مجله مدل‌سازی پیشرفته ریاضی، دوره ۴، شماره ۱، بهار و تابستان ۱۳۹۳

انتخاب متغیر در مدل‌های آمیخته متناهی تعمیم‌یافته نیم پارامتری

فرزاد اسکندری[1*]، احسان ارمز**، رحمان فرنوش***

* گروه آمار، دانشگاه علامه طباطبایی

** گروه آمار، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران

*** گروه ریاضی کاربردی، دانشگاه علم و صنعت ایران

تاریخ دریافت: ۱۳۹۲/۱۱/۲۲ تاریخ پذیرش ۱۳۹۳/۹/۳۰

چکیده: هدف این مقاله به دست آوردن بهترین متغیرهای کمکی تأثیرگذار بر متغیر پاسخ در مدل‌های نیم پارامتری تحت شرایطی است که تابع تاوانیده نیز در مسئله وجود دارد. لازم به ذکر است که ترکیبی از پارامترها به‌عنوان ضرایب متغیرهای موجود در هر مدل ارائه‌شده که برخی از آن‌ها به‌صورت خطی و بعضی دیگر به‌صورت تابعک بر متغیر پاسخ تأثیرگذارند. لاجرم روش نیم پارامتری به‌عنوان یک‌راه حل بهینه در حل مسئله مدنظر قرارگرفته است. لذا در این مقاله مسئله انتخاب متغیر را در مدل‌های تعمیم‌یافته نیم پارامتری آمیخته متناهی موردبررسی قرار خواهیم داد. این مسئله خود شامل انتخاب مدل در مؤلفه ناپارامتری مدل و انتخاب متغیر در بخش پارامتری می‌شود. بنابراین با انتخاب مدل‌های مجزایی برای هر مؤلفه ناپارامتری هر زیر مدل مواجه خواهیم بود. به‌منظور غلبه بر این بار محاسباتی بالا، کلاسی از رویه‌های انتخاب متغیر را برای مدل‌های تعمیم‌یافته نیم پارامتری آمیخته متناهی معرفی خواهیم کرد. نشان خواهیم داد رویکرد جدید برای انتخاب متغیر، سازگار است. همچنین به کمک شبیه‌سازی نشان خواهیم داد روش پیشنهادی از کارایی مناسبی برخوردار بوده و روش‌های موجود را بهبود می‌بخشد و علاوه بر این به توان محاسباتی کمتری نیاز دارد.

واژه‌های کلیدی: انتخاب متغیر، درستنمایی تاوانیده، مدل نیم پارامتری، مدل آمیخته متناهی.

رده‌بندی ریاضی (۲۰۱۰): ۶۲J۱۲

1- آدرس الکترونیکی نویسنده مسئول مقاله: aeskandari@atu.ac.ir

۱-مقدمه

مدل‌سازی یکی از مهم‌ترین فن‌های آمار است و انتخاب متغیر نقشی اساسی در آن ایفا می‌کند. در مسائل کاربردی به‌منظور کاستن از شدت اریبی‌های ممکن مدل‌سازی، معمولاً تعداد زیادی از متغیرهای کمکی در مراحل اولیه مدل‌سازی معرفی می‌شوند. از سوی دیگر، به‌منظور افزایش قابلیت پیش‌بینی و انتخاب متغیرهای معنی‌دار، آماردانان اغلب از حذف گام‌به‌گام و انتخاب بهترین زیرمجموعه استفاده می‌کنند.

اگرچه کاربرد این روش‌ها در عمل مفید واقع می‌شوند، اما خطاهای تصادفی ذاتی را در مرحله انتخاب متغیرها نادیده می‌گیرند. بنابراین خواص نظری آن‌ها تقریباً غیرقابل درک است. علاوه بر این، انتخاب بهترین زیرمجموعه متغیرها چندین نقطه‌ضعف دارند، که مهم‌ترین آن‌ها فقدان پایداری است (به‌عنوان مثال به [۱] مراجعه کنید). همچنین می‌توان به زمان‌بر بودن و مشکلات محاسباتی این روش‌ها نیز اشاره کرد.

در تلاشی به‌منظور انتخاب خودکار و همزمان متغیرها، فن و لی [۲] روش یکپارچه‌ای را بر اساس کمترین توان‌های دوم تاوانیده پیشنهاد داده‌اند. «روش‌های کمترین توان‌های دوم تاوانیده که می‌توانند به‌صورت خودکار و همزمان متغیرهای معنی‌دار را انتخاب کنند، در سال‌های اخیر از اقبال گسترده‌ای برخوردار شده‌اند. [۳]»

توابع تاوان متنوعی درزمینهٔ انتخاب متغیر معرفی شده‌اند. تاوان L_1 در LASSO[1] توسط تیبشیرانی [۴] معرفی شد و تبدیل به یکی از محبوب‌ترین تاوان‌ها شد. فن و لی [۲] روش انحراف مطلق مختصر شده هموار[2] (SCAD) را معرفی کردند که دارای خاصیت اراکل[3] بود. خصوصیات SCAD توسط [۲ و ۵-۸] مطالعه گردیده و الگوریتم‌های جدیدی توسط آنان تولیدشده است. زو [۹] LASSO سازوار[4] (ALASSO) را با استفاده از وزن‌های سازوار برای ضرایب مختلف تاوان در LASSO معرفی و خصوصیت اراکل آن را بررسی کرده است. منابع [۱۰و ۱۱] انتخاب متغیر را در حالتی مطالعه کرده‌اند که بعد داده‌ها بزرگ‌تر از اندازه نمونه است. همچنین گستره‌ای از توابع تاوان در مدل‌های رگرسیون آمیخته متناهی توسط [۱۲] بررسی‌شده است.

از سوی دیگر، مدل‌های آمیخته متناهی ابزار انعطاف‌پذیری برای مدل‌بندی داده‌های حاصل از جوامع ناهمگون فراهم ساخته‌اند. این مدل‌ها، در زمینه‌های بسیاری از جمله زیست‌شناسی،

1- Least Absolute Shrinkage and Selection Operator
2- Smoothly Clipped Absolute Deviation
3- Oracle
4- Adaptive Least Absolute Shrinkage and Selection Operator

ژنتیک، مهندسی و بازاریابی مورداستفاده قرارگرفته‌اند. کتاب [۱۳] مرور جامعی از مدل‌های آمیخته متناهی را در برگرفته است. زمانی که یک متغیر تصادفی با یک توزیع آمیخته متناهی به متغیرهای کمکی خاصی وابسته است، یک مدل رگرسیون آمیخته متناهی[۱] (FMR) به‌دست می‌آید. بنابراین مدل‌های رگرسیون آمیخته متناهی زمانی مناسب هستند که داده‌های رگرسیون به دو یا چند گروه مشاهده نشده تعلق داشته و یا ناهمگن باشند. این وضعیت زمانی رخ خواهد داد که اعتقاد داریم مدل‌های رگرسیون در هر گروه متمایزند.

مسئله انتخاب متغیر در مدل‌های رگرسیون آمیخته متناهی در سال‌های اخیر توجه ویژه‌ای را به خود اختصاص داده است. روش‌های انتخاب بهترین زیرمجموعه، مانند معیار اطلاعات آکاییک یا AIC [۱۴] و معیار اطلاعات بیزی یا BIC [۱۵] و حالت‌های اصلاح‌شده آن‌ها درزمینهٔ مدل‌های رگرسیون آمیخته متناهی موردمطالعه قرارگرفته‌اند. به‌هرحال حتی برای مدل‌های رگرسیون آمیخته متناهی با تعداد متوسط مؤلفه‌ها و متغیرهای کمکی، روش‌های انتخاب بهترین زیرمجموعه ازنظر محاسباتی مشکل‌اند.

روش‌های جدید انتخاب متغیر، مانند LASSO [۴ و ۱۶] و SCAD [۲ و ۵] به‌طور خاص برای انتخاب متغیر در ساختار مدل‌های رگرسیون آمیخته متناهی مفید هستند. از میان سایر روش‌ها می‌توان به گروت[۲] غیر منفی در یوان و لین [۱۷]، الگوریتم تفاضل محدب[۳] وو و لیو [۸]، رویه غربال مستقل[۴] ناپارامتری فن و همکاران [۱۸] و روش تیلتینگ[۵] چو و فریزلویسز [۱۹] اشاره کرد.

LASSO و SCAD از این نظر متفاوت از روش‌های انتخاب متغیر سنتی هستند که متغیرهای کمکی غیر معنی‌دار را با برآورد اثرات آن‌ها به‌عنوان صفر از مدل حذف می‌کنند و برخلاف روش‌های انتخاب بهترین زیرمجموعه، می‌توانند در مسائلی که بعد آن‌ها به‌صورت منطقی بالاست، به کار گرفته شوند. خلیلی و چن [۱۲]، یک‌رویه جدید انتخاب متغیر در مدل‌های رگرسیون آمیخته متناهی را بر اساس این روش‌ها طراحی کرده‌اند.

به‌هرحال گاهی اوقات نمی‌توان یک ارتباط خطی را برای تمام متغیرها فرض کرد. یک‌راه مقابله با چنین وضعیتی استفاده از مدل‌های نیم پارامتری است که مزایای مدل‌سازی پارامتری و ناپارامتری را حفظ می‌کنند و در سال‌های اخیر موردتوجه ویژه‌ای قرارگرفته‌اند.

1- Finite Mixture of Regression
2- Garrote
3- Difference Convex Algorithm
4- Independence Screening
5- Titling

خلیلی [۲۰] نیز به‌مرور جامع مسئله انتخاب متغیر در مدل‌های رگرسیون آمیخته متناهی پرداخته است.

دو و همکاران [۲۱] آمیخته متناهی مدل‌های خطی با اثرات آمیخته[1] را بررسی کرده‌اند که برای مدل‌سازی رگرسیون طولی در صورت وجود ناهمگنی در اثرات تصادفی و ثابت مفید است. با توجه به اینکه با افزایش تعداد متغیرها روش‌های انتخاب متغیر معمول پاسخگو نخواهد بود، آنان رویکرد درستنمایی تاوانیده را مدنظر قرار داده و جهت کارایی محاسبات از الگوریتم EM آشیانی[2] استفاده کرده‌اند.

ارمز و اسکندری [۲۲] مسئله انتخاب متغیر را در مدل‌های نیم پارامتری آمیخته متناهی مطالعه کرده‌اند. توجه داشته باشید چارچوب در نظر گرفته‌شده توسط آنان متفاوت از اینجاست. موارد اختلاف بین [۲۲] و مقاله حاضر را می‌توان به شرح زیر بیان نمود.

۱- مفهوم شناسایی پذیری لزوماً در تمام خانواده‌های ارائه‌شده در [۲۲] وجود ندارد، اما در این مقاله با توجه به انتخاب خانواده نمایی طبق قضایایی که در ادامه آورده خواهد شد، تضمین می‌شود.

۲- تابع ربط $h(.)$ مورداستفاده در این مقاله یک تابع معرفی‌شده بر اساس تعریف نلدر و وادربرن [۲۳] است. درصورتی‌که در [۲۲] لزوماً چنین نیست. این مطلب اثر خود را در نتایج قضایای حدی اثبات‌شده نشان می‌دهد. زیرا در این مقاله می‌توانیم برآورد پارامتر $\theta_k(u,\mathbf{x},\mathbf{z}) = h^{-1}\{\mathbf{x}^t\alpha_k(u) + \mathbf{z}^t\boldsymbol{\beta}_k\}$ را به دست آوریم درصورتی‌که این موضوع لزوماً در [۲۲] امکان‌پذیر نیست.

بنابراین هدف این مقاله مطالعه مسئله انتخاب متغیر در مدل‌های تعمیم‌یافته نیم پارامتری آمیخته متناهی به‌جای مدل‌های نیم پارامتری آمیخته متناهی است. لذا در این مقاله علاوه بر معرفی مدلی جدید و در نظر گرفتن مدل‌های تعمیم‌یافته نیم پارامتری به‌جای مدل‌های نیم پارامتری، قادر خواهیم بود مسئله شناسایی پذیری مدل‌های آمیخته را به‌صورت دقیق‌تری بررسی کنیم. بدین منظور ادامه مقاله به شکل زیر تنظیم‌شده است.

در بخش ۲ مدل‌های تعمیم‌یافته نیم پارامتری آمیخته متناهی را معرفی می‌کنیم. شناسایی پذیری مدل در بخش ۳ بررسی می‌شود. در بخش ۴ رویکرد انتخاب متغیر را ارائه خواهیم کرد. بدین منظور در بخش ۴-۱ درستنمایی تاوانیده و در بخش ۴-۲ الگوریتم EM و حل عددی برآوردگرها را موردبررسی قرار خواهیم داد. بخش ۴-۳ به انتخاب پارامترهای منظم سازی

1- Finite Mixtures of Linear Mixed-Effects Models
2- Nested

اختصاص خواهد یافت و خواص مجانبی روش پیشنهادی در بخش ۵ مطالعه می‌شوند و درنهایت، در بخش ۶ خصوصیات روش پیشنهادی را به کمک شبیه‌سازی مطالعه خواهیم کرد.

۲- مدل‌های آمیخته متناهی تعمیم‌یافته نیم پارامتری

فرض کنید Y متغیر پاسخ موردنظر و $(u,\mathbf{x},\mathbf{z})$ بردار متغیرهای کمکی دارای اثر احتمالی بر Y باشد که $\mathbf{x}=(x_1,x_2,\ldots,x_d)^t$، $\mathbf{z}=(z_1,z_2,\ldots,z_p)^t$ و u یک متغیر تکی است. مدل آمیخته متناهی تعمیم‌یافته نیم پارامتری به شکل زیر تعریف می‌شود.

تعریف ۱: فرض کنید $G=\{f(y;\theta,\phi);(\theta,\phi)\in\Theta\times(0,\infty)\}$ خانواده‌ای از چگالی‌های احتمال پارامتری Y نسبت به‌اندازه σـمتناهی ν باشد که $\Theta\subset R$ و ϕ پارامتر پراکندگی است. گوییم $(\mathbf{x},\mathbf{z},u,Y)$ از مدل آمیخته متناهی تعمیم‌یافته نیم پارامتری از مرتبه K تبعیت می‌کند هرگاه تابع چگالی Y به‌شرط $(\mathbf{x},\mathbf{z},u)$ دارای فرم

$$f(y;\mathbf{x},\boldsymbol{\Psi})=\sum_{k=1}^{K}\pi_k f_k(y;\theta_k(u,\mathbf{x},\mathbf{z}),\phi_k) \qquad (۱)$$

باشد که میانگین هر مؤلفه به ازای یک تابع ربط مشخص $h(.)$ به شکل

$$\theta_k(u,\mathbf{x},\mathbf{z})=h^{-1}\{\mathbf{x}^t\alpha_k(u)+\mathbf{z}^t\boldsymbol{\beta}_k\} \qquad k=1,2,\ldots,K$$

است. $\boldsymbol{\alpha}(.)$ بردار توابع ضرایب رگرسیون هموار مجهول است و $\boldsymbol{\beta}_k=(\beta_{k1},\ldots,\beta_{kP})^t$، $\boldsymbol{\Psi}=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K,\phi,\boldsymbol{\pi})$، $\boldsymbol{\alpha}_k=(\alpha_{1k},\ldots,\alpha_{dk})^t$، $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_K)^t$ و $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_{K-1})^t$.

لازم به ذکر است که در مدل (۱) ترکیبی از پارامترها به‌عنوان ضرایب متغیرهای موجود در هر مدل ارائه‌شده که برخی از آن‌ها به‌صورت خطی و بعضی دیگر به‌صورت تابعک بر متغیر پاسخ تأثیرگذارند و لذا این مدل، مدلی نیم پارامتری است.

همچنین فرض کنید $\pi_k>0$ و $\sum_{k=1}^{K}\pi_k=1$. f_k تابع چگالی مشخص مؤلفه است و در اینجا عضو خانواده نمایی در نظر گرفته می‌شود. همچنین در اینجا فرض می‌کنیم که چگالی‌های مؤلفه‌ها به ازای هر مؤلفه از خانواده پارامتری یکسانی پیروی می‌کنند، یعنی $f_k\equiv f$. در این مقاله چگالی توزیعی از خانواده نمایی را به فرم

$$f(y;\theta,\phi)=d(y,\phi)exp\left(\frac{T(y)\theta-A(\theta)}{c(\phi)}\right)$$

در نظر می‌گیریم که در آن $A(.)$، $c(.)$ و $d(.)$ توابع مشخصی هستند.

اگرچه می‌توانیم احتمال‌های آمیختگی را وابسته به متغیرهای کمکی در نظر گرفته و آن‌ها را برآورد کنیم، بااین‌وجود به جهت سادگی ارائه مطلب ازاین‌پس فرض می‌کنیم که احتمال‌ها به متغیرهای کمکی وابسته نیستند. لذا حالت نخست به‌عنوان یک مسئله باز باقی خواهد ماند که در حال حاضر مشغول مطالعه آن هستیم.

این مدل مشخصه‌های مدل‌های نیم پارامتری و مدل‌های آمیخته متناهی را ترکیب می‌کند. بنابراین مشابه هر مدل آمیخته متناهی یا مدل‌های رگرسیون، رابطه بین متغیر پاسخ و مجموعه‌ای از متغیرهای کمکی را مطالعه می‌کند. از سوی دیگر، توزیع شرطی متغیر پاسخ به‌شرط متغیرهای کمکی، یک آمیخته متناهی است.

۳- شناسایی پذیری

همان‌گونه که می‌دانیم یکی از مسائل مدل‌های آمیخته متناهی، شناسایی پذیری آن‌هاست. یک مدل را شناسایی پذیر گویم هرگاه هیچ دو مجموعه‌ای از مقادیر پارامترها توزیع مشابهی را به دست ندهند. برای شناسایی پذیری لازم است که نگاشت فضای پارامتر به فضای مدل، یک‌به‌یک باشد. یعنی برای هر مدل در فضای مدل، یک بردار پارامتر یکتا در فضای پارامتر وجود داشته باشد که به مدل نگاشته شده باشد.

تعریف۲: یک مدل آمیخته متناهی تعمیم‌یافته نیم پارامتری با تابع چگالی شرطی (۱) را در نظر بگیرید. به ازای یک ماتریس طرح معلوم، مدل آمیخته متناهی تعمیم‌یافته نیم پارامتری را شناسایی پذیر گوییم هرگاه به ازای هر دو پارامتر $\boldsymbol{\Psi}$ و $\boldsymbol{\Psi}^*$ رابطه

$$\sum_{k=1}^{K}\pi_k f(y;\theta_k(u_i,\mathbf{x}_i,\mathbf{z}_i),\phi_k)=\sum_{k=1}^{K^*}\pi_k^* f(y;\theta_k^*(u_i,\mathbf{x}_i,\mathbf{z}_i),\phi_k^*)$$

به ازای $i=1,\ldots,n$ و تمامی مقادیر ممکن y به دست دهد $K=K^*$ و $\boldsymbol{\Psi}=\boldsymbol{\Psi}^*$. شناسایی پذیری بدان معناست که هیچ دو مجموعه‌ای از مقادیر متفاوت پارامترها، تابع چگالی یکسانی ندارند. به‌هرحال در تعریف فوق $\boldsymbol{\Psi}=\boldsymbol{\Psi}^*$ را تا یک جایگشت تفسیر می‌کنیم ([۲۴ و ۲۶] را مشاهده کنید).

در ادامه فرض کنید Ω فضای پارامترهای پذیرفتنی برای آمیخته‌های K مؤلفه‌ای باشد که در شرایط زیر صدق می‌کنند.

$$\pi_k>0 \qquad \forall k=1,\ldots,K \qquad (۲)$$

$$\forall k,l \in \{1,\ldots,K\} : k \neq l \Rightarrow \theta_k \neq \theta_l \qquad (۳)$$

دو شرط (۲) و (۳) از بیش برازندگی و مشکلات شناسایی پذیری که به‌واسطه مؤلفه‌های تهی که در آن‌ها θ_k نمی‌تواند به‌صورت یکتا تعیین شود و همچنین به‌واسطه مؤلفه‌های دارای بردار پارامترهای برابر که مقادیر متفاوت π_k برای آن‌ها امکان‌پذیر است، جلوگیری می‌کنند.

مسائل شناسایی پذیری عام[1] برای آمیخته‌های متناهی توزیع‌ها توسط [۲۶] مطالعه شده‌اند. شناسایی پذیری عام برای توزیع‌های پیوسته مهم مانند گاما و نرمال و همچنین توزیع پواسون اثبات‌شده است. یک حالت خاص آمیخته متناهی توزیع‌های دوجمله‌ای است که تنها در حالتی شناسایی پذیر است که تعداد مؤلفه‌ها محدود باشد. برای کلاس مدل آمیخته متناهی توزیع‌های دوجمله‌ای $Bi(T,\pi)$ با احتمال موفقیت π و پارامتر تکرار T یک شرط لازم و کافی شناسایی پذیری عبارت است از $T \geq 2K-1$.

گرون و لیسیچ [۲۷] قضیه‌ای پیرامون شرایط کافی شناسایی پذیری مدل‌های آمیخته متناهی خطی تعمیم‌یافته ارائه کرده‌اند که اگرچه در قالب مدل‌های پارامتری ارائه گردیده است اما با اندکی تغییرات جزئی می‌توان نشان داد که در چارچوب کار ما نیز قابل‌استفاده است. لذا از بیان مجدد قضیه خودداری کرده و خوانندگان علاقه‌مند را به مرجع [۲۷] ارجاع می‌دهیم.

توجه داشته باشید که برای شرط (ا-الف) گرون و لیسچ [۲۷] شناسایی پذیری عام آمیخته‌های متناهی با توزیع مؤلفه معلوم، الزامی است. چنانچه توزیع مؤلفه‌ها، هر یک از توزیع‌های نرمال، گاما یا پواسون باشد این شرط محدودیتی ایجاد نخواهد کرد، زیرا آمیخته این توزیع‌ها شناسایی پذیر عام‌اند.

در مورد توزیع دوجمله‌ای بایستی پارامتر تکرار را به ازای هر مشاهده موردبررسی قرار دهیم تا تعیین شود که آیا می‌توان آن را در $\tilde{I}$ قرار داد یا خیر. شرط (۱-ب) گرون و لیسچ [۲۷] نشان می‌دهد که به ازای هر آزمودنی t بایستی یکی از q ابر صفحه گذرنده از مبدأ H_j وجود داشته باشد که تمام مشاهدات شناسایی پذیر این آزمودنی را بپوشاند.

۴- روش انتخاب متغیر

با توجه به اینکه مایلیم استنباط آماری صرفاً بر اساس تابع چگالی شرطی مشخص‌شده در تعریف (۱) صورت پذیرد، در حالتی که $(u,\mathbf{x},\mathbf{z})$ یک بردار تصادفی است، فرض می‌کنیم که چگالی آن به‌صورت تابعی مستقل از پارامترهای مدل است. فرض کنید $(u_1,\mathbf{x}_1,\mathbf{z}_1,y_1),\ldots,(u_n,\mathbf{x}_n,\mathbf{z}_n,y_n)$ نمونه‌ای از مشاهدات از مدل (۱) باشند. با در نظر

1- Generic Identifiability

گرفتن فرم متعارف خانواده نمایی یک پارامتری، تابع لگاریتم درستنمایی شرطی $\boldsymbol{\Psi}$ عبارت خواهد بود از

$$l_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} log\left\{\sum_{k=1}^{K} \pi_k exp\left[\frac{\theta_k(u_i, \mathbf{x}_i, \mathbf{z}_i)T(y_i) - A(\theta_k(u_i, \mathbf{x}_i, \mathbf{z}_i))}{c(\phi_k)}\right] d(y_i, \phi_k)\right\}$$

۴-۱- درستنمایی تاوانیده

زمانی که اثر یک مؤلفه معنی‌دار نیست، برآورد ماکسیمم درستنمایی یا کمترین توان‌های دوم متناظر آن اغلب نزدیک صفر است، اما برابر آن نیست. لذا این متغیر کمکی از مدل حذف نمی‌شود. همان‌گونه که پیش‌تر اشاره شد، یک راه‌حل مؤثر برای مقابله با مشکلات محاسباتی این وضعیت استفاده از برآوردهای ماکسیمم درستنمایی تاوانیده است. تابع ماکسیمم درستنمایی تاوانیده به شکل

$$\tilde{l}_n(\boldsymbol{\Psi}) = l_n(\boldsymbol{\Psi}) - p_n(\boldsymbol{\Psi}) \qquad (۵)$$

تعریف می‌شود، که در آن

$$p_n(\boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \left\{\sum_{j=1}^{P} p_{nk}(\beta_{kj})\right\} \qquad (۶)$$

و $p_{nk}(\beta_{kj})$ توابعی غیر منفی و غیر نزولی برحسب $|\beta_{kj}|$ هستند. با ماکسیمم کردن $\tilde{l}_n(\boldsymbol{\Psi})$ شانس این وجود دارد که برخی از مقادیر برآورد شده β برابر صفر شوند و لذا می‌توانیم به‌صورت خودکار یک زیر مدل را انتخاب کنیم. بنابراین همان‌گونه که در بخش یک اشاره شد، در این روش انتخاب متغیر و برآورد پارامترها همزمان صورت پذیرفته و مشکلات محاسباتی به نحو چشمگیری کاهش می‌یابند. مطابق معمول تابع تاوان را از طریق متناسب بودن تاوان ضریب رگرسیون بخش پارامتری مؤلفه kام با π_k در (۶) در ارتباط با حجم نمونه در نظر می‌گیریم.

به‌طورکلی بایستی تابع تاوان مناسبی را که متناسب با نیازهای کاربردی مسئله و مطالب نظری باشد، انتخاب کنیم. بااین‌وجود سه تابع تاوان زیر به‌دفعات در ادبیات انتخاب متغیر مورداستفاده قرارگرفته‌اند و در اینجا نیز آن‌ها را مطالعه می‌کنیم.

۱- تابع تاوان نرم L_1: $p_{nk}(\beta) = \gamma_{nk}\sqrt{n}\,|\beta|$

۲- تاوان HARD: $p_{nk}(\beta) = \gamma_{nk}^2 - \left(\sqrt{n}\,|\beta| - \gamma_{nk}\right)^2 I\left(\sqrt{n}\,|\beta| < \gamma_{nk}\right)$

۳- تابع تاوان SCAD: فرض کنید $(.)_+$ بخش مثبت یک کمیت باشد.

$$p'_{nk}(\beta)=\gamma_{nk}\sqrt{n}I\left(\sqrt{n}\,|\beta|\le\gamma_{nk}\right)+\frac{\sqrt{n}\left(a\gamma_{nk}-\sqrt{n}\,|\beta|\right)_+}{a-1}I\left(\sqrt{n}\,|\beta|>\gamma_{nk}\right)$$

در این توابع تاوان $\gamma_{nk}>0$ و $a>2$ بر اساس درجه‌ای که روش موردنظر تلاش می‌کند تا متغیرهای کمکی را از مدل حذف کند، انتخاب می‌شوند. در مسائل کاربردی این ثابت‌ها می‌توانند به‌صورت ذهنی توسط تحلیل‌گر داده‌ها و یا با استفاده از روش‌های مبتنی برداده‌ها انتخاب شوند. انتخاب پارامترهای منظم سازی[1] اغلب توسط معیارهایی مانند اعتبارسنجی متقابل[2] یا اعتبارسنجی متقابل تعمیم‌یافته[3] صورت می‌پذیرد.

با پیروی از [۱۲ و ۲۸]، توابع تاوان $p_n(.)$ در رابطه (۶) را که از LASSO، HARD و SCAD تولیدشده‌اند، به ترتیب تاوان‌های MIXLASSO، MIXHARD و IXSCAD می‌نامیم. ماکسیمم کردن درستنمایی تاوانیده، معادل ماکسیمم کردن مقید است. تابع تاوان LASSO محدب بوده و از این نظر برای محاسبات عددی مناسب‌تر است. این تابع تاوان تمامی اثرات را به‌اندازه مشابهی کاهش می‌دهد تا زمانی که اثرات برآورد شده به صفر کاهش می‌یابند. زمانی که تاوان صعودی است، تابع تاوان SCAD اثرات کوچک‌تر را سریع‌تر از اثرات بزرگ‌تر کاهش می‌دهد. همچنین انتظار می‌رود که HARD مشابه SCAD عمل کند، اما با هموارسازی کمتر [۱۲]. بنابراین می‌توان برآورد درستنمایی تاوانیده $\boldsymbol{\beta}$ را به دست آورد. با انتخاب تابع تاوان خاص، برآورد حاصل $\boldsymbol{\beta}$ شامل برخی ضرایب صفر دقیق خواهد بود. این مطلب متناظر با حذف متغیرهای متناظر از مدل نهایی است و درنتیجه به هدف انتخاب متغیر دست‌یافته‌ایم.

به‌هرحال رابطه (۵) برای بهینه‌سازی مناسب نیست، زیرا $\boldsymbol{\alpha}(.)$ از توابع ناپارامتری تشکیل‌شده است. مشابه [۲۸] ابتدا $\alpha_j(v)$ را به ازای v در یک همسایگی u توسط

$$\alpha_{kj}(v)\approx\alpha_{kj}(u)+\alpha'_{kj}(u)(v-u)\equiv a_{kj}+b_{kj}(v-u). \qquad (۷)$$

تقریب می‌زنیم. حال تابع لگاریتم درستنمایی

$$\sum_{i=1}^{n}log\left\{\sum_{k=1}^{K}\pi_k exp\left[\frac{\tilde{\theta}_k(u_i,\mathbf{x}_i,\mathbf{z}_i)T(y_i)-A(\tilde{\theta}_k(u_i,\mathbf{x}_i,\mathbf{z}_i))}{c(\phi_k)}\right]d(y_i,\phi_k)\mathrm{K}_h(u_i-u)\right\} \qquad (۸)$$

1- Tuning Parameter
2- Cross Validation
3- Generalized Cross Validation

را ماکسیمم خواهیم کرد که $K(.)$ یک تابع ربط و $K_h(t)=h^{-1}K(t/h)$ یک بازمقیاس تابع ربط توسط پهنای باند h است. در اینجا لازم است توجه خوانندگان را به تفاوت نمادهای h و $h(.)$ جلب کنیم. در این مقاله نماد نخست اسکالر بوده و نمایانگر پهنای باند است درحالی‌که $h(.)$ تابع بوده و به‌عنوان تابع ربط (تعریف ۱) استفاده‌شده است. همچنین

$$\tilde{\theta}_k(u_i,\mathbf{x}_i,\mathbf{z}_i)=\mathbf{x}_i^t\mathbf{a}_k+\mathbf{x}_i^t\mathbf{b}_k(u_i-u)+\mathbf{z}_i^t\boldsymbol{\beta}_k\,.$$

فرض کنید $\{\tilde{\mathbf{a}},\tilde{\mathbf{b}},\boldsymbol{\beta}\}$ پاسخ ماکسیمم کردن رابطه (۸) باشد. تعریف می‌کنیم

$$\tilde{\boldsymbol{\alpha}}(u)=\tilde{\mathbf{a}}\,.$$

حال $\boldsymbol{\beta}$ را توسط لگاریتم درستنمایی تاوانیده یعنی

$$l(\boldsymbol{\beta})=\sum_{i=1}^{n}log\left\{\sum_{k=1}^{K}\pi_k exp\left[\frac{\theta_k^*(u_i,\mathbf{x}_i,\mathbf{z}_i)T(y_i)-A(\theta_k^*(u_i,\mathbf{x}_i,\mathbf{z}_i))}{c(\phi_k)}\right]d(y_i,\phi_k)\right\}-p_n(\boldsymbol{\Psi}) \quad (۹)$$

برآورد می‌کنیم که در آن $\theta_k^*(u_i,\mathbf{x}_i,\mathbf{z}_i)$ از جایگزین کردن $\tilde{\boldsymbol{\alpha}}(u)$ در $\tilde{\theta}_k(u_i,\mathbf{x}_i,\mathbf{z}_i)$ توسط $\tilde{\mathbf{a}}$ به دست می‌آید. ماکسیمم کردن $l(\boldsymbol{\beta})$ به برآوردگر ماکسیم درستنمایی تاوانیده $\boldsymbol{\beta}$ منجر خواهد شد.

۴-۲- الگوریتم EM

همان‌گونه که می‌دانیم الگوریتم EM رویکردی ساده را برای بهینه‌سازی مدل‌های آمیخته متناهی فراهم می‌آورد. به دلیل شرط لازم تنک بودن (P_0)(به‌ضمیمه یک مراجعه نمایید.) $p_{nk}(\beta)$ ها در $\boldsymbol{\beta}=0$ مشتق‌پذیر نیستند. بنابراین با پیروی از [۲] جهت استفاده از الگوریتم نیوتن- رافسون در مرحله M الگوریتم EM، $p_{nk}(\beta)$ را با تقریب موضعی مرتبه دوم آن در یک همسایگی β_0 یعنی

$$p_{nk}(\beta)\approx p_{nk}(\beta_0)+\frac{p'_{nk}(\beta)}{2\beta_0}\left(\beta^2-\beta_0^2\right)^2$$

جایگزین می‌کنیم. فرض کنید $\boldsymbol{\Psi}^{(m)}$ مقدار پارامتر پس از تکرار mام باشد، $p_n(\boldsymbol{\Psi})$ در رابطه (۹) را با تابع

$$\tilde{p}_n(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(m)})=\sum_{k=1}^{K}\left\{\pi_k\sum_{j=1}^{P}p_{nk}(\beta_{jk}^{(m)})+\frac{p'_{nk}(\beta_{jk}^{(m)})}{2\beta_{jk}^{(m)}}(\beta_{jk}^2-\beta_{jk}^{(m)^2})\right\}$$

جایگزین می‌کنیم. الگوریتم EM اصلاح‌شده به شکل زیر خواهد بود. فرض کنید تابع لگاریتم درستنمایی کامل به شکل

$$l_n^c(\boldsymbol{\Psi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} v_{ik}\left[log\pi_k + log\{d(y_i,\phi_k)\} + \frac{\theta_k^*(u_i,\mathbf{x}_i,\mathbf{z}_i)T(y_i) - A(\theta_k^*(u_i,\mathbf{x}_i,\mathbf{z}_i))}{c(\phi_k)}\right]$$

باشد که v_{ik}ها متغیرهای نشانگر موهومی مشاهده‌نشده‌ای هستند که تعلق مشاهده iام به مؤلفه kام مدل آمیخته متناهی را نشان می‌دهند. الگوریتم EM تابع لگاریتم درستنمایی کامل تاوانیده $\tilde{l}_n^c(\boldsymbol{\Psi}) = l_n^c(\boldsymbol{\Psi}) - p_n(\boldsymbol{\Psi})$ را به‌صورت تکراری در دو مرحله زیر ماکسیمم می‌کند.

گام E: در گام E امید ریاضی شرطی $\tilde{l}_n^c(\boldsymbol{\Psi})$ را نسبت به v_{ik} به‌شرط داده‌ها $(u_i,\mathbf{x}_i,\mathbf{z}_i,y_i)$ و با فرض اینکه برآورد جاری $\boldsymbol{\Psi}^{(m)}$ پارامترهای حقیقی مدل را به دست می‌دهد، محاسبه می‌کنیم. امید ریاضی شرطی عبارت است از

$$Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(m)}\left[\frac{\theta_k^{*(m)}(u_i,\mathbf{x}_i,\mathbf{z}_i)T(y_i) - A(\theta_k^{*(m)}(u_i,\mathbf{x}_i,\mathbf{z}_i))}{c(\phi_k)}\right]$$
$$+\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(m)} log\{d(y_i,\phi_k)\} + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(m)} log\pi_k - p_n(\boldsymbol{\Psi})$$

که

$$w_{ik}^{(m)} = \frac{\pi_k^{(m)} exp\left(\frac{\theta_k^{*(m)}(u_i,\mathbf{x}_i,\mathbf{z}_i)T(y_i) - A(\theta_k^{*(m)}(u_i,\mathbf{x}_i,\mathbf{z}_i))}{c(\phi_k)}\right)}{\sum_{l=1}^{K}\pi_l^{(m)} exp\left(\frac{\theta_l^{*(m)}(u_i,\mathbf{x}_i,\mathbf{z}_i)T(y_i) - A(\theta_l^{*(m)}(u_i,\mathbf{x}_i,\mathbf{z}_i))}{c(\phi_l)}\right)} \qquad (۱۰)$$

امید ریاضی‌های شرطی v_{ik} هستند.

گام M: گام M در تکرار $(m+1)$ام تابع $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(m)})$ را نسبت به $\boldsymbol{\Psi}$ ماکسیمم می‌کند. احتمال‌های آمیختگی توسط

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ik}^{(m)} \qquad k = 1,\ldots,K$$

به‌روزرسانی می‌شوند. حال با در نظر گرفتن π_k به‌عنوان مقادیر ثابت در تابع $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(m)})$، این تابع را نسبت به $\boldsymbol{\beta}$ ماکسیمم می‌کنیم. ضرایب پارامتری رگرسیون به ازای $k = 1, \ldots, K$ توسط معادله

$$\sum_{i=1}^{n} w_{ik}^{(m)} z_{ij} \frac{T(y_i)}{c(\phi_k)} = \sum_{i=1}^{n} w_{ik}^{(m)} \frac{z_{ij}}{c(\phi_k)} \frac{\partial}{\partial \theta_k^{*(m)}} A(\theta_k^{*(m)}(u_i, \mathbf{x}_i, \mathbf{z}_i)) + \pi_k \frac{\partial}{\partial \beta_{kj}} \left\{ \tilde{p}_{nk}(\beta_{kj}) \right\} \quad (۱۱)$$

به‌روزرسانی می‌شوند که در آن $\tilde{p}_{nk}(\beta_{kj})$ جمله متناظر در $\tilde{p}_n(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(m)})$ است. این برآوردگر را $\boldsymbol{\beta}$ می‌نامیم. برآوردهای ضرایب ناپارامتری نیز به ازای $k = 1, \ldots, K$ از حل

$$\sum_{i=1}^{n} w_{ik}^{(m)} x_{ij} \frac{T(y_i)}{c(\phi_k)} K_h(u_i - u) = \sum_{i=1}^{n} w_{ik}^{(m)} \frac{x_{ij}}{c(\phi_k)} \frac{\partial}{\partial \theta_k^{*(m)}} A(\theta_k^{*(m)}(u_i, \mathbf{x}_i, \mathbf{z}_i)) K_h(u_i - u)$$

به دست می‌آیند. حال در معادله (۸)، $\boldsymbol{\beta}$ را با برآورد آن یعنی $\boldsymbol{\beta}$ جایگزین کرده و تابع درستنمایی حاصل را نسبت به $\mathbf{a}$ و $\mathbf{b}$ ماکسیمم می‌کنیم. فرض کنید پاسخ‌ها به ترتیب $\hat{\mathbf{a}}$ و $\hat{\mathbf{b}}$ باشند، خواهیم داشت $\hat{\boldsymbol{\alpha}}(u) = \hat{\mathbf{a}}$. لی و لیانگ [۲۸] ثابت کرده‌اند که این برآوردگر کاراتر از $\tilde{\boldsymbol{\alpha}}(u)$ است. بنابراین الگوریتم برآورد به شکل زیر است:

گام۱: به دست آوردن $\tilde{\boldsymbol{\alpha}}(u)$ از طریق ماکسیمم کردن رابطه (۸) نسبت به $\mathbf{a}$ و $\mathbf{b}$.

گام۲: به دست آوردن $\boldsymbol{\beta}$ با استفاده از الگوریتم EM و رابطه (۱۱).

گام۳: جایگذاری $\boldsymbol{\beta}$ در رابطه (۸) و به دست آوردن $\hat{\boldsymbol{\alpha}}(u)$.

پس از به دست آوردن برآورد ناپارامتری $\boldsymbol{\alpha}(.)$ بایستی متغیرهای کمکی معنی‌دار را انتخاب کنیم. با دنبال کردن ایده انتخاب متغیر در مدل‌های رگرسیون خطی، از روش حذف پس‌رو استفاده خواهیم کرد. بنابراین، در هر مرحله آزمون

$$H_0 : \alpha_{kj_1}(u) = \ldots = \alpha_{kj_d}(u) = 0 \quad \text{در برابر} \quad H_1 : not\ all\ \alpha_{kj_l}(u) = 0$$

را به ازای $\{j_1, \ldots, j_d\}$ زیرمجموعه‌ای از $\{1, \ldots, p\}$ و $k = 1, \ldots, K$ انجام می‌دهیم. چنین آزمونی را می‌توان با استفاده از آزمون نسبت درستنمایی تعمیم‌یافته انجام داد. برای مشاهده جزئیات بیشتر به [۲ و ۲۸] مراجعه کنید.

۴-۳- انتخاب پارامتر منظم سازی

هنگام استفاده از توابع تاوان بایستی مقادیر پارامترهای منظم سازی را انتخاب کرد. بدین منظور معمولاً از روش‌های مبتنی بر داده مانند اعتبارسنجی متقابل تعمیم‌یافته استفاده می‌شود. خلیلی و چن [۱۲] ملاک اعتبارسنجی متقابل تعمیم‌یافته مؤلفه به مؤلفه‌ای را برای مدل‌های رگرسیون آمیخته متناهی معرفی کرده‌اند که بر اساس انحراف طراحی‌شده است و در تنظیمات ما نیز می‌تواند مورداستفاده قرار گیرد.

فرض کنید $(\hat{\beta}_k, \hat{\phi})$ برآوردهای ماکسیمم درستنمایی تاوانیده پارامترهای kامین مؤلفه مدل باشند که با فرض ثابت بودن سایر مؤلفه‌های $\boldsymbol{\Psi}$ در برآورد ماکسیمم درستنمایی خود تحت مدل کامل به‌دست‌آمده‌اند. همچنین تابع انحراف متناظر با kامین مؤلفه مدل را به شکل

$$\hat{D}_k(\hat{\beta}_k, \hat{\phi}) = \sum_{i=1}^{n} w_{ik} \left[log\left\{ f(y_i; y_i, \hat{\phi}_k) \right\} - log\left\{ f(y_i; \hat{\theta}_k^*(u_i, \mathbf{x}_i, \mathbf{z}_i), \hat{\phi}_k) \right\} \right]$$

تعریف کنید که w_{ik}ها در معادله (۱۰) معرفی شدند. با استفاده از این مقادیر، خلیلی و چن [۱۲] ملاک اعتبارسنجی متقابل تعمیم‌یافته خود برای مؤلفه kام مدل را به شکل

$$GCV_k(\gamma_{nk}) = \frac{D_k(\hat{\boldsymbol{\beta}}_k, \hat{\phi})}{n(1 - e(\gamma_{nk})/n)^2} \qquad k = 1, \ldots, K$$

تعریف کرده‌اند که در آن $e(\gamma_{nk})$ تعداد پارامترهای مؤثر مدل است و به شکل زیر تعریف می‌شود:

$$e(\gamma_{nk}) = tr\left\{ \left[l_k''(\hat{\boldsymbol{\beta}}_k, \hat{\phi}) - \Sigma_k(\hat{\boldsymbol{\beta}}_k) \right]^{-1} l_k''(\hat{\boldsymbol{\beta}}_k, \hat{\phi}) \right\}.$$

همچنین $l_k''(\hat{\boldsymbol{\beta}}_k, \hat{\phi})$ مشتق دوم لگاریتم تابع درستنمایی نسبت به $\boldsymbol{\beta}_k$ است که در $(\hat{\boldsymbol{\beta}}_k, \hat{\phi})$ ارزیابی‌شده، tr نمایانگر اثر ماتریس است و داریم:

$$\Sigma_k(\hat{\boldsymbol{\beta}}_k) = \hat{\pi}_k diag\left\{ p_{nk}'(\hat{\beta}_{k1}) / \hat{\beta}_{k1}, \ldots, p_{nk}'(\hat{\beta}_{kP}) / \hat{\beta}_{kP} \right\}.$$

پارامترهای منظم سازی از طریق مینیمم‌سازی $GCV(\gamma_{nk})$ انتخاب می‌شوند. برای مشاهده جزییات بیشتر به [۱۲] مراجعه کنید.

۵- خواص مجانبی

بردار ضرایب رگرسیون $\boldsymbol{\beta}_k$ در مؤلفه kام را به $\boldsymbol{\beta}_k^t=\{\boldsymbol{\beta}_{1k}^t,\boldsymbol{\beta}_{2k}^t\}$ به‌گونه‌ای تجزیه کنید که $\boldsymbol{\beta}_{2k}$ عناصر صفر را در برگیرد. به‌طورکلی ممکن است بردار عناصر غیر صفر $\boldsymbol{\beta}_{1k}$ به k وابسته باشد. بدون از دست دادن کلیت و به‌منظور سادگی نمادها از این واقعیت چشم‌پوشی می‌کنیم. به‌طور طبیعی پارامترهای $\boldsymbol{\Psi}^t=\{\boldsymbol{\Psi}_1^t,\boldsymbol{\Psi}_2^t\}$ را به‌گونه‌ای تقسیم می‌کنیم که $\boldsymbol{\Psi}_2^t$ شامل تمامی اثرات صفر یعنی $\boldsymbol{\beta}_{2k}:k=1,\ldots,K$ شود. بردار پارامترهای حقیقی را نیز با $\boldsymbol{\Psi}_\circ$ نشان خواهیم داد. عناصر $\boldsymbol{\Psi}_\circ$ را با یک زیرنویس مانند β_{kj}^0 نشان خواهیم داد. برای ارائه نتایج مجانبی به نمادهای زیر نیاز خواهیم داشت:

$$a_n=\max_{k,j}\{p_{nk}(\beta_{kj}^\circ)/\sqrt{n}:\beta_{kj}^\circ\neq\circ\}$$

$$b_n=\max_{k,j}\{|\,p_{nk}^{'}(\beta_{kj}^\circ)\,|/\sqrt{n}:\beta_{kj}^\circ\neq\circ\}$$

$$c_n=\max_{k,j}\{|\,p_{nk}^{''}(\beta_{kj}^\circ)\,|/n:\beta_{kj}^\circ\neq\circ\}$$

که $p_{nk}^{'}(\beta)$ و $p_{nk}^{''}(\beta)$ مشتقات مراتب اول و دوم تابع $p_{nk}(\beta)$ نسبت به β است. همچنین نتایج مجانبی بر اساس شرایطی روی توابع تاوان $p_{nk}(.)$ هستند که در ضمیمه یک ارائه شدند. علاوه بر این، به‌منظور توسعه نظریه مجانبی به برخی شرایط نظم معمول بر روی تابع چگالی توأم $f(\mathbf{w};\boldsymbol{\Psi})$ متعلق به $\mathbf{W}=(\mathbf{x},\mathbf{z},u,Y)$ نیاز داریم که در ضمیمه یک آورده شده‌اند.

قضیه۲: فرض کنید $\mathbf{W}_i=(\mathbf{x}_i,\mathbf{z}_i,u_i,Y_i),\ i=1,2,\ldots,n$ نمونه‌ای تصادفی از تابع چگالی $f(\mathbf{w};\boldsymbol{\Psi})$ باشد که در شرایط نظم صدق می‌کند. همچنین فرض کنید توابع تاوان $p_{nk}(.)$ در شرایط $P_\circ$ و P_1 صدق کنند. آنگاه یک ماکسیمم کننده موضعی $\hat{\boldsymbol{\Psi}}_n$ برای تابع لگاریتم درستنمایی تاوانیده $\tilde{l}_n(\boldsymbol{\Psi})$ وجود دارد که نرخ همگرایی آن برابر $O_p\{n^{-1/2}(1+b_n)\}$ است.

اثبات. اثبات این قضیه مشابه اثبات قضیه ۱ [۲۲] بوده و از بیان آن صرف‌نظر می‌کنیم. ■

خلیلی و چن [۱۲] نشان داده‌اند که به ازای MIXHARD، MIXSCAD و MIXLASSO با انتخاب مناسب پارامترهای منظم سازی $b_n=O(1)$ و $\hat{\boldsymbol{\Psi}}$ دارای نرخ همگرایی معمول $n^{-1/2}$ است. خصوصیت مهم دیگر تنک بودن است که انتخاب متغیر سازگار را ممکن می‌سازد. قضیه بعد خاصیت تنک بودن را تحت شرایط ضعیفی ثابت می‌کند.

قضیه۳: شرایط اشاره‌شده در قضیه (۲) را در نظر بگیرید، همچنین فرض کنید که توابع تاوان $p_{nk}(.)$ در شرایط P_0 - P_2 صدق می‌کنند و مقدار K در بخش‌های (الف) و (ب) زیر معلوم است. داریم:

الف) به ازای هر $\boldsymbol{\Psi}$ که $\| \boldsymbol{\Psi} - \boldsymbol{\Psi}_0 \| = O(n^{-1/2})$ با احتمال متمایل به یک داریم

$$\tilde{l}_n\{(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)\} - \tilde{l}_n\{(\boldsymbol{\Psi}_1, 0)\} < 0$$

ب) به ازای هر برآوردگر ماکسیمم درستنمایی تاوانیده $\sqrt{n}$-سازگار $\hat{\boldsymbol{\Psi}}_n$ برای پارامتر $\boldsymbol{\Psi}$ داریم:

i. تنک بودن. هنگامی‌که $n \to \infty$ داریم $P\{\boldsymbol{\beta}_{2k} = 0\} \to 1,\ k = 1, \ldots, K$

ii. نرمال مجانبی بودن.

$$\sqrt{n}\{\left[I_1(\boldsymbol{\Psi}_{01}) - \frac{p_n''(\boldsymbol{\Psi}_{01})}{n} \right](\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) + \frac{p_n'(\boldsymbol{\Psi}_{01})}{n}\} \xrightarrow{d} N(0, I_1(\boldsymbol{\Psi}_{01}))$$

که $I_1(\boldsymbol{\Psi}_{01})$ اطلاع فیشر محاسبه‌شده تحت مدل کاهش‌یافته، یعنی پس از حذف تمام اثرات صفر است.

ج) چنانچه K به‌صورت جداگانه و سازگار توسط $\hat{K}_n$ برآورد شده باشد، تا زمانی که از $\hat{K}_n$ در انتخاب متغیر استفاده شود، نتایج (الف) و (ب) همچنان برقرار است.

اثبات. اثبات این قضیه مشابه اثبات قضیه ۲ [۲۲] است. ■

بایستی به این نکته توجه داشت که مشتقات $p_n(.)$ در (ب- ii) به ازای برخی انتخاب‌های تابع تاوان قابل‌چشم‌پوشی می‌شوند و تصحیح‌های حجم نمونه متناهی را به همراه خواهند داشت. این مطلب برآوردگر واریانس $\hat{\boldsymbol{\Psi}}_1$ به فرم

$$\hat{\text{var}}(\hat{\boldsymbol{\Psi}}_1) = \{l_n''(\hat{\boldsymbol{\Psi}}_1) - p''(\boldsymbol{\Psi})\}^{-1} \hat{\text{var}}\{l_n'(\hat{\boldsymbol{\Psi}}_1)\}\{l_n''(\hat{\boldsymbol{\Psi}}_1) - p''(\boldsymbol{\Psi})\}^{-1} \qquad (۱۲)$$

را پیشنهاد می‌کند.

۶- مطالعه شبیه‌سازی

در این بخش روش ارائه‌شده را بر روی دو توزیع یک پارامتری خانواده نمایی اعمال می‌کنیم. نتایج شبیه‌سازی به‌صورت زیر است.

مثال۱: شبیه‌سازی اول آمیخته متناهی مدل نیم پارامتری تعمیم‌یافته پواسون است. در اینجا دو جامعه را با احتمال‌های آمیختگی (0/7,0/3) مدنظر قرار خواهیم داد. فرض کنید X و

Z دارای توزیع نرمال چندمتغیره به ترتیب با میانگین‌های $\mu_1^t=(1,0)$ و $\mu_2^t=(0,0,0,0)$ و ماتریس کوواریانس مشترک، دارای ساختار $\sigma_{ij}=0/5^{|i-j|}$ باشند. همچنین U را دارای توزیع $U(0,1)$ در نظر گرفته‌ایم. بردارهای $\boldsymbol{\alpha}$ و $\boldsymbol{\beta}$ را نیز متناظر مقادیر زیر در نظرخواهیم گرفت.

$$\boldsymbol{\alpha}_1=exp(2u-1)\,,\ \boldsymbol{\alpha}_2=8u(1-u)\,,\ \boldsymbol{\beta}_1=(2,0,1,0)\,,\ \boldsymbol{\beta}_2=(0,2,0,0)$$

با استفاده از این کمیت‌ها در k امین جامعه پواسون تعریف می‌کنیم:

$$\log(\lambda_k)=\mathbf{x}^t\boldsymbol{\alpha}_k(u)+\mathbf{z}^t\boldsymbol{\beta}_k\,.$$

بنابراین با در نظر گرفتن اندازه ۱۰۰ برای هر جامعه و متوسط‌گیری بر روی ۲۰۰۰ اجرا با دو جامعه پواسون P(0/9997) و P(1/0006) مواجه هستیم. ۲۰۰۰ مشاهده از این مدل آمیخته شبیه‌سازی‌شده و روش پیشنهادی روی آن اعمال‌شده است. این عمل ۲۰۰۰ بار تکرار شده و نتایج زیر به‌دست‌آمده‌اند.

به‌منظور ارزیابی عملکرد رویه انتخاب متغیر پیشنهادی از شاخص RGMSE استفاده خواهیم کرد که به‌عنوان نسبت GMSE مدل انتخاب‌شده نهایی به GMSE مدل کامل تعریف می‌شود:

$$GMSE(\boldsymbol{\beta})=(\boldsymbol{\beta}-\boldsymbol{\beta})E(\mathbf{ZZ}^t)(\boldsymbol{\beta}-\boldsymbol{\beta})\,.$$

جهت مشاهده نحوه معرفی این شاخص می‌توانید به [۲۸] مراجعه کنید. در جداول بعد از میانگین $\boldsymbol{\beta}$ برآورد شده در ۲۰۰۰ اجرا به‌عنوان $\boldsymbol{\beta}$ استفاده کرده‌ایم. متوسط تعداد ضرایب صفر در جدول ۱ گزارش‌شده است. در این جدول ستون "C" تعداد متوسط ضرایب صفر حقیقی را که در طول ۲۰۰۰ اجرا به‌صورت صحیح برابر صفر قرار داده شده‌اند به دست می‌دهد. توجه داشته باشید که در فرایند شبیه‌سازی، ۵ پارامتر برابر صفر در نظر گرفته‌شده بودند. همان‌گونه که مشاهده می‌شود تنها برای تابع تاوان *MIXSCAD* اعداد گزارش‌شده نزدیک ۵ است و لذا در اغلب موارد ضرایبی را که در حقیقت صفر بوده‌اند، به‌عنوان صفر برآورد کرده‌ایم. همچنین مشاهده می‌شود که به‌طورکلی دقت توابع تاوان *MIXHARD* و *MIXSCAI* بیش از تابع تاوان *MIXLASSO* است. ستون "I" تعداد متوسط ضرایب غیر صفری را که به‌صورت غیر صحیح برابر صفر گزارش‌شده‌اند نشان می‌دهد. ملاحظه می‌کنیم که در شبیه‌سازی‌های انجام‌شده هرگز ضریب غیر صفری به‌عنوان صفر برآورد نشده است. در تاوان *MIXSCAD* از مقدار $a=3/7$ استفاده کرده‌ایم؛ همچنین همواره از پهنای باند $h=0/125$ و تابع ربط اپانچینکوف[1] $K(u)=0/75(1-u^2)I\left(|u|\le 1\right)$ استفاده کرده‌ایم. به‌علاوه الگوریتم را تا زمانی

1- Epanechnikov Kernel

تکرار می‌کنیم که نرم تفاوت بین بردار پارامترهای حقیقی و بردار پارامترهای برآوردشده کمتر از ۰/۰۰۱ باشد.

جدول ۱: مقایسه انتخاب متغیر برای آمیخته متناهی پواسون

تاوان	RGMSE	C	I
MIXLASSO	0/6911	3/5148	۰
MIXHARD	0/5772	3/5030	۰
MIXSCAD	0/4434	4/3956	۰

میانگین و انحراف استاندارد ۲۰۰۰ تکرار برآوردها و خطاهای استاندارد ارائه‌شده در جدول ۲ را به دست می‌دهد. به‌عنوان مثال برآورد $\hat{\beta}_{11}$ تحت توابع تاوان *MIXLASSO*، *MIXHARD* و *MIXSCAD* به ترتیب برابر 2/0803 ، 2/038 و 1/995 است. همچنین انحراف استاندارد این برآوردها نیز به ترتیب 0/1366، 0/1360 و 0/1339 است. این در حالی است که مقدار واقعی این پارامتر برابر 2 بوده است.

جدول۲: برآوردها و انحراف استاندارد آن‌ها در آمیخته متناهی پواسون براساس ۲۰۰۰ تکرار. انحرافات استاندارد داخل پرانتز داده‌شده‌اند.

β	β		
	MIXLASSO	*MIXHARD*	*MIXSCAD*
۲	$2/0803_{(0/1366)}$	$2/0381_{(0/1360)}$	$1/9957_{(0/1339)}$
۰	$0/3469_{(0/1158)}$	$0/3478_{(0/1136)}$	$0/3484_{(0/1113)}$
۱	$1/0375_{(0/1112)}$	$1/0166_{(0/1089)}$	$0/9964_{(0/1063)}$
۰	$0/0643_{(0/0535)}$	$0/0629_{(0/0524)}$	$0/0613_{(0.0507)}$
۰	$0/3069_{(0/0700)}$	$0/3103_{(0/0692)}$	$0/3137_{(0/0695)}$
۲	$2/0275_{(0/0955)}$	$2/0117_{(0/0954)}$	$2/0180_{(0/0972)}$
۰	$0/1534_{(0/0595)}$	$0/1551_{(0/0596)}$	$0/1574_{(0/0600)}$
۰	$0/0400_{(0/0333)}$	$0/0401_{(0/0330)}$	$0/0402_{(0/0328)}$

در جدول ۲ مشاهده می‌شود که تابع تاوان *MIXSCAD* نسبت به تاوان‌های *MIXHARD* و *MIXLASSO* نتایج دقیق‌تری به دست می‌دهد. به‌عنوان مثال می‌دانیم که مؤلفه چهارم این جامعه برابر صفر در نظر گرفته‌شده است. تاوان *MIXSCAD* برآورد 0/0613 را به دست می‌دهد که نسبت به برآوردهای حاصل تحت توابع تاوان *MIXHARD* و *MIXLASSO* (به ترتیب 0/0629 و 0/0643) به صفر نزدیک‌تر است. به‌هرحال برای جامعه دوم نتایج متفاوت است. در این جامعه تابع تاوان *MIXLASSO* برآوردهای دقیق‌تری را به دست می‌دهد. به‌عنوان مثال، برای عنصر اول جامعه دوم (مؤلفه پنجم جدول ۲) که مجدداً برابر صفر است برآورد حاصل تحت تابع تاوان *MIXLASSO* برابر 0/3069 است، درحالی‌که برآورد حاصل از تاوان *MIXSCAD* برابر 0/3137 است؛ واضح است که برآورد حاصل از تاوان *MIXLASSO* دقیق‌تر است.

جدول ۳: فواصل اطمینان همزمان مبتنی بر T^2 هتلینگ برای آمیخته متناهی پواسون

β					
MIXLASSO		*MIXHARD*		*MIXSCAD*	
۲/۰۶۸۲	۲/۰۹۲۳	۲/۰۲۶۱	۲/۰۵۰۱	۱/۹۸۳۹	۲/۰۰۷۶
۰/۳۳۶۷	۰/۳۵۷۲	۰/۳۳۷۸	۰/۳۵۷۹	۰/۳۳۸۶	۰/۳۵۸۲
۱/۰۲۷۶	۱/۰۴۷۳	۱/۰۰۷۰	۱/۰۲۶۲	۰/۹۸۷۰	۱/۰۰۵۸
۰/۰۵۹۵	۰/۰۶۹۰	۰/۰۵۸۲	۰/۶۷۵۰	۰/۰۵۶۸	۰/۰۶۵۸
۰/۳۰۰۸	۰/۳۱۳۱	۰/۳۰۴۲	۰/۳۱۶۴	۰/۳۰۷۵	۰/۳۱۹۸
۲/۰۱۹۱	۲/۰۳۵۹	۲/۰۰۳۳	۲/۰۲۰۲	۲/۰۰۹۴	۲/۰۲۶۶
۰/۱۴۸۱	۰/۱۵۸۶	۰/۱۴۹۹	۰/۱۶۰۴	۰/۱۵۲۱	۰/۱۶۲۷
۰/۰۳۷۰	۰/۰۴۲۹	۰/۰۳۷۲	۰/۰۴۳۰	۰/۰۳۷۳	۰/۰۴۳۱

جدول ۳ فواصل اطمینان همزمان ۹۵٪ پارامترهای برآورد شده را بر اساس T^2 هتلینگ نشان می‌دهد. بر اساس این جدول، فاصله اطمینان $\hat{\beta}_{11}$ تحت تابع تاوان *MIXLASSO* برابر (2/0682,2/0923) است. شایان‌ذکر است اعداد محاسبه‌شده بر اساس توزیع توأم بردار پارامترها (و نه به‌صورت تکی) محاسبه شده‌اند.

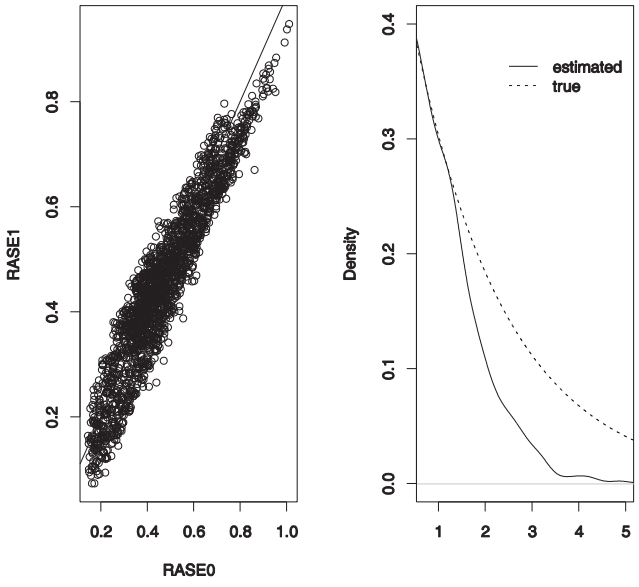
به‌منظور ارزیابی دقیق‌تر عملکرد الگوریتم برآورد پیشنهادی برای مؤلفه ناپارامتری از جذر متوسط توان‌های دوم خطا[1] (*RASE*) استفاده خواهیم کرد به شکل زیر تعریف می‌شود.

1- Square Root of Average Square Errors

$$RASE = \left\{ \frac{1}{n_{grid}} \sum_{k=1}^{n_{grid}} \left\| \hat{\alpha}(\mathbf{z}_k) - \alpha(\mathbf{z}_k) \right\|^2 \right\}^{\frac{1}{2}}$$

و در آن $\{\mathbf{z}_k, k = 1, \ldots, n_{grid}\}$ نقاط شبکه‌ای هستند که توابع $\{\hat{\alpha}_j(.)\}$ در آن‌ها ارزیابی شده‌اند. در اینجا از $n_{grid} = 200$ استفاده کرده‌ایم. در شکل ۱ $RASE$ مربوط به $\hat{\alpha}(\mathbf{z}_k)$ با استفاده از الگوریتم پیشنهادی ($RASE1$) را در برابر کمیت مشابه با استفاده از $\boldsymbol{\beta}$ حقیقی ($RASE0$) رسم کرده‌ایم. با توجه به اینکه نقاط حول خط $x = y$ تغییر می‌کنند می‌توان دقت مناسب برآورد بخش ناپارامتری مدل را نتیجه گرفت. حال کارایی رویه آزمون پیشنهادی در بخش ۵ را ارزیابی می‌کنیم. بدین منظور فرض زیر را در نظر بگیرید.

$$H_0 : \boldsymbol{\alpha}_2(u) = 0 \quad \text{در برابر} \quad H_1 : \boldsymbol{\alpha}_2(u) \neq 0.$$



شکل ۱: توزیع صفر و χ^2 در آمیخته متناهی پواسون

اجازه دهید بررسی کنیم که آیا توزیع صفر نمونه متناهی آزمون نسبت درستنمایی تعمیم‌یافته پیشنهادی نزدیک توزیع کای‌اسکور است یا خیر. بدین منظور توزیع صفر آزمون نسبت درستنمایی تعمیم‌یافته پیشنهادی را در هر یک از ۲۰۰۰ اجرا به دست آورده‌ایم.

برآورد هسته چگالی توزیع صفر در شکل ۱ نشان داده‌شده است. در این شکل خط پررنگ، تابع چگالی برآورد شده و خط نقطه‌چین توزیع کای‌اسکور را با دو درجه آزادی (که تعداد مؤلفه‌های مشخص‌شده تحت فرض صفر است) نشان می‌دهد. مشاهده می‌شود که توزیع صفر نمونه متناهی کاملاً نزدیک توزیع کای‌اسکور است. این مطلب توسط آزمون کولموگروف – اسمیرنوف ($D=0/5575,\ p-value=0/8856$) نیز اثبات می‌شود.

جهت مشاهده سایر خواص آزمون نسبت درستنمایی تعمیم‌یافته پیشنهادی می‌توانید به [۲۸] مراجعه نمایید.

مثال۲: شبیه‌سازی دوم بر اساس آمیخته متناهی مدل نیم پارامتری تعمیم‌یافته دوجمله‌ای صورت پذیرفته است. در اینجا دو جامعه را با احتمال‌های آمیختگی $(0/7,0/3)$ مدنظر قرار خواهیم داد. متغیرهای کمکی را مشابه مثال قبل در نظر می‌گیریم. بردارهای $\boldsymbol{\alpha}$ و $\boldsymbol{\beta}$ را نیز متناظر مقادیر زیر در نظرخواهیم گرفت.

$$\boldsymbol{\alpha}_1=exp(2u-1),\ \boldsymbol{\alpha}_2=2(sin(2\pi u))^2,\ \boldsymbol{\beta}_1=(2,0,1,0),\ \boldsymbol{\beta}_2=(0,2,0,0)$$

با استفاده از این کمیت‌ها در kامین جامعه دوجمله‌ای تعریف می‌کنیم:

$$logit(p_k)=\mathbf{x}^t\boldsymbol{\alpha}_k(u)+\mathbf{z}^t\boldsymbol{\beta}_k\ .$$

بنابراین با در نظر گرفتن اندازه ۱۰۰ برای هر جامعه و متوسط‌گیری بر روی ۲۰۰۰ اجرا با دو جامعه دوجمله‌ای $B(100,0/4999)$ و $B(100,0/5000)$ مواجه هستیم. ۲۰۰۰ مشاهده از این مدل آمیخته شبیه‌سازی شده و روش پیشنهادی روی آن اعمال‌شده است. این عمل ۲۰۰۰ بار تکرار شده و نتایج زیر به‌دست‌آمده‌اند.

شاخص RGMSE و تعداد متوسط ضرایب صفر در جدول ۴ داده‌شده‌اند. همان‌گونه که مشاهده می‌شود اعداد ستون "C" برای تاوان *MIXSCAD* نزدیک ۵ است و لذا در اغلب موارد ضرایبی را که در حقیقت صفر بوده‌اند برابر صفر برآورد کرده‌ایم. به‌علاوه مشاهده می‌کنیم که دقت توابع تاوان *MIXLASSO* و *MIXHARL* کمتر از تابع تاوان *MIXSCAD* است. همچنین مشاهده می‌شود که هرگز ضریب غیر صفری را برابر صفر برآورد نکرده‌ایم. مجدداً الگوریتم تا زمانی تکرار می‌شود که نرم تفاضل بین بردار و پارامتر حقیقی و برآوردشده کمتر از ۰/۰۰۱ باشد. میانگین و انحراف استاندارد ۲۰۰۰ تکرار برآوردها و خطاهای استاندارد ارائه‌شده در جدول ۵ را به دست می‌دهد. به‌عنوان مثال برآورد $\hat{\beta}_{11}$ تحت توابع تاوان *MIXLASSO*، *MIXHARD* و *MIXSCA* به ترتیب برابر $2/0204$ ، $2/021$ و $1/985$ است. همچنین انحراف استاندارد این برآوردها نیز به ترتیب $0/2680$، $0/2682$ و $0/2640$ است. این در حالی است که مقدار واقعی این پارامتر برابر 2 بوده است.

جدول ۴: مقایسه انتخاب متغیر برای آمیخته متناهی دوجمله‌ای

I	C	RGMSE	تاوان
0	3/5060	0/7232	*MIXLASSO*
0	3/5080	0/6238	*MIXHARD*
0	4/6974	0/4881	*MIXSCAD*

جدول ۵: برآوردها و انحراف استاندارد آن‌ها در آمیخته متناهی دوجمله‌ای بر اساس ۲۰۰۰ تکرار. انحرافات استاندارد داخل پرانتز داده‌شده‌اند.

β			β
MIXSCAD	*MIXHARD*	*MIXLASSO*	
$1/9858_{(0/2640)}$	$2/0212_{(0/2682)}$	$2/0204_{(0/2680)}$	۲
$0/1635_{(0/0932)}$	$0/1574_{(0/0938)}$	$0/1576_{(0/0938)}$	۰
$0/9925_{(0/1418)}$	$1/0103_{(0/1435)}$	$1/0097_{(0/1434)}$	۱
$-0/0002_{(0/0550)}$	$-0/0001_{(0/0550)}$	$0/0000_{(0/0550)}$	
$0/1901_{(0/0735)}$	$0/1909_{(0/0732)}$	$0/1910_{(0/0732)}$	
$2/0151_{(0/2566)}$	$2/0112_{(0/2542)}$	$2/0108_{(0/2542)}$	۰
$0/0976_{(0/0717)}$	$0/0978_{(0/0718)}$	$0/0979_{(0/0717)}$	۰
$-0/0009_{(0/0607)}$	$-0/0011_{(0/0608)}$	$-0/0011_{(0/0608)}$	۰

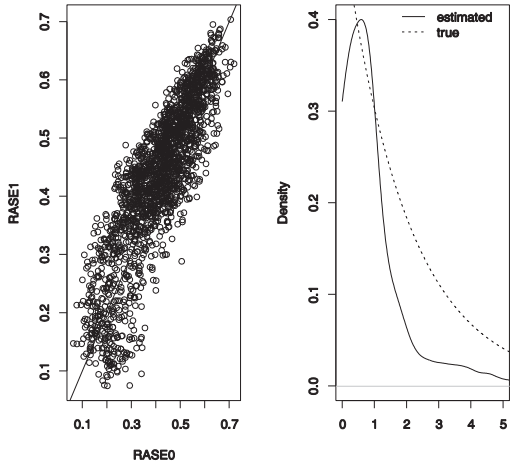
جدول ۶ فواصل اطمینان همزمان ۹۵٪ پارامترهای برآوردشده را بر اساس T^2 هتلینگ نشان می‌دهد. بر اساس این جدول فاصله اطمینان $\hat{\beta}_{11}$ تحت تابع تاوان *MIXLASSO* برابر $(1/9967, 2/0440)$ است. شایان‌ذکر است اعداد محاسبه‌شده بر اساس توزیع توأم بردار پارامترها (و نه به‌صورت تکی) محاسبه شده‌اند.

نمودارهای RASE و توزیع آماره آزمون در شکل ۲ رسم گردیده و نتایجی مشابه مثال قبل به دست می‌دهند. در این شکل، خط پررنگ، تابع چگالی برآوردشده و خط نقطه‌چین توزیع کای‌اسکور را با دو درجه آزادی (که تعداد مؤلفه‌های مشخص‌شده تحت فرض صفر است) نشان می‌دهد. مشاهده می‌شود که توزیع صفر نمونه متناهی نسبتاً نزدیک توزیع کای‌اسکور است. این

مطلب توسط آزمون کلموگروف اسمیرنوف ($D=0/5065,\ p-value=0/9875$) نیز تائید می‌شود.

جدول ۶: فواصل اطمینان همزمان مبتنی بر T^2 هتلینگ برای آمیخته متناهی دوجمله‌ای

β					
MIXLASSO		*MIXHARD*		*MIXSCAD*	
۱/۹۹۶۷	۲/۰۴۴۰	۱/۹۹۷۵	۲/۰۴۴۹	۱/۹۶۲۵	۲/۰۰۹۱
۰/۱۴۹۴	۰/۱۶۵۹	۰/۱۴۹۱	۰/۱۶۵۷	۰/۱۵۵۲	۰/۱۷۱۷
۰/۹۹۷۰	۱/۰۲۲۴	۰/۹۹۷۶	۱/۰۲۳۰	۰/۹۸۰۰	۱/۰۰۵۰
-۰/۰۰۴۸	۰/۰۰۴۹	-۰/۰۰۴۹	۰/۰۰۴۸	-۰/۰۰۵۱	۰/۰۰۴۶
۰/۱۸۴۵	۰/۱۹۷۵	۰/۱۸۴۵	۰/۱۹۷۴	۰/۱۸۳۶	۰/۱۹۶۶
۱/۹۸۸۴	۲/۰۳۳۳	۱/۹۸۸۷	۲/۰۳۳۶	۱/۹۹۲۴	۲/۰۳۷۸
۰/۰۹۱۵	۰/۱۰۴۲	۰/۰۹۱۵	۰/۱۰۴۲	۰/۰۹۱۳	۰/۱۰۴۰
-۰/۰۰۶۴	۰/۰۰۴۳	-۰/۰۰۶۴	۰/۰۰۴۳	-۰/۰۰۶۳	۰/۰۰۴۴



شکل ۲: توزیع صفر و χ^2 در آمیخته متناهی دوجمله‌ای

مراجع

[1] Breiman, L. (1996), Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**, 2350–2383.

[2] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

[3] Ma, S., Song, Q. and Wang, L. (2013), Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data, *Bernoulli*, **19**, 252-274.

[4] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc., Series B.*, **58**, 267-288.

[5] Fan, J. and Li, R. (2002), Variable selection for cox's proportional hazards model and frailty model, *Ann. Statist.*, **30**, 74-99.

[6] Fan, J. and Peng, H. (2004), Nonconcave penalized likelihood with a diverging number of parameters, *Amer. Statist.*, **32**, 928-961.

[7] Hunter, D.R. and Li, R. (2005), Variable selection using MM algorithm, *Amer. Statist.*, **33**, 1617-1642.

[8] Wu, Y. and Liu, Y. (2009), Variable selection in quantile regression, *Statist. Sinica.*, **19**, 801-817.

[9] Zou, H. (2006), The adaptive lasso and its oracle properties, *J. Amer. Statist. Assoc.*, **101**, 1418-1429.

[10] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n, *Amer. Statist.*, **6**, 2313-2351.

[11] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. R. Statist. Soc. B.*, **70**, Part 5, 849-911.

[12] Khalili, A. and Chen, J. (2007), Variable selection in finite mixture of regression models, *J. Amer. Statist. Assoc.*, **102**, 1025-1038.

[13] McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

[14] Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, *in Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski. Budapest: Akademiai Kiado, page 267.

[15] Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461-464.

[16] Tibshirani, R. (1997), The Lasso method for variable selection in the Cox model, *Statistics in Medicine*, **16**, 385-395.

[17] Yuan, M. and Lin, Y. (2007), On the nonnegative garrote estimator, *J. Roy. Statist. Soc. Ser. B.*, **69**, 143-161.

[18] Fan, J, Feng, Y. and Song, R. (2011), Nonparametric independence screening in ultra-high dimensional additive models, *J. Amer. Statist. Assoc.*, **106**, 544-557.

[19] Cho, H. and Fryzlewicz, P. (2012), High-dimensional variable selection via tilting, *J. Roy. Stat. Soc. Ser. B Stat. Metodol.*, **74**, 593-622.

[20] Khalili , A. (2011), An overview of the new feature selection methods in finite mixture of regression models, *JIRSS*, **10**, 201-235.

[21] Du, Y., Khalili, A., Neslehova, J.G., and Steele, R.J., (2013), Simultaneous fixed and random effects selection in finite mixture of linear mixed-effects models, *The Canadian Journal of Statistics*, **41**, 596-616.

[22] Ormoz, E., and Eskandari, F., (2013), Variable selection in finite mixture of semi-parametric regression models, *COMMUN STAT-THEOR M*, to appear.

[23] Nelder, J., and Wedderburn, R.W.M, (1972), Generalized linear models, *J. Roy. Statist. Soc. Ser. A.*, **135**, 370-384.

[24] Hennig, C. (2000), Identifiability of models for clusterwise linear regression, *Journal of Classification*, **17**, 273-296.

[25] Wang, P., Puterman, M. L., Cockburn, I. and Le, N. (1996), Mixed Poisson regression models with covariate dependent rates, *Biometrics*, **52**, 381-400.

[26] Titterington, D. M., Smith, A. F. M., and Markov, U. E. (1985), *Statistical analysis of finite mixture distributions*, New York: Wiley.

[27] Grun, B. and Leisch, F. (2008), *Finite mixtures of generalized linear regression models*. In Shalabh and Christian Heumann, editors, Recent Advances in Linear Models and Related Areas, pages 205-230. Springer.

[28] Li, R. and Liang, H. (2008), Variable selection in semiparametric regression modeling, *Ann. Statist.*, **36**, 261-286.

ضمیمه یک:

نتایج مجانبی مطرح‌شده در مقاله بر اساس شرایط زیر برای توابع تاوان $p_{nk}(.)$ است.

P_0: به ازای تمام مقادیر n و k، $p_{nk}(0)=0$ و $p_{nk}(\beta)$ متقارن و غیر منفی است. به‌علاوه غیرنزولی بوده و به ازای تمامی β، در $(0,\infty)$ به‌استثنای حداکثر چند استثنا، دو مرتبه مشتق‌پذیر است.

P_1: با میل کردن n به بی‌نهایت، $a_n=o(1+b_n)$ و $c_n=o(1)$

P_2: به ازای $N_n=\{\beta;0<\beta\le n^{-\frac{1}{2}}\log n\}$ داریم $\lim\limits_{n\to\infty}\inf\limits_{\beta\in N_n}\dfrac{p'_{nk}(\beta)}{\sqrt{n}}=\infty$

شرایط P_0 و P_2 برای تنک بودن یا انتخاب سازگار متغیرها لازم هستند. شرط P_1 برای حفظ خصوصیات مجانبی برآوردگرهای اثرات غیر صفر مدل مورداستفاده قرار می‌گیرد.

علاوه بر این، به‌منظور توسعه نظریه مجانبی به برخی شرایط نظم معمول بر روی تابع چگالی توأم $f(\mathbf{w};\boldsymbol{\Psi})$ متعلق به $\mathbf{W}=(\mathbf{x},\mathbf{z},u,Y)$ نیاز داریم که به شرح زیر هستند:

A_1: چگالی $f(\mathbf{w};\boldsymbol{\Psi})$ به ازای تمام $\boldsymbol{\Psi}\in\boldsymbol{\Omega}$ دارای دامنه مشترکی در $\mathbf{w}$ بوده و $f(\mathbf{w};\Psi)$ تا جایگشتی از مؤلفه‌های آمیخته برحسب $\boldsymbol{\Psi}$ شناسایی پذیر باشد.

A_2: به ازای هر $\boldsymbol{\Psi}\in\boldsymbol{\Omega}$ چگالی $f(\mathbf{w};\boldsymbol{\Psi})$ تقریباً به ازای تمام $\mathbf{w}$ نسبت به $\boldsymbol{\Psi}$ تا مرتبه سوم مشتق‌پذیر است.

A_3: به ازای هر $\boldsymbol{\Psi}_0\in\boldsymbol{\Omega}$ توابع $M_1(\mathbf{w})$ و $M_2(\mathbf{w})$ (احتمال وابسته به $\boldsymbol{\Psi}_0$) وجود دارند، به‌طوری‌که به ازای $\boldsymbol{\Psi}$ در یک همسایگی $N(\boldsymbol{\Psi}_0)$ داریم

$$\left|\frac{\partial f(\mathbf{w};\boldsymbol{\Psi})}{\partial\psi_j}\right|\le M_1(\mathbf{w})\qquad \left|\frac{\partial^2 f(\mathbf{w};\boldsymbol{\Psi})}{\partial\psi_j\partial\psi_l}\right|\le M_1(\mathbf{w})\qquad \left|\frac{\partial^3 logf(\mathbf{w};\boldsymbol{\Psi})}{\partial\psi_j\partial\psi_l\partial\psi_m}\right|\le M_2(\mathbf{w})$$

به قسمی که $\int M_1(\mathbf{w})d\mathbf{w}<\infty$ و $\int M_2(\mathbf{w})f(\mathbf{w};\boldsymbol{\Psi})d\mathbf{w}<\infty$

A_4: ماتریس اطلاع فیشر $I(\boldsymbol{\Psi})=E\left\{\left[\dfrac{\partial}{\partial\Psi}\log f(\mathbf{W};\boldsymbol{\Psi})\right]\left[\dfrac{\partial}{\partial\boldsymbol{\Psi}}\log f(\mathbf{W};\boldsymbol{\Psi})\right]^t\right\}$

به ازای هر $\boldsymbol{\Psi}\in\boldsymbol{\Omega}$ متناهی و معین مثبت است.

Variable Selection of Generalized Semi-Parametric Mixture Models

Farzad Eskandari[*], Ehsan Armaz[**], Rahman Farnoosh[***]

[*]Department of Statistics, Allameh Tabataba'i University, Tehran, Iran

[**]Department of Statistics, Mashhad branch, Islamic Azad University, Mashhad, Iran.

[**]School of Mathematics, Iran University of Science and Technology, Tehran, Iran.

Abstract

The purpose of this paper is identifying best covariates of a semi-parametric model in the presence of penalized coefficients. It should be noted that in each model, coefficients of the existing variables is considered as a combination of parameters where some of them affect the response variable linearly and some of them functionally. So, semi-parametric method was considered as an optimum solution. In this paper, we concerned with variable selection in finite mixture of generalized semi-parametric models. This task consists of model selection for nonparametric component and variable selection for parametric part. Thus, we encounter with separate model selection for each nonparametric component of each sub model. To overcome to this computational burden, we introduce a class of variable selection procedures for finite mixture of generalized semi-parametric models. It is shown that the new method is consistent for variable selection. Simulations show that the performance of proposed method is good and improve pervious works in this area and requires much less computing power than existing methods.

Keywords: Variable selection, Semi-parametric model, Finite mixture model, Penalized likelihood.

Mathematics Subject Classification (2010): 62J12