
بسم الله
الرحمن الرحیم

دانشکده علوم ریاضی

رساله دکتری آمار

انتخاب متغیر و تشخیص ساختار در مدل‌های نیمه‌پارامتری

نگارنده: محمد کاظمی

اساتید راهنما

دکتر داود شاهسونی
دکتر محمد آرشی

بهمن ۱۳۹۷

شماره:
تاریخ:
ویرایش:

باسمه تعالی

مدیریت تحصیلات تکمیلی

فرم شماره ۱۲: صورت جلسه نهایی دفاع از رساله دکتری (Ph.D)

(ویژه دانشجویان ورودی های ۹۴ و ما قبل)

بدینوسیله گواهی می شود آقای محمد کاظمی دانشجوی دکتری رشته آمار- استنباط به شماره دانشجویی ۹۳۰۱۱۸۵ ورودی مهر ماه سال ۱۳۹۳ در تاریخ ۹۷/۱۱/۲۸ از رساله **نظری** ■ **عملی** □ خود با عنوان: **انتخاب متغیر و تشخیص ساختار در مدل‌های نیمه‌پارامتری** دفاع و با اخذ نمره ۱۹/۸۶ به درجه : عالی نائل گردید.

الف) درجه عالی: نمره ۲۰-۱۹ □ ب) درجه بسیار خوب: نمره ۱۸/۹۹ – ۱۷ □

ج) درجه خوب: نمره۱۶/۹۹– ۱۵ □ د) غیر قابل قبول و نیاز به دفاع مجدد دارد □

ه) رساله نیاز به اصلاحات دارد □

ردیف	هیئت داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱	دکتر داود شاهسونی	استاد راهنمای اول	دانشیار	
۲	دکتر محمد آرشی	استاد راهنمای دوم	دانشیار	
۳	دکتر مهدی روزبه	استاد مدعو خارجی	دانشیار	
۴	دکتر حسین باغیشنی	استاد مدعو داخلی	استادیار	
۵	دکتر محمدرضا ربیعی	استاد مدعو داخلی	استادیار	
۶	دکتر ابراهیم هاشمی	سرپرست (نماینده) تحصیلات تکمیلی دانشکده	استاد	

مدیر محترم تحصیلات تکمیلی دانشگاه:

ضمن تأیید مراتب فوق مقرر فرمائید اقدامات لازم در خصوص انجام مراحل دانش آموختگی آقای محمد کاظمی بعمل آید.

نام و نام خانوادگی رئیس دانشکده: دکتر ابراهیم هاشمی

تاریخ و امضاء و مهر دانشکده: ۹۷/۱۱/۲۸

وزارت علوم، تحقیقات و فناوری
دانشکده ریاضی

تقدیم به:

پدرم، که همواره حامی و پشتیبان من بوده است

مادرم، که هیچ واژه‌ای نمی‌تواند مهر او را معنی کند

همسرم، که مسیح وار با صبرش در تمامی لحظات رفیق راه بود

دختر نازم، امید بخش جانم که آسایش او آرامش من است.

سپاس‌گزاری

سپاس ایزد منان را که در تمام مراحل زندگی مرا قوت قلب بود و توفیق را رفیق راهم ساخت تا این رساله را به پایان برسانم. اکنون که با یاری او توانسته‌ام قدمی هرچند ناچیز در راه کسب علم و دانش بردارم، برخود لازم می‌دانم از پدر و مادر عزیزم، که همواره بر کوتاهی و درشتی من، قلم عفو کشیده و با عاطفه سرشار و گرمای امیدبخش وجودشان، در این سردترین روزگاران، بهترین پشتیبان من بوده‌اند، تشکر و قدردانی نمایم. از همسر عزیزم که همواره رفیق راه و مشوق اصلی من در امر تحصیل و تحقیق بوده و موفقیت خود را از آغاز این دوره تحصیلی تاکنون مرهون حمایت‌ها و صبر و شکیبایی بی‌دریغش هستم، بی‌نهایت سپاس‌گزارم.

تقدیر به گونه‌ای رقم خورد که مقطع دکتری را در دانشگاه صنعتی شاهرود به پایان برسانم و از محضر اساتید خوب گروه آمار این دانشگاه بهره‌جویم. سپاس از استاد عزیزم جناب آقای دکتر داود شاهسونی، استادی دلسوز و با اخلاق، که درس ''فنون آماری'' را با سبکی جذاب و عشق و علاقه‌ای وصف‌ناپذیر تدریس می‌کرد و بارها از وی درس زندگی آموخته‌ام. بی‌شک کلاس درس ایشان از بهترین خاطرات من در این دوره تحصیلی است. کاش می‌شد دوباره دانشجوی ترم اول کارشناسی آمار بودم و آمار را جور دیگری یاد می‌گرفتم، اما افسوس نه راهی برای بازگشت مانده و نه پایی برای رفتن! سپاس از استاد خوبم جناب آقای دکتر محمد آرشی، که هنوز اولین جلسه کلاس درس ''استنباط آماری پیشرفته'' ایشان در خاطرم هست. استادی جوان، با اخلاق و دانا که آن روز آرزو کردم روزی مثل او باشم و هنوز هم از تلاش برای تحقق این آرزویم دست نکشیده‌ام. انگار بخت با من یار بود که این دو استاد بزرگوار زحمت راهنمایی مرا در این دوره تحصیلی برعهده گیرند. بدون شک جایگاه و منزلت معلم، والاتر از آن است که در مقام قدردانی از زحمات بی شائبه او، با زبان قاصر و دست ناتوان، چیزی بنگارم، اما بر حسب وظیفه و از باب ''مَنْ لَمْ یَشْکُرِ الْمُنْعِمَ مِنَ الْمَخْلُوقِینَ لَمْ یَشْکُرِ اللَّهَ عَزَّ وَ جَلَّ'' از اساتید ارجنمدم جناب آقای دکتر داود شاهسونی و جناب آقای دکتر محمد آرشی که با صبر و تلاش فراوان، از هیچ کمکی در این عرصه بر من دریغ ننمودند و زحمت راهنمایی بنده را در تهیه و تنظیم این رساله بر عهده گرفتند، کمال تشکر و قدردانی را دارم و از درگاه خداوند متعال برای آنها توفیق روزافزون آرزومندم.

از اساتید محترم، جناب آقای دکتر مهدی روزبه (دانشگاه سمنان)، جناب آقای دکتر حسین باغیشنی (دانشگاه صنعتی شاهرود) و جناب آقای دکتر محمدرضا ربیعی (دانشگاه صنعتی شاهرود)، که زحمت تصحیح و داوری این رساله را متقبل شدند، سپاس‌گزارم.

از استاد و دوست خوبم دکتر پائولو کنس رُدری‌گوئز[1]، که در طول دوره فرصت مطالعاتی من در دانشگاه تامپره فنلاند زحمت راهنمایی و میزبانی مرا برعهده گرفت و از هیچ کمکی به من دریغ نکرد، سپاس‌گزارم.

[1] Paulo Canas Rodrigues

از مدیر دلسوز و پرتلاش گروه آمار دانشگاه صنعتی شاهرود، جناب آقای دکتر محمدرضا ربیعی به دلیل خدمات، راهنمایی‌ها و زحمات‌شان سپاس‌گزارم. از دوستان و همکلاسی‌های خوبم که مرا در رسیدن به اهدافم کمک کرده‌اند، تشکر کرده و برایشان آرزوی موفقیت و سربلندی می‌کنم.

در نهایت، از همه عزیزانی که تا بدین‌جا مرا یاری کرده‌اند، به‌ویژه خانواده مهربانم که همواره پشتیبان من بوده‌اند، سپاس‌گزارم. امیدوارم بتوانم از عهده ادای حق این عزیزان برآیم.

یقیناً این مجموعه بدون عیب و نقص نیست. لذا از خوانندگان عزیز تقاضا دارم عیب‌ها و کاستی‌های این رساله را از طریق پست الکترونیکی[۲] به اطلاع اینجانب برسانند.

محمد کاظمی
بهمن ۱۳۹۷

[2] m.kazemie64@yahoo.com

تعهد نامه

اینجانب **محمد کاظمی** دانشجوی دکتری رشته **آمار** دانشکده **علوم ریاضی** دانشگاه‌صنعتی شاهرود، نویسنده رساله با عنوان **انتخاب متغیر و تشخیص ساختار در مدل‌های نیمه‌پارامتری** ، تحت راهنمایی **دکتر داود شاهسونی** و **دکتر محمد آرشی** متعهد می‌شوم:

- تحقیقات در این رساله توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این رساله، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ’’ دانشگاه صنعتی شاهرود ‘‘ یا ’’ Shahrood University of Technology ‘‘ به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی رساله تاثیرگذار بوده‌اند، در مقالات مستخرج از رساله رعایت می‌گردد.
- در تمام مراحل انجام این رساله، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این رساله، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

محمد کاظمی
بهمن ۱۳۹۷

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این رساله بدون ذکر منبع مجاز نمی‌باشد.

چکیده

مدل‌های نیمه‌پارامتری، به دلیل برخورداری از انعطاف‌پذیری مناسب و تفسیرپذیری، اغلب در تحلیل داده‌های واقعی مورد استفاده قرار می‌گیرند. دو مورد از رایج‌ترین مدل‌های نیمه‌پارامتری، مدل جمعی خطی–جزئی و مدل ضریب متغیر خطی–جزئی می‌باشند. استنباط آماری این مدل‌ها بر این شرط استوار است که اجزای پارامتری و ناپارامتری مدل از پیش تعیین‌شده باشند. اما در عمل چنین اطلاعات پیشین به ندرت در دسترس است، به ویژه هنگامی که تعداد متغیرهای توضیحی زیاد است. بنابراین لازم است روشی ارائه شود که بتواند مؤلفه‌های پارامتری و ناپارامتری را به درستی تشخیص دهد. در این رساله، به معرفی یک روش دو مرحله‌ای برای انتخاب متغیرهای مهم و تشخیص اثرات خطی و غیرخطی در مدل جمعی خطی–جزئی با بعد بسیار بالا پرداخته می‌شود. علاوه بر این، با استفاده از رگرسیون نما، یک روش نیرومند برای انتخاب متغیرهای مهم و تشخیص اثرات ثابت و متغیر در مدل ضریب متغیر خطی–جزئی ارائه می‌شود که نسبت به داده‌های پرت در متغیر پاسخ و توزیع غیر نرمال خطاها نیرومند است. کارایی روش های ارائه‌شده با مطالعه شبیه‌سازی و تحلیل داده‌های واقعی بررسی شده‌است. نتایج به‌دست آمده عملکرد مطلوب روش‌های مذکور را در مقایسه با برخی روش‌های موجود نشان می‌دهند.

کلمات کلیدی: انتخاب متغیر، تشخیص ساختار، رگرسیون نما، غربالگری مستقل مطمئن، کاهش بعد، مدل جمعی خطی–جزئی، مدل ضریب متغیر خطی–جزئی.

لیست مقالات مستخرج از رساله

۱. کاظمی، م.، شاهسونی، د. و آرشی، م. (۱۳۹۷). انتخاب متغیر و تشخیص ساختار در بعد بالا برای مدل‌های جمعی خطی–جزئی، *مجله علوم آماری*، ۱۲(۲)، ۴۸۵–۵۱۲.

۲. کاظمی، م.، شاهسونی، د. و آرشی، م. (۱۳۹۶). غربالگری متغیرها در داده‌های بزرگ با استفاده از همبستگی فاصله‌ای، کنفرانس روش‌های مدرن در قیمت‌گذاری‌های بیمه‌ای و آمارهای صنعتی، دانشگاه بوعلی‌سینا، همدان

۳. کاظمی، م.، شاهسونی، د. و آرشی، م. (۱۳۹۶). حل مسئله بهینه‌سازی برای انتخاب متغیرهای مهم در داده‌های با بعد بالا، دهمین کنفرانس بین‌المللی انجمن ایرانی تحقیق در عملیات، دانشگاه مازندران، بابلسر

4. Kazemi, M., Shahsavani, D. and Arashi, M. (2019). A sure independence screening procedure for ultra-high dimensional partially linear additive models, *Journal of Applied Statistics*, 46(8), 1385-1403.

5. Kazemi, M., Shahsavani, D., Arashi, M. and Rodrigues, P. C. (2019). Estimation in partial linear model with spline modal function, *Communications in Statistics - Simulation and Computation* (Revised).

6. Kazemi, M., Shahsavani, D. and Arashi, M. (2018). Variable selection and structure identification for ultrahigh-dimensional partially linear additive models with application to cardiomyopathy microarray data, *Statistics, Optimization and Information Computing*, DOI: 10.19139/soic.v6i3.576.

7. Kazemi, M., Arashi, M. and Shahsavani, D. (2017), Gene identification using independence screening procedure, The 11th Seminar on Probability and Stochastic Processes, Imam Khomeini International University, Qazvin, Iran.

8. Kazemi, M., Shahsavani, D. and Arashi, M. (2017). Group LASSO for high-dimensional partially linear errors-in-variables models, Second National and First International Conference on Soft Computing, University of Guilan, Faculty of Engineering and Technology East of Guilan, Rudsar, Iran.

پیش‌گفتار

مدل‌های رگرسیونی نیمه‌پارامتری در سال‌های اخیر مورد توجه بسیاری از محققین قرار گرفته‌اند. یک رده پرکاربرد از مدل‌های نیمه‌پارامتری، رده مدل‌های خطی ـ جزئی، مانند مدل جمعی خطی‌ـ جزئی و مدل ضریب متغیر خطی‌ـ جزئی، است. در مدل جمعی خطی‌ـ جزئی، برخی از متغیرهای توضیحی دارای اثر خطی و برخی دیگر دارای اثر غیر خطی بر متغیر پاسخ هستند و در مدل ضریب متغیر خطی‌ـ جزئی، اثرات بعضی از متغیرهای توضیحی ثابت و اثرات برخی دیگر از متغیرها تابعی از یک متغیر شاخص است. علاوه بر این برخی از متغیرهای توضیحی ارتباطی با متغیر پاسخ ندارند. در برازش مدل‌های خطی‌ـ جزئی، معمولاً فرض بر این است که مؤلفه‌های خطی و غیر خطی یا اثرات ثابت و متغیر از پیش تعیین شده‌اند. اما در عمل معمولا تعیین این مؤلفه‌ها بسیار مشکل است. بنابراین موضوعاتی از قبیل انتخاب متغیرهای مهم، تشخیص مؤلفه‌های خطی و غیر خطی در مدل جمعی خطی‌ـ جزئی و تعیین اثرات ثابت و متغیر در مدل ضریب‌ـ متغیر خطی‌ـ جزئی بسیار حائز اهمیت است. از طرفی امروزه با افزایش توان محاسباتی کامپیوترها و روش‌های نوین جمع‌آوری داده‌ها، داده‌های با بعد بسیار بالا در زمینه‌های مختلف علمی و تحقیقاتی با هزینه نسبتاً کمی در اختیار محققین قرار می‌گیرد. در این نوع داده‌ها فقط تعداد اندکی از متغیرهای توضیحی واقعاً با متغیر پاسخ مرتبط هستند و سایر متغیرها تأثیری بر پاسخ ندارند. هنگامی که تعداد متغیرهای توضیحی در مدل زیاد است، تفسیر آن مدل بسیار مشکل است. همچنین حضور تعداد زیادی از متغیرهای توضیحی باعث ایجاد همخطی بین متغیرها شده و منجر به برآوردگری با واریانس بزرگ می‌شود. بنابراین در این نوع داده‌ها شناسایی متغیرهای مؤثر بر پاسخ اهمیت بسزایی دارد.

لیان[3] (۲۰۱۲c)، لیان و همکاران (۲۰۱۲) و لیان و همکاران (۲۰۱۵) به موضوع انتخاب متغیر و تشخیص ساختار در مدل جمعی خطی‌ـ جزئی پرداختند. مطالعات مذکور برای انتخاب متغیر و تشخیص ساختار در داده‌های با بعد بسیار بالا، به دلایل هزینه محاسباتی، دقت آماری و ناپایداری الگوریتمی، مناسب نیستند. لذا شناسایی متغیرهای مهم در این نوع داده‌ها و تشخیص نوع تاثیر آنها بر متغیر پاسخ یک چالش است. در این حالت لازم است قبل از انتخاب متغیر و تشخیص ساختار، از یک روش غربالگری برای کاهش بعد استفاده شود. در این راستا، فن و لیو[4] (۲۰۰۸) روش غربالگری مستقل مطمئن را برای کاهش بعد در مدل خطی پیشنهاد دادند. در این روش، متغیرهای توضیحی براساس همبستگی پیرسن آنها با متغیر پاسخ رتبه‌بندی شده و سپس تعدادی متغیر با بیشترین همبستگی انتخاب می‌شوند.

[3] Lian
[4] Fan and LV

برای انتخاب متغیر و تشخیص ساختار در مدل ضریب متغیر خطی–جزئی، زیا[5] و همکاران (۲۰۰۴)، لنگ[6] (۲۰۰۹)، نه و کیلگوم[7] (۲۰۱۲)، هو[8] و زیا (۲۰۱۲) روش‌های مختلفی را برای تشخیص اثرات ثابت و متغیر در مدل ضریب متغیر خطی–جزئی معرفی کردند، اما این روش‌ها متغیرهای مهم را انتخاب نمی‌کنند. تانگ[9] و همکاران (۲۰۱۲) یک روش واحد برای انتخاب متغیر و تشخیص ساختار در مدل ضریب متغیر خطی–جزئی معرفی کردند، اما این روش مستلزم استفاده از یک الگوریتم دو مرحله‌ای تکراری است که در عمل چندان کارا نیست. علاوه بر این روش معرفی‌شده بر اساس رگرسیون میانگین نسبت به مشاهدات پرت و توزیع دم‌سنگین خطاها بسیار حساس است. بنابراین ارائه یک روش کارا و نیرومند برای انتخاب متغیر و تشخیص اثرات ثابت و متغیر در این مدل ضروری به نظر می‌رسد.

در این رساله به مسئله انتخاب متغیر و تشخیص ساختار در مدل‌های نیمه‌پارامتری خطی–جزئی پرداخته می‌شود. این مجموعه شامل ۴ فصل و ۴ پیوست است:

- **فصل اول**، به معرفی مفاهیم مقدماتی در تحلیل رگرسیونی و مدل‌های مورد مطالعه اختصاص دارد.

- در **فصل دوم**، به بیان جزئیات روش‌های انتخاب متغیر، مانند روش‌های انتخاب متغیر کلاسیک، روش کمترین توان‌های دوم جریمه‌شده و روش غربالگری پرداخته می‌شود.

- در **فصل سوم**، یک روش دو مرحله‌ای برای انتخاب متغیر و تشخیص مؤلفه‌های خطی و غیرخطی در مدل جمعی با بعد بسیار بالا معرفی می‌شود. در مرحله اول، از یک روش غربالگری برای کاهش بعد فضای متغیرهای توضیحی استفاده می‌شود. این روش غربالگری بر اساس همبستگی فاصله‌ای بین متغیرهای توضیحی و تابع توزیع حاشیه‌ای متغیر پاسخ ساخته شده و زمانی که متغیر پاسخ دم سنگین یا دارای مقادیر فرین باشد، عملکرد خوبی را از خود نشان می‌دهد. در مرحله دوم، از روشی مبتنی بر دو تابع جریمه برای انتخاب همزمان مؤلفه‌های غیرصفر و خطی استفاده می‌شود.

- در **فصل چهارم**، یک روش نیرومند برای انتخاب متغیر و تشخیص ساختار در مدل ضریب متغیر خطی–جزئی براساس رگرسیون نما[10] ارائه می‌شود. رگرسیون نما توسط یائو[11] و همکاران (۲۰۱۲) و یائو و لی[12] (۲۰۱۴) براساس نمای شرطی متغیر پاسخ معرفی شده است.

- **پیوست آ**، اثبات قضایای بیان‌شده در فصل‌های ۳ و ۴ را در برمی‌گیرد.

[5] Xia
[6] Leng
[7] Noh and Keilegom
[8] Hu
[9] Tang
[10] Modal regression
[11] Yao
[12] Li

- **پیوست ب**، مرور مختصری بر توابع اسپلاین و B_ اسپلاین را شامل می‌شود.
- **پیوست پ**، نامساوی‌ها و تعاریف مورد استفاده در این رساله را یادآوری می‌نماید.
- **پیوست ت**، شامل گزیده‌ای از برنامه‌های نوشته‌شده با نرم‌افزار R است.

فهرست مطالب

فهرست تصاویر

فهرست جداول

فهرست نمادها

ماتریس طرح ..... $\mathbf{X}$
بردار پاسخ ..... $\mathbf{Y}$
بردار ضرایب ..... $\boldsymbol{\beta}$
مقدار درست بردار $\boldsymbol{\beta}$ ..... $\boldsymbol{\beta}_0$
بردار خطا ..... $\boldsymbol{\varepsilon}$
متغیر شاخص در مدل ضریب متغیر ..... U
تابع نشانگر مجموعه A ..... $I(A)$
ماتریس واحد با بعد m ..... $\mathbf{I}_m$
ترانهاده ماتریس $\mathbf{A}$ ..... $\mathbf{A}^T$
معکوس ماتریس $\mathbf{A}$ ..... $\mathbf{A}^{-1}$
معکوس ماتریس $\mathbf{X}^T\mathbf{X}$ ..... $\mathbf{C}$
ضریب تعیین ..... R^2
ضریب تعیین تعدیل‌شده ..... R^2_{adj}
نرم اقلیدسی ..... $\|\cdot\|$
واریانس ..... Var
کوواریانس ..... Cov
کوواریانس فاصله‌ای ..... $dcov$
ضریب همبستگی فاصله‌ای ..... $dcorr$
همگرایی در توزیع ..... $\xrightarrow{D}$
همگرایی در احتمال ..... $\xrightarrow{P}$
امید ریاضی ..... $E[.]$
برآوردگر کمترین توان‌های دوم معمولی ..... $\hat{\boldsymbol{\beta}}^{OLS}$
برآوردگر ریج ..... $\hat{\boldsymbol{\beta}}^{Ridge}$
برآوردگر لاسو ..... $\hat{\boldsymbol{\beta}}^{Lasso}$
برآوردگر لاسوی تطبیقی ..... $\hat{\boldsymbol{\beta}}^{aLasso}$
برآوردگر SCAD ..... $\hat{\boldsymbol{\beta}}^{SCAD}$
برآوردگر MCP ..... $\hat{\boldsymbol{\beta}}^{MCP}$

فهرست کلمات اختصاری

aLasso(Adaptive Least Absolute Shrinkage and Selection Operator)

AIC(Akaike's Information Criterion)

BIC(Bayesian Information Criterion)

DC-SIS(Sure Independence Screening Procedure based on the Distance Correlation)

GCV(Generalized Cross-Validation)

ISIS(Iterative Sure Independence Screening)

Lasso(Least Absolute Shrinkage and Selection Operator)

LLA(Local Linear Approximation)

LQA(Local Quadratic Approximation)

MARS(Multivariate Adaptive Regression Splines)

MCP(Minimax Concave Penalty)

MEM(Modal Expectation-Maximization)

NIS(Nonparametric Independence Screening)

OLS(Ordinary Least Squares)

PMR(Penalized Modal Regression)

PQR(Penalized Quantile Regression)

RDC-SIS(Robust Sure Independence Screening Procedure based on Distance Correlation)

RMSE(Square Root of Mean Squared Errors)

RSS(Residual Sum of Squares)

SCAD(Smoothly Clipped Absolute Deviation)

SIRS(Sure Independent Ranking and Screening)

SIS(Sure Independence Screening)

VIF(Variance Inflation Factor)

فصل ۱

مقدمات و مفاهیم اولیه

در این فصل، انواع مدل‌های رگرسیونی و مدل‌های مورد بررسی در این رساله، به اختصار بیان می‌شوند. مطالب عمده این فصل از هاردل و همکاران (۲۰۰۰)، هاردل و همکاران (۲۰۰۴) و هستی و همکاران (۲۰۰۹) گرفته شده است.

۱.۱ مدل‌بندی رگرسیونی

تحلیل رگرسیونی روشی ساده برای بررسی روابط تابعی میان متغیرهاست. به‌عنوان مثال، فرض کنید بخواهیم بررسی کنیم که آیا میزان دستمزد با تحصیلات و سابقه کار در بازار مرتبط است یا خیر. این رابطه به شکل یک معادله یا الگویی است که متغیر وابسته یا پاسخ را به یک یا چند متغیر توضیحی یا پیشگو مرتبط می‌کند. در این مثال، میزان حقوق متغیر پاسخ و متغیرهای تحصیل و سابقه کار متغیرهای توضیحی (پیشگو یا کمکی) هستند.

در ابتدا، رگرسیون خطی چندگانه را در نظر بگیرید. در این مدل، میانگین متغیر پاسخ Y به‌صورت

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \qquad (۱.۱)$$

با متغیرهای توضیحی $\mathbf{X} = (X_1, ..., X_p)^T$ در ارتباط است که $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ بردار ضرایب رگرسیونی نامعلوم است.

با تعریف ε به‌عنوان انحراف Y از میانگین شرطی $E(Y|\mathbf{X})$، یعنی $\varepsilon = Y - E(Y|\mathbf{X})$، مدل خطی را می‌توان به‌صورت زیر تعریف کرد

$$Y = \beta_۰ + \beta_۱ X_۱ + ... + \beta_p X_p + \varepsilon \qquad (۲.۱)$$

که در آن ε می‌تواند نشان‌دهنده عواملی باشد که در مدل‌بندی تصادفی Y تأثیر داشته، اما در $\mathbf{X}$ در نظر گرفته نشده‌اند.

مثال ۱.۱.۱. (هاردل[1] و همکاران، ۲۰۰۴) فرض کنید Y لگاریتم دستمزد باشد. متغیرهای توضیحی تحصیل $(School)$ برحسب سال، سابقه کار در بازار (Exp) و توان دوم متغیر Exp را در نظر بگیرید که متغیر Exp از رابطه $Exp = Age - School - ۶$ محاسبه می‌شود و Age بیانگر سن است. اگر فرض کنیم لگاریتم دستمزد به‌صورت خطی با این متغیرهای توضیحی مرتبط است، در این‌صورت مدل رگرسیون خطی عبارتست از

$$E(Y|School, Exp) = \beta_۰ + \beta_۱ School + \beta_۲ Exp + \beta_۳ Exp^۲. \qquad (۳.۱)$$

معادله (۳.۱) به معادله درآمد سرمایه انسانی[2] یا معادله در آمد مینسر[3] معروف است و نقش مهمی را در اقتصاد تجربی بازی می‌کند. ویژگی مهم مدل (۳.۱) صورت پارامتری آن است، یعنی شکل تابع رگرسیونی به ضرایب رگرسیونی نامعلوم $\boldsymbol{\beta} = (\beta_۰, \beta_۱, \beta_۲, \beta_۳)^T$ بستگی دارد. بنابراین، برای تعیین تابع رگرسیونی، لازم است این ضرایب مجهول برآورد شوند. این مدل، روابط غیر خطی احتمالی بین متغیرهای توضیحی و پاسخ را نادیده می‌گیرد. لذا هنگامی که رابطه میان متغیر پاسخ و متغیرهای توضیحی نامعلوم است، استفاده از رگرسیون خطی ممکن است منجر به نتایج نادرست شود.

حال فرض کنید شکل درست (اما نامعلوم) تابع رگرسیونی لگاریتم دستمزد روی متغیرهای توضیحی تحصیل و سابقه کار به‌صورت

$$E(Y|School, Exp) = m(School, Exp) \qquad (۴.۱)$$

باشد و شما بخواهید تابع رگرسیونی $m(School, Exp)$ را تنها در یک *بار* کوشش و تا حد ممکن دقیق برآورد کنید. بدین معنی که اگر پی بردید مدل اولیه داده‌ها را به‌خوبی برازش نمی‌کند، دیگر مجاز به تغییر مدل نیستید. در ابتدا ممکن است فرض کنید، تابع رگرسیونی که مسئولیت برآورد آن را به عهده گرفته‌اید، دارای صورت مشخص (۳.۱) است، یعنی فرض می‌کنید که

$$m(School, Exp) = \beta_۰ + \beta_۱ School + \beta_۲ Exp + \beta_۳ Exp^۲$$

و سپس ضرایب نامعلوم را با استفاده از روش کمترین توان‌های دوم برآورد کنید. اما اگر به شما گفته شود که راه‌های دیگری برای برآورد تابع رگرسیونی فوق وجود دارند که به‌جز پذیره هموار بودن، هیچ

[1] Härdle
[2] Human capital earnings equation
[3] Mincer earnings equation

پذیره پیشین دیگری برای شکل تابع رگرسیونی در نظر نمی‌گیرند، شاید دیگر مدل پارامتری فوق را برازش نکنید. لازم به ذکر است که در اینجا منظور از هموار بودن، مشتق‌پذیر بودن تابع $m(\cdot)$ تا یک مرتبه خاص و کوچک است. به خاطر داشته‌باشید که در این مسئله، شما تنها یک فرصت دارید و اگر شکل درست $m(School, Exp)$ خیلی متفاوت از (۳.۱) باشد، برآورد مدل پارامتری ممکن است نتایج گمراه‌کننده‌ای را به دنبال داشته باشد. روش‌های متعددی برای برآورد تابع $m(\cdot)$ وجود دارند که تنها فرض می‌شود $m(\cdot)$ یک تابع هموار است. به این روش‌ها، رگرسیون ناپارامتری گویند.

برآوردگرهای رگرسیون ناپارامتری بسیار انعطاف‌پذیر هستند، اما با افزایش تعداد متغیرهای توضیحی در مدل، دقت آماری این برآوردگرها به شدت کاهش می‌یابد (هاردل و همکاران، ۲۰۰۴). این مشکل به مشقت بُعد[۴] معروف است. در نتیجه، محققان سعی کرده‌اند به مدل‌ها و برآوردگرهایی بپردازند که انعطاف‌پذیری بیشتری نسبت به رگرسیون پارامتری استاندارد دارند، و در عین حال بر مشکل مشقت بعد غلبه می‌کنند. چنین روش‌هایی معمولاً خصوصیات روش‌های پارامتری و ناپارامتری را با هم ترکیب می‌کنند و لذا به روش های نیمه‌پارامتری معروف هستند. مزیت‌های دیگر روش‌های نیمه‌پارامتری عبارتند از: امکان ورود متغیرهای طبقه‌ای به مدل، سهولت تفسیر نتایج و امکان مشخص کردن بخشی از مدل با استفاده از اطلاعات پیشین. بنابراین، رگرسیون را بر حسب نوع تابعی که رابطه بین متغیرهای توضیحی و متغیر پاسخ را تعیین می‌کند، می‌توان به سه نوع پارامتری، ناپارامتری و نیمه‌پارامتری تقسیم کرد.

در ادامه، برای درک بهتر مطلب و پیش‌درآمدی بر مطالب کلیدی این رساله، از معادله درآمد و مثال‌های دیگر برای نشان دادن تفاوت‌های بین روش‌های پارامتری، ناپارامتری و نیمه‌پارامتری استفاده می‌کنیم.

۱.۱.۱ رگرسیون پارامتری

در این زیر بخش، برای بررسی روابط بین لگاریتم درآمد و متغیرهای توضیحی از مدل پارامتری (۳.۱) استفاده می‌کنیم. این مدل، بیش از هر مدل دیگری در اقتصاد تجربی توسط محققین استفاده شده‌است.

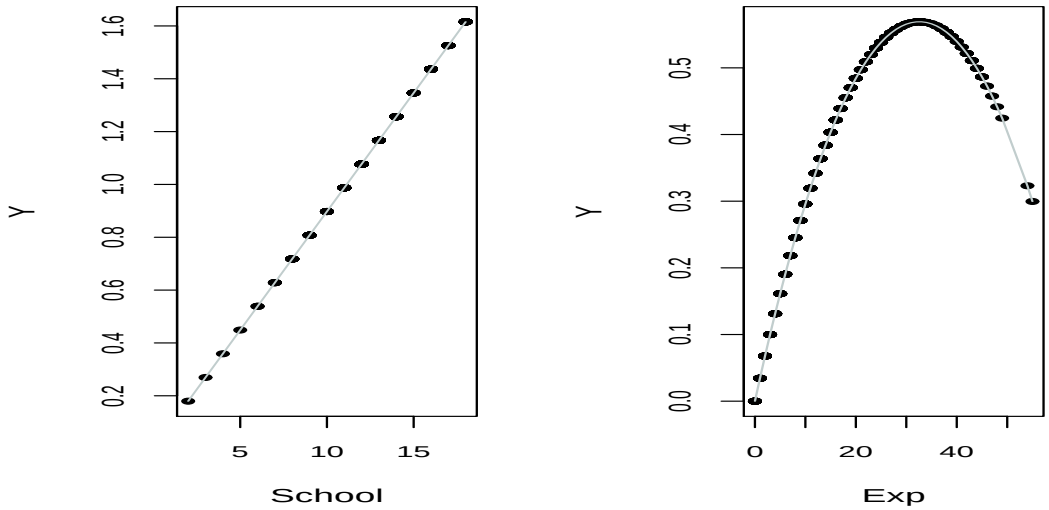
طبق نظریه سرمایه انسانی، اثر سابقه کار بر دستمزد، یا نیمرخ دستمزد ـ سابقه کار، یک تابع مقعر است، چون مراحل اولیه شروع کار حرفه‌ای در بازار با افزایش دستمزد و انباشت سریع سرمایه انسانی همراه است و در یک نقطه‌ای در میانسالی به اوج می‌رسد. بعد از آن، میزان درآمد به دلیل کاهش ساعات کار و انگیزه برای سرمایه‌گذاری کاهش می‌یابد. این دلیل استفاده از Exp و Exp^2 در مدل است. برای به‌دست آوردن مدلی که با نظریه سرمایه انسانی مطابقت داشته باشد، باید علامت β_2 مثبت و β_3 منفی باشد.

ضرایب مدل (۳.۱) را با استفاده از روش کمترین توان‌های دوم و براساس زیرنمونه‌ای به حجم $n = ۵۳۴$ از نظرسنجی جمعیت جاری سال ۱۹۸۵ که توسط اداره آمار ایالات متحده امریکا ارائه شده است، برآورد شده و نتایج در جدول ۱.۱ ارائه شده است. این داده‌ها در بسته "AER" نرم افزار R با نام "CPS۱۹۸۵" قابل دسترسی است.

[4] Curse of dimensionality

جدول ۱.۱: نتایج برآورد کمترین توان‌های دوم در مثال ۱.۱.۱

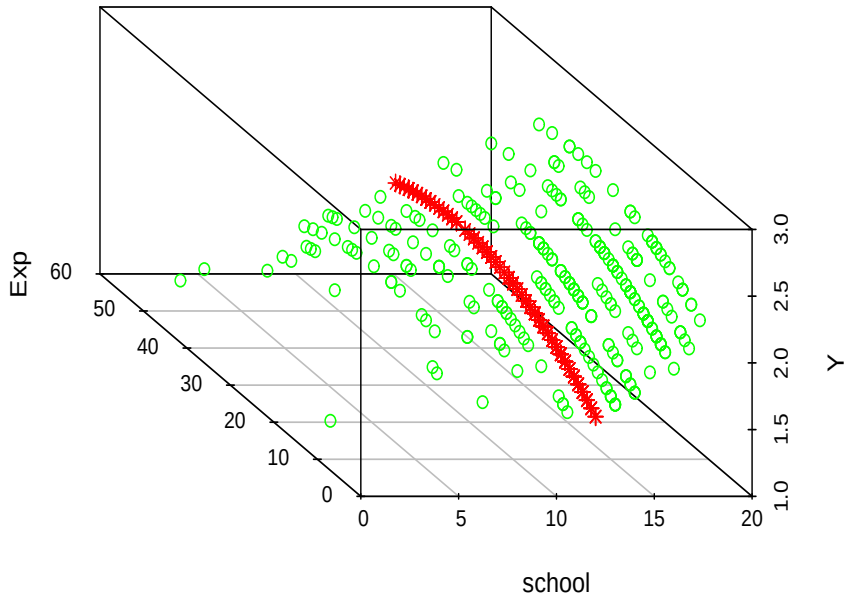
متغیر وابسته: لگاریتم درآمد			
متغیر	ضرایب	خطای معیار	آماره t
ثابت	۰٫۵۲۰۲	۰٫۱۲۳۶	۴٫۲۰۹
$School$	۰٫۰۸۹۸	۰٫۰۰۸۳	۱۰٫۷۸۸
Exp	۰٫۰۳۴۹	۰٫۰۰۵۶	۶٫۱۸۵
Exp^2	−۰٫۰۰۰۵	۰٫۰۰۰۱	−۴٫۳۰۷



شکل ۱.۱: قاب سمت چپ اثر متغیر تحصیل و قاب سمت راست اثر متغیر سابقه کار بر پاسخ را نشان می‌دهد.

توجه داشته باشید که ضرایب برآوردشده Exp و Exp^2 دارای همان علامت پیش‌بینی‌شده توسط نظریه سرمایه انسانی است. شکل ۱.۱ نیمرخ دستمزد–تحصیل ($School$ در مقابل ۰٫۰۸۹۸$School$) و نیمرخ دستمزد–سابقه کار (نمودار Exp در مقابل ۰٫۰۳۴۹Exp − ۰٫۰۰۰۵Exp^2) را نشان می‌دهد. شکل ۱.۱ پیش‌بینی نظریه سرمایه انسانی درباره مقعر بودن نیمرخ دستمزد–سابقه کار را تایید می‌کند.

رویه رگرسیونی برآوردشده در شکل ۲.۱ نشان داده شده است. در این شکل، محور عمودی مقادیر تابع رگرسیونی برآوردشده و محورهای افقی متغیرهای $School$ و Exp را نشان می‌دهند. همه منحنی‌های عضو این رویه در جهت Exp شبیه نمودار ۱.۱ (قاب سمت راست) و در جهت $School$ شبیه نمودار ۱.۱ (قاب سمت چپ) هستند. برای درک بهتر این نمودار سه بعدی، نیمرخ دستمزد–سابقه کار را، با ثابت گرفتن تحصیل در ۱۲ سالگی، نشان داده‌ایم، یعنی منحنی قرمز رنگ نیمرخ دستمزد–سابقه کار را برای فارغ‌التحصیلان دبیرستان نشان می‌دهد. نتایج شکل ۲.۱ با شکل ۱.۱ همخوانی دارند.



شکل ۲.۱: تابع رگرسیونی برآورد شده به صورت پارامتری. منحنی قرمز رنگ، نیمرخ دستمزد–سابقه کار را برای فارغ‌التحصیلان دبیرستان نشان می‌دهد.

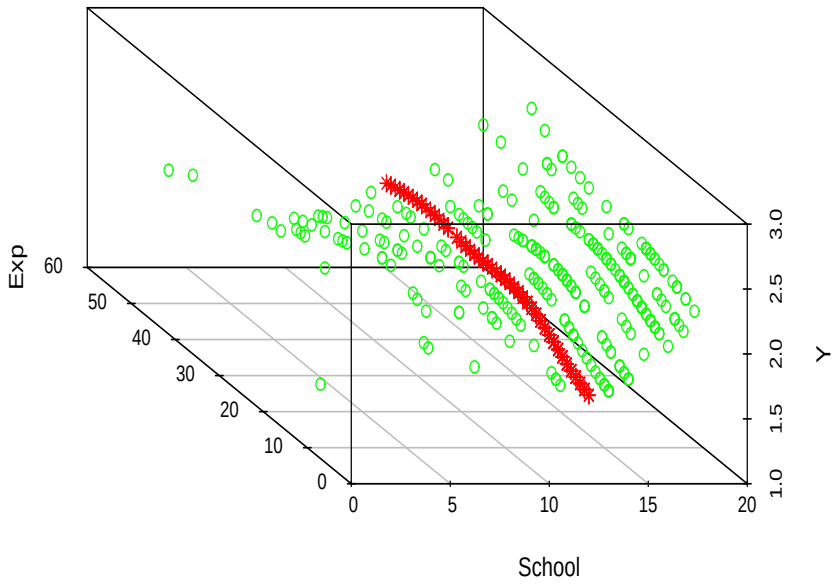
۲.۱.۱ رگرسیون ناپارامتری

در رگرسیون پارامتری رابطه بین متغیرهای پیشگو و متغیر پاسخ از طریق یک تابع پارامتری معین تعریف می‌شود. بنابراین برای برازش یک مدل پارامتری نیاز به اطلاعات پیشین درباره ساختار داده‌ها است. با افزایش بعد داده‌ها و پیچیدگی مسائل مورد مطالعه، تعیین شکل درست مدل‌های پارامتری چندان ساده نیست. تعیین نادرست شکل مدل‌های پارامتری ممکن است منجر به اریبی برآورد و استنباط نادرست شود. در مقابل، مدل‌های رگرسیونی ناپارامتری، بدون نیاز به اطلاعات پیشین، انعطاف‌پذیری زیادی را برای شناخت ساختار داده‌ها فراهم آورده‌اند.

فرض کنید

$$E(Y|School, Exp) = m(School, Exp) \qquad (۵.۱)$$

که $m(\cdot)$ یک تابع هموار است. هدف برآورد تابع $m(\cdot)$ بر اساس داده‌هاست. یک روش رایج در رگرسیون ناپارامتری میانگین وزنی موضعی است. فرض کنید $x = (s, e)$ مقدار دلخواهی باشد که $(School, Exp)$ می تواند اختیار کند. در روش میانگین وزنی موضعی، برای برآورد تابع $m(\cdot)$ در نقطه $x_\circ = (s_\circ, e_\circ)$، به آن مقادیری از Y که متناظر با مقادیر x نزدیک‌تر به نقطه $x_\circ$ باشند، وزن بیشتر و



شکل ۳.۱: تابع رگرسیونی برآوردشده با روش ناپارامتری. منحنی قرمز رنگ، نیمرخ دستمزد–سابقه کار را برای فارغ‌التحصیلان دبیرستان نشان می‌دهد.

به مقادیر Y متناظر با نقاط دورتر وزن کمتری نسبت داده می‌شود. نادارایا و واتسون[۵] (۱۹۶۴) پیشنهاد دادند که وزن‌ها در رگرسیون ناپارامتری می‌توانند به‌وسیله یک تابع متقارن تک‌مدی حول صفر که مقادیر آن در دو طرف صفر توسط یک پارامتر مقیاس کنترل می‌شود، انتخاب شود. نخستین نامزد آن‌ها برای چنین توابعی که هسته نامیده شدند، توابع چگالی احتمال بودند. سپس با استفاده از توابع کراندار هسته، برآوردگر زیر را پیشنهاد دادند

$$\hat{m}(x_\circ) = \sum_{i=1}^{n} W_i Y_i, \qquad W_i = \frac{K_h(x_\circ, x_i)}{\sum_{i=1}^{n} K_h(x_\circ, x_i)} \tag{۶.۱}$$

که در آن $K_h(.)$ یک تابع هسته بوده و پارامتر h پهنای باند[۶] است. برآوردگر (۶.۱) به هموارساز هسته[۷] معروف است. شکل وزن‌ها توسط تابع هسته و مقادیر آن‌ها بر اساس فاصله از نقطه $x_\circ$ و پهنای باند h تعیین می‌شود. بزرگ بودن پهنای باند باعث می‌شود که به نقاط نزدیک‌تر وزن‌های بزرگتری داده شود. ساده‌ترین هسته، هسته یکنواخت و پرکاربردترین آن‌ها، هسته نرمال می‌باشد. برآورد ناپارامتری تابع $m(\cdot)$ با استفاده از هموارساز (۶.۱) با هسته نرمال در شکل ۳.۱ آمده است.

[5] Nadaraya and Watson
[6] Bandwidth
[7] Kernel smoothing

سایر روش‌ها برای برازش مدل ناپارامتری عبارتند از: رگرسیون اسپلاین، هموارسازی اسپلاین، اسپلاین رگرسیونی تطبیقی چندمتغیره (MARS) و رگرسیون موضعی. برای مطالعه بیشتر به هستی[8] و همکاران (۲۰۰۹) و گرین و سیلورمن[9] (۱۹۹۳) مراجعه کنید.

۳.۱.۱ رگرسیون نیمه پارامتری

اگرچه رگرسیون ناپارامتری از انعطاف‌پذیری بالایی برخوردار است، اما افزایش بعد فضای متغیرهای توضیحی، استفاده از مدل‌های ناپارامتری را با مشکل مواجه می‌کند. استون[10] (۱۹۸۲) نشان داد که اگر مشتق تابع $m(\cdot)$ تا مرتبه r موجود باشد، سریع‌ترین نرخ همگرایی قابل حصول برآوردگر $\hat{m}(\cdot)$ به $m(\cdot)$ برابر با $(\log n/n)^{r/(2r+p)}$ است که در آن p بعد متغیرهای توضیحی است. این نرخ همگرایی نقش بعد متغیرهای توضیحی را در کارایی برآوردگر نشان می‌دهد، به عبارت دیگر، افزایش p منجر به کاهش عملکرد برآوردگر $\hat{m}(\cdot)$ می‌شود. این مسئله به ''مشقت بعد'' معروف است و به دلیل تنک بودن داده‌ها در فضای با ابعاد بالاتر اتفاق می‌افتد. برای اطلاعات بیشتر به گیننس[11] (۲۰۱۱) مراجعه کنید.

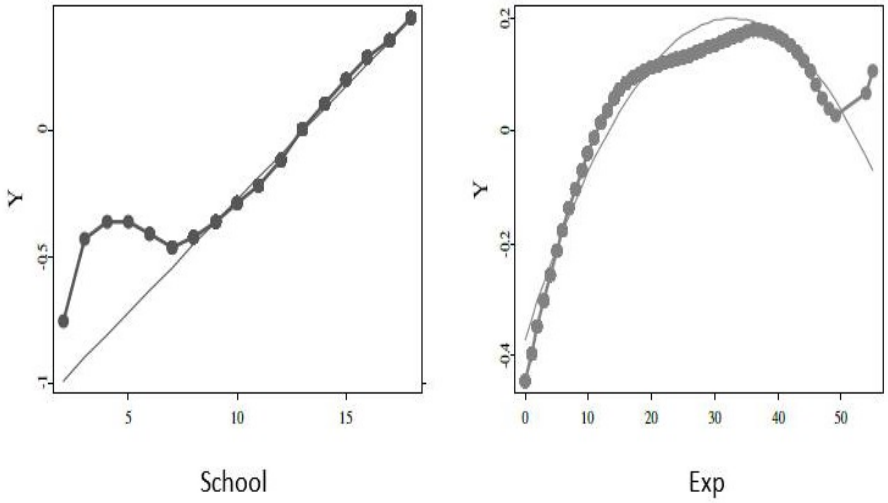
برای اجتناب از اریبی زیاد مدل‌های پارامتری و مشقت بعد مدل‌های ناپارامتری می‌توان از مدل‌های نیمه‌پارامتری استفاده کرد. برای شرح روش نیمه‌پارامتری، دوباره مدل درآمد سرمایه انسانی (۳.۱) را در نظر بگیرید. فرض کنید تابع رگرسیونی به شکل مدل جمعی زیر باشد

$$E(Y|School, Exp) = \alpha + g_1(School) + g_2(Exp) \qquad (۷.۱)$$

که $g_1(\cdot)$ و $g_2(\cdot)$ توابع هموار نامعلوم و α پارامتر نامعلوم است. این مدل ساختار جمعی رگرسیون پارامتری را با انعطاف‌پذیری رگرسیون ناپارامتری ترکیب می‌کند. در این مدل هیچ فرض محدود کننده قوی برای شکل توابع تعیین کننده اثر متغیرهای توضیحی $School$ و Exp بر لگاریتم درآمد در نظر گرفته نمی‌شود. بنابراین برای برآورد توابع نامعلوم $g_1(\cdot)$ و $g_2(\cdot)$ باید از روش‌های ناپارامتری استفاده شود. شکل ۴.۱ برآوردهای توابع $g_1(\cdot)$ و $g_2(\cdot)$ را در مقابل نیمرخ دستمزد ـ تحصیل و دستمزد ـ سابقه کار برآورد شده با روش پارامتری نشان می‌دهد. در این شکل، نیمرخ‌های برآوردشده با روش پارامتری با خطوط نازک و برآوردهای $g_1(\cdot)$ و $g_2(\cdot)$ با خطوط ضخیم نشان داده شده است. با توجه به نمودار، شباهت زیادی بین این دو برآورد وجود دارد، اما در نقاط پایانی تفاوت مشاهده می‌شود. همچنین، رویه رگرسیونی برآوردشده با روش نیمه پارامتری در شکل ۵.۱ نشان داده شده است. این رویه برآورد شده همخوانی خوبی بین برآوردهای پارامتری و برازش مدل جمعی نشان می‌دهد.

لذا می‌توانیم نتیجه بگیریم که در این مثال خاص، مدل پارامتری توسط روش‌های ناپارامتری و نیمه‌پارامتری تایید می‌شود. بنابراین روش‌های ناپارامتری و نیمه‌پارامتری را می‌توان برای بررسی کفایت مدل پارامتری نیز به‌کار برد.

[8] Hasti
[9] Green and Silverman
[10] Stone
[11] Geenens



شکل ۴.۱: برازش مدل جمعی در مقابل برازش مدل پارامتری. قاب سمت چپ اثر متغیر تحصیل و قاب سمت راست اثر متغیر سابقه کار بر پاسخ را نشان می‌دهد.

حال با دقت بیشتری به مدل‌های (۵.۱) و (۷.۱) نگاه کنید. مشاهده می‌کنید که در مدل (۵.۱) لازم است یک تابع نامعلوم شامل دو متغیر را برآورد کنیم، اما در مدل (۷.۱) باید دو تابع نامعلوم برآورد شود که هریک از این دو تابع شامل یک متغیر است. این بدین معنی است که با برازش مدل (۷.۱) توانسته‌ایم بعد مسئله برآورد را کاهش دهیم. اگرچه همه محققین بر این باورند که برازش مدل‌های جمعی مانند (۷.۱) باعث اعمال کاهش بعد بر مدل‌های کاملاً ناپارامتری می شود، اما برخی از آن‌ها مدل (۷.۱) رابه‌عنوان مدل نیمه‌پارامتری قبول ندارند. زیرا در این مدل هیچ پارامتری، به جز عرض از مبدأ، برای برآورد وجود ندارد. تاکنون مدل‌های رگرسیونی نیمه‌پارامتری متعددی توسط افراد مختلف معرفی شده‌اند. رایج‌ترین مدل‌های نیمه‌پارامتری عبارتند از: مدل جمعی خطی‌ـجزئی[۱۲]، مدل ضریب متغیر خطی‌ـجزئی[۱۳] و مدل تک‌شاخص[۱۴] که در ادامه به معرفی این مدل‌ها خواهیم پرداخت.

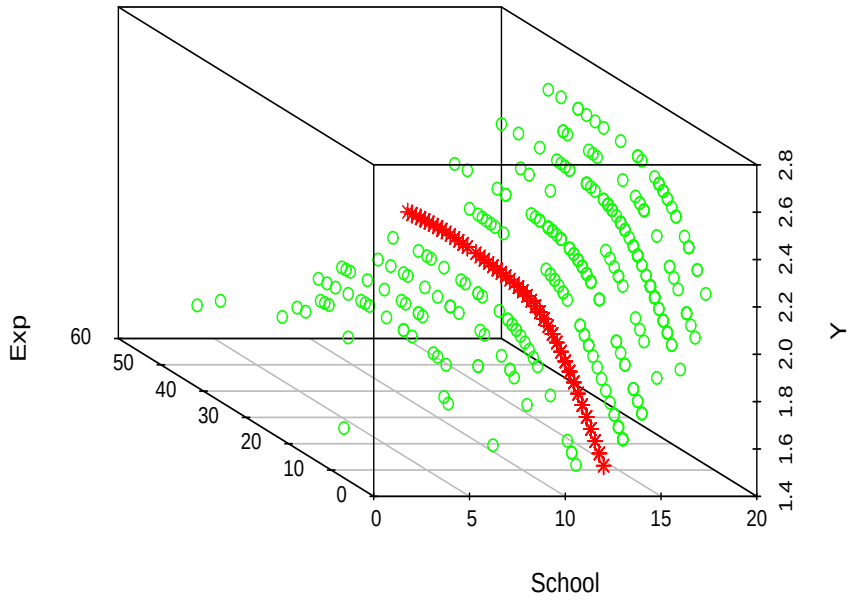
مدل جمعی خطی‌ـجزئی

مدل جمعی خطی‌ـجزئی به صورت

$$Y = \sum_{j\epsilon S_1} \beta_j X_j + \sum_{j\epsilon S_2} f_j(X_j) + \varepsilon \qquad (۸.۱)$$

تعریف می‌شود که در آن $f_j(\cdot)$ ها توابع هموار نامعلوم و S_1 و S_2 دو زیر مجموعه مجزا از $\{1,\ldots,p\}$ می‌باشند. این مدل تفسیرپذیری مدل‌های خطی و انعطاف‌پذیری مدل‌های ناپارامتری را با هم ترکیب می‌کند و یکی از رایج‌ترین مدل‌های نیمه‌پارامتری است. هاردل و همکاران (۲۰۰۰) بررسی دقیقی از این

[12] Partially linear additive models
[13] Partially linear varying coefficient model
[14] Single index model



شکل ۵.۱: تابع رگرسیونی برآوردشده با روش نیمه‌پارامتری. منحنی قرمز رنگ، نیمرخ دستمزد–سابقه کار را برای فارغ‌التحصیلان دبیرستان نشان می‌دهد.

مدل نیمه‌پارامتری ارائه داده‌اند. مدل جمعی خطی–جزئی دارای کاربردهای متعددی است. انگل[۱۵] و همکاران (۱۹۸۶) از نخستین افرادی بودند که این مدل را در نظر گرفتند. آن‌ها از این مدل برای بررسی رابطه بین دما و مصرف برق استفاده کردند. در ادامه به بیان چند مثال واقعی می‌پردازیم که محققین مدل جمعی خطی–جزئی را به‌کار برده‌اند.

مثال ۲.۱.۱. هاردل و همکاران (۲۰۰۴) از مدل جمعی خطی–جزئی برای بررسی تأثیر عوامل مختلف بر عرضه نیروی کار زن در شرق آلمان استفاده کردند. در این داده‌ها، عرضه نیروی کار زن با تعداد ساعت های کاری هفتگی (Y) اندازه‌گیری می‌شود. مجموعه داده‌های مورد بررسی یک زیر نمونه از پنل اقتصادی اجتماعی آلمان (GSOEP) شامل $n = ۶۰۷$ زن است که دارای شغل و در حال زندگی با همسر خود هستند. متغیرهای توضیحی عبارتند از: نرخ بیکاری در ایالتی که زن در آن زندگی می‌کند $(U_۱)$، زن دارای فرزند کمتر از ۱۶ ساله است، بله=۱ و خیر=۰ $(U_۲)$، سن برحسب سال $(T_۱)$، دستمزد هر ساعت $(T_۲)$، شاخص اعتبار شغل $(T_۳)$، اجاره ماهانه آپارتمان یا محل اقامت $(T_۴)$، تعداد سال‌های تحصیل $(T_۵)$، درآمد خالص ماهانه شوهر $(T_۶)$.

[15] Engle

هاردل و همکاران (۲۰۰۴) مدل جمعی خطی–جزئی زیر را برای تحلیل این مجموعه داده به کار بردند

$$Y = \alpha + \sum_{j=۱}^{۲} \beta_j U_j + \sum_{j=۱}^{۶} g_j(T_j) + \varepsilon.$$

آن‌ها نشان دادند که میانگین متغیر پاسخ به‌طور خطی با متغیرهای $U_۱$ و $U_۲$ در ارتباط است، در حالی که نحوه ارتباط آن با سایر متغیرهای توضیحی نامعلوم است.

مثال ۳.۱.۱. انگل و همکاران (۱۹۸۶) از مدل جمعی خطی–جزئی برای بررسی رابطهٔ بین میزان برق مصرفی ماهیانه (Y) با متغیرهای قیمت برق $(X_۱)$، درآمد $(X_۲)$ و متوسط دمای هوا (t) استفاده کردند. آن‌ها متغیر پاسخ Y را به صورت مجموع یک تابع هموار g از دمای ماهیانه t و متغیرهای $X_۱$ و $X_۲$ و همچنین متغیرهای نشانگر $X_۳, ..., X_{۱۳}$، که معرف ماه‌های میلادی هستند، به صورت

$$Y = \sum_{j=۱}^{۱۳} \beta_j X_j + g(t) + \varepsilon$$

مدل‌سازی کردند. آن‌ها دریافتند که متغیر پاسخ Y دارای رابطه خطی با متغیرهای $X_۱$ و $X_۲$ و غیر خطی با متغیر t است.

مثال ۴.۱.۱. اشمالنسی و استوکر[۱۶] (۱۹۹۹) مدل جمعی خطی–جزئی را برای بررسی رابطهٔ بین مصرف گازوئیل خانواده با متغیرهایی مانند درآمد، سن، تعداد افراد خانوار، نوع خانه و محل اقامت در ایالات متحدهٔ آمریکا به کار بردند. آن‌ها دریافتند که لگاریتم مصرف گازوئیل بر حسب گالن دارای رابطهٔ غیرخطی با لگاریتم متغیرهای درآمد و سن، و رابطهٔ خطی با لگاریتم تعداد افراد خانوار، نوع خانه و محل اقامت است.

مثال ۵.۱.۱. هوانگ[۱۷] و همکاران $(۲۰۱۲a)$ یک مجموعه داده مربوط به دیابت را، که توسط ویلمز[۱۸] (۱۹۹۷) گزارش شده است، مورد بررسی قرار دادند. در این داده ها، ۱۶ متغیر بر روی ۴۰۳ نفر اندازه‌گیری شده است تا میزان شیوع چاقی، دیابت و سایر عوامل خطر بیماری قلبی و عروقی را در ویرجینیای مرکزی بررسی کنند. دیابت نوع دوم با چاقی همراه است. مقدار هموگلوبین گلیکوزیله[۱۹] بیشتر از ۰/۷ معمولاً به‌عنوان یک تشخیص مثبت از این بیماری در نظر گرفته می‌شود.

هوانگ و همکاران $(۲۰۱۲a)$ هموگلوبین گلیکوزیله رابه‌عنوان متغیر پاسخ و ۱۵ متغیر دیگر رابه‌عنوان متغیرهای توضیحی در نظر گرفتند. این ۱۵ متغیر شامل کلسترول $(X_۱)$، گلوکز تثبیت شده $(X_۲)$، لیپوپروتئین با چگالی بالا $(X_۳)$، نسبت کلسترول به لیپوپروتئین با چگالی بالا $(X_۴)$، مکان $(X_۵)$، سن $(X_۶)$، جنسیت $(X_۷)$، قد $(X_۸)$، وزن $(X_۹)$، اسکلت $(X_{۱۰})$، اولین فشار خون سیستولیک $(X_{۱۱})$، اولین فشار خون دیاستولیک $(X_{۱۲})$، دور کمر $(X_{۱۳})$، دور ران $(X_{۱۴})$، و زمان طراحی آزمایشگاه $(X_{۱۵})$

[16] Schmalensee and Stoker
[17] Huang
[18] Willems
[19] Glycosylated hemoglobin

است. در بین ۱۵ متغیر توضیحی، ۳ متغیر مکان، جنسیت و اسکلت متغیرهای طبقه‌ای و ۱۲ متغیر دیگر پیوسته هستند. آن‌ها با به‌کارگیری روشی موسوم به رگرسیون تعقیب نیمه‌پارامتری[۲۰] دریافتند که علاوه بر ۳ متغیر طبقه‌ای، متغیرهای پیوسته $X_۳$، $X_۹$، $X_{۱۱}$، $X_{۱۲}$، $X_{۱۳}$، $X_{۱۴}$، دارای اثر خطی و سایر متغیرها دارای رابطه غیرخطی با متغیر پاسخ هستند.

مدل ضریب متغیر خطی‌ـ‌جزئی

مدل ضریب متغیر تعمیمی از مدل خطی است که برای افزایش انعطاف‌پذیری مدل، ضرایب رگرسیونی با توابع ناپارامتری هموار جایگزین می‌شوند (کلولند[۲۱] و همکاران، ۱۹۹۱؛ هستی و تیبشیرانی[۲۲]، ۱۹۹۳). صورت کلی این مدل به‌صورت

$$Y = \sum_{j=۱}^{p} \beta_j(U) X_j + \varepsilon \qquad (۹.۱)$$

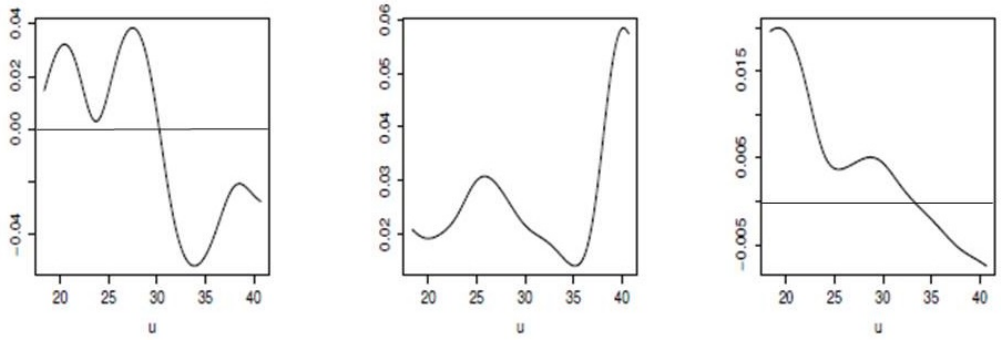
است که U متغیر شاخص[۲۳] یا تعدیل‌کننده[۲۴]، و $\boldsymbol{\beta}(U) = (\beta_۱(U), ..., \beta_p(U))^T$ بردار توابع ضرایب، شامل توابع هموار نامشخص از U، است و به کمک نمونه برآورد می‌شوند. در این مدل، متغیر شاخص معمولاً سن یا زمان در نظر گرفته می‌شود و هر یک از متغیرهای پیشگو، برای مقادیر داده‌شده متغیر U، دارای رابطه خطی با میانگین متغیر پاسخ می‌باشند. در عمل، برخی از توابع ضرایب در مدل (۹.۱) ثابت و بقیه ضرایب توابعی از متغیر شاخص U هستند. در این حالت خاص می‌توان مدل را به‌صورت

$$Y = \sum_{j \in S_۱} \beta_j X_j + \sum_{j \in S_۲} \beta_j(U) X_j + \varepsilon \qquad (۱۰.۱)$$

نوشت. مدل (۱۰.۱) را مدل ضریب متغیر خطی‌ـ‌جزئی گویند. در نظر گرفتن ضرایب ثابت به صورت تابعی از متغیر شاخص باعث افزایش اریبی و کاهش دقت برآورد می‌شود. بنابراین، مدل ضریب متغیر خطی‌ـ‌جزئی در تحلیل داده‌های واقعی بسیار مفید است.

مثال ۶.۱.۱. ژائو[۲۵] و همکاران (۲۰۱۴) با استفاده از مدل ضریب متغیر به بررسی رابطه بین سطح بتاکاروتن پلاسما و متغیرهایی مانند جنسیت، وضعیت استعمال سیگار، شاخص توده بدنی، وضعیت مصرف ویتامین، تعداد کالری، میزان چربی، میزان فیبر، تعداد نوشیدنی‌های الکلی، کلسترول و سن پرداختند. ابتدا متغیرهای طبقه‌ای وضعیت استعمال سیگار و مصرف ویتامین را با متغیرهای ساختگی[۲۶] جایگزین کردند. سپس با در نظر گرفتن سن به‌عنوان متغیر شاخص نشان دادند که متغیر سن دارای اثر غیر خطی و متغیرهای جنسیت، وضعیت استعمال سیگار، وضعیت مصرف ویتامین و نوشیدنی های الکلی دارای اثر خطی روی پاسخ هستند و اثر سایر متغیرهای توضیحی تابعی از سن است.

[20] Semiparametric regression pursuit
[21] Cleveland
[22] Tibshirani
[23] Index variable
[24] Modifying predictor
[25] Zhao
[26] Dummy variables



شکل ۶.۱: توابع ضرایب برآورد شده مثال ۷.۱.۱. نمودارها، به‌ترتیب، از چپ به راست $\hat{\alpha}_1(u)$، $\hat{\alpha}_2(u)$ و $\hat{\alpha}_3(u)$ را نشان می‌دهند.

مثال ۷.۱.۱. لی[۲۷] و همکاران (۲۰۱۶) از مدل ضریب متغیر برای مدل‌بندی داده‌های وزن نوزاد هنگام تولد استفاده کردند. بدین منظور، ۶۴ مادر که در یک بیمارستان خاص از جمهوری چک در فاصله زمانی نوامبر ۲۰۰۸ تا مارس ۲۰۰۹ زایمان کرده بودند، مورد بررسی قرار گرفتند. در این راستا، وزن نوزاد برحسب کیلوگرم (BW) همراه با متغیرهایی مانند سن مادر (MOA)، مدت زمان بارداری (GEA) و شاخص توده بدنی مادر (BMI) ثبت شد. علاوه بر این، سطح بیان ۵۸۶۹ ژن این مادران نیز در دسترس است. هدف شناسایی ژن‌های مؤثر بر وزن نوزاد است. آن‌ها با در نظر گرفتن سن مادر به‌عنوان متغیر شاخص، مدل ضریب متغیر خطی–جزئی زیر را برای برازش داده‌های وزن هنگام تولد به صورت زیر به‌کار بردند

$$BW = \alpha_1(U) + \alpha_2(U)GEA + \alpha_3(U)BMI + \sum_{j=1}^{5869} jGE_j + \varepsilon$$

که $U = MOA$ و jGE_j بیان j–امین ژن است. پس از برازش مدل فوق، ۱۰ ژن مؤثر شناسایی شدند. توابع ضرایب برآوردشده در شکل ۶.۱ نشان داده شده‌اند. با توجه به این شکل، می‌توان دید که تا قبل از ۳۰ سالگی مادران، سن مادر دارای تاثیر مثبت بر وزن نوزاد و برای سن بیش از ۳۰ سال دارای اثر منفی است. مدت زمان بارداری دارای اثر مثبت بر وزن نوزاد است و میزان تاثیر بعد از ۳۵ سالگی به سرعت افزایش می‌یابد. همچنین، متغیر BMI دارای اثر مثبت تا سن ۳۵ سالگی، و بعد از ۳۵ سالگی دارای اثر منفی بر وزن نوزاد است.

مدل تک شاخص

مدل رگرسیونی تک‌شاخص به شکل

$$Y = g(\mathbf{X}^T\boldsymbol{\beta}) + \varepsilon \qquad (۱۱.۱)$$

تعریف می‌شود که g یک تابع نامعلوم است. تابع g را معمولاً تابع پیوند و کمیت $X^T\beta$ را شاخص گویند. مدل (۱۱.۱) بسیاری از مدل‌های پارامتری پرکاربرد را به‌عنوان یک حالت خاص شامل می‌شود. اگر g

[27] Li

تابع همانی باشد، پس (۱۱.۱) یک مدل خطی است. همچنین اگر g تابع توزیع تجمعی نرمال یا تابع توزیع لجستیک باشد، مدل فوق به ترتیب به مدل پروبیت یا مدل لوجیت تبدیل می‌شود. علاوه‌براین، مدل‌های توبیت و پواسن نیز حالت خاصی از مدل تک‌شاخص می‌باشند. برای اطلاعات بیشتر به هاردل و همکاران (۱۹۹۳) و هاردل و استوکر (۱۹۸۹) مراجعه کنید.

۲.۱ اهداف رساله

همان‌طور که قبلاً اشاره شد، مدل‌های نیمه‌پارامتری علاوه بر کاهش اریبی مدل‌های پارامتری، با مشکل مشقت بعد مدل‌های ناپارامتری نیز مقابله می‌کنند. یکی از این مدل‌های پرکاربرد، مدل جمعی خطی–جزئی

$$Y = \sum_{j\epsilon S_1} \beta_j X_j + \sum_{j\epsilon S_2} f_j(X_j) + \varepsilon \qquad (۱۲.۱)$$

است که در آن S_1 و S_2 دو زیر مجموعه مجزا از $\{1, ..., p\}$ است. فرض کنید تعداد متغیرهای توضیحی مؤثر بر پاسخ برابر با p باشد. در این‌صورت، مجموعه $S_1 \cup S_2$ دارای اندازه p_1 است. به‌عنوان مثال، اگر $p = 10$، $p_1 = 5$، $S_1 = \{1, 2, 3\}$ و $S_2 = \{4, 5\}$ باشد، آنگاه ۳ متغیر توضیحی اثر خطی و ۲ متغیر اثر غیرخطی در مدل‌بندی متغیر پاسخ دارند و ۵ متغیر دیگر بی‌اثر هستند. برای استفاده از مدل (۱۲.۱) لازم است اجزای خطی و غیرخطی مدل مشخص باشند. در تحلیل داده‌های واقعی، از پیش تعیین کردن مؤلفه‌های خطی و غیرخطی چندان ساده به نظر نمی‌رسد، به‌ویژه زمانی که تعداد متغیرهای توضیحی زیاد است. بنابراین، علاوه بر تشخیص مؤلفه‌های غیرصفر، شناسایی مؤلفه‌های خطی و غیرخطی نیز بسیار حائز اهمیت است. تاکنون بسیاری از محققین، با تکیه بر روش‌های جریمه، به موضوع انتخاب متغیر و تشخیص ساختار[۲۸] در مدل‌های جمعی خطی–جزئی پرداخته‌اند. از آن جمله می‌توان به ژانگ[۲۹] و همکاران (۲۰۱۱)، هوآنگ و همکاران (۲۰۱۲)، لیان[۳۰] (۲۰۱۲b)، لیان (۲۰۱۲c)، لیان و همکاران (۲۰۱۲)، لیان و همکاران (۲۰۱۵) اشاره کرد. لازم به ذکر است که منظور از تشخیص ساختار در مدل (۱۲.۱)، شناسایی متغیرهایی است که به صورت خطی یا غیرخطی وارد مدل می‌شوند. به عبارت دیگر، هدف تشخیص مجموعه‌های S_1 و S_2 است.

از طرفی، با پیشرفت فناوری و افزایش توان ذخیره‌سازی کامپیوترها، محققین می‌توانند با هزینه نسبتاً کمی داده‌هایی با "بعد بسیار بالا"[۳۱] در اختیار داشته باشند. تحلیل این نوع داده‌ها فرصت‌ها و چالش‌های جدیدی را پیش‌روی آماردانان قرار داده است. به دلیل وجود تعداد زیادی از متغیرها در مدل‌های با بعد بسیار بالا، تفسیر این مدل‌ها بسیار مشکل است. لذا مسئله انتخاب متغیر نقش بسیار مهمی را در تحلیل این نوع داده‌ها ایفا می‌کند.

[28] Structure identification
[29] Zhang
[30] Lian
[31] Ultrahigh dimensional

در این خصوص، مطالعات انجام شده فوق، که مبتنی بر توابع جریمه هستند، به دلایل هزینه محاسباتی، دقت آماری و ناپایداری الگوریتمی از کارآیی کافی برخوردار نیستند (فن[۳۲] و همکاران، ۲۰۰۹). لذا شناسایی متغیرهای مهم و تشخیص نوع تاثیر آن‌ها در متغیر پاسخ به لحاظ خطی یا غیرخطی بودن، کماکان مورد توجه متخصصین علوم مذکور و آماردانان است.

در این رساله با تمرکز بر انتخاب متغیر و تشخیص ساختار در داده‌های با بعد بسیار بالا، ابتدا از یک روش غربالگری مستقل مطمئن برای کاهش بعد مدل استفاده می‌شود. این روش بر اساس همبستگی فاصله‌ای بین متغیرهای توضیحی و تابع توزیع حاشیه‌ای متغیر پاسخ تعریف می‌شود و انتظار می‌رود وقتی که متغیر پاسخ دم سنگین، چوله یا دارای مقادیر فرین باشد، عملکرد خوبی داشته باشد. سپس از دو تابع جریمه برای تشخیص همزمان متغیرهای مهم و مؤلفه‌های خطی در زیر مدل غربال‌شده در مرحله قبل استفاده می‌شود. علاوه بر این ویژگی غربالگری مطمئن را برای روش غربالگری معرفی‌شده اثبات کرده و با استفاده از شبیه‌سازی و تحلیل داده‌های واقعی عملکرد این روش را با روش‌های موجود مقایسه می‌کنیم.

در بعضی از مسائل، اثر برخی از متغیرهای توضیحی بر پاسخ به صورت خطی و اثر متغیرهای دیگر تابعی از متغیر شاخص است. بنابراین لازم است به جای مدل (۱۲.۱) از مدل ضریب متغیر خطی–جزئی (۱۰.۱) استفاده شود. مشابه مدل جمعی خطی–جزئی، استفاده از مدل (۱۰.۱) بر فرض معلوم بودن اثرات ثابت و متغیر استوار است. در عمل، افراز این ضرایب، به دو جزء ثابت و متغیر، بسیار مشکل است. بنابراین لازم است روشی ارائه شود تا ضرایب صفر، ثابت غیر صفر و متغیر شناسایی شوند. مسئله انتخاب متغیر و تشخیص ساختار در مدل‌های ضریب متغیر خطی–جزئی تاکنون توسط افراد مختلف مورد بررسی قرار گرفته است. به‌عنوان مثال می‌توان به لنگ[۳۳] (۲۰۰۹)، هو و زیا[۳۴] (۲۰۱۲)، نه و کیلگوم[۳۵] (۲۰۱۲)، وانگ و سانگ[۳۶] (۲۰۱۳) اشاره کرد. با این حال، همه این روش‌ها بر رگرسیون میانگین یا تابع درستنمایی استوار هستند، که به شدت تحت تاثیر داده‌های پرت یا توزیع دم‌سنگین خطاها قرار می‌گیرند. بنابراین لازم است یک روش نیرومند برای انتخاب متغیر و تشخیص ساختار در مدل‌های ضریب متغیر خطی–جزئی ارائه شود. از طرفی، یائو[۳۷] و همکاران (۲۰۱۲) رگرسیون نما را معرفی کردند که، به جای میانگین، از نمای متغیر پاسخ به منظور مدل‌سازی استفاده می‌کند. بنابراین با توجه به ویژگی‌های شاخص نما، این روش در مقایسه با رگرسیون میانگین از نیرومندی بیشتری نسبت به داده‌های پرت و توزیع غیر نرمال خطاها برخوردار است.

در این رساله با استفاده از رگرسیون نما یک روش نیرومند برای انتخاب متغیر و تشخیص ساختار در مدل‌های ضریب متغیر خطی–جزئی معرفی می‌شود. عملکرد این روش با استفاده از مطالعه شبیه‌سازی و مثال واقعی مورد ارزیابی قرار می‌گیرد. علاوه بر این، ویژگی‌های مجانبی برآوردگر ارائه شده نیز ثابت می‌شود.

---

[32] Fan
[33] Leng
[34] Hu and Xia
[35] Noh and Keilegom
[36] Wang and Song
[37] Yao

۳.۱ ساختار رساله

- در فصل ۲ به معرفی برخی از روش‌های انتخاب متغیر مبتنی بر جریمه می‌پردازیم. در ادامه، با توجه به ناکارآمدی روش‌های جریمه برای انتخاب متغیر در داده‌های با بعد بسیار بالا، به توصیف برخی از روش‌های غربالگری برای کاهش بعد در این نوع داده‌ها خواهیم پرداخت. مسئله انتخاب متغیر با استفاده از توابع جریمه مستلزم حل یک مسئله بهینه‌سازی است. در انتهای این فصل حل این مسئله بهینه‌سازی مورد بحث قرار می‌گیرد.

- در فصل ۳، یک روش دو مرحله‌ای برای انتخاب متغیر و تشخیص ساختار در مدل‌های جمعی خطی–جزئی معرفی می‌شود. در مرحله اول، از یک روش غربالگری آزاد–مدل برای کاهش بعد فضای متغیرهای توضیحی استفاده می‌شود. این روش براساس ضریب همبستگی فاصله‌ای بین تابع توزیع متغیر پاسخ و متغیرهای توضیحی تعریف می‌شود. سپس در مرحله دوم، از دو تابع جریمه برای انتخاب متغیر و تشخیص ساختار استفاده می‌شود.

- در فصل ۴، مسئله انتخاب متغیر و تشخیص ساختار در مدل ضریب متغیر خطی–جزئی مورد بررسی قرار می‌گیرد. در این فصل با ترکیب رگرسیون نما و توابع جریمه، یک روش نیرومند برای برازش مدل ضریب متغیر خطی–جزئی ارائه می‌شود.

- در بخش ضمائم، اثبات قضایای مطرح‌شده در فصل‌های ۳ و ۴، مروری بر توابع اسپلاین و B –اسپلاین، گزیده‌ای از کدهای R و برخی از تعاریف و مفاهیم مورد نیاز در این رساله بیان می‌شوند.

فصل ۲

انتخاب متغیر در مدل‌های رگرسیونی

در این فصل، به معرفی رایج‌ترین روش‌های انتخاب متغیر پرداخته می‌شود. مطالب این فصل از منابع متعددی گرفته شده‌است که از آن جمله می‌توان به مونت‌گومری و همکاران (۲۰۱۲)، تیبشیرانی (۱۹۹۶)، فن و لی (۲۰۰۱)، هوانگ و همکاران (۲۰۱۲b) و فن و لیو (۲۰۰۸) اشاره کرد.

۱.۲ مقدمه

مدل رگرسیونی چندگانه زیر را در نظر بگیرید

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon \qquad (۱.۲)$$

که در آن $\mathbf{X} = (X_1, ..., X_p)^T$ بردار متغیرهای توضیحی، $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ بردار ضرایب رگرسیونی و ε خطای تصادفی با میانگین صفر و واریانس ثابت است. در این فصل، بدون از دست دادن کلیت، فرض می‌شود عرض از مبدا برابر با صفر است. یک روش متداول برآورد ضرایب رگرسیونی در مدل (۱.۲)، روش کمترین توان‌های دوم است. هنگامی‌که بین متغیرهای توضیحی همخطی وجود دارد، استنباط بر اساس این روش می‌تواند گمراه‌کننده باشد. در صورت وجود همخطی، برآوردگر کمترین توان‌های دوم منجر به برآوردهایی با واریانس بزرگ می‌شود که دقت پیشگویی را کاهش می‌دهد. یکی از عوامل ایجاد همخطی، زیاد بودن تعداد متغیرهای توضیحی در مدل است. در مدل‌های با بعد بالا، مسئله همخطی بسیار شدید است، یعنی هر یک از متغیرهای توضیحی را می‌توان به صورت ترکیب خطی

از سایر متغیرها نوشت (جیمز[1] و همکاران، ۲۰۱۳؛ فن و لی، ۲۰۰۸). روش پیشنهادی در این حالت، انتخاب زیر مجموعه‌ای از متغیرهای توضیحی برای ورود به مدل است. در این رهیافت علاقه‌مندیم زیرمجموعه‌ای از متغیرها را طوری انتخاب کنیم که رابطه بین متغیر پاسخ و متغیرهای توضیحی را به خوبی توصیف کند. یافتن زیر مجموعه‌ای مناسب از متغیرهای توضیحی برای ورود به مدل ''انتخاب متغیر'' نامیده می‌شود.

ساختن یک مدل رگرسیونی که تنها شامل زیرمجموعه‌ای از متغیرهای توضیحی است، دو واقعیت ناسازگار را در بر دارد: الف) علاقه‌مندیم که مدل، تا حد امکان، متغیرهای توضیحی بیشتری را شامل شود، به‌نحوی که اطلاعات موجود در این متغیرها بتواند در پیشگویی متغیر پاسخ مؤثر باشد. ب) می‌خواهیم مدل حتی الامکان دارای متغیرهای رگرسیونی کمتری باشد، چون با افزایش تعداد متغیرهای توضیحی در مدل، هزینه محاسبات افزایش می‌یابد و تفسیر مدل دشوارتر می‌شود. علاوه براین، وجود تعداد زیادی متغیر توضیحی ممکن است باعث بیش‌برازش[2] شود، یعنی مدل برازش‌شده برای داده‌های آموزشی[3]، که برای ساخت مدل از آن‌ها استفاده شده است، مناسب باشد، اما برای پیشگویی مشاهدات آینده بسیار ضعیف باشد. بنابراین مسئله انتخاب متغیر امری بسیار مهم در تحلیل داده‌هاست. فرآیند به‌دست آوردن یک مدل که سازگاری و توافق بین این دو واقعیت است، انتخاب بهترین معادله رگرسیونی نامیده می‌شود.

به‌عنوان رهیافت دیگر، می‌توان از روش‌هایی مانند رگرسیون مؤلفه‌های اصلی[4] یا رگرسیون کمترین توان‌های دوم جزیی[5] برای کاهش بعد استفاده کرد. در این روش‌ها، ابتدا p ترکیب خطی ناهمبسته از متغیرهای توضیحی تولید کرده و سپس M تا از این ترکیب‌ها به عنوان جایگزین متغیرهای اصلی استفاده می‌شود که $M < p$. به‌وضوح، اگر $M = p$، آنگاه برآورد به‌دست آمده با برآورد کمترین توان‌های دوم معادل است. استفاده از این مدل‌ها به منظور پیشگویی منطقی است، اما معمولاً تفسیر این مدل‌ها دشوار است (هستی و همکاران، ۲۰۰۹).

برای تعیین مدل پیشگو، لازم است پارامترهای مدل (ضرایب رگرسیونی) برآورد شوند. لذا فرآیندی که علاوه بر انتخاب متغیر و بهترین معادله رگرسیونی، برآورد ضرایب را نیز انجام دهد، مورد نظر است. روش کمترین توان‌های دوم جریمه‌شده[6] یک روش کارا برای برآورد پارامترها است که همزمان انتخاب متغیر را نیز انجام می‌دهد.

در این فصل، ابتدا خواص برآوردگر کمترین توان‌های دوم و تاثیر همخطی بر عملکرد این برآوردگر بررسی می‌شود. سپس به معرفی رایج‌ترین روش‌های انتخاب متغیر می‌پردازیم. این روش‌ها شامل روش‌های انتخاب متغیر کلاسیک، روش کمترین توان‌های دوم جریمه‌شده و روش غربالگری[7] هستند.

---

[1] James
[2] Overfitting
[3] Training data set
[4] Principal component regression
[5] Partial least squares regression
[6] Penalized least squares
[7] Screening

۲.۲ روش کمترین توان‌های دوم

فرض کنید $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ یک نمونه تصادفی از مدل (۱.۲) باشند. بنابراین مدل رگرسیون خطی چندگانه را می‌توان با نماد ماتریسی

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{۲.۲}$$

نوشت که در آن $\mathbf{Y} = (y_1, ..., y_n)^T$ بردار متغیر پاسخ، $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ ماتریس طرح در اندازه $n \times p$ با رتبه کامل ستونی، $\boldsymbol{\beta}$ بردار ضرایب رگرسیونی و $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$ بردار خطا است.

در رگرسیون خطی چندگانه فرض بر این است که رابطه بین متغیر پاسخ و متغیرهای توضیحی خطی است. بنابراین برای پیشگویی، لازم است ضرایب رگرسیونی برآورد شوند. این روش پارامتری هنگامی مناسب است که فرض خطی بودن رابطه میان متغیر پاسخ و متغیرهای توضیحی لااقل به صورت تقریبی برقرار باشد. اما هنگامی که رابطه میان متغیر پاسخ و متغیرهای توضیحی نامعلوم است، استفاده از رگرسیون خطی ممکن است منجر به نتایج گمراه کننده‌ای شود.

روش کمترین توان‌های دوم متداول‌ترین روش برآورد برای ضرایب رگرسیونی است که در سال ۱۸۰۹ توسط گاوس معرفی شد. به منظور یافتن برآوردگر کمترین توان‌های دوم معمولی (OLS)، لازم است برخی پذیره‌های زیربنایی در مدل (۲.۲) برقرار باشند که عبارتند از: الف) خطاها ناهمبسته‌اند. ب) $E(\boldsymbol{\varepsilon}) = 0$، ج) $E(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$ که σ مجهول، مثبت و متناهی است. برآوردگر OLS بردار ضرایب $\boldsymbol{\beta}$ به‌صورت

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{OLS} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|^2\} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\end{aligned} \tag{۳.۲}$$

است. به سادگی می‌توان نشان داد که اگر $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$، آنگاه $\hat{\boldsymbol{\beta}}^{OLS} \sim \mathcal{N}_p(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$. بنا به قضیه گاوس‌ـ مارکوف، $\hat{\boldsymbol{\beta}}^{OLS}$ بهترین برآوردگر نااریب خطی (BLUE) است. همچنین برآوردگر نااریب واریانس σ^2 بر اساس این برآوردگر به‌صورت

$$S^2 = \frac{1}{m}\| \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{OLS} \|^2, \qquad m = n - p$$

به‌دست می‌آید. با داشتن برآورد ضرایب رگرسیونی، بردار مقادیر برازش شده به‌صورت

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{OLS} = \mathbf{H}\mathbf{Y} \tag{۴.۲}$$

محاسبه می‌شود که در آن $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ ماتریس هَت نامیده می‌شود و یک ماتریس متقارن و خودتوان است. بنابراین $Var(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$ و بردار باقیمانده‌ها و واریانس آن، به ترتیب، به‌صورت $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ و $Var(\mathbf{e}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$ به‌دست می‌آیند. همان‌طور که ملاحظه می‌شود ماتریس $\mathbf{H}$ نقش بسیار مهمی در نتایج روش کمترین توان‌های دوم بازی می‌کند. برای اطلاعات بیشتر در این‌باره به نوروزی‌راد (۱۳۹۶) مراجعه کنید.

اگر داده‌ها از یک مدل خطی با پذیره‌های مطرح‌شده پیروی کنند، روش کمترین توان‌های دوم برآوردهای خوبی را ارائه می‌دهد. هنگامی‌که حداقل یکی از این پذیره‌ها برقرار نباشد، برآوردهای OLS بهینه نیستند.

۱.۲.۲ همخطی در رگرسیون چندگانه

یک مسئله جدی که کاربرد برآوردگر OLS را با مشکل مواجه می‌کند، وجود همخطی بین متغیرهای توضیحی است، بدین معنی که متغیرهای توضیحی به صورت کامل یا تقریباً کامل وابسته خطی باشند. اگر یکی از متغیرهای توضیحی تابعی خطی از یک یا چند متغیر توضیحی دیگر باشد، گوییم رگرسیون دارای همخطی کامل است. همخطی ناقص هنگامی اتفاق می‌افتد که یکی از متغیرهای توضیحی به‌طور تقریبی یک تابع خطی از یک یا چند متغیر توضیحی دیگر باشد. به‌عبارت دیگر، همخطی کامل هنگامی رخ می‌دهد که حداقل به‌ازای یک j $(j = ۱, ..., p)$ داشته باشیم $R_j^۲ = ۱$ و همخطی ناقص زمانی پیش می‌آید که حداقل به‌ازای یک j داشته باشیم $R_j^۲ \approx ۱$، که $R_j^۲$ ضریب تعیین در رگرسیون خطی متغیر X_j بر سایر متغیرهای توضیحی است. طبق تعریف می‌دانیم که ضریب تعیین $(R^۲)$ برابر با نسبت تغییرات بیان‌شده به تغییرات کل است. لذا در مدل رگرسیونی (۲.۲) ضریب تعیین برابر است با

$$R^۲ = \frac{\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{Y} - n\bar{Y}^۲}{\mathbf{Y}^T \mathbf{Y} - n\bar{Y}^۲}, \qquad \bar{Y} = \frac{۱}{n}\sum_{i=۱}^{n} Y_i. \tag{۵.۲}$$

وجود همخطی، تاثیرات جدی بر برآوردهای OLS دارد. برای درک بهتر موضوع، مدل رگرسیون خطی با دو متغیر توضیحی را در نظر بگیرید

$$Y = \beta_۱ X_۱ + \beta_۲ X_۲ + \varepsilon. \tag{۶.۲}$$

فرض کنید $X_۱$، $X_۲$ و Y استاندارد شده‌اند، به‌طوری‌که $\bar{X}_j = ۰$، $\sum_{i=۱}^{n}(X_{ij} - \bar{X}_j)^۲ = ۱$، $\bar{Y} = ۰$ و $\sum_{i=۱}^{n}(Y_i - \bar{Y})^۲ = ۱$. در این‌صورت معادلات نرمال روش کمترین توان‌های دوم به‌صورت زیر است

$$(\mathbf{X}^T \mathbf{X})\hat{\boldsymbol{\beta}}^{OLS} = \mathbf{X}^T \mathbf{Y}$$

یا به‌طور معادل

$$\begin{bmatrix} ۱ & r_{۱۲} \\ r_{۱۲} & ۱ \end{bmatrix} \begin{bmatrix} \hat{\beta}_۱ \\ \hat{\beta}_۲ \end{bmatrix} = \begin{bmatrix} r_{۱y} \\ r_{۲y} \end{bmatrix}$$

که $r_{۱۲}$ همبستگی ساده بین $X_۱$ و $X_۲$ و r_{jy} همبستگی ساده بین Y و X_j به‌ازای $j = ۱, ۲$ می‌باشند. معکوس $\mathbf{X}^T\mathbf{X}$ عبارتست از

$$C = (\mathbf{X}^T \mathbf{X})^{-۱} = \begin{bmatrix} \frac{۱}{۱-r_{۱۲}^۲} & \frac{-r_{۱۲}}{۱-r_{۱۲}^۲} \\ \frac{-r_{۱۲}}{۱-r_{۱۲}^۲} & \frac{۱}{۱-r_{۱۲}^۲} \end{bmatrix} \tag{۷.۲}$$

و برآوردهای ضرایب رگرسیونی به‌صورت

$$\hat{\beta}_۱ = \frac{r_{۱y} - r_{۱۲}r_{۲y}}{۱ - r_{۱۲}^۲}, \qquad \hat{\beta}_۲ = \frac{r_{۲y} - r_{۱۲}r_{۱y}}{۱ - r_{۱۲}^۲} \tag{۸.۲}$$

است. اگر همخطی ناقص بین $X_۱$ و $X_۲$ به گونه‌ای باشد که $r_{۱۲}$ بزرگ شود، آنگاه با توجه به روابط (۷.۲) و (۸.۲)، برآورد ضرایب و واریانس برآوردگرها بسیار بزرگ خواهد بود. به عبارت دیگر، اگر $|r_{۱۲}| \to ۱$، آنگاه، به‌ازای $j = ۱, ۲$، $Var(\hat{\beta}_j) = C_{jj}\sigma^۲ \to \infty$ و $Cov(\hat{\beta}_۱, \hat{\beta}_۲) = C_{۱۲}\sigma^۲ \to \infty$ که C_{jj} مؤلفه jام روی قطر اصلی ماتریس C است. بنابراین همخطی زیاد بین $X_۱$ و $X_۲$ منجر به واریانس‌ها و کوواریانس‌های بزرگ برای برآوردگرهای OLS خواهد شد. برای بیش از دو متغیر توضیحی نیز همخطی اثرات مشابهی ایجاد می‌کند. می‌توان نشان داد که در این حالت اعضای قطر اصلی ماتریس $C = (\mathbf{X}^T\mathbf{X})^{-۱}$ عبارتند از

$$C_{jj} = \frac{۱}{۱ - R_j^۲}, \qquad j = ۱, ..., p \tag{۹.۲}$$

که در آن $R_j^۲$ ضریب تعیین حاصل از رگرسیون خطی X_j بر $p - ۱$ متغیر توضیحی دیگر است. اگر همخطی شدید بین X_j و هر زیر مجموعه‌ای از $p - ۱$ متغیر دیگر وجود داشته باشد، در این‌صورت مقدار $R_j^۲$ نزدیک به یک خواهد بود. چون $Var(\hat{\beta}_j) = C_{jj}\sigma^۲ = \sigma^۲(۱ - R_j^۲)^{-۱}$، همخطی شدید موجب می‌شود که واریانس برآوردگر OLS ضرایب رگرسیونی $\boldsymbol{\beta}$ بسیار بزرگ شود. در این‌صورت $\hat{\boldsymbol{\beta}}^{OLS}$ برآورد دقیقی از $\boldsymbol{\beta}$ نیست. به‌عبارت دیگر، دقت برآورد ضرایب رگرسیونی بسیار کم می‌شود. همچنین همخطی باعث می‌شود که بردار $\hat{\boldsymbol{\beta}}^{OLS}$ دارای اعضایی باشد که از لحاظ قدر مطلق بسیار بزرگ هستند. علاوه‌براین، $\hat{\boldsymbol{\beta}}^{OLS}$ ناپایدار است، یعنی با حذف یا اضافه کردن یک متغیر توضیحی، یا با یک تغییر کوچک در مقادیر متغیرهای توضیحی، مقادیر $\hat{\boldsymbol{\beta}}^{OLS}$ به‌طور قابل توجهی تغییر می‌کنند.

در صورت وجود همخطی کامل، ماتریس $\mathbf{X}^T\mathbf{X}$ وارون‌پذیر نیست، زیرا در این حالت پررتبه نیست. بنابراین نمی‌توان با استفاده از معادلات نرمال، جواب یکتا برای برآورد ضرایب به‌دست آورد. لازم به ذکر است که به دلیل وجود خطای اندازه‌گیری، همخطی کامل معمولاً در عمل رخ نمی‌دهد، مگر هنگامی که کاربر به اشتباه در مدل رگرسیونی یکی از متغیرها را تابعی خطی از متغیر دیگر قرار دهد یا تعداد متغیرهای توضیحی بیشتر از تعداد مشاهدات باشد. برای اطلاعات بیشتر درباره منابع پیدایش همخطی به مونت‌گومری[۸] و همکاران (۲۰۱۲) مراجعه کنید.

بنابراین لازم است بدانیم که چگونه وجود همخطی را کشف کنیم تا مواظب پیامدهای ممکن باشیم. روش‌های مختلفی برای تشخیص همخطی وجود دارند. در اینجا به برخی از این روش‌ها اشاره می‌کنیم.

۱. عامل تورم واریانس (VIF): یکی از روش‌های معروف برای تشخیص همخطی استفاده از کمیت VIF است که به‌صورت

$$\text{VIF}_j = \frac{۱}{۱ - R_j^۲}, \qquad j = ۱, ..., p \tag{۱۰.۲}$$

[8] Montgomery

محاسبه می‌شود. واضح است که اگر X_j رابطه خطی قوی با سایر متغیرهای توضیحی داشته باشد، آنگاه R_j^2 نزدیک به یک بوده و VIF_j بزرگ می‌شود. اگر X_j بر $p-1$ متغیر توضیحی دیگر عمود باشد، آنگاه R_j^2 صفر بوده و VIF_j برابر با یک خواهد شد. بنابراین انحراف VIF_j از مقدار یک، انحراف از متعامد بودن و گرایش به همخطی را نشان می‌دهد. اگر هر یک از مقادیر VIF بیشتر از ۱۰ باشد، همخطی شدید بین متغیرهای توضیحی وجود دارد و ممکن است موجب بروز مشکلاتی در برآورد شود.

از آنجا که $Var(\hat{\beta}_j) = VIF_j\sigma^2$، لذا VIF_j میزان افزایش در واریانس $\hat{\beta}_j$ به‌واسطه ارتباط خطی X_j با بقیه متغیرهای توضیحی را نسبت به واریانس آن در صورت ناهمبسته بودن متغیر X_j با سایر متغیرهای توضیحی اندازه می‌گیرد. این مطلب دلیل نامگذاری این کمیت خاص را نشان می‌دهد.

۲. عدد شرطی[۹]: مقادیر ویژه ماتریس $\mathbf{X}^T\mathbf{X}$ می‌توانند برای اندازه‌گیری میزان همخطی مورد استفاده قرار گیرند. یک روش متداول برای کشف وجود همخطی، محاسبه عدد شرطی به‌صورت

$$\kappa = \sqrt{\frac{\max_{1\le j\le p}\lambda_j}{\min_{1\le j\le p}\lambda_j}} \qquad (۱۱.۲)$$

است که در آن $\lambda_1, ..., \lambda_p$ مقادیر ویژه ماتریس $\mathbf{X}^T\mathbf{X}$ هستند.

بلسلی[۱۰] و همکاران (۱۹۸۰) پیشنهاد کردند که اگر عدد شرطی برای ماتریس $\mathbf{X}^T\mathbf{X}$ بزرگتر از ۱۰ باشد، آنگاه در ماتریس طرح همخطی وجود دارد. اگر $۳۰ < \kappa < ۱۰۰$ باشد، همخطی شدید و چنانچه $\kappa > ۱۰۰$ باشد همخطی بسیار جدی است.

تاکنون روش‌های متعددی برای مقابله با همخطی بین متغیرهای توضیحی معرفی شده است. برای اطلاعات بیشتر در این باره به مونت‌گومری و همکاران (۲۰۱۲) و چاترجی و هادی[۱۱](۲۰۱۲) مراجعه کنید.

یک رهیافت مناسب برای مقابله با همخطی، استفاده از برآوردگرهای اریب مانند ریج[۱۲] (هورل و کنارد[۱۳]، ۱۹۷۰) است. در این روش با پذیرفتن مقدار کمی اریبی، برآوردگری حاصل می‌شود که واریانس آن در مقایسه با برآوردگر OLS کمتر است. این کاهش واریانس، کاهش MSE را به دنبال خواهد داشت. در نتیجه برآوردگر ریج در مقایسه با برآوردگر OLS پایدارتر است.

هنگامی که تعداد متغیرهای توضیحی زیاد است، همخطی امری اجتناب‌ناپذیر است. امروزه مجموعه داده‌هایی با تعداد زیادی از متغیرهای توضیحی در زمینه‌های مختلف علمی و تحقیقاتی مانند پزشکی، ژنتیک، بیوانفورماتیک، پردازش تصویر و مالی به وفور مورد استفاده قرار می‌گیرند. در این نوع داده‌ها، اطلاعات تعداد زیادی از متغیرهای توضیحی روی تعداد اندکی از واحدهای نمونه ثبت می‌شود $(p > n)$. ورود تعداد زیادی متغیر توضیحی به مدل اگرچه انعطاف‌پذیری مدل را افزایش می‌دهد، اما علاوه بر ایجاد همخطی شدید، باعث کاهش توان پیشگویی و مشکل‌شدن تفسیر مدل می‌شود. برآوردگر ریج اگرچه

[9] Condition number
[10] Belsley
[11] Chatterjee and Hadi
[12] Ridge
[13] Hoerl and Kennard

توان پیشگویی بالایی دارد، اما به دلیل تعدد ضرایب، مدل پیچیده‌ای را نتیجه می‌دهد که تفسیر آن بسیار مشکل است. این در حالی‌است که در مدل‌های با بعد بالا، معمولاً تنها تعداد کمی از متغیرهای توضیحی با متغیر پاسخ مرتبط هستند. بهترین روش در این حالت، حذف متغیرهای زائد و مدل‌بندی با زیرمجموعه‌ای از مؤثرترین متغیرهای توضیحی است. تاکنون روش‌های مختلفی برای انتخاب متغیر معرفی شده‌اند که در ادامه به معرفی مهمترین آنها می‌پردازیم.

۳.۲ انتخاب متغیر با معیارهای کلاسیک

تاکنون الگوریتم‌های متعددی برای انتخاب متغیر در مدل خطی (۲.۲) معرفی شده‌اند. این الگوریتم‌ها بر اساس معیارهایی مانند معیار اطلاع آکائیک (AIC)، معیار اطلاع بیزی (BIC)، اعتبارسنجی متقابل تعمیم‌یافته (GCV)، ضریب تعیین تعدیل‌شده (R^2_{adj})، مجموع توان‌های دوم باقیمانده‌ها (RSS) و C_p – مالو[۱۴] زیرمجموعه‌های متفاوتی از متغیرهای توضیحی را به‌عنوان بهترین زیرمجموعه انتخاب می‌کنند. از جمله این روش‌ها می‌توان انتخاب بهترین زیرمجموعه[۱۵]، انتخاب پیش‌رو[۱۶]، روش حذفی پس‌رو[۱۷] و روش گام به گام[۱۸] را نام برد.

در روش انتخاب بهترین زیرمجموعه، تمام زیرمجموعه‌های ممکن را در نظر گرفته و سپس بر اساس معیارهای فوق، بهترین زیر مجموعه از مدل کامل انتخاب می‌شود. در این حالت، اگر p تعداد کل متغیرهای توضیحی باشد، تعداد 2^p زیرمجموعه باید مورد بررسی قرار گیرند. به‌عنوان مثال، اگر $p = 10$، آنگاه باید ۱۰۲۴ مدل ممکن در نظر گرفته شوند و برای $p = 20$ باید بیش از یک میلیون مدل ممکن بررسی شوند. بنابراین با افزایش p، تعداد زیرمجموعه‌های ممکن به سرعت افزایش می‌یابد. لذا روش انتخاب بهترین زیر مجموعه از نظر محاسباتی، حتی در صورت استفاده از کامپیوترهای با سرعت بالا، برای $p > 40$ قابل استفاده نیست. در این راستا، دراپر و اسمیت[۱۹] (۱۹۸۱) روش‌های انتخاب پیش‌رو، حذفی پس‌رو و گام به گام را معرفی کردند. روش انتخاب پیش‌رو با مدلی که تنها شامل عرض از مبدا است، شروع می‌شود و سپس در هر گام متغیری که بیشترین تأثیر را در بهبود مدل داشته باشد، انتخاب می‌شود. این رویه تا زمانی ادامه می‌یابد که تأثیر هیچ یک از متغیرهای باقیمانده در بهبود مدل معنی‌دار نباشد. برای اندازه‌گیری میزان تأثیر هر یک از متغیرها در بهبود مدل، از آماره‌هایی مانند R^2، مجموع توان‌های دوم باقیمانده‌ها (RSS)، آماره F و آماره t استفاده می‌شود. به‌طور دقیق‌تر، در روش انتخاب پیش‌رو، برای انتخاب بهترین مدل بر اساس آماره F، به‌صورت زیر عمل می‌کنیم. اولین متغیر توضیحی که برای ورود به مدل نامزد می‌شود، متغیری است که دارای بیشترین همبستگی ساده با متغیر پاسخ است. این متغیر توضیحی همان متغیری است که بیشترین مقدار آماره F را برای آزمون معنی‌داری رگرسیون نتیجه می‌دهد. اگر آماره F متناظر با این متغیر از مقدار F از پیش تعیین‌شده F^*

[14] Mallows's C_p
[15] Best subset selection
[16] Forward selection
[17] Backward elimination
[18] Stepwise
[19] Draper and Smith

تجاوز کند، این متغیر توضیحی وارد مدل می‌شود. فرض کنیم اولین متغیر توضیحی وارد شده به مدل $X_۱$ باشد. دومین متغیر توضیحی منتخب برای ورود به مدل آن است که پس از تعدیل اثر $X_۱$ روی Y، اکنون بیشتر همبستگی را با Y دارد. به عبارت دیگر، دومین متغیر منتخب متغیری است که دارای بیشترین همبستگی جزیی با Y باشد. فرض کنیم دومین متغیر انتخاب‌شده $X_۲$ باشد. در این‌صورت بزرگترین آماره F جزیی عبارتست از

$$F = \frac{SSR(X_۲|X_۱)}{MSE(X_۱, X_۲)}$$

که $MSE(X_۱, X_۲)$ میانگین مربعات خطا و $SSR(X_۲|X_۱) = SSR(X_۱, X_۲) - SSR(X_۱)$ نشان‌دهنده میزان افزایش در مجموع مربعات رگرسیونی است که با افزودن متغیر توضیحی $X_۲$ به مدل حاصل می‌شود. این افزایش در مجموع مربعات رگرسیونی سهم اثرگذاری متغیر توضیحی $X_۲$ را به شرط حضور $X_۱$ در مدل اندازه‌گیری می‌کند. اگر آماره F جزیی از F^* بزرگتر باشد، $X_۲$ به مدل اضافه می‌شود. به‌طور کلی، در هر گام، متغیر توضیحی که دارای بیشترین همبستگی جزیی با Y است، یا به‌طور معادل دارای بیشترین آماره F جزیی به شرط حضور دیگر متغیرهای توضیحی در مدل است، برای ورود به مدل نامزد می‌شود. اگر آماره F جزیی آن متغیر از مقدار F^* تجاوز کند، متغیر مورد نظر وارد مدل می‌شود. این رویه هنگامی پایان می‌یابد که بزرگترین آماره F از مقدار F^* تجاوز نکند یا همه متغیرهای توضیحی وارد مدل شده باشند. در عمل معمولاً مقدار F^* برابر $F_{\alpha,۱,n-p}$ در نظر گرفته می‌شود که α بیانگر خطای نوع اول است.

در روش حذفی پس‌رو، فرآیند با مدل کامل شروع می‌شود و هر بار یک متغیر از مدل حذف می‌شود. متغیر حذف شده همان متغیری است که کمترین تأثیر را در بهبود مدل دارد. به عبارت دیگر، در هر گام، متغیر توضیحی با کوچکترین آماره F جزیی برای حذف از مدل نامزد می‌شود. اگر آماره F جزیی متغیر منتخب کمتر از مقدار F^* باشد، متغیر مورد نظر از مدل حذف می‌شود. فرآیند حذف متغیرها زمانی متوقف می‌شود که کوچکترین مقدار آماره F جزیی کوچکتر از مقدار از پیش تعیین شده F^* نباشد یا همه متغیرهای توضیحی از مدل حذف شده باشند.

در روش‌های انتخاب پیش‌رو و حذفی پس‌رو، هنگامی که یک متغیر حذف یا اضافه می‌شود، در مراحل بعدی معنی‌داری این متغیرها مجددا بررسی نمی‌شود. روش گام به گام تعدیل روش انتخاب پیش‌رو می‌باشد که در هر گام همه متغیرهای توضیحی وارد شده به مدل در گام‌های قبلی، مجددا ارزیابی می‌شوند. بنابراین یک متغیر توضیحی اضافه شده در گام قبلی ممکن است در گام بعدی از مدل حذف شود.

برای اطلاعات بیشتر درباره سایر روش‌های انتخاب متغیر می‌توان به دزیاک[۲۰] و همکاران (۲۰۰۵)، هاینز[۲۱] و همکاران (۲۰۱۸)، دسبولتس[۲۲] (۲۰۱۸) و دینگ[۲۳] (۲۰۱۸) مراجعه کرد.

اگرچه روش‌های انتخاب زیرمجموعه با معیارهای کلاسیک از ویژگی‌های نمونه‌ای مطلوبی برخوردارند، اما این روش‌ها پایدار نیستند، بدین معنی که یک تغییر بسیار کوچک در داده‌ها ممکن است مدل‌های

[20] Dziak
[21] Heinze
[22] Desboulets
[23] Ding

کاملاً متفاوت را نتیجه دهد. لذا ممکن است این روش‌ها پیشگویی نامناسبی را ایجاد کنند (بریمن، ۱۹۹۶). مشکل دیگر این روش‌ها این است که تمامی مدل‌های ممکن را در نظر نمی‌گیرند و این موضوع ممکن است منجر به از دست دادن یک مدل خوب شود. در روش‌های انتخاب پیش‌رو و حذفی پس‌رو، حداکثر تعداد $p+(p-۱)+\cdots+۱$ مدل مورد بررسی قرار می‌گیرند. به‌عنوان مثال، برای $p=۱۵$، نسبت ماکزیمم تعداد مدل‌های مورد بررسی به تعداد کل مدل‌های ممکن برابر با

$$\frac{۱۵+۱۴+\cdots+۱}{۲^{۱۵}-۱}=\frac{۱۲۰}{۳۲۷۶۷}\approx ۰/۴٪ \qquad (۱۲.۲)$$

است، یعنی ۹۹/۶٪ از مدل‌های ممکن برای انتخاب بهترین مدل در نظر گرفته نمی‌شوند. بنابراین مدل نهایی انتخاب شده ممکن است بهترین مدل نباشد. از طرفی در صورت وجود همخطی بین متغیرهای توضیحی، استفاده از روش‌های مذکور ممکن است منجر به انتخاب مدل مناسبی نشود (دزیاک و همکاران، ۲۰۰۵). علاوه بر این، زمانی که p بزرگ باشد، هزینه محاسباتی این روش‌ها بسیار بالاست. برای رفع این چالش‌ها، محققین رگرسیون جریمه‌شده را معرفی کردند که، با انتخاب یک تابع جریمه مناسب، دو عمل انتخاب متغیر و برآورد ضرایب را به‌طور همزمان انجام می‌دهد.

۴.۲ انتخاب متغیر با کمترین توان‌های دوم جریمه‌شده

در این بخش، به توصیف روش کمترین توان‌های دوم جریمه‌شده می‌پردازیم. بدیهی است وقتی تعداد متغیرهای توضیحی p زیاد است، اندازه مجموع قدر مطلق ضرایب β_j، $j=۱,...,p$، افزایش می‌یابد. از آنجایی که در روش کمترین توان‌های دوم به دنبال مینیمم کردن $\|\varepsilon\|^۲=\varepsilon^T\varepsilon$ نسبت به $\boldsymbol{\beta}$ هستیم، پس چنانچه بخواهیم علاوه بر برآورد ضرایب رگرسیونی، تعداد متغیرهای توضیحی مدل را نیز کاهش دهیم، کافی‌است بزرگی قدرمطلق ضرایب یا تابعی از آن را نیز به‌طور همزمان مینیمم کنیم. بنابراین، در رگرسیون کمترین توان‌های دوم جریمه‌شده، محک $\|\varepsilon\|^۲$ را تحت این قید که تابعی از قدرمطلق ضرایب رگرسیونی مانند $\sum_{j=۱}^{p}p(|\beta_j|)$ نباید بزرگ باشد، مینیمم می‌کنند. این یک مسئله بهینه‌سازی مقید[۲۴] است که به‌صورت زیر بیان می‌شود

$$\min_{\boldsymbol{\beta}}\|\varepsilon\|^۲ \qquad \text{s.t} \qquad \sum_{j=۱}^{p}p(|\beta_j|)\le s. \qquad (۱۳.۲)$$

یکی از راه‌های حل مسئله بهینه‌سازی مقید، استفاده از روش لاگرانژ[۲۵] است که با استفاده از این روش، مسئله بهینه‌سازی (۱۳.۲) را می‌توان به صورت زیر بیان کرد

$$\begin{aligned}\hat{\boldsymbol{\beta}}&=\arg\min_{\boldsymbol{\beta}}\left\{\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^۲+\lambda\sum_{j=۱}^{p}p(|\beta_j|)\right\}\\&=\arg\min_{\boldsymbol{\beta}}\left\{\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^۲+\sum_{j=۱}^{p}p_\lambda(|\beta_j|)\right\}.\end{aligned} \qquad (۱۴.۲)$$

[24] Constraint optimization

[25] Lagrange method

با دقت در ساختار مسئله (۱۴.۲) دیده می‌شود که به تابع مجموع توان‌های دوم خطا در رگرسیون معمولی، عبارت جریمه $\sum_{j=1}^{p} p_\lambda(|\beta_j|)$ اضافه شده است. در اینجا، $p_\lambda(\cdot)$ تابع جریمه و $\lambda \geq 0$ پارامتر جریمه است که میزان جریمه را برحسب بزرگی ضرایب رگرسیونی کنترل می‌کند. در این حالت بین λ در (۱۴.۲) و s در (۱۳.۲) رابطه‌ای معکوس وجود دارد. به‌طور دقیق‌تر، اگر بخواهیم تعداد متغیرهای بیشتری از مدل حذف شوند، کافی است s را کوچک یا λ را بزرگ انتخاب کنیم و برعکس. پارامتر تنظیم‌کننده λ موازنه‌ای بین خطای پیشگویی و پیچیدگی مدل برقرار می‌کند. جواب مسئله بهینه‌سازی (۱۴.۲) به شدت به مقدار λ وابسته است و لذا تعیین مقدار مناسب برای λ یک مسئله چالش‌برانگیز است. روش‌های مختلف انتخاب مقدار بهینه λ در زیربخش ۳.۴.۲ آمده‌اند.

بنابراین با انتخاب یک مقدار بهینه برای λ، مقادیر برخی از پارامترها ممکن است صفر شوند که در این‌صورت متغیر توضیحی متناظر با آن پارامتر، به عنوان متغیر بی‌اثر شناخته شده و بدین‌ترتیب مجموعه متغیرهای مهم و بی‌اهمیت از یکدیگر تشخیص داده می‌شوند و در اصطلاح ''انتخاب متغیر'' صورت می‌گیرد.

۱.۴.۲ انواع توابع جریمه

تا کنون توابع جریمه زیادی توسط افراد مختلف معرفی شده‌اند که برخی از آن‌ها دو عمل انتخاب متغیر و برآورد را به‌طور همزمان انجام می‌دهند. در این بخش به معرفی برخی از توابع جریمه معروف می‌پردازیم.

۱. جریمه L_q: این تابع جریمه، رگرسیون بریج را نتیجه می‌دهد که توسط فرانک و فریدمن[۲۶] در سال ۱۹۹۳ معرفی شد. برآوردگر بریج به صورت

$$\hat{\boldsymbol{\beta}}^{Bridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid^q \right\}, \qquad 0 < q \leq 2 \qquad (۱۵.۲)$$

تعریف می‌شود. این تابع جریمه، برخی از توابع جریمه مهم را به عنوان یک حالت خاص شامل می‌شود.

حالت خاص $q = 2$

مدل رگرسیونی حاصل را در این حالت رگرسیون ریج گویند و برآورد پارامترها به‌صورت

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{Ridge} &= \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid^2 \right\} \\ &= \left(\mathbf{X}^T\mathbf{X} + K\mathbf{I}_p\right)^{-1} \mathbf{X}^T\mathbf{Y} \end{aligned}$$

به‌دست می‌آید که $K = 2\lambda$ را پارامتر تنظیم‌کننده یا پارامتر ریج گویند. یک مزیت رگرسیون ریج پیاده‌سازی آسان و شکل صریح جواب آن است.

[26] Frank and Friedman

در برآورد کمترین توان‌های دوم معمولی، اگر ماتریس $\mathbf{X}$ دارای رتبه کامل ستونی نباشد، ماتریس $\mathbf{X}^T\mathbf{X}$ وارون‌پذیر نخواهد بود و لذا جواب یکتایی برای برآوردگر OLS وجود نخواهد داشت. اما برای هر ماتریس دلخواه $\mathbf{X}$ و $K > 0$، ماتریس $\mathbf{X}^T\mathbf{X} + K\mathbf{I}_p$ همواره وارون‌پذیر است و در نتیجه پاسخ یکتا برای برآوردگر ریج وجود دارد. برآوردگر ریج اریب است، اما در مقایسه با برآوردگر OLS دارای واریانس کوچکتری است. همچنین همواره λای وجود دارد به‌طوری‌که $MSE(\hat{\boldsymbol{\beta}}^{Ridge}) < MSE(\hat{\boldsymbol{\beta}}^{OLS})$ (برهنی و هوانگ[۲۷]، ۲۰۱۹). بنابراین رگرسیون ریج پیشگویی بهتری از رگرسیون کمترین توان‌های دوم معمولی ارائه می‌کند. با این وجود، برآوردگر ریج اگرچه ضرایب رگرسیونی را منقبض می‌کند، اما هیچ یک از برآوردهای ضرایب دقیقاً برابر صفر نیستند، یعنی این روش توانایی انتخاب متغیر ندارد. این موضوع خللی در پیشگویی ایجاد نمی‌کند، اما با توجه به حضور همه متغیرها در مدل، تفسیر مدل برازش‌شده به‌سادگی امکان‌پذیر نیست.

در حالتی که $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$ ماتریس طرح متعامد خواهد بود. در این حالت، برآوردگر ریج به‌صورت

$$\hat{\boldsymbol{\beta}}^{Ridge} = \frac{1}{1+K}\hat{\boldsymbol{\beta}}^{OLS} \qquad (۱۶.۲)$$

در می‌آید. این رابطه ماهیت انقباضی رگرسیون جریمه‌شده را نشان می‌دهد، چون با افزایش K ضرایب به سمت صفر منقبض می‌شوند. هنگامی‌که $K = 0$، برآوردگرهای OLS و ریج معادل‌اند.

حالت خاص $q = 1$

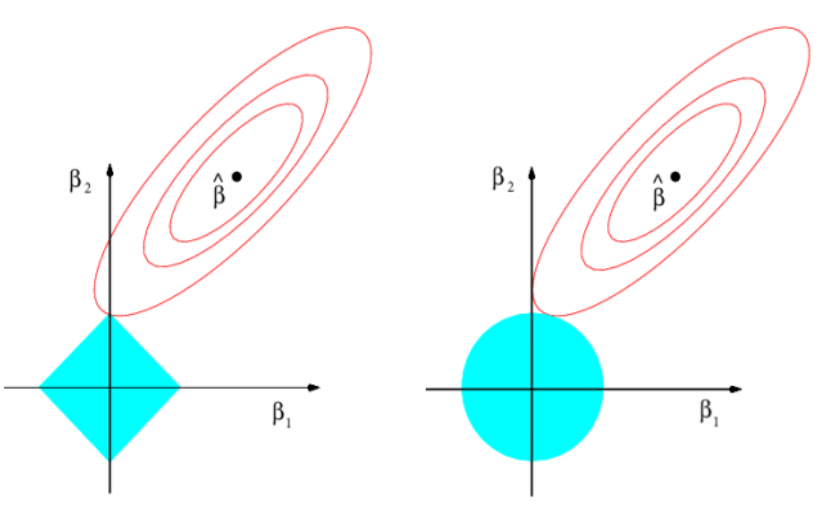
مدل رگرسیونی حاصل در این حالت را رگرسیون لاسو (Lasso) گویند که توسط تیبشیرانی[۲۸] در سال ۱۹۹۶ معرفی شد. از دیدگاه آماری، لاسو به معنی عملگر انتخاب کننده، منقبض کننده و مینیمم‌کننده از جنس قدرمطلق است. در رگرسیون لاسو برآورد ضرایب رگرسیونی به صورت

$$\hat{\boldsymbol{\beta}}^{Lasso} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \right\} \qquad (۱۷.۲)$$

به‌دست می‌آید که در آن λ پارامتر جریمه است و سطح تنکی (تعداد ضرایب صفر) را کنترل می‌کند. وقتی $\lambda \to 0$، آنگاه $\hat{\boldsymbol{\beta}}^{Lasso} \to \hat{\boldsymbol{\beta}}^{OLS}$ و وقتی $\lambda \to \infty$، آنگاه $\hat{\boldsymbol{\beta}}^{Lasso} \to 0$. به‌عبارت دیگر، هرچه جریمه بزرگتری به‌کار رود، تعداد بیشتری از ضرایب به سمت صفر منقبض می‌شوند. برآورد به‌دست آمده با رگرسیون لاسو تنک است، یعنی برآوردهای برخی از ضرایب برابر صفر هستند. به‌عبارت دیگر، این روش، انتخاب متغیر و برآورد پارامتر را به‌طور همزمان انجام می‌دهد.

تابع جریمه قدرمطلق باعث می‌شود که جواب صریحی برای برآوردگر لاسو وجود نداشته باشد. تیبشیرانی این برآوردگر را از طریق برنامه‌ریزی درجه دوم[۲۹] به دست آورده بود. سپس افرون[۳۰] و همکاران

[27] Breheny and Huang
[28] Tibshirani
[29] Quadaratic programming
[30] Efron



شکل ۱.۲: منحنی‌های تراز و نواحی محدودیت رگرسیون لاسو (قاب سمت چپ) و رگرسیون ریج (قاب سمت راست). ناحیه محدودیت لاسو به‌صورت $|\beta_۱| + |\beta_۲| \leq s$ و ناحیه محدودیت ریج به‌صورت $\beta_۱^۲ + \beta_۲^۲ \leq s$ است. بیضی‌ها، منحنی‌های تراز تابع خطای برآورد کمترین توان‌های دوم هستند.

(۲۰۰۴) رگرسیون کمترین زاویه[۳۱] را معرفی کردند که برآوردهای لاسو را نتیجه می‌دهد و هزینه محاسباتی آن با برآوردهای کمترین توان‌های دوم برابر است.

برای روشن‌تر شدن مفهوم انتخاب متغیر توسط تابع جریمه لاسو، حالت خاصی از مدل رگرسیونی را به صورت زیر در نظر بگیرید

$$Y = \beta_۱ X_۱ + \beta_۲ X_۲ + \varepsilon.$$

بنابراین مجموع توان‌های دوم خطا تابعی از دو پارامتر $\beta_۱$ و $\beta_۲$ است

$$Q(\boldsymbol{\beta}) = \| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|^۲ = \sum_{i=۱}^{n} e_i^۲ = \sum_{i=۱}^{n} (Y_i - \beta_۱ X_{i۱} - \beta_۲ X_{i۲})^۲. \qquad (۱۸.۲)$$

نمودار تابع $Q(.)$، یک رویه درجه دوم به صورت

$$Q(\boldsymbol{\beta}) = a_۰ + a_۱\beta_۱ + a_۲\beta_۲ + a_۳\beta_۱^۲ + a_۴\beta_۲^۲ + a_۵\beta_۱\beta_۲ \qquad (۱۹.۲)$$

است که منحنی‌های تراز آن در شکل ۱.۲ نشان داده شده است. اکنون شرط‌های مینیمم‌سازی ریج و لاسو را به صورت

$$|\beta_۱| + |\beta_۲| = s \qquad \text{و} \qquad \beta_۱^۲ + \beta_۲^۲ = s \qquad (۲۰.۲)$$

در نظر بگیرید که شکل هندسی آن‌ها به‌صورت دایره برای رگرسیون ریج (قاب سمت راست) و لوزی برای رگرسیون لاسو (قاب سمت چپ) نشان داده شده است.

[31] Least angle regression

با توجه به قاب سمت چپ شکل ۱.۲، در رگرسیون لاسو حالتی می‌تواند رخ دهد که یکی از ضرایب صفر شده و دیگری همواره برابر s باشد، در صورتی‌که در قاب سمت راست این امکان وجود ندارد. به عبارت دقیق‌تر، در قاب سمت راست، اگر یکی از ضرایب صفر شود، آنگاه دیگری می‌تواند $\sqrt{s}$، و نه s، باشد. بنابراین برای لاسو امکان صفر برآورد شدن ضرایب وجود دارد، اما برای ریج این اتفاق هرگز رخ نمی‌دهد. در صورت صفر بودن برآوردهای هر یک از ضرایب در رگرسیون لاسو، متغیر متناظر با آن ضریب به عنوان متغیر بی‌اهمیت شناخته می‌شود.

لازم به ذکر است که در هر دو روش لاسو و ریج، با انتخاب مقدار کوچک s، هر دو پارامتر β_1 و β_2 (در حالت دو بعدی) تا حد ممکن می‌توانند کوچک شوند. بنابراین هر دو روش ریج و لاسو انقباضی هستند.

اگر ماتریس طرح متعامد باشد، آنگاه می‌توان برآوردگر لاسو را به عنوان تابعی از برآوردگر کمترین توان‌های دوم معمولی β_j به‌صورت

$$\hat{\beta}_j^{Lasso} = sgn(z_j)(|z_j| - \lambda)_+, \qquad j = 1, ..., p \tag{۲۱.۲}$$

به‌دست آورد که در آن $sgn(x) = I(x > 0) - I(x < 0)$، $I(\cdot)$ تابع نشانگر، $x_+ = \max\{0, x\}$ و $z_j = \mathbf{X}_j^T\mathbf{Y}$ برآورد OLS ضریب β_j است که $\mathbf{X}_j$، ستون j ام ماتریس طرح $\mathbf{X}$ است (فن و لی، ۲۰۰۱). برآوردگر (۲۱.۲) را به صورت

$$\hat{\beta}_j^{Lasso} = S(z_j|\lambda) \tag{۲۲.۲}$$

نشان می‌دهیم که $S(.|\lambda)$ عملگر آستانه ای نرم نامیده می‌شود. رابطه (۲۱.۲) در بررسی ویژگی‌های مطلوب رگرسیون لاسو مفید است که در پایان این بخش به این موضوع خواهیم پرداخت.

رگرسیون لاسو با اینکه در بسیاری از مسایل انتخاب متغیر عملکرد خوبی از خود نشان داده، اما دارای برخی محدودیت‌ها است. به‌عنوان مثال، لاسو برای یک مدل رگرسیونی خطی با p متغیر پیشگو و n مشاهده، حداکثر n متغیر را انتخاب می‌کند. بنابراین اگر تعداد متغیرهای توضیحی معنی‌دار در مدل بیشتر از n باشد، برخی از آن‌ها توسط لاسو انتخاب نمی‌شوند. لاسو در بین یک مجموعه از متغیرهای به شدت همبسته، فقط یک متغیر را انتخاب می‌کند (زو و هستی، ۲۰۰۵). برای حالت معمولی $n > p$، اگر همبستگی بالایی بین متغیرهای پیشگو وجود داشته باشد، کارآیی پیشگویی برآوردگر لاسو به خوبی برآوردگر ریج نیست (تیبشیرانی، ۱۹۹۶). رگرسیون لاسو برآوردگری را نتیجه می‌دهد که اریبی آن زیاد است. در حالتی که ماتریس طرح متعامد است، می‌توان نشان داد که برای $j = 1, ..., p$ داریم

$$\begin{cases} E|\hat{\beta}_j^{Lasso} - \beta_j| = 0, & \beta_j = 0, \\ E|\hat{\beta}_j^{Lasso} - \beta_j| \approx \beta_j, & \beta_j \in [0, \lambda], \\ E|\hat{\beta}_j^{Lasso} - \beta_j| \approx \lambda, & |\beta_j| > \lambda. \end{cases} \tag{۲۳.۲}$$

بنابراین مقدار اریبی برآوردگر لاسو برای ضرایب بزرگ، نزدیک به λ است (برهنی و هوانگ، ۲۰۱۹).

برای از بین بردن این محدودیت‌ها، توابع جریمه متعددی معرفی شدند که در ادامه به برخی از آن‌ها اشاره می‌کنیم. قبل از معرفی این توابع جریمه، ابتدا ویژگی پیشگویی را تعریف می‌کنیم.

ویژگی پیشگویی: فرض کنید $\mathcal{A} = \{j : \beta_j \neq ۰\}$ مجموعه اندیس ضرایب غیرصفر باشد، به‌طوری که $|\mathcal{A}| = p_۰ < p$. در این‌صورت، مدل درست تنها به زیرمجموعه‌ای از متغیرها (با اندازه $p_۰$) وابسته است. منظور از مدل درست مدلی است که تنها شامل متغیرهای مؤثر بر پاسخ باشد. بر اساس پیشنهاد فن و لی (۲۰۰۱)، یک روش برآورد ضرایب دارای ویژگی پیشگویی[۳۲] است، اگر به‌طور مجانبی دارای خواص زیر باشد:

۱. بتواند ضرایب غیرصفر را به‌درستی شناسایی کند، یعنی اگر $\hat{\mathcal{A}} = \{j : \hat{\beta}_j \neq ۰\}$، آنگاه $\lim_{n\to\infty} P(\hat{\mathcal{A}} = \mathcal{A}) = ۱$.

۲. دارای نرخ برآورد بهینه باشد. به‌عبارت دیگر

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{D} \mathcal{N}(۰, \Sigma^*),$$

که Σ^* ماتریس کوواریانس زیر مدل درست است.

بنابراین مطلوب است که یک برآوردگر دارای ویژگی پیشگویی باشد. زو[۳۳] (۲۰۰۶) نشان داد که برآوردگر لاسو دارای ویژگی پیشگویی نیست. با توجه به رابطه (۲۳.۲) ملاحظه می‌شود که مقدار اریبی برآورد لاسو توسط λ تعیین می‌شود. بنابراین یک روش کاهش اریبی لاسو، استفاده از جریمه وزنی به‌صورت $\lambda_j = w_j\lambda$ است. اگر بتوانیم وزن‌ها را طوری انتخاب کنیم که به ضرایب بزرگتر جریمه کوچکتری تخصیص داده شود، آنگاه با حفظ ویژگی تنکی، اریبی برآوردگر لاسو کاهش می‌یابد. بنابراین زو (۲۰۰۶) لاسوی تطبیقی را معرفی کرد و نشان داد که این روش از ویژگی پیشگویی برخوردار است.

۲. جریمه لاسوی تطبیقی: صورت کلی تابع جریمه لاسوی تطبیقی مانند لاسو است، اما برخلاف لاسو برای ضرایب مختلف، جریمه‌های مختلفی در نظر می‌گیرد. برآوردگر لاسوی تطبیقی به صورت زیر بیان می‌شود

$$\hat{\boldsymbol{\beta}}^{aLasso} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{۱}{۲} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^۲ + \lambda \sum_{j=۱}^{p} w_j \mid \beta_j \mid \right\} \qquad (۲۴.۲)$$

که w_j وزن β_j است تا ضرایب کوچک را بیشتر از ضرایب بزرگ جریمه کند. معمولاً در رگرسیون لاسوی تطبیقی، وزن $w_j = \frac{۱}{|\hat{\beta}_j^\circ|^\gamma}$ را انتخاب می‌کنند که $\gamma > ۰$ و $\hat{\beta}_j^\circ$ یک برآوردگر $\sqrt{n}$ سازگار از β_j مانند برآوردگر OLS، ریج یا لاسو است.

۳. جریمه SCAD: فن و لی (۲۰۰۱) یک تابع جریمه محدب به نام SCAD را معرفی کردند و نشان دادند که این تابع جریمه، از همه ویژگی‌های مطلوب یک تابع جریمه برخوردار است (زیربخش ۲.۴.۲ را ببینید). آن‌ها همچنین ویژگی پیشگویی را نیز برای این روش ثابت کردند. برآوردگر SCAD به صورت

[32] Oracle property
[33] Zou

$$\hat{\boldsymbol{\beta}}^{SCAD} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{۱}{۲} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^{۲} + \sum_{j=۱}^{p} p_{\lambda}(|\beta_j|) \right\} \qquad (۲۵.۲)$$

تعریف می‌شود که $p_{\lambda}(\cdot)$ تابع جریمه SCAD است و به صورت

$$p_{\lambda}(t) = \begin{cases} \lambda t, & ۰ \leq t \leq \lambda \\ \frac{۲a\lambda t - t^{۲} - \lambda^{۲}}{۲(a-۱)}, & \lambda < t < a\lambda \\ \frac{\lambda^{۲}(a+۱)}{۲}, & t \geq a\lambda \end{cases} \qquad (۲۶.۲)$$

تعریف می‌شود که $t \geq ۰$، $p_{\lambda}(۰) = ۰$، λ پارامتر جریمه و $a > ۲$ مقداری ثابت است. فن و لی (۲۰۰۱) مقدار $a = ۳/۷$ را پیشنهاد کردند و نشان دادند که این مقدار در شرایط مختلف عملکرد خوبی را نتیجه می‌دهد.

شکل ۲.۲ نمودار تابع SCAD را به‌ازای $\lambda = ۱$ نشان می‌دهد. با توجه به شکل ۲.۲ و رابطه (۲۶.۲) ملاحظه می‌شود که در فاصله $|\beta| \leq \lambda$ تابع جریمه SCAD بر لاسو منطبق است، و پس از آن، تا زمانی‌که $|\beta| \leq a\lambda$، جریمه SCAD یک تابع درجه دوم است. سپس به ازای $|\beta| > a\lambda$ به یک تابع ثابت تبدیل می‌شود. همچنین، با توجه به (۲۶.۲)، وقتی $a \to \infty$، تابع جریمه SCAD با تابع جریمه لاسو معادل است، یعنی تابع جریمه SCAD در حالت خاص $a = \infty$ برآوردگر لاسو را نتیجه می‌دهد.

با فرض متعامد بودن ماتریس طرح، برآوردهای SCAD ضرایب را می توان به صورت

$$\hat{\beta}_j^{SCAD}(a, \lambda) = \begin{cases} S(z_j|\lambda), & |z_j| \leq ۲\lambda \\ \frac{a-۱}{a-۲} S(z_j|\frac{a\lambda}{a-۱}), & ۲\lambda < |z_j| \leq a\lambda \\ z_j, & |z_j| > a\lambda \end{cases} \qquad (۲۷.۲)$$

نشان داد که z_j برآورد OLS ضریب β_j و $S(.|\lambda)$ عملگر آستانه‌ای نرم است (فن و لی، ۲۰۰۱).

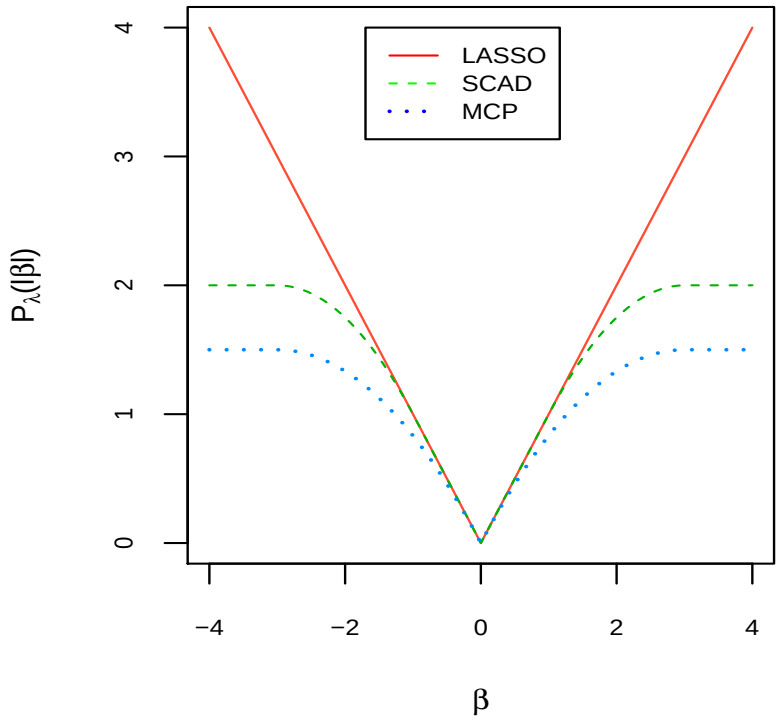
۴. جریمه MCP: ژانگ (۲۰۱۰) تابع جریمه دیگری به نام MCP را به صورت

$$p_{\lambda}(t) = \lambda \int_{۰}^{t} \left(۱ - \frac{x}{\gamma\lambda} \right)_{+} dx, \qquad t \geq ۰ \qquad (۲۸.۲)$$

معرفی کرد که در آن λ پارامتر جریمه و $\gamma > ۰$ پارامتر تنظیم‌کننده است که میزان مقعر بودن تابع $p_{\lambda}(\cdot)$ را کنترل می کند. در این حالت، برآوردگر MCP به‌صورت

$$\hat{\boldsymbol{\beta}}^{MCP} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{۱}{۲} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^{۲} + \sum_{j=۱}^{p} p_{\lambda}(|\beta_j|) \right\} \qquad (۲۹.۲)$$

تعریف می‌شود. تابع جریمه MCP به ازای $\gamma = \infty$ تابع جریمه لاسو را نتیجه می‌دهد.



شکل ۲.۲: نمودار توابع جریمه لاسو، SCAD و MCP.

تحت فرض متعامد بودن ماتریس طرح، مشابه برآوردهای لاسو و SCAD، برآوردهای MCP ضرایب را می‌توان به صورت

$$\hat{\beta}_j^{MCP}(\gamma,\lambda) = \begin{cases} \frac{\gamma}{\gamma-1} S(z_j|\lambda), & |z_j| \le \gamma\lambda \\ z_j, & |z_j| > \gamma\lambda \end{cases}$$

نشان داد (برهنی و هوانگ، ۲۰۱۹). شکل ۲.۲ نمودار تابع MCP را به‌ازای $\lambda = ۱$ و $\gamma = ۳$ نشان می‌دهد. تابع جریمه MCP در فاصله $|\beta| \le \gamma\lambda$ یک تابع درجه دوم و پس از آن یک تابع ثابت است.

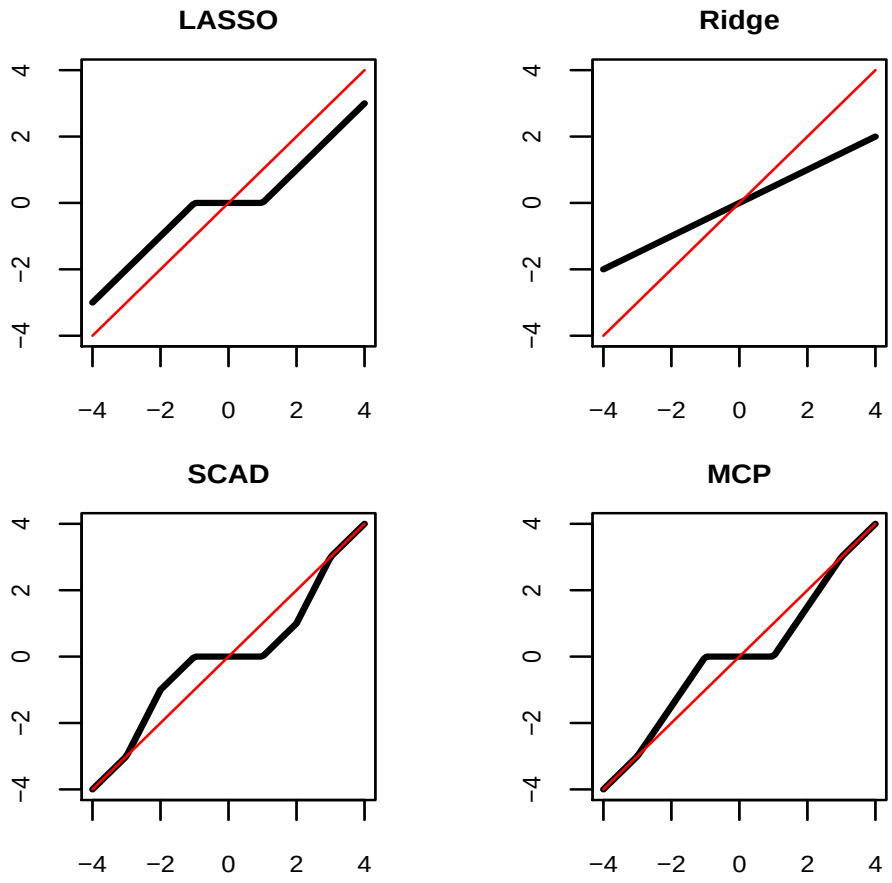
۲.۴.۲ ویژگی‌های یک تابع جریمه مناسب

برای استفاده از رگرسیون جریمه‌شده، سوالی که پیش می‌آید این است که از چه نوع تابع جریمه باید استفاده کرد. فن و لی (۲۰۰۱) نشان دادند که یک تابع جریمه خوب باید برآوردگری با سه ویژگی مطلوب زیر را نتیجه دهد:

۱. تنکی: برآوردگر نتیجه‌شده به‌طور خودکار ضرایب برآورد شده‌ای که مقدار کوچکی دارند را برابر صفر قرار دهد تا متغیرهای مناسب انتخاب شوند. این کار پیچیدگی مدل را کاهش می‌دهد.

۲. نا اریبی: برآوردگر به دست آمده برای ضرایب رگرسیونی که مقادیر واقعی آن‌ها بزرگ است، تقریباً ناریب باشد. این ویژگی اریبی مدل را کاهش می‌دهد.

۳. پیوستگی: برآوردگر نتیجه‌شده پیوسته باشد تا بتواند ناپایداری در پیشگویی مدل را کاهش دهد.



شکل ۳.۲: نمودار برآوردهای کمترین توان‌های دوم جریمه‌شده (PLS) در مقابل برآورد کمترین توان‌های دوم معمولی (OLS) وقتی که ماتریس طرح متعامد است. محور افقی OLS و محور عمودی PLS است.

شکل ۳.۲ ارتباط برآورد OLS را با برآوردهای لاسو، ریج، MCP و SCAD با فرض متعامد بودن ماتریس طرح X نشان می دهد. محور افقی نشان‌دهنده برآورد OLS و محور عمودی برآوردهای کمترین توان‌های دوم جریمه‌شده است. چون برآوردگر OLS نااریب است، انتظار داریم که برای ضرایب بزرگ، برآورد جریمه‌شده برابر یا نزدیک به برآورد OLS باشد تا شرط نااریبی برقرار شود. از طرفی برای برقراری شرط تنکی، لازم است روش کمترین توان‌های دوم جریمه‌شده، ضرایب رگرسیونی کوچک را صفر برآورد کند. همچنین لازم است برآوردگر موردنظر پیوسته باشد. با توجه به شکل، پیوستگی برآوردگر لاسو بدیهی است. همچنین برای ضرایب با مقادیر کوچک، برآورد لاسو برابر صفر است. در نتیجه شرط تنکی برقرار است. اما با بزرگ شدن مقادیر ضرایب، برآورد لاسو از برآورد OLS فاصله می‌گیرد، یعنی شرط نااریبی برقرار نیست. در نتیجه عیب روش لاسو این است که برای ضرایب با مقادیر بزرگ، برآوردهای اریب را نتیجه می‌دهد.

برآوردگر ریج برخلاف لاسو، علی‌رغم پیوسته بودن، ضرایب کوچک را صفر برآورد نمی‌کند. همچنین با افزایش قدر مطلق ضرایب، مقدار اریبی به شدت افزایش می‌یابد. بنابراین تابع جریمه ریج یک جریمه مناسب برای انتخاب متغیر نیست.

با توجه به نمودار برآورد SCAD در مقابل برآورد OLS، برقراری ویژگی های تنکی و پیوستگی واضح است. علاوه براین، در ابتدا میزان اریبی برآوردگر جریمه شده با اریبی برآوردگر لاسو معادل است، اما با افزایش قدر مطلق ضرایب، اریبی به سمت صفر تنزل می‌یابد. لذا ویژگی نااریبی نیز برقرار است.

همچنین با توجه به شکل، تابع جریمه MCP نیز برآوردگری با ویژگی‌های مطلوب را نتیجه می‌دهد. رفتار تابع جریمه MCP بسیار شبیه SCAD است، با این تفاوت که با فاصله گرفتن ضرایب از صفر، اریبی برآوردگر MCP فورا به سمت صفر می‌رود، اما اریبی برآوردگر SCAD در یک فاصله معین برابر با اریبی لاسو است و سپس به سرعت کاهش می‌یابد. بنابراین از توابع جریمه MCP و SCAD می‌توان به عنوان توابع جریمه مناسب نام برد.

فن و لی (۲۰۰۱) همچنین نشان دادند که تحت شرایط کافی زیر برآوردگر کمترین توان‌های دوم جریمه‌شده (۱۴.۲) دارای ویژگی‌های مطلوب فوق است:

۱. تنکی: برآوردگر دارای ویژگی تنکی است اگر $\min_{t\geq ۰}\{t+p'_\lambda(t)\} > ۰$.

۲. نا اریبی: برآوردگر به‌دست آمده نااریب است اگر، برای مقادیر بزرگ t، $p'_\lambda(t) = ۰$.

۳. پیوستگی: برآوردگر پیوسته است اگر و فقط اگر $\arg\min_{t\geq ۰}\{t+p'_\lambda(t)\} = ۰$.

همان‌طور که ملاحظه می شود، شرایط کافی برای داشتن ویژگی‌های مطلوب بیان‌شده توسط فن و لی (۲۰۰۱) وابسته به مشتق تابع جریمه هست. علاوه بر این، شکل مشتق تابع جریمه در حل مسئله بهینه‌سازی جریمه‌شده تاثیر بسزایی دارد. مشتق تابع جریمه لاسو به صورت $p'_\lambda(t) = \lambda$ و مشتق توابع جریمه SCAD و MCP، به ترتیب، به صورت

$$p'_\lambda(t) = \lambda\left\{I(t\leq\lambda) + \frac{(a\lambda - t)_+}{(a-۱)\lambda}I(t>\lambda)\right\}, \qquad t\geq ۰,\ a>۲ \tag{۳۰.۲}$$

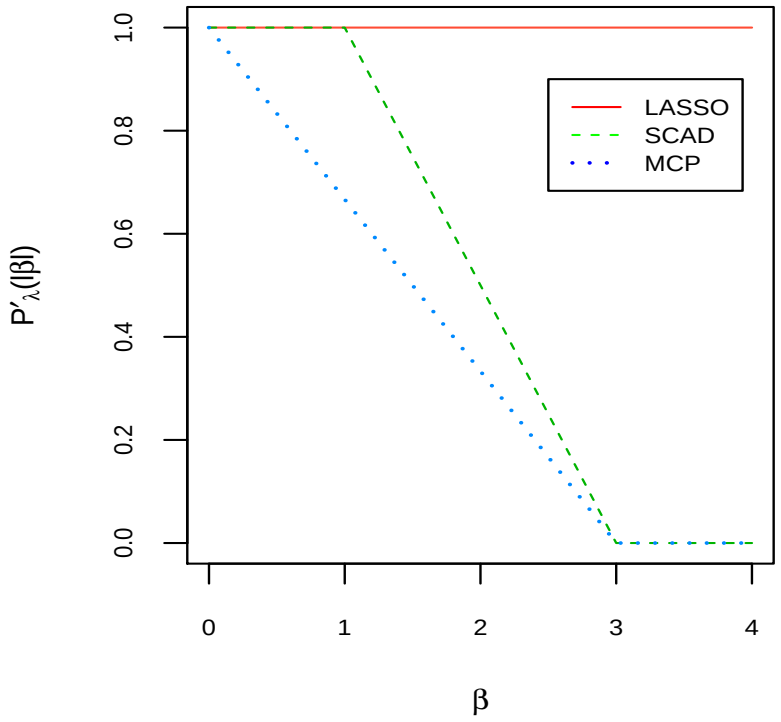
و

$$p'_\lambda(t) = \lambda\left(۱ - \frac{t}{\gamma\lambda}\right)_+, \qquad t\geq ۰,\ \gamma>۰ \tag{۳۱.۲}$$

به‌دست می‌آیند. بررسی شرایط کافی ویژگی‌های مطلوب توابع جریمه بسیار ساده است. به عنوان مثال، برای تابع جریمه لاسو، $\arg\min_{t\geq ۰}\{t+p'_\lambda(t)\} = ۰$ و $\min_{t\geq ۰}\{t+p'_\lambda(t)\} = \lambda > ۰$، اما $p'_\lambda(t) = \lambda > ۰$. بنابراین تابع جریمه لاسو دارای ویژگی نااریبی نیست.

نمودار مشتق توابع جریمه لاسو، SCAD و MCP به‌ازای $\lambda = ۱$ و $a = \gamma = ۳$ در شکل ۴.۲ داده شده است. با توجه به شکل می‌بینیم که نرخ جریمه توابع SCAD و MCP در مبدا با نرخ جریمه تابع لاسو برابر است، اما با افزایش قدرمطلق ضرایب، نرخ جریمه توابع SCAD و MCP به سمت صفر نزول می‌کند. لازم به ذکر است که نرخ جریمه تابع SCAD در ابتدا با فاصله گرفتن ضرایب از صفر معادل نرخ جریمه لاسو است و سپس با افزایش قدرمطلق ضرایب به سرعت کاهش می‌یابد، در حالی‌که تابع جریمه MCP به محض خارج شدن از مبدا، نرخ جریمه آن فورا به سمت صفر تنزل می‌یابد.

شکل ۴.۲ همچنین نشان می‌دهد که روش MCP مقعر مینیماکس است، یعنی در بین همه توابع جریمه پیوسته مشتق‌پذیر بر بازه $(۰,\infty)$ که برای هر $t\geq\gamma\lambda$ در دو شرط $p'_\lambda(t) = ۰$ و $p'_\lambda(۰_+) = \lambda$



شکل ۴.۲: نمودارهای مشتق توابع جریمه لاسو، SCAD و MCP.

صدق می‌کنند، MCP ماکزیمم تقعر زیر را مینیمم می‌کند

$$\mathcal{K} = \sup_{0<t_1<t_2} \frac{p'_\lambda(t_1) - p'_\lambda(t_2)}{t_2 - t_1}. \qquad (۳۲.۲)$$

به خاطر داشته باشید که اغلب توابع جریمه‌ای که شرایط مطلوب تنکی، ناااریبی و پیوستگی را دارا هستند، غیر محدب هستند. این موضوع، مسئله بهینه‌سازی را با چالش مواجه می‌کند. مقعر مینماکس بودن تابع MCP بدین معنی است که در بین همه توابعی که دارای سه ویژگی مطلوب فوق هستند، MCP اندازه ماکزیمم تقعر (۳۲.۲) را مینیمم می‌کند. در شکل ۴.۲، مشتق‌های توابع SCAD و MCP در نقاط ۰ و $\gamma\lambda = ۳$ برابر هستند، اما MCP در این ناحیه دارای تقعر $\frac{۱}{۳} = \frac{۱}{\gamma}$ است، در حالی‌که SCAD دارای ماکزیمم تقعر $\frac{۱}{۲} = \frac{۱}{\gamma - ۱}$ است. برای اطلاعات بیشتر به برهنی و هوانگ (۲۰۱۹) مراجعه شود.

۳.۴.۲ انتخاب پارامتر جریمه

انتخاب پارامتر جریمه در رگرسیون جریمه‌شده نقش بسیار مهمی را در کنترل پیچیدگی مدل بازی می‌کند. به عنوان مثال، به‌ازای $\lambda = ۰$ برآوردگر جریمه شده به برآوردگر OLS تبدیل می‌شود و همه متغیرهای توضیحی وارد مدل می‌شوند. اگر $\lambda = \infty$، جزء دوم تابع (۱۴.۲) برابر با ∞ می‌شود، و در نتیجه تمام ضرایب صفر شده و هیچ متغیری وارد مدل نمی‌شود. برای انتخاب بهترین مقدار بین این دو حالت فرین، از معیارهای کلاسیک مانند AIC، BIC و GCV استفاده می‌شود. به‌طور مشخص، برای

λ داده شده، برآورد کمترین توان‌های دوم جریمه شده، $\hat{\beta}_\lambda$، را به‌دست آورده و مقدار معیار کلاسیک انتخاب مدل را بر اساس $\hat{\beta}_\lambda$ محاسبه می‌کنیم. این معیارهای کلاسیک به صورت زیر بیان می‌شوند

$$AIC_\lambda = \| \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda \|^2 + 2df_\lambda\hat{\sigma}^2$$
$$BIC_\lambda = \| \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda \|^2 + \log(n)df_\lambda\hat{\sigma}^2$$
$$GCV_\lambda = \frac{\frac{1}{n}\| \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda \|^2}{(1 - df_\lambda/n)^2}$$

که $\hat{\sigma}^2 = \frac{RSS_p}{n-p}$ و RSS_p مجموع توان‌های دوم باقیمانده‌ها تحت مدل کامل است. منظور از مدل کامل مدلی است که همه متغیرهای توضیحی در مدل حضور دارند. در معیارهای فوق، درجه آزادی مدل برآوردشده df_λ معمولاً با استفاده از تعداد ضرایب غیر صفر به‌صورت

$$df_\lambda = \sum_{j=1}^{p} I(\hat{\beta}_{j,\lambda} \neq 0)$$

به‌دست می‌آید که در آن $I(.)$ تابع نشانگر است. راه دیگر محاسبه df_λ استفاده از ماتریس طرح به‌صورت

$$df_\lambda = tr\left(\mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda + n\boldsymbol{\Sigma}_\lambda)^{-1}\mathbf{X}_\lambda^T\right)$$

است که در آن $\mathbf{X}_\lambda$ ماتریس طرح مدل متناظر با λ داده شده و $\boldsymbol{\Sigma}_\lambda$ ماتریس بلوکی به‌صورت

$$\boldsymbol{\Sigma}_\lambda = \text{diag}_{\hat{\beta}_{j,\lambda}\neq 0}\{P'_\lambda(|\hat{\beta}_{j,\lambda}|)/|\hat{\beta}_{j,\lambda}|\}$$

است که $\hat{\beta}_{j,\lambda}$، j ــامین مولفه از برآورد کمترین توان‌های دوم جریمه‌شده $\hat{\boldsymbol{\beta}}_\lambda$ است. در پایان، برای یک مجموعه از نقاط داده شده $\lambda_1, ..., \lambda_M$، مقداری از λ را به عنوان مقدار بهینه انتخاب می‌کنیم که دارای کمترین مقدار معیار کلاسیک باشد.

یک روش دیگر برای انتخاب پارامتر جریمه استفاده از روش اعتبارسنجی متقابل k گروهی[۳۴] است. در این روش، برای یک λ داده شده، مجموعه مشاهدات را به طور تصادفی به k گروه با اندازه یکسان افراز می‌کنیم. سپس گروه اول را به‌عنوان داده‌های آزمون[۳۵] و سایر گروه‌ها را به‌عنوان داده‌های آموزشی[۳۶] در نظر می‌گیریم. پس از برازش مدل با استفاده از داده‌های آموزشی، مقدار متغیر پاسخ به ازای داده‌های آزمون پیشگویی می‌شود و خطای پیشگویی به‌صورت $PE_1 = \sum_{i\in I_1}(y_i - \hat{y}_i)^2$ به‌دست می‌آید که I_1 اندیس مشاهدات گروه اول است. این روند k بار تکرار می‌شود و در هر تکرار یکی از گروه‌ها به عنوان داده‌های آزمون و بقیه گروه‌ها به عنوان داده‌های آموزشی در نظر گرفته می‌شوند. در نتیجه k برآورد خطای آزمون $PE_k, ..., PE_1$ به‌دست آمده و در نهایت، امتیاز اعتبارسنجی متقابل به‌صورت

$$CV(\lambda) = \frac{1}{k}\sum_{i=1}^{k} PE_i$$

[34] k-fold cross validation
[35] Test dataset
[36] Training dataset

محاسبه می‌شود. این کار برای λهای مختلف تکرار شده و سپس λای انتخاب می‌شود که مقدار $CV(\lambda)$ را مینیمم کند. در عمل معمولاً مقدار k برابر ۵ یا ۱۰ در نظر گرفته می‌شود.

یکی از مشکلات روش لاسو، که قبلا به آن اشاره شد، این است که در بین یک گروه از متغیرهای توضیحی با همبستگی بالا، تنها یک متغیر را انتخاب می‌کند. بنابراین، این روش توانایی انتخاب متغیر گروهی را ندارد. در عمل ممکن است داده‌ها ساختار گروهی داشته باشند و لازم است متغیرها به‌صورت گروهی انتخاب شوند. در ادامه، چند روش انتخاب متغیر گروهی را معرفی می‌کنیم.

۵.۲ انتخاب متغیر گروهی

مدل رگرسیون خطی با p متغیر توضیحی را در نظر بگیرید. فرض کنید متغیرهای توضیحی به‌طور ذاتی به J گروه مجزا افراز شده باشند و بتوان مدل را به شکل

$$Y = \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \varepsilon \qquad (۳۳.۲)$$

نوشت که $\mathbf{Y} = (y_1, ..., y_n)^T$ بردار مشاهدات متغیر پاسخ، $\mathbf{X}_j$، با بعد $n \times p_j$، ماتریس طرح p_j متغیر توضیحی در گروه jـام، $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jp_j})^T$ ضرایب رگرسیونی گروه jـام و ε بردار خطا است. هنگام برآورد و انتخاب متغیرهای مهم، هر گروه از متغیرها را به‌عنوان یک واحد درنظر می‌گیرند و این عمل فرآیند انتخاب متغیر را تسهیل می‌کند. در نتیجه متغیرها به صورت گروهی از مدل حذف شده یا در مدل باقی می‌مانند. ساختار گروهی متغیرها می‌تواند دلایل و اهداف متفاوتی داشته باشد. نمایش یک متغیر کیفی چند سطحی با استفاده از متغیرهای ظاهری و نمایش اثر یک متغیر تصادفی پیوسته با استفاده از توابع پایه‌ای، نمونه‌هایی از متغیرهای گروه‌بندی‌شده هستند. گروه‌بندی همچنین می‌تواند به‌منظور بهره‌گیری از دانش پیشین، که از نظر علمی معنی‌دار است، مورد استفاده قرار گیرد. به‌عنوان مثال، در تحلیل داده‌های بیان ژنی[۳۷]، ژن‌های متعلق به یک مسیر بیولوژیکی یکسان می‌توانند در یک گروه قرار گیرند. در مطالعات پیوند ژنتیکی، نشانگرهای ژنتیکی از ژن یکسان می‌توانند به‌عنوان یک گروه در نظر گرفته شوند. در عمل مطلوب است ساختار گروهی را در تحلیل این نوع داده‌ها مورد توجه قرار دهیم. در ادامه چند تابع جریمه برای انتخاب متغیرهای گروهی معرفی می‌کنیم.

۱.۵.۲ جریمه لاسوی گروهی

تابع جریمه لاسوی گروهی توسط یوآن و لین[۳۸] (۲۰۰۶) معرفی شد و زمانی به کار می‌رود که متغیرهای توضیحی از قبل گروه‌بندی شده باشند. برآوردگر لاسوی گروهی با مینیمم کردن تابع هدف جریمه‌شده

$$\mathcal{L}(\beta) = \frac{1}{2} \parallel \mathbf{Y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j \parallel^2 + \lambda \sum_{j=1}^{J} \parallel \boldsymbol{\beta}_j \parallel_{\mathbf{R}_j} \qquad (۳۴.۲)$$

[37] Gene expression data

[38] Yuan and Lin

به‌دست می‌آید که λ پارامتر جریمه، $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_J^T)^T$ بردار ضرایب، $\mathbf{R}_j$ یک ماتریس معین مثبت با بعد $p_j \times p_j$ و $\| \boldsymbol{\beta}_j \|_{\mathbf{R}_j} = (\boldsymbol{\beta}_j^T \mathbf{R}_j \boldsymbol{\beta}_j)^{\frac{1}{2}}$ است.

یک مسئله مهم، نحوه انتخاب ماتریس $\mathbf{R}_j$ است. در حالتی که $\mathbf{X}_j^T \mathbf{X}_j = n\mathbf{I}_{p_j}$، یوآن و لین (۲۰۰۶)، $\mathbf{R}_j = \mathbf{I}_{p_j}$ را پیشنهاد دادند. با این وجود، انتخاب $\mathbf{R}_j = \mathbf{I}_{p_j}$ چندان مناسب نیست، زیرا مقیاس متغیرهای توضیحی ممکن است یکسان نباشد. در حالت کلی، یک انتخاب مناسب $\mathbf{R}_j$، ماتریس گرام[۳۹] بر اساس $\mathbf{X}_j$ یعنی $\mathbf{X}_j^T \mathbf{X}_j / n$ است. این انتخاب $\mathbf{R}_j$ معادل با استاندارد کردن متغیرها در هر گروه است.

برای برازش رگرسیون لاسوی گروهی با الگوریتم ارائه‌شده توسط یوآن و لین (۲۰۰۶)، لازم است که $\mathbf{X}_j$ ها متعامد باشند. بدین منظور روش متعامدسازی گرام‌ـاشمیت می‌تواند مورد استفاده قرار گیرد. ویژگی متعامد بودن، توسعه الگوریتم‌های محاسباتی برای برازش رگرسیون جریمه‌شده را تسهیل می‌کند و باعث کاهش بار محاسباتی می‌شود (برهنی و هوانگ، ۲۰۱۵). بعلاوه، این ویژگی امکان استفاده از الگوریتم‌های ارائه شده برای لاسوی گروهی را برای برازش رگرسیون SCAD گروهی و MCP گروهی فراهم می‌کند. لازم به ذکر است که بدون از دست دادن کلیت می‌توانیم فرض متعامد بودن متغیرهای درون هر گروه را بپذیریم، زیرا پس از برآورد ضرایب رگرسیونی مدل استاندارد، می‌توانیم با تبدیل ضرایب رگرسیونی مدل اصلی را به‌دست آوریم. در ادامه به بیان جزئیات این موضوع می‌پردازیم.

با استفاده از تجزیه چولسکی، ماتریس گرام را به صورت $\mathbf{R}_j = \mathbf{U}_j^T \mathbf{U}_j$ در نظر می‌گیریم که $\mathbf{U}_j$ یک ماتریس بالا مثلثی با بعد $p_j \times p_j$ است. فرض کنید $\tilde{\mathbf{X}}_j = \mathbf{X}_j \mathbf{U}_j^{-1}$ و $\mathbf{b}_j = \mathbf{U}_j \boldsymbol{\beta}_j$. بنابراین، می‌توان تابع جریمه‌شده (۳۴.۲) را به صورت

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2} \| \mathbf{Y} - \sum_{j=1}^{J} \tilde{\mathbf{X}}_j \mathbf{b}_j \|^2 + \lambda \sum_{j=1}^{J} \| \mathbf{b}_j \| \qquad (۳۵.۲)$$

نوشت. پس از به دست آوردن جواب با مینیمم کردن تابع (۳۵.۲)، می‌توانیم جواب مسئله اصلی (۳۴.۲) را با تبدیل $\boldsymbol{\beta}_j = \mathbf{U}_j^{-1} \mathbf{b}_j$ به‌دست آوریم. با توجه به تعریف $\mathbf{U}_j$ نتیجه می‌شود

$$\begin{aligned}
\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_j &= \mathbf{U}_j^{T^{-1}} \mathbf{X}_j^T \mathbf{X}_j \mathbf{U}_j^{-1} \\
&= \mathbf{U}_j^{T^{-1}} n \mathbf{R}_j \mathbf{U}_j^{-1} \\
&= n \mathbf{U}_j^{T^{-1}} \mathbf{U}_j^T \mathbf{U}_j \mathbf{U}_j^{-1} \\
&= n \mathbf{I}_{p_j}.
\end{aligned}$$

بنابراین با این انتخاب $\mathbf{R}_j$ و بدون از دست دادن کلیت، فرض می‌کنیم که به‌ازای $1 \leq j \leq J$، $\mathbf{X}_j^T \mathbf{X}_j = n\mathbf{I}_{p_j}$. توجه کنید که در اینجا فرض نکرده‌ایم که برای $j \neq k$، $\mathbf{X}_j$ و $\mathbf{X}_k$ متعامد هستند. هنگامی که $p_j = 1$ $(1 \leq j \leq J)$، لاسوی گروهی به لاسوی استاندارد تبدیل می‌شود. در این حالت، $\mathbf{R}_j = \frac{\|\mathbf{X}_j\|^2}{n}$ معادل با واریانس نمونه‌ای jـامین متغیر توضیحی است. بنابراین، در نظر گرفتن $\mathbf{R}_j$ به صورت ماتریس گرام معادل استاندارد کردن متغیرهای توضیحی قبل از تحلیل است که اغلب برای برازش رگرسیون لاسو توصیه می‌شود.

[39] Gram matrix

۲.۵.۲ لاسوی گروهی تطبیقی

همان‌طور که در بخش ۴.۲ بیان شد، برآوردگر لاسو دارای ویژگی پیشگویی نیست. لنگ[۴۰] و همکاران (۲۰۰۶) نشان دادند که اگر پارامتر جریمه با مینیمم کردن خطای پیشگویی انتخاب شود، لاسو دارای ویژگی سازگاری در انتخاب نیست. دلیل این مشکل این است که لاسو برای تمام ضرایب رگرسیونی میزان انقباض یکسانی را در نظر می‌گیرد. مطلوب این است که برای ضرایب بزرگتر انقباض کوچکتر و برای ضرایب نزدیک به صفر انقباض بزرگتری را در نظر بگیرد. این موضوع باعث می‌شود که لاسو مدلی بزرگتر از مدل درست را انتخاب کند، چون قادر به تشخیص ضرایب کوچک از ضرایب صفر نیست. تابع جریمه لاسوی گروهی هم رفتاری شبیه لاسو دارد و گروه‌های بی‌اهمیت زیادی را وارد مدل می‌کند. برای بهبود عملکرد لاسوی گروهی، تابع مجموع توان‌های دوم جریمه‌شده به‌صورت

$$\mathcal{L}(\beta) = \frac{1}{2} \parallel \mathbf{Y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j \parallel^2 + \lambda \sum_{j=1}^{J} \omega_j \parallel \boldsymbol{\beta}_j \parallel_{\mathbf{R}_j} \qquad (۳۶.۲)$$

تعریف می‌شود که $\omega_j = (\parallel \tilde{\boldsymbol{\beta}}_j \parallel_{\mathbf{R}_j})^{-1}$ و $\tilde{\boldsymbol{\beta}}_j$ یک مقدار اولیه برای $\boldsymbol{\beta}_j$ است. برآورد لاسوی گروهی می‌تواند به‌عنوان مقدار اولیه در این روش استفاده شود. همان‌طور که می‌بینید لاسوی گروهی تطبیقی برای گروه‌های مختلف، جریمه‌های متفاوتی را در نظر می‌گیرد.

۳.۵.۲ سایر توابع جریمه گروهی

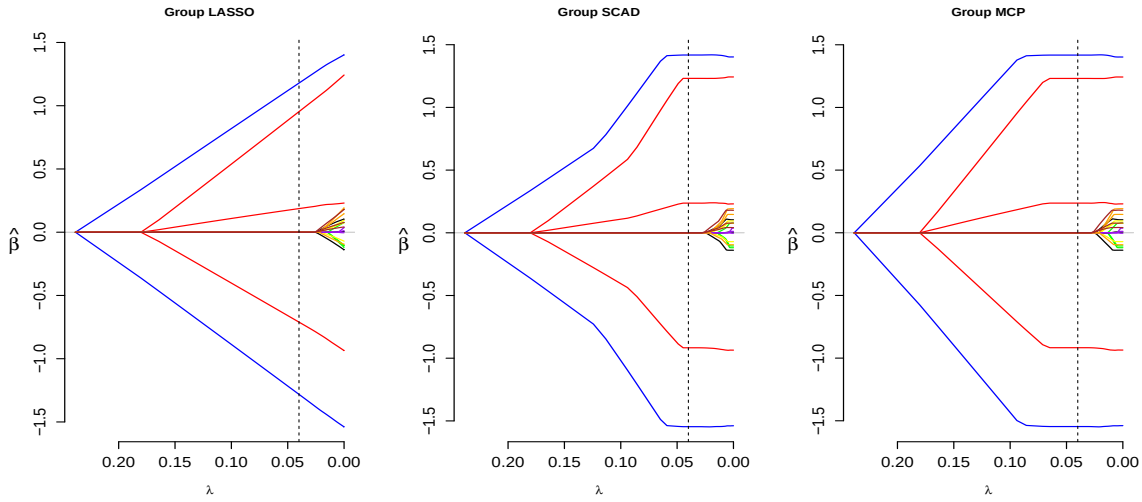
واضح است که تابع جریمه گروهی، با قرار دادن $\parallel \boldsymbol{\beta}_j \parallel_{\mathbf{R}_j}$ به جای $\mid \beta_j \mid$ در تابع جریمه معمولی به دست می‌آید. حال اگر $p_\lambda(t)$ یک تابع جریمه باشد، یک روش انتخاب متغیر گروهی بر اساس معیار

$$\mathcal{L}(\beta) = \frac{1}{2} \parallel \mathbf{Y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j \parallel^2 + \sum_{j=1}^{J} p_{\lambda,\gamma} \left(\parallel \boldsymbol{\beta}_j \parallel_{\mathbf{R}_j} \right) \qquad (۳۷.۲)$$

به‌دست می‌آید. اینجا γ یک پارامتر اضافی است که می‌تواند برای اصلاح $p_\lambda(\cdot)$ استفاده شود. بنابراین می‌توان مشابه لاسوی گروهی، جریمه‌های SCAD گروهی، MCP گروهی و بریج گروهی را نیز تعریف کرد.

برای درک بهتر مفهوم انتخاب متغیر گروهی، نمونه‌ای با $J = ۲۰$ گروه در نظر بگیرید به طوری که دو گروه اول دارای ضرایب غیر صفر و ضرایب مربوط به سایر گروه‌ها صفر می‌باشند. فرض کنید $\boldsymbol{\beta}_۱ = (-\sqrt{۲}, \sqrt{۲})^T$ و $\boldsymbol{\beta}_۲ = (-۰/۵, ۱, ۰/۵)^T$. همچنین فرض کنید اندازه گروه‌های صفر برابر ۳ باشد. شکل ۵.۲ برآوردهای لاسوی گروهی، SCAD گروهی و MCP گروهی را برای مقادیر مختلف λ نشان می‌دهد. با توجه به شکل می‌توان دید که، برای هر مقدار λ، همه ضرایب متعلق به یک گروه، یا برابر صفر برآورد می‌شوند، یا همه آن ضرایب دارای مقداری غیر صفر هستند. در نتیجه، همه متغیرهای یک گروه یا در مدل قرار می‌گیرند، یا همگی از مدل حذف می‌شوند. این موضوع، مفهوم انتخاب گروهی را بیان می‌کند. علاوه براین برای SCAD گروهی و MCP گروهی، بازه‌ای از λ وجود دارد به‌طوری که

[40] Leng

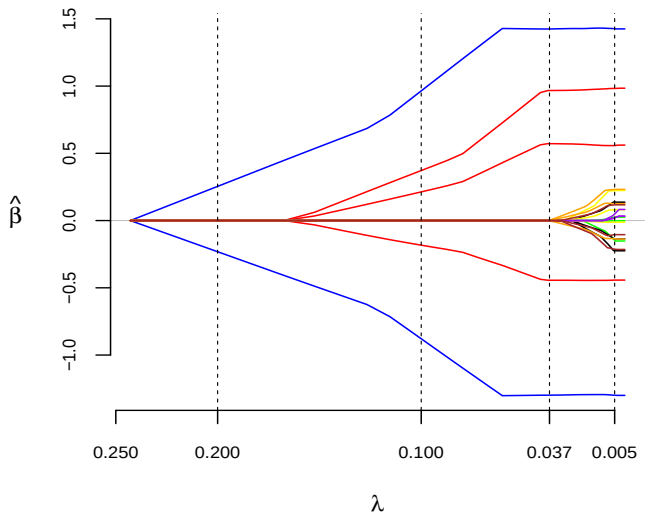


شکل ۵.۲: برآوردهای لاسوی گروهی، SCAD گروهی و MCP گروهی برای مقادیر مختلف λ. خط عمودی نقطه‌چین مقدار $\lambda = ۰/۰۴$ را نشان می‌دهد. نمودارهای آبی و قرمز، به ترتیب، برآوردهای ضرایب گروه اول و دوم و بقیه نمودارها مربوط به گروه‌های صفر هستند.

در آن بازه، برآوردها به مقادیر واقعی پارامترها بسیار نزدیک هستند. به‌عنوان مثال، به ازای $\lambda = ۰/۰۴$، MCP گروهی و SCAD گروهی همان برآوردهای کمترین توان‌های دوم را تولید می‌کنند و مدلی را نتیجه می‌دهند که تنها شامل متغیرهای مهم است. این مدل را مدل پیشگو[41] گویند. اما برای لاسوی گروهی، برآوردها همواره اریب هستند، مگر به ازای $\lambda = ۰$. در این حالت برآورد لاسو با برآورد کمترین توان‌های دوم معادل است.

برای بررسی دقیق‌تر، شکل ۶.۲ را برای تابع جریمه SCAD گروهی در نظر بگیرید. با توجه به این شکل، برای $\lambda = ۰/۰۰۵$، تقریبا متغیرهای توضیحی همه گروه‌ها در مدل حضور دارند، سپس با افزایش λ، همه ضرایب به سمت صفر منقبض می‌شوند. به‌عنوان مثال، به‌ازای $\lambda = ۰/۱۰$ فقط متغیرهای گروه‌های مهم در مدل حضور دارند، اما اریبی برآوردهای به‌دست آمده زیاد است. به‌ازای $\lambda = ۰/۲۰$ تنها متغیرهای گروه اول در مدل حضور دارند و متغیرهای گروه دوم، که متغیرهای مؤثر بر پاسخ هستند، از مدل حذف می‌شوند. در حالت خاص $\lambda = ۰/۰۳۷$ همه ضرایب گروه‌های بی‌اهمیت صفر می‌شوند، درحالی‌که برای گروه‌های مهم، برآوردهای به‌دست آمده با SCAD گروهی با برآوردهای کمترین توان‌های دوم معادل هستند. به‌عبارت دیگر، به‌ازای $\lambda = ۰/۰۳۷$ برآوردگر SCAD گروهی به خوبی برآوردگر پیشگو است. بنابراین $\lambda = ۰/۰۳۷$ می‌تواند یک مقدار بهینه برای λ باشد. البته در عمل مقدار بهینه λ با استفاده از روش‌های معرفی‌شده در زیربخش ۳.۴.۲ به‌دست می‌آید.

[41] Oracle model



شکل ۶.۲: برآوردهای SCAD گروهی برای مقادیر مختلف λ.

۶.۲ انتخاب متغیر در داده‌های با بعد بسیار بالا

به‌طور کلی، داده‌ها را از نظر بعد می‌توان به دو گروه تقسیم کرد: ۱ـ بعد پایین[۴۲] ۲ـ بعد بالا[۴۳]. در داده‌های با بعد پایین، تعداد متغیرها، p، بسیار کوچکتر از تعداد مشاهدات، n، است. به‌عنوان مثال، مجموعه داده با حجم $n = ۱۰۰$ و بعد $p = ۱۰$ یک مجموعه داده با بعد پایین است. بعلاوه، در این داده‌ها، برای نظریه مجانبی فرض می‌شود که با افزایش n، مقدار p ثابت می‌ماند. در داده‌های با بعد بالا، p از n بزرگتر است. به‌عنوان مثال، $n = ۱۰۰$ و $p = ۲۰۰$ تحققی از بعد بالا است. در این نوع داده‌ها با افزایش n، مقدار p می‌تواند افزایش یابد. امروزه محققین در زمینه‌های مختلف علمی و تحقیقاتی مانند ژنومیک، علوم اعصاب، علوم زمین و مالی با این نوع داده‌ها مواجه هستند. اگرچه محققین در زمینه‌های مختلف اهداف متفاوتی را دنبال می‌کنند، اما یک موضوع مشترک دارند: تحقیق آن‌ها به شدت به استخراج اطلاعات مفید از داده‌های حجیم وابسته است و تعداد متغیرهای توضیحی در مقایسه با حجم نمونه می‌تواند بزرگ باشد. مسئله "p بزرگ، n کوچک" چالش‌های زیادی را پیش روی آماردانان قرار داده است و تحلیل این نوع داده‌ها مستلزم به کارگیری روش‌های نوین آماری است. یک فرض اساسی که استنباط پیرامون این نوع داده‌ها را ممکن می‌سازد، فرض تنکی[۴۴] است که بر اساس این فرض، فقط مجموعه کوچکی از متغیرهای توضیحی بر متغیر پاسخ تاثیرگذارند و بقیه متغیرهای توضیحی ارتباطی با متغیر پاسخ ندارند. بنابراین کاهش بعد و انتخاب متغیر نقش مهمی را در تحلیل این نوع داده‌ها ایفا می‌کند.

تاکنون روش‌های آماری مختلفی برای انتخاب متغیر در داده‌های با بعد بالا معرفی شده‌اند که از

[42] Low dimension
[43] High dimension
[44] Sparsity assumption

جمله می‌توان لاسو، لاسوی تطبیقی، SCAD، الاستیک نت، گاروت نامنفی[۴۵] و انتخاب‌گر دَن‌زیگ[۴۶] را نام برد. همه این روش‌ها می‌توانند برای تحلیل داده‌هایی به کار روند که تعداد متغیرهای پیشگو، p، بزرگتر از حجم نمونه، n، است.

در حالت خاص، ممکن است p با نرخ نمایی رشد کند، به عبارت دیگر، رابطه $\log(p) = O(n^a)$ برای برخی مقادیر مثبت a برقرار باشد. داده‌های بیان ژنی، ریز آرایه‌ها، داده‌های تابعی، داده‌های مالی، داده‌های پردازش تصویر، داده‌های تصویربرداری پزشکی و طبقه‌بندی تومور نمونه‌هایی از این نوع داده‌ها هستند. به‌عنوان مثال، در مطالعه رابطه بین ویژگی‌های ظاهری[۴۷] و ویژگی‌های ژنتیکی[۴۸] از میلیون‌ها SNP[۴۹] استفاده می‌کنند، یا در طبقه‌بندی نوعی تومور از داده‌های ریز آرایه استفاده می‌کنند که شامل هزاران عامل ژنتیکی است و هدف شناسایی ژن‌های موثر بر بیماری است. در این مثال‌ها، تعداد متغیرهای توضیحی بسیار بزرگتر از تعداد مشاهدات است. برای تمایز این نوع داده‌ها از داده‌های با بعد بالای با نرخ رشد چندجمله‌ای، آن‌ها را داده‌های با بعد بسیار بالا[۵۰] می‌نامند. بنابراین، در این رساله، عبارت ''بعد بالا'' برای نرخ رشد چندجمله‌ای ($p = O(n^b)$) و عبارت ''بعد بسیار بالا'' برای نرخ رشد نمایی ($p = \exp\{O(n^a)\}$) استفاده می‌شود. برای درک بهتر مطلب، دو حالت $p = n^b$ و $p = \exp\{n^a\}$ را به ترتیب برای بعد بالا و بعد بسیار بالا در نظر گرفته و نسبت $\frac{p}{n}$ را برای مقادیر مختلف (n, a, b) در جدول ۱.۲ گزارش کرده‌ایم. با توجه به جدول ۱.۲ ملاحظه می‌شود که در بعد بسیار بالا، مقدار p با افزایش n به سرعت رشد می‌کند و به سمت بینهایت می‌رود. برای آگاهی بیشتر درباره داده‌های با بعد بالا و چالش‌های تحلیل این نوع داده‌ها به فن و لی (۲۰۰۶)، فن و لیو[۵۱] (۲۰۰۸)، فن و همکاران (۲۰۰۹)، فن و همکاران (۲۰۱۴)، جان‌استون و تیترینگتون[۵۲] (۲۰۰۹)، بولمن و فن دِگیر[۵۳] (۲۰۱۱) و گیراد[۵۴] (۲۰۱۴) مراجعه کنید.

تحلیل داده‌های با بعد بسیار بالا با سه چالش همزمان هزینه محاسباتی، دقت آماری و پایداری الگوریتمی روبروست (فن و لیو، ۲۰۰۹). این چالش‌ها کاربرد مستقیم روش‌های انتخاب متغیر اشاره شده را برای داده‌های با بعد بسیار بالا با مشکل مواجه می‌کند. قالب کلی پیشنهاد شده برای رفع این مشکل یک طرح دو مرحله‌ای است که در مرحله اول، بعد مدل توسط یک روش غربالگری کاهش یافته و در مرحله دوم از روش‌های انتخاب متغیر مبتنی بر جریمه، که در بخش‌های ۴.۲ و ۵.۲ به آن‌ها اشاره شد، برای انتخاب متغیر از زیر مدل غربال شده استفاده می‌شود. بدیهی است یکی از این روش‌ها که هزینه محاسباتی چندانی ندارد، غربالگری بر اساس همبستگی متغیرهای توضیحی با متغیر پاسخ است. در این راستا، یک راهکار ساده می‌تواند این باشد که اگر همبستگی یک متغیر توضیحی با پاسخ کمتر

[45] Nonnegative garrote
[46] Dantzig selector
[47] Phenotypes
[48] Genotypes
[49] Single-nucleotide polymorphism
[50] Ultrahigh-dimensional data
[51] Lv
[52] Johnstone and Titterington
[53] Buhlmann and van de Geer
[54] Giraud

جدول ۱.۲: مقادیر $\frac{p}{n}$ برای داده‌های با ابعاد بالا و بسیار بالا.

	بعد بسیار بالا (مقادیر مختلف a)			بعد بالا (مقادیر مختلف b)				
n	۱۵/۱	۲	۲۵/۱	۳	$\frac{3}{8}$	$\frac{5}{8}$	$\frac{7}{8}$	۱
۱۰	۳/۱۶	۱۰	۳۱/۶۲	۱۰۰	۱/۰۷	۶/۷۸	۱۸۰/۶۱	۲۲۰۲/۶۴
۳۰	۵/۴۷	۳۰	۱۶۴/۳۱	۹۰۰	۱/۱۹	۱۴۵/۱۷	۱۰۹۵۱۲۱۴	۳۵۶۲۱۵۸۱۹۳۸۴
۱۰۰	۱۰	۱۰۰	۱۰۰۰	۱۰۰۰۰	۲/۷۶	۵۲۸۴۰۸	$۲/۶۴۳۴۶۴e+۲۲$	$۲/۶۸۸۱۱۷e+۴۱$
۲۰۰	۱۴/۱۴	۲۰۰	۲۸۲۸/۴۲	۴۰۰۰۰	۷/۳۴	۴۰۶۸۳۳۷۱۹۴	$۳/۰۸۶۲۶۸e+۴۲$	$۳/۶۱۲۹۸۷e+۸۴$
۵۰۰	۲۲/۳۶	۵۰۰	۱۱۱۸۰/۳۴	۲۵۰۰۰۰	۵۸/۴۵	$۲/۶۲۰۸۱۱e+۱۸$	$۱/۴۴۲۳۶۹e+۹۷$	$۲/۸۰۷۱۸۴e+۲۱۴$

از یک حد آستانه از پیش تعیین‌شده باشد، آن متغیر توضیحی را از مدل حذف می‌کنیم. براساس این راهکار، فن و لیو (۲۰۰۸) نخستین روش غربالگری را با استفاده از همبستگی پیرسن برای مدل‌های خطی معرفی کردند و آن را غربالگری مستقل مطمئن (SIS) نامگذاری کردند. تاکنون، علاوه بر SIS، چندین روش غربالگری دیگر نیز برای مدل‌های آماری مختلف معرفی شده‌اند که در ادامه به توصیف دو روش SIRS و DC-SIS می‌پردازیم.

۱.۶.۲ غربالگری مستقل مطمئن

مدل خطی زیر را در نظر بگیرید

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{۳۸.۲}$$

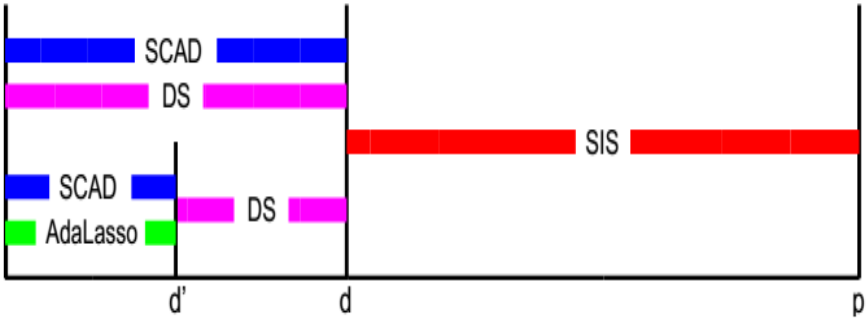
که $\mathbf{Y} = (y_1, ..., y_n)^T$ متغیر پاسخ، $\mathbf{X} = (x_1, ..., x_n)^T$ ماتریس طرح و $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$ خطاهای تصادفی می‌باشند. فرض کنید $\mathbb{M} = \{k : \beta_k \neq 0\}$ مجموعه اندیس متغیرهای مهم (مدل درست) با اندازه $s = |\mathbb{M}|$ و $\widehat{\mathbb{M}}$ مجموعه اندیس متغیرهای انتخاب‌شده با اندازه $|\widehat{\mathbb{M}}|$ باشد. برای رتبه‌بندی متغیرها، ابتدا بردار p بعدی $\boldsymbol{\omega} = (\omega_1, ..., \omega_p)^T$ را به‌صورت

$$\omega_k = \frac{1}{n}\mathbf{X}_k^T\mathbf{Y}, \qquad k = 1, ..., p \tag{۳۹.۲}$$

در نظر بگیرید که متغیرهای توضیحی و پاسخ استاندارد شده‌اند، به‌طوری‌که، برای $k = 1, ..., p$، $\frac{1}{n}\sum_{i=1}^n y_i^2 = 1, \frac{1}{n}\sum_{i=1}^n y_i = 0$ و $\frac{1}{n}\sum_{i=1}^n x_{ik} = 0, \frac{1}{n}\sum_{i=1}^n x_{ik}^2 = 1$. بنابراین، ω_k همان ضریب همبستگی نمونه‌ای بین kـامین متغیر توضیحی و متغیر پاسخ است.

روش غربالگری مستقل مطمئن، با نام اختصاری SIS، متغیرهای توضیحی را بر اساس مقادیر $|\omega_k|$ رتبه‌بندی می‌کند و متغیرهایی را انتخاب می‌کند که دارای همبستگی بالایی با متغیر پاسخ باشند. به‌طور واضح‌تر، برای هر مقدار $\gamma \epsilon (0, 1)$ مولفه‌های $|\omega_k|$ را به ترتیب صعودی مرتب کرده و سپس زیر مدل $\widehat{\mathbb{M}}$ را به صورت

$$\widehat{\mathbb{M}} = \{k : \text{در بین} |\omega_k| \text{مقدار بزرگ} [\gamma n] |\omega| \text{ قرار دارد.}\} \tag{۴۰.۲}$$



شکل ۷.۲: روش‌های انتخاب متغیر در ابعاد خیلی بالا

انتخاب می‌کنیم که γ یک مقدار ثابت از پیش تعیین شده، $[\gamma n]$ جزء صحیح γn و $R_k = |\omega_k|$ معیار مطلوبیت حاشیه‌ای[۵۵] نامیده می‌شود. به وضوح، انتخاب مقدار بزرگتری برای γ منجر به انتخاب تعداد بیشتری از متغیرها می‌شود. در عمل، انتخاب یک مقدار مناسب برای γ چندان ساده به نظر نمی‌رسد. در این راستا، فن و لیو (۲۰۰۸) انتخاب d متغیر با بزرگترین معیار مطلوبیت حاشیه‌ای را به‌عنوان یک روش جایگزین پیشنهاد دادند که d به‌صورت $d = [n/\log(n)]$ یا $d = ۲[n/\log(n)]$ تعیین می‌شود.

فن و لیو (۲۰۰۸) تحت قضیه‌ای نشان دادند که روش SIS دارای ویژگی غربالگری مطمئن[۵۶] است، یعنی این روش با احتمال نزدیک به یک همه متغیرهای مهم را انتخاب می‌کند. به‌عبارت دیگر، به ازای برخی مقادیر γ، داریم

$$P(\mathbb{M} \subseteq \widehat{\mathbb{M}}) \longrightarrow ۱, \qquad n \longrightarrow \infty. \tag{۴۱.۲}$$

مطلب مهم در این خصوص، شناسایی شرایطی است که تحت آن‌ها (۴۱.۲) برقرار باشد. شرایط لازم برای برقراری این ویژگی در قضیه مذکور مطرح شده است. این روش غربالگری، بعد متغیرهای توضیحی را از مقیاس بسیار بزرگ p به مقیاس نسبتا متوسط $d = [\gamma n] < p$ کاهش می‌دهد. بعد از مرحله غربالگری، روش‌های انتخاب متغیر مبتنی بر جریمه بر روی زیر مدل $\widehat{\mathbb{M}}$ برای انتخاب مدل نهایی به کار می‌روند. شکل ۷.۲ می‌تواند در فهم مطلب موثر باشد.

اگر پس از غربالگری، در مرحله دوم از جریمه SCAD استفاده شود، این روش دو مرحله ای را با SIS-SCAD نشان می‌دهیم. علاوه بر این، می‌توان پس از به‌کارگیری SIS، دوبار به‌طور متوالی از روش‌های جریمه استفاده کرد. به عنوان مثال، ابتدا با استفاده از SIS بعد مدل را از p به $d < n$ کاهش داده و سپس از انتخاب‌گر دَن‌زیگ برای انتخاب متغیر روی مدل غربال‌شده و کاهش بعد آن به $d' < d$ استفاده شود. در پایان، برای انتخاب مدل نهایی، روش لاسوی تطبیقی یا SCAD روی مدل کاهش یافته به کار می‌رود. در این حالت، این روش سه مرحله ای را با SIS-DS-SCAD یا SIS-DS-AdaLasso نشان می‌دهند.

پس از معرفی روش SIS توسط فن و لیو (۲۰۰۸)، این روش به مدل‌های آماری مختلف مانند مدل

[55] Marginal utility measure
[56] Sure screening property

خطی تعمیم‌یافته (فن و همکاران، ۲۰۰۹؛ فن و سانگ، ۲۰۱۰)، مدل خطرات متناسب کاکس[۵۷] (فن و همکاران، ۲۰۱۰)، مدل جمعی ناپارامتری (فن و همکاران، ۲۰۱۱) و مدل ضریب متغیر (فن و همکاران، ۲۰۱۴؛ لیو[۵۸] و همکاران، ۲۰۱۴؛ سانگ[۵۹] و همکاران، ۲۰۱۴؛ زیا و همکاران، ۲۰۱۶) تعمیم داده شد.

روش SIS را می توان برای کاهش بعد فضای متغیرها در مسائل رده‌بندی هم به‌کار برد. بدین منظور، فرض کنید n_0 و n_1 تعداد مشاهداتی از X_k باشند که رده آن‌ها به ترتیب ۰ و ۱ است. همچنین فرض کنید $\bar{X}_{k,0}$ و $\bar{X}_{k,1}$ به ترتیب میانگین متغیر توضیحی X_k در رده های ۰ و ۱، و $Se(X_k)$ انحراف معیار متغیر توضیحی X_k باشد. تحت این شرایط kـامین مولفه بردار ω به صورت

$$\omega_k = \frac{n_1 \bar{X}_{k,1} - n_0 \bar{X}_{k,0}}{Se(X_k)}$$

محاسبه می‌شود. بقیه مراحل انتخاب متغیر مشابه قبل انجام می‌شود.

- **غربالگری مستقل مطمئن تکراری**

ایده اصلی روش SIS رتبه‌بندی متغیرها بر اساس یک معیار مطلوبیت حاشیه‌ای است. بنابراین، سه مسئله مهم زیر ممکن است عملکرد این روش را با مشکل مواجه کنند:

۱. متغیر مهمی که به صورت حاشیه‌ای با متغیر پاسخ ناهمبسته است، توسط این روش انتخاب نمی‌شود.

۲. برخی از متغیرهای بی‌اهمیت که همبستگی بالایی با متغیرهای مهم دارند، نسبت به متغیرهای مهمی که دارای همبستگی نسبتاً ضعیفی با متغیر پاسخ هستند، در انتخاب متغیر توسط SIS، برتری دارند.

۳. این روش، همخطی بین متغیرهای توضیحی را در نظر نمی‌گیرد.

برای درک بهتر مشکلات روش SIS، دو مثال زیر را در نظر بگیرید.

مثال ۱.۶.۲. مدل خطی زیر را در نظر بگیرید

$$Y = \mathbf{X}^T \boldsymbol{\beta}^* + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1) \tag{۴۲.۲}$$

که $p = 2000$ و $\boldsymbol{\beta}^* = (3, 3, 3, 3, 3, -7/5, 0, ..., 0)^T$. بنابراین، مجموعه اندیس متغیرهای مهم به‌صورت $\mathbb{M} = \{j : \beta_j^* \neq 0\} = \{1, 2, 3, 4, 5, 6\}$ است.

فرض کنید متغیرهای توضیحی دارای توزیع نرمال استاندارد و ضریب همبستگی بین همه متغیرهای توضیحی برابر ۰/۵ باشد. بنابراین $Cov(X_6, Y) = 0$. لذا، X_6 و Y به‌طور حاشیه‌ای ناهمبسته هستند. در نتیجه متغیر مهم X_6 توسط SIS انتخاب نمی‌شود.

[57] Cox's proportional hazards model
[58] Liu
[59] Song

مثال ۲.۶.۲. دوباره مدل خطی (۴۲.۲) را با $\boldsymbol{\beta}^* = (۱۰, ۰, ..., ۰, ۱)^T$ در نظر بگیرید. در این مدل، مجموعه اندیس متغیرهای مهم شامل $\mathbb{M} = \{۱, ۲۰۰۰\}$ است. فرض کنید ضریب همبستگی بین تمام متغیرهای توضیحی، به جز $X_{۲۰۰۰}$، برابر ۰٫۹ و $X_{۲۰۰۰}$ مستقل از بقیه متغیرها است. در نتیجه

$$Cov(X_۱, Y) = ۱۰, \;\; Cov(X_{۲۰۰۰}, Y) = ۱, \qquad Cov(X_j, Y) = ۱۰, \;\; j \neq ۱, ۲۰۰۰.$$

در این حالت، ضریب همبستگی حاشیه‌ای همه متغیرهای بی‌اهمیت بزرگتر از ضریب همبستگی متغیر مهم $X_{۲۰۰۰}$ با پاسخ است. بنابراین تمام متغیرهای بی‌اهمیت در فرآیند انتخاب توسط SIS نسبت به $X_{۲۰۰۰}$ برتری دارند.

اگر ماتریس طرح داده‌ها متعامد باشد، روش SIS یک روش انتخاب متغیر قابل اعتماد بوده و نتایج آن مطلوب خواهد بود. اما در داده‌های با بعد بسیار بالا، موضوع همخطی امری اجتناب‌ناپذیر است. بنابراین با اعمال روش SIS، ممکن است سه مشکل فوق، عملکرد این روش را تحت تاثیر قرار دهند و برخی از متغیرهای مهم را از دست بدهیم. برای رفع این مشکل، فن و لیو (۲۰۰۸) روش SIS را به‌طور مکرر به کار بردند و این روش را غربالگری مستقل مطمئن تکراری (ISIS) نامیدند. الگوریتم روش ISIS به صورت زیر است:

- مرحله اول: یک زیرمجموعه $k_۱$ تایی از متغیرها، مانند $\mathcal{A}_۱ = \{X_{i۱}, ..., X_{ik_۱}\}$، را با استفاده از یک روش غربالگری مبتنی بر SIS، مانند SIS-SCAD یا SIS-Lasso انتخاب می‌کنیم. سپس یک مدل خطی با این مجموعه از متغیرهای توضیحی بر روی متغیر پاسخ برازش می‌دهیم.

- مرحله دوم: باقیمانده‌های مدل برازش شده در مرحله قبل را به‌عنوان متغیر پاسخ جدید در نظر گرفته و مرحله اول را برای $p - k_۱$ متغیر باقیمانده اجرا می‌کنیم تا زیرمجموعه جدید با $k_۲$ متغیر توضیحی و جدا از $\mathcal{A}_۱$ به صورت $\mathcal{A}_۲ = \{X_{j۱}, ..., X_{jk_۲}\}$ حاصل شود. این کار تا زمانی ادامه می‌یابد که l مجموعه جدا از هم $\mathcal{A}_۱, ..., \mathcal{A}_l$ حاصل شوند تا مجموعه $\mathcal{A} = \cup_{i=۱}^{l} \mathcal{A}_i$ با اندازه d را تشکیل دهند. در عمل l را می توان طوری انتخاب کرد که $|\mathcal{A}| < n$ باشد. در نهایت مجموعه $\mathcal{A}$ شامل همه متغیرهای مهم است که با روش ISIS انتخاب می‌شوند. سپس می‌توان روش‌های انتخاب متغیر مبتنی بر جریمه را برای متغیرهای مجموعه $\mathcal{A}$ به کار برد.

باید توجه داشته باشیم که برازش باقیمانده‌های مرحله اول بر روی متغیرهای توضیحی باقیمانده، می‌تواند به‌طور قابل توجهی اولویت انتخاب آن دسته از متغیرهای بی‌اهمیتی که همبستگی بالایی با متغیر پاسخ از طریق متغیرهای توضیحی $\{X_{i۱}, ..., X_{ik_۱}\}$ دارند را تضعیف کند، زیرا باقیمانده‌ها با این متغیرهای توضیحی ناهمبسته‌اند و این امر موجب رفع مشکل همخطی می‌شود. این عمل، همچنین باعث می‌شود متغیرهای توضیحی مهم از دست رفته در مرحله اول، احتمالا، حفظ شوند.

روش‌های غربالگری معرفی شده فوق مدل مبنا بوده و در صورت نادرست بودن مدل آماری مفروض، ممکن است منجر به حذف متغیرهای مهم شوند. در مدل‌سازی با بعد بالا، بدون داشتن اطلاعات پیشین، تعیین ساختار مدل واقعی بسیار مشکل است. بنابراین، برای اجتناب از تشخیص نادرست ساختار مدل، معرفی یک روش غربالگری آزاد ـ مدل[۶۰] ، که به نوع مدل بستگی نداشته باشد، ضروری به

[60] Model free

نظر می رسد. در این راستا، ژو[۶۱] و همکاران (۲۰۱۱) یک روش غربالگری آزاد- مدل به نام ''غربالگری و رتبه بندی مستقل مطمئن'' (SIRS) را معرفی کردند که در ادامه به بیان جزئیات این روش می‌پردازیم.

۲.۶.۲ غربالگری و رتبه بندی مستقل مطمئن

فرض کنید Y متغیر پاسخ با تکیه‌گاه Ψ_y و $\mathbf{X} = (X_1, ..., X_p)^T$ بردار متغیرهای توضیحی باشد. تابع توزیع شرطی Y به شرط $\mathbf{X}$ را به صورت $F(y|\mathbf{X}) = P(Y < y|\mathbf{X})$ در نظر گرفته و فرض کنید، اندیس متغیرهای مهم به صورت زیر تعریف شود

$$\mathbb{M} = \{k : \text{است} X_k \text{ به‌طور تابعی وابسته به } F(y|\mathbf{X}) \text{ تابع } ،y \in \Psi_y \text{ برای برخی مقادیر} \}.$$

اگر $k \in \mathbb{M}$، آنگاه X_k یک متغیر مهم و در غیر این‌صورت X_k یک متغیر بی‌اهمیت است. فرض کنید به‌ازای $1 \le k \le p$، $E(X_k) = 0$ و $Var(X_k) = 1$. حال $\Omega(y) = E\{\mathbf{X}F(y|\mathbf{X})\}$ را در نظر بگیرید. با استفاده از رابطه $F(y|\mathbf{X}) = E\{I(Y < y)|\mathbf{X}\}$ و قانون امید ریاضی مکرر داریم

$$\begin{aligned}\Omega(y) &= E[\mathbf{X}E\{I(Y < y)|\mathbf{X}\}] \\ &= E[E\{\mathbf{X}I(Y < y)|\mathbf{X}\}] \\ &= E[\mathbf{X}I(Y < y)] \\ &= Cov(\mathbf{X}, I(Y < y)) \end{aligned} \quad (۴۳.۲)$$

که $I(.)$ تابع نشانگر است. فرض کنید $\Omega_k(y)$، k -امین مولفه از $\Omega(y)$ باشد و

$$\omega_k = E\{\Omega_k^2(Y)\}, \qquad k = 1, ..., p \quad (۴۴.۲)$$

را تعریف کنید. ژو و همکاران (۲۰۱۱) از ω_k به‌عنوان معیار مطلوبیت حاشیه‌ای برای رتبه‌بندی متغیرها استفاده کردند، یعنی برای اندازه‌گیری اهمیت متغیر X_k بر پاسخ، همبستگی بین X_k و $I(Y < y)$ را به‌کار بردند. به‌سادگی می‌توان دید که، اگر X_k و Y مستقل باشند، آنگاه X_k و $I(Y < y)$ نیز مستقل‌اند. بنابراین، برای هر $y \in \Psi_y$، $\Omega_k(y) = 0$ و در نتیجه $\omega_k = 0$. از طرف دیگر، اگر X_k و Y مرتبط باشند، آنگاه برای برخی مقادیر $y \in \Psi_y$، $\Omega_k(y) \neq 0$، و بنابراین ω_k مقداری مثبت است. بنابراین (۴۴.۲) می تواند معیار مناسبی برای رتبه‌بندی متغیرها باشد.

برای پیاده‌سازی این روش غربالگری، لازم است معیار (۴۴.۲) بر اساس یک نمونه تصادفی برآورد شود. فرض کنید $(\mathbf{X}_i, Y_i)_{i=1}^n$ یک نمونه تصادفی از $(\mathbf{X}, Y)$ باشد. برای سهولت، فرض می‌کنیم متغیرهای توضیحی استاندارد شده‌اند، به‌طوری که، $\frac{1}{n}\sum_{i=1}^n X_{ij} = 0$ و $\frac{1}{n}\sum_{i=1}^n X_{ij}^2 = 1$. برای هر y داده شده، برآوردگر گشتاوری نمونه‌ای $\Omega_k(y)$ به صورت

$$\hat{\Omega}_k(y) = \frac{1}{n}\sum_{i=1}^n X_{ik} I(Y_i < y), \qquad k = 1, ..., p \quad (۴۵.۲)$$

[61] Zhu

محاسبه می‌شود که k، X_{ik} ــ امین مولفه از $\mathbf{X}_i$ است. در نتیجه ω_k به صورت

$$\hat{\omega}_k = \frac{1}{n}\sum_{j=1}^{n}\Big\{\frac{1}{n}\sum_{i=1}^{n} X_{ik} I(Y_i < Y_j)\Big\}^2, \qquad k = 1, ..., p \tag{۴۶.۲}$$

برآورد می‌شود. سپس با رتبه‌بندی متغیرهای توضیحی بر اساس $\hat{\omega}_k$، متغیرهای توضیحی با مقدار $\hat{\omega}_k$ بزرگتر از یک مقدار آستانه را به‌عنوان مجموعه متغیرهای مهم انتخاب می‌کنیم. برای اطلاعات بیشتر درباره انتخاب این مقدار آستانه به ژو و همکاران (۲۰۱۱) مراجعه کنید.

ژو و همکاران (۲۰۱۱) تحت قضیه‌ای نشان دادند

$$\max_{k\in\mathbb{M}^c}\omega_k < \min_{k\in\mathbb{M}}\omega_k$$

و $\omega_k = 0$ اگر و فقط اگر $k \in \mathbb{M}^c$. علاوه بر این، آن‌ها ویژگی سازگاری در رتبه‌بندی[۶۲] را برای روش ارائه شده ثابت کردند. این ویژگی بدین معنی است که، در احتمال، معیار ω_k همواره یک متغیر مهم را در رتبه بالاتری از یک متغیر بی اهمیت قرار می‌دهد. این ویژگی، سازگاری در انتخاب را نتیجه می‌دهد. قضایای فوق نشان می‌دهند که ω_k معیار حاشیه‌ای مناسبی برای غربالگری متغیرها در داده‌های با بعد بسیار بالاست.

۳.۶.۲ غربالگری با همبستگی فاصله‌ای

روش دیگر برای رسیدن به یک غربالگری آزادــ مدل، استفاده از یک معیار وابستگی است که به‌طور کارا بتواند روابط خطی و غیرخطی بین متغیرهای توضیحی و متغیر پاسخ را شناسایی کند. در این راستا، لی و همکاران (۲۰۱۲) یک روش غربالگری آزادــ مدل بر اساس همبستگی فاصله‌ای ارائه دادند. برخلاف ضریب همبستگی پیرسن که برای دو متغیر تصادفی تعریف می‌شود، ضریب همبستگی فاصله‌ای برای دو بردار تصادفی تعریف می‌شود که ممکن است طول این دو بردار متفاوت باشد.

تعریف ۳.۶.۲. کوواریانس فاصله‌ای بین بردارهای تصادفی $\mathbf{u} \in \mathbb{R}^q$ و $\mathbf{v} \in \mathbb{R}^r$، با گشتاورهای مرتبه اول متناهی، یک عدد نامنفی به صورت

$$dcov(\mathbf{u}, \mathbf{v}) = \Big[\int_{\mathbb{R}^{q+r}} \| \phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s}) - \phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s}) \|^2 \, \omega(\mathbf{t}, \mathbf{s}) d\mathbf{t} d\mathbf{s}\Big]^{\frac{1}{2}} \tag{۴۷.۲}$$

است که $\phi_{\mathbf{u}}(\mathbf{t})$ و $\phi_{\mathbf{v}}(\mathbf{s})$، به ترتیب، توابع مشخصه حاشیه‌ای $\mathbf{u}$ و $\mathbf{v}$، $\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s})$ تابع مشخصه توام $(\mathbf{u}, \mathbf{v})$ و $\omega(\mathbf{t}, \mathbf{s}) = \{c_q c_r \| \mathbf{t} \|_q^{q+1} \| \mathbf{s} \|_r^{r+1}\}^{-1}$ که $c_d = \pi^{(1+d)/2}/\Gamma\{\pi^{(1+d)/2}\}$. همچنین، $\| \mathbf{a} \|_d$ نرم اقلیدسی بردار $\mathbf{a} \in \mathbb{R}^d$ و $\| \phi \|^2 = \phi\bar{\phi}$ است که ϕ یک تابع مختلط و $\bar{\phi}$ مزدوج ϕ می‌باشد.

تعریف ۴.۶.۲. همبستگی فاصله‌ای بین دو بردار تصادفی $\mathbf{u} \in \mathbb{R}^q$ و $\mathbf{v} \in \mathbb{R}^r$، با گشتاورهای مرتبه اول متناهی به صورت

$$dcorr(\mathbf{u}, \mathbf{v}) = \begin{cases} \frac{dcov(\mathbf{u},\mathbf{v})}{\sqrt{dcov(\mathbf{u},\mathbf{u})dcov(\mathbf{v},\mathbf{v})}}, & dcov(\mathbf{u}, \mathbf{u})dcov(\mathbf{v}, \mathbf{v}) > 0 \\ 0, & dcov(\mathbf{u}, \mathbf{u})dcov(\mathbf{v}, \mathbf{v}) = 0. \end{cases}$$

[62] Consistency in ranking

تعریف می‌شود و در شرط $۰ \leq dcorr(\mathbf{u},\mathbf{v}) \leq ۱$ صدق می‌کند. در حالت خاص، اگر X و Y دو متغیر تصادفی نرمال یک متغیره با ضریب همبستگی پیرسن ρ و ضریب همبستگی فاصله‌ای $\mathcal{R}$ باشند، آنگاه $\mathcal{R} \leq |\rho|$. تساوی هنگامی برقرار است که $|\rho| = ۱$.

سزکلی[۶۳] و همکاران (۲۰۰۷) نشان دادند

$$dcov^{۲}(\mathbf{u},\mathbf{v}) = S_{۱} + S_{۲} - ۲S_{۳} \qquad (۴۸.۲)$$

به‌طوری‌که $S_{۱} = E\{\| \mathbf{u}-\tilde{\mathbf{u}} \|_q \| \mathbf{v}-\tilde{\mathbf{v}} \|_r\}$، $S_{۲} = E\{\| \mathbf{u}-\tilde{\mathbf{u}} \|_q\} E\{\| \mathbf{v}-\tilde{\mathbf{v}} \|_r\}$، $S_{۳} = E\{E(\| \mathbf{u}-\tilde{\mathbf{u}} \|_q |\mathbf{u}) E(\| \mathbf{v}-\tilde{\mathbf{v}} \|_r |\mathbf{v})\}$ و $(\tilde{\mathbf{u}},\tilde{\mathbf{v}})$ بردارهای تصادفی مستقل از $(\mathbf{u},\mathbf{v})$ و هم‌توزیع با آن‌ها هستند. بنابراین کوواریانس فاصله‌ای بین $\mathbf{u}$ و $\mathbf{v}$ را می‌توان به صورت

$$\widehat{dcov}^{۲}(\mathbf{u},\mathbf{v}) = \hat{S}_{۱} + \hat{S}_{۲} - ۲\hat{S}_{۳}$$

برآورد کرد که

$$\begin{aligned}
\hat{S}_{۱} &= \tfrac{۱}{n^{۲}} \textstyle\sum_{i=۱}^{n}\sum_{j=۱}^{n} \| \mathbf{u}_i-\mathbf{u}_j \|_q \| \mathbf{v}_i-\mathbf{v}_j) \|_r \\
\hat{S}_{۲} &= \tfrac{۱}{n^{۲}} \textstyle\sum_{i=۱}^{n}\sum_{j=۱}^{n} \| \mathbf{u}_i-\mathbf{u}_j \|_q \tfrac{۱}{n^{۲}} \sum_{i=۱}^{n}\sum_{j=۱}^{n} \| \mathbf{v}_i-\mathbf{v}_j \|_r \\
\hat{S}_{۳} &= \tfrac{۱}{n^{۳}} \textstyle\sum_{i=۱}^{n}\sum_{j=۱}^{n}\sum_{l=۱}^{n} \| \mathbf{u}_i-\mathbf{u}_l \|_q \| \mathbf{v}_j-\mathbf{v}_l \|_r .
\end{aligned}$$

همچنین کوواریانس‌های فاصله‌ای نمونه‌ای $dcov(\mathbf{u},\mathbf{u})$ و $dcov(\mathbf{v},\mathbf{v})$ را نیز می‌توان به‌طور مشابه برآورد کرد. در نتیجه همبستگی فاصله‌ای نمونه‌ای بین $\mathbf{u}$ و $\mathbf{v}$ برابر است با

$$\widehat{dcorr}(\mathbf{u},\mathbf{v}) = \frac{\widehat{dcov}(\mathbf{u},\mathbf{v})}{\sqrt{\widehat{dcov}(\mathbf{u},\mathbf{u})}\sqrt{\widehat{dcov}(\mathbf{v},\mathbf{v})}}. \qquad (۴۹.۲)$$

ضریب همبستگی فاصله‌ای دارای دو ویژگی مطلوب زیر است. ویژگی اول، رابطه بین ضریب همبستگی فاصله‌ای و ضریب همبستگی پیرسن است. ضریب همبستگی فاصله‌ای دو متغیر تصادفی نرمال یک متغیره تابعی اکیداً صعودی از قدر مطلق ضریب همبستگی پیرسن آن دو متغیر نرمال است (سزکلی و همکاران، ۲۰۰۷). ویژگی دوم بیان می‌کند که بردارهای تصادفی $\mathbf{u}$ و $\mathbf{v}$ مستقل‌اند اگر و تنها اگر $dcorr(\mathbf{u},\mathbf{v}) = ۰$ (سزکلی و همکاران، ۲۰۰۷؛ سزکلی و ریزو[۶۴]، ۲۰۰۹).

به دلیل دو ویژگی فوق، لی و همکاران (۲۰۱۲) از همبستگی فاصله‌ای برای رتبه‌بندی متغیرهای توضیحی استفاده کردند و یک روش غربالگری مستقل مطمئن موسوم به DC-SIS را ارائه دادند. آن‌ها از ضریب همبستگی فاصله‌ای بین متغیرهای توضیحی و متغیر پاسخ برای رتبه‌بندی متغیرها استفاده کردند و معیار مطلوبیت حاشیه‌ای را به صورت زیر تعریف کردند

$$\omega_k = dcorr^{۲}(X_k, Y), \qquad k = ۱, ..., p. \qquad (۵۰.۲)$$

[63] Székely
[64] Rizzo

ویژگی اول ضریب همبستگی فاصله‌ای نتیجه می‌دهد که در صورت نرمال بودن توزیع خطا و توزیع متغیرهای توضیحی، DC-SIS معادل با SIS است، یعنی در مدل خطی گاوسی در نظر گرفته‌شده توسط فن و لیو (۲۰۰۸)، DC-SIS به خوبی SIS عمل می‌کند. ویژگی دوم، آزاد_مدل بودن روش غربالگری DC-SIS را نتیجه می‌دهد، یعنی در این روش نیازی به در نظر گرفتن یک مدل مفروض برای متغیرهای توضیحی و پاسخ نیست. ویژگی آزاد_مدل بودن بسیار حائز اهمیت است، چون تعیین یک مدل پارامتری معین برای داده‌های با بعد بسیار بالا یک چالش است.

برای پیاده‌سازی این روش غربالگری، کافی است معیار مطلوبیت حاشیه‌ای (۵۰.۲) بر اساس یک نمونه تصادفی برآورد شود. فرض کنید $(\mathbf{X}_i, Y_i)_{i=1}^n$ یک نمونه تصادفی از $(\mathbf{X}, Y)$ باشد. بنابراین معیار (۵۰.۲) را به‌صورت

$$\widehat{\omega_k} = \widehat{dcorr}^2(X_k, Y)$$

برآورد می‌کنیم. سپس، مجموعه‌ای از متغیرهای توضیحی که دارای مقادیر بزرگ $\widehat{\omega_k}$ هستند، به‌عنوان مجموعه متغیرهای مهم انتخاب می‌شوند.

لی و همکاران (۲۰۱۲) نشان دادند که روش DC-SIS دارای ویژگی غربالگری مطمئن است. این روش دارای چندین مزیت است: الف) برای پیاده‌سازی این روش نیازی به استفاده از هیچ الگوریتم بهینه‌سازی نیست. ب) این روش را می‌توان مستقیماً برای پاسخ چندگانه یا متغیرهای توضیحی با ساختار گروهی به کار برد. ج) این روش برای هر نوع متغیر پاسخ پیوسته، گسسته یا شمارشی قابل استفاده است. بنابراین DC-SIS یک روش آزاد_مدل مناسب برای داده‌های با بعد بسیار بالا است.

در سال‌های اخیر مطالعات زیادی در زمینه غربالگری متغیرها انجام شده است. از جمله این کارها می‌توان هال و میلر[۶۵] (۲۰۰۹)، بولمن و همکاران (۲۰۱۰)، بارت[۶۶] و همکاران (۲۰۱۶)، وانگ[۶۷] و لنگ (۲۰۱۶)، کنگ[۶۸] و همکاران (۲۰۱۷)، چن[۶۹] و همکاران (۲۰۱۸)، هوانگ و ژو (۲۰۱۶)، لیو[۷۰] (۲۰۱۶)، لی و همکاران (۲۰۱۸)، چن (۲۰۱۸) و سانگ و همکاران (۲۰۱۸) را نام برد.

۷.۲ الگوریتم محاسباتی روش کمترین توان‌های دوم جریمه‌شده

تابع کمترین توان‌های دوم جریمه‌شده زیر را در نظر بگیرید

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{۵۱.۲}$$

با یک تابع جریمه محدب، مانند لاسو، تابع هدف (۵۱.۲) یک تابع محدب است و لذا برآوردهای کمترین توان‌های دوم جریمه‌شده را می‌توان با بهینه‌سازی محدب به‌دست آورد. اما اغلب توابع جریمه ای که

[65] Hall and Miller
[66] Barut
[67] Wang
[68] Kong
[69] Chen
[70] Liu

دارای سه ویژگی مطلوب تنکی، نااریبی و پیوستگی می‌باشند، توابعی غیر محدب بوده و این موضوع باعث می‌شود که تابع هدف (۵۱.۲) محدب نباشد. محدب نبودن این تابع، مسئله مینیمم‌سازی را با مشکل مواجه می‌کند. به هر حال، می‌توان توابع جریمه غیر محدب را با برخی توابع محدب تقریب زد و مسئله بهینه‌سازی غیر محدب را با استفاده از الگوریتم‌های بهینه‌سازی محدب حل کرد. در این بخش دو تقریب زیر را برای حل مشکل فوق معرفی می‌کنیم.

۱.۷.۲ تقریب درجه دو موضعی

چون عبارت اول در تابع کمترین توان‌های دوم جریمه‌شده (۵۱.۲) یک تابع محدب است، لذا کافی است تابع جریمه غیر محدب $P_\lambda(|\beta_j|)$ را با یک تابع محدب تقریب کنیم. در این راستا، فن و لی (۲۰۰۱) الگوریتم تقریب درجه دو موضعی (LQA) را برای توابع جریمه غیر محدب ارائه دادند. فرض کنید $\boldsymbol{\beta}^{(0)} = (\beta_1^{(0)}, ..., \beta_p^{(0)})^T$ یک مقدار اولیه برای $\boldsymbol{\beta}$ باشد. اگر $\beta_j^{(0)}$ نزدیک به صفر باشد، پس قرار دهید $\hat{\beta}_j = 0$، در غیر اینصورت با استفاده از تقریب موضعی داریم

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\frac{\beta_j}{|\beta_j|} \approx p'_\lambda(|\beta_j^{(0)}|)\frac{\beta_j}{|\beta_j^{(0)}|}, \qquad \beta_j \approx \beta_j^{(0)}. \tag{۵۲.۲}$$

با انتگرال گیری از طرفین نسبت به β_j از $\beta_j^{(0)}$ تا β_j داریم

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}\left\{\beta_j^2 - \beta_j^{(0)2}\right\}. \tag{۵۳.۲}$$

بنابراین

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}) &= \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \\
&\approx \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{p} \left\{ p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}\left(\beta_j^2 - \beta_j^{(0)2}\right)\right\} \\
&= \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{p} \frac{1}{2}\frac{p'_\lambda(|\beta_j^{(0)}|)}{|\beta_j^{(0)}|}\beta_j^2 + c
\end{aligned} \tag{۵۴.۲}$$

که ثابت c تابعی از $\boldsymbol{\beta}^{(0)}$ است. با استفاده از تقریب فوق و حذف جملات ثابت، تابع هدف (۵۱.۲) به یک تابع درجه دو تبدیل می‌شود. در نتیجه، برآورد کمترین توان‌های دوم جریمه‌شده را می‌توان با محاسبه تکراری رگرسیون ریج به صورت

$$\boldsymbol{\beta}^{(1)} = \left\{\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(0)})\right\}^{-1}\mathbf{X}^T\mathbf{Y} \tag{۵۵.۲}$$

به‌دست آورد که

$$\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(0)}) = \text{diag}\left\{\frac{p'_\lambda(|\beta_1^{(0)}|)}{|\beta_1^{(0)}|}, \ldots, \frac{p'_\lambda(|\beta_p^{(0)}|)}{|\beta_p^{(0)}|}\right\}.$$

با توجه به این واقعیت که رگرسیون ریج نمی‌تواند متغیرهای معنی‌دار را انتخاب کند، ضرایب رگرسیونی برآورد شده را که بسیار نزدیک به صفر می‌باشند، برابر صفر قرار می‌دهیم تا متغیرهای بی‌اهمیت از مدل حذف شوند.

یکی از مزایای تقریب LQA تعمیم‌پذیری آن به هر تابع زیان هموار دیگری (غیر از تابع زیان کمترین توان‌های دوم) است. با در نظر گرفتن تابع زیان $l(\boldsymbol{\beta})$ تابع هدف جریمه‌شده به صورت

$$l(\boldsymbol{\beta}) + \sum_{j=1}^{p} P_\lambda(|\beta_j|) \qquad (۵۶.۲)$$

تعریف می‌شود. در این حالت، ابتدا تابع زیان $l(\boldsymbol{\beta})$ را با یک تابع درجه دو به صورت

$$l(\boldsymbol{\beta}) \approx l(\boldsymbol{\beta}^{(0)}) + \nabla l(\boldsymbol{\beta}^{(0)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^T \nabla^2 l(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) \qquad (۵۷.۲)$$

تقریب می‌کنیم که

$$\nabla l(\boldsymbol{\beta}^{(0)}) = \frac{\partial l(\boldsymbol{\beta}^{(0)})}{\partial \boldsymbol{\beta}}, \qquad \nabla^2 l(\boldsymbol{\beta}^{(0)}) = \frac{\partial^2 l(\boldsymbol{\beta}^{(0)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}.$$

با حذف جمله ثابت، تابع هدف (۵۱.۲) را می‌توان به صورت

$$\nabla l(\boldsymbol{\beta}^{(0)})^T(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^T \nabla^2 l(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \frac{1}{2}\boldsymbol{\beta}^T \Sigma_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta} \qquad (۵۸.۲)$$

نوشت و سپس با استفاده از الگوریتم نیوتن ـ رافسون مینیم تابع (۵۱.۲) را به‌دست آورد. به‌طور دقیق‌تر، با اجرای مراحل زیر، الگوریتم LQA را می‌توان برای تقریب تابع (۵۶.۲) به‌کار برد:

- مرحله اول: مقدار اولیه $\boldsymbol{\beta}^{(0)}$ را برای $\boldsymbol{\beta}$ انتخاب کنید. این مقدار اولیه می‌تواند برآورد OLS باشد.
- مرحله دوم: تابع هدف (۵۶.۲) را بر اساس (۵۸.۲) تقریب کنید.
- مرحله سوم: الگوریتم نیوتن ـ رافسون را برای محاسبه مقدار مینیمم‌کننده تابع درجه دو به‌دست آمده در مرحله دوم به کار برید تا مقدار $\beta_j^{(0)}$ به‌روزرسانی شود. سپس اگر $|\beta_j^{(0)}| < \epsilon$ باشد، متغیر X_j را از مدل حذف کنید.
- مرحله چهارم: مراحل دوم و سوم را تا رسیدن به همگرایی تکرار کنید.

فن و لی (۲۰۰۱) نشان دادند که در صورت انتخاب یک مقدار مناسب برای $\boldsymbol{\beta}^{(0)}$، الگوریتم LQA پس از تعداد کمی تکرار به همگرایی می رسد. علاوه براین، با یک مقدار اولیه مناسب، کارایی الگوریتم LQA با یک تکرار معادل با تکرار کامل الگوریتم تا رسیدن به هگرایی است. اما ایراد وارد بر این الگوریتم این است که اگر یک متغیر در هر تکرار حذف شود، در تکرار بعدی وارد مدل نمی‌شود و در نتیجه از مدل نهایی حذف می‌شود. بنابراین، این روش مشابه روش انتخاب پیشرو عمل می کند.

۲.۷.۲ تقریب خطی موضعی

برای غلبه بر مشکل بیان شده در تقریب LQA، زو[۷۱] و لی (۲۰۰۸) الگوریتم تقریب خطی موضعی (LLA) را برای تقریب توابع جریمه غیر محدب پیشنهاد دادند. با استفاده از این تقریب برای تابع جریمه

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + p'_\lambda(|\beta_j^{(0)}|)\left\{|\beta_j| - |\beta_j^{(0)}|\right\}, \qquad \beta_j \approx \beta_j^{(0)} \tag{۵۹.۲}$$

می‌توان تابع هدف (۵۱.۲) را به‌صورت

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}) &= \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \\ &\approx \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{p} \left\{p_\lambda(|\beta_j^{(0)}|) + p'_\lambda(|\beta_j^{(0)}|)\big(|\beta_j| - |\beta_j^{(0)}|\big)\right\} \\ &= \frac{1}{2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)|\beta_j| + c \end{aligned} \tag{۶۰.۲}$$

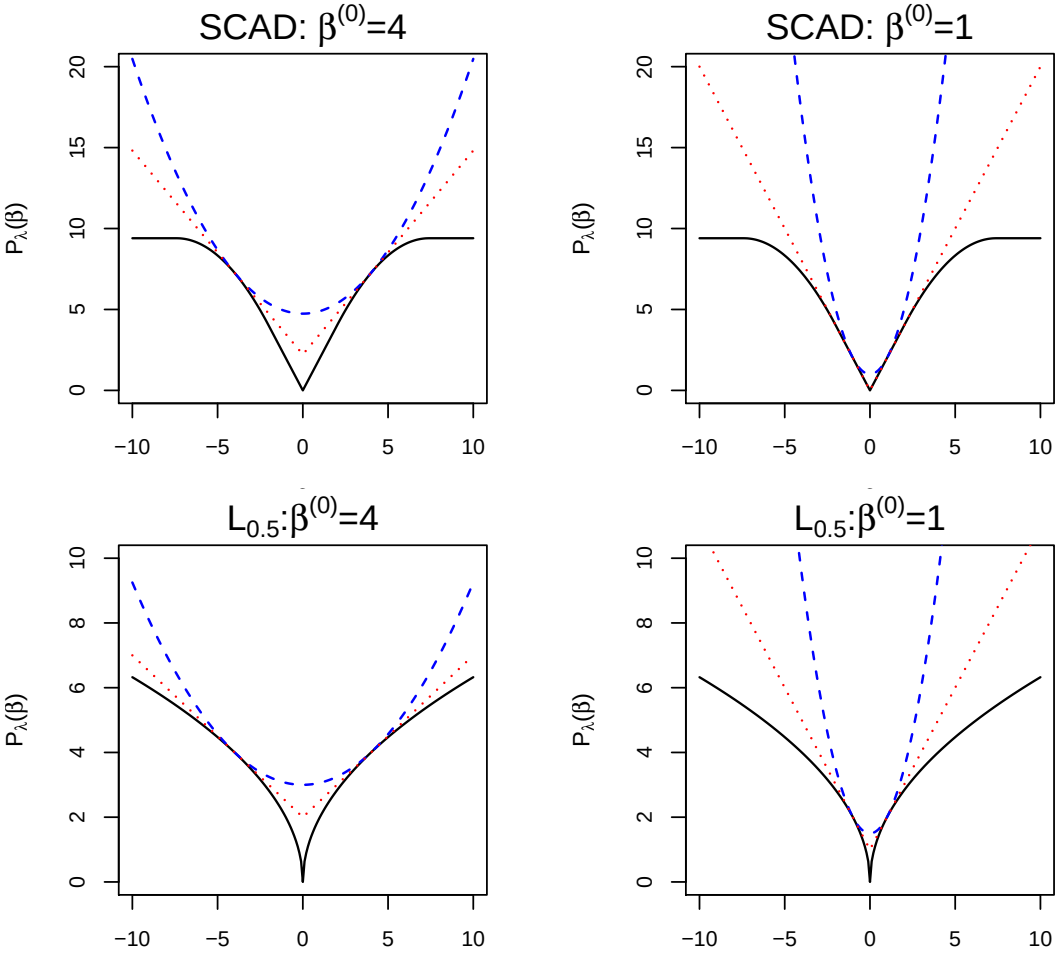
تقریب کرد که ثابت c تابعی از $\boldsymbol{\beta}^{(0)}$ است. بنابراین تقریب LLA تابع هدف جریمه‌شده غیر محدب را به تابع هدف لاسوی وزنی تبدیل می‌کند. این تابع هدف تبدیل‌یافته را می‌توان با استفاده از برنامه‌ریزی درجه دو یا الگوریتم کمترین زاویه حل کرد.

شکل ۸.۲ تقریب LQA و LLA را برای توابع جریمه $L_{0.5}$ و SCAD نشان می‌دهد. در این شکل برای هر دو تابع جریمه $\lambda = 2$ و دو مقدار اولیه مختلف در نظر گرفته شده‌است.

تاکنون، علاوه بر دو تقریب فوق، الگوریتم‌های دیگری نیز برای حل مسئله کمترین توان‌های دوم جریمه‌شده معرفی شده‌اند. الگوریتم پیش‌برنده[۷۲] (فو[۷۳]، ۱۹۹۸)، الگوریتم مختصات نزولی[۷۴] برای لاسوی گروهی (وو و لانگ[۷۵]، ۲۰۰۸)، الگوریتم ماکزیمم‌سازی شرطی تکراری[۷۶] (ژانگ و لی، ۲۰۰۹)، الگوریتم مختصات نزولی برای توابع جریمه غیرمحدب (برهنی و هوانگ، ۲۰۱۱) و الگوریتم مختصات نزولی گروهی (برهنی و هوانگ، ۲۰۱۵) نمونه‌هایی از این الگوریتم‌ها می‌باشند.

---

[71] Zou
[72] Shooting algorithm
[73] Fu
[74] Coordinate descent algorithm
[75] Wu and Lang
[76] Iterative conditional maximization algorithm



شکل ۸.۲: تقریب درجه دو موضعی و تقریب خطی موضعی برای توابع جریمه $L_{0.5}$ و SCAD. نمودار آبی تقریب درجه دو موضعی، نمودار قرمز تقریب خطی موضعی و نمودار مشکی تابع جریمه را نشان می‌دهد.

فصل ۳

انتخاب متغیر و تشخیص ساختار در مدل جمعی خطی- جزئی

در این فصل، یک روش دو مرحله‌ای برای انتخاب متغیر و تشخیص مؤلفه‌های خطی و غیرخطی در مدل جمعی با بعد بسیار بالا معرفی می‌شود. در مرحله اول، با استفاده از یک روش غربالگری آزاد- مدل، بعد فضای متغیرهای توضیحی را کاهش داده و سپس در مرحله دوم، از روشی مبتنی بر دو تابع جریمه برای انتخاب همزمان مؤلفه‌های غیرصفر و خطی استفاده می‌شود. از این فصل، مقاله‌های کاظمی و همکاران (۱۳۹۷)، کاظمی و همکاران (۲۰۱۸) و کاظمی و همکاران (۲۰۱۹a) استخراج شده است.

۱.۳ مقدمه

مدل جمعی ناپارامتری را به‌صورت

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \qquad 1 \le i \le n \tag{۱.۳}$$

در نظر بگیرید که در آن Y_i متغیر پاسخ، $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ بردار متغیرهای توضیحی و f_jها توابع یک متغیره هموار[1] نامعلوم هستند. همچنین $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ بردار خطای تصادفی با میانگین

[1] تابع هموار تابعی است که بر روی یک بازه مشخص، تا مرتبه معینی مشتق‌پذیر باشد.

صفر و $E(\varepsilon\varepsilon^T) = \sigma^2\mathbf{I}_n$ است. از آنجا که در عمل ممکن است تعداد زیادی از متغیرهای توضیحی دارای اثر خطی یا حتی بدون اثر روی متغیر پاسخ باشند و بقیه متغیرها به‌صورت غیرخطی وارد مدل شوند، اپسومر و راپرت[۲] (۱۹۹۹) مدل جمعی خطی–جزئی را به‌صورت

$$Y_i = \sum_{j\epsilon S_1} \beta_j X_{ij} + \sum_{j\epsilon S_2} f_j(X_{ij}) + \varepsilon_i, \qquad 1 \le i \le n \tag{۲.۳}$$

معرفی کردند، که در آن S_1 و S_2 دو زیر مجموعه مجزا از $\{1, ..., p\}$ است. این مدل دو ویژگی مثبت تفسیرپذیری مدل‌های خطی و انعطاف‌پذیری مدل‌های ناپارامتری را با هم ترکیب می‌کند. از نظر آماری، مؤلفه‌های پارامتری دارای همگرایی سریع‌تر از مؤلفه‌های ناپارامتری هستند. در نتیجه غیرخطی در نظر گرفتن مؤلفه‌هایی که ذاتاً خطی هستند، کارآیی برآورد را کاهش می‌دهد. بنابراین استفاده از مدل‌های جمعی خطی–جزئی در برخی از مسائل، نسبت به مدل‌های جمعی ناپارامتری، مناسب‌تر است. تاکنون بسیاری از محققین به مسئله برآورد و انتخاب متغیر در مدل جمعی خطی–جزئی پرداخته‌اند که از جمله می‌توان به لیو[۳] و همکاران (۲۰۱۱)، لیان (۲۰۱۲a)، گو[۴] و همکاران (۲۰۱۳)، دو[۵] و همکاران (۲۰۱۵)، روزبه (۲۰۱۵)، روزبه (۲۰۱۶)، آرشی و روزبه (۲۰۱۶)، لیو[۶] و همکاران (۲۰۱۷) و اکدنیز[۷] و همکاران (۲۰۱۸) اشاره کرد.

مهمترین پذیره در استفاده از مدل (۲.۳) این است که اجزای خطی و غیرخطی مدل، از پیش تعیین شده باشند. اما در عمل چنین اطلاعات پیشین به ندرت در دسترس است، به ویژه وقتی که تعداد متغیرهای توضیحی زیاد است، یا به عبارتی مدلی با بعد بالا داشته باشیم. بنابراین، علاوه بر تشخیص مؤلفه‌های غیرصفر، لازم است مؤلفه‌های خطی و غیرخطی نیز شناسایی شوند.

یک روش ممکن برای تشخیص ساختار خطی–جزئی در مدل جمعی، استفاده از آزمون فرض است، یعنی ابتدا مدل را به‌صورت جمعی ناپارامتری (۱.۳) در نظر گرفته و سپس با انجام آزمون مؤلفه به مؤلفه اثرات خطی و غیرخطی را شناسایی می‌کنیم (چن[۸] و همکاران، ۲۰۱۱). روش دیگر برای تشخیص اثرات ثابت و متغیر، به‌کارگیری دستورالعمل زیر است: ابتدا همه متغیرهای توضیحی پیوسته را در قسمت ناپارامتری و همه متغیرهای توضیحی گسسته را در قسمت پارامتری قرار دهید. اگر نتایج برآورد نشان دهد که اثر برخی از متغیرهای پیوسته را می‌توان با توابع پارامتری مشخص، مانند مدل خطی، توصیف کرد، با قرار دادن آن متغیرهای پیوسته در قسمت پارامتری، مدل جدید را برازش دهید (لیو[۹] و همکاران، ۲۰۱۱). به‌کارگیری روش‌های مذکور در داده‌های با بعد بالا بسیار طاقت‌فرسا است. روش متداول برای انتخاب متغیر و تشخیص ساختار در مدل‌های جمعی با بعد بالا استفاده از روش‌های جریمه است.

[2] Opsomer and Ruppert
[3] Liu
[4] Guo
[5] Du
[6] Lv
[7] Akdeniz
[8] Chen
[9] Liu

ژانگ و همکاران (۲۰۱۱) از دو تابع جریمه برای شناسایی همزمان مؤلفه‌های خطی و غیرخطی در مدل جمعی استفاده کردند، اما نتوانستند ویژگی سازگاری در انتخاب را برای روش ارائه شده ثابت کنند. سپس هوانگ و همکاران (۲۰۱۲a) روشی موسوم به تعقیب رگرسیون نیمه پارامتری را برای تشخیص مؤلفه‌های خطی و غیرخطی با استفاده از تابع جریمه MCP ارائه دادند و ویژگی سازگاری را برای روش ارائه شده ثابت کردند، اما به موضوع انتخاب متغیر نپرداختند. لیان (۲۰۱۲b) روشی برای تشخیص مؤلفه‌های خطی با استفاده از تابع جریمه SCAD معرفی کرد، اما این روش انتخاب متغیر را انجام نمی‌دهد. لیان (۲۰۱۲c) با به کاربردن دو تابع جریمه SCAD توانست مؤلفه‌های خطی و غیرصفر را به طور همزمان تشخیص دهد. سپس لیان و همکاران (۲۰۱۲) دو تابع جریمه SCAD را برای رگرسیون چندکی به‌کار بردند و یک روش نیرومند برای انتخاب متغیر و تشخیص ساختار ارائه دادند. همچنین لیان و همکاران (۲۰۱۵) با استفاده از دو تابع جریمه LASSO گروهی تطبیقی به شناسایی مؤلفه‌های خطی و غیرصفر پرداختند.

مطالعات انجام شده بر اساس توابع جریمه که در بالا به آن‌ها اشاره شد، برای انتخاب متغیر در داده‌های با بعد بسیار بالا، به دلایل هزینه محاسباتی، دقت آماری و ناپایداری الگوریتمی، مناسب نیستند و لذا شناسایی متغیرهای مهم در این نوع داده‌ها و تشخیص نوع تاثیر آن‌هابر متغیر پاسخ به لحاظ خطی یا غیرخطی بودن، همواره مورد توجه محققین است.

به‌طور کلی، مسئله انتخاب متغیر و تشخیص ساختار در بعد بالا، بر اساس رهیافت توابع جریمه، یک روش دو مرحله‌ای است که در مرحله اول، ابتدا بعد مدل توسط یک روش غربالگری مستقل کاراَ کاهش یافته و در مرحله دوم از روشی مبتنی بر دو تابع جریمه برای شناسایی مؤلفه‌های غیرصفر و خطی در زیر مدل غربال‌شده استفاده می‌شود.

در این فصل ضمن رعایت الگوی متعارف دو مرحله‌ای ذکر شده، ابتدا ایدهٔ خود را برای کاهش بعد مدل، در قالب یک روش SIS بر اساس همبستگی فاصله‌ای بین متغیرهای توضیحی و تابع توزیع حاشیه‌ای متغیر پاسخ ارائه می‌دهیم. این روش، آزاد ــ مدل بوده و نشان می‌دهیم وقتی که متغیر پاسخ دم سنگین، چوله یا دارای مقادیر فرین است، عملکرد خوبی دارد. سپس دو تابع جریمه SCAD گروهی را برای تشخیص همزمان متغیرهای مهم و مؤلفه‌های خطی در زیر مدل غربال‌شده در مرحله قبل به‌کار می‌بریم. سرانجام عملکرد این روش را با مطالعات شبیه‌سازی و مثال واقعی مورد بررسی قرار می‌دهیم.

۲.۳ غربالگری مستقل مطمئن نیرومند بر اساس همبستگی فاصله‌ای

در این فصل، اگرچه هدف اصلی ما تمرکز بر مدل (۲.۳) است، اما از آنجایی که در مسائل واقعی مؤلفه‌های خطی و غیرخطی مشخص نیستند، ابتدا فرض می‌شود که مدل به‌صورت جمعی ناپارامتری (۱.۳) است، سپس روشی ارائه خواهد شد تا مؤلفه‌های خطی در مدل (۱.۳) تعیین شوند و صورت کلی رابطه (۲.۳) بدست آید. برای این منظور، ابتدا با استفاده از یک روش غربالگری نیرومند، بعد فضای متغیرها به بعد کوچکتر از n کاهش می‌یابد. این روش غربالگری بر اساس همبستگی فاصله‌ای بین متغیرهای توضیحی و تابع توزیع حاشیه‌ای متغیر پاسخ تعریف می‌شود. استفاده از تابع توزیع متغیر پاسخ به جای خود متغیر

پاسخ باعث می‌شود که روش ارائه شده نیرومند باشد. دلیل استفاده از روش غربالگری این است که کاربرد مستقیم روش‌های جریمه برای انتخاب متغیر و برآورد در مدل‌های با بعد بسیار بالا کارایی لازم را ندارند (فن و همکاران، ۲۰۰۹). بنابراین لازم است ابتدا با یک روش غربالگری بعد مدل را کاهش داده، سپس از توابع جریمه برای تشخیص ساختار مدل استفاده شود.

سزکلی و همکاران (۲۰۰۷) ضریب همبستگی فاصله‌ای را به عنوان یک معیار وابستگی بین دو بردار تصادفی معرفی کردند. به دلیل ویژگی های مطلوب ضریب همبستگی فاصله ای، که در زیر بخش ۳.۶.۲ به آن‌ها اشاره شد، لی و همکاران (۲۰۱۲) یک روش غربالگری مستقل مطمئن موسوم به DC-SIS را برای رتبه‌بندی متغیرهای توضیحی با استفاده از همبستگی فاصله‌ای آن‌هابا متغیر پاسخ ارائه دادند. در زیر بخش ۳.۶.۲ به جزئیات این روش غربالگری پرداخته شده است.

فرض کنید $\mathbf{X} = (X_1, ..., X_p)^T$ بردار متغیرهای توضیحی و Ψy تکیه‌گاه متغیر پاسخ Y باشد. مجموعه اندیس متغیرهای مهم را به‌صورت

$$\mathbb{M} = \{j : \text{است} X_j \text{ به‌طور تابعی وابسته به } F(y|\mathbf{X}) \text{، } y \in \Psi y \text{ برای برخی مقادیر}\}$$

در نظر بگیرید که $F(y|\mathbf{X}) = P(Y < y|\mathbf{X})$ تابع توزیع شرطی Y به شرط $\mathbf{X}$ است. تعریف کنید $s_n = |\mathbb{M}|$ و فرض کنید $s_n < p$ که مقدار s_n ممکن است با n تغییر کند. در این فصل برای اندازه‌گیری همبستگی بین متغیرهای توضیحی و متغیر پاسخ، از $F(Y)$ به جای Y در روش لی و همکاران (۲۰۱۲) استفاده می‌شود، یعنی معیار مطلوبیت حاشیه‌ای برای رتبه‌بندی متغیرها به‌صورت

$$\omega_j = dcorr\big(X_j, F(Y)\big) = \frac{dcov\big(X_j, F(Y)\big)}{\sqrt{dcov\big(X_j, X_j\big)}\sqrt{dcov\big(F(Y), F(Y)\big)}}, \quad ۱ \le j \le p \quad (۳.۳)$$

تعریف می‌شود، که در آن $F(y)$ تابع توزیع حاشیه‌ای Y و $dcov\big(U, V\big)$ کوواریانس فاصله‌ای بین U و V است که نحوه محاسبه آن در زیر بخش ۳.۶.۲ داده شده است. این روش غربالگری نسبت به روش‌های موجود دارای دو مزیت است:

الف) بنا به ویژگی همبستگی فاصله‌ای، X_j و $F(Y)$ مستقل‌اند، اگر و تنها اگر $dcorr(X_j, F(Y)) = ۰$. بنابراین، این روش آزاد–مدل بوده و برای غربالگری متغیرها نیازی به مشخص کردن ساختار مدل نیست.

ب) چون $F(Y)$ برای هر نوع متغیر پاسخ یک تابع کراندار است، می‌توان انتظار داشت که این روش برای متغیر پاسخ دم سنگین یا دارای مقادیر فرین، به دلیل جایگزینی Y با $F(Y)$، نسبت به DC-SIS عملکرد بهتری داشته باشد.

برای پیاده‌سازی این روش غربالگری، کافی است معیار مطلوبیت حاشیه‌ای (۳.۳) بر اساس یک نمونه تصادفی برآورد شود. فرض کنید $\big(\mathbf{X}_i, Y_i\big)_{i=۱}^n$ یک نمونه تصادفی از مدل جمعی ناپارامتری (۱.۳) باشد. با استفاده از رابطه (۴۸.۲)، نتیجه می شود

$$\widehat{dcov}^۲(X_j, F(Y)) = \hat{S}_{j,۱} + \hat{S}_{j,۲} - ۲\hat{S}_{j,۳}$$

که در آن

$$\hat{S}_{j,۱} = \frac{۱}{n^۲}\sum_{i=۱}^{n}\sum_{k=۱}^{n}|X_{ij}-X_{kj}||F_n(Y_i)-F_n(Y_k)|$$
$$\hat{S}_{j,۲} = \frac{۱}{n^۲}\sum_{i=۱}^{n}\sum_{k=۱}^{n}|X_{ij}-X_{kj}|\frac{۱}{n^۲}\sum_{i=۱}^{n}\sum_{k=۱}^{n}|F_n(Y_i)-F_n(Y_k)|$$
$$\hat{S}_{j,۳} = \frac{۱}{n^۳}\sum_{i=۱}^{n}\sum_{k=۱}^{n}\sum_{l=۱}^{n}|X_{ij}-X_{lj}||F_n(Y_k)-F_n(Y_l)|$$

و $F_n(y) = \frac{۱}{n}\sum_{i=۱}^{n} I(Y_i \le y)$ تابع توزیع تجربی Y است. به‌طور مشابه، می‌توان کوواریانس‌های فاصله‌ای نمونه‌ای $dcov\big(X_j, X_j\big)$ و $dcov\big(F(Y), F(Y)\big)$ را نیز برآورد کرد. در نتیجه

$$\widehat{\omega}_j = \widehat{dcorr}\big(X_j, F(Y)\big) = \frac{\widehat{dcov}\big(X_j, F(Y)\big)}{\sqrt{\widehat{dcov}\big(X_j, X_j\big)}\sqrt{\widehat{dcov}\big(F(Y), F(Y)\big)}}.$$

برای کاهش بعد فضای متغیرها، مجموعه‌ای از متغیرهای توضیحی را که دارای مقادیر بزرگ $\widehat{\omega}_j$ هستند، به عنوان مجموعه متغیرهای مهم انتخاب می‌شوند. به عبارت دیگر

$$\widehat{\mathbb{M}} = \big\{ ۱ \le j \le p : \widehat{\omega}_j \ge cn^{-\kappa} \big\} \qquad (۴.۳)$$

که در آن $c > ۰$ و $۰ \le \kappa < ۱/۲$ مقادیر ثابت از پیش تعیین شده و در شرط C۲ از زیر بخش ۱.۲.۳ صدق می کنند. این روش غربالگری که بعد فضای متغیرها را از مقیاس بسیار بزرگ p به یک فضای بسیار کوچکتر با اندازه $|\widehat{\mathbb{M}}|$ کاهش می‌دهد، غربالگری مستقل مطمئن نیرومند بر اساس همبستگی فاصله‌ای (RDC-SIS) نامیده می‌شود. استفاده از روش غربالگری RDC-SIS مستلزم تعیین یک مقدار آستانه است که در عمل، تعیین مقدار آن معمولاً مشکل است. در راستای یافتن راهکاری مناسب برای این مسئله، یک روش جایگزین، انتخاب d $(d < n)$ متغیر با بیشترین مطلوبیت حاشیه‌ای یا بیشترین همبستگی است. انتخاب مقدار d نقش مهمی را در مرحله غربالگری ایفا می‌کند. فن و لیو (۲۰۰۸) ضریبی از $[n/\log(n)]$، مانند $d_۱ = [n/\log(n)]$، $d_۲ = ۲[n/\log(n)]$ یا $d_۳ = ۳[n/\log(n)]$ را به عنوان یک مقدار مناسب پیشنهاد دادند. البته یک انتخاب محافظه‌کارانه می‌تواند $d = n - ۱$ باشد.

۱.۲.۳ بررسی خواص نظری روش RDC-SIS

در این بخش به اثبات سازگاری برآوردگر $\widehat{\omega}_j$ و ویژگی غربالگری مطمئن روش RDC-SIS می‌پردازیم. بدین منظور، فرض می‌کنیم دو شرط زیر برقرار هستند.

C۱. ثابت های مثبت C و $t_۰$ موجودند به‌طوری‌که

$$\max_{۱\le j\le p} E\{\exp(t|X_j|)\} \le C < \infty, \qquad ۰ < t \le t_۰.$$

C۲. معیار مطلوبیت حاشیه‌ای (۳.۳) به ازای برخی مقادیر ثابت $c > ۰$ و $۰ \le \kappa < ۱/۲$ در نامساوی زیر صدق می‌کند

$$\min_{j\in\mathbb{M}} \omega_j \ge ۲cn^{-\kappa}.$$

قضیه ۱.۲.۳ (ویژگی سازگاری). تحت شرط C۱، برای هر $0 < \gamma \leq \frac{1}{2} - \kappa$، ثابت‌های مثبت c_1 و c_2 موجودند به‌طوری‌که

$$P(\max_{1\leq j\leq p} |\widehat{\omega}_j - \omega_j| \geq cn^{-\kappa}) \leq O\Big(p\Big[\exp\big\{-c_1 n^{1-2(\kappa+\gamma)}\big\} + n\exp(-c_2 n^{\gamma})\Big]\Big).$$

این قضیه نشان می‌دهد که $\widehat{\omega}_j$ یک برآوردگر سازگار برای ω_j است، بدین معنی که با افزایش n، برآورد بدست آمده بسیار نزدیک به پارامتر است، یا احتمال اینکه اختلاف بین برآورد و پارامتر بزرگ باشد، برابر با صفر است.

قضیه ۲.۲.۳ (ویژگی غربالگری مطمئن). تحت شرط‌های C۱ و C۲، داریم

$$P(\mathbb{M} \subseteq \widehat{\mathbb{M}}) \geq 1 - O\Big(s_n\Big[\exp\big\{-c_1 n^{1-2(\kappa+\gamma)}\big\} + n\exp(-c_2 n^{\gamma})\Big]\Big)$$

که $s_n = |\mathbb{M}|$. بنابراین نتیجه می‌شود

$$P(\mathbb{M} \subseteq \widehat{\mathbb{M}}) \to 1 \quad \text{as} \quad n \to \infty.$$

روش‌های غربالگری مستقل با مطلوبیت حاشیه‌ای، تنها از اطلاعات حاشیه‌ای متغیرها به جای مدل کامل استفاده می‌کنند. این موضوع ممکن است عملکرد این روش‌ها را با مشکل مواجه کند. لذا روش RDC-SIS نیز مشابه سایر روش‌های غربالگری مستقل، در صورت عدم برقراری شرایط لازم برای داشتن ویژگی غربالگری مطمئن، ممکن است منجر به حذف متغیرهای مهم از مدل شود. برای رفع این چالش، یک روش غربالگری تکراری را در ادامه معرفی می‌کنیم.

۲.۲.۳ غربالگری مستقل مطمئن تکراری

همانطور که در زیربخش ۱.۶.۲ بیان شد، در برخی از موارد، به کاربردن SIS ممکن است منجر به حذف برخی از متغیرهای مهم شود. فن و لیو (۲۰۰۸) برای رفع مشکل و افزایش کارایی انتخاب متغیر، روش SIS را به‌طور مکرر برای مدل‌های خطی به کار گرفتند و این روش را ISIS نامگذاری کردند. تاثیر روش‌های SIS و ISIS به فرض خطی بودن مدل وابسته است، لذا ایده بهبود SIS توسط ISIS را نمی‌توان مستقیماً به روش غربالگری آزاد–مدل RDC-SIS تعمیم داد، مگر اینکه یک مدل مفروض برای $\mathbf{X}$ و Y در نظر گرفته شود. برای ارائه یک روش تکراری به منظور افزایش کارایی RDC-SIS از یک مثال استفاده می‌کنیم. مدل

$$Y = 5X_1 + 5X_2 + 5X_3 - 15\sqrt{\rho}X_4 + \varepsilon \qquad (5.3)$$

را در نظر بگیرید که خطای ε و هر کدام از متغیرهای توضیحی آن از توزیع $\mathcal{N}(0,1)$ تولید می‌شوند. در اینجا فرض می‌شود که به استثنای X_4، ضریب همبستگی هر متغیر با سایر متغیرها یکسان و برابر با $\rho \neq 0$ است و X_4 دارای همبستگی $\sqrt{\rho}$ با سایر متغیرها است. این مثال از فن و لی (۲۰۰۸) گرفته شده است. در این مثال، متغیر X_4 به‌طور توأم مهم اما به‌طور حاشیه‌ای با Y ناهمبسته است.

بنابراین روش غربالگری RDC-SIS در شناسایی متغیر مهم $X_۴$ ناتوان است. یک روش ممکن برای از بین بردن ناهمبستگی حاشیه‌ای بین $X_۴$ و Y و تقویت تأثیر حاشیه‌ای $X_۴$ روی Y، حذف همبستگی بین $X_۴$ و $(X_۱, X_۲, X_۳)$ است. روش رایج برای از بین بردن این همبستگی، مدل کردن X_j روی $(X_۱, X_۲, X_۳)$ به‌صورت خطی برای $۴ \leq j \leq p$ است. باقیمانده‌های بدست آمده از این رگرسیون‌های خطی با $(X_۱, X_۲, X_۳)$ ناهمبسته هستند. سپس روش RDC-SIS برای این باقیمانده‌ها (به جای متغیر X_j) و Y به کار برده می‌شود. باقیمانده‌های متناظر با $X_۴$ با Y ناهمبسته نیستند. بنابراین $X_۴$ به‌عنوان متغیر مهم انتخاب می‌شود.

با توجه به بحث فوق، یک روش تکراری کلی برای RDC-SIS ارائه می‌شود که شامل سه گام زیر است. فرض کنید $\mathbf{Y} = (Y_۱, ..., Y_n)^T$ و $\mathbf{X}$ ماتریس طرح با بعد $n \times p$ باشد:

- گام ۱: ابتدا روش RDC-SIS را برای $\mathbf{Y}$ و $\mathbf{X}$ به کار برید. فرض کنید در این مرحله $d_۱$ متغیر توضیحی به‌صورت $X_{\mathcal{M}_۱} = \{X_j : j \in \mathcal{M}_۱\}$ انتخاب می‌شوند که $\mathcal{M}_۱$ مجموعه اندیس متغیرهای انتخاب شده با اندازه $d_۱ < d$ و d یک مقدار از پیش تعیین شده است. در اینجا برای سهولت از $d = ۲[n/\log(n)]$ استفاده می‌شود.

- گام ۲: فرض کنید $\mathbf{X}_۱$ ماتریس طرح متناظر با متغیرهای مجموعه $\mathcal{M}_۱$ و $\mathbf{X}_۱^c$ ماتریس طرح متناظر با متغیرهای مجموعه $\mathcal{M}_۱^c$ است. ماتریس $\mathbf{X}_{new} = \{\mathbf{I}_n - \mathbf{X}_۱(\mathbf{X}_۱^T\mathbf{X}_۱)^{-۱}\mathbf{X}_۱^T\}\mathbf{X}_۱^c$ را محاسبه کنید و سپس روش RDC-SIS را برای $\mathbf{Y}$ و تمام ستونهای ماتریس $\mathbf{X}_{new}$ به‌کار برید. فرض کنید در این مرحله $d_۲$ متغیر توضیحی انتخاب می‌شوند و مجموعه اندیس متغیرهای انتخاب شده را با $\mathcal{M}_۲$ نشان می‌دهیم. مجموعه $\mathcal{M}_۱$ را با $\mathcal{M}_۱ \cup \mathcal{M}_۲$ به‌روز رسانی کنید.

- گام ۳: گام ۲ را $k - ۱$ بار تکرار کنید تا تعداد متغیرهای توضیحی انتخاب شده از d تجاوز کند، یعنی $d_۱ + ... + d_k \geq d$. سرانجام، مجموعه متغیرهای انتخاب شده $\mathcal{M}_۱ \cup \cdots \cup \mathcal{M}_k$ است.

از آنجا که متغیرهای حذف شده در مرحله پیشین، مجدداً در مرحله کنونی برای ورود به مدل بررسی می‌شوند، این الگوریتم قادر است احتمال حذف متغیرهای مهم را کاهش دهد. روش غربالگری بالا از ایده تصویرسازی متعامد استفاده می‌کند که در ژو و همکاران (۲۰۱۱) برای غربالگری آزاد ـ مدل استفاده شده است. در اینجا مقدار $d_۱$ در گام ۱ توسط کاربر تعیین می‌شود. در عمل، $d_۱$ به‌عنوان یک پارامتر تنظیم‌کننده در نظر گرفته می‌شود و مقدار بهینه آن با مینیمم کردن میانگین توان دوم خطای پیش‌بینی تعیین می‌گردد.

انتخاب یک مقدار بزرگ‌تر برای d، احتمال انتخاب متغیرهای بی‌اهمیت را برای ورود به مدل افزایش می‌دهد. در ادامه، یک روش مبتنی بر جریمه برای حذف متغیرهای اضافی از مدل و تشخیص مؤلفه‌های خطی و غیرخطی ارائه می‌شود.

۳.۳ انتخاب متغیر و تشخیص ساختار

فرض کنید d متغیر در مرحله غربالگری برای ورود به مدل انتخاب شده‌اند. مدل جمعی ناپارامتری

$$Y_i = \sum_{j=۱}^{d} f_j(X_{ij}) + \epsilon_i, \qquad ۱ \le i \le n$$

را در نظر بگیرید. برای تقریب توابع هموار نامعلوم از توابع پایه‌ای B ـ اسپلاین به‌صورت

$$f_j(x) \approx \sum_{k=۱}^{K} \beta_{jk} B_{jk}(x) \qquad j = ۱, ..., d$$

استفاده می‌شود که K تعداد توابع پایه‌ای استفاده شده است. برای اطلاعات بیشتر درباره توابع B ـ اسپلاین به دی بور[10] (۲۰۰۱) و پیوست پ مراجعه کنید.

یک روش رایج برای برآورد و انتخاب متغیر به‌طور همزمان، استفاده از رگرسیون جریمه‌شده است. در اینجا برای انتخاب متغیر و تشخیص ساختار مدل، از دو تابع جریمه به طور همزمان استفاده می‌شود، یعنی بردار ضرایب $\boldsymbol{\beta} = (\boldsymbol{\beta}_۱^T, \dots \boldsymbol{\beta}_d^T)^T$، که $\boldsymbol{\beta}_j = (\beta_{j۱}, \dots \beta_{jK})^T$، با حل مسئله بهینه‌سازی

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \Big\{ \frac{۱}{۲} \sum_{i=۱}^{n} \big(Y_i - \sum_{j=۱}^{d} \sum_{k=۱}^{K} \beta_{jk} B_{jk}(X_{ij})\big)^۲ + n \sum_{j=۱}^{d} p_{\lambda_۱}(\|\boldsymbol{\beta}_j\|_{\mathbf{A}_j}) + n \sum_{j=۱}^{d} p_{\lambda_۲}(\|\boldsymbol{\beta}_j\|_{\mathbf{D}_j}) \Big\} \qquad (۶.۳)$$

برآورد می‌شود که در آن $p_{\lambda_۱}(.)$ و $p_{\lambda_۲}(.)$ توابع جریمه SCAD با پارامترهای تنظیم کننده $\lambda_۱$ و $\lambda_۲$ به ترتیب برای شناسایی مؤلفه‌های صفر و خطی به کار می‌روند. علاوه براین، $\mathbf{A}_j$ و $\mathbf{D}_j$ دو ماتریس $K \times K$، $\|\boldsymbol{\beta}_j\|_{\mathbf{A}_j} = \big(\boldsymbol{\beta}_j^T \mathbf{A}_j \boldsymbol{\beta}_j\big)^{\frac{۱}{۲}}$ و $\|\boldsymbol{\beta}_j\|_{\mathbf{D}_j} = \big(\boldsymbol{\beta}_j^T \mathbf{D}_j \boldsymbol{\beta}_j\big)^{\frac{۱}{۲}}$ هستند. انتخاب ماتریس‌های $\mathbf{A}_j$ و $\mathbf{D}_j$ بسیار حائز اهمیت است. این ماتریس‌ها باید طوری انتخاب شوند که $\|\boldsymbol{\beta}_j\|_{\mathbf{A}_j} = ۰$، اگر وفقط اگر $\sum_k \beta_{jk} B_{jk}(x) \equiv ۰$ و به‌طور مشابه $\|\boldsymbol{\beta}_j\|_{\mathbf{D}_j} = ۰$، اگر و فقط اگر $\sum_k \beta_{jk} B_{jk}(x)$ یک تابع خطی باشد، یعنی دو تابع جریمه بتوانند برای شناسایی مؤلفه‌های صفر و خطی به کار روند. در اینجا $\mathbf{A}_j$ و $\mathbf{D}_j$ به‌صورت $\mathbf{A}_j = \{\int_۰^۱ B_{jk}(x) B_{jk'}(x) dx\}_{k,k'=۱}^{K}$ و $\mathbf{D}_j = \{\int_۰^۱ B''_{jk}(x) B''_{jk'}(x) dx\}_{k,k'=۱}^{K}$ انتخاب می‌شوند. انتخاب $\mathbf{D}_j$ به‌صورت فوق از این واقعیت نتیجه می شود که مشتق دوم یک تابع خطی برابر صفر است. فرض کنید $\mathbf{Y} = (Y_۱, ..., Y_n)^T$ و $\mathbf{Z} = (\mathbf{Z}_۱, ..., \mathbf{Z}_d)$ که در آن

$$\mathbf{Z}_j = \begin{pmatrix} B_{j۱}(X_{۱j}) & B_{j۲}(X_{۱j}) & \cdots & B_{jK}(X_{۱j}) \\ \vdots & \vdots & & \vdots \\ B_{j۱}(X_{nj}) & B_{j۲}(X_{nj}) & \cdots & B_{jK}(X_{nj}) \end{pmatrix}.$$

[10] De Boor

پس رابطه (۶.۳) را می‌توان به‌صورت ماتریسی

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{d} p_{\lambda_1}(\|\boldsymbol{\beta}_j\|_{\mathbf{A}_j}) + n\sum_{j=1}^{d} p_{\lambda_2}(\|\boldsymbol{\beta}_j\|_{\mathbf{D}_j}) \right\}. \qquad (۷.۳)$$

نوشت. به دلیل سهولت، تابع هدف سمت راست رابطه (۷.۳) با $Q(\boldsymbol{\beta})$ نشان داده می‌شود.

برای حل مسئله بهینه‌سازی (۷.۳) از تقریب موضعی درجه دو (LQA) معرفی شده در زیر بخش ۱.۷.۲ استفاده می‌کنیم. با استفاده از بسط تیلور و داشتن یک مقدار اولیه $\boldsymbol{\beta}_j^{\circ}$، توابع جریمه را می‌توان به‌صورت

$$p_{\lambda_1}(\|\boldsymbol{\beta}_j\|_{\mathbf{A}_j}) \approx p_{\lambda_1}(\|\boldsymbol{\beta}_j^{(\circ)}\|_{\mathbf{A}_j}) + \frac{1}{2}\frac{p'_{\lambda_1}(\|\boldsymbol{\beta}_j^{(\circ)}\|_{\mathbf{A}_j})}{\|\boldsymbol{\beta}_j^{(\circ)}\|_{\mathbf{A}_j}}\{\|\boldsymbol{\beta}_j\|^2_{\mathbf{A}_j} - \|\boldsymbol{\beta}_j^{(\circ)}\|^2_{\mathbf{A}_j}\}$$

و

$$p_{\lambda_2}(\|\boldsymbol{\beta}_j\|_{\mathbf{D}_j}) \approx p_{\lambda_2}(\|\boldsymbol{\beta}_j^{(\circ)}\|_{\mathbf{D}_j}) + \frac{1}{2}\frac{p'_{\lambda_2}(\|\boldsymbol{\beta}_j^{(\circ)}\|_{\mathbf{D}_j})}{\|\boldsymbol{\beta}_j^{(\circ)}\|_{\mathbf{D}_j}}\{\|\boldsymbol{\beta}_j\|^2_{\mathbf{D}_j} - \|\boldsymbol{\beta}_j^{(\circ)}\|^2_{\mathbf{D}_j}\}$$

تقریب زد. با به کار بردن تقریب‌های فوق و حذف جملات ثابت می‌توان رابطه (۷.۳) را به‌صورت

$$Q(\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\|^2 + \frac{1}{2}\boldsymbol{\beta}^T(\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2)\boldsymbol{\beta} \qquad (۸.۳)$$

خلاصه کرد، که در آن $\boldsymbol{\Omega}_1$ و $\boldsymbol{\Omega}_2$ دو ماتریس بلوکی با بعد $dK \times dK$ به‌صورت

$$\boldsymbol{\Omega}_1 = \operatorname{diag}\left(\frac{p'_{\lambda_1}(\|\boldsymbol{\beta}_1^{(\circ)}\|_{\mathbf{A}_1})}{\|\boldsymbol{\beta}_1^{(\circ)}\|_{\mathbf{A}_1}}\mathbf{A}_1, \ldots, \frac{p'_{\lambda_1}(\|\boldsymbol{\beta}_d^{(\circ)}\|_{\mathbf{A}_d})}{\|\boldsymbol{\beta}_d^{(\circ)}\|_{\mathbf{A}_d}}\mathbf{A}_d \right)$$

و

$$\boldsymbol{\Omega}_2 = \operatorname{diag}\left(\frac{p'_{\lambda_2}(\|\boldsymbol{\beta}_1^{(\circ)}\|_{\mathbf{D}_1})}{\|\boldsymbol{\beta}_1^{(\circ)}\|_{\mathbf{D}_1}}\mathbf{D}_1, \ldots, \frac{p'_{\lambda_2}(\|\boldsymbol{\beta}_d^{(\circ)}\|_{\mathbf{D}_d})}{\|\boldsymbol{\beta}_d^{(\circ)}\|_{\mathbf{D}_d}}\mathbf{D}_d \right)$$

هستند. توجه شود که رابطه (۸.۳) یک تابع درجه دو برحسب $\boldsymbol{\beta}$ بوده و در نتیجه دارای جواب صریح

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}^T\mathbf{Z} + n(\boldsymbol{\Omega}_1 + \boldsymbol{\Omega}_2)\right)^{-1}\mathbf{Z}^T\mathbf{Y} \qquad (۹.۳)$$

است. برآورد بدست آمده $\hat{\boldsymbol{\beta}}$، به‌عنوان برآورد اولیه در تکرار بعدی استفاده می‌شود. این الگوریتم تا رسیدن به همگرایی تکرار می‌شود. بر اساس پیشنهاد فن و لی (۲۰۰۱)، در هر تکرار، به محض اینکه برخی از مؤلفه‌های $\|\boldsymbol{\beta}_j\|_{\mathbf{A}_j}$ یا $\|\boldsymbol{\beta}_j\|_{\mathbf{D}_j}$ کوچکتر از 10^{-6} شود، مولفه $f_j(x)$ به ترتیب به‌عنوان تابع صفر یا تابع خطی تشخیص داده می‌شود.

همانطور که در بخش ۷.۲ اشاره شد، تقریب LQA یک الگوریتم پیش‌رونده است، یعنی اگر متغیری در یک تکرار از مدل حذف شود، در تکرار بعدی وارد مدل نمی‌شود و در نتیجه آن متغیر توضیحی از مدل نهایی حذف می‌شود. بنابراین، با بکاربردن تقریب LQA، ممکن است برخی متغیرهای توضیحی مهم را از دست بدهیم. برای رفع این مشکل، هانتر و لی[11] (۲۰۰۵) اعمال برخی تغییرات جزیی را در

[11] Hunter and Li

این الگوریتم پیشنهاد دادند. به‌طور دقیق‌تر، برای $\varepsilon_1 > 0$ و $\varepsilon_2 > 0$، رابطه (۹.۳) را، با اندکی تغییر، می‌توان بصورت زیر نوشت

$$\boldsymbol{\beta} = \left(\mathbf{Z}^T\mathbf{Z} + n(\boldsymbol{\Omega}_{\varepsilon_1} + \boldsymbol{\Omega}_{\varepsilon_2})\right)^{-1}\mathbf{Z}^T\mathbf{Y} \qquad (۱۰.۳)$$

که

$$\boldsymbol{\Omega}_{\varepsilon_1} = \text{diag}\left(\frac{p'_{\lambda_1}(\|\boldsymbol{\beta}_1^{(0)}\|_{\mathbf{A}_1})}{\|\boldsymbol{\beta}_1^{(0)}\|_{\mathbf{A}_1} + \varepsilon_1}\mathbf{A}_1, \ldots, \frac{p'_{\lambda_1}(\|\boldsymbol{\beta}_d^{(0)}\|_{\mathbf{A}_d})}{\|\boldsymbol{\beta}_d^{(0)}\|_{\mathbf{A}_d} + \varepsilon_1}\mathbf{A}_d\right)$$

و

$$\boldsymbol{\Omega}_{\varepsilon_2} = \text{diag}\left(\frac{p'_{\lambda_2}(\|\boldsymbol{\beta}_1^{(0)}\|_{\mathbf{D}_1})}{\|\boldsymbol{\beta}_1^{(0)}\|_{\mathbf{D}_1} + \varepsilon_2}\mathbf{D}_1, \ldots, \frac{p'_{\lambda_2}(\|\boldsymbol{\beta}_d^{(0)}\|_{\mathbf{D}_d})}{\|\boldsymbol{\beta}_d^{(0)}\|_{\mathbf{D}_d} + \varepsilon_2}\mathbf{D}_d\right).$$

بر اساس پیشنهاد هانتر و لی (۲۰۰۵)، در پیاده‌سازی این الگوریتم، ε_1 و ε_2 را به‌صورت زیر در نظر می‌گیریم

$$\varepsilon_1 = \frac{\tau}{2n\lambda_1}\min\{\|\boldsymbol{\beta}_j^{(0)}\|_{\mathbf{A}_j}, j = 1, \ldots, d\}, \quad \varepsilon_2 = \frac{\tau}{2n\lambda_2}\min\{\|\boldsymbol{\beta}_j^{(0)}\|_{\mathbf{D}_j}, j = 1, \ldots, d\}$$

که $\tau = 10^{-8}$. قرار دهید $k = 0$ و با استفاده از الگوریتم زیر جواب نهایی را بدست آورید:

- گام ۱: $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$ را به‌عنوان مقدار اولیه در نظر بگیرید.
- گام ۲: با استفاده از رابطه (۱۰.۳)، مقدار $\boldsymbol{\beta}$ را به‌روز رسانی کرده تا $\boldsymbol{\beta}^{(k+1)}$ حاصل شود.
- گام ۳: قرار دهید $k = k + 1$ و گام‌های ۱ و ۲ را تا رسیدن به همگرایی تکرار کنید.

در ابتدا، برآورد کمترین توان‌های دوم به‌عنوان برآورد اولیه در نظر گرفته می‌شود. در گام ۳ تکرار تا زمانی ادامه می‌یابد که

$$\max_{1\le j\le p}|\|\boldsymbol{\beta}_j^{(k)}\|_{\mathbf{A}_j} - \|\boldsymbol{\beta}_j^{(k+1)}\|_{\mathbf{A}_j}| + \max_{1\le j\le p}|\|\boldsymbol{\beta}_j^{(k)}\|_{\mathbf{D}_j} - \|\boldsymbol{\beta}_j^{(k+1)}\|_{\mathbf{D}_j}| < 10^{-6}.$$

پس از همگرایی الگوریتم فوق، برآورد ضرایب دقیقاً برابر با صفر نیست. لذا لازم است از یک مقدار آستانه استفاده شود. در مطالعات شبیه‌سازی و تحلیل داده‌های واقعی، اگر $\|\boldsymbol{\beta}_j^{(k)}\|_{\mathbf{A}_j} < 10^{-6}$، قرار دهید $\hat{\boldsymbol{\beta}}_j = 0$، و به‌طور مشابه، اگر $\|\boldsymbol{\beta}_j\|_{\mathbf{D}_j} < 10^{-6}$، مولفه به‌عنوان تابع خطی انتخاب می‌شود. نتایج شبیه‌سازی نشان می‌دهند که این انتخاب مناسب است.

۴.۳ مطالعه شبیه‌سازی

در این بخش، عملکرد روش دو مرحله‌ای مذکور با انجام شبیه‌سازی بررسی می‌شود. ابتدا، به‌عنوان توضیحی در خصوص مثال‌های عددی، در مثال ۱.۴.۳ حساسیت روش RDC-SIS نسبت به پارامتر d تحلیل می‌شود. سپس در مثال ۲.۴.۳، عملکرد RDC-SIS با DC-SIS (لی و همکاران، ۲۰۱۲)، SIRS

(ژو و همکاران، ۲۰۱۱)، SIS (فن و لیو، ۲۰۰۸) و NIS (فن و همکاران، ۲۰۱۱) مقایسه می‌کنیم. در مثال‌های ۳.۴.۳ و ۴.۴.۳، عملکرد روش تکراری RDC-ISIS را با روش‌های مذکور و همچنین روش تکراری ISIS (فن و لیو، ۲۰۰۸) مقایسه می‌شود. عملکرد این روش‌ها را با سه معیار S, P_j, M ارزیابی می‌کنیم که M حداقل اندازه مدل برای در برگرفتن تمام متغیرهای مهم، P_j احتمال تجربی انتخاب متغیر مهم X_j و S احتمال انتخاب تمام متغیرهای مهم برای اندازه مدل داده شده است. به منظور استنباط بهتر، در مثال ۲.۴.۳ چندک‌های ٪۵, ٪۲۵, ٪۵۰, ٪۷۵ و ٪۹۵ معیار M در ۵۰۰ تکرار نیز ارائه شده است. توجه شود که معیار M نیازی به مشخص کردن مقدار آستانه ندارد. اگر مقادیر M مربوط به یک روش نزدیک به تعداد متغیرهای مهم باشند، آن روش از عملکرد مطلوبی برخوردار است. همچنین در یک روش غربالگری مناسب مقادیر P_j و S باید نزدیک به یک باشند.

برای پیاده‌سازی RDC-ISIS، مطالعات تجربی نشان می‌دهند که تعداد کمی از تکرارها کافی است و می‌تواند هزینه محاسبات را کاهش دهد. با تکرار بیشتر این الگوریتم ممکن است احتمال حذف متغیرهای مهم کاهش یابد، اما هزینه محاسبات افزایش می‌یابد. در این رساله، برای شبیه‌سازی با انتخاب $d_1 = 5$ و $d_2 = d - 5$ الگوریتم فقط یکبار تکرار می‌شود.

در مثال ۵.۴.۳ کارایی روش دو مرحله‌ای در شناسایی مؤلفه‌های غیرصفر و همچنین تشخیص مؤلفه‌های خطی و غیرخطی در یک مدل جمعی خطی–جزئی با استفاده از دو تابع جریمه بررسی می‌شود. برای انتخاب بهینه پارامترهای λ_1 و λ_2 از معیار BIC به‌صورت

$$\log(\frac{1}{n}\|\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_\lambda\|^2) + d_1\frac{\log(n/K)}{n/K} + d_2\frac{\log n}{n}$$

استفاده می‌شود، که در آن $\hat{\boldsymbol{\beta}}_\lambda$ مقدار مینیمم کننده (۷.۳) برای $\lambda = (\lambda_1, \lambda_2)$ داده شده است. همچنین d_1 تعداد مؤلفه‌های ناپارامتری و d_2 تعداد مؤلفه‌های پارامتری به ازای مقدار داده شده λ است.

مثال ۱.۴.۳. برای تحلیل حساسیت عملکرد RDC-SIS نسبت به مقادیر مختلف d، داده‌ها از مدل

$$Y = 5g_1(X_1) + 3g_2(X_2) + 4g_3(X_3) + 6g_4(X_4) + \sqrt{1.74}\varepsilon$$

تولید شده‌اند، که در آن

$$g_1(x) = x \qquad g_2(x) = (2x-1)^2 \qquad g_3(x) = \sin(2\pi x)/(2-\sin(2\pi x))$$

$$g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$$

و متغیرهای توضیحی دارای توزیع حاشیه‌ای نرمال استاندارد و همبستگی $Cov(X_i, X_j) = 0.8^{|i-j|}$ هستند. سه مقدار مختلف $d_1 = [n/\log(n)]$, $d_2 = 2[n/\log(n)]$, $d_3 = 3[n/\log(n)]$ را برای d در نظر گرفته و شبیه‌سازی برای مقادیر مختلف (n, p) انجام شده است. همچنین برای خطا دو توزیع نرمال استاندارد و تی–استودنت با یک درجه آزادی در نظر گرفته شده است. نتایج شبیه‌سازی پس از ۵۰۰ بار تکرار در جدول ۱.۳ گزارش شده است. همان‌طور که ملاحظه می‌شود، به‌ازای هر (n, p)، با

جدول ۱.۳: احتمال تجربی P_j و احتمال S در مثال ۱.۴.۳

$\varepsilon \sim t(1)$					$\varepsilon \sim \mathcal{N}(0,1)$					d	n	p
S	P_4	P_3	P_2	P_1	S	P_4	P_3	P_2	P_1			
۰٫۴۱	۰٫۴۵	۰٫۹۲	۱٫۰۰	۰٫۸۴	۰٫۵۳	۰٫۶۰	۰٫۹۶	۱٫۰۰	۰٫۹۱	d_1		
۰٫۵۸	۰٫۶۲	۰٫۹۶	۱٫۰۰	۰٫۹۴	۰٫۷۳	۰٫۷۵	۰٫۹۸	۱٫۰۰	۰٫۹۸	d_2	۱۰۰	
۰٫۶۷	۰٫۷۰	۰٫۹۷	۱٫۰۰	۰٫۹۶	۰٫۸۱	۰٫۸۱	۰٫۹۹	۱٫۰۰	۱٫۰۰	d_3		
۰٫۹۳	۰٫۹۳	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۹	۰٫۹۹	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_1		
۰٫۹۷	۰٫۹۷	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_2	۲۰۰	۱۰۰۰
۰٫۹۸	۰٫۹۸	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_3		
۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_1		
۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_2	۴۰۰	
۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_3		
۰٫۴۲	۰٫۵۵	۰٫۹۰	۱٫۰۰	۰٫۷۷	۰٫۵۵	۰٫۶۱	۰٫۹۶	۱٫۰۰	۰٫۹۲	d_1		
۰٫۵۸	۰٫۶۶	۰٫۹۶	۱٫۰۰	۰٫۸۹	۰٫۷۱	۰٫۷۲	۰٫۹۹	۱٫۰۰	۰٫۹۶	d_2	۱۰۰	
۰٫۶۵	۰٫۷۰	۰٫۹۸	۱٫۰۰	۰٫۹۲	۰٫۷۸	۰٫۷۹	۰٫۹۹	۱٫۰۰	۰٫۹۹	d_3		
۰٫۹۲	۰٫۹۷	۱٫۰۰	۱٫۰۰	۰٫۹۹	۰٫۹۷	۰٫۹۷	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_1		
۰/۹۷	۰/۹۷	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_2	۲۰۰	۲۰۰۰
۰/۹۸	۰/۹۸	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_3		
۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_1		
۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_2	۴۰۰	
۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	d_3		

افزایش مقدار d عملکرد RDC-SIS بهبود می‌یابد. برای $n = ۱۰۰$، عملکرد RDC-SIS به ازای d_1 و d_2 چندان مطلوب نیست و امکان حذف شدن متغیر مهم X_4 وجود دارد. لذا برای حجم نمونه کوچک مقدار d_3 پیشنهاد می‌شود. اما برای $n = ۲۰۰, ۴۰۰$ و هر دو نوع توزیع خطا، هر یک از مقادیر d_1، d_2 و d_3 را می‌توان استفاده کرد.

در مثال‌های ۲.۴.۳ تا ۴.۴.۳، حجم نمونه $n = ۲۰۰$، تعداد متغیرهای توضیحی را $p = ۱۰۰۰$ و $d = ۲\lfloor n/\log(n) \rfloor$ در نظر گرفته و شبیه‌سازی ۵۰۰ بار تکرار شده است.

مثال ۲.۴.۳. دو مدل به‌صورت

مدل ۱: $Y = c\boldsymbol{\beta}^T\mathbf{X} + \sigma\varepsilon$ مدل۲: $Y = X_1 + ۲X_2^2 + ۳X_3^3 + ۴X_4^4 + \varepsilon$

را در نظر بگیرید، که در آن $\boldsymbol{\beta} = (۱, ۰٫۸, ۰٫۶, ۰٫۴, ۰٫۲, ۰, \ldots, ۰)^T$ و $\sigma^2 = ۶٫۸۳$. در مدل ۱، به منظور کنترل نسبت سیگنال به نوفه[12]، مقادیر $c = ۱, ۲$ در نظر گرفته شده است. در هر دو مدل، بردار

[12] Signal-to-noise ratio

جدول ۲.۳: چندک‌های M، احتمال تجربی P_j و احتمال S در مدل ۱

خطا	c	روش	M					P_j					S
			٪۵	٪۲۵	٪۵۰	٪۷۵	٪۹۵	$P_۱$	$P_۲$	$P_۳$	$P_۴$	$P_۵$	
$\mathcal{N}(۰,۱)$	۱	RDC-SIS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
		DC-SIS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
		SIS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
		SIRS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
	۲	RDC-SIS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
		DC-SIS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
		SIS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
		SIRS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
$t(۳)$	۱	RDC-SIS	۵	۵	۵	۵	۱۹	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۹	۰٫۹۹	۰٫۹۹
		DC-SIS	۵	۷	۱۹	۱۰۹	۵۸۴	۰٫۸۲	۰٫۸۴	۰٫۸۶	۰٫۷۷	۰٫۷۱	۰٫۶۷
		SIS	۳۵	۴۶۷	۸۰۶	۹۱۶	۹۶۶	۰٫۲۰	۰٫۲۱	۰٫۲۱	۰٫۲۰	۰٫۱۶	۰٫۱۰
		SIRS	۵	۵	۵	۶	۲۵	۱٫۰۰	۱٫۰۰	۰٫۹۹	۰٫۹۹	۰٫۹۸	۰٫۹۸
	۲	RDC-SIS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
		DC-SIS	۵	۵	۵	۶	۱۵۶	۰٫۹۵	۰٫۹۴	۰٫۹۴	۰٫۹۴	۰٫۹۰	۰٫۸۷
		SIS	۵	۹۰	۴۹۴	۸۶۵	۹۸۱	۰٫۴۴	۰٫۴۳	۰٫۴۴	۰٫۴۲	۰٫۳۰	۰٫۲۳
		SIRS	۵	۵	۵	۵	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰

متغیرهای توضیحی از توزیع نرمال چند متغیره با میانگین صفر و ماتریس کوواریانس $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ تولید می‌شود که $\sigma_{ij} = ۰٫۵^{|i-j|}$. در مدل ۱، برای خطا دو توزیع نرمال استاندارد و توزیع تی‌ـاستودنت با سه درجه آزادی در نظر گرفته شده است. در مدل ۲، علاوه بر دو توزیع فوق، توزیع نرمال چوله با پارامترهای $\alpha = ۲, \sigma = ۱, \mu = ۰$ نیز در نظر گرفته شده و نتایج شبیه‌سازی پس از ۵۰۰ تکرار در جداول ۲.۳ و ۳.۳ ارائه شده‌اند.

با نگاه کلی به مقادیر جدول ۲.۳، می‌توان کارایی روش غربالگری پیشنهادی را دریافت. با توجه به جدول ۲.۳، هنگامی که خطا دارای توزیع نرمال استاندارد است، برای هر دو حالت $c = ۱$ و $c = ۲$ هر چهار روش در شناسایی متغیرهای مهم $X_۵, X_۴, X_۳, X_۲, X_۱$ بسیار خوب عمل می‌کنند و همواره هر ۵ متغیر مهم را به درستی شناسایی می‌کنند، زیرا برای همه روش‌ها، مقادیر P_j $(۱ \le j \le ۵)$ و S برابر با یک است. اما برای خطای غیر نرمال نتایج متفاوت است. برای توزیع خطای تی‌ـاستودنت، عملکرد روش‌های DC-SIS و SIS بسیار ضعیف است، در حالی‌که روش RDC-SIS با احتمال تجربی تقریباً ۱۰۰٪ متغیرهای مهم را به درستی تشخیص می‌دهد. به‌عنوان مثال، برای $c = ۱$، مقدار $P_۱ = ۱$ بدین معنی است که RDC-SIS در ۱۰۰٪ دفعات، متغیر $X_۱$ را به‌عنوان متغیر مهم انتخاب می‌کند و همچنین مقدار $S = ۹۹٪$ بدین معنی است که روش مذکور در ۹۹٪ اوقات ۵ متغیر مهم را به درستی تشخیص می‌دهد. روش SIRS نیز در شناسایی متغیرهای مهم خوب عمل می‌کند و عملکرد دو روش

جدول ۳.۳: چندک‌های M، احتمال تجربی P_j و احتمال S در مدل ۲

خطا	روش	M					P_j				S
		٪۵	٪۲۵	٪۵۰	٪۷۵	٪۹۵	P_1	P_2	P_3	P_4	
$\mathcal{N}(0,1)$	RDC-SIS	۴	۴	۴	۴	۴	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
	DC-SIS	۴	۵	۵	۶	۹	۰٫۹۹	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۹
	NIS	۵	۵	۶	۱۱	۱۰۲	۰٫۹۳	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۳
	SIRS	۴	۴	۵	۱۹	۲۹۹	۱٫۰۰	۱٫۰۰	۰٫۸۵	۰٫۸۵	۰٫۸۵
t(۳)	RDC-SIS	۴	۴	۴	۴	۵	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
	DC-SIS	۴	۵	۵	۷	۶۴	۰٫۹۴	۰٫۹۶	۰٫۹۸	۰٫۹۸	۰٫۹۴
	NIS	۵	۶	۱۱	۸۲	۸۶۴	۰٫۶۵	۰٫۸۱	۰٫۸۵	۰٫۸۹	۰٫۵۵
	SIRS	۴	۴	۷	۳۴	۴۶۸	۰٫۹۹	۱٫۰۰	۱٫۰۰	۰٫۷۸	۰٫۶۹
SN	RDC-SIS	۴	۴	۴	۴	۴	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
	DC-SIS	۴	۵	۵	۶	۸	۰٫۹۹	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۹
	NIS	۵	۵	۶	۱۱	۱۰۴	۰٫۹۲	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۲
	SIRS	۴	۴	۵	۲۳	۳۴۵	۱٫۰۰	۱٫۰۰	۰٫۸۷	۰٫۸۷	۰٫۸۷

RDC-SIS و SIRS تقریبا یکسان و بسیار رضایت‌بخش است. با توجه به مقادیر P_j و S در جدول ۲.۳، برای خطای تی–استودنت، روش SIS شانس بسیار کمی برای انتخاب متغیرهای مهم دارد. این روش در حالت $c = ۱$ با احتمال تجربی ۱۰٪ و در حالت $c = ۲$ با احتمال ۲۳٪ همه متغیرهای مهم را انتخاب می‌کند.

جدول ۳.۳ که حاوی نتایج مدل ۲ است، نشان می‌دهد که برای هر سه نوع توزیع خطا، عملکرد روش RDC-SIS عالی است و نسبت به روش‌های دیگر بهتر عمل می‌کند. برای $۱ \le j \le ۴$، مقادیر P_j و S مربوط به روش RDC-SIS برابر یک است. بنابراین، برای هر سه نوع توزیع خطا، روش RDC-SIS هر چهار متغیر مهم را برای ورود به مدل انتخاب می کند. همچنین در این مدل، عملکرد DC-SIS بسیار مشابه روش RDC-SIS است. این دو روش نسبت به NIS و SIRS برتری دارند.

همانطور که اشاره شد، در بسیاری از مدل‌ها، روش‌های غربالگری حاشیه‌ای در شناسایی متغیرهای مهم با شکست مواجه می‌شوند. در مثال ۳.۴.۳ متغیر مهم $X_۴$ به‌طور حاشیه‌ای با متغیر پاسخ ناهمبسته است و لذا توسط روش‌های غربالگری حاشیه‌ای انتخاب نمی‌شود. بنابراین از روش تکراری RDC-ISIS برای بازگرداندن متغیرهای مهم به مدل استفاده می‌شود.

مثال ۳.۴.۳. در این مثال، عملکرد روش RDC-ISIS را با روش‌های SIS، ISIS، SIRS، DC-SIS و RDC-SIS در مدل خطی (۵.۳)، یعنی

$$\text{مدل ۱:} \qquad Y = ۵X_۱ + ۵X_۲ + ۵X_۳ - ۱۵\sqrt{\rho}X_۴ + \varepsilon$$

مقایسه می‌شود. علاوه براین، یک مدل خطی دیگر، با نسبت سیگنال به نوفه ضعیف‌تر، به‌صورت

$$\text{مدل ۲:} \qquad Y = ۲٫۵X_۱ + ۲٫۵X_۲ + ۲٫۵X_۳ - ۷٫۵\sqrt{\rho}X_۴ + \varepsilon$$

جدول ۴.۳: احتمال تجربی P_j و احتمال S در مثال ۳

ρ	روش	مدل ۱					مدل ۲				
		P_1	P_2	P_3	P_4	S	P_1	P_2	P_3	P_4	S
۰٫۲	SIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۱	۰٫۰۱
	ISIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۵	۰٫۹۵
	SIRS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۳	۰٫۰۳	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۳	۰٫۰۳
	DC-SIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۲	۰٫۰۲
	RDC-SIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۳	۰٫۰۳	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۴	۰٫۰۴
	RDC-ISIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
۰٫۵	SIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۱	۰٫۰۱	۱٫۰۰	۰٫۹۸	۰٫۹۹	۰٫۰۰	۰٫۰۰
	ISIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۴	۰٫۹۴
	SIRS	۱٫۰۰	۰٫۹۹	۱٫۰۰	۰٫۰۰	۰٫۰۰	۱٫۰۰	۰٫۹۹	۱٫۰۰	۰٫۰۱	۰٫۰۱
	DC-SIS	۱٫۰۰	۰٫۹۸	۰٫۹۹	۰٫۰۳	۰٫۰۳	۱٫۰۰	۰٫۹۸	۰٫۹۹	۰٫۰۱	۰٫۰۱
	RDC-SIS	۱٫۰۰	۰٫۹۹	۱٫۰۰	۰٫۰۲	۰٫۰۲	۱٫۰۰	۰٫۹۹	۱٫۰۰	۰٫۰۱	۰٫۰۱
	RDC-ISIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰
۰٫۸	SIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۸۵	۰٫۸۷	۰٫۸۶	۰٫۰۰	۰٫۰۰
	ISIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۶۰	۰٫۶۰
	SIRS	۰٫۹۵	۰٫۹۵	۰٫۹۷	۰٫۰۰	۰/۰۰	۰٫۹۳	۰٫۹۴	۰٫۹۳	۰٫۰۰	۰٫۰۰
	DC-SIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۸۴	۰٫۸۳	۰٫۸۴	۰٫۰۰	۰٫۰۰
	RDC-SIS	۰٫۹۵	۰٫۹۴	۰٫۹۵	۰٫۰۰	۰٫۰۰	۰٫۹۳	۰٫۹۵	۰٫۹۴	۰٫۰۰	۰٫۰۰
	RDC-ISIS	۱٫۰۰	۱٫۰۰	۱٫۰۰	۰٫۹۵	۰٫۹۵	۰٫۹۹	۱٫۰۰	۱٫۰۰	۰٫۹۳	۰٫۹۱

در نظر گرفته شده است. در این مثال، با فرض $\varepsilon \sim \mathcal{N}(۰, ۱)$ و $\rho = ۰٫۲, ۰٫۵, ۰٫۸$، شبیه‌سازی ۵۰۰ بار تکرار شده است. در این مدل‌ها، متغیر $X_۴$ به‌طور توأم مهم، اما به‌طور حاشیه‌ای با Y ناهمبسته است. بنابراین روش‌های غربالگری حاشیه‌ای SIS، SIRS، DC-SIS و RDC-SIS به سختی می‌توانند متغیر مهم $X_۴$ را شناسایی کنند. اما همان‌طور که در جدول ۴.۳ ملاحظه می‌شود روش RDC-ISIS در هر دو مدل خطی و به ازای مقادیر مختلف ρ در انتخاب متغیر $X_۴$ بسیار تواناست. ما همچنین RDC-ISIS را با ISIS، نسخه تکراری SIS، مقایسه کرده‌ایم. نتایج نشان می‌دهد که در برخی موارد روش RDC-ISIS در انتخاب $X_۴$ حتی از ISIS نیز بهتر عمل می‌کند. به‌عنوان مثال، در مدل ۲ به ازای $\rho = ۰٫۸$ روش RDC-ISIS با احتمال تجربی ۹۱٪ همه متغیرهای مهم را به درستی انتخاب می‌کند، در حالیکه ISIS تنها با احتمال ۶۰٪ عمل انتخاب متغیر را به درستی انجام می دهد. به خاطر داشته باشید که RDC-ISIS از اطلاعات ساختاری مدل رگرسیونی استفاده نمی‌کند، در حالیکه ISIS برای غربالگری در مدل‌های خطی معرفی شده و از اطلاعات درست مدل خطی استفاده می‌کند. روش RDC-ISIS یک روش غربالگری آزاد مدل است و می‌تواند برای مدل‌های رگرسیونی مختلف جهت تشخیص روابط خطی و غیر خطی به کار رود.

در مثال زیر برخی از متغیرهای بی‌اهمیت، به دلیل همبستگی بالا با متغیر مهم X_1، نسبت به برخی از متغیرهای مهم دیگر در فرآیند انتخاب برتری دارند. بنابراین لازم است از روش‌های تکراری برای انتخاب متغیرهای مهم استفاده شود. در این مثال عملکرد روش RDC-ISIS با روش‌های دیگر مقایسه می‌شود.

مثال ۴.۴.۳. در این مثال، مدل رگرسیونی جمعی ناپارامتری به‌صورت

$$Y = ۲f_۱(X_۱) + \sqrt{۶}f_۲(X_{۱۰۱}) + ۳f_۳(X_{۲۰۱}) - ۰/۶f_۴(X_{۲۰۲}) + \varepsilon$$

است، که در آن توابع پایه‌ای به‌صورت

$$f_۱(x) = \exp(۲x/۳) \qquad f_۳(x) = \frac{\sin(۳\pi x/۴ + ۳/۲)}{۲ - \sin(۳\pi x/۴ + ۳/۲)}$$

$$f_۲(x) = \begin{cases} x + ۴ & x < -۲ \\ |x| & |x| \le ۲ \\ ۴ - x & x > ۲ \end{cases} \qquad f_۴(x) = \log(x^۲)$$

تعریف شده‌اند. با توجه به مدل، متغیرهای $(X_۱, X_{۱۰۱}, X_{۲۰۱}, X_{۲۰۲})$ متغیرهای مؤثر بر پاسخ هستند. متغیرهای توضیحی از توزیع نرمال چند متغیره با میانگین صفر و همبستگی $Cov(X_i, X_j) = \rho^{|i-j|}$ تولید شده‌اند. سپس متغیر X_k با متغیر جدید $X_k = ۰/۸X_۱ + \zeta_k$ جایگزین می‌شود، که در آن $\zeta_k \sim \mathcal{N}(۰, ۱)$ و $k = ۲, \ldots, ۱۰۰$. بنابراین ۹۹ متغیر بی‌اهمیت $(X_۲, ..., X_{۱۰۰})$ دارای همبستگی بالایی با متغیر مهم $X_۱$ هستند. همچنین با در نظر گرفتن $\rho = ۰/۵, ۰/۸$ دو نوع ماتریس کوواریانس برای متغیرهای توضیحی تعریف شده است. به منظور بررسی نیرومندی هر روش، با در نظر گرفتن دو نوع توزیع خطا، نتایج شبیه‌سازی در جدول ۵.۳ گزارش شده‌اند.

همان‌طور که ملاحظه می‌شود روش‌های غربالگری مستقل حاشیه‌ای SIS، NIS، SIRS و DC-SIS در شناسایی متغیرهای مهم $X_{۱۰۱}$، $X_{۲۰۱}$ و $X_{۲۰۲}$ ناتوان هستند، چون متغیرهای بی‌اهمیت $(X_۲, \ldots, X_{۱۰۰})$ دارای همبستگی بالایی با متغیر مهم $X_۱$ هستند و در فرآیند انتخاب متغیرها توسط این سه روش نسبت به سه متغیر مهم $X_{۱۰۱}$، $X_{۲۰۱}$ و $X_{۲۰۲}$ اولویت دارند. بنابراین یک روش غربالگری تکراری مناسب برای حذف ۹۹ متغیر بی‌اهمیت و بازگرداندن سه متغیر مهم $X_{۱۰۱}$، $X_{۲۰۱}$ و $X_{۲۰۲}$ به مدل نیاز است. با توجه به جدول ۵.۳ ملاحظه می‌شود که روش ISIS نمی‌تواند روابط غیرخطی بین متغیرهای توضیحی و متغیر پاسخ را شناسایی کند، در حالی‌که روش RDC-ISIS تمام متغیرهای مهم را در مدل جمعی ناپارامتری به درستی انتخاب می‌کند. بنابراین روش RDC-ISIS به دلیل آزاد-مدل بودن و نیرومند بودن یک روش قابل قبول برای مدل‌های غیرخطی با بعد خیلی بالا است.

مثال‌های بیان‌شده کارایی روش RDC-ISIS را برای غربالگری در داده‌های با بعد بسیار بالا نشان می‌دهند. در مثال زیر، پس از کاهش بعد اولیه با RDC-ISIS، از دو تابع جریمه SCAD به‌طور همزمان برای برازش مدل جمعی خطی-جزئی استفاده می‌شود.

جدول ۵.۳: احتمال تجربی P_j و احتمال S در مثال ۴

ρ	روش	$\varepsilon \sim \mathcal{N}(۰,۱)$					$\varepsilon \sim t(۲)$				
		$P_۱$	$P_{۱۰۱}$	$P_{۲۰۱}$	$P_{۲۰۲}$	S	$P_۱$	$P_{۱۰۱}$	$P_{۲۰۱}$	$P_{۲۰۲}$	S
۰٫۵	SIS	۱٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۹۹	۰٫۰۱	۰٫۰۰	۰٫۰۱	۰٫۰۰
	ISIS	۰٫۹۸	۰٫۰۷	۰٫۰۶	۰٫۰۵	۰٫۰۰	۰٫۹۸	۰٫۰۶	۰٫۰۵	۰٫۰۵	۰٫۰۰
	NIS	۱٫۰۰	۰٫۸۲	۰٫۸۳	۰٫۷۴	۰٫۵۰	۰٫۹۸	۰٫۷۵	۰٫۷۱	۰٫۶۱	۰٫۳۲
	SIRS	۱٫۰۰	۰٫۰۱	۰٫۰۱	۰٫۰۰	۰٫۰۰	۱٫۰۰	۰٫۰۰	۰٫۰۱	۰٫۰۰	۰٫۰۰
	DC-SIS	۱٫۰۰	۰٫۲۵	۰٫۲۹	۰٫۲۴	۰٫۰۲	۱٫۰۰	۰٫۳۱	۰٫۳۴	۰٫۲۷	۰٫۰۳
	RDC-SIS	۱٫۰۰	۰٫۳۴	۰٫۴۰	۰٫۳۳	۰٫۱۱	۱٫۰۰	۰٫۳۰	۰٫۳۶	۰٫۳۵	۰٫۱۱
	RDC-ISIS	۱٫۰۰	۰٫۹۸	۰٫۹۹	۰٫۹۹	۰٫۹۶	۱٫۰۰	۰٫۹۴	۰٫۹۹	۰٫۸۸	۰٫۸۲
۰٫۸	SIS	۱٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۰۰	۰٫۹۹	۰٫۰۰	۰٫۰۱	۰٫۰۰	۰٫۰۰
	ISIS	۰٫۹۳	۰٫۰۷	۰٫۰۴	۰٫۰۴	۰٫۰۰	۰٫۹۷	۰٫۰۹	۰٫۰۵	۰٫۰۵	۰٫۰۰
	NIS	۱٫۰۰	۰٫۸۵	۰٫۹۹	۰٫۹۶	۰٫۸۱	۰٫۹۹	۰٫۷۲	۰٫۹۳	۰٫۸۸	۰٫۵۹
	SIRS	۱٫۰۰	۰٫۰۰	۰٫۰۱	۰٫۰۱	۰٫۰۰	۱٫۰۰	۰٫۰۰	۰٫۰۱	۰٫۰۰	۰٫۰۰
	DC-SIS	۱٫۰۰	۰٫۲۸	۰٫۷۸	۰٫۶۱	۰٫۱۳	۱٫۰۰	۰٫۳۰	۰٫۷۲	۰٫۶۲	۰٫۱۴
	RDC-SIS	۱٫۰۰	۰٫۳۵	۰٫۷۸	۰٫۷۲	۰٫۳۱	۱٫۰۰	۰٫۳۶	۰٫۷۹	۰٫۷۲	۰٫۳۲
	RDC-ISIS	۱٫۰۰	۱٫۰۰	۰٫۹۸	۰٫۹۳	۰٫۹۰	۰٫۹۸	۰٫۹۱	۰٫۹۸	۰٫۹۵	۰٫۸۳

مثال ۵.۴.۳. داده‌ها از مدل

$$Y = \sum_{j=۱}^{p} f_j(X_j) + \varepsilon$$

تولید شده‌اند، که در آن $f_۳(x) = ۲x, f_۲(x) = ۶x(۱-x), f_۱(x) = ۳\sin(۲\pi x)/(۲-\sin(۲\pi x))$, $f_۴(x) = x, f_۵(x) = -x$ و به ازای $j > ۵$، $f_j(x) = ۰$. بنابراین تعداد مؤلفه‌های ناپارامتری در این مدل ۲ و تعداد مؤلفه‌های پارامتری برابر ۳ است. متغیرهای توضیحی ابتدا همانند مثال ۲.۴.۳ تولید شده‌اند. سپس با استفاده از تابع توزیع تجمعی نرمال استاندارد به توزیع حاشیه‌ای یکنواخت روی بازه $[۰,۱]$ تبدیل شده‌اند. برای نشان دادن کارایی روش ارائه شده، برای خطا دو توزیع نرمال استاندارد و توزیع تی‌ـاستودنت با ۲ درجه آزادی که برای تولید خطای دم سنگین استفاده می‌شود، در نظر گرفته‌شده و شبیه‌سازی با $n = ۷۰, ۲۰۰, ۴۰۰$ و $p = ۱۰۰۰, ۲۰۰۰$ به تعداد ۲۰۰ بار تکرار شده است. نتایج انتخاب متغیر و تشخیص ساختار مدل در جدول ۶.۳ خلاصه شده‌اند، که در آن N نشان دهنده متوسط تعداد مولفه‌های غیرصفر انتخاب شده، NN متوسط تعداد مولفه‌های ناپارامتری انتخاب شده، NNT متوسط تعداد مولفه‌های ناپارامتری انتخاب شده که واقعاً ناپارامتری هستند، NL متوسط تعداد مولفه‌های خطی انتخاب شده و NLT متوسط تعداد مولفه‌های خطی است که درست انتخاب شده‌اند.

برای بررسی دقت برآورد، ریشه دوم میانگین توان‌های دوم خطا (RMSE) برای پنج مؤلفه اول در

جدول ۶.۳: نتایج تشخیص ساختار مدل در مثال ۵

NLT(۳)*	NL(۳)*	NNT(۲)*	NN(۲)*	N(۵)*	خطا	n	p
۲٫۶۹(۰٫۶۳)	۲٫۷۴(۰٫۸۶)	۲(۰)	۲٫۶۸(۱٫۰۴)	۵٫۴۲	$\mathcal{N}(۰,۱)$	۷۰	
۲٫۶۲(۰٫۸۵)	۳٫۶۵(۱٫۱۵)	۱٫۹۵(۰٫۱۹)	۲٫۹۶(۱٫۱۰)	۶٫۶۱	t(۲)		
۲٫۷۴(۰٫۳۲)	۲٫۹۳(۰٫۸۸)	۲(۰)	۲٫۳۱(۰٫۹۶)	۵٫۲۴	$\mathcal{N}(۰,۱)$	۲۰۰	۱۰۰۰
۲٫۷۱(۰٫۷۱)	۳٫۳۷(۱٫۰۳)	۱٫۹۷(۰٫۱۲)	۲٫۵۲(۰٫۸۸)	۵٫۸۹	t(۲)		
۲٫۹۶(۰٫۲۶)	۳٫۱۲(۰٫۵۹)	۲(۰)	۲٫۲۰(۰٫۶۴)	۵٫۳۲	$\mathcal{N}(۰,۱)$	۴۰۰	
۲٫۷۴(۰٫۷۵)	۳٫۲۸(۰٫۶۶)	۱٫۹۹(۰٫۰۸)	۲٫۴۹(۰٫۵۴)	۵٫۷۷	t(۲)		
۲٫۸۵(۱٫۰۵)	۳٫۲۳(۱٫۰۱)	۲(۰)	۲٫۷۱(۱٫۰۹)	۵٫۹۴	$\mathcal{N}(۰,۱)$	۷۰	
۲٫۷۲(۰٫۹۸)	۳٫۳۸(۱٫۴۸)	۱٫۹۰(۰٫۵۸)	۳٫۲۶(۱٫۳۱)	۶٫۶۴	t(۲)		
۲٫۸۴(۰٫۶۴)	۳٫۱۲(۰٫۹۴)	۲(۰)	۲٫۴۶(۰٫۸۴)	۵٫۵۸	$\mathcal{N}(۰,۱)$	۲۰۰	۲۰۰۰
۲٫۷۶(۰٫۸۲)	۳٫۲۴(۱٫۰۱)	۱٫۹۸(۰٫۱۷)	۲٫۷۷(۰٫۷۹)	۶٫۰۱	t(۲)		
۲٫۹۲(۰٫۶۴)	۳٫۰۴(۰٫۹۴)	۲(۰)	۲٫۳۳(۰٫۸۱)	۵٫۳۷	$\mathcal{N}(۰,۱)$	۴۰۰	
۲٫۸۱(۰٫۶۳)	۲٫۹۷(۰٫۸۵)	۱٫۹۹(۰٫۰۶)	۲٫۵۴(۰٫۷۴)	۵٫۵۱	t(۲)		

*اعداد داخل پرانتز، مقدار درست آن معیار در مدل واقعی است.

جدول ۷.۳ گزارش شده است و برای مؤلفه jام بصورت

$$\text{RMSE}(f_j) = \sqrt{\frac{۱}{۵۰۰}\sum_{i=۱}^{۵۰۰}(\hat{f}_j(t_i) - f_j(t_i))^۲}, \qquad j = ۱,..,۵$$

محاسبه می‌شود که $(t_۱,...,t_{۵۰۰})$ نقاط شبکه‌ای، شامل ۵۰۰ نقطه با فواصل یکسان، روی بازه $[۰,۱]$ هستند و تابع $f_j(.)$ در این نقاط ارزیابی می‌شود. همچنین RMSE تابع رگرسیون $f = \sum_{j=۱}^{p} f_j$ نیز ارائه شده است. اعداد داخل پرانتز در جدول ۷.۳ خطای استاندارد متناظر هستند.

نتایج شبیه‌سازی نشان می‌دهد که روش دو مرحله‌ای ارائه شده در برآورد و تشخیص مؤلفه‌های غیرصفر و همچنین شناسایی مؤلفه‌های خطی و غیرخطی تواناست. برای خطای نرمال، این روش در هر ۲۰۰ تکرار هر دو مؤلفه ناپارامتری مدل را به درستی به‌عنوان مؤلفه غیرصفر ناپارامتری تشخیص می‌دهد، اما برای خطای غیرنرمال در برخی از تکرارها بعضی از مؤلفه‌های ناپارامتری به‌عنوان مؤلفه صفر یا خطی تشخیص داده می‌شوند. همچنین در برخی از تکرارها، مؤلفه‌های خطی به‌عنوان مولفه‌های صفر یا غیرخطی انتخاب می‌شوند. با توجه به جدول، هنگامی که $n = ۷۰$ و خطا دارای توزیع تی–استودنت است، این روش عملکرد چندان مطلوبی ندارد. در این حالت، مدل انتخاب شده معمولاً بزرگ‌تر از مدل واقعی است. به ویژه برای $p = ۲۰۰۰$ این مشکل بسیار واضح است. اما با افزایش حجم نمونه عملکرد بهبود می‌یابد. علاوه براین، دقت برآورد روش ارائه‌شده، تحت تاثیر توزیع خطاها قرار می‌گیرد. با توجه به جدول ۷.۳، در بیشتر حالت‌ها، مقدار RMSE متناظر با خطای نرمال کوچکتر از RMSE متناظر با خطای تی–استودنت است. این بدین دلیل است که در مرحله دوم برای انتخاب متغیر و

جدول ۷.۳: مجذور میانگین توان‌های دوم خطا برای پنج مؤلفه اول و تابع رگرسیون.

n	خطا	f_1	f_2	f_3	f_4	f_5	f
				$p = ۱۰۰۰$			
۷۰	$\mathcal{N}(۰, ۱)$	۰٫۲۴۷۸(۰٫۱۱۸)	۰٫۲۱۵۴(۰٫۱۳۵)	۰٫۱۲۴۴(۰٫۱۴۲)	۰٫۱۴۵۶(۰٫۱۰۲)	۰٫۱۳۵۴(۰٫۱۰۳)	۰٫۳۷۷۳(۰٫۱۸۶)
	t(۲)	۰٫۳۴۰۱(۰٫۲۶۹)	۰٫۲۵۶۵(۰٫۱۵۶)	۰٫۱۸۶۴(۰٫۲۰۰۸)	۰٫۱۵۹۷(۰٫۱۱۳)	۰٫۱۵۸۰(۰٫۱۱۲)	۰٫۴۷۲۰(۰٫۳۲۰)
۲۰۰	$\mathcal{N}(۰, ۱)$	۰٫۲۱۴۲(۰٫۰۲۵)	۰٫۲۰۳۴(۰٫۰۸۱)	۰٫۰۴۷۰(۰٫۰۲۷)	۰٫۱۴۳۷(۰٫۰۴۵)	۰٫۱۴۲۴(۰٫۰۴۶)	۰٫۳۸۶۷(۰٫۰۷۴)
	t(۲)	۰٫۲۳۳۰(۰٫۰۳۰)	۰٫۲۸۰۰(۰٫۰۶۴)	۰٫۰۵۱۰(۰٫۰۲۸)	۰٫۱۶۲۸(۰٫۰۶۱)	۰٫۱۵۵۹(۰٫۰۶۴)	۰٫۴۰۲۴(۰٫۰۶۵)
۴۰۰	$\mathcal{N}(۰, ۱)$	۰٫۱۴۶۴(۰٫۰۳۲)	۰٫۱۶۷۱(۰٫۰۸۰)	۰٫۰۳۱۲(۰٫۰۲۶)	۰٫۰۶۱۲(۰٫۰۵۵)	۰٫۰۶۱۹(۰٫۰۵۳)	۰٫۲۴۶۱(۰٫۰۷۹)
	t(۲)	۰٫۱۸۰۴(۰٫۰۶۴)	۰٫۱۷۵۰(۰٫۰۶۷)	۰٫۰۳۹۹(۰٫۰۵۷)	۰٫۱۰۰۱(۰٫۰۷۵)	۰٫۰۸۱۰(۰٫۰۷۱)	۰٫۲۷۴۸(۰٫۰۹۱)
				$p = ۲۰۰۰$			
۷۰	$\mathcal{N}(۰, ۱)$	۰٫۲۷۱۸(۰٫۲۱۵)	۰٫۲۲۵۰(۰٫۱۵۳)	۰٫۰۹۴۹(۰٫۱۰۷)	۰٫۱۴۴۵(۰٫۱۰۸)	۰٫۱۳۲۳(۰٫۱۰۵)	۰٫۳۷۷۵(۰٫۲۵۶)
	t(۲)	۰٫۳۰۵۷(۰٫۲۴۶)	۰٫۲۱۸۷(۰٫۱۴۰)	۰٫۰۹۵۲(۰٫۱۳۴)	۰٫۱۵۰۶(۰٫۱۰۵)	۰٫۱۴۴۹(۰٫۱۰۸)	۰٫۴۲۷۴(۰٫۲۷۸)
۲۰۰	$\mathcal{N}(۰, ۱)$	۰٫۲۱۵۹(۰٫۰۲۶)	۰٫۲۸۲۸(۰٫۰۷۱)	۰٫۰۴۷۶(۰٫۰۲۴)	۰٫۱۳۵۹(۰٫۰۴۰)	۰٫۱۳۳۶(۰٫۰۳۷)	۰٫۳۷۶۰(۰٫۰۶۰)
	t(۲)	۰٫۲۴۵۳(۰٫۰۴۲)	۰٫۲۳۲۱(۰٫۰۸۳)	۰٫۰۶۳۴(۰٫۰۴۳)	۰٫۱۸۴۷(۰٫۰۶۵)	۰٫۱۴۸۹(۰٫۰۶۴)	۰٫۴۳۳۱(۰٫۰۸۴)
۴۰۰	$\mathcal{N}(۰, ۱)$	۰٫۱۵۲۸(۰٫۰۴۵)	۰٫۱۵۹۹(۰٫۰۶۵)	۰٫۰۲۸۳(۰٫۰۲۴)	۰٫۰۵۴۸(۰٫۰۴۳)	۰٫۰۶۵۴(۰٫۰۵۲)	۰٫۲۳۸۹(۰٫۰۶۷)
	t(۲)	۰٫۱۶۸۵(۰٫۰۴۹)	۰٫۱۸۴۹(۰٫۰۸۰)	۰٫۰۳۶۱(۰٫۰۲۶)	۰٫۰۶۹۸(۰٫۰۵۸)	۰٫۰۸۲۷(۰٫۰۵۵)	۰٫۲۷۶۳(۰٫۰۸۴)

تشخیص ساختار از تابع زیان توان دوم استفاده شده است. این تابع زیان نسبت به داده‌های پرت و توزیع دم‌سنگین خطاها بسیار حساس است.

۵.۳ تحلیل داده‌های مسمومیت کبد

در این بخش، با استفاده از روش دو مرحله‌ای پیشنهادی، به تحلیل یک مجموعه داده واقعی می پردازیم. این مجموعه داده در بسته ''mixOmics'' نرم افزار R قابل دسترسی است و نخستین بار توسط بشل[۱۳] و همکاران (۲۰۰۷) مورد تحلیل قرار گرفته است. این داده‌ها مربوط به بیماری مسمومیت کبد بوده و شامل سطوح بیان ۳۱۱۶ ژن و ۹ اندازه‌گیری بالینی برای ۶۴ موش است. بنابراین، ۳۱۱۶ متغیر توضیحی و ۹ متغیر پاسخ کمی وجود دارند. متغیرهای پاسخ به اختصار عبارتند از: BUN، نیتروژن اوره؛ TP، پروتئین کل؛ ALB، آلبومین؛ ALT، آلانین آمینوترانسفراز؛ SDH، سوربیتول دهیدروژناز؛ AST، آسپارتات آمینوترانسفراز؛ ALP، آلکالین فسفاتاز؛ TBA، کل اسیدهای صفراوی؛ و CHOL، کلسترول. در این بخش ALT به‌عنوان متغیر پاسخ است و هدف پیدا کردن مؤثرترین ژن‌ها در پیش‌بینی ALT و نوع تاثیر این ژن‌هاست.

جدول ۸.۳ نشان دهنده ژن‌هایی است که توسط روش‌های غربالگری مختلف در رتبه‌های اول تا چهارم قرار می‌گیرند. به‌عنوان مثال، RDC-SIS و SIRS رتبه اول را به ژن $A\_۴۲\_P۴۹۶۶۲۲$ اختصاص می‌دهند. در مقابل، DC-SIS ژن $A\_۴۳\_P۱۴۱۳۱$ و NIS ژن $A\_۴۲\_P۸۲۵۲۹۰$ را به‌عنوان مهمترین ژن انتخاب می‌کنند.

[13] Bushel

جدول ۸.۳: چهار ژن بسیار مهم انتخاب شده توسط روش‌های مختلف

روش	ژن‌های انتخاب شده			
RDC-SIS	$A\_۴۲\_P۴۹۶۶۲۲$	$A\_۴۳\_P۱۴۱۶۳$	$A\_۴۳\_P۲۰۴۳۸$	$A\_۴۲\_P۴۶۹۵۵۱$
SIRS	$A\_۴۲\_P۴۹۶۶۲۲$	$A\_۴۳\_P۲۰۴۳۸$	$A\_۴۳\_P۱۲۷۲۴$	$A\_۴۲\_P۴۶۹۵۵۱$
DC-SIS	$A\_۴۳\_P۱۴۱۳۱$	$A\_۴۲\_P۶۲۰۹۱۵$	$A\_۴۲\_P۸۴۰۷۷۶$	$A\_۴۳\_P۱۱۷۲۴$
NIS	$A\_۴۲\_P۸۲۵۲۹۰$	$A\_۴۳\_P۱۰۶۰۶$	$A\_۴۳\_P۱۱۷۲۴$	$A\_۴۳\_P۱۴۱۳۱$

جدول ۹.۳: ژن‌های مهم انتخاب شده و نوع تأثیر آن‌ها

نوع تأثیر	ژن			
غیرخطی	$A\_۴۳\_P۱۲۷۲۴$	$A\_۴۲\_P۶۹۴۱۰۵$	$A\_۴۳\_P۱۴۸۶۴$	$A\_۴۳\_P۱۴۱۶۳$
	$A\_۴۲\_P۴۶۹۵۵۱$	$A\_۴۳\_P۲۰۴۳۸$		
خطی	$A\_۴۲\_P۶۱۹۲۸۸$	$A\_۴۳\_P۲۰۹۶۲$	$A\_۴۲\_P۴۹۶۶۲۲$	$A\_۴۲\_P۶۷۷۶۲۸$

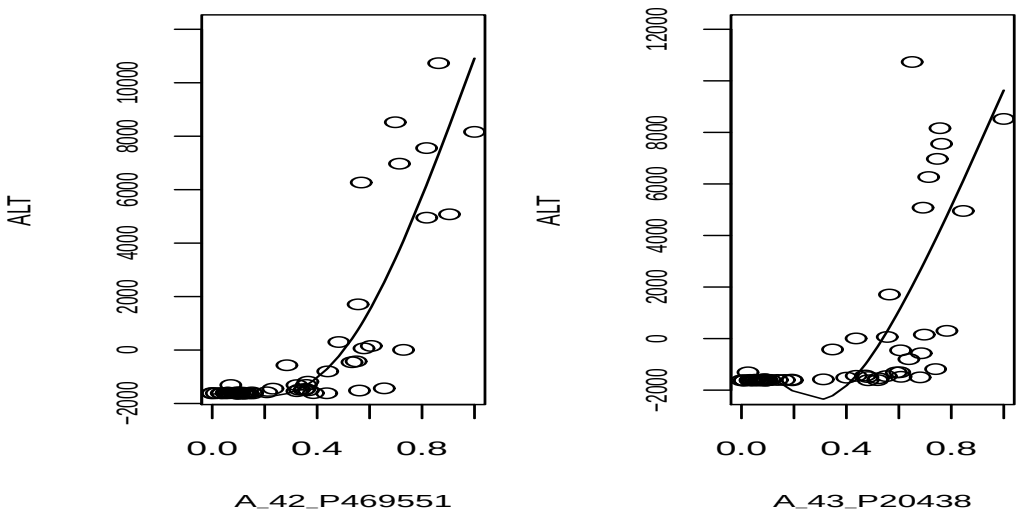
برای به‌کارگیری روش دو مرحله‌ای پیشنهادی در تحلیل داده‌های مسمومیت کبد، ابتدا نسخه تکراری RDC-SIS، یعنی RDC-ISIS با $d = ۲[n/\log(n)] = ۲۸$ و $d_۱ = d/۲$ را به کار برده و بعد داده‌ها را به $d = ۲۸$ کاهش داده‌ایم. دلیل استفاده از RDC-ISIS این است که اگر متغیر مهمی توسط RDC-SIS شناسایی نشود، آن متغیر مهم مجدداً برای ورود به مدل کاندید می‌شود و اهمیت آن مورد بررسی قرار می‌گیرد. علاوه براین، نمودارهای پراکنش و منحنی‌های برازش اسپلاین مکعبی متغیرهای $A\_۴۲\_P۴۶۹۵۵۱$ و $A\_۴۳\_P۲۰۴۳۸$ را در مقابل متغیر پاسخ Ro1 در شکل ۱.۳ نشان داده‌ایم. این نمودار وجود یک الگوی غیرخطی بین متغیرهای توضیحی و متغیر پاسخ را نشان می دهد.

پس از کاهش بعد اولیه، از دو تابع جریمه SCAD برای انتخاب متغیر و تشخیص ساختار مدل استفاده شده است. بعد از به کارگیری این روش دو مرحله‌ای، ۶ ژن دارای اثر غیرخطی و ۴ ژن دارای اثر خطی تشخیص داده شدند. ژن‌های انتخاب شده و نوع تأثیر آن‌هادر جدول ۹.۳ گزارش شده است.

برای ارزیابی روش دو مرحله‌ای ارائه شده، از اعتبارسنجی متقابل استفاده می‌کنیم. ابتدا یک مشاهده را به‌عنوان داده آزمون و سایر مشاهدات را به‌عنوان داده‌های آموزشی در نظر بگیرید. پس از برازش مدل با استفاده از داده‌های آموزشی، مقدار متغیر پاسخ به ازای داده آزمون پیش بینی می‌شود. این روند برای تمام مشاهدات تکرار شده و نمودار مقادیر مشاهده شده متغیر پاسخ در مقابل مقادیر پیش بینی شده در نمودار ۲.۳ رسم شده است. با توجه به نمودار، روش معرفی شده در پیش بینی مقادیر ALT عملکرد خوبی دارد.

بحث و نتیجه‌گیری

در این فصل، یک روش دو مرحله‌ای برای انتخاب متغیر و تشخیص ساختار در مدل‌های جمعی با بعد بسیار بالا ارائه شد. در مرحله اول، از یک روش غربالگری مستقل مطمئن برای کاهش بعد مدل استفاده



شکل ۱.۳: نمودارهای پراکنش متغیرپاسخ در مقابل دو ژن مهم و منحنی‌های اسپلاین مکعبی.

شد که در آن متغیرهای توضیحی براساس میزان همبستگی فاصله‌ای آنها با تابع توزیع حاشیه‌ای متغیر پاسخ رتبه‌بندی می‌شوند. کارایی این روش غربالگری با مطالعه شبیه‌سازی و تحلیل یک مجموعه داده واقعی مورد ارزیابی قرار گرفت که نتایج حاکی از عملکرد مطلوب روش ارائه شده است. در داده‌های با بعد بسیار بالا، معمولاً برخی از متغیرهای توضیحی دارای اثر خطی و برخی دیگر دارای اثر غیر خطی هستند. بنابراین تشخیص بخش‌های پارامتری و ناپارامتری بسیار حائز اهمیت است. لذا در مرحله دوم، برای انتخاب مدل نهایی و تشخیص مؤلفه‌های خطی و غیرخطی از دو تابع جریمه به‌طور همزمان استفاده شد. در اینجا نیز مطالعات شبیه‌سازی نشان دادند که این روش دو مرحله‌ای در برازش مدل جمعی خطی–جزیی کاراست.

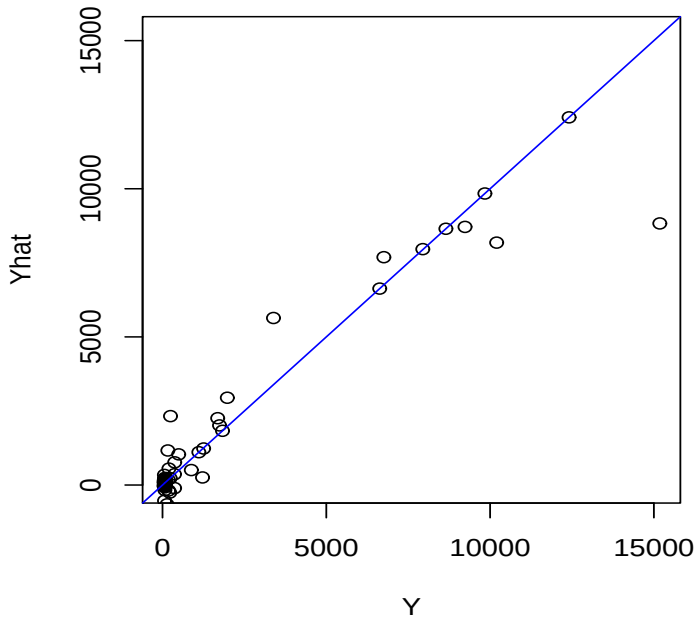
اگر چه در این فصل، مدل جمعی مد نظر بوده است، اما روش ارائه شده برای غربالگری و تشخیص ساختار را می‌توان به مدل‌های دیگر مانند مدل ضریب متغیر و مدل جمعی تعمیم یافته تعمیم داد. همچنین لازم به ذکر است که در این فصل، متغیرهای توضیحی و پاسخ صرفاً از نوع کمی در نظر گرفته شده‌اند. موضوع غربالگری در حالتی که این متغیرها از نوع کیفی و چند سطحی است، می‌تواند موضوعی برای تحقیقات آینده در نظر گرفته شود.

به‌عنوان روش دیگر برای غربالگری در مدل جمعی ناپارامتری، می‌توان از کوواریانس توابع ناپارامتری و تابع توزیع متغیر پاسخ استفاده کرد. بطور واضح‌تر، فرض کنید $m_j(X_j) = E(Y|X_j)$ تصویر Y بر X_j و $G(y) = P(Y \leq y)$ تابع توزیع غیر شرطی Y باشد. سپس از معیار مطلوبیت حاشیه‌ای

$$\omega_j = \big[Cov\big(m_j(X_j), G(Y)\big)\big]^2, \qquad 1 \leq j \leq p \qquad (۱۱.۳)$$

برای رتبه‌بندی متغیرهای توضیحی استفاده می‌کنیم. برای پیاده‌سازی این روش غربالگری کافی‌است توابع ناپارامتری $m_j(X_j)$ را با توابع B– اسپلاین و تابع توزیع $G(y)$ را با تابع توزیع تجربی برآورد کنیم.

در خصوص تقلیل بعد بالای فضای متغیر های توضیحی به بعد مرتبه d توسط روش غربالگری به‌کار



شکل ۲.۳: نمودار مقادیر مشاهده شده متغیر پاسخ در مقابل مقادیر پیش بینی شده

رفته در این فصل، می‌توان گفت که انتخاب d، همانند انتخاب پارامتر جریمه در روش‌های انقباضی، دارای اهمیت بسزائی است. تاکنون اغلب روش‌های معرفی شده برای تشخیص مقدار d وابسته به ساختار مدل هستند، بنابراین معرفی روشی جدید برای انتخاب d در روش‌های غربالگری آزاد–مدل بسیار حائز اهمیت است.

نوع تابع جریمه در تشخیص ساختار مدل، مسئله دیگری است که بایستی به آن توجه شود. اگر چه ما از تابع جریمه SCAD استفاده کرده‌ایم، اما می‌توان از ترکیب روش غربالگری با سایر توابع جریمه نظیر Lasso، MCP یا aLasso نیز استفاده کرد.

فصل ۴

انتخاب متغیر و تشخیص ساختار در مدل ضریب متغیر خطی‌ـ جزئی

در این فصل، با استفاده از رگرسیون نما، یک روش جدید برای انتخاب متغیر و تشخیص اثرات ثابت و متغیر در مدل ضریب متغیر ارائه می‌شود. این روش نسبت به خطاهای غیر نرمال و داده‌هات پرت در پاسخ نیرومند است. از این فصل، مقاله کاظمی و همکاران (۲۰۱۹b) استخراج شده است.

۱.۴ مقدمه

مدل نیمه‌پارامتری ضریب متغیر خطی‌ـ جزئی، تعمیمی از مدل ضریب متغیر و مدل خطی‌ـ جزئی است که در آن برخی از ضرایب رگرسیونی، ثابت و برخی دیگر تابعی از یک متغیر شاخص، مانند سن یا زمان، هستند. این مدل علاوه بر بهره‌گیری از انعطاف‌پذیری مدل‌های ناپارامتری و تفسیرپذیری مدل‌های پارامتری، اثرات متقابل بین متغیر شاخص و متغیرهای توضیحی را نیز به‌صورت ناپارامتری در نظر می‌گیرد. تاکنون بسیاری از محققین به مسئله برآورد در مدل‌های ضریب متغیر خطی‌ـ جزئی پرداخته‌اند. به عنوان مثال، می‌توان به احمد[1] و همکاران (۲۰۰۵)، فن و هوانگ (۲۰۰۵)، ژو و لیانگ[2]

[1] Ahmad
[2] Zhou and Liang

(۲۰۰۹)، وانگ و همکاران (۲۰۰۹)، سان و لین[۳] (۲۰۱۳) و شن[۴] و همکاران (۲۰۱۴) اشاره کرد.

در همه مطالعات مذکور، فرض بر این است که زیرمجموعه متغیرهایی که دارای اثر ثابت یا اثر متغیر بر پاسخ هستند، از پیش تعیین شده‌اند. این یک فرض زیربنایی است که در ساخت برآوردگرها و بررسی ویژگی‌های مجانبی آنها اهمیت بسزایی دارد. با این وجود، در عمل، تشخیص اینکه متغیری دارای اثر ثابت یا متغیر بر پاسخ است، بسیار حائز اهمیت است. علاوه براین، اغلب تعداد زیادی از متغیرهای توضیحی در مدل حضور دارند. با توجه به ضرورت حذف متغیرهای بی‌اهمیت از مدل، لازم است که در مدل‌های نیمه‌پارامتری ضریب متغیر، انتخاب متغیر و تشخیص ساختار به‌طور همزمان انجام شود. تاکنون تعداد معدودی از محققین به موضوع انتخاب متغیر در مدل ضریب متغیر خطی–جزئی پرداخته‌اند. از جمله این کارها می توان به وانگ و همکاران (۲۰۰۸)، وانگ و زیا (۲۰۰۹)، ژائو و زوئی[۵] (۲۰۰۹)، ژائو و زوئی (۲۰۱۰)، کی[۶] و همکاران (۲۰۱۱)، ژائو و همکاران (۲۰۱۳)، ژانگ و همکاران (۲۰۱۳) اشاره کرد. اما تاکنون انتخاب متغیر به‌طور همزمان با تشخیص ساختار مورد بررسی قرار نگرفته است. لذا ارائه یک روش جدید که بتواند انتخاب متغیرهای مهم و تشخیص ضرایب ثابت و متغیر را بطور همزمان انجام دهد، مورد توجه محققین است که در این فصل به آن می‌پردازیم. زیا و همکاران (۲۰۰۴)، لنگ (۲۰۰۹)، نه و کیلگوم[۷] (۲۰۱۲)، هو[۸] و زیا (۲۰۱۲) روش‌های مختلفی را برای تشخیص اثرات ثابت در این مدل معرفی کردند، اما این روش‌ها متغیرهای مهم را انتخاب نمی‌کنند. همچنین، تانگ[۹] و همکاران (۲۰۱۲) با استفاده از رگرسیون‌های چندکی و میانگین، یک روش واحد برای مدل‌های ضریب متغیر خطی–جزئی معرفی کردند که علاوه بر انتخاب متغیرهای مهم، اثرات ثابت یا متغیر را نیز شناسایی می‌کند، اما این روش مستلزم استفاده از یک الگوریتم دو مرحله‌ای تکراری است که در عمل چندان کارا نیست، زیرا علاوه بر هزینه محاسباتی زیاد، هیچ تضمینی برای همگرایی این الگوریتم به جواب درست وجود ندارد. همچنین روش معرفی شده بر اساس رگرسیون میانگین نسبت به مشاهدات پرت یا توزیع دم‌سنگین خطاها بسیار حساس است. علاوه‌براین رگرسیون چندکی نیز در صورت نرمال بودن توزیع خطاها و عدم وجود مشاهدات پرت، ممکن است از کارایی لازم برخوردار نباشد. بنابراین ارائه یک روش نیرومند برای انتخاب متغیر و تشخیص ساختار در مدل‌های ضریب متغیر بسیار حائز اهمیت است.

اخیراً، یائو[۱۰] و همکاران (۲۰۱۲) و یائو و لی (۲۰۱۴)، براساس نمای شرطی متغیر پاسخ Y، یک روش نیرومند به نام ”رگرسیون نما“[۱۱] را معرفی کردند. این روش در تحلیل داده‌های چوله کاربرد زیادی دارد. در این فصل، یک روش نیرومند برای انتخاب متغیر و تشخیص ساختار در مدل‌های ضریب متغیر خطی–جزئی براساس رگرسیون نما ارائه می‌شود. ویژگی‌های مجانبی برآوردگرها مورد بررسی قرار

[3] Sun and Lin
[4] Shen
[5] Zhao and Zue
[6] Kai
[7] Noh and Keilegom
[8] Hu
[9] Tang
[10] Yao
[11] Modal regression

گرفته و عملکرد این روش با مطالعات شبیه‌سازی و تحلیل داده‌های واقعی ارزیابی می‌شود.

۱.۱.۴ رگرسیون نما

رگرسیون نما، نمای متغیر پاسخ Y را، به شرط بردار متغیرهای توضیحی $\mathbf{X} = (X_1, ..., X_p)^T$، به‌صورت یک تابع خطی از $\mathbf{X}$ مدل‌سازی می‌کند. این مدل رگرسیونی، متفاوت از رگرسیون خطی معمولی است، که فرض می‌شود میانگین شرطی Y تابعی خطی از $\mathbf{X}$ است. رگرسیون نما، به جای میانگین شرطی، از ''محتمل‌ترین'' مقادیر شرطی برای سنجش مرکز توزیع متغیر پاسخ استفاده می‌کند. بنابراین، در مقایسه با رگرسیون میانگین، رگرسیون نما دارای چندین مزیت است (یائو و همکاران، ۲۰۱۴). برخی از مزیت‌ها عبارتند از:

۱. به دلیل نیرومندی نما، رگرسیون نما نسبت به مشاهدات پرت و توزیع دم‌سنگین خطا نیرومند است.

۲. هنگامی که چولگی توزیع خطا بسیار زیاد است، رگرسیون نما پیشگویی نقطه‌ای معنی‌دارتری از رگرسیون میانگین نتیجه می‌دهد.

۳. با توجه به تمرکز رگرسیون نما بر ناحیه با بزرگترین چگالی شرطی، انتظار می‌رود که یک فاصله کوتاه حول برآورد رگرسیون نما نسبت به یک فاصله با همان طول حول برآورد رگرسیون میانگین، دارای احتمال پوشش بزرگتری باشد.

بنابراین رگرسیون نما یک ابزار مفید برای تحلیل داده‌های چوله است. برای برآورد رگرسیون نما یک تابع هدف هسته مبنا به‌کار می‌رود. برای ماکزیمم کردن این تابع هدف از الگوریتم MEM استفاده می‌شود. فرض کنید $f(y|\mathbf{x})$ چگالی شرطی Y به شرط $\mathbf{X}$ باشد و نمای تابع $f(y|\mathbf{x})$، که به‌صورت $\text{Mode}(Y|\mathbf{X}) = \arg\max_y(f(y|\mathbf{x}))$ تعریف می‌شود، یکتاست. در رگرسیون نما، فرض می‌شود $\text{Mode}(Y|\mathbf{X})$ تابعی خطی از $\mathbf{X}$ است، یعنی

$$\text{Mode}(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta} \qquad (۱.۴)$$

که در آن $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ بردار ضرایب رگرسیونی است. فرض کنید $\varepsilon = Y - \mathbf{X}^T\boldsymbol{\beta}$ خطای تصادفی و تابع $g(\varepsilon|\mathbf{X})$ چگالی شرطی ε به شرط $\mathbf{X}$ باشد. اگر $g(\varepsilon|\mathbf{X})$ حول صفر متقارن باشد، بردار ضرایب $\boldsymbol{\beta}$ در (۱.۴) همان ضرایب رگرسیون خطی میانگین است. اما اگر $g(\varepsilon|\mathbf{X})$ چوله باشد، این دو بردار ضرایب متفاوت خواهند بود.
برای درک بهتر تفاوت بین تابع رگرسیون نما و تابع رگرسیون میانگین، مثال زیر را در نظر بگیرید.

مثال ۱.۱.۴. مدل زیر را در نظر بگیرید

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\varepsilon \qquad (۲.۴)$$

که ε دارای تابع چگالی چوله $h(.)$ با میانگین صفر و مد ۱ است. اکنون دو حالت زیر را در نظر بگیرید.

۱. اگر $m(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$ و $\sigma(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\alpha}$، آنگاه

$$E(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}, \qquad \text{Mode}(Y|\mathbf{X}) = \mathbf{X}^T(\boldsymbol{\alpha}+\boldsymbol{\beta}).$$

بنابراین، در هر دو رگرسیون میانگین و نما، Y بطور خطی به $\mathbf{X}$ وابسته است، اگرچه دو رگرسیون دارای ضرایب رگرسیونی متفاوت هستند.

۲. اگر $m(\mathbf{X}) = ۰$ و $\sigma(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\alpha}$، پس

$$E(Y|\mathbf{X}) = ۰, \qquad \text{Mode}(Y|\mathbf{X}) = \mathbf{x}^T\boldsymbol{\alpha}.$$

بنابراین، از منظر رگرسیون میانگین، Y وابسته به $\mathbf{X}$ نیست، اما براساس رگرسیون نما، Y دارای رابطه خطی با $\mathbf{x}$ است. از این مثال می‌توان نتیجه گرفت که روش‌های انتخاب متغیر براساس رگرسیون نما، در مقایسه با رگرسیون میانگین، ممکن است در شناسایی برخی متغیرهای توضیحی مهم مفیدتر باشند.

فرض کنید $\{(\mathbf{X}_i, Y_i), i = ۱, ..., n\}$ یک نمونه تصادفی از $(\mathbf{X}, Y)$ باشد. برای برازش رگرسیون نما، بردار ضرایب $\boldsymbol{\beta}$ در (۱.۴) با ماکزیمم کردن

$$Q_h(\boldsymbol{\beta}) = \sum_{i=۱}^{n} \phi_h\big(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}\big) \tag{۳.۴}$$

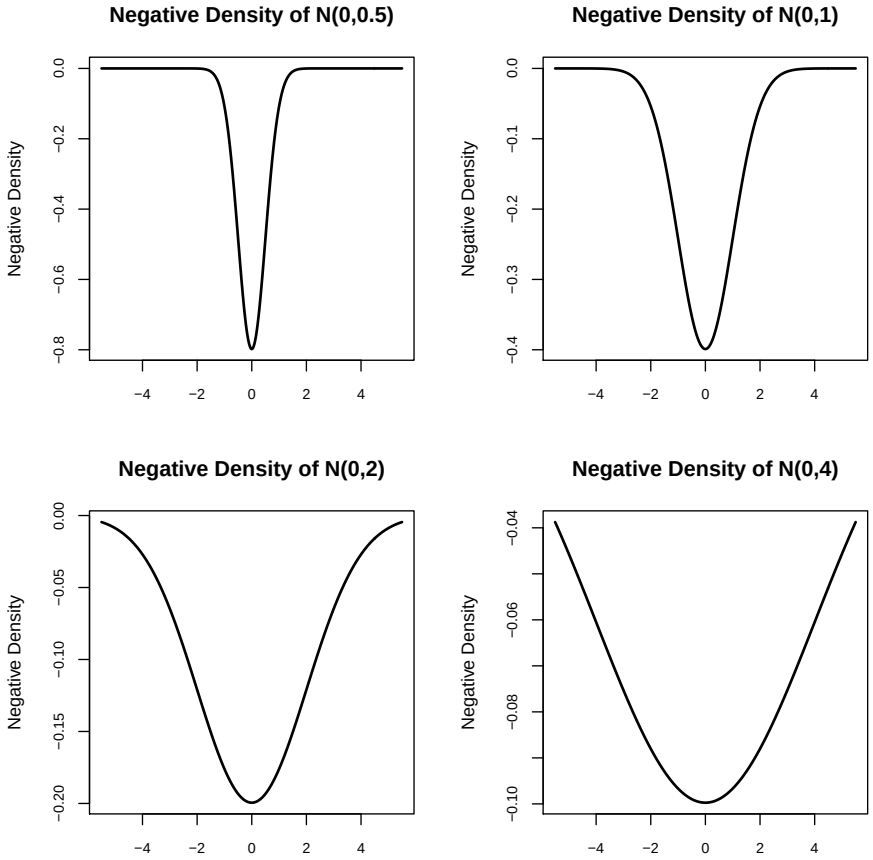
برآورد می‌شود که در آن $\phi_h(t) = \frac{۱}{h}\phi(t/h)$، $\phi(.)$ یک تابع چگالی هسته متقارن حول صفر و h یک پارامتر هموارساز است که پهنای باند نامیده می‌شود.

به منظور درک علت استفاده از (۳.۴) برای برآورد رگرسیون نما، مدل تنها شامل عرض از مبدأ را در نظر بگیرید. در این مدل، هیچ متغیر توضیحی در آن حضور ندارد، به عبارتی $\boldsymbol{\beta} = \boldsymbol{\beta}_۰$. در این‌صورت تابع هدف عبارتست از

$$Q_h(\boldsymbol{\beta}_۰) = \sum_{i=۱}^{n} \phi_h\big(Y_i - \boldsymbol{\beta}_۰\big). \tag{۴.۴}$$

$Q_h(\boldsymbol{\beta}_۰)$، به عنوان یک تابع از $\boldsymbol{\beta}_۰$، برآورد هسته تابع چگالی Y است. بنابراین ماکزیمم‌کننده (۴.۴)، نمای تابع چگالی هسته براساس $(Y_۱, ..., Y_n)$ است. هنگامی که $n \to \infty$ و $h \to ۰$، نمای تابع چگالی هسته به نمای توزیع Y همگرا می‌شود. این برآوردگر نما توسط پارزن[۱۲] (۱۹۶۲) معرفی شد. هنگامی که متغیرهای توضیحی در مدل حضور دارند، برای هر $\boldsymbol{\beta}$ ثابت، $Q_h(\boldsymbol{\beta})$ در (۳.۴) مقدار تابع چگالی هسته براساس باقیمانده‌های $\varepsilon_i = Y_i - \mathbf{X}_i^T\boldsymbol{\beta}$ در $\varepsilon = ۰$ است. ماکزیمم کردن (۳.۴) نسبت به $\boldsymbol{\beta}$ خط $\mathbf{X}^T\hat{\boldsymbol{\beta}}$ را نتیجه می‌دهد به‌طوری که تابع چگالی هسته باقیمانده‌های ε_i دارای بیشترین مقدار در ۰ است. در حالت خاص، اگر هسته یکنواخت $\phi_h(t) = \frac{۱}{۲h}I(|t| \le h)$ استفاده شود، ماکزیمم کردن (۳.۴) خط $\mathbf{X}^T\hat{\boldsymbol{\beta}}$ را نتیجه می‌دهد به‌طوری‌که باند $\mathbf{X}^T\hat{\boldsymbol{\beta}} \pm h$ بیشترین تعداد از مقادیر پاسخ Y_i را شامل می‌شود.

[12] Parzen



شکل ۱.۴: نمودار $-\phi_h(.)$ به ازای پارامترهای مختلف.

یائو و لی (۲۰۱۴) ثابت کردند اگر وقتی که $n \to \infty$ داشته باشیم $h \to 0$، آنگاه برای هر تابع چگالی خطا (بدون فرض متقارن بودن توزیع خطا)، $\hat{\boldsymbol{\beta}}$ به‌دست آمده با ماکزیمم کردن تابع هدف (۳.۴)، یک برآوردگر سازگار از بردار ضرایب $\boldsymbol{\beta}$ در (۱.۴) است.

در حالت کلی، نمای نمونه‌ای، به عنوان برآوردگری از نمای جامعه، نسبت به داده‌های پرت حساس نیست (یائو و همکاران، ۲۰۱۲). علاوه براین، اگر $-\phi_h(.)$ را به‌عنوان یک تابع زیان در نظر بگیریم، رگرسیون نما یک M- برآوردگر[۱۳] را نتیجه می‌دهد که برآوردگری نیرومند است. پهنای باند h میزان نیرومندی برآوردگر را تعیین می‌کند. فرض کنید $\phi(.)$ چگالی نرمال استاندارد باشد. شکل ۱.۴ چگونگی نیرومندی برآوردگر رگرسیون نما را نشان می‌دهد. توجه کنید که h^2 متناظر با واریانس چگالی نرمال است. از شکل ۱.۴ می‌توان دید که $-\phi_h(.)$ برای h کوچک، مانند $h = 0.5$، به‌صورت یک تابع زیان پرت‌ـمقاوم[۱۴] به نظر می‌رسد، در حالی‌که شکل این تابع برای h بزرگ، مانند $h = 4$، شبیه به تابع زیان توان دوم است. در عمل، مقدار بهینه h با یک روش مناسب مبتنی بر داده تعیین می‌شود که در ادامه به آن خواهیم پرداخت.

[13] M-estimator
[14] Outlier-resistant

۲.۴ روش نیرومند انتخاب متغیر و تشخیص ساختار

فرض کنید مجموعه داده‌های مشاهده‌شده به‌صورت

$$\mathbf{D}_n = \{(Y_i, \mathbf{X}_i, U_i), i = ۱, \ldots, n\}$$

باشد که Y_i پاسخ i امین مشاهده، $\mathbf{X}_i \in \mathbb{R}^p$ بردار متغیرهای توضیحی i امین مشاهده و U_i متغیر شاخص که، بدون از دست دادن کلیت، فرض می‌شود دارای توزیع یکنواخت $[۰, ۱]$ است. لذا مدل ضریب متغیر خطی–جزئی با ساختار درست به‌صورت زیر بیان می‌شود

$$Y_i = \sum_{j \in \mathcal{I}_V} \theta_j(U_i) X_{ij} + \sum_{j \in \mathcal{I}_C} \beta_j X_{ij} + \sum_{j \in \mathcal{I}_Z} ۰(U_i) X_{ij} + \varepsilon_i \qquad (۵.۴)$$

که $\{\theta_j(.), j \in \mathcal{I}_V\}$ توابع هموار نامعلوم، $\{\beta_j, j \in \mathcal{I}_C\}$ ضرایب رگرسیونی نامعلوم و $۰(.)$ نشان‌دهنده ضرایب صفر است. همچنین مجموعه‌های $\mathcal{I}_V, \mathcal{I}_C$ و $\mathcal{I}_Z$ مجموعه‌های اندیس، به ترتیب، برای X های با اثرات متغیر، اثرات ثابت غیر صفر و اثرات صفر هستند. این مجموعه‌ها دو به دو مجزا و در شرط $\mathcal{I}_V \cup \mathcal{I}_C \cup \mathcal{I}_Z = \{۱, \ldots, p\}$ صدق می‌کنند.

نخستین گام برای استفاده از مدل (۵.۴)، تعیین مجموعه‌های $\mathcal{I}_V, \mathcal{I}_C$ و $\mathcal{I}_Z$ می‌باشد. شناسایی متغیرهای توضیحی با اثر ثابت و اثر متغیر در عمل بسیار حائز اهمیت است و لذا در این فصل، هدف ما تشخیص این مجموعه‌ها و همچنین برآورد ضرایب متغیر هموار $\{\theta_j(U), j \in \mathcal{I}_V\}$ و ضرایب ثابت $\{\beta_j, j \in \mathcal{I}_C\}$ بطور کارا و نیرومند است.

در ابتدای تحلیل، چون ساختار خطی–جزئی نامعلوم است، مدل (۵.۴) را به‌صورت مدل ضریب متغیر زیر در نظر می‌گیریم

$$Y_i = \sum_{j=۱}^{p} \alpha_j(U_i) X_{ij} + \varepsilon_i, \qquad i = ۱, \ldots, n. \qquad (۶.۴)$$

حال اگر، برای هر $U \in [۰, ۱]$، داشته باشیم $\alpha_j(U) \equiv ۰$، آنگاه X_j یک متغیر بی اهمیت یا نامرتبط است. اگر $\alpha_j(U)$ یک مقدار ثابت غیر صفر باشد، X_j دارای اثر ثابت و در غیر این‌صورت، $\alpha_j(U)$ یک تابع متغیر و X_j دارای اثر متغیر است. بنابراین، می‌خواهیم تعیین کنیم که کدام‌یک از ضرایب $\alpha_j(.)$ توابعی متغیر، صفر یا ثابت غیر صفر هستند.

بدین منظور، ابتدا $\alpha_j(U)$ را به‌صورت زیر به دو جزء ثابت و متغیر تجزیه می‌کنیم

$$\alpha_j(U) = \eta_j + g_j(U), \qquad j = ۱, \ldots, p \qquad (۷.۴)$$

که برای تضمین شناسایی‌پذیری[15] فرض می‌شود $E\{g_j(U)\} = ۰$. به‌وضوح تجزیه (۷.۴) یکتاست و

$$\eta_j = E\{\alpha_j(U)\}, \qquad g_j(U) = \alpha_j(U) - E\{\alpha_j(U)\}.$$

[15] Identifiability

اکنون می‌توانیم از اسپلاین‌های چند جمله‌ای برای تقریب توابع ناپارامتری $g_j(.)$ استفاده کنیم. فرض کنید $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K_n} < \tau_{K_n+1} = 1$ به $K_n + 1$ افرازی از بازه $[0, 1]$ زیر بازه $I_j = [\tau_j, \tau_{j+1}), j = 0, 1, \ldots, K_n - 1,$ و $I_{K_n} = [\tau_{K_n}, 1]$ باشد که K_n تعداد گره‌های داخلی است. مجموعه توابع اسپلاین درجه q، دارای پایه‌های B ـ اسپلاین $\mathbf{B}(.) = \big(B_1(.), \ldots, B_{K'}(.)\big)^T$ می‌باشند که $K' = K_n + q + 1$. با در نظر گرفتن شرط محدودکننده $E\{g_j(U)\} = 0$، تنها فضای توابع اسپلاین $S_j^0 = \{s : s = \sum_{k=1}^{K'} \gamma_{jk} B_k(x), \sum_{i=1}^n s(X_{ij}) = 0\}$ با پایه‌های مرکزی شده $\{B_{jk}(x) = B_k(x) - \sum_{i=1}^n B_k(X_{ij})/n, k = 1, 2, \ldots, K\}$ را در نظر می‌گیریم که $K = K' - 1$. لذا تابع $g_j(U)$ را به‌صورت

$$g_j(U) \approx \sum_{k=1}^{K} B_{jk}(U)\gamma_{j,k} = \mathbf{B}(U)^T \boldsymbol{\gamma}_j, \qquad j = 1, \ldots, p \tag{۸.۴}$$

تقریب می‌زنیم که $\boldsymbol{\gamma_j} = (\gamma_{j,1}, \ldots, \gamma_{j,K})^T$. فرض کنید $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^T$، $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ و $\boldsymbol{\Pi}_i^T = \big(X_{i1}\mathbf{B}(U_i)^T, \ldots, X_{ip}\mathbf{B}(U_i)^T\big)$. با استفاده از تقریب (۸.۴)، مدل (۶.۴) را می‌توان به‌صورت

$$Y_i = \mathbf{X}_i^T \boldsymbol{\eta} + \boldsymbol{\Pi}_i^T \boldsymbol{\gamma} + \varepsilon \tag{۹.۴}$$

نوشت. مدل (۹.۴) یک مدل خطی استاندارد است و هر تابع ناپارامتری $\alpha_j(.)$ در مدل (۶.۴)، با استفاده از η_j و $\boldsymbol{\gamma}_j$ در مدل (۹.۴) مشخص می‌شود. حال برای برآورد نیرومند پارامترهای $\boldsymbol{\eta}$ و $\boldsymbol{\gamma}$ از رگرسیون نما استفاده می‌کنیم. این روش علاوه بر نیرومند بودن، برآوردی را نتیجه می‌دهد که بسیار کارا است. بنابراین، با ترکیب رگرسیون نما و توابع اسپلاین، تابع هدف زیر را در نظر می‌گیریم

$$Q(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \phi_h\big(Y_i - \mathbf{X}_i^T \boldsymbol{\eta} - \boldsymbol{\Pi}_i^T \boldsymbol{\gamma}\big) \tag{۱۰.۴}$$

که $\phi_h(t) = \phi(t/h)/h$ و $\phi(.)$ تابع چگالی هسته با پهنای باند h است. برای سهولت در محاسبات، سرتاسر این فصل، از تابع چگالی نرمال استاندارد به عنوان تابع چگالی هسته استفاده می‌شود. با ماکزیمم کردن تابع هدف (۱۰.۴)، برآوردهای ضرایب $\hat{\boldsymbol{\eta}}$ و $\hat{\boldsymbol{\gamma}}$ به‌دست می‌آیند. در نتیجه ضرایب $\alpha_j(U)$ به‌صورت

$$\hat{\alpha}_j(U) = \hat{\eta}_j + \sum_{k=1}^{K} B_{jk}(U)\hat{\gamma}_{j,k}$$

برآورد می‌شود. به خاطر داشته باشید که هدف اصلی ما تعیین متغیرهای بی‌اهمیت $(\alpha_j(U) \equiv 0)$ و مؤلفه‌های پارامتری $(\alpha_j(U) \equiv \alpha_j)$ است. اگر $\| \boldsymbol{\gamma}_j \|_1 = \sum_{k=1}^{K} |\gamma_{j,k}| \neq 0$، آنگاه X_j دارای اثر متغیر است. اگر $\| \boldsymbol{\gamma}_j \|_1 = 0$، و $\eta_k \neq 0$ باشد، X_j دارای اثر ثابت غیر صفر است. در غیر این صورت X_j بی‌اثر است.

تابع هدف (۱۰.۴) قابلیت انتخاب متغیرهای مهم و تشخیص اثرات ثابت و متغیر را ندارد. لذا، به جای (۱۰.۴) از تابع هدف جریمه‌شده با دو تابع جریمه به‌صورت

$$L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \phi_h\big(Y_i - \mathbf{X}_i^T \boldsymbol{\eta} - \boldsymbol{\Pi}_i^T \boldsymbol{\gamma}\big) - n\sum_{j=1}^{p} p_{\lambda_1}(|\eta_j|) - n\sum_{j=1}^{p} p_{\lambda_2}(\|\boldsymbol{\gamma}_j\|_{\mathbf{R}}) \tag{۱۱.۴}$$

استفاده می‌شود که $\|\gamma_j\|_{\mathbf{R}} = \left(\gamma_j^T \mathbf{R} \gamma_j\right)^{\frac{1}{2}}$، $\mathbf{R} = \int_{[0,1]} \mathbf{B}(u)\mathbf{B}(u)^T du$ و $p_\lambda(.)$ تابع جریمه SCAD (فن و لی، ۲۰۰۱) و $\lambda_1, \lambda_2 \geq 0$ پارامترهای جریمه هستند. تابع جریمه اول در (۱۱.۴)، ضرایب ثابت η_j در $\alpha_j(U)$ را به سمت صفر منقبض می‌کند. تابع جریمه دوم منقبض کردن بخش تابعی، $g_j(U)$، را انجام می‌دهد. بنابراین اگر X_j دارای اثر ثابت باشد، همه مولفه های γ_j دقیقا صفر برآورد می شوند. از طرفی اگر X_j دارای هیچ اثری روی متغیر پاسخ نباشد، η_j و همه مولفه های γ_j به سمت صفر منقبض می شوند. در غیر از این دوحالت، X_j دارای اثر متغیر است.

فرض کنید

$$(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}) = \arg\max_{\boldsymbol{\eta},\boldsymbol{\gamma}} L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma}).$$

بنابراین، مجموعه اندیس اثرات متغیر، اثرات ثابت و اثرات صفر، به ترتیب، به‌صورت

$$\hat{\mathcal{I}_V} = \{j : \|\hat{\gamma}_j\|_1 \neq 0, j = 1, \ldots, p\}$$
$$\hat{\mathcal{I}_C} = \{j : \|\hat{\gamma}_j\|_1 = 0, \hat{\eta}_j \neq 0, j = 1, \ldots, p\}$$
$$\hat{\mathcal{I}_Z} = \{j : \|\hat{\gamma}_j\|_1 = 0, \hat{\eta}_j = 0, j = 1, \ldots, p\}$$

تعیین می‌شوند. در نتیجه اگر $\alpha_j(U)$ ضریب متغیر باشد، به‌صورت $\hat{\alpha}_j(U) = \hat{\eta}_j + \mathbf{B}(U)^T\hat{\gamma}_j$ و در صورت ثابت بودن به‌صورت $\hat{\eta}_j$ برآورد می‌شود.

۳.۴ ویژگی‌های مجانبی

در این بخش، به بررسی برخی ویژگی‌های مجانبی برآوردگرهای رگرسیون نمای جریمه‌شده در رابطه (۲.۴)، $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}})$، می‌پردازیم. فرض کنید $\alpha_{0j}(\cdot)$ مقدار درست $\alpha_j(\cdot)$ باشد. با توجه به تجزیه (۷.۴)، $\eta_{0j} = E\{\alpha_{0j}(U)\}$ و $g_{0j}(\cdot) = \alpha_{0j}(\cdot) - E\{\alpha_{0j}(U)\}$، به ترتیب، مقادیر درست η_j و $g_j(\cdot)$ هستند. فرض کنید بردار $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{01}^T, \ldots, \boldsymbol{\gamma}_{0p}^T)^T$ ضریب بهترین تقریب برای $g_j(U)$ در فضای B- اسپلاین باشد، به‌طوری که

$$\| \mathbf{B}(U)^T\boldsymbol{\gamma}_{0j} - g_{0j}(U) \|_\infty = O(K_n^{-r}),\ j \in \mathcal{I}_V, \qquad \| \boldsymbol{\gamma}_{0j} \|_1 = 0,\ j \in \mathcal{I}_C \cup \mathcal{I}_Z$$

که r در شرط C۲ زیر تعیین می‌شود. تعریف کنید

$$G(x, u, h) = E\{\phi_h'^2(\varepsilon)|\mathbf{X} = x, U = u\}$$
$$F(x, u, h) = E\{\phi_h''(\varepsilon)|\mathbf{X} = x, U = u\}$$

و

$$a_n = \max_{k,j}\{|p'_{\lambda_1}(|\eta_{0k}|)|, |p'_{\lambda_2}(\| \boldsymbol{\gamma}_{0j} \|_{\mathbf{R}})| : |\eta_{0k}| \neq 0, \| \boldsymbol{\gamma}_{0j} \|_1 \neq 0\}$$
$$b_n = \max_{k,j}\{|p''_{\lambda_1}(|\eta_{0k}|)|, |p''_{\lambda_2}(\| \boldsymbol{\gamma}_{0j} \|_{\mathbf{R}})| : |\eta_{0k}| \neq 0, \| \boldsymbol{\gamma}_{0j} \|_1 \neq 0\}.$$

برای اثبات ویژگی‌های مجانبی برآوردگرهای ارائه‌شده، به شرایط نظم زیر نیاز داریم:

C۱. متغیر شاخص U دارای تکیه‌گاه کراندار Ω و تابع چگالی آن، $f_U(.)$، مثبت و دارای مشتق دوم پیوسته است. بدون از دست دادن کلیت، فرض می‌کنیم Ω روی فاصله واحد $[۰, ۱]$ باشد.

C۲. به‌ازای هر $j \in \mathcal{I}_v$ داریم $\alpha_{\circ j}(u) \in \mathcal{H}_r$، که $\mathcal{H}_r$ مجموعه همه توابع g روی $[۰, ۱]$ است، به‌طوری که مشتق مرتبه ν ام g، g^ν، موجود و در شرط لیپ‌شیتز مرتبه α صدق می‌کند، یعنی

$$|g^\nu(s) - g^\nu(t)| \leq C|s-t|^\alpha, \qquad ۰ \leq s,t \leq ۱$$

که C یک مقدار ثابت مثبت، ν یک عدد صحیح مثبت و $\alpha \in (۰, ۱]$ است به‌طوری که $r = \nu + \alpha > ۲$.

C۳. فرض کنید $\mathbf{\Sigma}(u) = E\{\mathbf{X}\mathbf{X}^T | U = u\}$ نسبت به u پیوسته است. علاوه براین، برای u داده‌شده، $\mathbf{\Sigma}(u)$ ماتریس معین مثبت است و مقادیر ویژه آن کراندارند. همچنین فرض می‌کنیم $\max_i \parallel \mathbf{X}_i \parallel / \sqrt{n} = o_p(۱)$.

C۴. فرض کنید $\xi_j = \tau_j - \tau_{j-۱}$ و $\xi = \max_j\{\xi_j\}$. یک مقدار ثابت مانند $C_\circ$ وجود دارد به‌طوری که $\dfrac{\xi}{\min_j\{\xi_j\}} \leq C_\circ$ و $\max_j\{|\xi_{j+۱} - \xi_j|\} = o(K_n^{-۱})$.

C۵. $F(x,u,h)$ و $G(x,u,h)$ نسبت به (x,u) پیوسته هستند.

C۶. برای هر $h > ۰$، $F(x,u,h) < ۰$.

C۷. $E\{\phi_h'(\varepsilon)|\mathbf{X} = x, U = u\} = ۰$. همچنین میانگین‌های شرطی $E\{\phi_h''^{۲}(\varepsilon)|\mathbf{X} = x, U = u\}$، $E\{\phi_h'^{۳}(\varepsilon)|\mathbf{X} = x, U = u\}$ و $E\{\phi_h'''(\varepsilon)|\mathbf{X} = x, U = u\}$ نسبت به x پیوسته هستند.

C۸. به‌ازای $j \in \mathcal{I}_Z \cup \{l : E\alpha_{\circ l}(U) = ۰, l \in \mathcal{I}_V\}$ داریم $\liminf_{n\to\infty} \liminf_{\eta_j \to ۰^+} p'_{\lambda ۱}(\eta_j)/\lambda_۱ > ۰$ و به‌ازای $j \in \mathcal{I}_C \cup \mathcal{I}_Z$ داریم $\liminf_{n\to\infty} \liminf_{\|\gamma_j\|_{\mathbf{R}} \to ۰^+} p'_{\lambda ۲}(\parallel \gamma_j \parallel_{\mathbf{R}})/\lambda_۲ > ۰$.

در ادامه، قضیه ۱.۳.۴ نرخ همگرایی برآوردگرهای ضرایب متغیر و قضیه ۲.۳.۴ توزیع مجانبی برآوردگرهای ضرایب ثابت غیر صفر را ارائه می‌دهد.

قضیه ۱.۳.۴. فرض کنید شرایط C۱ − C۸ برقرار هستند و تعداد گره‌ها $K_n = O(n^{\frac{۱}{۲r+۱}})$ باشد. اگر وقتی $n \to \infty$، داشته باشیم $b_n \to ۰$، آنگاه

$$\parallel \hat{\alpha}_k(u) - \alpha_{\circ k}(u) \parallel_۲ = O_p(n^{\frac{-r}{۲r+۱}} + a_n), \quad k \in \mathcal{I}_V.$$

قضیه ۲.۳.۴. تحت شرایط C۱–C۸، فرض کنید تعداد گره‌ها $K_n = O(n^{\frac{۱}{۲r+۱}})$، $\lambda_{\max} = \max\{\lambda_۱, \lambda_۲\}$ و $\lambda_{\min} = \min\{\lambda_۱, \lambda_۲\}$ باشد. اگر $\lambda_{\max} \to ۰$ و $n^{\frac{r}{۲r+۱}}\lambda_{\min} \to \infty$، وقتی $n \to \infty$، آنگاه

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C} - \boldsymbol{\eta}_{۰\mathcal{I}_C}) \xrightarrow{d} \mathcal{N}\big(۰, G(h)F^{-۲}(h)\boldsymbol{\Sigma}_{\mathcal{I}_C}^{-۱}\big)$$

که $\boldsymbol{\eta}_{۰_{\mathcal{I}_C}}$ مقدار درست بردار ضرایب ثابت غیر صفر، $\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}$ برآوردگرآن، $G(h)$ و $F(h)$ به‌صورت

$$G(h) = \frac{۱}{۲\pi h^۶}E\{\varepsilon \exp(-\varepsilon^۲/۲h^۲)\}^۲, F(h) = \frac{۱}{\sqrt{۲\pi}h^۳}E\{\exp(-\varepsilon^۲/۲h^۲)(\varepsilon^۲/h^۲ - ۱)\}$$

و $\boldsymbol{\Sigma}_{\mathcal{I}_C}$ در رابطه (آ۲۹.) تعریف می‌شود.

۴.۴ انتخاب پهنای باند و الگوریتم پیاده‌سازی

در این بخش، ابتدا به موضوع انتخاب پهنای باند، h، می‌پردازیم. سپس از الگوریتم MEM (لی و همکاران، ۲۰۰۷) و تقریب LQA (فن و لی، ۲۰۰۱) برای پیاده‌سازی روش ارائه‌شده استفاده می‌کنیم.

۱.۴.۴ انتخاب پهنای باند بهینه

نیرومندی و کارایی برآوردگرهای نما، به پهنای باند h بستگی دارد. بنابراین، مقدار بهینه آن (h_{opt}) را طوری انتخاب می‌کنیم که واریانس مجانبی برآوردگرهای ارائه‌شده را، در مقایسه با روش کمترین توان‌های دوم، تا حد ممکن کوچک کند. برای سهولت، فرض می‌شود که ε از $\mathbf{X}$ و U مستقل است. ژائو و زوئی (۲۰۰۹) نشان دادند که برآوردگر کمترین توان‌های دوم ضرایب رگرسیونی ثابت در مدل ضریب متغیر خطی–جزئی دارای توزیع مجانبی زیر است

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C} - \boldsymbol{\eta}_{۰\mathcal{I}_C}) \xrightarrow{d} \mathcal{N}\big(۰, \sigma^۲\boldsymbol{\Sigma}_{\mathcal{I}_C}^{-۱}\big)$$

که $\boldsymbol{\Sigma}_{\mathcal{I}_C}$ در قضیه ۲.۳.۴ تعریف شده‌است. بنابراین نسبت واریانس مجانبی برآورد رگرسیون نما به واریانس مجانبی برآوردگر کمترین توان‌های دوم به‌صورت

$$r(h) = \frac{G(h)F^{-۲}(h)}{\sigma^۲}$$

به‌دست می‌آید که $\sigma^۲ = E(\varepsilon^۲)$، $G(h)$ و $F(h)$ در قضیه ۲.۳.۴ داده شده‌اند. نسبت $r(h)$ فقط به h وابسته است. بنابراین، مقدار بهینه h به‌صورت

$$h_{opt} = \arg\min_h r(h) = \arg\min_h G(h)F^{-۲}(h) \qquad (۱۲.۴)$$

حاصل می‌شود. از رابطه (۱۲.۴) می‌توان دید که h_{opt} به n بستگی نداشته و تنها به توزیع شرطی ε وابسته است.

در عمل، به دلیل نامعلوم بودن توزیع خطاها، محاسبه مستقیم $G(h)$ و $F(h)$ امکان‌پذیر نیست. از این رو، ابتدا باقیمانده‌های $\hat{\varepsilon}_i = Y_i - \sum_{j=1}^{p} \hat{\alpha}_j(U_i)X_{ij}$ را محاسبه می‌کنیم که ضرایب $\hat{\alpha}_j(.)$ با یک روش ساده برآورد می‌شوند. بهتر است برای محاسبه این ضرایب از یک روش نیرومند ساده مانند رگرسیون چندکی استفاده شود. پس از محاسبه باقیمانده‌ها، $G(h)$ و $F(h)$ را به‌صورت

$$\hat{G}(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2\pi h^6}\{\hat{\varepsilon}\exp(-\hat{\varepsilon}^2/2h^2)\}^2$$

$$\hat{F}(h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h^3}\{\exp(-\hat{\varepsilon}^2/2h^2)(\hat{\varepsilon}^2/h^2 - 1)\}$$

برآورد می‌کنیم. بنابراین، $r(h)$ به‌صورت

$$\hat{r}(h) = \frac{\hat{G}(h)\hat{F}^{-2}(h)}{\hat{\sigma}^2}$$

برآورد می‌شود. با این حال، چون هیچ جواب صریحی برای h وجود ندارد، از روش جستجوی شبکه‌ای[16] استفاده می‌کنیم. براساس پیشنهاد یائو و همکاران (۲۰۱۲) و ژانگ و همکاران (۲۰۱۳)، نقاط ممکن برای h را به‌صورت

$$h = 0.5\hat{\sigma} \times 1.02^j, \quad j = 0, \ldots k$$

در نظر می‌گیریم که k یک مقدار ثابت مانند $k = 50$ یا $k = 100$ است. در نتیجه، با استفاده از روش جستجوی شبکه‌ای، پهنای باند براساس (۱۲.۴) به‌دست می‌آید.

۲.۴.۴ الگوریتم MEM

برای حل مسئله بهینه‌سازی (۱۱.۴)، ابتدا از تقریب LQA (فن و لی، ۲۰۰۱)، برای تقریب توابع جریمه $p_{\lambda_1}(.)$ و $p_{\lambda_2}(.)$ استفاده می‌کنیم. براساس این تقریب، برای مقادیر اولیه داده‌شده η_j° و γ_j°، توابع جریمه را به‌صورت

$$p_{\lambda_1}(|\eta_j|) \approx p_{\lambda_1}\left(|\eta_j^{(\circ)}|\right) + \frac{1}{2}\frac{p'_{\lambda_1}\left(|\eta_j^{(\circ)}|\right)}{|\eta_j^{(\circ)}|}\left\{|\eta_j|^2 - |\eta_j^{(\circ)}|^2\right\}$$

و

$$p_{\lambda_2}\left(\|\gamma_j\|_{\mathbf{R}}\right) \approx p_{\lambda_2}\left(\|\gamma_j^{(\circ)}\|_{\mathbf{R}}\right) + \frac{1}{2}\frac{p'_{\lambda_2}\left(\|\gamma_j^{(\circ)}\|_{\mathbf{R}}\right)}{\|\gamma_j^{(\circ)}\|_{\mathbf{R}}}\left\{\|\gamma_j\|_{\mathbf{R}}^2 - \|\gamma_j^{(\circ)}\|_{\mathbf{R}}^2\right\}$$

تقریب می‌زنیم. فرض کنید $\boldsymbol{Z}_i = \left(\mathbf{X}_i^T, \mathbf{\Pi}_i^T\right)^T$. قرار دهید $m = 0$ و $\boldsymbol{\theta}^{(m)} = (\boldsymbol{\eta}^{(m)^T}, \boldsymbol{\gamma}^{(m)^T})^T$. برای $\varepsilon_1 > 0$ و $\varepsilon_2 > 0$، ماتریس بلوکی $\mathbf{\Sigma}_\lambda(\boldsymbol{\theta}^{(m)})$ را به‌صورت زیر تعریف کنید

$$\mathbf{\Sigma}_\lambda(\boldsymbol{\theta}^{(m)}) = \text{diag}\left(\frac{p'_{\lambda_1}\left(|\eta_1^{(m)}|\right)}{|\eta_1^{(m)}| + \varepsilon_1}, \ldots, \frac{p'_{\lambda_1}\left(|\eta_p^{(m)}|\right)}{|\eta_p^{(m)}| + \varepsilon_1}, \frac{p'_{\lambda_2}\left(\|\gamma_1^{(m)}\|_{\mathbf{R}}\right)}{\|\gamma_1^{(m)}\|_{\mathbf{R}} + \varepsilon_2}\mathbf{R}, \ldots, \frac{p'_{\lambda_2}\left(\|\gamma_p^{(m)}\|_{\mathbf{R}}\right)}{\|\gamma_p^{(m)}\|_{\mathbf{R}} + \varepsilon_2}\mathbf{R}\right).$$

[16] Grid search method

براساس پیشنهاد هانتر و لی (۲۰۰۵)، ε_1 و ε_2 را به‌صورت

$$\varepsilon_1 = \frac{\tau}{2n\lambda_1}\min\{|\eta_j^{(0)}|, j = 1,\ldots,p\}, \qquad \varepsilon_2 = \frac{\tau}{2n\lambda_2}\min\{\|\gamma_j^{(0)}\|_{\mathbf{R}}, j = 1,\ldots,p\}$$

در نظر می‌گیریم که $\tau = 10^{-8}$. اکنون می‌توانیم با در نظر گرفتن تابع چگالی هسته نرمال استاندارد و با استفاده از الگوریتم MEM، پارامترهای مجهول را به‌صورت زیر برآورد کنیم:

- گام ۱ (E): $\pi(i|\boldsymbol{\theta}^{(m)})$ را به‌صورت زیر به‌روز رسانی کنید

$$\pi(i|\boldsymbol{\theta}^{(m)}) = \frac{\phi_h(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\theta}^{(m)})}{\sum_{i=1}^n \phi_h(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\theta}^{(m)})} \propto \phi_h(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\theta}^{(m)}), \quad i = 1,...,n.$$

- گام ۲ (M): مقدار $\boldsymbol{\theta}$ را به‌صورت زیر به‌روز رسانی کرده تا $\boldsymbol{\theta}^{(m+1)}$ به‌دست آید

$$\begin{aligned}\boldsymbol{\theta}^{(m+1)} &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n \pi(i|\boldsymbol{\theta}^{(m)}) \log\phi_h(Y_i - \boldsymbol{Z}_i^T\boldsymbol{\theta}^{(m)}) + \frac{n}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}_\lambda(\boldsymbol{\theta}^{(m)})\boldsymbol{\theta} \\ &= \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\theta}^{(m)})\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Y}\end{aligned}$$

که $\boldsymbol{W} = \mathrm{diag}\{\pi(1|\boldsymbol{\theta}^{(m)}), \ldots, \pi(n|\boldsymbol{\theta}^{(m)})\}$ است.

- گام ۳: قرار دهید $m = m + 1$ و گام های اول و دوم را تا رسیدن به همگرایی تکرار کنید.

در این الگوریتم، شرط توقف تکرار، $\max\{\| \hat{\boldsymbol{\theta}}^{(m+1)} - \hat{\boldsymbol{\theta}}^{(m)} \|\} < 10^{-8}$ است. برآورد حاصل را با $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\eta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ نشان می‌دهیم. در عمل مؤلفه‌های $\hat{\boldsymbol{\theta}}$ دقیقاً برابر با صفر نیستند. لذا با استفاده از یک مقدار آستانه‌ای، مؤلفه‌های کوچک را برابر صفر قرار می‌دهیم. در مطالعات شبیه‌سازی، هنگامی که $|\hat{\eta}_j| < 10^{-4}$، آنگاه $\hat{\eta}_j$ را برابر صفر، و هنگامی که $\| \hat{\gamma}_j \|_2 < 10^{-4}$، آنگاه قرار می‌دهیم $\hat{\gamma}_j = 0$. علاوه بر این از برآورد رگرسیون چندکی به عنوان برآورد اولیه ضرایب استفاده می‌شود.

در الگوریتم فوق، وقتی از هسته نرمال برای $\phi(\cdot)$ استفاده می‌شود، تابع بهینه‌شده در گام ۲ مجموع وزنی لگاریتم تابع درستنمایی متناظر با رگرسیون خطی معمولی است. در این حالت، یک عبارت بسته‌شکل برای ماکزیمم‌کننده حاصل می‌شود. اگر در گام ۲ هسته دیگری، غیر از نرمال، به‌کار رود، برای ماکزیمم‌کردن لازم است از برخی الگوریتم‌های بهینه‌سازی استفاده شود.

۳.۴.۴ انتخاب پارامترهای تنظیم‌کننده

برای پیاده‌سازی روش ارائه‌شده، لازم است مقدار مناسبی برای پارامترهای تنظیم‌کننده انتخاب شود. این پارامترها عبارتند از: مرتبه B-اسپلاین q، تعداد گره‌های داخلی K_n و پارامترهای جریمه λ_1 و λ_2. در مطالعات شبیه‌سازی از اسپلاین مکعبی ($q = 3$) و $K_n = [n^{1/(2q+3)}]$ گره داخلی با فواصل یکسان استفاده می‌شود. این روش انتخاب گره‌ها توسط زوئی و کیو[17] (۲۰۱۲) پیشنهاد شده‌است.

[17] Qu

همچنین مقدار بهینه (λ_1, λ_2) را با اعتبارسنجی متقابل ۵ گروهی انتخاب می‌کنیم. بطور واضح‌تر، داده‌های $\mathbf{D}_n$ را بطور تصادفی به ۵ گروه با اندازه تقریباً یکسان افراز می‌کنیم. این گروه‌ها را، برای $j = ۱, ۲, ۳, ۴, ۵$ به‌صورت $\{\mathbf{D}_i, i \in \mathcal{I}_j\}$ نشان می‌دهیم که $\mathcal{I}_j$ مجموعه اندیس مشاهدات متناظر با j امین گروه می‌باشد. برای هر j، مجموعه $\{\mathbf{D}_i, i \in \mathcal{I}_j\}$ را به عنوان داده‌های آزمون و ۴ گروه دیگر را به عنوان داده‌های آموزشی در نظر می‌گیریم. برای هر زوج مشخص (λ_1, λ_2) و هر $\mathcal{I}_j$، ابتدا از داده‌های آموزشی برای برآورد ضرایب متغیر براساس (۱۱.۴) استفاده کرده و متغیر پاسخ را برای داده‌های آزمون به‌صورت $\hat{Y}_i = \mathbf{X}_i^T \hat{\boldsymbol{\eta}} + \boldsymbol{\Pi}_i^T \hat{\boldsymbol{\gamma}},\ i \in \mathcal{I}_j$ پیشگویی می‌کنیم. سپس امتیاز اعتبارسنجی متقابل متناظر با این زوج (λ_1, λ_2) را به‌صورت

$$CV(\lambda_1, \lambda_2) = \sum_{j=1}^{5} \sum_{i \in \mathcal{I}_j} \phi_{h_{opt}}\left(Y_i - \hat{Y}_i\right) \qquad (۱۳.۴)$$

محاسبه می‌کنیم. در پایان، λ_1 و λ_2 بهینه با ماکزیمم‌کردن $CV(\lambda_1, \lambda_2)$ به‌دست می‌آید.

۵.۴ مطالعات شبیه‌سازی

در این بخش، با استفاده از شبیه‌سازی به بررسی عملکرد رگرسیون نمای جریمه‌شده می‌پردازیم. برای سادگی، رگرسیون نمای جریمه‌شده را با PMR نشان می‌دهیم. عملکرد این روش را با دو روش تکراری ارائه‌شده توسط تانگ و همکاران (۲۰۱۲)، رگرسیون کمترین توان‌های دوم جریمه‌شده (PLS) و رگرسیون چندکی جریمه‌شده با چندک $\tau = ۰/۵$ (PQR)، مقایسه می‌کنیم. تانگ و همکاران (۲۰۱۲) از روش‌های تکراری دومرحله‌ای برای انتخاب متغیر و تشخیص ساختار در مدل ضریب متغیر خطی‌ـ جزئی استفاده کردند. آنها در مرحله اول به تشخیص ضرایب ثابت و متغیر، و در مرحله دوم به تشخیص ضرایب ثابت غیر صفر و صفر پرداختند. برای بررسی نیرومندی و کارایی روش ارائه‌شده، سه نوع توزیع مختلف برای خطا در نظر می‌گیریم: توزیع نرمال استاندارد، توزیع تی‌ـاستودنت با ۳ درجه آزادی و توزیع آمیخته نرمال (MN) که به‌صورت $\text{MN} : ۰/۹\mathcal{N}(۰, ۱) + ۰/۱\mathcal{N}(۰, ۲۵)$ تعریف می‌شود. برای بررسی میزان دقت روش ارائه شده در انتخاب مدل درست، از معیارهای زیر استفاده می‌کنیم:

- NV: متوسط تعداد ضرایب انتخاب شده به‌عنوان ضرایب متغیر
- NC: متوسط تعداد ضرایب انتخاب شده به‌عنوان ضرایب ثابت
- NVT: متوسط تعداد ضرایب متغیر انتخاب شده که واقعاً متغیر هستند
- NCT: متوسط تعداد ضرایب ثابت غیر صفر انتخاب شده که واقعاً ثابت غیر صفر هستند
- CF: درصد دفعاتی که ساختار مدل، دقیقاً به درستی تشخیص داده می‌شود.

برای بررسی دقت برآوردگر $\hat{\alpha}(.)$، از ریشه دوم میانگین توان‌های دوم خطا (RMSE) استفاده می‌کنیم که به‌صورت

$$\mathrm{RMSE} = \Big\{ \frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} \sum_{j=1}^{p} \big(\hat{\alpha}_j(u_i) - \alpha_j(u_i)\big)^2 \Big\}^{\frac{1}{2}} \qquad (۱۴.۴)$$

تعریف می‌شود و $\{u_1, \ldots, u_{grid}\}$ نقاط شبکه‌ای هستند که تابع $\hat{\alpha}_j(.)$ در این نقاط ارزیابی می‌شود.

مثال ۱.۵.۴. فرض کنید $p = 15$ و داده‌ها را از مدل زیر تولید کنید

$$Y_i = \sum_{j=1}^{p} \alpha_j(U_i) X_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n \qquad (۱۵.۴)$$

که $\alpha_1(u) = 2\sin(2\pi u)$، $\alpha_2(u) = 8u(1-u)$، $\alpha_3(u) = 2.5$، $\alpha_4(u) = 1$، $\alpha_5(u) = 1.5$ و برای $6 \le j \le p$، $\alpha_j(u) = 0$. لذا در این مدل $\mathrm{NVT} = 2$ و $\mathrm{NCT} = 3$ است. متغیر U از توزیع $U[0, 1]$ و متغیرهای توضیحی $(X_1, \ldots, X_p)$ از توزیع نرمال چندمتغیره با میانگین صفر و ماتریس کوواریانس $\Sigma = (\sigma_{ij})_{p\times p}$ تولید می‌شود که $\sigma_{ij} = 0.5^{|i-j|}$. نتایج شبیه‌سازی برای $n = 200, 400$، پس از ۵۰۰ تکرار، در جدول ۱.۴ گزارش شده‌اند.

در جدول ۱.۴ ملاحظه می‌شود که عملکرد هر سه روش تحت تاثیر توزیع خطاها قرار می‌گیرد. همانطور که انتظار داشتیم، برای هر نوع توزیع خطا، عملکرد هر یک از روش‌ها با افزایش حجم نمونه بهبود می‌یابد. برای هر دو مقدار حجم نمونه، رگرسیون نما، در مقایسه با دو روش دیگر، بسیار بهتر عمل می‌کند. هنگامی که توزیع خطاها نرمال است، مقادیر $\mathrm{NVT} = 1.63, 1.82$ و $\mathrm{NCT} = 2.99, 2.99$ برای $n = 200$، و $\mathrm{NVT} = 1.98, 1.98$ و $\mathrm{NCT} = 2.97, 2.99$ برای $n = 400$، نشان می‌دهد که رگرسیون نمای جریمه‌شده در تشخیص اثرات ثابت و متغیر نسبت به دو روش دیگر کاراتر است. علاوه‌براین، مقدار CF متناظر با رگرسیون نما بسیار بزرگتر از مقادیر CF متناظر با دو روش دیگر است. مشاهدات فوق از جدول ۱.۴، دلالت بر قابلیت بالای روش ارائه شده در شناسایی ساختار مدل درست دارند.

با توجه به مقادیر RMSE در جدول ۱.۴، می‌توان دید که در حضور داده‌های پرت یا توزیع دم‌سنگین خطاها، برآوردگرهای حاصل از رگرسیون نمای جریمه‌شده نسبت به رگرسیون کمترین توان‌های دوم جریمه‌شده و رگرسیون چندکی جریمه‌شده از دقت برآورد بیشتری برخوردار است، اما برای خطای نرمال، رگرسیون کمترین توان‌های دوم از رگرسیون چندکی و رگرسیون نما اندکی بهتر است.

برای بررسی بیشتر دقت برآوردگرها، ریشه دوم میانگین توان‌های دوم خطای متناظر با هر روش را با ریشه دوم میانگین توان‌های دوم برآوردگر پیشگو مقایسه کرده‌ایم. برآوردگر پیشگو برآوردگری است که با استفاده از مدل درست محاسبه می‌شود. منظور از مدل درست، مدلی است که ساختار خطی–جزئی آن از پیش معلوم است، یعنی ضرایب متغیر، ضرایب ثابت غیر صفر و صفر قبل از تحلیل داده‌ها معلوم‌اند و لذا برای به‌دست آوردن برآوردگر پیشگو از هیچ تابع جریمه‌ای استفاده نمی‌شود. نتایج شبیه‌سازی نشان می‌دهد که مقادیر RMSE برآوردگرهای جریمه‌شده و برآوردگر پیشگو بسیار نزدیک است.

جدول ۱.۴: نتایج تشخیص ساختار روش‌های مختلف در مثال ۱.۵.۴

n	خطا	روش	NV(۲)*	NVT(۲)*	NC(۳)*	NCT(۳)*	CF%	RMSE	
								جریمه‌شده	پیشگو
۲۰۰	$\mathcal{N}(۰,۱)$	PLS	۲٫۵۷	۱٫۸۸	۲٫۸۴	۲٫۷۳	۵۱	۰٫۳۹۲	۰٫۲۹۷
		PQR	۲٫۳۷	۱٫۸۹	۳٫۲۹	۲٫۸۵	۴۵	۰٫۴۵۵	۰٫۳۴۱
		PMR	۱٫۹۴	۱٫۹۴	۳٫۰۵	۳٫۰۰	۹۲	۰٫۴۲۰	۰٫۳۱۶
	t(۳)	PLS	۳٫۰۹	۱٫۷۲	۲٫۹۷	۲٫۵۴	۲۲	۱٫۰۸۶	۰٫۴۲۴
		PQR	۲٫۲۸	۱٫۷۳	۳٫۲۹	۲٫۷۹	۳۹	۰٫۶۸۳	۰٫۴۱۹
		PMR	۱٫۶۴	۱٫۶۳	۳٫۳۶	۲٫۹۹	۶۳	۰٫۴۷۷	۰٫۳۷۱
	MN	PLS	۳٫۳۳	۱٫۷۳	۲٫۹۲	۲٫۴۷	۱۷	۱٫۰۹۰	۰٫۴۹۳
		PQR	۲٫۲۶	۱٫۷۸	۳٫۲۹	۲٫۸۴	۴۸	۰٫۶۴۲	۰٫۳۶۴
		PMR	۱٫۸۳	۱٫۸۲	۳٫۱۷	۲٫۹۹	۸۱	۰٫۴۵۸	۰٫۳۳۲
۴۰۰	$\mathcal{N}(۰,۱)$	PLS	۱٫۹۳	۱٫۹۲	۳٫۰۷	۳٫۰۰	۹۱	۰٫۲۴۱	۰٫۲۲۲
		PQR	۱٫۹۸	۱٫۹۶	۳٫۱۶	۲٫۹۷	۸۲	۰٫۲۹۶	۰٫۲۴۶
		PMR	۲٫۰۰	۲٫۰۰	۳٫۲۵	۳٫۰۰	۹۴	۰٫۲۵۹	۰٫۲۰۱
	t(۳)	PLS	۲٫۳۳	۱٫۸۶	۲٫۹۲	۲٫۷۹	۶۰	۰٫۵۱۰	۰٫۳۲۳
		PQR	۱٫۹۷	۱٫۹۱	۳٫۱۴	۲٫۹۶	۷۸	۰٫۳۸۷	۰٫۳۰۸
		PMR	۲٫۰۰	۱٫۹۸	۳٫۳۵	۲٫۹۷	۸۷	۰٫۳۱۳	۰٫۲۵۴
	MN	PLS	۲٫۴۴	۱٫۸۹	۲٫۸۷	۲٫۷۸	۵۵	۰٫۸۲۴	۰٫۴۰۷
		PQR	۲٫۰۱	۱٫۹۳	۳٫۱۰	۲٫۹۶	۸۲	۰٫۳۷۷	۰٫۳۱۲
		PMR	۱٫۹۸	۱٫۹۸	۳٫۱۸	۲٫۹۹	۹۱	۰٫۳۰۱	۰٫۲۴۵

*اعداد داخل پرانتز، مقدار درست آن معیار در مدل واقعی است.

هرچند که برای خطاهای نرمال، دقت برآورد رگرسیون کمترین توان‌های دوم جریمه‌شده اندکی بیشتر از دقت رگرسیون نمای جریمه‌شده است، اما مقادیر $\text{NVT} = ۱٫۹۴,۲$، $\text{NCT} = ۳,۳$ و $\text{CF} = ۰٫۹۲, ۰٫۹۴$ برای $n = ۲۰۰, ۴۰۰$ نشان می‌دهد که رگرسیون نمای جریمه شده، در مقایسه با دو روش رقیب، بهترین روش برای تشخیص ساختار مدل است.

مثال ۲.۵.۴. در این مثال، روش انتخاب متغیر معرفی‌شده را برای یک مدل ضریب متغیر خطی–جزئی با بعد بالاتر به‌کار می‌بریم. بدین منظور، مدل زیر را با ابعاد $p = ۳۰, ۵۰$ در نظر بگیرید

$$Y_i = \sum_{j=۱}^{p} \alpha_j(U_i) X_{ij} + \varepsilon_i, \qquad i = ۱, \ldots, n \tag{۱۶.۴}$$

که $\alpha_۱(u) = ۳\cos(\pi(۶u - ۵)/۳)$، $\alpha_۲(u) = -۲u(۲ - ۳u)^۳$، $\alpha_۳(u) = u^۲ \exp(۱ + u)$،

جدول ۲.۴: درصد انتخاب درست ضرایب توسط رگرسیون نما در مثال ۲.۵.۴.

p	خطا	α_1	α_2	α_3	α_4	α_5	α_6	CF%	RMSE	
									جریمه‌شده	پیشگو
۳۰	$\mathcal{N}(0,1)$	۱۰۰	۹۸	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۹۸	۰٫۳۴۲۶	۰٫۳۳۶۷
	t(۳)	۱۰۰	۹۰	۱۰۰	۱۰۰	۱۰۰	۹۵	۸۹	۰٫۴۲۸۲	۰٫۳۸۹۲
	MN	۱۰۰	۹۱	۱۰۰	۱۰۰	۱۰۰	۹۷	۸۸	۰٫۴۳۴۱	۰٫۳۵۸۳
۵۰	$\mathcal{N}(0,1)$	۱۰۰	۹۶	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۹۶	۰٫۳۶۵۲	۰٫۳۳۶۱
	t(۳)	۱۰۰	۸۸	۱۰۰	۱۰۰	۱۰۰	۹۵	۸۵	۰٫۴۴۶۷	۰٫۳۹۳۵
	MN	۱۰۰	۸۸	۱۰۰	۱۰۰	۱۰۰	۹۴	۸۷	۰٫۴۱۳۰	۰٫۳۶۶۳

$\alpha_4(u) = 1.5$، $\alpha_5(u) = -1$، $\alpha_6(u) = 0.5$ و برای $7 \le j \le p$، $\alpha_j(u) = 0$. متغیرهای توضیحی همانند مثال ۱.۵.۴ تولید می‌شوند. حجم نمونه تصادفی را $n = 400$ در نظر گرفته و شبیه‌سازی را ۲۰۰ بار تکرار می‌کنیم. در این مثال، بعد ضرایب اسپلاین و بعد ماتریس طرح، به ترتیب، برابر با $p(K_n + q)$ و $p + p(K_n + q + 1)$ می‌باشند. درصد انتخاب درست هریک از ضرایب متغیر و ثابت غیر صفر براساس ۲۰۰ تکرار در جدول ۲.۴ گزارش شده‌است. به‌عنوان مثال، عدد ۹۸ متناظر با α_1 بدین معنی است که ضریب $\alpha_1(u)$ در ۹۸٪ اوقات به عنوان ضریب متغیر انتخاب می‌شود، یا عدد ۱۰۰ متناظر با α_6 بدین معنی است که در ۱۰۰٪ اوقات ضریب $\alpha_6(u)$ به عنوان ثابت غیر صفر انتخاب می‌شود. همچنین مقدار RMSE و CF متناظر با هر روش را نیز در جدول ۲.۴ ارائه داده‌ایم.

نتایج شبیه‌سازی نشان می‌دهد که حتی برای مدل‌های ضریب متغیر با ابعاد بزرگتر، رگرسیون نمای جریمه‌شده با احتمال بالا می‌تواند ضرایب متغیر و ثابت غیر صفر را به درستی شناسایی کند.

۶.۴ تحلیل داده‌های قیمت مسکن

در این بخش، روش معرفی‌شده را برای تحلیل داده‌های مسکن بوستون[18]، که توسط فن و هوانگ (۲۰۰۵) تحلیل شده‌است، به‌کار می‌بریم. این مجموعه داده در بسته "mlbench" نرم‌افزار R قابل دسترسی است و شامل ۱۴ متغیر زیر است:

- MEDV: میانه قیمت خانه
- LSTAT: درصد وضعیت پایین‌تر جامعه
- CRIM: نرخ سرانه جرم
- ZN: نسبت سطح ناحیه‌بندی شده با قطعات بزرگ

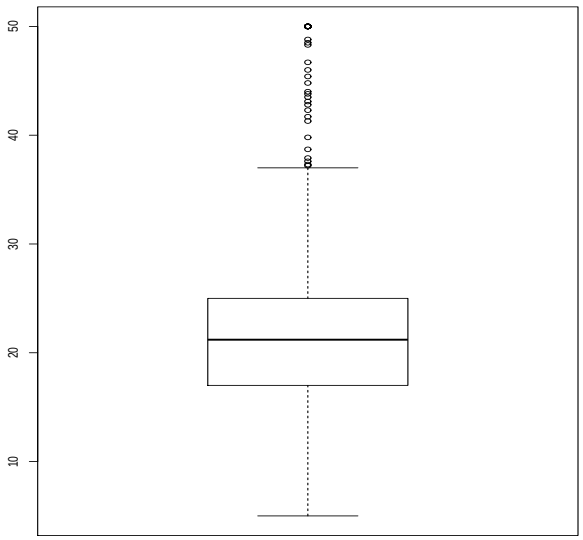
[18] Boston housing data

- INDUS: نسبت زمین‌های غیرتجاری
- CHAS: متغیر ساختگی رودخانه چارلز: اگر رودخانه مرزی است ۱، در غیر این‌صورت ۰
- NOX: غلظت اکسید نیتروژن
- RM: متوسط تعداد اتاق در هر خانه
- AGE: نسبت واحدهای ساخته‌شده قبل از سال ۱۹۴۰، که توسط مالک استفاده می‌شود
- DIS: میانگین وزنی فاصله تا پنج مرکز اشتغال بوستون
- RAD: شاخص دسترسی به بزرگ‌راه‌های محوری
- TAX: نرخ مالیاتی
- PTRATIO: نسبت دانش آموز به معلم
- Black: نسبت جمعیت سیاه‌پوست.

این مجموعه داده شامل ۵۰۶ مشاهده است. هدف از تحلیل این داده‌ها، بررسی ارتباط بین متغیر پاسخ MEDV و سایر متغیرها است. پیرو پیشنهاد فن و هوانگ (۲۰۰۵)، $U = \sqrt{LSTAT}$ را به عنوان متغیر شاخص در نظر می‌گیریم. این انتخاب متغیر U این امکان را به ما می‌دهد که برای درصدهای مختلف از وضعیت درآمد پایین جامعه، مدل‌های خطی مختلفی را برازش کنیم. علاوه بر این اثر سایر متغیرهای توضیحی با متغیر U در نظر گرفته می‌شود. مدل زیر را در نظر بگیرید

$$\begin{aligned} MEDV = {} & \alpha_0(U) + \alpha_1(U)CRIM + \alpha_2(U)ZN + \alpha_3(U)INDUS + \alpha_4(U)CHAS \\ & + \alpha_5(U)NOX + \alpha_6(U)RM + \alpha_7(U)AGE + \alpha_8(U)DIS + \alpha_9(U)RAD \\ & + \alpha_{10}(U)TAX + \alpha_{11}(U)PTRATIO + \alpha_{12}(U)Black + \epsilon. \end{aligned} \tag{۱۷.۴}$$

نمودار جعبه‌ای متغیر پاسخ MEDV در شکل ۳.۴ رسم شده‌است. این نمودار، وجود داده‌های پرت را در متغیر پاسخ نشان می‌دهد. از این رو، روش ارائه‌شده، به عنوان یک روش انتخاب متغیر نیرومند، برای تحلیل این داده‌ها به کار می‌رود. برای برازش مدل ضریب متغیر خطی–جزئی با روش پیشنهادی، از توابع اسپلاین مکعبی و ۴ گره با فواصل یکسان استفاده می‌شود. نتایج انتخاب متغیر و تشخیص ساختار در جدول ۳.۴ آورده شده‌است. جدول ۳.۴ نشان می‌دهد که متغیرهای INTERCEPT، CRIM، CHAS، NOX و RM دارای اثر متغیر بر پاسخ (نماد V در جدول) و متغیرهای DIS، RAD، PTRATIO، Black دارای اثر ثابت غیر صفر می‌باشند. همچنین متغیرهای ZN، INDUS، AGE و TAX به‌عنوان متغیرهای بی‌اثر بر پاسخ (نماد ۰ در جدول) انتخاب شده‌اند. به منظور مقایسه، نتایج انتخاب متغیر و تشخیص ساختار با رگرسیون میانگین جریمه‌شده (PLS) و رگرسیون چندکی جریمه‌شده (PQR) نیز در جدول ۳.۴ گزارش شده‌است.



شکل ۲.۴: نمودار جعبه‌ای متغیر پاسخ MDEV.

همانطور که در بخش ۲.۴ بیان شد، پس از تقریب ضرایب با توابع B-اسپلاین، مدل ضریب متغیر به مدل خطی تبدیل می‌شود. لذا برای ارزیابی عملکرد روش پیشنهادی از ضریب تعیین تعدیل شده استفاده می‌کنیم. این ضریب به‌صورت $R^{۲}_{adj} = ۱ - [(۱ - R^{۲})(n - ۱)/(n - k - ۱)]$ تعریف می‌شود که k تعداد متغیرهای توضیحی موجود در مدل و $R^{۲} = ۱ - RSS/\sum_{i=۱}^{n}(y_i - \bar{y})^{۲}$ که RSS مجموع توان‌های دوم باقیمانده‌هاست. در جدول ۳.۴، ضریب تعیین تعدیل شده، $R^{۲}_{adj}$، و انحراف معیار باقیمانده‌های متناظر با هر روش، $\hat{\sigma}$، گزارش شده‌است. با توجه به مقادیر $R^{۲}_{adj}$ و $\hat{\sigma}$، مدل برازش‌شده با استفاده از رگرسیون نما می‌تواند ۸۳٫۶۲٪ از تغییرات کل قیمت مسکن را تبیین کند و انحراف استاندارد باقیمانده‌ها برابر با $\hat{\sigma} = ۳٫۸۰۷$ است. مقدار $R^{۲}_{adj}$ متناظر با رگرسیون نما بزرگتر از مقدار $R^{۲}_{adj}$ متناظر با دو روش دیگر است، یعنی مدل حاصل از رگرسیون نما تغییرات بیشتری از قیمت مسکن را تبیین می‌کند. حال می‌خواهیم بدانیم آیا متغیرهای انتخاب شده با ضرایب متغیر واقعاً با متغیر پاسخ مرتبط هستند. بدین منظور، برآوردهای ضرایب متغیر برازش شده و فواصل اطمینان نقطه به نقطه ۹۵٪ آنها، که با روش بوت‌استرپ به‌دست آمده، در شکل ۳.۴ نشان داده‌شده‌است. به وضوح، همه نمودارها نشان می‌دهند که صفر بودن این ضرایب بسیار نامحتمل است، زیرا برای هریک از این متغیرها، خط $\alpha(u) = ۰$ کاملا در داخل فاصله اطمینان قرار نمی‌گیرد. با توجه به شکل ۳.۴، ملاحظه می‌شود هنگامی که $\sqrt{LSTAT} > ۰٫۳$، ضریب متغیر CRIM بسیار نزدیک به صفر است، یعنی تاثیر میزان جرم و جنایت بر قیمت مسکن، وابسته به درصد وضعیت درآمد پایین جامعه است. این تعبیر، مانند تعبیر ضرایب رگرسیون چند متغیره، با فرض ثابت بودن سایر متغیرهای توضیحی به‌دست می‌آید.

لازم به ذکر است، برای محاسبه فواصل اطمینان نقطه به نقطه هریک از ضرایب متغیر، ابتدا با استفاده از B نمونه بوت‌استرپ مستقل، B برآورد $\tilde{\alpha}_j(u)$ به‌دست می‌آوریم. حال یک فاصله اطمینان

جدول ۳.۴: نتایج انتخاب متغیر و برآورد ضرایب ثابت غیر صفر در داده‌های مسکن بوستون. نمادهای ◦ و V، به ترتیب، بیانگر اثر صفر و اثر متغیر است. مقادیر سطر آخر جدول، ضریب تعیین تعدیل شده و انحراف معیار باقیمانده‌ها را نشان می‌دهند.

PMR	PQR	PLS	متغیرهای توضیحی
V	V	V	INTERCEPT
V	◦	◦	CRIM
◦	◦	V	ZN
◦	−۰٫۰۳۹	◦	INDUS
V	V	V	CHAS
V	−۱۰٫۳۹	V	NOX
V	V	V	RM
◦	−۰٫۳۰۷	−۰٫۲۱۲	AGE
−۰٫۵۶۸	−۱٫۱۰۸	−۰٫۸۵۴	DIS
۰٫۲۴۵	V	V	RAD
◦	◦	V	TAX
−۰٫۶۷۸	−۰٫۷۸۴	−۱٫۰۳۲	PTRATIO
۰٫۰۱۹	V	V	Black
(٪۸۳٫۶۲, ۳٫۸۰۷)	(٪۸۱٫۳۴, ۴٫۱۱۲)	(٪۷۹٫۴۴, ۴٫۳۵۷)	$(R^2_{adj}, \hat{\sigma})$

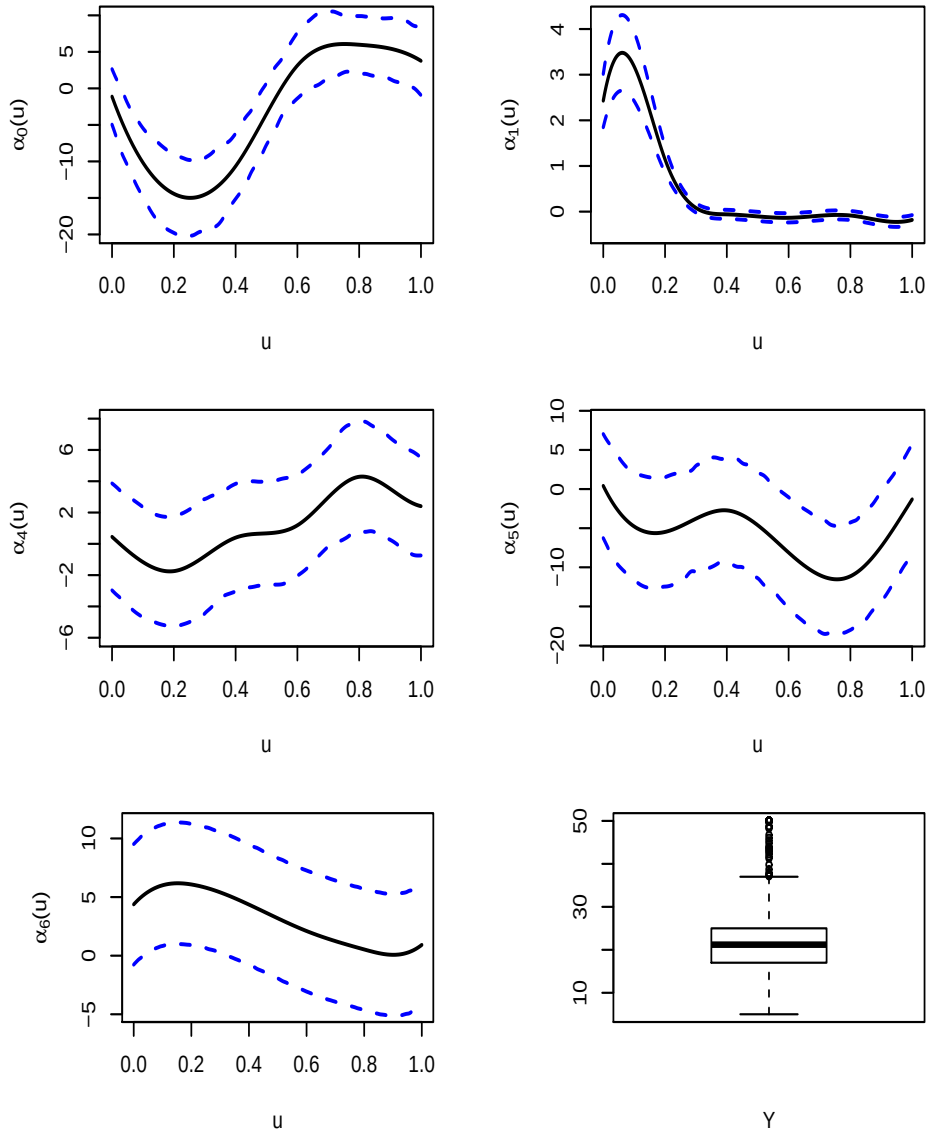
$1-\alpha$ برای $\alpha_j(u)$ براساس صدک‌های بوت‌استرپ به‌صورت

$$\left(L_{j,\alpha/2}(u), U_{j,\alpha/2}(u)\right), \qquad j \in \hat{\mathcal{I}_V}$$

به‌دست می‌آید که $L_{j,\alpha/2}(u)$ و $U_{j,\alpha/2}(u)$، به ترتیب، صدک‌های $100 \times \alpha/2$ و $100 \times (1-\alpha/2)$ برآوردهای بوت‌استرپ $\tilde{\alpha}_j(u)$ می‌باشند. علاوه‌براین، براساس تقریب نرمال، فاصله

$$\hat{\alpha}_j(u) \pm Z_{1-\alpha/2}\hat{\sigma}_{j,B}(u), \qquad j \in \hat{\mathcal{I}_V}$$

یک فاصله اطمینان در سطح $1-\alpha$ برای $\alpha_j(u)$ است که $\hat{\sigma}_{j,B}(u)$ خطای استاندارد نمونه‌ای برآوردهای $\tilde{\alpha}_j(u)$ و $Z_{(1-\alpha/2)}$، صدک $100 \times (1-\alpha/2)$ توزیع نرمال استاندارد است. اگر اریبی $\hat{\alpha}_j(u)$ نسبت به واریانس آن بطور مجانبی قابل اغماض باشد، آنگاه $\hat{\alpha}_j(u) \pm Z_{1-\alpha/2}\hat{\sigma}_{j,B}(u)$ یک فاصله اطمینان مجانبی در سطح $1-\alpha$ برای $\alpha_j(u)$ است. برای اطلاعات بیشتر می‌توان به هوانگ و همکاران (۲۰۰۲) و وانگ و همکاران (۲۰۰۸) مراجعه کرد.



شکل ۳.۴: برآوردهای ضرایب متغیر مدل (۱۷.۴) و فواصل اطمینان نقطه‌ای ۹۵٪ این ضرایب با استفاده از رگرسیون نمای جریمه‌شده (PMR).

بحث و نتیجه‌گیری

در این فصل، با به‌کارگیری رگرسیون نما، یک روش نیرومند برای برآورد ضرایب در مدل ضریب متغیر ارائه شد. این روش قادر است علاوه بر برآورد ضرایب، اثرات صفر، ثابت غیر صفر و متغیر را نیز بطور همزمان تشخیص دهد. کارایی روش ارائه‌شده با مثال شبیه‌سازی و تحلیل مجموعه داده واقعی مورد بررسی قرار گرفت. نتایج مطالعات عددی عملکرد رضایت‌بخش روش پیشنهادی را در حضور داده‌های پرت و خطاهای دم‌سنگین نشان می‌دهند.

اگرچه در این فصل مدل ضریب متغیر مورد بررسی قرار گرفته است، اما می‌توان روش ارائه‌شده را به مدل‌های دیگر مانند مدل ضریب متغیر تعمیم‌یافته، مدل تک‌ـشاخص و مدل تک‌ـشاخص ضریب متغیر تعمیم داد.

این روش پیشنهادی، مشابه سایر روش‌های انتخاب متغیر مبتنی بر جریمه، ممکن است برای داده‌های با بعد بسیار بالا دارای عملکرد مناسبی نباشد. برای رفع این چالش، می‌توان ابتدا با استفاده از روش‌های غربالگری بعد مسئله را از مقیاس بسیار بزرگ p به یک بعد کوچکتر از n کاهش داده و سپس از روش نیرومند ارائه شده برای انتخاب متغیر و تشخیص ساختار استفاده کرد. تاکنون چندین روش غربالگری برای کاهش بعد در مدل‌های ضریب متغیر معرفی شده است. به عنوان مثال، لیو و همکاران (۲۰۱۴) یک روش غربالگری برای مدل‌های ضریب متغیر براساس همبستگی شرطی (CC-SIS) معرفی کردند. همچنین، فن و همکاران (۲۰۱۴) و سانگ و همکاران (۲۰۱۴) روش‌های غربالگری دیگری برای این مدل‌ها ارائه دادند. علاوه‌براین تاکنون چندین روش آزاد‌ـمدل برای غربالگری معرفی شده‌است که در فصل ۲ به معرفی برخی از این روش‌ها پرداختیم. این روش‌ها می‌توانند برای غربالگری در مدل‌های ضریب متغیر مورد استفاده قرار گیرند.

پیوست آ

اثبات قضایا

آ.۱ قضایای فصل ۳

اثبات قضیه ۱.۲.۳ در اثبات این قضیه، c_1 و c_2 مقادیر ثابت مثبت هستند که ممکن است، در جاهای مختلف، دارای مقادیر متفاوت باشند.

ابتدا فرض کنید $F(y)$ معلوم باشد. بنابراین، $\widehat{dcov}^{*2}\{X_j, F(Y)\} = \hat{S}^*_{j,1} + \hat{S}^*_{j,2} - 2\hat{S}^*_{j,3}$ که در آن

$$\hat{S}^*_{j,1} = \frac{1}{n^2}\sum_{i=1}^n\sum_{k=1}^n |X_{ij} - X_{kj}||F(Y_i) - F(Y_k)|$$
$$\hat{S}^*_{j,2} = \frac{1}{n^2}\sum_{i=1}^n\sum_{k=1}^n |X_{ij} - X_{kj}|\frac{1}{n^2}\sum_{i=1}^n\sum_{k=1}^n |F(Y_i) - F(Y_k)|$$
$$\hat{S}^*_{j,3} = \frac{1}{n^3}\sum_{i=1}^n\sum_{k=1}^n\sum_{l=1}^n |X_{ij} - X_{lj}||F(Y_k) - F(Y_l)|.$$

تعریف کنید

$$\widehat{\omega}^*_k = \widehat{dcorr}^{*2}\{X_k, F(Y)\}.$$

قضیه ۱ لی، ژانگ و ژو (۲۰۱۲) بیان می‌کند که برای هر $0 < \gamma < 1/2 - \kappa$، ثابت‌های مثبت c_1 و c_2 وجود دارند به‌طوری‌که

$$P(|\widehat{\omega}^*_j - \omega_j| \geq cn^{-\kappa}) \leq O\Big(\Big[\exp\big\{-c_1 n^{1-2(\kappa+\gamma)}\big\} + n\exp(-c_2 n^\gamma)\Big]\Big). \qquad (۱.آ)$$

بنابراین برای اثبات قضیه، کافی است نشان دهیم که برای مقادیر بزرگ n، اختلاف بین $\widehat{\omega}^*_j$ و $\widehat{\omega}_j$ قابل اغماض است، که معادل با بررسی اختلاف بین $\hat{S}^*_{j,m}$ و $\hat{S}_{j,m}$ به ازای $m = 1, 2, 3$ است. در اینجا

قضیه را تنها برای $m = ۱$ ثابت می‌کنیم، زیرا برای $m = ۲, ۳$ نیز به‌طور مشابه ثابت می‌شود. به خاطر بیاورید که

$$\hat{S}^*_{j,۱} = \frac{۱}{n^۲} \sum_{i=۱}^{n} \sum_{k=۱}^{n} |X_{ij} - X_{kj}||F(Y_i) - F(Y_k)|$$

و

$$\hat{S}_{j,۱} = \frac{۱}{n^۲} \sum_{i=۱}^{n} \sum_{k=۱}^{n} |X_{ij} - X_{kj}||F_n(Y_i) - F_n(Y_k)|.$$

بنابراین

$$\begin{aligned}
&P\Big(|\hat{S}^*_{j,۱} - \hat{S}_{j,۱}| \geq \varepsilon\Big) \\
&= P\Big(\frac{۱}{n^۲} \sum_{i=۱}^{n} \sum_{k=۱}^{n} |X_{ij} - X_{kj}| \, \big| \, |F(Y_i) - F(Y_k)| - |F_n(Y_i) - F_n(Y_k)| \, \big| \geq \varepsilon\Big) \\
&\leq P\big((A_n B_n)^{\frac{۱}{۲}} \geq \varepsilon\big) \\
&\leq P\Big((A_n B_n)^{\frac{۱}{۲}} \geq \varepsilon, |X_j| \leq M\Big) + P\Big((A_n B_n)^{\frac{۱}{۲}} \geq \varepsilon, |X_j| > M\Big) \\
&=: I_۱ + I_۲
\end{aligned}$$

که M یک عدد مثبت بوده و بعدا مقدار آن مشخص می‌شود، $A_n = \frac{۱}{n^۲} \sum_{i=۱}^{n} \sum_{k=۱}^{n} (X_{ij} - X_{kj})^۲$ و $B_n = \frac{۱}{n^۲} \sum_{i=۱}^{n} \sum_{k=۱}^{n} \{|\, F(Y_i) - F(Y_k) \,| - |\, F_n(Y_i) - F_n(Y_k) \,|\}^۲$. حال با استفاده از نامساوی $\big|\, |x| - |y| \,\big| \leq |x - y| \leq |x| + |y|$، داریم

$$\begin{aligned}
&\big|\, |F(Y_i) - F(Y_k)| - |F_n(Y_i) - F_n(Y_k)| \,\big| \\
&\leq |F_n(Y_i) - F(Y_i)| + |F_n(Y_k) - F(Y_k)| \leq ۲ \max_{۱ \leq i \leq n} |F_n(Y_i) - F(Y_i)| .
\end{aligned}$$

علاوه براین، چون $A_n \leq \frac{۱}{n^۲} \sum_{i=۱}^{n} \sum_{k=۱}^{n} ۲(X_{ij}^۲ + X_{kj}^۲) \leq ۴M^۲$، می‌توان نتیجه گرفت

$$\begin{aligned}
I_۱ &\leq P\Big(۲Mn^{-۱}\Big\{\sum_{i=۱}^{n} \sum_{k=۱}^{n} \big(\, |F(Y_i) - F(Y_k)| - |F_n(Y_i) - F_n(Y_k)| \,\big)^۲\Big\}^{\frac{۱}{۲}} \geq \varepsilon\Big) \\
&\leq P\Big(۴M \max_{۱ \leq i \leq n} |F_n(Y_i) - F(Y_i)| \geq \varepsilon\Big) \leq P\Big(\max_{y \in \mathbb{R}} |F_n(Y_i) - F(Y_i)| \geq \varepsilon/۴M\Big) \\
&\leq ۲ \exp\{-۲n(\varepsilon/۴M)^۲\} = ۲ \exp(-n\varepsilon^۲/۸M^۲) \qquad (۲.آ)
\end{aligned}$$

که نامساوی آخر از نامساوی دوُرتزکی‌ـ کیفرـ وُلفوویتز[1] نتیجه می‌شود (پیوست پ را ببینید).

از طرفی، به ازای هر $۰ < t \leq t_۰$، که $t_۰$ در شرط C۱ تعریف شده‌است، داریم

$$\begin{aligned}
I_۲ &\leq P\big(|X_j| > M\big) = P\big(\exp(t|X_j|) > \exp(tM)\big) \\
&\leq E\{\exp(t|X_j|)\} \exp(-tM) \leq C \exp(-tM) \qquad (۳.آ)
\end{aligned}$$

[1] Dvoretzky-Kiefer-Wolfowitz inequality

که C یک ثابت مثبت است. نامساوی دوم از نامساوی مارکف و نامساوی آخر از شرط C۱ نتیجه می‌شود.
پس با انتخاب $M = O(n^{\gamma})$ که $0 < \gamma < \frac{1}{2} - \kappa$، و نامساوی‌های (۲.آ) و (۳.آ) برای برخی مقادیر ثابت مثبت c_1 و c_2 داریم

$$P\Big(|\hat{S}^{*}_{j,1} - \hat{S}_{j,1}| \geq \varepsilon\Big) \leq 2\exp(-n\varepsilon^{2}/8M^{2}) + C\exp(-tM)$$
$$\leq 2\exp(-c_1\varepsilon^{2}n^{1-2\gamma}) + C\exp(-c_2 n^{\gamma}). \qquad (۴.آ)$$

بنابراین، نرخ همگرایی صورت $\widehat{\omega}^{*}_{j}$ بدست آمد. به‌طور مشابه، می توان نرخ همگرایی مخرج آن را نیز بدست آورد. در پایان، می‌توان نشان داد که نرخ همگرایی $\widehat{\omega}^{*}_{j}$ همان (۴.آ) است. بنابراین، با فرض $\varepsilon = \nu n^{-\kappa}$، داریم

$$P\Big(|\widehat{\omega}^{*}_{j} - \widehat{\omega}_{j}| \geq \nu n^{-\kappa}\Big) \leq O\Big(\exp\big\{-c_1 n^{1-2(\kappa+\gamma)}\big\} + \exp(-c_2 n^{\gamma})\Big). \qquad (۵.آ)$$

با ترکیب (۱.آ) و (۵.آ) نتیجه می‌شود

$$P\Big(|\widehat{\omega}_{j} - \omega_{j}| \geq 2\nu n^{-\kappa}\Big) \leq P\Big(|\widehat{\omega}^{*}_{j} - \widehat{\omega}_{j}| \geq \nu n^{-\kappa}\Big) + P\Big(|\widehat{\omega}_{j} - \omega_{j}| \geq \nu n^{-\kappa}\Big)$$
$$\leq O\Big(\exp\big\{-c_1 n^{1-2(\kappa+\gamma)}\big\} + n\exp(-c_2 n^{\gamma})\Big). \qquad (۶.آ)$$

بنابراین می‌توان نوشت

$$P\Big(\max_{1\leq j\leq p}|\widehat{\omega}_{j} - \omega_{j}| \geq 2\nu n^{-\kappa}\Big) = P\Big(\bigcup_{j=1}^{p}\Big\{|\widehat{\omega}_{j} - \omega_{j}| \geq 2\nu n^{-\kappa}\Big\}\Big)$$
$$\leq \sum_{j=1}^{p} P\Big(|\widehat{\omega}_{j} - \omega_{j}| \geq 2\nu n^{-\kappa}\Big)$$
$$\leq p\max_{1\leq j\leq p} P\Big(|\widehat{\omega}_{j} - \omega_{j}| \geq 2\nu n^{-\kappa}\Big)$$
$$\leq O\Big(p\Big[\exp\big\{-c_1 n^{1-2(\kappa+\gamma)}\big\} + n\exp(-c_2 n^{\gamma})\Big]\Big).$$

با انتخاب $c = 2\nu$ اثبات کامل است.

اثبات قضیه ۲.۲.۳ اگر $\mathbb{M} \subsetneq \hat{\mathbb{M}}$، پس به ازای برخی مقادیر $j \in \mathbb{M}$ داریم $\widehat{\omega}_{j} < cn^{-\kappa}$. بنابراین از شرط C۲ نتیجه می‌شود که برای برخی مقادیر $j \in \mathbb{M}$ داریم $|\hat{r}_j - r_j| > cn^{-\alpha}$. بنابراین

$$\{\mathbb{M} \subsetneq \hat{\mathbb{M}}\} \subseteq \{|\hat{r}_j - r_j| > cn^{-\alpha} : j \in \mathbb{M} \text{ برای برخی مقادیر}\}. \qquad (۷.آ)$$

از رابطه (۷.آ) نتیجه می‌شود $\{\max_{j\in\mathbb{M}}|\hat{r}_j - r_j| \leq cn^{-\alpha}\} \subseteq \{\mathbb{M} \subseteq \hat{\mathbb{M}}\}$. با استفاده از قضیه ۱.۲.۳ داریم

$$P(\mathbb{M} \subseteq \hat{\mathbb{M}}) \geq P\Big(\max_{j\in\mathbb{M}}|\widehat{\omega}_{j} - \omega_{j}| \leq cn^{-\kappa}\Big)$$
$$\geq 1 - O\Big(s_n\Big[\exp\big\{-c_1 n^{1-2(\kappa+\gamma)}\big\} + n\exp(-c_2 n^{\gamma})\Big]\Big).$$

اثبات قضیه کامل است.

آ.۲ قضایای فصل ۴

فرض کنید C یک ثابت مثبت باشد که در جاهای مختلف این بخش، ممکن است دارای مقدار مختلف باشد. لم‌های آ.۱.۲ و آ.۲.۲ در اثبات قضایای فصل ۴ استفاده می‌شود.

لم آ.۱.۲. فرض کنید شرایط C۱-C۸ برقرار است و $K_n = O(n^{\frac{۱}{۲r+۱}})$ باشد. اگر وقتی $n \to \infty$، $b_n \to ۰$، داشته باشیم، آنگاه

$$\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_۰ \|_۲ = O_p(n^{\frac{-r}{۲r+۱}} + a_n) \quad \text{(i)}$$

$$\| \hat{g}_j(u) - g_{۰j}(u) \|_۲ = O_p(n^{\frac{-r}{۲r+۱}} + a_n), \quad j = ۱, \ldots, p \quad \text{(ii)}$$

اثبات. فرض کنید $\delta_n = n^{\frac{-r}{۲r+۱}} + a_n$، $\boldsymbol{\eta} = \boldsymbol{\eta}_۰ + \delta_n \boldsymbol{u}_۱$، $\boldsymbol{\gamma} = \boldsymbol{\gamma}_۰ + \delta_n \boldsymbol{u}_۲$ و $\boldsymbol{u} = (\boldsymbol{u}_۱^T, \boldsymbol{u}_۲^T)^T$، که $\boldsymbol{u}_۱$ و $\boldsymbol{u}_۲$ دو بردار، به ترتیب، با اندازه‌های p و $p \times K$ هستند. ابتدا نشان می‌دهیم که برای هر $\varepsilon > ۰$، یک ثابت بزرگ C وجود دارد، به‌طوری که

$$P\Big(\sup_{\|u\|_۲ = C} L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma}) < L_\lambda(\boldsymbol{\eta}_۰, \boldsymbol{\gamma}_۰) \Big) \geq ۱ - \varepsilon. \tag{۸.آ}$$

با استفاده از بسط تیلور حول $\varepsilon_i + \mathbf{X}_i^T \mathbf{R}(U_i)$ داریم

$$\begin{aligned} K_n^{-۱}\{L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma}) - L_\lambda(\boldsymbol{\eta}_۰, \boldsymbol{\gamma}_۰)\} \leq & \frac{\delta_n}{K_n} \sum_{i=۱}^{n} \phi_h'(\varepsilon_i + \mathbf{X}_i^T \mathbf{R}(U_i))(\mathbf{X}_i^T \boldsymbol{u}_۱ + \boldsymbol{\Pi}_i^T \boldsymbol{u}_۲) \\ & + \frac{\delta_n^۲}{۲K_n} \sum_{i=۱}^{n} \phi_h''(\varepsilon_i + \mathbf{X}_i^T \mathbf{R}(U_i))(\mathbf{X}_i^T \boldsymbol{u}_۱ + \boldsymbol{\Pi}_i^T \boldsymbol{u}_۲)^۲ \\ & + \frac{\delta_n^۳}{۶K_n} \sum_{i=۱}^{n} \phi_h'''(\zeta_i))(\mathbf{X}_i^T \boldsymbol{u}_۱ + \boldsymbol{\Pi}_i^T \boldsymbol{u}_۲)^۳ \\ & - \frac{n}{K_n} \sum_{j \in \mathcal{S}} \{p_{\lambda_۱}(|\eta_j|) - p_{\lambda_۱}(|\eta_{۰j}|)\} \\ & - \frac{n}{K_n} \sum_{j \in \mathcal{I}_V} \{p_{\lambda_۲}(\| \boldsymbol{\gamma}_j \|_{\mathbf{R}}) - p_{\lambda_۲}(\| \boldsymbol{\gamma}_{۰j} \|_{\mathbf{R}})\} \\ & := \Delta_۱ + \Delta_۲ + \Delta_۳ + \Delta_۴ + \Delta_۵ \end{aligned}$$

که در آن $\mathcal{S} = \mathcal{I}_C \cup \{l : E\alpha_{۰l}(U) \neq ۰, l \in \mathcal{I}_V\}$ و ζ_i دارای مقداری بین $\varepsilon_i + \mathbf{X}_i^T \mathbf{R}(U_i)$ و $\varepsilon_i + \mathbf{X}_i^T \mathbf{R}(U_i) - \delta_n(\mathbf{X}_i^T \boldsymbol{u}_۱ + \boldsymbol{\Pi}_i^T \boldsymbol{u}_۲)$ است . همچنین، $\mathbf{R}(U_i) = (R_۱(U_i), ..., R_p(U_i))^T$ که مؤلفه jام آن به‌صورت $R_j(U_i) = g_{۰j}(U_i) - \mathbf{B}(U_i)^T \boldsymbol{\gamma}_{۰j}$ تعریف می‌شود. حال با استفاده از بسط تیلور حول ε_i، داریم

$$\begin{aligned} & \sum_{i=۱}^{n} \phi_h'(\varepsilon_i + \mathbf{X}_i^T \mathbf{R}(U_i))(\mathbf{X}_i^T \boldsymbol{u}_۱ + \boldsymbol{\Pi}_i^T \boldsymbol{u}_۲) \\ & = \sum_{i=۱}^{n} \{\phi_h'(\varepsilon_i) + \frac{۱}{۲}\phi_h''(\varepsilon_i)\mathbf{X}_i^T \mathbf{R}(U_i) + \frac{۱}{۶}\phi_h'''(\zeta_i^*)[\mathbf{X}_i^T \mathbf{R}(U_i)]^۲\}(\mathbf{X}_i^T \boldsymbol{u}_۱ + \boldsymbol{\Pi}_i^T \boldsymbol{u}_۲) \end{aligned}$$

که ζ_i^* بین ε_i و $\varepsilon_i + \mathbf{X}_i^T\mathbf{R}(U_i)$ قراردارد. از طرف دیگر،

$$\begin{aligned} \| \mathbf{X}_i^T\mathbf{R}(U_i)\mathbf{R}(U_i)^T\mathbf{X}_i \| &\leq \| \mathbf{X}_i^T\mathbf{X}_i \| \| \mathbf{R}(U_i) \|_\infty \\ &= \| \mathbf{X}_i^T\mathbf{X}_i \| O(K_n^{-r}). \end{aligned} \tag{۹.آ}$$

بنابراین نتیجه می‌شود

$$\begin{aligned} \sum_{i=1}^{n} \phi_h'(\varepsilon_i + \mathbf{X}_i^T\mathbf{R}(U_i))(\mathbf{X}_i^T\boldsymbol{u}_1 + \boldsymbol{\Pi}_i^T\boldsymbol{u}_2) &= \sum_{i=1}^{n} (\mathbf{X}_i^T\boldsymbol{u}_1 + \boldsymbol{\Pi}_i^T\boldsymbol{u}_2) \| \mathbf{X}_i^T\mathbf{X}_i \| O(K_n^{-r}) \\ &= O_p(nK_n^{-r} \| \boldsymbol{u} \|_2). \end{aligned}$$

در نتیجه، داریم $\Delta_1 = O_p(n\delta_n K_n^{-(r+1)} \| \boldsymbol{u} \|_2) = O_p(n\delta_n^2 K_n^{-1} \| \boldsymbol{u} \|_2)$.

به‌طور مشابه، می‌توان ثابت کرد که

$$\Delta_2 = E\big(F(\mathbf{X}, U, h)\big) O_p(nK_n^{-1}\delta_n^2 \| \boldsymbol{u} \|_2^2), \qquad \Delta_3 = O_p(nK_n^{-1}\delta_n^3 \| \boldsymbol{u} \|_2^3).$$

لذا با انتخاب یک ثابت C، که بقدر کافی بزرگ است، Δ_2 به‌طور یکنواخت Δ_1 را در $\| \boldsymbol{u} \|_2 = C$ مغلوب می‌کند. با توجه به شرط $a_n \to 0$، می‌توان نوشت $\delta_n \to 0$. بنابراین $\delta_n \| \boldsymbol{u} \|_2 \to 0$ که نتیجه می‌دهد $\Delta_3 = o_p(\Delta_2)$. پس Δ_3 به‌طور یکنواخت توسط Δ_2 در $\| \boldsymbol{u} \|_2 = C$ مغلوب می‌شود. همچنین با استفاده از بسط تیلور داریم

$$\begin{aligned} p_{\lambda_1}(|\eta_j|) - p_{\lambda_1}(|\eta_{0j}|) &= (|\eta_j| - |\eta_{0j}|)p_{\lambda_1}'(|\eta_{0j}|) + \frac{1}{2}(|\eta_j| - |\eta_{0j}|)^2 p_{\lambda_1}''(|\eta_{0j}|) \\ &\leq (|\eta_j| - |\eta_{0j}|) \max_{j\in\mathcal{S}}\{p_{\lambda_1}'(|\eta_{0j}|)\} + \frac{1}{2}(|\eta_j| - |\eta_{0j}|)^2 \max_{j\in\mathcal{S}}\{p_{\lambda_1}''(|\eta_{0j}|)\} \\ &= (|\eta_{0j} + \delta_n\boldsymbol{u}_1| - |\eta_{0j}|)a_n + \frac{1}{2}(|\eta_{0j} + \delta_n\boldsymbol{u}_1| - |\eta_{0j}|)^2 b_n \\ &\leq |\delta_n\boldsymbol{u}_1|a_n + \frac{1}{2}|\delta_n\boldsymbol{u}_1|^2 b_n \\ &= O(\delta_n a_n \| \boldsymbol{u} \|_2) + O(\delta_n^2 b_n \| \boldsymbol{u} \|_2^2). \end{aligned}$$

بنابراین نتیجه می‌شود

$$\Delta_4 = O_p(nK_n^{-1}\delta_n a_n \| \boldsymbol{u} \|_2 + nK_n^{-1}\delta_n^2 b_n \| \boldsymbol{u} \|_2^2).$$

سپس با توجه به شرط $b_n \to 0$، به سادگی می‌توان نشان داد که Δ_4 توسط Δ_2 به‌طور یکنواخت در $\| \boldsymbol{u} \|_2 = C$ مغلوب می‌شود.

به‌طور مشابه، برای $\Delta_۵$، داریم

$$\begin{aligned}
p_{\lambda_۲}(\| \boldsymbol{\gamma}_j \|_{\mathbf{R}}) - p_{\lambda_۲}(\| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}}) &= (\| \boldsymbol{\gamma}_j \|_{\mathbf{R}} - \| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}}) p'_{\lambda_۲}(\| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}}) + \\
&\quad \frac{۱}{۲}(\| \boldsymbol{\gamma}_j \|_{\mathbf{R}} - \| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}})^۲ p''_{\lambda_۲}(\| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}}) \\
&\leq (\| \boldsymbol{\gamma}_j \|_{\mathbf{R}} - \| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}}) \max_{j \in \mathcal{I}_V} \{p'_{\lambda_۲}(\| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}})\} + \\
&\quad \frac{۱}{۲}(\| \boldsymbol{\gamma}_j \|_{\mathbf{R}} - \| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}})^۲ \max_{j \in \mathcal{I}_V} \{p''_{\lambda_۲}(\| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}})\} \\
&= (\| \boldsymbol{\gamma}_{\circ j} + \delta_n \boldsymbol{u}_۲ \|_{\mathbf{R}} - \| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}}) a_n + \\
&\quad \frac{۱}{۲}(\| \boldsymbol{\gamma}_{\circ j} + \delta_n \boldsymbol{u}_۲ \|_{\mathbf{R}} - \| \boldsymbol{\gamma}_{\circ j} \|_{\mathbf{R}})^۲ b_n \\
&\leq \| \delta_n \boldsymbol{u}_۲ \|_{\mathbf{R}} a_n + \frac{۱}{۲} \| \delta_n \boldsymbol{u}_۲ \|_{\mathbf{R}}^۲ b_n \\
&= a_n \delta_n (\boldsymbol{u}_۲^T \mathbf{R} \boldsymbol{u}_۲)^{\frac{۱}{۲}} + \frac{۱}{۲} b_n \delta_n^۲ (\boldsymbol{u}_۲^T \mathbf{R} \boldsymbol{u}_۲) \\
&= O(a_n \delta_n \| \boldsymbol{u} \|_۲) + O(b_n \delta_n^۲ \| \boldsymbol{u} \|_۲^۲).
\end{aligned}$$

تساوی آخر از این واقعیت نتیجه می‌شود که $\| \mathbf{R} \| = O(۱)$. بنابراین داریم

$$\Delta_۵ = O_p(n K_n^{-۱} a_n \delta_n \| \boldsymbol{u} \|_۲ + n K_n^{-۱} b_n \delta_n^۲ \| \boldsymbol{u} \|_۲^۲).$$

حال با شرط $b_n \to ۰$، می‌توان نتیجه گرفت که $\Delta_۵$ نیز توسط $\Delta_۲$ مغلوب می‌شود.

با توجه به شرط C۶، می‌دانیم که $F(x, u, h) < ۰$. لذا با انتخاب یک ثابت C به‌قدر کافی بزرگ، نامساوی $K_n^{-۱}\{L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma}) - L_\lambda(\boldsymbol{\eta}_\circ, \boldsymbol{\gamma}_\circ)\} < ۰$ با احتمال نزدیک به یک برقرار است، که برقرار بودن (آ.۸) را نتیجه می‌دهد. در نتیجه، با احتمال نزدیک به یک، یک ماکزیمم موضعی در گوی $\{(\boldsymbol{\eta}_\circ^T, \boldsymbol{\gamma}_\circ^T)^T + \delta_n \boldsymbol{u} : \| \boldsymbol{u} \|_۲ \leq C\}$ وجود دارد، یعنی یک ماکزیمم‌کننده موضعی مانند $(\hat{\boldsymbol{\eta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ موجود است به‌طوری که

$$(\hat{\boldsymbol{\eta}}^T, \hat{\boldsymbol{\gamma}}^T)^T = (\boldsymbol{\eta}_\circ^T, \boldsymbol{\gamma}_\circ^T)^T + \delta_n \boldsymbol{u}. \tag{آ.۱۰}$$

از رابطه (آ.۱۰) نتیجه می‌شود

$$\| \frac{\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_\circ}{\delta_n} \|_۲ = \| \boldsymbol{u}_۱ \|_۲ < C, \qquad \| \frac{\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_\circ}{\delta_n} \|_۲ = \| \boldsymbol{u}_۲ \|_۲ < C. \tag{آ.۱۱}$$

اکنون با جایگذاری $\delta_n = n^{\frac{-r}{۲r+۱}} + a_n$ در (آ.۱۱) داریم

$$\| \hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_\circ \|_۲ = O_p(n^{\frac{-r}{۲r+۱}} + a_n), \qquad \| \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_\circ \|_۲ = O_p(n^{\frac{-r}{۲r+۱}} + a_n). \tag{آ.۱۲}$$

بنابراین اثبات (i) کامل است.

حال به اثبات (ii) می‌پردازیم. در ابتدا، توجه کنید که

$$\begin{aligned} \| \hat{g}_j(u) - g_{\circ j}(u) \|_2^2 &= \int_{[\circ,1]} [\hat{g}_j(u) - g_{\circ j}(u)]^2 du \\ &= \int_{[\circ,1]} [\mathbf{B}(u)^T \hat{\boldsymbol{\gamma}}_j - \mathbf{B}(u)^T \boldsymbol{\gamma}_{\circ j} + R_j(u)]^2 du \\ &\leq 2 \int_{[\circ,1]} [\mathbf{B}(u)^T \hat{\boldsymbol{\gamma}}_j - \mathbf{B}(u)^T \boldsymbol{\gamma}_{\circ j}]^2 du + 2 \int_{[\circ,1]} [R_j(u)]^2 du \\ &= 2(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{\circ j})^T \mathbf{R} (\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{\circ j}) + 2 \int_{[\circ,1]} [R_j(u)]^2 du \qquad (۱۳.آ) \end{aligned}$$

که $\mathbf{R} = \int_{[\circ,1]} \mathbf{B}(u)\mathbf{B}(u)^T du$. اکنون با در نظر گرفتن $\| \mathbf{R} \| = O(1)$ (ژائو و همکاران، ۲۰۱۴) و $\| R_j(u) \| = O(K_n^r) = O(n^{\frac{-r}{2r+1}})$ و با استفاده از رابطه (۱۲.آ)، داریم

$$(\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{\circ j})^T \mathbf{R} (\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{\circ j}) = O_p(n^{\frac{-2r}{2r+1}} + a_n^2), \qquad \int_{[\circ,1]} [R_j(u)]^2 du = O_p(n^{\frac{-2r}{2r+1}}). \quad (۱۴.آ)$$

بنابراین، با ترکیب روابط (۱۳.آ) و (۱۴.آ)، می‌توان (ii) را نتیجه گرفت. اثبات لم کامل است.

لم ۲.۲.آ. تحت شرایط بیان‌شده در قضیه ۲.۳.۴، داریم

(i) برای $j \in \{1, ..., p\} \backslash \mathcal{S}$، رابطه $\hat{\eta}_j = \circ$ با احتمال نزدیک به یک برقرار است.

(ii) برای $j \in \mathcal{I}_C \cup \mathcal{I}_Z$، رابطه $\hat{g}_j(u) \equiv \circ$ با احتمال نزدیک به یک برقرار است.

اثبات. ابتدا (i) را ثابت می‌کنیم. بنا به ویژگی‌های تابع جریمه SCAD، برای مقادیر بزرگ n، وقتی $\lambda^{max} \to \circ$، داریم $a_n = \circ$. پس با توجه به لم ۱.۲.آ، کافی است نشان دهیم، برای هر $\boldsymbol{\gamma}$ که در شرط $\| \boldsymbol{\gamma} - \boldsymbol{\gamma}_\circ \|_2 = O_p(n^{\frac{-r}{2r+1}})$ و η_j که در شرط $|\eta_j - \eta_{\circ j}| = O_p(n^{\frac{-r}{2r+1}}),\ j \in \mathcal{S}$ صدق کند، نامساوی‌های

$$\frac{\partial L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma})}{\partial \eta_j} < \circ \quad \text{for} \quad \circ < \eta_j < \varepsilon, \quad j \in \{1, ..., p\} \backslash \mathcal{S} \qquad (۱۵.آ)$$

و

$$\frac{\partial L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma})}{\partial \eta_j} > \circ \quad \text{for} \quad -\varepsilon < \eta_j < \circ, \quad j \in \{1, ..., p\} \backslash \mathcal{S} \qquad (۱۶.آ)$$

برای برخی مقادیر کوچک $\varepsilon = Cn^{-r/(2r+1)}$، با احتمال نزدیک به یک برقرار است. در نتیجه جواب ماکزیمم‌کننده $L_\lambda(\boldsymbol{\eta}, \boldsymbol{\gamma})$ در $\hat{\eta}_j = \circ$ اتفاق می‌افتد که $j \in \{1, ..., p\} \backslash \mathcal{S}$.

به طریق مشابه مباحث بیان شده در اثبات لم آ.۱.۲، داریم

$$\begin{aligned}\frac{\partial L_\lambda(\boldsymbol{\eta},\boldsymbol{\gamma})}{\partial \eta_j} &= \frac{\partial Q(\boldsymbol{\eta},\boldsymbol{\gamma})}{\partial \eta_j} - np'_{\lambda_1}(|\eta_j|)\mathrm{sgn}(\eta_j)\\ &= -\sum_{i=1}^{n} X_{ij}\phi'_h\big(Y_i - \mathbf{X}_i^T\boldsymbol{\eta} - \boldsymbol{\Pi}_i^T\boldsymbol{\gamma}\big) - np'_{\lambda_1}(|\eta_j|)\mathrm{sgn}(\eta_j)\\ &= -\sum_{i=1}^{n}\Big\{X_{ij}\phi'_h\big(\varepsilon_i + \mathbf{X}_i^T\mathbf{R}(U_i)\big) - X_{ij}\phi''_h\big(\varepsilon_i + \mathbf{X}_i^T\mathbf{R}(U_i)\big)[\boldsymbol{\Pi}_i^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}_\circ)\\ &+ \mathbf{X}_i^T(\boldsymbol{\eta}-\boldsymbol{\eta}_\circ)] + X_{ij}\phi'''_h\big(\zeta_i^*\big)[\boldsymbol{\Pi}_i^T(\boldsymbol{\gamma}-\boldsymbol{\gamma}_\circ) + \mathbf{X}_i^T(\boldsymbol{\eta}-\boldsymbol{\eta}_\circ)]^2\Big\} - np'_{\lambda_1}(|\eta_j|)\mathrm{sgn}(\eta_j)\\ &= n\lambda_1\Big\{\frac{-1}{\lambda_1}p'_{\lambda_1}(|\eta_j|)\mathrm{sgn}(\eta_j) + O_p(\lambda_1^{-1}n^{\frac{-r}{2r+1}})\Big\} \qquad (۱۷.آ)\end{aligned}$$

که ζ_i^* مقداری بین $Y_i - \mathbf{X}_i^T\boldsymbol{\eta} - \boldsymbol{\Pi}_i^T\boldsymbol{\gamma}$ و $\varepsilon_i + \mathbf{X}_i^T\mathbf{R}(U_i)$.

بنا به شرط C۸، $\lambda_1 n^{\frac{r}{2r+1}} > \lambda_{\min} n^{\frac{r}{2r+1}} \to \infty$ و $\liminf_{n\to\infty}\liminf_{\eta_j\to 0^+} p'_{\lambda_1}(\eta_j)/\lambda_1 > 0$، لذا علامت مشتق فوق، کاملاً با علامت η_j تعیین می‌شود. پس (آ.۱۵) و (آ.۱۶) برقرار است. بنابراین اثبات (i) کامل است.

با استفاده از روش به کار رفته در اثبات قسمت (i) این لم، برای $j \in \mathcal{I}_C \cup \mathcal{I}_Z$، با احتمال نزدیک به یک داریم $\| \hat{\gamma}_j \|_1 = 0$. سپس با در نظرگرفتن $\sup_u \| \mathbf{B}(u) \|_2 = O(1)$، قسمت (ii) به سادگی از $\hat{g}_j(u) = \mathbf{B}(u)^T\hat{\gamma}_j$ نتیجه می‌شود.

اثبات قضیه ۱.۳.۴ برای $j \in \mathcal{I}_V$، با تعریف $\hat{G}_j(u)$ و لم آ.۱.۲، داریم

$$\begin{aligned}\| \hat{\alpha}_j(u) - \alpha_{\circ j}(u) \|_2^2 &= \int_{[0,1]} [\hat{\alpha}_j(u) - \alpha_{\circ j}(u)]^2 du\\ &\le 2|\hat{\eta}_j - \eta_{\circ j}|^2 + 2 \| \hat{g}_j(u) - g_{\circ j}(u) \|_2^2\\ &= O_p(n^{\frac{-2r}{2r+1}} + a_n^2). \qquad (۱۸.آ)\end{aligned}$$

اثبات کامل است.

اثبات قضیه ۲.۳.۴ فرض کنید $\boldsymbol{\gamma}_{\mathcal{I}_V} = \big((\eta_j, \boldsymbol{\gamma}_j^T), j \in \mathcal{I}_V\big)^T$، $\boldsymbol{\gamma}_{\circ\mathcal{I}_V} = \big((\eta_{\circ j}, \boldsymbol{\gamma}_{\circ j}^T), j \in \mathcal{I}_V\big)^T$ و $\boldsymbol{\eta}_{\mathcal{I}_C} = \big(\eta_j, j \in \mathcal{I}_C\big)^T$. با استفاده از لم آ.۲.۲، با احتمال نزدیک به یک، $(\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}, \hat{\boldsymbol{\eta}}_{\mathcal{I}_C})$ ماکزیمم کننده تابع

$$\mathcal{L}(\boldsymbol{\eta},\boldsymbol{\gamma}) = \sum_{i=1}^{n}\phi_h\big(Y_i - \mathbf{X}_i^{cT}\boldsymbol{\eta}_{\mathcal{I}_C} - \boldsymbol{\Pi}_i^{vT}\boldsymbol{\gamma}_{\mathcal{I}_V}\big) - n\sum_{j\in\mathcal{S}} p_{\lambda_1}(|\eta_j|) - n\sum_{j\in\mathcal{I}_V} p_{\lambda_2}(\|\boldsymbol{\gamma}_j\|_{\mathbf{R}}) \qquad (۱۹.آ)$$

است که در آن $\mathbf{X}_i^c = (X_{ij}, j \in \mathcal{I}_C)^T$ و $\boldsymbol{\Pi}_i^v = \Big((X_{ij}, X_{ij}\mathbf{B}(U_i)^T), j \in \mathcal{I}_V\Big)^T$.

فرض کنید $\mathcal{L}_{۱}(\boldsymbol{\eta}_{\mathcal{I}_C},\boldsymbol{\gamma}_{\mathcal{I}_V})=\dfrac{\partial\mathcal{L}(\boldsymbol{\eta},\boldsymbol{\gamma})}{\partial\boldsymbol{\eta}_{\mathcal{I}_C}}$ و $\mathcal{L}_{۲}(\boldsymbol{\eta}_{\mathcal{I}_C},\boldsymbol{\gamma}_{\mathcal{I}_V})=\dfrac{\partial\mathcal{L}(\boldsymbol{\eta},\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}_{\mathcal{I}_V}}$. پس $\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}$ و $\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}$ باید در معادلات زیر صدق کنند

(۲۰.آ)
$$\frac{۱}{n}\mathcal{L}_{۱}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C},\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V})=\frac{۱}{n}\sum_{i=۱}^{n}\mathbf{X}_i^c\phi_h'\big(Y_i-\mathbf{X}_i^{cT}\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\Pi}_i^{vT}\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}\big)-p_{\lambda_{۱}}'(|\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}|)\,\mathbf{o}\,\mathrm{sgn}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C})=۰$$

و

$$\frac{۱}{n}\mathcal{L}_{۲}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C},\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V})=\frac{۱}{n}\sum_{i=۱}^{n}\boldsymbol{\Pi}_i^{v}\phi_h'\big(Y_i-\mathbf{X}_i^{cT}\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\Pi}_i^{vT}\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}\big)-\kappa=۰ \qquad (۲۱.آ)$$

که $p_{\lambda_{۱}}'(|\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}|)\,\mathbf{o}\,\mathrm{sgn}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C})$ یک زیر-بردار از $p_{\lambda_{۱}}'(|\hat{\boldsymbol{\eta}}|)\,\mathbf{o}\,\mathrm{sgn}(\hat{\boldsymbol{\eta}})$ است و j-امین مؤلفه از $p_{\lambda_{۱}}'(|\hat{\boldsymbol{\eta}}|)$ و $\mathrm{sgn}(\hat{\boldsymbol{\eta}})$، به ترتیب، به‌صورت $p_{\lambda_{۱}}'(|\hat{\eta}_j|)$ و $\mathrm{sgn}(\hat{\eta}_j)$ می‌باشند که $j=۱,\ldots,p$ و "o" نشان‌دهنده ضرب مؤلفه به مؤلفه است. همچنین κ یک بردار با طول $K\times|\mathcal{I}_V|$ است که j-امین زیربردار بلوکی آن به‌صورت $\mathbf{R}\dfrac{\hat{\boldsymbol{\gamma}}_j}{\|\hat{\boldsymbol{\gamma}}_j\|_{\mathbf{R}}}p_{\lambda_{۲}}'(\|\hat{\boldsymbol{\gamma}}_j\|_{\mathbf{R}})$ می‌باشد. با استفاده از بسط تیلور $p_{\lambda_{۱}}'(|\hat{\eta}_j|)$، می‌توان نوشت

$$p_{\lambda_{۱}}'(|\hat{\eta}_j|)=p_{\lambda_{۱}}'(|\eta_{۰j}|)+(p_{\lambda_{۱}}''(|\eta_{۰j}|)+o_p(۱))(\hat{\eta}_j-\eta_{۰j}),\quad j\in\mathcal{I}_C. \qquad (۲۲.آ)$$

توجه کنید که براساس یادآوری ۱ فن و لی (۲۰۰۱)، وقتی $\lambda_{\max}\to۰$، آنگاه $p_{\lambda_{۱}}'(|\eta_{۰j}|)=۰$. بنابراین با توجه به شرط $b_n\to۰$، با ترکیب (۲۰.آ) و (۲۲.آ) داریم

$$\begin{aligned}\frac{۱}{n}\sum_{i=۱}^{n}\mathbf{X}_i^c\Big\{&\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)\Big[\mathbf{X}_i^{vT}\mathbf{R}^v(U_i)-[\mathbf{X}_i^{cT}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\eta}_{۰\mathcal{I}_C})+\boldsymbol{\Pi}_i^{vT}(\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}-\boldsymbol{\gamma}_{۰\mathcal{I}_V})]\Big]\\&+\phi_h'''(\varepsilon_i^*)\Big[\mathbf{X}_i^{vT}\mathbf{R}^v(U_i)-[\mathbf{X}_i^{cT}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\eta}_{۰\mathcal{I}_C})+\boldsymbol{\Pi}_i^{vT}(\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}-\boldsymbol{\gamma}_{۰\mathcal{I}_V})]\Big]^{۲}\Big\}\\&+o_p(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\eta}_{۰\mathcal{I}_C})=۰\end{aligned} \qquad (۲۳.آ)$$

که $\mathbf{R}^v(U_i)=(R_j(U_i),j\in\mathcal{I}_V)^T$ و ε_i^* بین ε_i و $Y_i-\mathbf{X}_i^{cT}\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\Pi}_i^{vT}\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}$ قرار دارد.

مشابه معادله (۲۳.آ)، با استفاده از بسط تیلور $p_{\lambda_{۲}}'(\|\hat{\boldsymbol{\gamma}}_j\|_{\mathbf{R}})$ و رابطه (۲۱.آ) می‌توان نوشت

$$\begin{aligned}\frac{۱}{n}\sum_{i=۱}^{n}\boldsymbol{\Pi}_i^{v}\Big\{&\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)\Big[\mathbf{X}_i^{vT}\mathbf{R}^v(U_i)-[\mathbf{X}_i^{cT}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\eta}_{۰\mathcal{I}_C})+\boldsymbol{\Pi}_i^{vT}(\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}-\boldsymbol{\gamma}_{۰\mathcal{I}_V})]\Big]\\&+\phi_h'''(\varepsilon_i^{**})\Big[\mathbf{X}_i^{vT}\mathbf{R}^v(U_i)-[\mathbf{X}_i^{cT}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\eta}_{۰\mathcal{I}_C})+\boldsymbol{\Pi}_i^{vT}(\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}-\boldsymbol{\gamma}_{۰\mathcal{I}_V})]\Big]^{۲}\Big\}\\&+o_p(\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}-\boldsymbol{\gamma}_{۰\mathcal{I}_V})=۰\end{aligned} \qquad (۲۴.آ)$$

که ε_i^{**} بین ε_i و $Y_i-\mathbf{X}_i^{cT}\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\Pi}_i^{vT}\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}$ قرار دارد. فرض کنید $\boldsymbol{\Phi}_n=\frac{۱}{n}\sum_{i=۱}^{n}\phi_h''(\varepsilon_i)\boldsymbol{\Pi}_i^{v}\boldsymbol{\Pi}_i^{vT}$ و $\boldsymbol{\Psi}_n=\frac{۱}{n}\sum_{i=۱}^{n}\phi_h''(\varepsilon_i)\boldsymbol{\Pi}_i^{v}\mathbf{X}_i^{cT}$. پس با استفاده از (۲۴.آ)، داریم

$$\hat{\boldsymbol{\gamma}}_{\mathcal{I}_V}-\boldsymbol{\gamma}_{۰\mathcal{I}_V}=(\boldsymbol{\Phi}_n+o_p(۱))^{-۱}\{-\boldsymbol{\Psi}_n(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C}-\boldsymbol{\eta}_{۰\mathcal{I}_C})+\boldsymbol{\Lambda}_n\} \qquad (۲۵.آ)$$

که $\mathbf{\Lambda}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Pi}_i^v \left[\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\mathbf{X}_i^{vT}\mathbf{R}^v(U_i)\right]$. با جایگزینی آ۲۵ در (آ۲۳)، می‌توان نوشت

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \phi_h''(\varepsilon_i)\mathbf{X}_i^c\{\mathbf{X}_i^c - \mathbf{\Psi}_n^T\Phi(n)^{-1}\mathbf{\Pi}_i^v\}^T(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C} - \boldsymbol{\eta}_{\circ\mathcal{I}_C}) + o_p(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C} - \boldsymbol{\eta}_{\circ\mathcal{I}_C}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^c\Big\{\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\mathbf{X}_i^{cT}\mathbf{R}^v(U_i) - \phi_h''(\varepsilon_i)\mathbf{\Pi}_i^{vT}\Big[\frac{1}{n}\sum_{k=1}^{n}\phi_h'(\varepsilon_k)\mathbf{\Phi}_n^{-1}\mathbf{\Pi}_k^v\Big]\Big\} \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \phi_h''(\varepsilon_i)\mathbf{X}_i^c\mathbf{\Pi}_i^{vT}\Big[\frac{1}{n}\sum_{k=1}^{n}\mathbf{X}_k^{vT}\mathbf{R}^v(U_k)\Big]. \qquad (۲۶.آ)
\end{aligned}$$

علاوه بر این داریم

$$\begin{aligned}
&E\Big\{\frac{1}{n}\sum_{i=1}^{n} \phi_h''(\varepsilon_i)\mathbf{\Psi}_n^T\mathbf{\Phi}_n^{-1}\mathbf{\Pi}_i^v[\mathbf{X}_i^{cT} - \mathbf{\Pi}_i^{vT}\mathbf{\Phi}_n^{-1}\mathbf{\Psi}_n]\Big\} = 0 \\
&Var\Big\{\frac{1}{n}\sum_{i=1}^{n} \phi_h''(\varepsilon_i)\mathbf{\Psi}_n^T\mathbf{\Phi}_n^{-1}\mathbf{\Pi}_i^v[\mathbf{X}_i^{cT} - \mathbf{\Pi}_i^{vT}\mathbf{\Phi}_n^{-1}\mathbf{\Psi}_n]\Big\} = o_p\Big(\frac{1}{n}\Big).
\end{aligned}$$

حال با استفاده از (آ۲۶)، به سادگی می‌توان نشان داد

$$\Big[\frac{1}{n}\sum_{i=1}^{n} \phi_h''(\varepsilon_i)\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^T + o_p(1)\Big]\sqrt{n}(\hat{\boldsymbol{\eta}}_{\mathcal{I}_C} - \boldsymbol{\eta}_{\circ_{\mathcal{I}_C}}) = I_1 + I_2 \qquad (۲۷.آ)$$

که در آن

$$\begin{aligned}
I_1 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_h'(\varepsilon_i)\tilde{\mathbf{X}}_i \\
I_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_h''(\varepsilon_i)\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^{vT}\mathbf{R}^v(U_i)
\end{aligned}$$

و $\tilde{\mathbf{X}}_i = \mathbf{X}_i^c - \mathbf{\Psi}_n^T\mathbf{\Phi}_n^{-1}\mathbf{\Pi}_i^v$. بنا به قضیه حد مرکزی، داریم

$$I_1 \xrightarrow{D} N\big(0, G(h)\mathbf{\Sigma}_{\mathcal{I}_C}\big) \qquad (۲۸.آ)$$

که در آن

$$\mathbf{\Sigma}_{\mathcal{I}_C} = E(\mathbf{X}_i^c\mathbf{X}_i^{cT}) - E(\mathbf{X}_i^c\mathbf{\Pi}_i^{vT})\big\{E(\mathbf{\Pi}_i^v\mathbf{\Pi}_i^{vT})\big\}^{-1}E(\mathbf{\Pi}_i^v\mathbf{X}_i^{cT}). \qquad (۲۹.آ)$$

توجه کنید که در رابطه (آ۲۸) برای محاسبه کوواریانس I_1 از این واقعیات استفاده شده است که

$$\begin{aligned}
&\mathbf{\Phi}_n^{-1} \to E(\phi_h''(\varepsilon_i)\mathbf{\Pi}_i^v\mathbf{\Pi}_i^{vT})^{-1} \\
&\mathbf{\Psi}_n \to E(\phi_h''(\varepsilon_i)\mathbf{\Pi}_i^v\mathbf{X}_i^{cT}) \\
&\mathbf{\Psi}_n\mathbf{\Phi}_n^{-1} \to E[\phi_h''(\varepsilon_i)^{-1}(\mathbf{\Pi}_i^v\mathbf{\Pi}_i^{vT})^{-1}\phi_h''(\varepsilon_i)\mathbf{\Pi}_i^v\mathbf{X}_i^{cT}] = E(\mathbf{\Pi}_i^v\mathbf{\Pi}_i^{vT})^{-1}E(\mathbf{\Pi}_i^v\mathbf{X}_i^{cT}) \\
&\mathbf{\Psi}_n^T\mathbf{\Phi}_n^{-1}\mathbf{\Pi}_i^v\mathbf{\Pi}_i^{vT} = \Psi_n^T E(\mathbf{\Pi}_i^v\mathbf{\Pi}_i^{vT})^{-1}\mathbf{\Pi}_i^v\mathbf{\Pi}_i^{vT} = \mathbf{\Psi}_n^T.
\end{aligned}$$

با استفاده از روابط فوق نتیجه می‌شود

$$\begin{aligned}
Cov(I_۲) &= \frac{۱}{n}\sum_{i=۱}^{n} Cov(\phi_h'(\varepsilon_i)\tilde{\mathbf{X}}_i) \\
&= \frac{۱}{n}\sum_{i=۱}^{n} G(h)Cov(\tilde{\mathbf{X}}_i) \\
&= E[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^T] - E(\tilde{\mathbf{X}}_i)E(\tilde{\mathbf{X}}_i^T) \\
&= E[\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^T] \\
&= E[(\mathbf{X}_i^c - \boldsymbol{\Psi}_n^T\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Pi}_i^v)(\mathbf{X}_i^c - \boldsymbol{\Psi}_n^T\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Pi}_i^v)^T] \\
&= E[\mathbf{X}_i^c\mathbf{X}_i^{cT} - \mathbf{X}_i^c\boldsymbol{\Pi}_i^{vT}\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^T\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Pi}_i^v\mathbf{X}_i^{cT} \\
&\qquad + \boldsymbol{\Psi}_n^T\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Pi}_i^v\boldsymbol{\Pi}_i^{vT}\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Psi}_n] \\
&= E[\mathbf{X}_i^c\mathbf{X}_i^{cT} - \boldsymbol{\Psi}_n^T\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Psi}_n - \boldsymbol{\Psi}_n^T\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Psi}_n + \boldsymbol{\Psi}_n^T\boldsymbol{\Phi}_n^{-۱}\boldsymbol{\Psi}_n] \\
&= E[\mathbf{X}_i^c\mathbf{X}_i^{cT}] - E(\mathbf{X}_i^c\boldsymbol{\Pi}_i^{vT})E(\boldsymbol{\Pi}_i^v\boldsymbol{\Pi}_i^{vT})^{-۱}E(\boldsymbol{\Pi}_i^v\mathbf{X}_i^c).
\end{aligned}$$

با استفاده از تعریف $\mathbf{R}^v(U_i)$، می‌توان ثابت کرد $I_۲ = o_p(۱)$ (ژائو و همکاران، ۲۰۱۴). همچنین با استفاده از قانون اعداد بزرگ، داریم

$$\frac{۱}{n}\sum_{i=۱}^{n}\phi_h''(\varepsilon_i)\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^T \xrightarrow{p} E(\phi_h''(\varepsilon)\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T) = F(h)\boldsymbol{\Sigma}_{\mathcal{I}_C}. \tag{۳۰.آ}$$

از طرفی با استفاده از قضیه اسلاتسکی می‌دانیم که اگر $A_n \xrightarrow{P} A$ و $X_n \xrightarrow{D} X$، آنگاه $\frac{X_n}{A_n} \xrightarrow{D} \frac{X}{A}$. بنابراین با در نظر گرفتن روابط (۲۷.آ)ـ(۳۰.آ) و قضیه اسلاتسکی، اثبات کامل می‌شود.

پیوست ب

توابع اسپلاین و Bـ اسپلاین

مدل ناپارامتری زیر را در نظر بگیرید

$$E(Y|X) = f(X)$$

که در آن $f(\cdot)$ یک تابع هموار نامعلوم است و باید بر اساس یک نمونه تصادفی از (X, Y) برآورد شود. تقریب این تابع رگرسیونی ''هموارسازی'' نامیده می‌شود. بنابراین یک هموارساز، تابعی برای خلاصه‌سازی متغیر پاسخ Y به‌عنوان تابعی از متغیر توضیحی X است.

تاکنون روش‌های متعددی برای هموارسازی معرفی شده است. یکی از متداول‌ترین این روش‌ها، اسپلاین‌ها هستند. اسپلاین توسط شخصی به نام شوئنبرگ[1] در سال ۱۹۴۶ معرفی شد. وی برای نخستین بار قضیه‌ای برای اثبات وجود اسپلاین‌های درون‌یاب بیان کرد و سپس مطالعه چندجمله‌ای‌های تکه‌ای برای توابع اسپلاین آغاز شد. در ادامه به بیان جزئیات این روش می‌پردازیم.

ب.۱ اسپلاین

اسپلاین مرتبه یک

بازه $[a, b]$ را به‌صورت اجتماع دو بازه $[a, t) \cup [t, b]$ در نظر بگیرید. یک اسپلاین از مرتبه یک، تابعی است که در هر یک از بازه‌های مذکور، یک تابع خطی و در نقطه t نیز پیوسته باشد. این تابع را می‌توان

[1] Schoenberg

به‌صورت زیر نوشت

$$f(x)=\begin{cases}\beta_۰+\beta_۱x & x\le t\\ \beta'_۰+\beta'_۱x & x>t.\end{cases} \qquad (ب.۱)$$

با اعمال شرط پیوستگی تابع f در نقطه‌ی $x=t$ داریم

$$\beta_۰+\beta_۱t=\beta'_۰+\beta'_۱t$$
$$\beta'_۰=\beta_۰+\beta_۱t-\beta'_۱t.$$

با جایگذاری $\beta'_۰$ در ضابطه‌ی دوم f داریم

$$\beta'_۰+\beta'_۱x=\beta_۰+\beta_۱t-\beta'_۱t+\beta'_۱x=\beta_۰+\beta_۱t+\beta'_۱(x-t).$$

لذا

$$f(x)=\begin{cases}\beta_۰+\beta_۱x & x\le t\\ \beta_۰+\beta_۱t+\beta'_۱(x-t) & x>t\end{cases}$$

که می‌توان این تابع دو ضابطه‌ای را به‌صورت زیر نوشت

$$f(x)=\beta_۰+\beta_۱x+\beta_۲(x-t)_+ \qquad (ب.۲)$$

که $(x-t)_+=(x-t)I(x>t)$ و $\beta_۲=\beta'_۱-\beta_۱$ تفاوت بین شیب خطوط اول و دوم رابطه (ب.۱) را بیان می‌کند. در واقع می‌توان گفت که تابع f، ترکیبی خطی از توابع پایه‌ای ۱، x و $(x-t)_+$ است. اسپلاین مرتبه اول را اسپلاین خطی نیز گویند.

به همین ترتیب، فرض کنید که بتوان بازه $[a,b]$ را به‌صورت اجتماع بازه‌های

$$[a,t_۱),[t_۱,t_۲),\cdots,[t_k,b]$$

نوشت به‌طوری که $t_۱<\cdots<t_k$. نقاط $\{t_i\}_{i=۱}^k$ را گره نامند. در این‌صورت اسپلاین مرتبه اول را می‌توان به‌صورت

$$f(x)=\beta_۰+\beta_۱x+\sum_{j=۱}^{k}\beta_{۱+j}(x-t_j)_+$$

بیان کرد که ترکیبی خطی از توابع پایه‌ای ۱، x، $(x-t_۱)_+$، $\cdots$ و $(x-t_k)_+$ به شمار می‌آید.

اسپلاین مرتبه p

توابع اسپلاین برای مراتب بالاتر را نیز می‌توان به‌طور مشابه تعریف کرد. فرض کنید بازه $[a,b]$ به $k+1$ زیربازه به‌صورت $[a,t_1),[t_1,t_2),\cdots,[t_k,b]$ افراز شده باشد به‌طوری که $a<t_1<\ldots<t_k<b$. تابع $f(\cdot)$ را یک اسپلاین مرتبه‌ی p نامیم اگر در هر یک از زیر بازه‌های فوق یک چندجمله‌ای از مرتبه p باشد به‌طوری‌که $f(\cdot)$ و مشتق‌های آن از مراتب ۱ تا $p-1$، در گره‌ها پیوسته باشند. لذا تابع اسپلاین مرتبه p را می‌توان به‌صورت زیر بیان کرد

$$f(x)=\beta_0+\beta_1 x+\beta_2 x^2+\cdots+\beta_p x^p+\sum_{j=1}^{k}\beta_{p+j}(x-t_j)^p \qquad (ب.۳)$$

که در آن

$$(x-t)_+^p=(x-t)^p I(x>t)$$

تابع توانی بریده‌شده[2] از مرتبه p است. یعنی تابع اسپلاین مرتبه p، ترکیبی خطی از یک چندجمله‌ای مرتبه‌ی p و توابع توانی بریده‌شده است. به‌عبارت دیگر مجموعه

$$\{1,x,x^2,\cdots,x^p,(x-t_1)^p,\cdots,(x-t_k)^p\}$$

یک پایه برای فضای اسپلاین‌های مرتبه‌ی p با گره‌های t_1، t_2، $\cdots$ و t_k است. این مجموعه را پایه‌های اسپلاین مرتبه p گویند.

اسپلاین مکعبی

اگر $p=3$، آن‌گاه تابع مورد اشاره در رابطه (ب.۳) را یک اسپلاین مکعبی[3] نامند که به‌صورت

$$f(x)=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\sum_{j=1}^{k}\beta_{3+j}(x-t_j)^3,$$

تعریف می‌شود. اسپلاین مکعبی یکی از متداول‌ترین نوع اسپلاین است که به دلیل انعطاف‌پذیری بالایی که دارد، در بسیاری از علوم مورد استفاده قرار می‌گیرد.

ب.۲ رگرسیون اسپلاین

مدل رگرسیونی

$$y=f(x)+\epsilon$$

[2] Truncated Power Function
[3] Cubic Spline

را در نظر بگیرید. اگر تابع $f(\cdot)$ را با تابع اسپلاین تعریف شده در رابطه (ب.۳) جایگزین کنیم، نتیجه حاصل را رگرسیون اسپلاین نامند.

یک برتری رگرسیون اسپلاین، برای روش‌های ناپارامتری و نیمه پارامتری این است که متعلق به رده رگرسیون خطی است. به عبارت دیگر، فرم تابع $f(\cdot)$ نسبت به پارامترها خطی است و از این‌رو هزینه‌های محاسباتی برای برآورد منحنی رگرسیون مربوطه، ناچیز است. بنابراین برازش یک اسپلاین به داده‌ها می‌تواند با استفاده از روش کمترین توان‌های دوم خطا صورت گیرد. استفاده از توابع پایه‌ای بریده‌شده، منجر به برازش مدلی به صورت زیر می‌شود

$$\begin{aligned} y_i &= \beta_۰ + \beta_۱ x_i + \cdots + \beta_p x_i^p \\ &\quad + \beta_{p+۱}(x_i - t_۱)_+^p + \cdots + \beta_{p+k}(x_i - t_k)_+^p + \epsilon_i, \qquad i = ۱, \cdots, n \end{aligned}$$

که ϵ_iها در پذیره‌های زیربنایی رگرسیون صدق می‌کنند. شکل ماتریسی این عبارت را می‌توان به‌صورت

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

بیان کرد که در آن $\boldsymbol{\beta} = (\beta_۰, \beta_۱, \cdots, \beta_{p+k})^T$ و $\boldsymbol{T}$ ماتریس طرح با اندازه $n \times (p+k+۱)$ است که درایه‌های ستون اول آن برابر ۱، درایه‌های p ستون بعدی، مربوط به مشاهدات متغیرهای x، $x^۲$، $\cdots$ و x^p و درایه‌های ستون‌های $p+۲$ تا $(p+k+۱)$ مربوط به مشاهدات حاصل از متغیرهای $(x-t_۱)^p$، $\cdots$ و $(x-t_k)^p$ به ازای گره‌های از قبل تعیین شده $t_۱$، $t_۲$، $\cdots$ و t_k هستند، یعنی

$$\boldsymbol{T} = \begin{bmatrix} ۱ & x_۱ & x_۱^۲ & \cdots & x_۱^p & (x_۱ - t_۱)^p & (x_۱ - t_۲)^p \cdots & (x_۱ - t_k)^p \\ ۱ & x_۲ & x_۲^۲ & \cdots & x_۲^p & (x_۲ - t_۱)^p & (x_۲ - t_۲)^p \cdots & (x_۲ - t_k)^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ ۱ & x_n & x_n^۲ & \cdots & x_n^p & (x_n - t_۱)^p & (x_n - t_۲)^p \cdots & (x_n - t_k)^p \end{bmatrix}.$$

با به‌کار بردن روش کمترین توان‌های دوم، برآوردگر زیر حاصل می‌شود

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{T}'\boldsymbol{T})^{-۱}\boldsymbol{T}'\boldsymbol{y}. \qquad (ب.۴)$$

مشکلی که در برآوردگر (ب.۴) وجود دارد، همخطی بین پایه‌های توانی بریده‌شده

$$(x - t_۱)_+^p, (x - t_۲)_+^p, \cdots, (x - t_k)^p$$

است که باعث عدم وارون‌پذیری یا بدشرطیدگی[۴] ماتریس $\boldsymbol{T}'\boldsymbol{T}$ خواهد شد که منجر به ناپایداری ضرایب رگرسیونی می‌شود. برای اطلاعات بیشتر به اریه (۱۳۹۴) و راپرت (۲۰۰۳) مراجعه کنید.

یک روش مناسب برای حل مشکل فوق، استفاده از پایه‌های Bـاسپلاین به جای پایه‌های توانی بریده‌شده، است که منجر به همان برآورد در بازه $[t_۱, t_k]$ می‌شود. با استفاده از پایه‌های Bـاسپلاین می‌توان چندجمله‌ای‌های متعامدی به‌دست آورد. در نتیجه، ضرایب رگرسیونی به‌دست آمده پایدار خواهند بود.

[4] Ill-Condition

ب.۳ Bـاسپلاین

بازه $[a,b]$ و تعداد q گره

$$a = t_۰ < t_۱ < \cdots < t_q < t_{q+۱} = b$$

را در نظر بگیرید. پایه‌های Bـاسپلاین، چندجمله‌ای‌های تکه‌ای در بازه $[a,b]$ هستند و دارای تکیه‌گاه کوچکی می‌باشند. متداول‌ترین روش محاسبه پایه‌های Bـاسپلاین مرتبه p، به‌وسیله پایه‌های Bـاسپلاین مرتبه $p-۱$ و بر اساس رابطه بازگشتی زیر است

$$B_{i,p}(x) = (\frac{x-t_i}{t_{i+p}-t_i})B_{i,p-۱}(x) + (\frac{t_{i+p+۱}-x}{t_{i+p+۱}-t_{i+۱}})B_{i+۱,p-۱}(x), \quad i = -p, -p+۱, \cdots, q$$

(ب.۵)

که در آن $t_{-p} = t_{-p+۱} = \cdots = t_{-۲} = t_{-۱} = t_۰ = a$ و $t_{q+۱} = b$. بنابراین تعداد پایه‌های Bـاسپلاین مرتبه‌ی p برابر با $p+q+۱$ است.

پایه‌های Bـاسپلاین مرتبه صفر

پایه‌های Bـاسپلاین مرتبه صفر $(p=۰)$، برای هر یک از گره‌ها، به‌صورت زیر تعریف می‌شوند

$$B_{i,۰}(x) = \begin{cases} ۱ & x \in [t_i, t_{i+۱}] \\ ۰ & \text{سایر نقاط} \end{cases}$$

که در آن، اندیس صفر بیانگر مرتبه و اندیس i $(i = ۰, ۱, ۲, \cdots, q)$ معرف iـامین تابع متناظر با بازه $[t_i, t_{i+۱}]$ است.

به‌عنوان مثال، این پایه‌ها برای گره‌های $t_۱ = ۱, t_۲ = ۲, t_۳ = ۳$ در بازه $[۰, ۴]$ به‌صورت زیر تعریف می‌شوند

$$B_{۰,۰}(x) = \begin{cases} ۱ & ۰ \le x < ۱ \\ ۰ & \text{سایر نقاط} \end{cases} \qquad B_{۱,۰}(x) = \begin{cases} ۱ & ۱ \le x < ۲ \\ ۰ & \text{سایر نقاط} \end{cases}$$

$$B_{۲,۰}(x) = \begin{cases} ۱ & ۲ \le x < ۳ \\ ۰ & \text{سایر نقاط} \end{cases} \qquad B_{۳,۰}(x) = \begin{cases} ۱ & ۳ \le x < ۴ \\ ۰ & \text{سایر نقاط} \end{cases}$$

پایه‌های Bـاسپلاین مرتبه یک

پایه‌های Bـاسپلاین خطی یا مرتبه یک $(p = ۱)$، با توجه به رابطه (ب.۵) بر حسب پایه‌های مرتبه صفر به‌صورت زیر تعریف می‌شوند

$$B_{i,۱}(x) = \frac{x - t_i}{t_{i+۱} - t_i} B_{i,۰}(x) + \frac{t_{i+۲} - x}{t_{i+۲} - t_{i+۱}} B_{i+۱,۰}(x), \quad i = -۱, ۰, ۱, \cdots, q.$$

که در آن، $t_{-۱} = t_۰ = a$ و $t_{q+۱} = b$. هم‌چنین تعداد این پایه‌ها برابر $p + q + ۱ = q + ۲$ است. با جایگذاری پایه‌های Bـاسپلاین مرتبه صفر، تابع زیر حاصل می‌شود

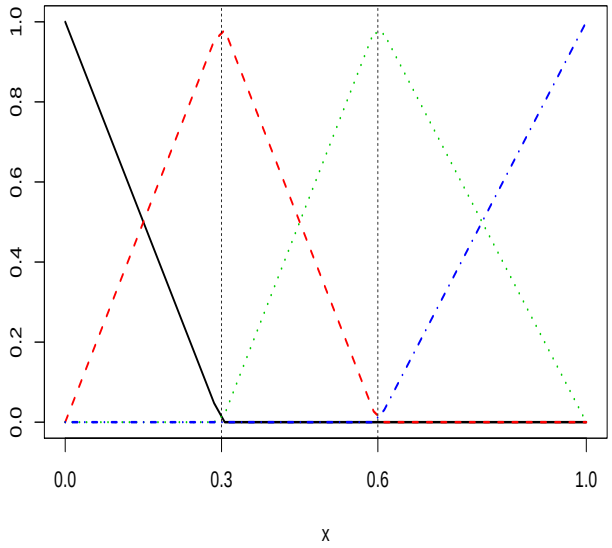
$$B_{i,۱}(x) = \begin{cases} \frac{x-t_i}{t_{i+۱}-t_i} & t_i \le x < t_{i+۱} \\ \frac{t_{i+۲}-x}{t_{i+۲}-t_{i+۱}} & t_{i+۱} \le x < t_{i+۲} \\ ۰ & \text{سایر نقاط} \end{cases}$$

به‌عنوان مثال، اگر در بازه $[۰, ۱]$، دو گره ۰٫۳ و ۰٫۶ را در نظر بگیریم، آن‌گاه پایه‌های Bـاسپلاین مرتبه اول، برای $i = -۱, ۰, ۱, ۲, ۳$، که تعداد آن‌ها برابر $p + q + ۱ = ۴$ است، به‌صورت زیر به‌دست می‌آیند

$$B_{-۱,۱}(x) = \begin{cases} \frac{۰٫۳-x}{۰٫۳} & ۰ \le x < ۰٫۳ \\ ۰ & \text{سایر نقاط} \end{cases} \qquad B_{۰,۱}(x) = \begin{cases} \frac{x}{۰٫۳} & ۰ \le x < ۰٫۳ \\ \frac{۰٫۶-x}{۰٫۳} & ۰٫۳ \le x < ۰٫۶ \\ ۰ & \text{سایر نقاط} \end{cases}$$

$$B_{۱,۱}(x) = \begin{cases} \frac{x-۰٫۳}{۰٫۳} & ۰٫۳ \le x < ۰٫۶ \\ \frac{۱-x}{۰٫۴} & ۰٫۶ \le x < ۱ \\ ۰ & \text{سایر نقاط} \end{cases} \qquad B_{۲,۱}(x) = \begin{cases} \frac{x-۰٫۶}{۰٫۴} & ۰٫۶ \le x < ۱ \\ ۰ & \text{سایر نقاط} \end{cases}$$

نمودار این پایه‌ها در شکل ب.۱ رسم شده‌اند. دو پایه وسطی دارای شکلی شبیه کلاه (۸) هستند، در حالی‌که پایه‌های اولی و آخری، کلاه‌های ناقصی هستند. توجه کنید که چون پایه‌های Bـاسپلاین در زیربازه‌ها غیر صفر هستند، لذا برخی از این پایه‌ها دو به دو متعامد هستند.



شکل ب.۱: نمودار پایه‌های Bـاسپلاین مرتبه یک که به ترتیب از سمت چپ نمودارهای $B_{-۱,۱}(x)$، $B_{۰,۱}(x)$، $B_{۱,۱}(x)$ و $B_{۲,۱}$ هستند.

پایه‌های Bـاسپلاین مرتبه‌ی دو

پایه‌های Bـاسپلاین مرتبه دو $(p = ۲)$ بر حسب مرتبه یک تعریف شده و به‌صورت زیر بیان می‌شوند

$$\begin{aligned} B_{i,۲}(x) &= \frac{x-t_i}{t_{i+۲}-t_i}B_{i,۱}(x) + \frac{t_{i+۳}-x}{t_{i+۳}-t_{i+۱}}B_{i+۱,۱}(x) \\ &= \frac{x-t_i}{t_{i+۲}-t_i}\left[\frac{x-t_i}{t_{i+۱}-t_i}B_{i,۰}(x) + \frac{t_{i+۲}-x}{t_{i+۲}-t_{i+۱}}B_{i+۱,۰}(x)\right] \\ &+ \frac{t_{i+۳}-x}{t_{i+۳}-t_{i+۱}}\left[\frac{x-t_{i+۱}}{t_{i+۲}-t_{i+۱}}B_{i+۱,۰}(x) + \frac{t_{i+۳}-x}{t_{i+۳}-t_{i+۲}}B_{i+۲,۰}(x)\right], i = -۲,-۱,۰,۱,\cdots,k \end{aligned}$$

که در آن $t_{-۲} = t_{-۱} = t_۰ = a$ و $t_{q+۱} = b$. با جایگذاری پایه‌های Bـاسپلاین مرتبه اول، تابع زیر نتیجه می‌شود

$$B_{i,۲} = \begin{cases} \frac{(x-t_i)^۲}{(t_{i+۲}-t_i)(t_{i+۱}-t_i)} & t_i \le x < t_{i+۱} \\ \frac{(x-t_i)(t_{i+۲}-x)}{(t_{i+۲}-t_i)(t_{i+۲}-t_{i+۱})} + \frac{(t_{i+۳}-x)(x-t_{i+۱})}{(t_{i+۳}-t_{i+۱})(t_{i+۲}-t_{i+۱})} & t_{i+۱} \le x < t_{i+۲} \\ \frac{(t_{i+۳}-x)^۲}{(t_{i+۳}-t_{i+۱})(t_{i+۳}-t_{i+۲})} & t_{i+۲} \le x < t_{i+۳} \\ ۰ & \text{سایر نقاط} \end{cases}$$

به‌عنوان مثال، در بازه $[۰, ۱]$ سه گره ۰٫۳، ۰٫۶ و ۰٫۹ را در نظر بگیرید. تعداد پایه‌های Bـاسپلاین مرتبه دوم با سه گره، برابر $p+q+۱=۶$ است که به‌صورت زیر به‌دست می‌آیند

$$B_{-۲,۲}(x)=\begin{cases}(۰٫۳-x)^۲ & ۰\le x<۰٫۳\\ ۰ & \text{سایر نقاط}\end{cases}$$

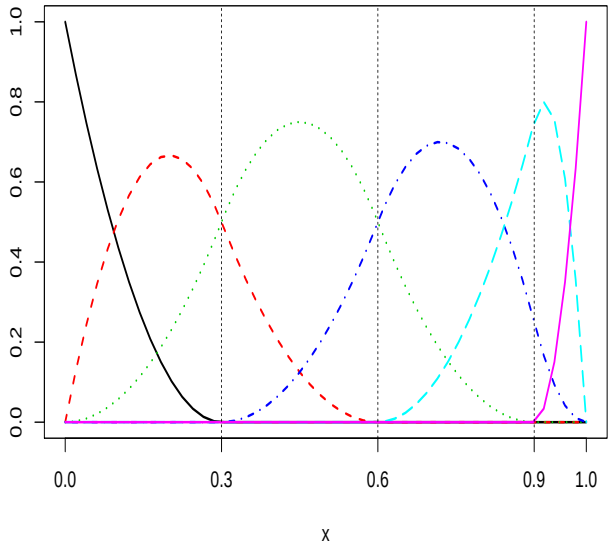
$$B_{-۱,۲}(x)=\begin{cases}\frac{(x)(۰٫۳-x)}{(۰٫۳)(۰٫۳)}+\frac{(۰٫۶-x)(x)}{(۰٫۶)(۰٫۳)} & ۰\le x<۰٫۳\\ \frac{(۰٫۶-x)^۲}{(۰٫۶)(۰٫۳)} & ۰٫۳\le x<۰٫۶\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۰,۲}(x)=\begin{cases}\frac{x^۲}{(۰٫۳)(۰٫۳)} & ۰\le x<۰٫۳\\ \frac{(x)(۰٫۶-x)}{(۰٫۶)(۰٫۳)}+\frac{(۰٫۹-x)(x-۰٫۳)}{(۰٫۶)(۰٫۳)} & ۰٫۳\le x<۰٫۶\\ \frac{(۰٫۹-x)^۲}{(۰٫۶)(۰٫۳)} & ۰٫۶\le x<۰٫۹\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۱,۲}(x)=\begin{cases}\frac{(x-۰٫۳)^۲}{(۰٫۶)(۰٫۳)} & ۰٫۳\le x<۰٫۶\\ \frac{(x-۰٫۳)(۰٫۹-x)}{(۰٫۶)(۰٫۳)}+\frac{(۰٫۹-x)(x-۰٫۶)}{(۰٫۴)(۰٫۳)} & ۰٫۶\le x<۰٫۹\\ \frac{(۱-x)^۲}{(۰٫۳)(۰٫۱)} & ۰٫۹\le x<۱\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۲,۲}(x)=\begin{cases}\frac{(x-۰٫۶)^۲}{(۰٫۴)(۰٫۳)} & ۰٫۶\le x<۰٫۹\\ \frac{(x-۰٫۶)(۱-x)}{(۰٫۴)(۰٫۱)}+\frac{(۱-x)(x-۰٫۹)}{(۰٫۱)(۰٫۱)} & ۰٫۹\le x<۱\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۳,۲}(x)=\begin{cases}\frac{(x-۰٫۹)^۲}{(۰٫۱)(۰٫۱)} & ۰٫۹\le x<۱\\ ۰ & \text{سایر نقاط}\end{cases}$$



شکل ب.۲: نمودار پایه‌های Bـاسپلاین مرتبه‌ی دو که به ترتیب از سمت چپ نمودارهای $B_{-۲,۲}(x)$، $B_{-۱,۲}(x)$، $B_{۰,۲}(x)$، $B_{۱,۲}(x)$، $B_{۲,۲}(x)$ و $B_{۳,۲}(x)$ هستند.

پایه‌های Bـاسپلاین مرتبه‌ی سه

پایه‌های Bـاسپلاین مرتبه سه $(p = ۳)$ به‌صورت سلسله‌مراتبی بر حسب درجات پایین‌تر تعریف شده و به‌صورت زیر بیان می‌شوند

$$
\begin{aligned}
B_{i,۳}(x) &= \frac{(x-t_i)}{(t_{i+۳}-t_i)}B_{i,۲}(x) + \frac{(t_{i+۴}-x)}{(t_{i+۴}-t_{i+۱})}B_{i+۱,۲}(x) \\
&= \frac{(x-t_i)}{(t_{i+۳}-t_i)}\left[\frac{x-t_i}{t_{i+۲}-t_i}B_{i,۱}(x) + \frac{t_{i+۳}-x}{t_{i+۳}-t_{i+۱}}B_{i+۱,۱}(x)\right] \\
&+ \frac{(t_{i+۴}-x)}{(t_{i+۴}-t_{i+۱})}\left[\frac{x-t_{i+۱}}{t_{i+۳}-t_{i+۱}}B_{i+۱,۱}(x) + \frac{t_{i+۴}-x}{t_{i+۴}-t_{i+۲}}B_{i+۲,۱}(x)\right] \\
&= \frac{(x-t_i)}{(t_{i+۳}-t_i)}\left[\frac{x-t_i}{t_{i+۲}-t_i}\left(\frac{x-t_i}{t_{i+۱}-t_i}B_{i,۰}(x) + \frac{t_{i+۲}-x}{t_{i+۲}-t_{i+۱}}B_{i+۱,۰}(x)\right)\right. \\
&\left.+\frac{t_{i+۳}-x}{t_{i+۳}-t_{i+۱}}\left(\frac{x-t_{i+۱}}{t_{i+۲}-t_{i+۱}}B_{i+۱,۰}(x) + \frac{t_{i+۳}-x}{t_{i+۳}-t_{i+۲}}B_{i+۲,۰}(x)\right)\right] \\
&+ \frac{(t_{i+۴}-x)}{(t_{i+۴}-t_{i+۱})}\left[\frac{x-t_{i+۱}}{t_{i+۳}-t_{i+۱}}\left(\frac{x-t_{i+۱}}{t_{i+۲}-t_{i+۱}}B_{i+۱,۰}(x) + \frac{t_{i+۳}-x}{t_{i+۳}-t_{i+۲}}B_{i+۲,۰}(x)\right)\right. \\
&\left.+\frac{t_{i+۴}-x}{t_{i+۴}-t_{i+۲}}\left(\frac{x-t_{i+۲}}{t_{i+۳}-t_{i+۲}}B_{i+۲,۰}(x) + \frac{t_{i+۴}-x}{t_{i+۴}-t_{i+۳}}B_{i+۳,۰}(x)\right)\right]
\end{aligned}
$$

$$B_{i,۳}(x)=\begin{cases}\frac{(t_{i+۴}-x)^۳}{(t_{i+۴}-t_{i+۱})(t_{i+۴}-t_{i+۲})(t_{i+۴}-t_{i+۳})} & t_i\le x<t_{i+۱}\\ \frac{(x-t_i)^۲(t_{i+۲}-x)}{(t_{i+۳}-t_i)(t_{i+۲}-t_i)(t_{i+۲}-t_{i+۱})}+\frac{(x-t_i)(t_{i+۳}-x)(x-t_{i+۱})}{(t_{i+۳}-t_i)(t_{i+۳}-t_{i+۱})(t_{i+۲}-t_{i+۱})} & \\ +\frac{(t_{i+۴}-x)(x-t_{i+۱})^۲}{(t_{i+۴}-t_{i+۱})(t_{i+۳}-t_{i+۱})(t_{i+۲}-t_{i+۱})} & t_{i+۱}\le x<t_{i+۲}\\ \frac{(x-t_i)(t_{i+۳}-x)^۲}{(t_{i+۳}-t_i)(t_{i+۳}-t_{i+۱})(t_{i+۳}-t_{i+۲})}+\frac{(t_{i+۴}-x)(x-t_{i+۱})(t_{i+۳}-x)}{(t_{i+۴}-t_{i+۱})(t_{i+۳}-t_{i+۱})(t_{i+۳}-t_{i+۲})} & \\ +\frac{(t_{i+۴}-x)^۲(x-t_{i+۲})}{(t_{i+۴}-t_{i+۱})(t_{i+۴}-t_{i+۲})(t_{i+۳}-t_{i+۲})} & t_{i+۲}\le x<t_{i+۳}\\ \frac{(x-t_i)^۳}{(t_{i+۳}-t_i)(t_{i+۲}-t_i)(t_{i+۱}-t_i)} & t_{i+۳}\le x<t_{i+۴}\\ ۰ & \text{سایر نقاط}\end{cases}$$

به‌عنوان مثال، در بازه $[۰,۱]$ چهار گره ۰٫۲، ۰٫۴، ۰٫۶ و ۰٫۸ را در نظر بگیرید. تعداد پایه‌های Bـاسپلاین مرتبه سه با چهار گره برابر $p+q+۱=۸$ است که به‌صورت زیر به‌دست می‌آیند

$$B_{-۳,۳}(x)=\begin{cases}\frac{(۰٫۲-x)^۳}{(۰٫۲)(۰٫۲)(۰٫۲)} & ۰\le x<۰٫۲\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{-۲,۳}(x)=\begin{cases}\frac{(۰٫۴-x)^۳}{(۰٫۴)(۰٫۴)(۰٫۲)} & ۰\le x<۰٫۲\\ \frac{(x)(۰٫۲-x)^۲}{(۰٫۲)(۰٫۲)(۰٫۲)}+\frac{(۰٫۴-x)(x)(۰٫۲-x)}{(۰٫۴)(۰٫۲)(۰٫۲)}+\frac{(۰٫۴-x)^۲(x)}{(۰٫۴)(۰٫۴)(۰٫۲)} & ۰٫۲\le x<۰٫۴\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{-۱,۳}(x)=\begin{cases}\frac{(۰٫۶-x)^۳}{(۰٫۶)(۰٫۴)(۰٫۲)} & ۰\le x<۰٫۲\\ \frac{(x)^۲(۰٫۲-x)}{(۰٫۴)(۰٫۲)(۰٫۲)}+\frac{(x)(۰٫۴-x)(x)}{(۰٫۴)(۰٫۴)(۰٫۲)}+\frac{(۰٫۶-x)(x)^۲}{(۰٫۶)(۰٫۴)(۰٫۲)} & ۰٫۲\le x<۰٫۴\\ \frac{(x)(۰٫۴-x)^۲}{(۰٫۴)(۰٫۴)(۰٫۲)}+\frac{(۰٫۶-x)(x)(۰٫۴-x)}{(۰٫۶)(۰٫۴)(۰٫۲)}+\frac{(۰٫۶-x)^۲(x-۰٫۲)}{(۰٫۶)(۰٫۴)(۰٫۲)} & ۰٫۴\le x<۰٫۶\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۰,۳}(x)=\begin{cases}\frac{x^{۳}}{(۰/۶)(۰/۴)(۰/۲)} & ۰\le x<۰/۲\\ \frac{(x^{۲})(۰/۴-x)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(x)(۰/۶-x)(x-۰/۲)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۰/۸-x)(x-۰/۲)^{۲}}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۲\le x<۰/۴\\ \frac{(x)(۰/۶-x)^{۲}}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۰/۸-x)(x-۰/۲)(۰/۶-x)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۰/۸-x)^{۲}(x-۰/۴)}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۴\le x<۰/۶\\ \frac{(۰/۸-x)^{۳}}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۶\le x<۰/۸\\ ۰ & \text{سایر نقاط}\end{cases}$$

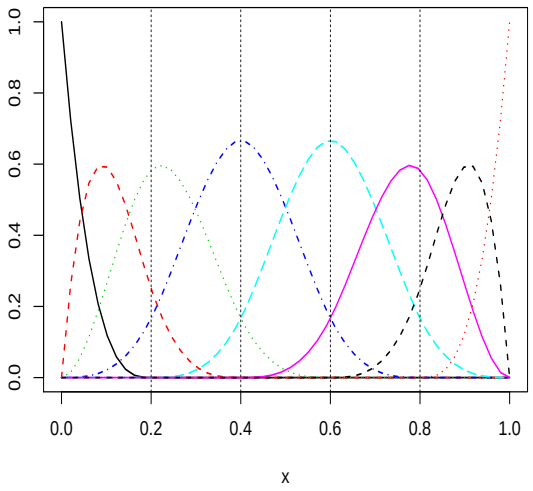
$$B_{۱,۳}(x)=\begin{cases}\frac{(۱-x)^{۳}}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۲\le x<۰/۴\\ \frac{(x-۰/۲)^{۲}(۰/۶-x)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(x-۰/۲)(۰/۸-x)(x-۰/۴)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۱-x)(x-۰/۴)^{۲}}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۴\le x<۰/۶\\ \frac{(x-۰/۲)(۰/۸-x)^{۲}}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۱-x)(x-۰/۴)(۰/۸-x)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۱-x)^{۲}(x-۰/۶)}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۶\le x<۰/۸\\ \frac{(x-۰/۲)^{۳}}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۸\le x<۱\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۲,۳}(x)=\begin{cases}\frac{(x-۰/۴)^{۲}(۰/۸-x)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(x-۰/۴)(۱-x)(x-۰/۶)}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۱-x)(x-۰/۶)^{۲}}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۴\le x<۰/۶\\ \frac{(x-۰/۴)(۱-x)^{۲}}{(۰/۶)(۰/۴)(۰/۲)}+\frac{(۱-x)(x-۰/۶)(۱-x)}{(۰/۴)(۰/۴)(۰/۲)}+\frac{(۱-x)^{۲}(x-۰/۸)}{(۰/۴)(۰/۲)(۰/۲)} & ۰/۶\le x<۰/۸\\ \frac{(x-۰/۴)^{۳}}{(۰/۶)(۰/۴)(۰/۲)} & ۰/۸\le x<۱\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۳,۳}(x)=\begin{cases}\frac{(x-۰/۶)^{۳}}{(۰/۴)(۰/۴)(۰/۲)} & ۰/۶\le x<۰/۸\\ \frac{(x-۰/۶)^{۲}(۱-x)}{(۰/۴)(۰/۴)(۰/۲)}+\frac{(x-۰/۶)(۱-x)(x-۰/۸)}{(۰/۴)(۰/۲)(۰/۲)}+\frac{(۱-x)(x-۰/۸)}{(۰/۲)(۰/۲)(۰/۲)} & ۰/۸\le x<۱\\ ۰ & \text{سایر نقاط}\end{cases}$$

$$B_{۴,۳}(x)=\begin{cases}\frac{(x-۰/۸)^{۳}}{(۰/۲)(۰/۲)(۰/۲)} & ۰/۸\le x<۱\\ ۰ & \text{سایر نقاط}\end{cases}$$

این توابع از جنس چندجمله‌ای‌های تکه‌ای مرتبه سه هستند و نمودارهای این توابع در شکل ب.۳ رسم شده‌اند.



شکل ب.۳: نمودار پایه‌های Bـاسپلاین مرتبه سه که به ترتیب از سمت چپ نمودارهای $B_{-۳,۳}(x)$، $B_{-۲,۳}(x)$، $B_{-۱,۳}(x)$، $B_{۰,۳}(x)$، $B_{۱,۳}(x)$ ، $B_{۲,۳}(x)$، $B_{۳,۳}(x)$ و $B_{۴,۳}(x)$ هستند.

ب.۴ کمترین توان‌های دوم Bـاسپلاین

ابتدا با $p+q+۱$ پایه Bـاسپلاین مرتبه p، رابطه (ب.۵) را با اندیس‌گذاری مجدد، به‌صورت

$$B_{۰,p}, B_{۱,p}, \cdots, B_{p+q,p}$$

در نظر می‌گیریم. مدل رگرسیونی Bـاسپلاین به‌صورت زیر است

$$s(x) = \sum_{j=۱}^{p+q} \beta_j B_{j,p}(x) + \epsilon. \qquad (ب.۶)$$

از آن‌جا که صورت ماتریسی این مدل به‌صورت

$$s = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

است، لذا برآورد کمترین توان‌های دوم $\boldsymbol{\beta}$ عبارت است از

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{B}'\boldsymbol{B})^{-۱}\boldsymbol{B}'\boldsymbol{y}.$$

متعامد بودن پایه‌های Bـاسپلاین موجب می‌شود که ماتریس $\boldsymbol{B}'\boldsymbol{B}$ وارون‌پذیر باشد و لذا مشکل موجود در رابطه (ب.۴) مرتفع گردد. لازم به ذکر است که علیرغم به‌دست آوردن این مزیت، درک پایه‌های Bـاسپلاین به آسانی درک توابع توانی بریده‌شده نیست و لذا ضرایب مدل مذکور غیر قابل تفسیر است. برای اطلاعات بیشتر در این‌باره می‌توانید به دی بور (۲۰۰۱)، راپرت (۲۰۰۳) و دایرکس[۵] (۱۹۹۵) مراجعه کنید.

[۵] Dierckx

پیوست پ

تعاریف و نامساوی‌ها

تعریف پ.۱.۰ (نامساوی دوُرتزکی– کیفر– وُلفوویتز)**.** فرض کنید $X_n, ..., X_1$ یک نمونه تصادفی از تابع توزیع $F(\cdot)$ بوده و $F_n(\cdot)$ تابع توزیع تجربی متناظر باشد. این نامساوی یک کران بالا برای احتمال اینکه اختلاف بین $F(\cdot)$ و $F_n(\cdot)$ از یک مقدار ثابت داده‌شده $\varepsilon > 0$ بزرگتر باشد، نتیجه می‌دهد. به‌طور دقیق‌تر، به ازای هر $\varepsilon > 0$، داریم

$$P\Big(\sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid \geq \varepsilon\Big) \leq 2\exp\{-2n\varepsilon^2\}.$$

تعریف پ.۲.۰ (نامساوی مارکوف)**.** اگر X یک متغیر تصادفی نامنفی باشد، آنگاه به‌ازای هر $a > 0$، داریم

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

تعریف پ.۳.۰ (تجزیه چولسکی)**.** هر ماتریس متقارن معین مثبت $\mathbf{A}$ را می‌توان به‌صورت زیر نوشت

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

که $\mathbf{L}$ یک ماتریس پایین مثلثی و $\mathbf{L}^T$ ترانهاده آن است. لازم به ذکر است که تجزیه چولسکی یکتاست.

به‌عنوان مثال، در یک ماتریس ۳ × ۳ باید دستگاه معادلات زیر حل شود:

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} a_{۱۱} & a_{۱۲} & x_{۱۳} \\ a_{۲۱} & a_{۲۲} & x_{۲۳} \\ a_{۳۱} & a_{۳۲} & x_{۳۳} \end{pmatrix} \\ &= \begin{pmatrix} l_{۱۱} & ۰ & ۰ \\ l_{۲۱} & l_{۲۲} & ۰ \\ l_{۳۱} & l_{۳۲} & l_{۳۳} \end{pmatrix} \begin{pmatrix} l_{۱۱} & l_{۲۱} & l_{۳۱} \\ ۰ & l_{۲۲} & l_{۳۲} \\ ۰ & ۰ & l_{۳۳} \end{pmatrix} = \mathbf{L}\mathbf{L}^T \\ &= \begin{pmatrix} l_{۱۱}^۲ & l_{۲۱}l_{۱۱} & l_{۳۱}l_{۱۱} \\ l_{۲۱}l_{۱۱} & l_{۲۱}^۲ + l_{۲۲}^۲ & l_{۳۱}l_{۲۱} + l_{۳۲}l_{۲۲} \\ l_{۳۱}l_{۱۱} & l_{۳۱}l_{۲۱} + l_{۳۲}l_{۲۲} & l_{۳۱}^۲ + l_{۳۲}^۲ + l_{۳۳}^۲ \end{pmatrix}.\end{aligned}$$

می‌توان نشان داد که برای اعضای قطر اصلی (l_{kk}) و اعضای زیر قطر اصلی $(l_{ik}, i > k)$ روابط زیر برقرار است

$$l_{kk} = a_k - \sum_{j=۱}^{k-۱} a_{kj}^۲, \qquad l_{ik} = \frac{۱}{l_{kk}}(a_{kk} - \sum_{j=۱}^{k-۱} l_{ij}l_{kj}).$$

تعریف پ.۴.۰ (پایه)**.** مجموعه بردارهای $\mathbf{u}_۱, \ldots, \mathbf{u}_n$ در فضای برداری $\mathbf{V}$ تشکیل یک پایه می‌دهند، اگر دو شرط زیر را داشته باشند:

۱. بردارهای $\mathbf{u}_۱, \ldots, \mathbf{u}_n$ مستقل خطی باشند.

۲. هر کدام از بردارهای فضای $\mathbf{V}$ را بتوان از ترکیب خطی این بردارها به دست آورد.

تعریف پ.۵.۰ (همگرایی در توزیع)**.** دنباله متغیرهای تصادفی $\{X_n\}$ در توزیع به متغیر تصادفی X همگراست اگر برای هر $x \in \mathbb{R}$ که تابع F در آن پیوسته است، داشته باشیم

$$\lim_{n\to\infty} F_n(x) = F(x)$$

که در آن F_n تابع توزیع تجمعی $\{X_n\}$ و F تابع توزیع تجمعی X است. این نوع همگرایی را به‌صورت $X_n \xrightarrow{D} X$ نشان می‌دهند.

تعریف پ.۶.۰ (همگرایی در احتمال)**.** دنباله متغیرهای تصادفی $\{X_n\}$ در احتمال به متغیر تصادفی X همگراست اگر برای هر $\varepsilon > ۰$ داشته باشیم

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = ۰.$$

این نوع همگرایی را با نماد $X_n \xrightarrow{P} X$ نشان می‌دهند.

تعریف پ.۷.۰ (سازگاری)**.** دنباله برآوردگرهای $\{T_n\}$ برای $a(\theta)$ سازگار است، اگر $\{T_n\}$ در احتمال به $a(\theta)$ همگرا باشد.

تعریف پ.۰.۸. گوییم $f(n) = O\big(g(n)\big)$ ($f(n)$ از مرتبه $g(n)$ است)، اگر وقتی $n \to \infty$، آنگاه $\frac{f(n)}{g(n)}$ متناهی باشد. به‌عبارت دیگر، $f(n) = O\big(g(n)\big)$، اگر عدد حقیقی مثبت M و عدد حقیقی n_0 موجود باشند، به‌طوری‌که

$$\forall n > n_0 \qquad |f(n)| \leq M|g(n)|$$

یعنی برای مقادیر به‌قدر کافی بزرگ n، قدرمطلق $f(n)$ حداکثر M برابر قدرمطلق $g(n)$ باشد.

تعریف پ.۰.۹. گوییم $f(n) = o\big(g(n)\big)$ ($f(n)$ از مرتبه کوچکتر از $g(n)$ است)، اگر وقتی $n \to \infty$، آنگاه داشته باشیم

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

عبارت $f(n) = o\big(g(n)\big)$ بیان می‌کند که $g(n)$ بسیار سریع‌تر از $f(n)$ رشد می‌کند. بنابراین o حالت قوی‌تری از O را بیان می‌کند. هر عبارتی که $o(g)$ باشد، $O(g)$ هم هست، اما عکس آن برقرار نیست.

تعریف پ.۰.۱۰. فرض کنید $\{X_n\}$ یک دنباله از متغیرهای تصادفی و $\{a_n\}$ یک دنباله از ثابت‌های متناظر باشد. نماد $X_n = o_p\big(a(n)\big)$ بدین معنی است که مجموعه مقادیر $\frac{X_n}{a_n}$ در احتمال به صفر همگرا است، یعنی برای هر $\varepsilon > 0$، داریم

$$\lim_{n\to\infty} P(|\frac{X_n}{a_n}| > \varepsilon) = 0.$$

تعریف پ.۰.۱۱. فرض کنید $\{X_n\}$ یک دنباله از متغیرهای تصادفی و $\{a_n\}$ یک دنباله از ثابت‌های متناظر باشد. نماد $X_n = O_p\big(a(n)\big)$ بدین معنی است که مقادیر $\frac{X_n}{a_n}$ در احتمال کراندار است، یعنی برای هر $\varepsilon > 0$، اعداد متناهی $M, N > 0$ وجود دارند، به‌طوری‌که

$$\forall n > N \qquad P(|\frac{X_n}{a_n}| > M) < \varepsilon.$$

تعریف پ.۰.۱۲. گوییم برآوردگر $\hat{\theta}_n$ یک برآوردگر $\sqrt{n}$ـ سازگار برای θ است، اگر رابطه زیر برقرار باشد

$$\hat{\theta_n} - \theta = O_p(n^{-1/2}).$$

تعریف پ.۰.۱۳ (بسط تیلور). فرض کنید تابع $f(x)$ بر بازه $[a, b]$ تعریف شده و $n+1$ بار مشتق‌پذیر باشد. نقطه $x_0 \in (a, b)$ را در نظر بگیرید. برای هر $x \in (a, b)$ که $x = x_0 + \Delta x$، عددی مانند ξ بین دو نقطه x و x_0 وجود دارد، به‌طوری که تساوی زیر برقرار است

$$f(x_0 + \Delta x) = f(x_0) + \Delta x f'(x_0) + \frac{\Delta x^2}{2!} f''(x_0) + ... + \frac{\Delta x^n}{n!} f^{(n)}(x_0) + R_n(x_0 + \Delta x)$$

که

$$R_n(x_0 + \Delta x) = \frac{\Delta x^{n+1}}{(n+1)!} f^{(n+1)}(\xi), \qquad x_0 \leq \xi \leq x.$$

منظور از $f^{(n)}$ در این روابط مشتق مرتبه nام می‌باشد. در عمل معمولاً از تعداد اندکی از جملات اول سری برای تقریب استفاده می‌شود. در این حالت، اگر مقدار Δx به‌قدر کافی کوچک باشد، تقریب

مناسبی از تابع در همسایگی $x_۰$ به‌دست خواهد آمد. در صورتی‌که تقریب تابع با سری تیلور برای n جمله اول انجام شود، مانده سری یا خطای تقریب متناسب با Δx^n خواهد بود و آن‌را با $O(\Delta x^n)$ نشان می‌دهند

$$f(x_۰ + \Delta x) = f(x_۰) + \Delta x f'(x_۰) + \frac{\Delta x^۲}{۲!} f''(x_۰) + ... + \frac{\Delta x^{n-۱}}{(n-۱)!} f^{(n-۱)}(x_۰) + O(\Delta x^n).$$

پیوست ت

گزیده‌ای از برنامه‌های رایانه‌ای

جدول ۱.۱

```
library(scatterplot3d)
library(AER)
data("CPS1985")
fit <- lm(log(wage)~education+experience+I(experience^2),data=CPS1985)
summary(fit)
```

شکل ۱.۱

```
b <- coef(fit)
m.school <- data.frame(CPS1985$education,CPS1985$education*b["education"])
m.school <- m.school[order(CPS1985$education),]
m.experi <- data.frame(CPS1985$experience,
CPS1985$experience*b["experience"]
+CPS1985$experience^2*b["I(experience^2)"])
m.experi <- m.experi[order(CPS1985$experience),]
par(mfrow=c(1,2))
par(mar = c(bottom=4, 2, top=2, 2))
```

```
plot(m.school,col="black",pch=19,cex=0.7,
xaxt="n",yaxt="n",xlab="school",ylab="")
axis(1,cex.axis=0.8)
axis(2,cex.axis=0.8)
lines(m.school,col="azure3")
plot(m.experi,col="black",pch=19,cex=0.7,xaxt="n",yaxt="n",
xlab="experience",ylab="")
axis(1,cex.axis=0.8)
axis(2,cex.axis=0.8)
lines(m.experi,col="azure3")
col <- rep("green",nrow(CPS1985)); col[CPS1985$education==12] <- "red"
pch <- rep(1,nrow(CPS1985)); pch[CPS1985$education==12] <- 8
```

شکل ۲.۱

```
scatterplot3d(CPS1985$education,CPS1985$experience, fitted(fit),
xlab="school",ylab="experience",zlab="",
color=col,pch=pch,angle=130)
```

شکل ۳.۱

```
library(np)
require(graphics)

education<- CPS1985$education
experience<-CPS1985$experience
X <- cbind(education,experience)
y <- log(CPS1985$wage)
bw <- npregbw(xdat=X, ydat=y)
model <- npreg(bws = bw, gradients = TRUE)
summary(model)
col <- rep("green",nrow(CPS1985)); col[CPS1985$education==12] <- "red"
pch <- rep(1,nrow(CPS1985)); pch[education==12] <- 8
scatterplot3d(education,experience, model$mean,
xlab="education",ylab="experience",zlab="",
color=col,pch=pch,angle=130)
```

شکل ۵.۱

```
library(mgcv)
b <- gam(y ~s(education) + s(experience))
plot(b,pages = 1)
fit3 <- fitted(b)
scatterplot3d(education,experience, fit3,
xlab="education",ylab="experience",zlab="",
color=col,pch=pch,angle=130)
```

شکل ۲.۲

```
S <- function(xx,l){
(xx - sign(xx))*(abs(xx) > l)
}
scad <- function(theta,l,a=3.7){
theta <- abs(theta)
(theta <= l) * theta * l + (theta > l & theta <= a*l)*
(a*l*theta-(theta^2+l^2)/2)/(a-1)+(theta>a*l)*(l^2*(a^2-1))/(2*(a-1))
}
Fs <- function(z,l,g=3.7){
(abs(z)>g*l)*z+(abs(z)>2*l & abs(z)<=g*l)*(S(z,g*l/(g-1))/(1-1/(g-1)))+
(abs(z) <= 2*l)*S(z,l)
}
mcp <- function(theta,l,a=3.7) {
T <- length(theta)
val <- numeric(T)
for (i in 1:T) {
x <- abs(theta[i])
val[i] <- (x < a*l)*(l*x - x^2/(2*a)) + (x >= a*l)*(1/2)*a*l^2
}
val
}
Fm <- function(z,l,g=3.7){
(abs(z) > g*l)*z + (abs(z) <= g*l)*(S(z,l)/(1-1/g))
```

```
}
lasso <- function(theta,l){l*abs(theta)}

res <- 101
x <- seq(-4,4,len=res)
xx <- seq(0.0001,4,len=res)
g <- 3
col <- c("#FF4E37FF", "#00B500FF", "#008DFFFF")
Y <- cbind(lasso(x, 1), scad(x, 1, g), mcp(x, 1, g))

require(latex2exp)
par(mar = c(bottom=4, 5, top=1, 2))
matplot(x,Y,type='l',xaxt="n",yaxt="n",lty=1:3,lwd=c(1,1,1.8),col=col,las=1,
xlab=expression(beta), ylab=TeX("$P_{\\lambda}(l\\beta l)$"),cex.lab=0.85)
axis(1,cex.axis=0.7)
axis(2,cex.axis=0.7)
legend(-1.70, 4, legend=c("LASSO", "SCAD","MCP"),
col=c("red","green", "blue"), lty=1:3,lwd=c(1,1,1.8), cex=0.7)
```

شکل ۳.۲

```
par(mfrow=c(2,2))
par(mar = c(bottom=2.2, 3, top=2, 2))
Y <- cbind(S(x,1), Fm(x,1,g), Fs(x,1,g))
plot(x,S(x,1),type='l',xaxt="n",yaxt="n",xlim=c(-4,4),
ylim=c(-4,4),lwd=2.5,xlab="",ylab="",main="LASSO",cex.main=0.8)
lines(x,x,col="red")
axis(1,cex.axis=0.7)
axis(2,cex.axis=0.7)
par(mar = c(bottom=2.2, 3, top=2, 2))
plot(x,0.5*x,type = 'l',xaxt="n",yaxt="n", xlim=c(-4,4), ylim=c(-4,4),
lwd=2.5,xlab="",ylab="",main="Ridge",cex.main=0.8)
lines(x,x,col="red")
axis(1,cex.axis=0.7)
axis(2,cex.axis=0.7)
```

```
plot(x,Fs(x,1,g),type='l',xaxt="n",yaxt="n",xlim=c(-4,4),ylim=c(-4,4),
lwd=2.5,xlab="",ylab="",main = "SCAD",cex.main=0.8)
lines(x,x,col="red")
axis(1,cex.axis=0.7)
axis(2,cex.axis=0.7)
plot(x,Fm(x,1,g),type='l',xaxt="n",yaxt="n",xlim=c(-4,4),ylim=c(-4,4),
lwd=2.5,xlab="",ylab="",main="MCP",cex.main=0.8)
lines(x,x,col="red")
axis(1,cex.axis=0.7)
axis(2,cex.axis=0.7)
```

شکل ۴.۲

```
dLasso <- function(theta,l){rep(l,length(theta))}
dMCP <- function(theta,l,a=3.7) {
theta = abs(theta)
(theta <= a*l) * (l - theta/a)}
dSCAD <- function(theta,l,a=3.7) {
theta = abs(theta)
(theta<l)*l+((theta>l)&(theta<a*l))*((a*l-theta)/(a-1))
}
Y<- cbind(dLasso(xx, 1), dSCAD(xx, 1, g), dMCP(xx, 1, g))
matplot(xx,Y,type='l',xaxt="n",yaxt="n",lty=1:3,lwd=c(1,1,1.8),
col=col, las=1,cex.lab=0.85, xlab=expression(beta),
ylab=TeX("$P'_{\\lambda}(l\\beta l)$"))
axis(1,cex.axis=0.7)
axis(2,cex.axis=0.7)
legend(2.5, 0.9, legend=c("LASSO", "SCAD","MCP"),
col=c("red","green","blue"), lty=1:3, lwd=c(1,1,1.8),cex=0.7)
```

شکل ۵.۲

```
library(splines)
library(grpreg)

n <- 2000; p <- 20; K < -2
```

```
X <- matrix(0,n,p)
for(j in 1:p ){
X[,j] <- runif(n)
}

obj0 <- function(xx,K)
{
n <- nrow(xx);p <- ncol(xx)
z <- c()
for (j in 1:p){
B1 <- splineDesign(knots=c(0,0,0,seq(0,1,length=K+2),1,1,1),x=xx[,j],
derivs=rep(0,n))
B1 <- B1[,1:(K+1)]
z <- cbind(z,B1)}
z
}
Z < -obj0(X,K)
Z <- Z[,-3]
group<-c(c(1,1),rep(2:p,each=K+1))
Y <- -sqrt(2)*Z[,1]+sqrt(2)*Z[,2]+0.5*Z[,3]+Z[,4]-0.5*Z[,5]+rnorm(n,0,1)
fit1 <- grpreg(Z, Y, group, penalty="grLasso")
fit2 <- grpreg(Z, Y, group, penalty="grSCAD")
fit3 <- grpreg(Z, Y, group, penalty="grMCP",gamma=2.7)
par(mfrow=c(1,3))
myColors<-c("blue","red","green","black","yellow","purple","orange","brown")
plot(fit1,col=myColors,main="Group LASSO", xaxt="n",yaxt="n",cex.main=1)
axis(1,cex.axis=1.2)
axis(2,cex.axis=1.2)
abline(v=.04,col="black",lty=2,lwd=0.75)
plot(fit2,col=myColors,main="Group SCAD", xaxt="n",yaxt="n",cex.main=1)
axis(1,cex.axis=1.2)
axis(2,cex.axis=1.2)
abline(v=.04,col="black",lty=2,lwd=0.75)
plot(fit3,col=myColors,main="Group MCP", xaxt="n",yaxt="n",cex.main=1)
```

```
axis(1,cex.axis=1.2)
axis(2,cex.axis=1.2)
abline(v=.04,col="black",lty=2,lwd=0.75)
```

شکل ۸.۲

```
res <- 201
x <- seq(-10,10,len=res)
xx <- seq(0.0001,4,len=res)
g <- 3
col <- c("#FF4E37FF", "#00B500FF", "#008DFFFF")
lambda <- 2
y <- scad(x, lambda) # a=3.7
#### SCAD
require(latex2exp)
par(mfrow=c(2,2))
par(mar = c(bottom=3, 4, top=2, 3))
# lambda=2,beta0=4
matplot(x, y,lwd=1.3, type='l', ylim=c(0,20), main=TeX("SCAD:
$\\beta^{(0)}$=4"), xaxt="n",yaxt="n", cex.main=1.5, lty=1,
las=1, xlab=expression(beta), ylab=TeX("$P_{\\lambda}(\\beta)$"))
axis(1,cex.axis=0.9)
axis(2,cex.axis=0.9)
beta0<-4
LLAapprox<-function(Beta){
scad(beta0,lambda)+dSCAD(beta0,lambda)*(abs(Beta)-beta0)
}
LQAapprox<-function(Beta){
scad(beta0,lambda)+0.5*(dSCAD(beta0,lambda)/beta0)*(Beta^2-beta0^2)
}
lines(x,LQAapprox(x),lty=2,lwd=1.4,col="blue")
lines(x,LLAapprox(x),lty=3,lwd=1.4,col="red")
# lambda=2,beta0=1
par(mar = c(bottom=3, 4, top=2, 3))
matplot(x, y, type='l', ylim=c(0,20), main=TeX("SCAD: $\\beta^{(0)}$=1"),
```

```
cex.main=1.5, xaxt="n",yaxt="n",lty=1,lwd=1.3,las=1,
xlab=expression(beta), ylab=TeX("$P_{\\lambda}(\\beta)$"))
axis(1,cex.axis=0.9)
axis(2,cex.axis=0.9)
beta0<-1
LLAapprox<-function(Beta){
scad(beta0,lambda)+dSCAD(beta0,lambda)*(abs(Beta)-beta0)
}
LQAapprox<-function(Beta){
scad(beta0,lambda)+0.5*(dSCAD(beta0,lambda)/beta0)*(Beta^2-beta0^2)
}
lines(x,LQAapprox(x),lty=2,lwd=1.4,col="blue")
lines(x,LLAapprox(x),lty=3,lwd=1.4,col="red")
############ Lq:q=0.5
# lambda=2,beta0=4
Lq <- function(theta,l){l*abs(theta)^0.5}
dLq <- function(theta,l){(l/2)*theta^(-0.5)}
lambda<-2
yLq <- Lq(x, lambda)
par(mar = c(bottom=3, 4, top=2, 3))
matplot(x, yLq,lwd=1.3, type='l', ylim=c(0,10), main=TeX("L$_{0.5}:
\\beta^{(0)}$=4"), xaxt="n",yaxt="n", cex.main=1.5, lty=1, las=1,
xlab=expression(beta), ylab=TeX("$P_{\\lambda}(\\beta)$"))
axis(1,cex.axis=0.9)
axis(2,cex.axis=0.9)
beta0<-4
LLAapprox<-function(Beta){
Lq(beta0,lambda)+dLq(beta0,lambda)*(abs(Beta)-beta0)
}
LQAapprox<-function(Beta){
Lq(beta0,lambda)+0.5*(dLq(beta0,lambda)/beta0)*(x^2-beta0^2)
}
lines(x,LQAapprox(x),lty=2,lwd=1.4,col="blue")
lines(x,LLAapprox(x),lty=3,lwd=1.4,col="red")
```

```
# lambda=2,beta0=1
par(mar = c(bottom=3, 4, top=2, 3))
matplot(x, yLq,lwd=1.3, type='l', ylim=c(0,10),main=TeX("L$_{0.5}:
\\beta^{(0)}$=1"), xaxt="n",yaxt="n", cex.main=1.5, lty=1, las=1,
xlab=expression(beta), ylab=TeX("$P_{\\lambda}(\\beta)$"))
axis(1,cex.axis=0.9)
axis(2,cex.axis=0.9)
beta0<-1
LLAapprox<-function(Beta){
Lq(beta0,lambda)+dLq(beta0,lambda)*(abs(Beta)-beta0)
}
LQAapprox<-function(Beta){
Lq(beta0,lambda)+0.5*(dLq(beta0,lambda)/beta0)*(x^2-beta0^2)
}
lines(x,LQAapprox(x),lty=2,lwd=1.4,col="blue")
lines(x,LLAapprox(x),lty=3,lwd=1.4,col="red")
```

مثال ۲.۴.۳

```
############ Load R Packages
library(splines)
library(MASS)
library(sn)
library(energy)

##### Three Screening Procedures: RDC-SIS, DC-SIS, SIS
Screening<-function(X=X,Y=Y)
{
p<-dim(X)[2]
n<- length(Y)
s1<- matrix(0,nrow=p,ncol=1)
s2<- matrix(0,nrow=p,ncol=1)
s3<- matrix(0,nrow=p,ncol=1)
Y<- Y-mean(Y)
OLS<- t(X) %*% Y # for ranking based on SIS
```

```
G.y<- c() # Empirical distribution function of Y
for(i in 1:n){
G.y[i]<- sum(Y<=Y[i])
}
for(j in 1:p){
s1[j,] <- (dcor(X[,j],G.y)) # RDC-SIS
s2[j,] <- (dcor(X[,j],Y))   # DC-SIS
s3[j,] <- abs(OLS[j])       # SIS
}
rank1=(p+1)-rank(s1) # RDC-SIS
rank2=(p+1)-rank(s2) # DC-SIS
rank3=(p+1)-rank(s3) # SIS
list(rank.ROSIS=rank1,rank.DCSIS=rank2,rank.SIS=rank3)
}
### Sure Independent Ranking and Screening (SIRS)
SIRS<-function(X=X,Y=Y){
p<- dim(X)[2]
n<- length(Y)
w<- matrix(0, nrow=p, ncol=1)
X<- apply(X,2, function(x)(x-mean(x))/sd(x))
for(k in 1:p){
w.k.j<-NULL
for(j in 1:n){
s<- (t(X[,k])%*%(1*(Y<Y[j]))/n)^2
w.k.j<- c(w.k.j,s)
}
w[k,]<- (n^3*(mean(w.k.j)))/(n*(n-1)*(n-2))
}
list(rank=(p+1)-rank(w))
}
###################################
n<- 200   # number of data points
p<- 1000  # number of covariates
ss<- 500  # number of repetitions
```

```
gg<- 5    #number of important variables
c<- 2     # c=1 or 2
nm<- 4    # number of methodes (RDC-SIS,DC-SIS,SIS,SIRS)
R<- array(0,c(ss,gg,nm))
Sigma1<- diag(p) ## covariance matrix
for (i in 1:p){
for (j in 1:p){
if (i<j) {
Sigma1[i,j]<- (0.5)^abs(i-j)
Sigma1[j,i]<- Sigma1[i,j] }
}
}

### Beginning of the iteration
for( s in 1:ss)
{
set.seed(12345+s)
X<- mvrnorm(n,mu=rep(0,p),Sigma=Sigma1)
beta<- c(1,0.8,0.6,0.4,0.2,rep(0,p-gg))
# error<- rsn(n,0,1,2) # different error distributions
# error<- rt(n,1)
error<- rnorm(n)
Y<- c*X%*%beta+sqrt(6.83)*error
#####
Q<- Screening(X,Y)
ex1b<- SIRS(X,Y)
R[s,,1]<- c(Q$rank.ROSIS[1:gg])
R[s,,2]<- c(Q$rank.DCSIS[1:gg])
R[s,,3]<- c(Q$rank.SIS[1:gg])
R[s,,4]<- c(ex1b$rank[1:gg])
print(s)
}
quantile(apply(R[,,1],1,max),c(0.05,0.25,0.50,0.75,0.95)) #  ROSIS
quantile(apply(R[,,2],1,max),c(0.05,0.25,0.50,0.75,0.95)) #  DC-SIS
```

```
quantile(apply(R[,,3],1,max),c(0.05,0.25,0.50,0.75,0.95)) #  SIS
quantile(apply(R[,,4],1,max),c(0.05,0.25,0.50,0.75,0.95)) #  SIRS
###############
h1<- rep(0,nm)
for (i in 1:nm)
{
for(s in 1:ss)
{
if (max(R[s,,i])<=2*floor(n/log(n))){
h1[i]<- h1[i]+1 }
}
}
propertion1<- h1/ss
propertion1  # propotion S
#################
h2<- rep(0,gg)
for(s in 1:ss)
{
for(k in 1:gg){
if ((R[s,k,3])<=2*floor(n/log(n))){
h2[k]<- h2[k]+1}
}
}
propertion2<- h2/ss
propertion2   # propotion P_{3}
```

مثال ۱.۵.۴

```
### Load R Packages
library(glmnet)
library(abind)
library(magic)
library(mvtnorm)
library(splines)
```

```
library(nor1mix)
library(MASS)
library(SparseM)
library(quantreg)
library(hqreg)

# First order derivative of SCAD penalty
scad.der<-function (x, lambda)
{
x<- abs(x)
a<- 3.7
u<- (x<=lambda)
penalty.derivative<- lambda*u+(pmax(a*lambda-x,0)/(a-1))*(1-u)
return(penalty.derivative)
}

###### Inputs
n<- 200    # sample size
p<- 15     # number of covariates
p1<- 2     # number of constant effects
p2<- 3     # number of varying effects
s<- p1+p2  # number of nonzero effects
degree<- 3 # degree of spline
K<- floor(n^{1/(2*degree+3)}) # number of internal knots

Sigma<- diag(p) ## covariance matrix
for (i in 1:p){
for (j in 1:p){
Sigma[i,j]<- 0.5^abs(i-j)
}
}
#### simulation
nsimu=500; NV=rep(0,nsimu); NC=rep(0,nsimu)
NVT=rep(0,nsimu);NCT=rep(0,nsimu);CF=rep(0,nsimu)
```

```
for (simu in 1:nsimu)
{
set.seed(1234+simu)
X<- mvrnorm(n,mu=rep(0,p),Sigma=Sigma)
u<- runif(n)
error<- rnorm(n)   # normal error
# error<- rt(n,3)  # t(3) error
# error<- rnorMix(n,norMix(mu=c(0,0),sigma=c(1,5),w=c(0.9,0.1))) # MN error
Y<- 2*sin(2*pi*u)*X[,1]+8*u*(1-u)*X[,2]+2.5*X[,3]+X[,4]+1.5*X[,5]+error

####### culculate R matrix
q=K+degree+1 # number of B-spline basis functions
grid<- seq(0.01,0.99,length=500) # grid points
evalspline<- function(K){
B<- bs(grid,df=q,intercept=TRUE,degree=degree)
R<- t(B)%*%B/length(grid)
list(R=R)
}
temp<- evalspline(K)
R<- temp$R

######## Z matrix
Zmatrix<- function(xx,y,B,K)
{
n<- nrow(xx);p<- ncol(xx)
z<- c()
for (j in 1:p)
{
Phi<- B*xx[,j]
z<- cbind(z,Phi)
}
Z<- cbind(xx,as.matrix(z))
return(Z)
```

```
}
Z<- Zmatrix(X,Y,B,K)

####### unpenalized quantile estimation as initial estimate
obj0=function(xx,y,B,K)
{
n<- nrow(xx);p<- ncol(xx)
Z<- Zmatrix(xx,y,B,K)
theta<- rq(y~Z-1)$coef
list(theta=theta)
}
object<- obj0(X,Y,B,K)
thetainit<- object$theta  # initial estimate
etainit<- thetainit[1:p]
gammainit<- thetainit[(p+1):length(thetainit)]
pgaminit<- rep(0,p)
for (j in 1:p){
gammaj<- gammainit[((j-1)*(K+4)+1):(j*(K+4))]
pgaminit[j]<- sqrt(gammaj%*%R%*%gammaj)
}
####### Find optimal tuning parameters using 5-fold CV
cvfolds<- function(n,folds)
{ split(sample(1:n),rep(1:folds,length=n)) }
x<- X
y<- Y
nK<- 5 # number of folds
folds_ans<- cvfolds(n,nK)
mseld<-c ()

## Possible values of parameters
lambda1s<- seq(lambda1.min,lambda1.max,length=20)
lambda2s<- seq(lambda2.min,lambda2.max,length=20)
for (lambda1 in lambda1s)
{
```

```
for (lambda2 in lambda2s)
{
mse<- rep(0,nK)
for(i in 1:nK)
{
omit<- folds_ans[[i]]
xcv<- x[-omit,];ycv=y[-omit]
xt<- x[omit,];yt=y[omit]
Bcv<- B[-omit,];ucv=u[-omit]
Bt<- B[omit,];ut=u[omit]
zcv<- Z[-omit,];zt=Z[omit,]
thetam<- thetainit
iter<- 1; diff<- 1

while(diff>1e-6 & iter<50){
## Optimal h
e<- Y-Z%*%thetam; sigm<- sqrt(var(e))
rh<- rep(0,100)
for (j in 1:100)
{
h<- as.numeric(0.5*sigm*1.02^(j-1))
fh<- mean((e^2/h^2-1)/h^3*dnorm(e/h))
gh<- mean(e^2*(dnorm(e/h))^2/h^6)
rh[j]<- gh/(fh^2*sigm^2)
}
index<- which.min(rh);hopt=0.5*sigm*1.02^(index-1)
ecv<- ycv-zcv%*%thetam
w<- dnorm(ecv,mean=0,sd=hopt);W=as.vector(w/sum(w))
etam<- thetam[1:p]
gammam<- thetam[(p+1):length(thetam)]
pgam<- rep(0,p)
for (j in 1:p){
gammamj<- gammam[((j-1)*(K+4)+1):(j*(K+4))]
pgam[j]<- sqrt(gammamj%*%R%*%gammamj)
```

```
}
mpw1<- rep(0,p)
mpw2<- rep(0,p)
for(j in 1:p){
mpw1[j]<- scad.der(etam[j],lambda1)/(abs(etam[j])+1e-6)
mpw2[j]<- scad.der(pgam[j],lambda2)/(pgam[j]+1e-6)
}
omega1<- diag(mpw1)
omega2<- matrix(0,nrow=(K+4)*p,ncol=(K+4)*p)
for (j in 1:p){
omega2[((j-1)*(K+4)+1):(j*(K+4)),((j-1)*(K+4)+1):(j*(K+4))]<-mpw2[j]*R
}
Gm<- adiag(omega1,omega2)
thetam0=thetam
thetam<- ginv(t(zcv)%*%diag(W)%*%zcv+n*Gm)%*%t(zcv)%*%diag(W)%*%ycv
diff<- max(abs(thetam-thetam0))
iter<- iter+1
}
mse[i]<- sum(dnorm((yt-zt%*%thetam),mean=0,sd=hopt))
}# for nK
mseld<- c(mseld,mean(mse))
} # for lambda2
} # for lambda1

mldindex<- which.max(mseld) # index of optimal tuning parameters
lambda1opt<- lambda1s[floor((mldindex-1)/length(lambda2s)+1)]
lambda2opt<- lambda2s[(mldindex-1)%%length(lambda2s)+1]

####### estimation by optimal tuning parameters ###############
thetam<- thetainit
lambda1<- lambda1opt
lambda2<- lambda2opt

while(diff>1e-6 & iter<50){
```

```
## Find optimal h
e<- Y-Z%*%thetam; sigm=sqrt(var(e))
rh<- rep(0,100)
for (j in 1:100)
{
h<- as.numeric(0.5*sigm*1.02^(j-1))
fh<- mean((e^2/h^2-1)/h^3*dnorm(e/h))
gh<- mean(e^2*(dnorm(e/h))^2/h^6)
rh[j]<- gh/(fh^2*sigm^2)
}
index<- which.min(rh);hopt=0.5*sigm*1.02^(index-1)
w<- dnorm(e,mean=0,sd=hopt);W=as.vector(w/sum(w))
etam<- thetam[1:p]
gammam<- thetam[(p+1):length(thetam)]
pgam<- rep(0,p)
for (j in 1:p){
gammamj<- gammam[((j-1)*(K+4)+1):(j*(K+4))]
pgam[j]<- sqrt(gammamj%*%R%*%gammamj)
}
mpw1<- rep(0,p)
mpw2<- rep(0,p)
for(j in 1:p){
mpw1[j]<- scad.der(etam[j],lambda1)/(abs(etam[j])+1e-6)
mpw2[j]<-scad.der(pgam[j],lambda2)/(pgam[j]+1e-6)
}
omega1<- diag(mpw1)
omega2<- matrix(0,nrow=(K+4)*p,ncol=(K+4)*p)
for (j in 1:p){
omega2[((j-1)*(K+4)+1):(j*(K+4)),((j-1)*(K+4)+1):(j*(K+4))]<-mpw2[j]*R
}
Gm<- adiag(omega1,omega2)
thtam0<- thetam
thetam<- ginv(t(Z)%*%diag(W)%*%Z+n*Gm)%*%t(Z)%*%diag(W)%*%Y
diff<- max(abs(thetam-thetam0))
```

```
iter<- iter+1
}
etam<- thetam[1:p]
etam[abs(etam)<1e-4]<- 0 # small eta coefficients set to be zero

gammam<- thetam[(p+1):length(thetam)]
for (j in 1:p){  # small gamma coefficients set to be zero
curgammam<-gammam[((j-1)*(K+4)+1):(j*(K+4))]
if(sqrt(sum(curgammam^2))<1e-4)
{gammam[((j-1)*(K+4)+1):(j*(K+4))]<-0}
}
pgam<- rep(0,p)
for (j in 1:p){
gammamj<- gammam[((j-1)*(K+4)+1):(j*(K+4))]
pgam[j]<- sqrt(gammamj%*%R%*%gammamj)
}

############## check the varying and constant effects
allk<-rep(K,p)
for (j in 1:p){
curgammam<-gammam[((j-1)*(K+4)+1):(j*(K+4))]
if(sum(abs(curgammam))<1e-4&abs(etam[j])<1e-4){allk[j]<- -K;next;}
   # zero effect
if(sum(abs(curgammam))<1e-4&abs(etam[j])>=1e-4){allk[j]<- -2} else
   {allk[j]<-K} # constant effect
}
############################
# NV: number of varying effects
# NC: number of constant effects
NV<- sum(allk>0); NC<- sum(allk==-2)

# NVT: number of varying coefficients that are correctly identified
# NCT: number of constant coefficients that are correctly identified
u=0;v=0
```

```
for (j in 1:p1)
{ if (allk[j]>0) v=v+1 }
for (k in (p1+1):s)
{ if (allk[k]==-2) u=u+1 }
NVT<- v; NCT<- u

# CF: the proportion of correctly fitted models
r=0
{if(sum(allk[1:p1])==p1*K & sum(allk[(p1+1):s])==-(p2)*2 &
    sum(allk[(s+1):p])==-(p-s)*K) r=1}
CF[simu]<- r

### RASE
u.new<- runif(200) # new grid points
Beta_t<- matrix(0,200,p) # true coefficients
Beta_t[,1]<- (2*sin(2*pi*u.new))
Beta_t[,2]<- (8*u.new*(1-u.new))
Beta_t[,3]<- rep(2.5,200)
Beta_t[,4]<- rep(1.2,200)
Beta_t[,5]<- rep(1.5,200)
B<- bs(u.new,df=q,intercept=TRUE,degree=degree)
Beta_hat<- matrix(0,nrow=200, ncol=p)
for(j in 1:p){
gammaj<- gammam[((j-1)*(K+4)+1):(j*(K+4))]
Beta_hat[,j]<- etam[j]+B%*%gammaj
}
res<- (Beta_t-Beta_hat)^2
RASE[simu]<- mean(apply(res,1,sum))
cat("result:","number varying:", sum(allk>0)," number constant",sum(allk==-2),
"true number varying:",NNT[simu],"true number constant:",NLT[simu],"CF:",
CF[simu],"\n")
}
```

مراجع

[۱] اریه، م. (۱۳۹۴). *برآورد مدل نیمه‌پارامتری به روش Bـاسپلاین*، پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود.

[۲] روزبه، م. (۱۳۹۰). *برآورد در مدل‌های خطی جزئی*، رساله دکتری، دانشگاه فردوسی مشهد.

[۳] کاظمی، م.، شاهسونی، د. و آرشی، م. (۱۳۹۷). انتخاب متغیر و تشخیص ساختار در بعد بالا برای مدل‌های جمعی خطی‌ـجزئی، *مجله علوم آماری*، ۱۲(۲)، ۴۸۵ـ۵۱۲.

[۴] نوروزی‌راد، م. (۱۳۹۶). *برآوردگرهای بهبودیافته در برخی مدل‌های رگرسیونی خطی جریمه‌شده*، رساله دکتری، دانشگاه صنعتی شاهرود.

[5] Ahmad, I., Leelahanon, S. and Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model, *The Annals of Statistics*, 39, 305-332.

[6] Akdeniz, E., Akdeniz, F. and Roozbeh, M. (2018). A new difference-based weighted mixed Liu estimator in partially linear models, *Statistics*, 52(6), 1309-1327.

[7] Arashi, M. and Roozbeh, M. (2016). Some improved estimation strategies in high-dimensional semiparametric regression models with application to riboflavin production data, *Statistical Papers*, DOI: 10.1007/s00362-016-0843-y.

[8] Barut, E., Fan, J. and Verhasselt, A. (2016). Conditional sure independence screening, *Journal of the American Statistical Association*, 111(515), 1266-1277.

[9] Breheny, P. and Huang, J. (2019). *High-Dimensional Regression Modeling: Methodology, Applications and Software*, Chapman and Hall/CRC.

[10] Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, *Statistics and Computing*, 25(2), 173-187.

[11] Breheny, P. and Huang, J.(2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, *Annals of Applied Statistics*, 5, 232-253.

[12] Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, 24(6), 2350–2383.

[13] Bushel, P., Wolfinger, R. D. and Gibson, G. (2007). Simultaneous Clustering of Gene Expression Data with Clinical Chemistry and Pathological Evaluations Reveals Phenotypic Prototypes, *BMC Systems Biology*, Doi:10.1186/1752-0509-1-15.

[14] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications,* Springer Science & Business Media.

[15] Bühlmann, P., Kalisch, M. and Maathuis, M. H. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm, *Biometrika*, 97(2), 261-278.

[16] Chatterjee, S. and Hadi, A. S. (2012). *Regression Analysis by Example, 5th Edition*, John Wiley.

[17] Chen, X. (2018). Model-free conditional feature screening for ultra-high dimensional right censored data, *Journal of Statistical Computation and Simulation*, 88(12), 2425-2446.

[18] Chen, R., Liang, H. and Wang, J. (2011). Determination of linear components in additive models, *Journal of Nonparametric Statistics*, 23(2), 367-383.

[19] Chen, J., Li, D., Linton, O. and Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series, *Journal of the American Statistical Association*, 113(522), 919-932.

[20] Cleveland, W. S., Grosse, E. and Shyu, W. M. (1991). Local regression models, In *Statistical Models in S* (Chambers, J. M. and Hastie, T. J., eds), 309-376. Wadsworth & Brooks, Pacific Grove.

[21] De Boor, C. (2001). *A Practical Guide to Splines*, Revised Edition, Springer-Verlag, New York.

[22] Deng, G. and Liang, H. (2010). Model averaging for semiparametric additive partial linear models, *Science China Mathematics*, 53(5), 1363-1376.

[23] Desboulets, L. (2018). A Review on Variable Selection in Regression Analysis, *Econometrics*, 6(4), 45.

[24] Dierckx, P. (1995). *Curve and Surface Fitting with Splines*, Oxford University Press.

[25] Ding, J., Tarokh, V. and Yang, Y. (2018). Model selection techniques: an overview, *IEEE Signal Processing Magazine*, 35(6), 16-34.

[26] Drapper, N. R. and Smith, H. (1981). *Applied Regresion Analysis, 2nd Edition*, John Wiley.

[27] Du, J., Li, G. and Peng, H. (2015). Variable selection for semiparametric partially linear covariate-adjusted regression models, *Communication in Statistics- Theory and Methods*, 44, 2809-2826.

[28] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator, *Annals of Mathematical Statistics*, 27, 642-669.

[29] Dziak, J., Li, R. and Collins, L. (2005). *Critical Review and Comparison of Variable Selection Procedures for Linear Regression*, State College, PA: Pennsylvania State University.

[30] Efron B., Hastie T., Johnstone I. and Tibshirani, R. (2004). Least angle regression, *The Annals of Statistics*, 32, 407-451.

[31] Engle R. F., Granger C. W. J., Rice J. and Weiss A. (1986). Semiparametric estimates of the relation between weather and electricity sales, *Journal of the American Statistical Association*, 81, 310-320.

[32] Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model, *In Borrowing Strength: Theory Powering Applications– A Festschrift for Lawrence D. Brown* (pp. 70-86). Institute of Mathematical Statistics.

[33] Fan J., Feng Y. and Song R. (2011). Nonparametric independence screening in sparse ultrahigh-dimensional additive models, *Journal of the American Statistical Association*, 106(494), 544-557.

[34] Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis, *National Science Review*, 1(2), 293-314.

[35] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying coefficient partially linear models, *Bernoulli*, 11, 1031-1057.

[36] Fan J. and Li R., (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 96, 1348-1360.

[37] Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery, In *Proceedings of the International Congress of Mathematicians*, 3, 595-622.

[38] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849-911.

[39] Fan, J., Ma, Y. and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models, *Journal of the American Statistical Association*, 109(507), 1270-1284.

[40] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model, *Journal of Machine Learning Research*, 10(Sep), 2013-2038.

[41] Fan J. and Song R. (2010). Sure independence screening in generalized linear models with NP-dimensionality, *The Annals of Statistics*, 6, 3567–3604.

[42] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools, *Technometrics*, 35(2), 109-135.

[43] Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso, *Journal of Computational and Graphical Statistics*, 7(3), 397-416.

[44] Geenens, G. (2011). Curse of dimensionality and related issues in nonparametric functional regression, *Statistics Surveys*, 5, 30-43.

[45] Giraud, C. (2014). *Introduction to high-dimensional statistics*, Chapman and Hall/CRC.

[46] Green, P. J. and Silverman, B. W. (1993). Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach, CRC Press.

[47] Gujrati, D. N. (2004). *Basic Econometrics (4th Edition)*, Tata McGrow- Hill.

[48] Gunst, R. F. and Webster, J. T. (1975). Regression analysis and problems of multicollinearity, *Communications in Statistics-Theory and Methods*, 4(3), 277-292.

[49] Guo, J., Tang, M., Tian, M. and Zhu, K. (2013). Variable selection in high-dimensional partially linear additive models for composite quantile regression, *Computational Statistics and Data Analysis*, 65, 56-67.

[50] Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems, *Journal of Computational and Graphical Statistics*, 18(3), 533-550.

[51] Härdle, W. (1990). *Applied Non-parametric Regression*, Economic Society Monographs, vol, 19. Cambridge University Press.

[52] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models, *The Annals of Statistics*, 21(1), 157-178.

[53] Härdle, W., Liang, H. and Gao, J. T. (2000). *Partially Linear Models*, Heidelberg: Springer Physica- Verlag.

[54] Härdle, W. K., Muller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*, Springer.

[55] Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association*, 84(408), 986-995.

[56] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall.

[57] Hastie, T. and Tibshirani, R. (1993). *Varying-coefficient models*, *Journal of the Royal Statistical Society, Series B (Methodological)*, 55(4), 757-796.

[58] Hastie, T., R. Tibshirani and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd Edition)*, Springer.

[59] Heinze, G., Wallisch, C. and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician, *Biometrical Journal*, 60(3), 431-449.

[60] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67.

[61] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications nonorthogonal problems, *Technometrics*, 12, 69–82.

[62] Hu, T., Xia, Y., (2012). Adaptive semi-varying coefficient model selection, *Statistica Sinica*, 22, 575-599.

[63] Huang, J., Wei, F. and Ma, S. (2012a). Semiparametric regression pursuit, *Statistica Sinica* , 22, 1403-1426.

[64] Huang, J., Breheny, P. and Ma, S. (2012b). A selective review of group selection in high-dimensional models, *Statistical Science*, 27(4), 481–499.

[65] Huang, J., Horowitz, J. L. and Wei, F., (2010). Variable selection in nonparametric additive models, *The Annals of Statistics*, 38, 2282-2313.

[66] Huang, J., Wu, C., Zhou, L. (2002). Varying-coefficient models and basis function approximation for the analysis of repeated measurements, *Biometrika*, 89, 111–128.

[67] Huang, Q. and Zhu, Y. (2016). Model-free sure screening via maximum correlation, *Journal of Multivariate Analysis*, 148, 89-106.

[68] Huber, P. (1981). *Robust Estimation*, New York: John Wiley.

[69] Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms, *The Annals of Statistics*, 33(4), 1617-1642.

[70] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning (Vol. 6)*, New York: springer.

[71] Jia, J. and Yu, B. (2010). On model selection consistency of elastic net when $p \gg n$, *Statistica Sinica*, 20, 595–611.

[72] Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data, *Philosophical Transactions of the Royal Society A*, 367, 4237–4253

[73] Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection method for semiparametric varying-coefficient partially linear models, *The Annals of Statistics*, 39, 305-332.

[74] Kazemi, M., Shahsavani, D. and Arashi, M. (2019a). A sure independence screening procedure for ultra-high dimensional partially linear additive models, *Journal of Applied Statistics*, 46(8), 1385-1403.

[75] Kazemi, M., Shahsavani, D. and Arashi, M. (2018). Variable selection and structure identification for ultrahigh-dimensional partially linear additive models with application to cardiomyopathy microarray data, *Statistics, Optimization and Information Computing*, DOI: 10.19139/soic.v6i3.576.

[76] Kazemi, M., Shahsavani, D., Arashi, M. and Rodrigues, P. C. (2019b). Estimation in partial linear model with spline modal function, *Communications in Statistics - Simulation and Computation* (Revised).

[77] Kong, X. B., Liu, Z., Yao, Y. and Zhou, W. (2017). Sure screening by ranking the canonical correlations, *Test*, 26(1), 46-70.

[78] Leng, C. (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference*, 139, 2138-2146.

[79] Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection, *Statistica Sinica*, 16, 1273–1284.

[80] Li, Q. and Lin, N. (2010). The Bayesian elastic net, *Bayesian Analysis*, 5(1), 151-170.

[81] Li, Y., Li, G., Lian, H. and Tong, T. (2017). Profile forward regression screening for ultra-high dimensional semiparametric varying coefficient partially linear models, *Journal of Multivariate Analysis*, 155, 133-150.

[82] Li, X., Ma, X. and Zhang, J. (2018). Conditional quantile correlation screening procedure for ultrahigh-dimensional varying coefficient models, *Journal of Statistical Planning and Inference*, 197, 69-92.

[83] Li, J., Ray, S. and Lindsay, B. (2007). A nonparametric statistical approach to clustering via mode identification, *Journal of Machine Learning Research*, 8, 1687–1723.

[84] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning, *Journal of the American Statistical Association*, 107(499), 1129-1139.

[85] Lian, H. (2012a). Variable selection in high-dimensional partly linear additive models, *Journal of Nonparametric Statistics*, 24, 825-839.

[86] Lian, H. (2012b). Shrinkage estimation for identification of linear components in additive models, *Statistics and Probability Letters*, 82, 225-231.

[87] Lian, H. (2012c). Semiparametric estimation of additive quantile regression models by two-fold penalty, *Journal of Business and Economic Statistics*, 30(3), 337-350.

[88] Lian, H., Chen, X. and Yang, J. Y. (2012). Identification of partially linear structure in additive models with an application to gene expression prediction from sequences, *Biometrics*, 68(2), 437-445.

[89] Lian, H, Du, P, Li, Y,. and Liang, H. (2014). Partially linear structure identification in generalized additive models with NP-dimensionality, *Computational Statistics and Data Analysis*, 80, 197–208.

[90] Lian, H., Liang, H. and Rpuppert, D. (2015). Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models, *Statistica Sinica*, 591-607.

[91] Liu, J. (2016). Feature screening and variable selection for partially linear models with ultrahigh-dimensional longitudinal data, *Neurocomputing*, 195, 202-210.

[92] Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates, *Journal of the American Statistical Association*, 109(505), 266-274.

[93] Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models, *Statistica Sinica*, 21(3), 1225-1248.

[94] Lv, J., Yang, H. and Guo, C. (2017). Variable selection in partially linear additive models for modal regression, *Communications in Statistics-Simulation and Computation*, 46(7), 5646-5665.

[95] Meier, L., Van de Geer, S. and Buhlmann, P. (2009). High-dimensional additive modeling, *The Annals of Statistics*, 37(6B), 3779-3821.

[96] Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (Vol. 821), John Wiley.

[97] Nadaraya, E. A. (1964). On Estimating Regression, *Theory of Probability and Its Applications*, 9(1), 141–142.

[98] Noh, H., Keilegom, V. (2012). Efficient model selection in semivarying coefficient models, *Electronic Journal of Statistics*, 6, 2519-2534.

[99] Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling, *Journal of Computational and Graphical Statistics*, 8(4), 715-732.

[100] Park, T. and Casella, G. (2008). The bayesian lasso, *Journal of the American Statistical Association*, 103(482), 681-686.

[101] Parzen, E. (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, 33, 1065–1076

[102] Qiao, X. (2014). Variable selection using L_q penalties, *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(3), 177-184.

[103] Roozbeh, M. (2015). Shrinkage ridge estimators in semiparametric regression models, *Journal of Multivariate Analysis*, 136, 56-74.

[104] Roozbeh, M. (2016). Robust ridge estimator in restricted semiparametric regression models, *Journal of Multivariate Analysis*, 147, 127-144.

[105] Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

[106] Schmalensee, R. and Stoker, T.M. (1999). Household gasoline demand in the United States, *Econometrica*, 67, 645-662.

[107] Schumaker, L. (1981). *Splines Function: Basic Theory*, New York: John Wiley.

[108] Shen, S., Cui, J., Mei, C., Wang, C. (2014). Estimation and inference of semi-varying coefficient models with heteroscedastic errors, *Journal of Multivariate Analysis*, 124, 70-93.

[109] Song, R., Yi, F. and Zou, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models, *Statistica Sinica*, 24(4), 1735.

[110] Song, A., Ma, T., Lv, S. and Lin, C. (2018). A model-free variable selection method for reducing the number of redundant variables, *Statistics*, 52(6), 1212-1248.

[111] Stone,C. (1982). Optimal global rates of convergence for nonparametric regression, *The Annals of Statistics*, 10, 1040-1053.

[112] Sun, J., Lin, L. (2014). Local rank estimation and related test for varying-coefficient partially linear models, *Journal of Nonparametric Statistics*, 26, 187-206.

[113] Szekely, G. J. and Rizzo, M. L. (2009). Brownian Distance Covariance, *The Annals of Applied Statistics*, 3, 1233–1303.

[114] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances, *The Annals of Statistics*, 35(6), 2769-2794.

[115] Tang, Y., Wang, H., Zhu, Z., Song, X. (2012). A unified variable selection approach for varying coefficient models, *Statistica Sinica*, 22, 601-628.

[116] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, 267-288.

[117] Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3), 589-611.

[118] Wang, L., Li, H., Huang, J. Z. (2008). Variable selection in nonparametric varying coefficient models for analysis of repeated measurements, *Journal of the American Statistical Association*, 103, 1556-1569.

[119] Wang, K., Lin, L. (2016). Robust structure identification and variable selection in partial linear varying coefficient models, *Journal of Statistical Planning and Inference*, 174, 153-168.

[120] Wang, M. and Song, L. (2013). Identification for semiparametric varying coefficient partially linear models, *Statistics and Probability Letters*, 83(5), 1311-1320.

[121] Wang, H., Xia, Y. (2009). Shrinkage estimation of the varying coefficient model, *Journal of the American Statistical Association*, 104, 747-757.

[122] Wang, Z., Xue, L., Li, G., Lu, F. (2018). Spline estimator for ultra-high dimensional partially linear varying coefficient models, *Annals of the Institute of Statistical Mathematics*, 1-21.

[123] Watson, G.S. (1964). Smooth regression analysis, *Sankhya A*, 26, 359–372.

[124] Wei, C. H. and Liu, C. (2012). Statistical inference on semi-parametric partial linear additive models, *Journal of Nonparametric Statistics*, 24(4), 809-823.

[125] Willems, J. P., Saunders, J. T., Hunt, D. E. and Schorling, J. B. (1997). Prevalence of coronary heart disease risk factors among rural blacks: a community-based study, *Southern Medical Journal*, 90(8), 814-820.

[126] Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression, *The Annals of Applied Statistics*, 2(1), 224-244.

[127] Xia, X., Yang, H. and Li, J. (2016). Feature screening for generalized varying coefficient models with application to dichotomous responses, *Computational Statistics and Data Analysis*, 102, 85-97.

[128] Xia, Y., Zhang, W., Tong, H. (2004). Efficient estimation for semivarying-coefficient models, *Biometrika*, 91, 661-681.

[129] Xue, L. (2009). Consistent variable selection in additive models, *Statistica Sinica*, 1281-1296.

[130] Xue, L., Qu, A. (2012). Variable selection in high-dimensional varying coefficient models with global optimality, *Journal of Machine Learning Research*, 13, 1973-1998

[131] Yao, W., Li, L. (2014). A new regression model: modal linear regression, *Scandinavian Journal of Statistics*, 41(3), 656-671.

[132] Yao, W., Lindsay, B., Li, R. (2012). Local modal regression, *Journal of Nonparametric Statistics*, 24, 647-663.

[133] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49-67.

[134] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model, *Biometrika*, 94(1), 19-35.

[135] Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2), 143-161.

[136] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, 38(2), 894-942.

[137] Zhang, H. H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models, *Journal of the American Statistical Association*, 106(495), 1099-1112.

[138] Zhang, Y. and Li, R. (2011). Iterative conditional maximization algorithm for nonconcave penalized likelihood, In *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P Hettmansperger*, 336-351.

[139] Zhang, J., Liu, Y., Wu, Y. (2017). Correlation rank screening for ultrahigh-dimensional survival data, *Computational Statististics and Data Analysis*, 108, 121-132.

[140] Zhang, R., Zhao, W., Liu, J. (2013). Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression, *Journal of Nonparametric Statistics*, 25, 523-544.

[141] Zhao, P., Xue, L. (2009). Variable selection in semiparametric regression analysis for longitudinal data, *Annals of the Institute of Statistical Mathematics*, 64, 213-231.

[142] Zhao, P., Xue, L. (2010). Variable selection for semiparametric varying coefficient partially linear errors-in-variables models, *Journal of Multivariate Analysis*, 101, 1872-1883.

[143] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso, *Journal of Machine Learning Research*, 7(Nov), 2541-2563.

[144] Zhao, W., Zhang, R., Liu, J., Lv. Y. (2014). Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression, *Annals of the Institute of Statistical Mathematics*, 66, 165-191.

[145] Zhao, W., Zhang, R., Lv, Y., Liu, J. (2013). Variable selection of the quantile varying coefficient regression models, *Journal of the Korean Statistical Society*, 42(3), 343-358.

[146] Zhou, Y. and Liang, H. (2009). Statistical inference for semiparametric varying-coefficient partially linear models with error-prone linear covariates, *The Annals of Statistics*, 37(1), 427-458.

[147] Zhu, L.P., Li, L., Li, R., Zhu, L.X., (2011). Model-free feature screening for ultra-high dimensional data, *Journal of the American Statistical Association*, 106, 1464-1475.

[148] Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, 101(476), 1418-1429.

[149] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

[150] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models, *The Annals of Statistics*, 36(4), 1509-1533.

واژه‌نامه فارسی به انگلیسی

اثر متقابل ..... Interaction
ارزیابی ..... Evaluation
اریبی ..... Bias
اسپلاین رگرسیونی تطبیقی چند متغیره ..... Multivariate adaptive regression splines
اسپلاین مکعبی ..... Cubic spline
اعتبارسنجی متقابل ..... Cross validation
انعطاف‌پذیری ..... Flexibility
باقیمانده ..... Residual
بدشرطیدگی ..... Ill-Condition
برآورد انقباضی ..... Shrinkage estimation
برآوردگر پیشگو ..... Oracle estimator
بهینه ..... Optimal
بیش‌برازش ..... Overfitting
پارامتر جریمه ..... Penalty parameter
پایه ..... Basis
پذیره ..... Assumption
پهنای باند ..... Bandwidth
پیشگویی ..... Prediction
پیوسته ..... Continuous
تابع چگالی ..... Density function
تابع مشتق‌پذیر ..... Differentiable function
تجزیه چولسکی ..... Cholesky decomposition
تشخیص ساختار ..... Structure identification
تصویرسازی متعامد ..... Orthogonal projection
تفسیرپذیری ..... Interpretability

تقریب درجه دو موضعی ..... Locally quadratic approximation
تک‌مدی ..... Unimodal
توزیع دم‌سنگین ..... Heavy-tailed distribution
توزیع مجانبی ..... Asymptotic distribution
تنکی ..... Sparsity
چندجمله‌ای تکه‌ای ..... Piecewise polynomial
داده‌های آزمایشی ..... Test dataset
داده‌های آموزشی ..... Training dataset
داده پرت ..... Outlier
داده‌های با بعد بالا ..... High-dimensional data
داده‌های با بعد بسیار بالا ..... Uultrahigh-dimensional data
درآمد سرمایه انسانی ..... Human capital earnings
رتبه کامل ..... Full rank
رگرسیون بریج ..... Bridge regression
رگرسیون جریمه‌شده ..... Penalized regression
رگرسیون چندکی ..... Quantile regression
رگرسیون ریج ..... Ridge regression
رگرسیون لاسو ..... Lasso regression
رگرسیون موضعی ..... Local regression
رگرسیون مؤلفه‌های اصلی ..... Principal component regression
رگرسیون نما ..... Modal regression
رویه ..... Surface
ریزآرایه ..... Microarray
سازگاری ..... Consistency
شناساپذیری ..... Identifiability
ضریب تعیین ..... Coefficient of determination
عامل تورم واریانس ..... Variance inflation factor
غربالگری ..... Screening
کارایی ..... Efficiency
گام به گام ..... Stepwise
گره داخلی ..... Internal knot
ماتریس قطری ..... Diagonal matrix

متعامدسازی گرام‌ـاشمیت ..... Gram-Schmidt orthonormalization
متغیر پاسخ ..... Response variable
متغیر توضیحی ..... Explanatory Variable
متغیر طبقه‌ای ..... Categorical variable
متغیر ظاهری ..... Dummy variable
متغیر مهم ..... Important variable
متقارن ..... Symmetric
مجانبی ..... Asymptotic
محدب ..... Convex
مدل پارامتری ..... Parametric model
مدل تک‌شاخص ..... Single index model
مدل جمعی ..... Additive model
مدل خطی‌ـجزئی ..... Partial linear model
مدل ضریب متغیر ..... Varying coefficient model
مدل ناپارامتری ..... Non-parametric model
مدل نیمه‌پارامتری ..... Semiparametric model
مشقت بعد ..... Curse of dimensionality
مطالعات پیوند ژنتیکی ..... Genetic association studies
مقدار آستانه ..... Threshold
مقدار اولیه ..... Initial value
مقدار درست ..... True value
مقعر ..... Concave
مقعر مینیماکس ..... Minimax concave
ناااریب ..... Unbiased
نرخ همگرایی ..... Convergence rate
نیرومند ..... Robust
نیمرخ ..... Profile
ویژگی غربالگری مطمئن ..... Sure screening property
هسته ..... Kernel
همبستگی فاصله‌ای ..... Distance correlation
همخطی ..... Collinearity
همخطی کامل ..... Exact collinearity

همخطی ناقص ..... Partial collinearity
هموار ..... Smooth
هموارساز هسته ..... Kernel smoothing

واژه‌نامه انگلیسی به فارسی

پذیره ..... Assumption
مجانبی ..... Asymptotic
توزیع مجانبی ..... Asymptotic distribution
مدل جمعی ..... Additive model
پهنای باند ..... Bandwidth
پایه ..... Basis
اریبی ..... Bias
رگرسیون بریج ..... Bridge regression
متغیر طبقه‌ای ..... Categorical variable
تجزیه چولسکی ..... Cholesky decomposition
همخطی ..... Collinearity
مقعر ..... Concave
سازگاری ..... Consistency
پیوسته ..... Continuous
محدب ..... Convex
مشقت بعد ..... Curse of dimensionality
نرخ همگرایی ..... Convergence rate
اعتبارسنجی متقابل ..... Cross validation
ضریب تعیین ..... Coefficient of determination
اسپلاین مکعبی ..... Cubic spline
تابع چگالی ..... Density function
ماتریس قطری ..... Diagonal matrix
تابع مشتق‌پذیر ..... Differentiable function
همبستگی فاصله‌ای ..... Distance correlation
متغیر ظاهری ..... Dummy variable

کارایی ..... Efficiency
همخطی کامل ..... Exact collinearity
متغیر توضیحی ..... Explanatory Variable
ارزیابی ..... Evaluation
انعطاف‌پذیری ..... Flexibility
رتبه کامل ..... Full rank
مطالعات پیوند ژنتیکی ..... Genetic association studies
متعامدسازی گرام_اشمیت ..... Gram-Schmidt orthonormalization
توزیع دم‌سنگین ..... Heavy-tailed distribution
داده‌های با بعد بالا ..... High-dimensional data
درآمد سرمایه انسانی ..... Human capital earnings
بدشرطیدگی ..... Ill-Condition
متغیر مهم ..... Important variable
شناساپذیری ..... Identifiability
اثر متقابل ..... Interaction
گره داخلی ..... Internal knot
تفسیرپذیری ..... Interpretability
مقدار اولیه ..... Initial value
هسته ..... Kernel
هموارساز هسته ..... Kernel smoothing
رگرسیون لاسو ..... Lasso regression
تقریب درجه دو موضعی ..... Locally quadratic approximation
رگرسیون موضعی ..... Local regression
ریزآرایه ..... Microarray
مقعر مینیماکس ..... Minimax concave
اسپلاین رگرسیونی تطبیقی چند متغیره ..... Multivariate adaptive regression splines
رگرسیون نما ..... Modal regression
مدل ناپارامتری ..... Non-parametric model
تصویرسازی متعامد ..... Orthogonal projection
بیش‌برازش ..... Overfitting
داده پرت ..... Outlier
برآوردگر پیشگو ..... Oracle estimator

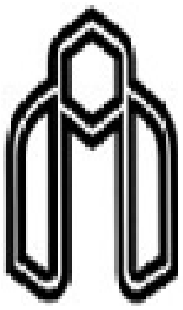
بهینه ..... Optimal
مدل پارامتری ..... Parametric model
همخطی ناقص ..... Partial collinearity
رگرسیون کمترین توان‌های دوم جزیی ..... Partial least squares regression
مدل خطی‌ـ‌جزئی ..... Partial linear model
پارامتر جریمه ..... Penalty parameter
رگرسیون جریمه‌شده ..... Penalized regression
چندجمله‌ای تکه‌ای ..... Piecewise polynomial
پیشگویی ..... Prediction
رگرسیون مؤلفه‌های اصلی ..... Principal component regression
نیمرخ ..... Profile
رگرسیون چندکی ..... Quantile regression
رگرسیون ریج ..... Ridge regression
باقیمانده ..... Residual
نیرومند ..... Robust
متغیر پاسخ ..... Response variable
متقارن ..... Symmetric
برآورد انقباضی ..... Shrinkage estimation
مدل نیمه‌پارامتری ..... Semiparametric model
گام به گام ..... Stepwise
تشخیص ساختار ..... Structure identification
ویژگی غربالگری مطمئن ..... Sure screening property
غربالگری ..... Screening
هموار ..... Smooth
تنکی ..... Sparsity
رویه ..... Surface
مدل تک‌شاخص ..... Single index model
داده‌های آزمایشی ..... Test dataset
مقدار آستانه ..... Threshold
داده‌های آموزشی ..... Training dataset
مقدار درست ..... True value
نااریب ..... Unbiased

تک‌مدی ..... Unimodal
داده‌های با بعد بسیار بالا ..... Uultrahigh-dimensional data
مدل ضریب متغیر ..... Varying coefficient model
عامل تورم واریانس ..... Variance inflation factor

Abstract

Semiparametric models are often used in real data analysis for their flexibility and parsimony. Two well-known examples of semiparametric models are the partially linear additive model and the partially linear varying coefficient model. Statistical inference of these models is restricted to a condition that the parametric and nonparametric parts are known in advance. However, such prior information is usually unavailable, especially when the number of covariates is large. Therefore, it is of great interest to develop some efficient methods to distinguish parametric components from nonparametric ones. In this dissertation, we introduce a two-step procedure, in the context of ultra-high dimensional additive models, which aims to reduce the size of covariates vector and distinguish linear and nonlinear effects among nonzero components. Also, we propose a robust method for simultaneously variable selection and parametric component identification in varying coefficient models based on modal regression, which is robust with respect to non-normal errors and outliers in the response. The performance of the two proposed methods is examined by simulation studies and real data analysis. Results of numerical studies demonstrate the superiority in comparison with the existing methods.

Keywords: Dimensionality reduction, Modal regression, Partially linear additive model, Partially linear varying coefficient model, Structure identification, Sure independence screening, Variable selection.



Shahrood University of Technology

Faculty of Mathematical Sciences

Ph.D. Thesis in Statistics

Variable Selection and Structure Identification in Semiparametric Models

By: Mohammad Kazemi

Supervisors

Dr. Davood Shahsavani
Dr. Mohammad Arashi

February 2019