

انتخاب متغیر در مدل‌های آمیخته متناهی تعمیم‌یافته نیم پارامتری

فرزاد اسکندری^{۱*}، 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 پذیرش ۱۳۹۳/۹/۳۰

تاریخ دریافت: ۱۳۹۲/۱۱/۲۲

چکیده: هدف این مقاله به دست آوردن بهترین متغیرهای کمکی تأثیرگذار بر متغیر پاسخ در مدل‌های نیم پارامتری تحت شرایطی است که تابع توانیده نیز در مسئله وجود دارد. لازم به ذکر است که ترکیبی از پارامترها به‌عنوان ضرایب متغیرهای موجود در هر مدل ارائه شده که برخی از آن‌ها به‌صورت خطی و بعضی دیگر به‌صورت تابع بر متغیر پاسخ تأثیرگذارند. لاجرم روش نیم پارامتری به‌عنوان یک‌راه حل بهینه در حل مسئله مدنظر قرار گرفته است. لذا در این مقاله مسئله انتخاب متغیر را در مدل‌های تعمیم‌یافته نیم پارامتری آمیخته متناهی مورد بررسی 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، درستی‌تایی توانیده، مدل نیم پارامتری، مدل آمیخته متناهی.

رده‌بندی ریاضی (۲۰۱۰): ۶۲J۱۲

۱-مقدمه

مدل سازی یکی از مهم ترین فن های آمار است و انتخاب متغیر نقشی اساسی در آن ایفا می کند. در مسائل کاربردی به منظور کاستن از شدت اریبی های ممکن مدل سازی، معمولاً تعداد زیادی از متغیرهای کمکی در مراحل اولیه مدل سازی معرفی می شوند. از سوی دیگر، به منظور افزایش قابلیت پیش بینی و انتخاب متغیرهای معنی دار، آماردانان اغلب از حذف گام به گام و انتخاب بهترین زیرمجموعه استفاده می کنند.

اگرچه کاربرد این روش ها در عمل مفید واقع می شوند، اما خطاهای تصادفی ذاتی را در مرحله انتخاب متغیرها نادیده می گیرند. بنابراین خواص نظری آن ها تقریباً غیرقابل درک است. علاوه بر این، انتخاب بهترین زیرمجموعه متغیرها چندین نقطه ضعف دارند، که مهم ترین آن ها فقدان پایداری است (به عنوان مثال به [۱] مراجعه کنید). همچنین می توان به زمان بر بودن و مشکلات محاسباتی این روش ها نیز اشاره کرد.

در تلاشی به منظور انتخاب خودکار و همزمان متغیرها، فن و لی [۲] روش یکپارچه ای را بر اساس کمترین توان های دوم تاوانیده پیشنهاد داده اند. «روش های کمترین توان های دوم تاوانیده که می توانند به صورت خودکار و همزمان متغیرهای معنی دار را انتخاب کنند، در سال های اخیر از اقبال گسترده ای برخوردار شده اند. [۳]»

توابع تاوان متنوعی در زمینه انتخاب متغیر معرفی شده اند. تاوان L_1 در $LASSO^1$ توسط تیبشیرانی [۴] معرفی شد و تبدیل به یکی از محبوب ترین تاوان ها 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 الگوریتم های جدیدی توسط آنان تولید شده است. زو [۹] $LASSO$ سازوار 4 (ALASSO) را با استفاده از وزن های سازوار برای ضرایب مختلف تاوان در $LASSO$ معرفی و خصوصیت اراکل آن را بررسی کرده است. منابع [۱۰ و ۱۱] انتخاب متغیر را در حالتی مطالعه کرده اند که بعد داده ها بزرگ تر از اندازه نمونه است. همچنین گستره ای از توابع تاوان در مدل های رگرسیون آمیخته متناهی توسط [۱۲] بررسی شده است.

از سوی دیگر، مدل های آمیخته متناهی ابزار انعطاف پذیری برای مدل بندی داده های حاصل از جوامع ناهمگون فراهم ساخته اند. این مدل ها، در زمینه های بسیاری از جمله زیست شناسی،

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 مورد استفاده قرار گرفته‌اند. کتاب [۱۳] مرور جامعی از مدل‌های آمیخته متناهی را در بر گرفته است. زمانی که یک متغیر تصادفی با یک توزیع آمیخته متناهی به متغیرهای کمکی خاصی وابسته است، یک مدل رگرسیون آمیخته متناهی^۱ (FMR) به دست 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 در زمینه مدل‌های رگرسیون آمیخته متناهی مورد مطالعه قرار گرفته‌اند. به هر حال حتی برای مدل‌های رگرسیون آمیخته متناهی با تعداد متوسط مؤلفه‌ها و متغیرهای کمکی، روش‌های انتخاب بهترین زیرمجموعه از نظر 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]، یک رویه جدید انتخاب متغیر در مدل‌های رگرسیون آمیخته متناهی را بر اساس این روش‌ها طراحی کرده‌اند.

به هر حال گاهی اوقات نمی‌توان یک ارتباط خطی را برای تمام متغیرها فرض کرد. یک راه مقابله با چنین وضعیتی استفاده از مدل‌های نیم‌پارامتری است که مزایای مدل‌سازی پارامتری و ناپارامتری را حفظ می‌کنند و در سال‌های اخیر مورد توجه ویژه‌ای قرار گرفته‌اند.

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

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

دو و همکاران [۲۱] آمیخته متناهی مدل‌های خطی با اثرات آمیخته^۱ 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 توانانیده را مدنظر قرار داده و جهت کارایی محاسبات از الگوریتم EM آشیانی^۲ استفاده کرده‌اند.

ارمز و اسکندری [۲۲] مسئله انتخاب متغیر را در مدل‌های نیم پارامتری آمیخته متناهی مطالعه کرده‌اند. توجه داشته باشید چارچوب در نظر گرفته شده 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 ارائه شده 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 $h(\cdot)$ مورد استفاده در این مقاله یک تابع معرفی شده بر اساس تعریف نلدرو و وادبرن [۲۳] است. در صورتی که در [۲۲] لزوماً چنین نیست. این مطلب اثر خود را در نتایج قضایای حدی اثبات شده نشان می‌دهد. زیرا در این مقاله می‌توانیم برآورد پارامتر $\theta_k(u, \mathbf{x}, \mathbf{z}) = h^{-1}\{\mathbf{x}'\alpha_k(u) + \mathbf{z}'\beta_k\}$ را به دست آوریم در صورتی که این موضوع لزوماً در [۲۲] امکان‌پذیر نیست.

بنابراین هدف این مقاله مطالعه مسئله انتخاب متغیر در مدل‌های تعمیم‌یافته نیم پارامتری آمیخته متناهی به جای مدل‌های نیم پارامتری آمیخته متناهی است. لذا در این مقاله علاوه بر معرفی مدلی جدید و در نظر گرفتن مدل‌های تعمیم‌یافته نیم پارامتری به جای مدل‌های نیم پارامتری، قادر خواهیم بود مسئله شناسایی پذیری مدل‌های آمیخته را به صورت دقیق‌تری بررسی کنیم. بدین منظور ادامه مقاله به شکل زیر تنظیم شده 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 توانانیده و در بخش ۴-۲ الگوریتم EM و حل عددی برآوردگرها را مورد بررسی قرار خواهیم داد. بخش ۴-۳ به انتخاب پارامترهای منظم سازی

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

۲- مدل‌های آمیخته متناهی تعمیم‌یافته نیم‌پارامتری

فرض کنید Y متغیر پاسخ موردنظر و $(u, \mathbf{x}, \mathbf{z})$ بردار متغیرهای کمکی دارای اثر احتمالی بر Y باشد که $\mathbf{z} = (z_1, z_2, \dots, z_p)^t$ ، $\mathbf{x} = (x_1, x_2, \dots, x_d)^t$ و u یک متغیر تکی است. مدل آمیخته متناهی تعمیم‌یافته نیم‌پارامتری 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 به اندازه σ متناهی ν باشد که $\Theta \subset R$ و ϕ پارامتر پراکندگی است. گوییم $(\mathbf{x}, \mathbf{z}, u, Y)$ از مدل آمیخته متناهی تعمیم‌یافته نیم‌پارامتری از مرتبه K تبعیت می‌کند هرگاه تابع چگالی Y به شرط $(\mathbf{x}, \mathbf{z}, u)$ دارای فرم

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

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

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

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

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

همچنین فرض کنید $\pi_k > 0$ و $\sum_{k=1}^K \pi_k = 1$. 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 $A(\cdot)$ ، $c(\cdot)$ و $d(\cdot)$ توابع مشخصی هستند.

اگرچه می‌توانیم احتمال‌های آمیخته‌های آمیخته را وابسته به متغیرهای کمکی در نظر گرفته و آن‌ها را برآورد کنیم، با این وجود به جهت سادگی ارائه مطلب از این پس 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 به شرط 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 گوئیم 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، یک به یک باشد. یعنی برای هر مدل در فضای مدل، یک بردار پارامتر یکتا در فضای پارامتر وجود داشته باشد که به مدل نگاشته شده باشد.

تعریف ۲: یک مدل آمیخته متناهی تعمیم یافته نیم پارامتری با تابع چگالی شرطی (۱) را در نظر بگیرید. به ازای یک ماتریس طرح معلوم، مدل آمیخته متناهی تعمیم یافته نیم پارامتری را شناسایی پذیر گوئیم هرگاه به ازای هر دو پارامتر Ψ و Ψ^* 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 $i=1, \dots, n$ و تمامی مقادیر ممکن y به دست دهد $K=K^*$ و $\Psi = \Psi^*$. شناسایی پذیری بدان معناست که هیچ دو مجموعه‌ای از مقادیر متفاوت پارامترها، تابع چگالی یکسانی ندارند. به‌هرحال در تعریف فوق $\Psi = \Psi^*$ 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 می‌کند.

$$\pi_k > 0 \quad \forall k = 1, \dots, K \quad (۲)$$

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

دو شرط (۲) و (۳) از بیش برزندگی و مشکلات شناسایی پذیری که به واسطه مؤلفه‌های تهی که در آن‌ها θ_k نمی‌تواند به صورت یکتا تعیین شود و همچنین به واسطه مؤلفه‌های دارای بردار پارامترهای برابر که مقادیر متفاوت π_k برای آن‌ها امکان پذیر است، جلوگیری می‌کنند.

مسائل شناسایی پذیری عام^۱ برای آمیخته‌های متناهی توزیع‌ها توسط [۲۶] مطالعه شده‌اند. شناسایی پذیری عام برای توزیع‌های پیوسته مهم مانند گاما و نرمال و همچنین توزیع پواسون اثبات شده 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 لیسچ [۲۷] قضیه‌ای پیرامون شرایط کافی شناسایی پذیری مدل‌های آمیخته متناهی خطی تعمیم یافته 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 قابل استفاده است. لذا از بیان مجدد قضیه خودداری کرده و خوانندگان علاقه‌مند را به مرجع [۲۷] ارجاع می‌دهیم.

توجه داشته باشید که برای شرط (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 مورد بررسی قرار دهیم تا تعیین شود که آیا می‌توان آن را در \bar{T} قرار داد یا خیر. شرط (ب-۱) 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 مشخص شده در تعریف (۱) صورت پذیرد، در حالتی که $(u, \mathbf{x}, \mathbf{z})$ یک بردار تصادفی است، فرض می‌کنیم که چگالی آن به صورت تابعی مستقل از پارامترهای مدل است. فرض کنید $(u_1, \mathbf{x}_1, \mathbf{z}_1, y_1), \dots, (u_n, \mathbf{x}_n, \mathbf{z}_n, y_n)$ نمونه‌ای از مشاهدات از مدل (۱) باشند. با در نظر

گرفتن فرم متعارف خانواده نمایی یک پارامتری، تابع لگاریتم درستنمایی شرطی Ψ عبارت خواهد بود از

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

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 کمی 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

$$\tilde{l}_n(\Psi) = l_n(\Psi) - p_n(\Psi) \quad (5)$$

تعریف می‌شود، که در آن

$$p_n(\Psi) = \sum_{k=1}^K \pi_k \left\{ \sum_{j=1}^P p_{nk}(\beta_{kj}) \right\} \quad (6)$$

و $p_{nk}(\beta_{kj})$ توابعی غیر منفی و غیر نزولی برحسب $|\beta_{kj}|$ هستند. با ماکسیمم کردن $\tilde{l}_n(\Psi)$ 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 k ام با π_k در (۶) در ارتباط با حجم نمونه در نظر می‌گیریم.

به‌طور کلی بایستی تابع تاوان مناسبی را که متناسب با نیازهای کاربردی مسئله و مطالب نظری باشد، انتخاب کنیم. باین‌وجود سه تابع تاوان زیر به‌دفعات در ادبیات انتخاب متغیر مورد استفاده قرار گرفته‌اند و در اینجا نیز آن‌ها را مطالعه می‌کنیم.

$$1- \text{تابع تاوان نرم } L_1: p_{nk}(\beta) = \gamma_{nk} \sqrt{n} |\beta|$$

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

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

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

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 بر داده‌ها انتخاب شوند. انتخاب پارامترهای منظم‌سازی^۱ اغلب توسط معیارهایی مانند اعتبارسنجی متقابل^۲ یا اعتبارسنجی متقابل تعمیم‌یافته^۳ صورت می‌پذیرد.

با پیروی از [۱۲ و ۲۸]، توابع تاوان $p_n(\cdot)$ در رابطه (۶) را که از HARD, LASSO و SCAD تولید شده‌اند، به ترتیب تاوان‌های MIXLASSO, MIXHARD و IXSCAD می‌نامیم. ماکسیمم کردن درست‌نمایی تاوانیده، معادل ماکسیمم کردن مقید است. تابع تاوان LASSO محذب 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 β را به دست آورد. با انتخاب تابع تاوان خاص، برآورد حاصل β شامل برخی ضرایب صفر دقیق خواهد بود. این مطلب متناظر با حذف متغیرهای متناظر از مدل نهایی است و در نتیجه به هدف انتخاب متغیر دست‌یافته‌ایم.

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

$$\alpha_{kj}(v) \approx \alpha_{kj}(u) + \alpha'_{kj}(u)(v-u) \equiv a_{kj} + b_{kj}(v-u). \quad (7)$$

تقریب می‌زنیم. حال تابع لگاریتم درست‌نمایی

$$\sum_{i=1}^n \log \left\{ \sum_{k=1}^K \pi_k \exp \left[\frac{\tilde{\theta}_k(u_i, \mathbf{x}_i, \mathbf{z}_i) T(y_i) - A(\tilde{\theta}_k(u_i, \mathbf{x}_i, \mathbf{z}_i))}{c(\phi_k)} \right] d(y_i, \phi_k) K_h(u_i - u) \right\} \quad (8)$$

1- Tuning Parameter

2- Cross Validation

3- Generalized Cross Validation

را ماکسیمم خواهیم کرد که $K(\cdot)$ 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 $h(\cdot)$ جلب کنیم. در این مقاله نماد نخست اسکالر بوده و نمایانگر پهنای باند است در حالی که $h(\cdot)$ تابع بوده و به عنوان تابع ربط (تعریف ۱) استفاده شده است. همچنین

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

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

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

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

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n \log \left\{ \sum_{k=1}^K \pi_k \exp \left[\frac{\theta_k^*(u_i, \mathbf{x}_i, \mathbf{z}_i) T(y_i) - A(\theta_k^*(u_i, \mathbf{x}_i, \mathbf{z}_i))}{c(\phi_k)} \right] d(y_i, \phi_k) \right\} \quad (9)$$

$$- p_n(\boldsymbol{\Psi})$$

برآورد می‌کنیم که در آن $\theta_k^*(u_i, \mathbf{x}_i, \mathbf{z}_i)$ از جایگزین کردن $\tilde{\mathbf{a}}(u)$ 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 ماکسیمم درست‌نمایی توانیده $\boldsymbol{\beta}$ 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 تنگ بودن (P_0) (به‌ضمیمه یک مراجعه نمایید). $p_{nk}(\boldsymbol{\beta})$ ها در $\boldsymbol{\beta} = \mathbf{0}$ مشتق‌پذیر نیستند. بنابراین با پیروی از [۲] جهت استفاده از الگوریتم نیوتن-رافسون در مرحله M الگوریتم EM، $p_{nk}(\boldsymbol{\beta})$ را با تقریب موضعی مرتبه دوم آن در یک همسایگی $\boldsymbol{\beta}_0$ یعنی

$$p_{nk}(\boldsymbol{\beta}) \approx p_{nk}(\boldsymbol{\beta}_0) + \frac{p'_{nk}(\boldsymbol{\beta})}{2\boldsymbol{\beta}_0} (\boldsymbol{\beta}^2 - \boldsymbol{\beta}_0^2)^2$$

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

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

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

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

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

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

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

که

$$w_{ik}^{(m)} = \frac{\pi_k^{(m)} \exp \left(\frac{\theta_k^{*(m)}(u_i, \mathbf{x}_i, \mathbf{z}_i)T(y_i) - A(\theta_k^{*(m)}(u_i, \mathbf{x}_i, \mathbf{z}_i))}{c(\phi_k)} \right)}{\sum_{l=1}^K \pi_l^{(m)} \exp \left(\frac{\theta_l^{*(m)}(u_i, \mathbf{x}_i, \mathbf{z}_i)T(y_i) - A(\theta_l^{*(m)}(u_i, \mathbf{x}_i, \mathbf{z}_i))}{c(\phi_l)} \right)} \quad (10)$$

امید ریاضی‌های شرطی v_{ik} هستند.

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

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

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

$$\sum_{i=1}^n w_{ik}^{(m)} z_{ij} \frac{T(y_i)}{c(\phi_k)} = \sum_{i=1}^n w_{ik}^{(m)} \frac{z_{ij}}{c(\phi_k)} \frac{\partial}{\partial \theta_k^{*(m)}} A(\theta_k^{*(m)}(u_i, \mathbf{x}_i, \mathbf{z}_i)) + \pi_k \frac{\partial}{\partial \beta_{kj}} \{ \tilde{p}_{nk}(\beta_{kj}) \} \quad (11)$$

به روزرسانی می‌شوند که در آن جمله متناظر در $\tilde{p}_n(\Psi, \Psi^{(m)})$ است. این برآوردگر را β می‌نامیم. برآوردهای ضرایب ناپارامتری نیز به ازای $k=1, \dots, K$ 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)، β را با برآورد آن یعنی $\hat{\beta}$ 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 $\hat{\mathbf{a}}(u) = \hat{\mathbf{a}}$. لی و لیانگ [۲۸] ثابت کرده‌اند که این برآوردگر کارتر از $\tilde{\mathbf{a}}(u)$ است. بنابراین الگوریتم برآورد به شکل زیر است:

گام ۱: به دست آوردن $\tilde{\mathbf{a}}(u)$ از طریق ماکسیمم کردن رابطه (۸) نسبت به \mathbf{a} و \mathbf{b} .

گام ۲: به دست آوردن β با استفاده از الگوریتم EM و رابطه (۱۱).

گام ۳: جایگذاری β در رابطه (۸) و به دست آوردن $\hat{\mathbf{a}}(u)$.

پس از به دست آوردن برآورد ناپارامتری $\mathbf{a}(\cdot)$ 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

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

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

۳-۴- 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 طراحی شده است و در تنظیمات ما نیز می‌تواند مورداستفاده قرار گیرد.

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

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

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

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

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

$$\Sigma_k(\hat{\beta}_k) = \hat{\pi}_k \text{diag} \left\{ p'_{nk}(\hat{\beta}_{k1}) / \hat{\beta}_{k1}, \dots, p'_{nk}(\hat{\beta}_{kp}) / \hat{\beta}_{kp} \right\}.$$

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

۵- خواص مجانبی

بردار ضرایب رگرسیون β_k در مؤلفه k ام را به $\beta_k^t = \{\beta_{1k}^t, \beta_{2k}^t\}$ به گونه‌ای تجزیه کنید که β_{2k}^t عناصر صفر را در برگیرد. به‌طور کلی ممکن است بردار عناصر غیر صفر β_{1k}^t به k وابسته باشد. بدون از دست دادن کلیت و به‌منظور سادگی نمادها از این واقعیت چشم‌پوشی می‌کنیم. به‌طور طبیعی پارامترهای $\Psi^t = \{\Psi_1^t, \Psi_2^t\}$ را به گونه‌ای تقسیم می‌کنیم که Ψ_2^t شامل تمامی اثرات صفر یعنی $\beta_{2k}^t: k=1, \dots, K$ شود. بردار پارامترهای حقیقی را نیز با Ψ_0 نشان خواهیم داد. عناصر Ψ_0 را با یک زیرنویس مانند β_{kj}^0 نشان خواهیم داد. برای ارائه نتایج مجانبی به نمادهای زیر نیاز خواهیم داشت:

$$a_n = \max_{k,j} \{p_{nk}(\beta_{kj}^0) / \sqrt{n} : \beta_{kj}^0 \neq 0\}$$

$$b_n = \max_{k,j} \{|p'_{nk}(\beta_{kj}^0)| / \sqrt{n} : \beta_{kj}^0 \neq 0\}$$

$$c_n = \max_{k,j} \{|p''_{nk}(\beta_{kj}^0)| / n : \beta_{kj}^0 \neq 0\}$$

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 $p_{nk}(\cdot)$ هستند که در ضمیمه یک ارائه شدند. علاوه بر این، به‌منظور توسعه نظریه مجانبی به برخی شرایط نظم معمول بر روی تابع چگالی توأم $f(\mathbf{w}; \Psi)$ متعلق به $\mathbf{W} = (\mathbf{x}, \mathbf{z}, u, Y)$ نیاز داریم که در ضمیمه یک آورده شده‌اند.

قضیه ۲: فرض کنید $\mathbf{W}_i = (\mathbf{x}_i, \mathbf{z}_i, u_i, Y_i)$, $i=1, 2, \dots, n$ نمونه‌ای تصادفی از تابع چگالی $f(\mathbf{w}; \Psi)$ باشد که در شرایط نظم صدق می‌کند. همچنین فرض کنید توابع تاوان $p_{nk}(\cdot)$ در شرایط P_0 و P_1 صدق کنند. آنگاه یک ماکسیمم کننده موضعی $\hat{\Psi}_n$ برای تابع لگاریتم درست‌نمایی توانیده $\bar{l}_n(\Psi)$ 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 $\hat{\Psi}$ 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.

قضیه ۳: شرایط اشاره شده در قضیه (۲) را در نظر بگیرید، همچنین فرض کنید که توابع تاوان $p_{nk}(\cdot)$ در شرایط $P_0 - P_2$ صدق می‌کنند و مقدار K در بخش‌های (الف) و (ب) زیر معلوم است. داریم:

(الف) به ازای هر Ψ که $\|\Psi - \Psi_0\| = O(n^{-1/2})$ با احتمال متمایل به یک داریم

$$\tilde{l}_n\{(\Psi_1, \Psi_2)\} - \tilde{l}_n\{(\Psi_1, 0)\} < 0$$

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

- i تنک بودن. هنگامی که $n \rightarrow \infty$ داریم $P\{\beta_{2k} = 0\} \rightarrow 1, k=1, \dots, K$
 ii نرمال مجانبی بودن.

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

که $I_1(\Psi_{01})$ اطلاع فیشر محاسبه شده تحت مدل کاهش یافته، یعنی پس از حذف تمام اثرات صفر است.

(ج) چنانچه K به صورت 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 $p_n(\cdot)$ در (ب-ii) به ازای برخی انتخاب‌های تابع تاوان قابل چشم‌پوشی می‌شوند و تصحیح‌های حجم نمونه متناهی را به همراه خواهند داشت. این مطلب برآوردگر واریانس $\hat{\Psi}_1$ به فرم

$$\widehat{\text{var}}(\hat{\Psi}_1) = \{l_n''(\hat{\Psi}_1) - p''(\Psi)\}^{-1} \widehat{\text{var}}\{l_n'(\hat{\Psi}_1)\} \{l_n''(\hat{\Psi}_1) - p''(\Psi)\}^{-1} \quad (12)$$

را پیشنهاد می‌کند.

۶- مطالعه شبیه‌سازی

در این بخش روش ارائه شده را بر روی دو توزیع یک پارامتری خانواده نمایی اعمال می‌کنیم. نتایج شبیه‌سازی به صورت زیر است.

مثال ۱: 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 α و β را نیز متناظر مقادیر زیر در نظر خواهیم گرفت.

$$\alpha_1 = \exp(2u-1), \alpha_2 = 8u(1-u), \beta_1 = (2,0,1,0), \beta_2 = (0,2,0,0)$$

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

$$\log(\lambda_k) = \mathbf{x}'\alpha_k(u) + \mathbf{z}'\beta_k.$$

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 شبیه‌سازی شده و روش پیشنهادی روی آن اعمال شده است. این عمل ۲۰۰۰ بار تکرار شده و نتایج زیر به دست آمده‌اند.

به منظور ارزیابی عملکرد رویه انتخاب متغیر پیشنهادی از شاخص RGMSE استفاده خواهیم کرد که به عنوان نسبت GMSE مدل انتخاب شده نهایی به GMSE مدل کامل تعریف می‌شود:

$$GMSE(\beta) = (\beta - \hat{\beta})E(ZZ')(\beta - \hat{\beta}).$$

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

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

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

میانگین و انحراف استاندارد ۲۰۰۰ تکرار برآوردها و خطاهای استاندارد ارائه شده در جدول ۲ را به دست می‌دهد. به عنوان 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.

جدول ۲: برآوردها و انحراف استاندارد آن‌ها در آمیخته متناهی پواسون براساس ۲۰۰۰ تکرار. انحرافات استاندارد داخل پرانتز داده شده‌اند.

β			β
<i>MIXSCAD</i>	<i>MIXHARD</i>	<i>MIXLASSO</i>	
1/9957 _(0/1339)	2/0381 _(0/1360)	2/0803 _(0/1366)	۲
0/3484 _(0/1113)	0/3478 _(0/1136)	0/3469 _(0/1158)	۰
0/9964 _(0/1063)	1/0166 _(0/1089)	1/0375 _(0/1112)	۱
0/0613 _(0/0507)	0/0629 _(0/0524)	0/0643 _(0/0535)	۰
0/3137 _(0/0695)	0/3103 _(0/0692)	0/3069 _(0/0700)	۰
2/0180 _(0/0972)	2/0117 _(0/0954)	2/0275 _(0/0955)	۲
0/1574 _(0/0600)	0/1551 _(0/0596)	0/1534 _(0/0595)	۰
0/0402 _(0/0328)	0/0401 _(0/0330)	0/0400 _(0/0333)	۰

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 گرفته شده 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. به‌رحال 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					
<i>MIXLASSO</i>		<i>MIXHARD</i>		<i>MIXSCAD</i>	
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	۰/۰۶۷۵	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 محاسبه شده بر اساس توزیع توأم بردار پارامترها (و نه به‌صورت تکی) محاسبه شده‌اند.

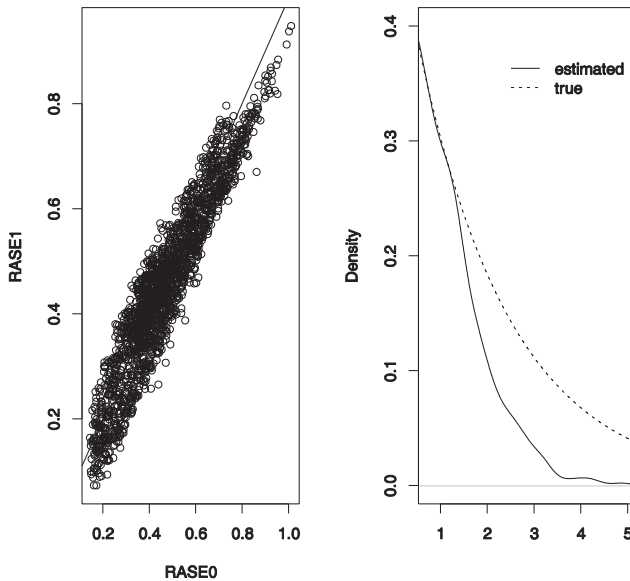
به‌منظور ارزیابی دقیق‌تر عملکرد الگوریتم برآورد پیشنهادی برای مؤلفه ناپارامتری از جذر متوسط توابع دوم خطا^۱ (*RASE*) استفاده خواهیم کرد به شکل زیر تعریف می‌شود.

1- Square Root of Average Square Errors

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

و در آن $\{\mathbf{z}_k, k=1, \dots, n_{grid}\}$ نقاط شبکه‌ای هستند که توابع $\{\hat{\alpha}_j(\cdot)\}$ 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 β 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.

$$H_0: \alpha_1(u) = 0 \text{ در برابر } H_1: \alpha_2(u) \neq 0.$$



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 کای اسکور 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 داده شده است. در این شکل خط پررنگ، تابع چگالی برآورد شده و خط نقطه چین توزیع کای اسکور را با دو درجه آزادی (که تعداد مؤلفه‌های مشخص شده تحت فرض صفر است) نشان می‌دهد. مشاهده می‌شود که توزیع صفر نمونه متناهی کاملاً نزدیک توزیع کای اسکور است. این مطلب توسط آزمون کولموگروف - اسمیرنوف ($D = 0/5575$, $p - value = 0/8856$) نیز اثبات می‌شود.

جهت مشاهده سایر خواص آزمون نسبت درستمایی تعمیم یافته پیشنهادی می‌توانید به [۲۸] مراجعه نمایید.

مثال ۲: شبیه‌سازی دوم بر اساس آمیخته متناهی مدل نیم پارامتری تعمیم یافته 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 α و β را نیز متناظر مقادیر زیر در نظر خواهیم گرفت.

$$\alpha_1 = \exp(2u - 1), \alpha_2 = 2(\sin(2\pi u))^2, \beta_1 = (2, 0, 1, 0), \beta_2 = (0, 2, 0, 0)$$

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

$$\text{logit}(p_k) = \mathbf{x}'\alpha_k(u) + \mathbf{z}'\beta_k.$$

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 $B(100, 0/500)$ مواجه هستیم. ۲۰۰۰ مشاهده از این مدل آمیخته شبیه‌سازی شده و روش پیشنهادی روی آن اعمال شده است. این عمل ۲۰۰۰ بار تکرار شده و نتایج زیر به دست آمده‌اند.

شاخص RGMSE و تعداد متوسط ضرایب صفر در جدول ۴ داده شده‌اند. 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. به علاوه مشاهده می‌کنیم که دقت توابع تاوان MIXLASSO و MIXHARD 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 برآورد شده کمتر از ۰/۰۰۱ باشد. میانگین و انحراف استاندارد ۲۰۰۰ تکرار برآوردها و خطاهای استاندارد ارائه شده در جدول ۵ را به دست می‌دهد. به عنوان مثال برآورد $\hat{\beta}_1$ تحت توابع تاوان MIXLASSO، MIXHARD و MIXSCA به ترتیب برابر ۲/۰۲۰۴، ۲/۰۲۱ و ۱/۹۸۵ است. همچنین انحراف استاندارد این برآوردها نیز به ترتیب ۰/۲۶۸۰، ۰/۲۶۸۲ و ۰/۲۶۴۰ است. این در حالی است که مقدار واقعی این پارامتر برابر ۲ 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 داده شده‌اند.

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

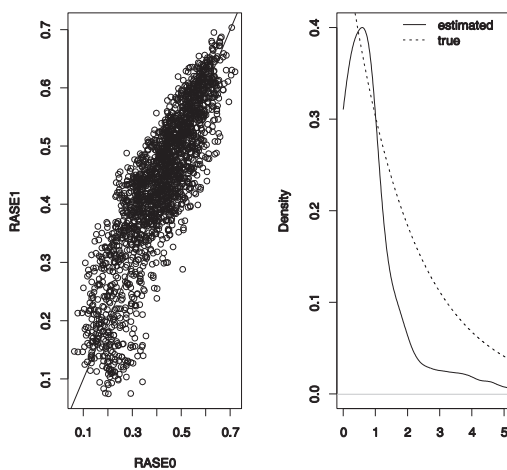
جدول ۶ فواصل اطمینان همزمان ۹۵٪ پارامترهای برآورده شده را بر اساس T^2 هتلینگ نشان می‌دهد. بر اساس این جدول فاصله اطمینان $\hat{\beta}_1$ تحت تابع تاوان MIXLASSO برابر (1/9967, 2/0440) است. شایان‌ذکر است اعداد محاسبه شده بر اساس توزیع توأم بردار پارامترها (و نه به صورت تکی) محاسبه شده‌اند.

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

مطلب توسط آزمون کلموگروف اسمیرنوف ($D=0/5065, p\text{-value}=0/9875$) 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] Candès, 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 $p_{nk}(\cdot)$ 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 استثناء، 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 $N_n = \{\beta; 0 < \beta \leq n^{-\frac{1}{2}} \log n\}$ داریم $\liminf_{n \rightarrow \infty} \inf_{\beta \in N_n} \frac{p'_{nk}(\beta)}{\sqrt{n}} = \infty$

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 مورد استفاده قرار می‌گیرد.

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

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

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

A_3 : به ازای هر $\Psi_0 \in \Omega$ توابع $M_1(\mathbf{w})$ و $M_2(\mathbf{w})$ (احتمال وابسته به Ψ_0) وجود دارند، به‌طوری‌که به ازای Ψ در یک همسایگی $N(\Psi_0)$ داریم

$$\left| \frac{\partial f(\mathbf{w}; \Psi)}{\partial \psi_j} \right| \leq M_1(\mathbf{w}) \quad \left| \frac{\partial^2 f(\mathbf{w}; \Psi)}{\partial \psi_j \partial \psi_l} \right| \leq M_1(\mathbf{w}) \quad \left| \frac{\partial^3 \log f(\mathbf{w}; \Psi)}{\partial \psi_j \partial \psi_l \partial \psi_m} \right| \leq M_2(\mathbf{w})$$

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

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

به ازای هر $\Psi \in \Omega$ 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