰٫۰۰	۰٫۸۹	۰٫۱۱
۳	۰٫۰۴	۱۱٫۹۶	۰٫۰۰	۰٫۰۰	۰٫۹۶	۰٫۰۴	۰٫۰۱۸	۱۹۹۶٫۹۳	۰٫۰۰	۰٫۰۰	۰٫۹۲	۰٫۰۸

جدول ۲. برآوردهای P-PLMM در حالت‌های $p_n < n$ و $p_n > n$.

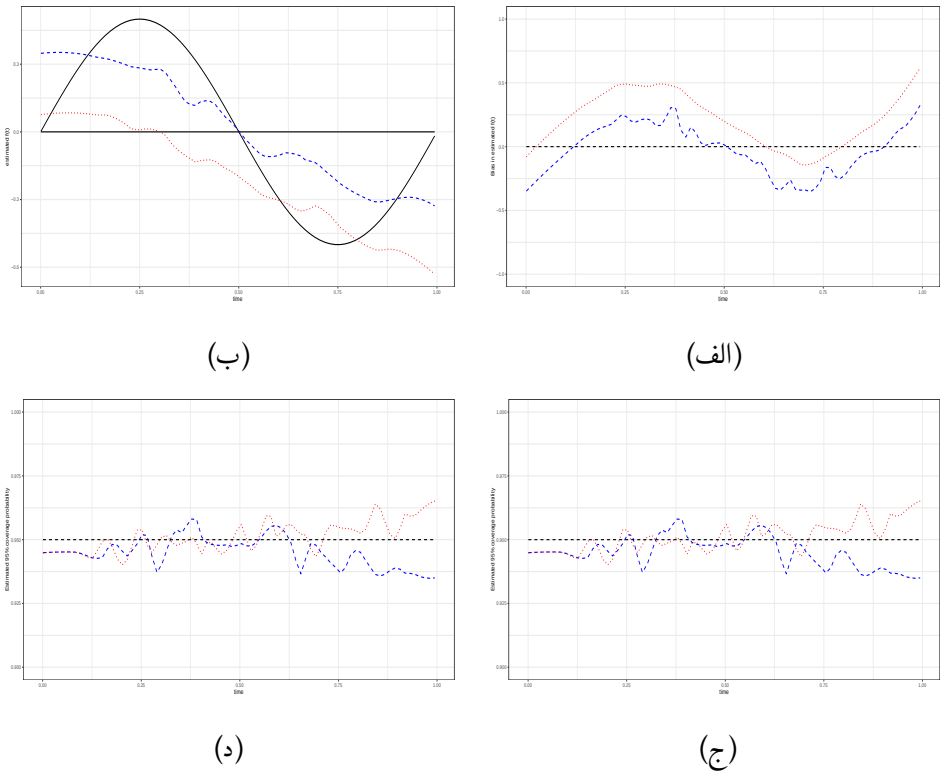
	حالت $p_n < n$					حالت $p_n > n$				
(n,p_n)	معیار	$\beta_۱$	$\beta_۲$	$\beta_۳$	$\beta_۴$	(n,p_n)	$\beta_۱$	$\beta_۲$	$\beta_۳$	$\beta_۴$
	$Bias$	۰٫۰۴	۰٫۰۱	۰٫۰۰	۰٫۰۱		۰٫۰۸	۰٫۱۱	۰٫۰۵	۰٫۰۶
(۵۰, ۱۱)	$SD۱$	۰٫۱۰	۰٫۱۱	۰٫۱۱	۰٫۱۰	(۵۰, ۱۰۰)	۰٫۰۹	۰٫۰۹	۰٫۰۸	۰٫۰۸
	$SD۲$	۰٫۰۷	۰٫۰۸	۰٫۰۸	۰٫۰۸		۰٫۰۹	۰٫۰۸	۰٫۰۷	۰٫۰۷
	CP	۰٫۹۵	۰٫۹۷	۰٫۹۵	۰٫۹۱		۰٫۹۶	۰٫۹۶	۰٫۹۹	۰٫۹۳
	$Bias$	۰٫۰۱	۰٫۰۵	۰٫۰۱	۰٫۰۲		۰٫۰۴	۰٫۰۳	۰٫۰۲	۰٫۰۳
(۱۰۰, ۱۴)	$SD۱$	۰٫۰۷	۰٫۰۸	۰٫۰۷	۰٫۰۸	(۱۰۰, ۵۰۰)	۰٫۰۷	۰٫۰۸	۰٫۰۷	۰٫۰۸
	$SD۲$	۰٫۰۵	۰٫۰۵	۰٫۰۵	۰٫۰۵		۰٫۰۵	۰٫۰۵	۰٫۰۵	۰٫۰۵
	CP	۰٫۹۱	۰٫۹۴	۰٫۹۵	۰٫۹۴		۰٫۹۳	۰٫۹۳	۰٫۹۵	۰٫۹۴
	$Bias$	۰٫۰۴	۰٫۰۰	۰٫۰۱	۰٫۰۱		۰٫۰۲	۰٫۰۴	۰٫۰۱	۰٫۰۱
(۲۰۰, ۱۶)	$SD۱$	۰٫۰۶	۰٫۰۶	۰٫۰۶	۰٫۰۶	(۲۰۰, ۲۰۰۰)	۰٫۰۸	۰٫۰۸	۰٫۰۷	۰٫۰۸
	$SD۲$	۰٫۰۴	۰٫۰۴	۰٫۰۴	۰٫۰۴		۰٫۰۵	۰٫۰۵	۰٫۰۵	۰٫۰۶
	CP	۰٫۹۶	۰٫۹۵	۰٫۹۵	۰٫۹۵		۰٫۹۳	۰٫۹۶	۰٫۹۲	۰٫۹۴

این وجود MSE روش پیشنهادی همواره کمتر از P-LMM است. از نظر انتخاب مدل مشاهده می‌شود که PLMM انتخاب مدل انجام نمی‌دهد اما روش‌های P-LMM و P-LMM تمامی متغیرهای مهم را انتخاب کرده است زیرا معیار I آن‌ها صفر است. همچنین روش پیشنهادی به دلیل داشتن مقادیر بالا در نرخ‌های C و CF و مقادیر کمتر در OF، نسبت به رقیب خود یعنی P-LMM عملکرد بهتری داشته است. با توجه به معیارهای CF و OF، هنگامی که $p_n > n$ نسبت به حالت $p_n < n$ روش‌ها در انتخاب مدل عملکرد پایینی دارند و ضرایب صفر تمایل بیشتری دارند که در مدل گنجانده شوند، اما با

افزایش حجم نمونه هر دو معیار بهبود می‌یابند. برای بررسی بیشتر عملکرد روش پیشنهادی، نتایج مربوط به اریبی، خطای استاندارد مجانبی، خطای استاندارد نمونه‌ای و احتمال پوشش نمونه‌ای برای بازه اطمینان ۹۵٪ در جدول ۲ گزارش شده است. نتایج نشان می‌دهد که خطاهای استاندارد نمونه‌ای و مجانبی نزدیک به هم برآورد شده‌اند و احتمال پوشش در اغلب موارد نزدیک به ۹۵٪ است که اینها بیانگر عملکرد خوب روش است. به عبارت دیگر برآوردگرها رفتار مجانبی خوبی را نشان داده‌اند که این امر با نتایج قضایای حدی بیان شده مطابقت دارد. همچنین رفتار تابع $\boldsymbol{g}(\boldsymbol{t})$ از طریق رسم نمودارهایی بررسی شده است. این نمودارها شامل منحنی‌های $\widehat{\boldsymbol{g}}(\boldsymbol{t})$، اریبی، انحراف استاندارد و احتمال پوشش هستند. شکل‌های ۱ و ۲ به ترتیب نتایج مربوط به حجم نمونه ۵۰ و ۱۰۰ هستند. در هر دو تصویر نتایج برای حالت‌های $p_n < n$ (منحنی خط‌چین) و $p_n > n$ (منحنی نقطه‌چین) رسم شده‌اند. ملاحظه می‌شود که منحنی‌های برآورد شده تقریبا دارای تابع سینوسی هستند و این امر در حجم نمونه بیشتر یعنی شکل ۲ مشهودتر است. منحنی اریبی برآوردگرها حول خط صفر در نوسان هستند و برای حجم نمونه ۱۰۰ میزان اریبی کمتر از حجم نمونه ۵۰ است. همچنین برای حجم نمونه ۵۰ میزان اریبی حالت $p_n < n$ کمتر از $p_n > n$ است در حالی‌که در در حجم نمونه بالا میزان اریبی در دو حالت رفتار یکسانی را نشان می‌دهد. در هر دو شکل مقادیر انحراف استاندارد حالت $p_n < n$ کمتر از $p_n > n$ است . همچنین منحنی احتمال پوشش در هر دو شکل حول خط ۹۵٪ در نوسان است.

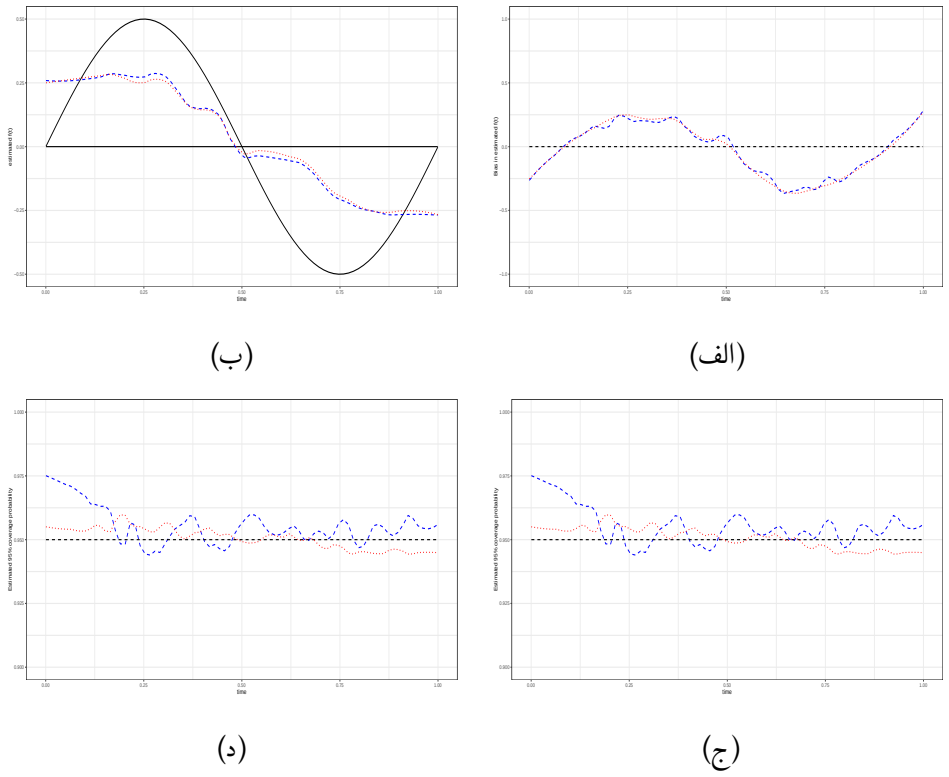
۶ تحلیل داده‌های HIV

دراین قسمت با استفاده از داده‌های مربوط به آزمایشات بیماری ایدز، رفتار برآوردگرهای پیشنهادی مطالعه می‌شود. این داده‌ها تعداد ۳۶۹ مرد مبتلا به HIV را، مورد بررسی قرار می‌دهد. در این مثال، متغیر پاسخ، تعداد سلول‌های $CD۴$ و متغیرهای تبیینی شامل استفاده یا عدم استفاده از مواد مخدر $(DRUG)$، تعداد شرکای جنسی $(SEXP)$، تعداد بسته‌های سیگار مورد استفاده در روز $(SMOKE)$، میزان افسردگی $(CESD)$، سن (AGE) و سال $(YEAR)$ است. به منظور اینکه توزیع داده‌ها، به توزیع نرمال نزدیک باشد، تبدیل ریشه دوم متغیر پاسخ درنظر گرفته شده است. زیگر و دیگل (۱۹۹۴) دریافتند که این داده‌ها دارای همبستگی درونی از نوع ساختار همبستگی تبادل پذیر با پارامتر همبستگی $\rho = ۰٫۵۰۹$ هستند. برای بهره‌مندی از انعطاف بالای مدل‌های خطی-جزئی متغیر $YEAR$ به عنوان جزء ناپارامتری و سایر متغیرها به صورت خطی وارد مدل شده‌اند. از آنجایی که اثرات متقابل نیز ممکن است در مدل اهمیت داشته باشند، تمامی اثرات متقابل متغیرها نیز در مدل وارد شده و از روش پیشنهادی برای



شکل ۱. به ازای $n = ۵۰$، منحنی‌های واقعی (خط پر)، برای $p = ۱۱$ (خط‌چین آبی) و برای $p = ۱۰۰$ (نقطه‌چین قرمز)، الف: برآورد $\boldsymbol{g}(\boldsymbol{t})$، ب: اریبی، ج: انحراف استاندارد و د: احتمال پوشش.

انتخاب متغیرهای مهم استفاده می‌شود. به منظور مقایسه عملکرد روش P-PLMM با دو روش دیگر یعنی P-LMM و PLMM از معیار انحراف استاندارد مجانبی (SD) و کمترین توان های دوم خطای تجربی (MSE) پارامترها استفاده گردید. در اینجا خطای استاندارد مجانبی، جذر واریانس مجانبی پارامترها یعنی جذر عناصر قطر اصلی ماتریس $\boldsymbol{I}^{-۱}_{\boldsymbol{\beta}_{n۰۱}}\boldsymbol{\beta}_{n۰۱}$ در قضیه ۲ بوده و میانگین توان های دوم خطای تجربی (MSE) از طریق میانگین گرفتن توان های دوم خطا از یک نمونه‌گیری بوت‌استرپ که به صورت $MSE(\widehat{\beta}_k) = \frac{۱}{S}\sum_{s=۱}^{S}(\widehat{\beta}_k - \widehat{\beta}_k^{(s)})^۲$ است، حاصل می‌شود که در آن $\widehat{\beta}_k$ مقدار برآورد شده پارامتر kام از کل نمونه و $\widehat{\beta}_k^{(s)}$ مقدار برآورد شده پارامتر از نمونه بوت‌استرپ sام است. همچنین برای شناسایی بهترین مدل پشتیبانی شده توسط داده‌ها، شاخص‌های مجموع توان های دوم رگرسیون (SSR)، مجموع توان های دوم خطا (SSE)، ضریب تعیین ($R^۲$) و میانگین توان های دوم خطای پیشگویی



شکل ۲. به ازای $n = ۱۰۰$، منحنی‌های واقعی (خط پر)، برای $p = ۱۴$ (خط‌چین آبی) و برای $p = ۵۰۰$ (نقطه‌چین قرمز)، الف: برآورد $\boldsymbol{g}(\boldsymbol{t})$، ب: اریبی، ج: انحراف استاندارد و د: احتمال پوشش.

$(MSPE)$ محاسبه شدند. نتایج در جدول ۳ نشان می‌دهد که برآوردگر P-PLMM در مقایسه با دو روش دیگر دارای مقادیر SD و MSE کمتری هستند. اما این تفاوت زیاد چشمگیر نیست، با این وجود در روش P-PLMM، شاخص‌های SSR و $R^۲$ دارای بیشترین مقدار بوده و شاخص‌های SSE و $MSPE$ کمترین مقدار را دارند که عملکرد بهتر این روش را تایید می‌کنند. همچنین مشاهده می‌شود که روش PLMM انتخاب متغیر انجام نمی‌دهد و در روش P-LMM به غیر از متغیر $CESD$ تمام متغیرها وارد مدل شده‌اند. تحت مدل P-PLMM، متغیرهای $SMOKE$، $SEXP$، $CESD$ و اثر متقابل $SMOKE * SEXP$ به عنوان متغیرهای مهم شناسایی شده‌اند.

جدول ۳. برآورد پارامترها به همراه SD و MSE تحت برازش سه مدل برای داده‌های بیماری ایدز.

P-PLMM $\widehat{\beta}$(SD،MSE)	P-LMM $\widehat{\beta}$(SD،MSE)	PLMM $\widehat{\beta}$(SD،MSE)	متغیرها
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	−۰٫۶۶۱ (۰٫۰۰۵, ۰٫۰۰۱)	۰٫۰۸۴ (۰٫۰۰۴, ۰٫۰۰۱)	AGE
۰٫۴۱۷(۰٫۰۱۴, ۰٫۰۰۱)	۲٫۴۹۸ (۰٫۰۳۰, ۰٫۰۰۱)	۰٫۷۳۰ (۰٫۰۲۴, ۰٫۱۶۳)	$SMOKE$
۰(۰, ۰٫۰۰۱)	۶٫۶۰۸ (۰٫۰۳۸, ۰٫۰۰۴)	۰٫۸۱۱ (۰٫۰۴۷, ۳٫۳۳۰)	$DRUG$
۰٫۰۸۴(۰٫۰۰۶, ۰٫۰۰)	۱٫۱۴۵ (۰٫۰۰۸, ۰٫۰۰۱)	۰٫۱۷۰ (۰٫۰۰۹, ۰٫۰۷۴)	$SEXP$
−۰٫۰۵۴(۰٫۰۰۳, ۰٫۰۰)	۰ (۰, ۰٫۰۰۴)	−۰٫۰۶۹ (۰٫۰۰۴, ۰٫۰۰۱)	$CESD$
۰(۰, ۰٫۰۰۱)	۰٫۰۱۶ (۰٫۰۰۱, ۰٫۰۰۲)	۰٫۰۱۲ (۰٫۰۰۱, ۰٫۰۰)	$AGE * SMOKE$
۰(۰, ۰٫۰۰)	۰٫۰۷۰ (۰٫۰۰۵, ۰٫۰۰)	−۰٫۰۲۰ (۰٫۰۰۴, ۰٫۰۰۱)	$AGE * DRUG$
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	۰٫۰۰ (۰٫۰۰, ۰٫۰۰)	−۰٫۰۱۲ (۰٫۰۰۱, ۰٫۰۰۱)	$AGE * SEXP$
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	۰٫۰۰ (۰٫۰۰, ۰٫۰۰)	−۰٫۰۰۵ (۰٫۰۰۲, ۰٫۰۰)	$AGE * CESD$
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	−۱٫۱۶۶ (۰٫۰۳۱, ۰٫۰۰۱)	−۰٫۲۷۸ (۰٫۰۲۲, ۰٫۲۹۴)	$SMOKE * DRUG$
۰٫۰۴۲(۰٫۰۰۳, ۰٫۰۰۱)	۰٫۰۰ (۰٫۰۰, ۰٫۰۰)	۰٫۰۳۸ (۰٫۰۰۳, ۰٫۰۰۲)	$SMOKE * SEXP$
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	۰٫۰۰ (۰٫۰۰, ۰٫۰۰)	−۰٫۰۰۵ (۰٫۰۰۱, ۰٫۰۰)	$SMOKE * CESD$
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	−۰٫۵۵۸ (۰٫۰۱۰, ۰٫۰۰۱)	−۰٫۰۷۹ (۰٫۰۰۹, ۰٫۰۷۴)	$DRUG * SEXP$
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	۰٫۰۰ (۰٫۰۰, ۰٫۰۰)	۰٫۰۲۰ (۰٫۰۰۳, ۰٫۰۰۱)	$DRUG * CESD$
۰٫۰۰(۰٫۰۰, ۰٫۰۰)	۰٫۰۰ (۰٫۰۰, ۰٫۰۰۱)	−۰٫۰۰۱ (۰٫۰۰۳, ۰٫۰۰)	$SEXP * CESD$
۶۷۵۷۱٫۸۹۰	۷۳۲۵۳٫۲۱۰	۶۶۵۶۱٫۶۸۰	SSR
۳۹۱۲۷٫۷۶۰	۵۴۳۶۸٫۸۸۰	۳۹۲۸۷٫۱۰۰	SSE
۰٫۶۳۳	۰٫۵۷۴	۰٫۶۲۸	R^2
۴۰٫۷۵۹	۷۷٫۰۲۴	۴۷٫۰۱۳	$MSPE$

بحث و نتیجه‌گیری

مسئله برآورد و انتخاب متغیر همزمان در مدل‌های نیمه پارامتری با اثرات آمیخته برای داده‌های طولی با بعد بالا در نظر گرفته شده است. مولفه ناپارامتری موجود در مدل با اسپلاین‌های رگرسیونی تقریب زده شده و از طریق بهینه‌سازی تابع هدف مبتنی بر تابع تاوان برآورد و انتخاب متغیر به طور همزمان انجام می‌شود. در ادامه، رفتار حدی برآوردگرهای حاصل در چارچوب داده‌های طولی با بعد بالا که در آن تعداد پارامترها متناسب با افزایش حجم نمونه افزایش می‌یابد، مورد مطالعه قرار گرفت. به منظور پیاده‌سازی روش برآورد پیشنهادی، یک الگوریتم تکراری مناسب برای انتخاب متغیرهای مهم و برآورد ضرایب غیر صفر ارائه گردیده است. در نهایت، عملکرد روش پیشنهادی با مطالعه شبیه‌سازی و تحلیل یک مجموعه داده واقعی مورد ارزیابی قرار گرفته است. نتایج نشان می‌دهند که مدل پیشنهادی در مقایسه با سایر مدل‌های رقیب، در برآورد ضرایب غیر صفر و انتخاب مدل عملکرد بهتری دارد. به عبارت دیگر وقتی داده‌ها دارای روند غیر خطی هستند، مدل های خطی-جزئی نسبت به مدل‌های خطی کارآمدتر هستند.

تقدیر و تشکر

نویسندگان از سردبیر، داوران و ویراستار محترم مجله که نظرات ارزشمند ایشان باعث بهبود مطالب ارائه شده در این مقاله گردید، کمال تشکر و قدردانی را دارند.

مراجع

کاظمی، م.، شاهسونی، د. و آرشی، م. (۱۳۹۷)، انتخاب متغیر و تشخیص ساختار در بعد بالا برای مدل‌های جمعی خطی-جزیی، *مجله علوم آماری*، ۱۲، ۴۸۵-۵۱۲.

Bondell, H. D., Krishna, A. and Ghosh, S. K. (2010), Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models, *Biometrics*, **66**, 1069-1077.

Cantoni, E., Mills, F. J. and Ronchetti, E. (2005), Variable Selection for Marginal Longitudinal Generalized Linear Models, *Biometrics*, **61**, 507-514.

Fan, J. Q., Huang, T. and Li, R. (2007), Analysis of Longitudinal Data with Semiparametric Estimation of Covariance Function, *Journal of the American Statistical Association*, **102**, 632-641.

Fan, J. Q. and Li, R. (2001), Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

Fan, J. Q. and Li, R. (2004), New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis, *Journal of the American Statistical Association*, **99**, 710-723.

Fu, W. J. (2003), Penalized Estimating Equations, *Biometrics*, **59**, 126-132.

He, X. M., Zhu, Z. Y. and Fung, W. K. (2002), Estimation in a Semiparametric Model for Longitudinal Data with Unspecified Dependence Structure, *Biometrika*, **89**, 579-590.

Hunter, D. and Li, R. (2005), Variable Selection Using MM Algorithm. *Annals of Statistics*, **33**, 1617-1642

Laird, N. M. and Ware, J. H. (1982), Random Effects Models for Longitudinal Data, *Biometrics*, **38**, 963-974.

Liang, K. Y. and Zeger, S. L. (1986), Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, **73**, 13-22.

Ma, S., Song, Q. and Wang, L. (2013), Simultaneous Variable Selection and Estimation in Semiparametric Mdeling of Longitudinal/Clustered Data, *Bernoulli*, **19**, 252–274.

Ni, X., Zhang, D. and Zhang, H. H. (2010), Variable Selection for Semiparametric Mixed Models in Longitudinal Studies, *Biometrics*, **66**, 79-88.

Qin, G. Y. and Zhu, Z. Y. (2007), Robust Estimation in Generalized Semiparametric Mixed Models for Longitudinal Data, *Journal of Multivariate Analysis*, **98**, 1658-1683.

Schumaker, L. L. (1981), *Spline Functions*, New York: Wiley.

Sinha, S. K. and Sattar, A. (2015), Inference in Semi-Parametric Spline Mixed Models for Longitudinal Data, *Metron*, **73**, 377-395.

Taavoni, M. and Arashi, M. (2019), Requlrization in Generalized Semiparametric Mixed-Effects Model for Longitudinal Data, Submitted.

Wang, L. and Qu, A. (2009), Consistent Model Selection and Data-Driven Smooth Tests for Longitudinal Data in the Estimating Equations Approach, *Journal of the Royal Statistical Society (Series B)*, **71**, 177-190.

Wang, W. L., Fan, T. H. (2011), Estimation in Multivariate t Linear Mixed Models for Multiple Longitudinal Data, *Statistica Sinica*, **21**, 1857-1880.

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis, *Biometrics*, **68**, 353-360.

Xu, P. R., Fu, W. and Zhu, L. X. (2013), Shrinkage Estimation Analysis of Correlated Binary Data with a Diverging Number of Parameters, *Science China Mathematics*, **56**, 359–377.

Xue, L., Qu, A. and Zhou, J. (2010), Consistent Model Selection for Marginal Generalized Additive Model for Correlated Data, *Journal of the American Statistical Association*, **105**, 1518-1530.

Zeger, S. L. and Diggle, P. J. (1994), Semi-Parametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters, *Biometrics*, **50**, 689-699.

Journal of Statistical Sciences, Autumn and Winter, 2020
Vol. 14, No. 2, pp 367-388
DOI: 10.29252/jss.13.2.367

Variable Selection in Semiparametric Mixed Effect Model for High–Dimension Longitudinal Data

Taavoni, M. and Arashi, M.
Department of Statistics, Shahrood University of Technology, Shahrood, Iran.

Abstract: This paper considers the problem of simultaneous variable selection and estimation in a semiparametric mixed-effects model for longitudinal data with normal errors. We approximate the nonparametric function by regression spline and simultaneously estimate and select the variables under the optimization of the penalized objective function. Under some regularity conditions, the asymptotic behaviour of the resulting estimators is established in a high-dimensional framework where the number of parametric covariates increases as the sample size increases. For practical implementation, we use an EM algorithm to selects the significant variables and estimates the nonzero coefficient functions. Simulation studies are carried out to assess the performance of our proposed method, and a real data set is analyzed to illustrate the proposed procedure.

Keywords: Longitudinal data, Penalized estimator, Smoothing spline, Semiparametric mixed model, Variable selection, HIV.

Mathematics Subject Classification (2010): 62H20, 62F12.