مجله علوم آماری، پاییز و زمستان ۱۳۹۷
جلد ۱۲، شماره ۲، ص ۴۸۵ – ۵۱۲
DOI: 10.29252/jss.12.2.485

انتخاب متغیر و تشخیص ساختار در بعد بالا برای مدل‌های جمعی خطی-جزیی

محمد کاظمی، داود شاهسونی، محمد آرشی

گروه آمار، دانشگاه صنعتی شاهرود

تاریخ دریافت: ۱۳۹۶/۱۰/۰۷ تاریخ آخرین بازنگری: ۱۳۹۷/۰۴/۲۹

چکیده: در این مقاله یک روش دو مرحله‌ای برای انتخاب متغیر و تشخیص مؤلفه‌های خطی و غیرخطی در مدل‌های جمعی با بعد بالا معرفی می‌شود. در مرحله اول، از یک روش غربالگری برای کاهش بعد فضای متغیرها 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 تاوان برای انتخاب همزمان مؤلفه‌های غیرصفر و خطی استفاده می‌شود. کارایی این روش دو مرحله‌ای با مطالعه شبیه‌سازی و تحلیل یک مجموعه داده واقعی بررسی شده است.

واژه‌های کلیدی: انتخاب متغیر، تشخیص ساختار، غربالگری، کاهش بعد، مدل جمعی خطی-جزیی.

۱ مقدمه

مدل جمعی ناپارامتری را به‌صورت

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

در نظر بگیرید، که در آن Y_i متغیر پاسخ، $X_i = (X_{i1}, \ldots, X_{ip})^T$ بردار متغیرهای توضیحی و f_j ها توابع یک متغیره هموار نامعلوم هستند. منظور از تابع هموار، تابعی است که بی‌نهایت بار مشتق‌پذیر باشد.

آدرس الکترونیکی نویسنده مسئول مقاله: محمد کاظمی، m.kazemie64@yahoo.com

کد موضوع‌بندی ریاضی (۲۰۱۰): 62J07، 62G05، 62G08

همچنین $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ بردار خطای تصادفی با میانگین صفر و $E(\varepsilon\varepsilon^T) = \sigma^2 I_n$ 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 به صورت غیرخطی وارد مدل شوند، اپسومر و راپرت (۱۹۹۹) مدل جمعی خطی-جزیی را به صورت

$$Y_i = \sum_{j \epsilon S_1} \beta_j X_{ij} + \sum_{j \epsilon S_2} f_j(X_{ij}) + \varepsilon_i, \qquad (۲)$$

معرفی کردند، که در آن S_1 و S_2 دو زیر مجموعه مکمل و دو به دو ناسازگار از $\{1, ..., p\}$ است. این مدل از دو ویژگی مثبت تفسیرپذیری مدل‌های خطی و انعطاف‌پذیری مدل‌های ناپارامتری استفاده می‌کند. از نظر آماری، مؤلفه‌های پارامتری دارای همگرایی سریع‌تر از مؤلفه‌های ناپارامتری هستند. در نتیجه در نظر گرفتن مؤلفه‌های خطی به صورت غیرخطی کارآیی برآورد را کاهش می‌دهد. بنابراین، استفاده از مدل‌های جمعی خطی-جزیی در برخی از مسائل، نسبت به مدل‌های جمعی ناپارامتری، مناسب‌تر است. مسئله برآورد و انتخاب متغیر در مدل جمعی خطی-جزیی تاکنون توسط محققین زیادی مورد مطالعه قرار گرفته است. از جمله می‌توان به لیو و همکاران (۲۰۱۱)، لیان (۲۰۱۲a)، گو و همکاران (۲۰۱۳)، دو و همکاران (۲۰۱۵) و لیو و همکاران (۲۰۱۶) اشاره کرد.

مهمترین فرض در استفاده از مدل (۲) این است که اجزای خطی و غیرخطی مدل از پیش تعیین شده‌اند. اما در عمل چنین اطلاعات پیشین به ندرت در دسترس است، به ویژه وقتی که تعداد متغیرهای توضیحی زیاد باشد، یا به عبارتی مدلی با بعد بالا باشد. بنابراین، علاوه بر تشخیص مؤلفه‌های غیرصفر، شناسایی مؤلفه‌های خطی و غیرخطی نیز بسیار حائز اهمیت است.

ژانگ و همکاران (۲۰۱۱) از دو تابع تاوان برای شناسایی همزمان مؤلفه‌های خطی و غیرخطی در مدل جمعی خطی-جزیی استفاده کردند، اما نتوانستند ویژگی سازگاری در انتخاب را برای روش ارائه شده ثابت کنند. لذا، هوآنگ و همکاران (۲۰۱۲) یک روش موسوم به «تعقیب رگرسیون نیمه پارامتری[1]» را برای تشخیص مؤلفه‌های خطی و غیرخطی با استفاده از تابع تاوان محدب مینیماکس[2] (MCP) ارائه دادند. لیان (۲۰۱۲b) روشی برای تشخیص مؤلفه‌های خطی با استفاده از تابع تاوان SCAD[3] معرفی کرد. سپس لیان و همکاران (۲۰۱۲c) با به کاربردن دو تابع تاوان SCAD توانستند مؤلفه‌های خطی و

[1] Semiparametric regression pursuit
[2] Minimax Concave Penalty
[3] Smoothly Clipped Absolute Deviation

غیرصفر را به طور همزمان تشخیص دهند. همچنین لیان و همکاران (۲۰۱۵) با استفاده از دو تابع تاوان LASSO[۴] گروهی تطبیقی به شناسایی مؤلفه‌های خطی و غیرصفر در یک مدل جمعی خطی-جزیی با بعد بالا [۵] پرداختند. در داده‌های با بعد بالا، تعداد متغیرهای توضیحی، p، می‌تواند بزرگ‌تر از حجم نمونه، n، باشد ($p > n$). در این نوع داده‌ها، معمولاً فقط تعداد اندکی از متغیرهای توضیحی واقعاً با متغیر پاسخ مرتبط هستند. به دلیل وجود تعداد زیادی از متغیرها در مدل‌های با بعد بالا، تفسیر این مدل‌ها بسیار مشکل است. لذا مسئله انتخاب متغیر نقش بسیار مهمی را در مدل‌سازی آماری با بعد بالا ایفا می‌کند.

موضوع بعد بالای داده‌ها، چالشی اجتناب ناپذیر در مسئله مدل‌سازی است که مرهون ارتقای فناوری تولید داده‌های بزرگ در علوم مختلف، از جمله ژنتیک و پزشکی و همچنین افزایش توان ذخیره‌سازی کامپیوترها است. فراتر از آن، داده‌های با «بعد بسیار بالا» [۶] فرصتی پدید آورده‌اند که روش‌های نوین انتخاب متغیر آماری و یادگیری ماشین بتوانند هنگامی که p بسیار بزرگتر از n است ($p >> n$) ، نقش خود را بطور مؤثر ایفا کنند. در این نوع داده‌ها p یک تابع نمایی از n است، به عبارت دیگر $\log(p) = O(n^{\alpha})$، که در آن $\alpha > 0$. در این خصوص، مطالعات انجام شده بر اساس توابع تاوان که در بالا بدان اشاره شد، به دلایل هزینه محاسباتی، دقت آماری و ناپایداری الگوریتمی از کارآیی کافی برخوردار نیستند و 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 راستا، قالب کلی راه حل موجود مبتنی بر رهیافت توابع تاوان، یک روش دو مرحله‌ای است که در مرحله اول ابتدا بعد مدل توسط یک روش غربالگری مستقل کارآ کاهش یافته و در مرحله دوم از روشی مبتنی بر دو تابع تاوان برای شناسایی مؤلفه‌های غیرصفر و خطی در زیر مدل غربال شده استفاده می‌شود. البته لازم به ذکر است که رهیافت استفاده از تابع تاوان، تنها راه موجود برای غربالگری و انتخاب متغیر نیست. به عنوان مثال، چانگ (۲۰۱۰) در قالب یک فن یادگیری ماشین به نام SLiM، رده‌بندی داده‌های بیان ژنی با بعد بالا را مورد بررسی قرار داد. همچنین غربالگری متغیرها در بعد بسیار بالا توسط وانگ (۲۰۰۹) و از منظر رگرسیون پیش رو مورد مطالعه قرار گرفت. فراتی و هال (۲۰۱۵) نیز به نقش آفرینی روش ناپارامتری رگرسیون موضعی در غربالگری و انتخاب متغیرها پرداختند. مسئله انتخاب متغیر در داده‌های با بعد بالا، از دیدگاه بیزی و توسط فرآیند پواسون – دیریکله نیز مورد بحث قرار گرفت (گوها و بلادان‌دیوزاپانی، ۲۰۱۶).

---

[4] Least Absolute Shrinkage and Selection Operator

[5] High dimensional

[6] Ultrahigh dimensional

در رابطه با مرحله اول روش دو مرحله‌ای مذکور، فن و لیو (۲۰۰۸) روش غربالگری مستقل مطمئن [۷] (SIS) را با استفاده از همبستگی پیرسن برای مدل‌های خطی با بعد بسیار بالا معرفی کردند. پس از آن، روش SIS به مدل‌های آماری مختلف مانند مدل خطی تعمیم‌یافته (فن و همکاران ۲۰۰۹، فن و سانگ ۲۰۱۰) و مدل جمعی ناپارامتری (فن و همکاران ۲۰۱۱، کاظمی و همکاران ۲۰۱۷) تعمیم داده شد. این روش‌ها مدل مبنا بوده و در صورت نادرست بودن مدل آماری مفروض، ممکن است دارای عملکرد مناسبی نباشند. بنابراین برای اجتناب از تشخیص نادرست ساختار مدل، معرفی یک روش غربالگری آزاد - مدل، که به نوع مدل بستگی نداشته باشد، ضروری به نظر می‌رسید. لذا، ژو و همکاران (۲۰۱۱) یک روش غربالگری آزاد- مدل به نام غربالگری و رتبه‌بندی مستقل مطمئن [۸] (SIRS) را معرفی کردند. این روش را می‌توان برای بسیاری از مدل‌های پارامتری و نیمه‌پارامتری مانند مدل خطی، مدل خطی تعمیم یافته، مدل شاخص و بسیاری از مدل‌های متداول دیگر به‌کار برد.

با توجه به اهمیت روش‌های غربالگری آزاد-مدل، یک روش دیگر موسوم به غربالگری مستقل مطمئن براساس همبستگی فاصله‌ای [۹] (DC-SIS) توسط لی و همکاران (۲۰۱۲) براساس همبستگی فاصله‌ای بین متغیرهای توضیحی و متغیر پاسخ ارائه شد. آنها نشان دادند که این روش دارای ویژگی غربالگری مطمئن [۱۰] است، یعنی با احتمال نزدیک به یک متغیرهای مهم را برای ورود به مدل انتخاب می‌کند. روش DC-SIS دارای چندین مزیت است: الف. برای پیاده سازی 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 این مقاله ضمن رعایت الگوی کلی و متعارف دو مرحله‌ای ذکر شده، ابتدا از یک روش غربالگری مستقل مطمئن بر اساس همبستگی فاصله‌ای بین متغیرهای توضیحی و تابع توزیع حاشیه‌ای متغیر پاسخ برای کاهش بعد مدل استفاده می‌شود. این روش آزاد - مدل بوده و انتظار می‌رود وقتی که متغیر پاسخ دم سنگین، چوله یا دارای مقادیر فرین باشد، عملکرد خوبی داشته باشد. این روش غربالگری تعمیمی از روش نااستوار لی و همکاران (۲۰۱۲) است که به جای Y از $F(Y)$، تابع توزیع حاشیه‌ای متغیر پاسخ، استفاده می‌شود. سپس دو تابع تاوان SCAD گروهی را برای تشخیص همزمان متغیرهای مهم و مؤلفه‌های خطی در زیر مدل غربال شده در مرحله قبل به‌کار برده می‌شود.

[7] Sure independence screening
[8] Sure independence ranking and screening
[9] Sure independence screening based on the distance correlation
[10] Sure screening property

در بخش ۲ تعمیمی از روش لی و همکاران (۲۰۱۲) برای کاهش بعد معرفی می شود. در بخش ۳ روش انتخاب متغیر و تشخیص ساختار با استفاده از دو تابع تاوان با جزئیات ارائه می‌شود. در بخش ۴ در مطالعه‌ای شبیه‌سازی به بررسی عملکرد روش معرفی‌شده پرداخته می‌شود. در بخش ۵ تحلیل یک مجموعه داده واقعی ارائه می‌شود. بحث و نتیجه‌گیری کلی بخش پایانی مقاله حاضر است.

۲ غربالگری مستقل مطمئن نیرومند بر اساس همبستگی فاصله‌ای

در این مقاله مدل (۲) بررسی می‌شود، اما از آنجا که در مسائل واقعی مؤلفه‌های خطی و غیرخطی مشخص نیستند، ابتدا فرض می‌شود که مدل به صورت جمعی ناپارامتری (۱) است، سپس روشی ارائه خواهد شد تا مؤلفه‌های خطی در مدل (۱) تعیین شوند و صورت کلی رابطه (۲) بدست آید. برای این منظور، ابتدا با استفاده از یک روش غربالگری نیرومند به کاهش بعد فضای متغیرها به یک بعد کوچکتر از n پرداخته می‌شود. این روش غربالگری براساس همبستگی فاصله‌ای بین متغیرهای توضیحی و تابع توزیع حاشیه‌ای متغیر پاسخ تعریف می شود. 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 روش‌های تاوان برای تشخیص اجزای خطی و غیرخطی در مدل جمعی ناپارامتری با بعد بسیار بالا مستقیماً قابل استفاده نیستند. لازم است ابتدا با یک روش غربالگری بعد مدل را کاهش داده، سپس از توابع تاوان برای تشخیص ساختار مدل استفاده شود.

سزکلی و همکاران (۲۰۰۷) همبستگی فاصله‌ای را به عنوان یک معیار وابستگی بین دو بردار تصادفی معرفی کردند. همبستگی فاصله‌ای بین بردارهای تصادفی $U \in \mathbb{R}^q$ و $V \in \mathbb{R}^r$، با گشتاورهای مرتبه اول متناهی، یک عدد نامنفی به صورت

$$dcorr(U,V) = \frac{dcov(U,V)}{\sqrt{dcov(U,U)dcov(V,V)}}$$

تعریف می‌شود، که در آن $dcov(U,V)$ کوواریانس فاصله‌ای U و V است و ثابت کردند

$$dcov^{۲}(U,V) = S_{۱} + S_{۲} - ۲S_{۳}$$

که در آن

$$S_1 = E\{\| U - \tilde{U} \|_q \| V - \tilde{V} \|_r\},$$
$$S_2 = E\{\| U - \tilde{U} \|_q\} E\{\| V - \tilde{V} \|_r\},$$
$$S_3 = E\{E(\| U - \tilde{U} \|_q | U) E(\| V - \tilde{V} \|_r | V)\},$$

و $\| \cdot \|$ نرم اقلیدسی و $(\tilde{U}, \tilde{V})$ بردارهای تصادفی مستقل از (U, V) و هم‌توزیع با آنها هستند. سزکلی و همکاران (۲۰۰۷) نشان دادند که U و V مستقل‌اند اگر و تنها اگر $dcorr(U, V) = 0$. همچنین $dcorr(U, V)$ تابعی اکیدا صعودی از قدر مطلق همبستگی پیرسن بین U و V است. به دلیل این دو ویژگی، لی و همکاران (۲۰۱۲) یک روش غربالگری مستقل مطمئن موسوم به DC-SIS را برای رتبه‌بندی متغیرهای توضیحی با استفاده از همبستگی فاصله‌ای آنها با متغیر پاسخ ارائه دادند. آنها همچنین نشان دادند که این روش دارای ویژگی غربالگری مطمئن است.

در این مقاله 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 به صورت

$$\omega_k = dcorr\big(X_k, F(Y)\big), \qquad k = 1, \ldots, p, \tag{۳}$$

تعریف می‌شود، که در آن $F(y)$ تابع توزیع حاشیه‌ای Y است. این روش غربالگری نسبت به روش‌های موجود دارای دو مزیت است: الف. با توجه به ویژگی‌های همبستگی فاصله‌ای، X_k و $F(Y)$ مستقل‌اند، اگر و تنها اگر $dcorr\big(X_k, F(Y)\big) = 0$. 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) براساس یک نمونه تصادفی برآورد شود. فرض کنید $\big(X_i, Y_i\big)_{i=1}^n$ یک نمونه تصادفی از مدل جمعی ناپارامتری (۱) باشد.

مقادیر $S_۳, S_۲, S_۱$ با روش گشتاوری به صورت

$$\hat{S}_{k,۱} = \frac{۱}{n^۲}\sum_{i=۱}^{n}\sum_{j=۱}^{n} \parallel X_{ik} - X_{jk} \parallel_q \parallel F_n(Y_i) - F_n(Y_j) \parallel_r,$$
$$\hat{S}_{k,۲} = \frac{۱}{n^۲}\sum_{i=۱}^{n}\sum_{j=۱}^{n} \parallel X_{ik} - X_{jk} \parallel_q \frac{۱}{n^۲}\sum_{i=۱}^{n}\sum_{j=۱}^{n} \parallel F_n(Y_i) - F_n(Y_j) \parallel_r,$$
$$\hat{S}_{k,۳} = \frac{۱}{n^۳}\sum_{i=۱}^{n}\sum_{j=۱}^{n}\sum_{l=۱}^{n} \parallel X_{ik} - X_{lk} \parallel_q \parallel F_n(Y_j) - F_n(Y_l) \parallel_r,$$

برآورد می‌شود، که در آن $F_n(y) = \frac{۱}{n}\sum_{i=۱}^{n} I(Y_i \leq y)$ تابع توزیع تجربی Y است. بنابراین، برآورد $dcov^۲\big(X_k, F(Y)\big)$ به صورت

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

است. همچنین کوواریانس‌های فاصله‌ای نمونه‌ای $dcov\big(X_k, X_k\big)$ و $dcov\big(F(Y), F(Y)\big)$ را نیز می‌توان به طور مشابه برآورد نمود. در نتیجه همبستگی فاصله‌ای نمونه‌ای بین X_k و $F(Y)$ برابر است با

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

بنابراین $\hat{\omega_k} = \widehat{dcorr}\big(X_k, F(Y)\big)$. برای کاهش بعد فضای متغیرها، مجموعه‌ای از متغیرهای توضیحی را که دارای مقادیر بزرگ $\hat{\omega_k}$ هستند، به عنوان مجموعه متغیرهای مهم انتخاب می‌شوند. فرض کنید $\mathbb{M}$ مجموعه اندیس متغیرهای مهم باشد، در نتیجه

$$\hat{\mathbb{M}}_{\nu_n} = \{۱ \leq k \leq p : \hat{\omega}_k \geq \nu_n\}, \qquad (۴)$$

که در آن ν_n یک عدد مثبت از پیش تعیین شده است و در زیربخش ۱.۲ به نحوه انتخاب آن پرداخته خواهد شد. این روش بعد فضای متغیرها را از p به یک فضای بسیار کوچکتر با اندازه $d = |\hat{\mathbb{M}}_{\nu_n}|$ کاهش می‌دهد. این روش پیشنهادی، غربالگری مستقل مطمئن نیرومند براساس همبستگی فاصله‌ای نامیده می‌شود. در ادامه برای سادگی، از نمایش اختصاری RDC-SIS استفاده می‌شود.

۱.۲ قانون آستانه

استفاده از روش غربالگری RDC-SIS مستلزم تعیین یک مقدار آستانه [11] معقول ν_n است که در عمل، تعیین مقدار آن معمولاً مشکل است. در راستای یافتن راهکاری مناسب برای این مسئله، یک روش جایگزین، انتخاب d 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 $d_1 = [n/\log(n)]$، $d_2 = 2[n/\log(n)]$ یا $d_3 = 3[n/\log(n)]$ را به عنوان یک مقدار مناسب پیشنهاد دادند. این مقادیر ممکن است در برخی شرایط خوب عمل کنند اما دارای دو عیب عمده‌اند:
الف. هنوز بطور واضح مقدار دقیق d مشخص نیست. برای یک مجموعه داده واقعی، دقیقاً مشخص نیست کدام‌یک از مقادیر d_1، d_2، d_3، یا حتی یک مقدار بزرگتر باید استفاده شود.
ب. فرمول $d = [n/\log(n)]$ تنها به اندازه نمونه، n، وابسته است و تعداد متغیرهای توضیحی را نادیده می‌گیرد. در این راستا، ژائو و لی (۲۰۱۲) روشی را برای انتخاب d در مدل کاکس معرفی کردند، اما روش آنها صرفاً برای روش‌های غربالگری مدل-مبنا قابل استفاده است.

ژو و همکاران (۲۰۱۱) با افزودن q متغیر کمکی به مجموعه داده‌ها، یک روش دیگر برای تعیین d در روش غربالگری SIRS پیشنهاد دادند. مشکل عمده این روش، انتخاب تعداد متغیرهای کمکی، q، است. آنها بطور تجربی مقدار $q = p$ را انتخاب کردند و در مطالعه‌ای شبیه‌سازی نشان دادند که این مقدار q مناسب است.

در عمل باید تعیین شود کدام یک از دو روش فوق برای انتخاب d مناسب‌تر است. ژو و همکاران (۲۰۱۱) نشان دادند وقتی مدل واقعی بسیار تنک است، یا به عبارتی دیگر تعداد متغیرهای توضیحی مهم بسیار اندک است، روش فن و لیو (۲۰۰۸) به روش ژو و همکاران (۲۰۱۱)برتری دارد، اما وقتی تعداد متغیرهای با اهمیت زیاد است، روش ژو و همکاران (۲۰۱۱) نسبت به روش فن و لیو (۲۰۰۸) عملکرد بهتری دارد.

۲.۲ غربال‌گری مستقل مطمئن تکراری

روش‌های غربالگری مستقل با مطلوبیت حاشیه‌ای، تنها از اطلاعات حاشیه‌ای متغیرها به جای مدل کامل استفاده می‌کنند. لذا دو مسئله مهم ممکن است عملکرد این روش‌ها را با مشکل مواجه کند:
الف. در این روش‌ها، برخی از متغیرهای بی اهمیت که همبستگی بالایی با متغیرهای با اهمیت دارند،

[11] Threshold

نسبت به سایر متغیرهای با اهمیت که همبستگی ضعیفی با متغیر پاسخ دارند، ارجحیت دارند.
ب. متغیری که به صورت حاشیه‌ای با متغیر پاسخ ناهمبسته اما به صورت توأم و از طریق سایر متغیرها با متغیر پاسخ همبسته است، توسط این روش‌ها انتخاب نمی‌شود. فن و لیو (۲۰۰۸) نشان دادند که در صورت وجود دو مشکل فوق، با به‌کاربردن SIS ممکن است برخی متغیرهای مهم از دست داده شوند. آنها برای رفع مشکل و افزایش کارایی انتخاب متغیر، روش SIS را به صورت مکرر برای مدل‌های خطی به‌کار گرفتند و این روش را ISIS نامگذاری کردند (فن و لیو، ۲۰۰۸). تاثیر روش‌های SIS و ISIS به فرض خطی بودن مدل بستگی دارد، لذا ایده بهبود SIS توسط ISIS را نمی‌توان مستقیماً به روش غربالگری آزاد-مدل RDC-SIS تعمیم داد، مگر اینکه یک مدل مفروض برای Y و X در نظر گرفته شود. ارائه یک روش تکراری برای افزایش کارایی RDC-SIS با یک مثال شروع می‌شود. بدین منظور، مدل

$$Y = ۵X_۱ + ۵X_۲ + ۵X_۳ - ۱۵\sqrt{\rho}X_۴ + \varepsilon \qquad (۵)$$

را در نظر بگیرید که هر کدام از متغیرهای توضیحی آن از توزیع $N(۰,۱)$ تولید می‌شوند. در اینجا فرض می‌شود که به استثنای $X_۴$، ضریب همبستگی هر متغیر با سایر متغیرها یکسان و برابر با $\rho \neq ۰$ است و $X_۴$ دارای همبستگی $\sqrt{\rho}$ با $p-۱$ متغیر دیگر است. در این مثال، متغیر $X_۴$ بطور توأم مهم اما بطور 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_k روی $(X_۱, X_۲, X_۳)$ به صورت خطی برای $k = ۴, ..., p$ است. باقیمانده‌های بدست آمده از این رگرسیون‌های خطی با $(X_۱, X_۲, X_۳)$ ناهمبسته هستند. سپس روش RDC-SIS برای این مانده‌ها (به جای متغیر X_k) و Y به کار برده می‌شود. مانده‌های متناظر با $X_۴$ با Y ناهمبسته نیستند. بنابراین $X_۴$ به عنوان 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 $Y = (Y_۱, ..., Y_n)^T$ و $\mathbf{X}$ ماتریس طرح با بعد $n \times p$ باشد:

- گام ۱: ابتدا روش RDC-SIS را برای Y و $\mathbf{X}$ به کار برید. فرض کنید در این مرحله $d_۱$ متغیر توضیحی به صورت $X_{\mathcal{M}_۱} = \{X_j : j \in \mathcal{M}_۱\}$ انتخاب می‌شود 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 $\mathbf{X}_{new} = \{I_n - \mathbf{X}_۱(\mathbf{X}_۱^T\mathbf{X}_۱)^{-۱}\mathbf{X}_۱^T\}\mathbf{X}_۱^c$ را محاسبه کنید و سپس روش RDC-SIS را برای Y و تمام ستونهای ماتریس $\mathbf{X}_{new}$ بکار 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 $\mathcal{M}_۱ \cup \mathcal{M}_۲$ بروز رسانی کنید.

- گام ۳: گام ۲ را $k-۱$ بار تکرار کنید تا تعداد متغیرهای توضیحی انتخاب شده از d تجاوز کند، یعنی $d_۱ + ... + d_k \geq d$. در پایان مجموعه متغیرهای توضیحی 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 [۱۲] استفاده می‌کند که در مقاله ژو و همکاران (۲۰۱۱) برای غربالگری آزاد-مدل استفاده شده است. در اینجا مقدار $d_۱$ در گام ۱ توسط کاربر تعیین می‌شود. در عمل، $d_۱$ به عنوان یک پارامتر تنظیم‌کننده در نظر گرفته می‌شود و مقدار بهینه آن با مینیمم کردن میانگین مربع خطای پیش‌بینی تعیین می‌گردد.

۳ انتخاب متغیر و تشخیص ساختار

انتخاب یک مقدار بزرگتر برای d، احتمال انتخاب متغیرهای بی‌اهمیت را برای ورود به مدل افزایش می‌دهد. در این بخش، یک روش مبتنی بر تاوان برای حذف متغیرهای اضافی از مدل و تشخیص مؤلفه‌های خطی و غیرخطی ارائه می‌شود. فرض کنید d متغیر در مرحله غربالگری برای ورود به مدل انتخاب شده‌اند. مدل جمعی ناپارامتری

$$Y = \sum_{j=۱}^{d} f_j(X_j) + \epsilon,$$

[12] Orthogonal projection

را در نظر بگیرید. برای تقریب توابع هموار نامعلوم از توابع پایه‌ای B -اسپلاین به صورت

$$f_j(x) \approx \sum_{k=1}^{K} b_{jk} B_{jk}(x) \qquad j = 1, ..., d,$$

استفاده می‌شود یک روش رایج برای برآورد و انتخاب متغیر بطور همزمان، استفاده از رگرسیون تاوانیده است. در اینجا برای انتخاب متغیر و تشخیص ساختار مدل، از دو تابع تاوان به طور همزمان استفاده می‌شود، یعنی بردار ضرایب $b = (b_1^T, \dots b_d^T)^T$، که $b_j = (b_{j1}, \dots b_{jK})^T$، با حل مسئله بهینه‌سازی

$$\begin{aligned} \hat{b} &= \arg\min_b \frac{1}{2} \sum_{i=1}^{n} (Y_i - \mu - \sum_{j=1}^{d} \sum_{k=1}^{K} b_{jk} B_{jk}(X_{ij}))^2 \\ &+ n \sum_{j=1}^{d} p_{\lambda_1}(\|b_j\|_{A_j}) + n \sum_{j=1}^{d} p_{\lambda_2}(\|b_j\|_{D_j}), \end{aligned} \qquad (۶)$$

برآورد می‌شود که در آن توابع تاوان $p_{\lambda_1}(.)$ و $p_{\lambda_2}(.)$ با پارامترهای تنظیم کننده λ_1 و λ_2 به ترتیب برای شناسایی مؤلفه‌های صفر و خطی به کار می‌روند. علاوه براین، A_j و D_j دو ماتریس $K \times K$ و $\|b_j\|_{A_j} = \left(b_j^T A_j b_j\right)^{\frac{1}{2}}$ و $\|b_j\|_{D_j} = \left(b_j^T D_j b_j\right)^{\frac{1}{2}}$ هستند. انتخاب ماتریس‌های A_j و D_j بسیار حائز اهمیت است. این ماتریس‌ها باید طوری انتخاب شوند که $\|b_j\|_{A_j} = 0$ باشد، اگر وفقط اگر $\sum_k b_{jk} B_{jk}(x) \equiv 0$ و بطور مشابه $\|b_j\|_{D_j} = 0$ باشد، اگر و فقط اگر $\sum_k b_{jk} B_{jk}(x)$ یک تابع خطی باشد، یعنی دو تابع تاوان بتوانند برای شناسایی مؤلفه‌های صفر و خطی به کار روند. در اینجا ماتریس‌های A_j و D_j را به صورت $A_j = \{\int_0^1 B_{jk}(x) B_{jk'}(x) dx\}_{k,k'=1}^{K}$ و $D_j = \{\int_0^1 B''_{jk}(x) B''_{jk'}(x) dx\}_{k,k'=1}^{K}$ انتخاب می‌کنند. انتخاب D_j به صورت فوق از این واقعیت نتیجه می شود که مشتق دوم یک تابع خطی برابر صفر است. فرض کنید

$$Z_j = \begin{pmatrix} B_{j1}(X_{1j}) & B_{j2}(X_{1j}) & \cdots & B_{jK}(X_{1j}) \\ \vdots & \vdots & & \vdots \\ B_{j1}(X_{nj}) & B_{j2}(X_{nj}) & \cdots & B_{jK}(X_{nj}) \end{pmatrix}_{n \times K},$$

$Z = (Z_۱, ..., Z_d)$ و $Y = (Y_۱, ..., Y_n)$ باشد. پس رابطه (۶) را می‌توان به صورت ماتریسی

$$\hat{b} = \arg\min_b \frac{۱}{۲}\|Y - Zb\|^۲ + n\sum_{j=۱}^{d} p_{\lambda_۱}(\|b_j\|_{A_j}) + n\sum_{j=۱}^{d} p_{\lambda_۲}(\|b_j\|_{D_j}). \qquad (۷)$$

نوشت. به دلیل سهولت، تابع هدف سمت راست رابطه (۷) با $Q(b)$ نشان داده می‌شود.

تاکنون توابع تاوان زیادی توسط افراد مختلف معرفی شده‌است. سوالی که اینجا ممکن است پیش بیاید این است که از چه نوع تابع تاوان باید استفاده کرد. فن و لی (۲۰۰۱) نشان دادند که یک تابع تاوان خوب باید برآوردگری با سه ویژگی مطلوب زیر را نتیجه دهد:

۱. تنکی: برآوردگر نتیجه شده باید به طور خودکار ضرایب برآورد شده‌ای که مقدار کوچکی دارند را برابر صفر قرار دهد تا متغیرهای مناسب را انتخاب کند. این کار پیچیدگی مدل را کاهش می دهد.

۲. ناااریبی: برآوردگر به دست آمده برای ضرایب رگرسیونی که در واقع مقادیر آنها بزرگ است، تقریبا ناااریب باشد. این ویژگی اریبی مدل را کاهش می دهد.

۳. پیوستگی: برآوردگر نتیجه شده دارای تغییرات پیوسته باشد و ناپایداری در پیش بینی مدل را کاهش دهد.

بنابراین، فن و لی (۲۰۰۱) یک تابع تاوان محدب به نام SCAD را معرفی کردند که دارای هر سه ویژگی مطلوب است. این تابع تاوان دارای مشتق مرتبه اول به صورت

$$p'_{a,\lambda}(x) = \lambda\left\{I(|x| \le \lambda) + \frac{(a\lambda - |x|)_+}{(a-۱)\lambda} I(|x| > \lambda)\right\}, \qquad x \ge ۰,$$

است، که در آن $a > ۲$ و $p'_{a,\lambda}(۰) = ۰$. فن و لی (۲۰۰۱) از دیدگاه بیزی مقدار $a = ۳/۷$ را به عنوان یک مقدار مناسب برای مسائل مختلف پیشنهاد دادند.

برای حل مسئله بهینه‌سازی (۷) از روش تکراری تقریب موضعی درجه دو [۱۳] (LQA) معرفی شده توسط فن و لی (۲۰۰۱) استفاده می‌کنیم. با استفاده از بسط تیلور و داشتن یک مقدار اولیه b_j°، توابع

[13] Local quadratic approximation

تاوان را می‌توان به صورت

$$p_{\lambda_1}(\|b_j\|_{A_j}) \approx p_{\lambda_1}(\|b_j^{(0)}\|_{A_j}) + \frac{1}{2}\frac{p'_{\lambda_1}(\|b_j^{(0)}\|_{A_j})}{\|b_j^{(0)}\|_{A_j}}\{\|b_j\|^2_{A_j} - \|b_j^{(0)}\|^2_{A_j}\},$$

$$p_{\lambda_2}(\|b_j\|_{D_j}) \approx p_{\lambda_2}(\|b_j^{(0)}\|_{D_j}) + \frac{1}{2}\frac{p'_{\lambda_2}(\|b_j^{(0)}\|_{D_j})}{\|b_j^{(0)}\|_{D_j}}\{\|b_j\|^2_{D_j} - \|b_j^{(0)}\|^2_{D_j}\}.$$

تقریب زد. با به کار بردن تقریب‌های فوق و حذف جملات ثابت می‌توان رابطه (۷) را به صورت

$$Q(b) = \frac{1}{n}\|Y - Zb\|^2 + \frac{1}{2}b^T(\Omega_1 + \Omega_2)b \qquad (۸)$$

خلاصه کرد، که در آن Ω_1 و Ω_2 دو ماتریس بلوکی با بعد $dK \times dK$ به صورت

$$\begin{aligned}
\Omega_1 &= \text{diag}\left(\frac{p'_{\lambda_1}(\|b_1^{(0)}\|_{A_1})}{\|b_1^{(0)}\|_{A_1}}A_1, \ldots, \frac{p'_{\lambda_1}(\|b_d^{(0)}\|_{A_d})}{\|b_d^{(0)}\|_{A_d}}A_d\right) \\
\Omega_2 &= \text{diag}\left(\frac{p'_{\lambda_2}(\|b_1^{(0)}\|_{D_1})}{\|b_1^{(0)}\|_{D_1}}D_1, \ldots, \frac{p'_{\lambda_2}(\|b_d^{(0)}\|_{D_d})}{\|b_d^{(0)}\|_{D_d}}D_d\right)
\end{aligned}$$

هستند. توجه شود که رابطه (۸) یک تابع درجه دو برحسب b بوده و در نتیجه دارای جواب صریح

$$\hat{b} = (Z^TZ + n(\Omega_1 + \Omega_2))^{-1}Z^TY,$$

است، به عنوان برآورد اولیه در تکرار بعدی استفاده می‌شود. این الگوریتم به طور مکرر مسئله بهینه‌سازی (۸) را حل می‌کند. یعنی برآورد $b^{(m)}$ توسط $b^{(m+1)}$ به روز می شود که $m = 0, 1, 2, \ldots$. به عبارت دیگر، در تکرار $m+1$ باید مسئله بهینه‌سازی (۸) حل شود، اما به جای استفاده از $b^{(0)}$، از برآورد $b^{(m)}$ به عنوان مقدار اولیه استفاده می‌شود. برآورد بدست آمده با مینیمم کردن (۸) برآورد جدید ($b^{(m+1)}$) است. این الگوریتم تا رسیدن به همگرایی تکرار می‌شود. فن و لی (۲۰۰۱) نشان دادند که الگوریتم LQA پس از تعداد کمی تکرار به همگرایی می‌رسد. در هر تکرار به محض اینکه برخی از مؤلفه‌های $\|b_j\|_{A_j}$ یا $\|b_j\|_{D_j}$ کوچکتر از 10^{-6} شود، مولفه $f_j(x)$ به ترتیب به عنوان تابع صفر یا تابع خطی تشخیص داده می‌شود.

۴ مطالعه شبیه‌سازی

در این بخش، با شبیه‌سازی عملکرد روش دو مرحله‌ای بررسی می‌شود. ابتدا، در مثال ۱ حساسیت روش RDC-SIS نسبت به پارامتر d تحلیل می‌شود. سپس در مثال ۲، عملکرد RDC-SIS را 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 و ۴، عملکرد روش تکراری RDC-ISISرا 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 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 می‌دهد 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 مقاله، برای شبیه‌سازی با انتخاب $d_1 = 5$ و $d_2 = p - 5$ الگوریتم فقط یکبار تکرار می‌شود.

در مثال ۵ کارایی روش دو مرحله‌ای را در شناسایی مؤلفه‌های غیرصفر و همچنین تشخیص مؤلفه‌های خطی و غیرخطی در یک مدل جمعی خطی-جزیی با استفاده از دو تابع تاوان بررسی می‌شود. برای انتخاب بهینه پارامترهای λ_1 و λ_2 از معیار BIC به صورت

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

استفاده می‌شود، که در آن $\hat{b}_\lambda$ مقدار مینیمم کننده (۷) 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: برای تحلیل حساسیت عملکرد RDC-SIS نسبت به مقادیر مختلف d، داده‌ها از مدل

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

جدول ۱: احتمال تجربی P_j و احتمال S در مثال ۱

$\varepsilon \sim t(۱)$					$\varepsilon \sim N(۰,۱)$					d	n	p
S	$p_۴$	$p_۳$	$p_۲$	$p_۱$	S	$p_۴$	$p_۳$	$p_۲$	$p_۱$			
۰/۴۱	۰/۴۵	۰/۹۲	۱/۰۰	۰/۸۴	۰/۵۳	۰/۶۰	۰/۹۶	۱/۰۰	۰/۹۱	$d_۱$	۱۰۰	
۰/۵۸	۰/۶۲	۰/۹۶	۱/۰۰	۰/۹۴	۰/۷۳	۰/۷۵	۰/۹۸	۱/۰۰	۰/۹۸	$d_۲$		
۰/۶۷	۰/۷۰	۰/۹۷	۱/۰۰	۰/۹۶	۰/۸۱	۰/۸۱	۰/۹۹	۱/۰۰	۱/۰۰	$d_۳$		
۰/۹۳	۰/۹۳	۱/۰۰	۱/۰۰	۱/۰۰	۰/۹۹	۰/۹۹	۱/۰۰	۱/۰۰	۱/۰۰	$d_۱$	۲۰۰	۱۰۰۰
۰/۹۷	۰/۹۷	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۲$		
۰/۹۸	۰/۹۸	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۳$		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۱$	۴۰۰	
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۲$		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۳$		
۰/۴۲	۰/۵۵	۰/۹۰	۱/۰۰	۰/۷۷	۰/۵۵	۰/۶۱	۰/۹۶	۱/۰۰	۰/۹۲	$d_۱$	۱۰۰	
۰/۵۸	۰/۶۶	۰/۹۶	۱/۰۰	۰/۸۹	۰/۷۱	۰/۷۲	۰/۹۹	۱/۰۰	۰/۹۶	$d_۲$		
۰/۶۵	۰/۷۰	۰/۹۸	۱/۰۰	۰/۹۲	۰/۷۸	۰/۷۹	۰/۹۹	۱/۰۰	۰/۹۹	$d_۳$		
۰/۹۲	۰/۹۷	۱/۰۰	۱/۰۰	۰/۹۹	۰/۹۷	۰/۹۷	۱/۰۰	۱/۰۰	۱/۰۰	$d_۱$	۲۰۰	۲۰۰۰
۰/۹۷	۰/۹۷	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۲$		
۰/۹۸	۰/۹۸	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۳$		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۱$	۴۰۰	
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۲$		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	$d_۳$		

تولید شده‌اند، که در آن

$$g_۱(x) = x, \qquad g_۲(x) = (۲x-۱)^۲, \qquad g_۳(x) = \sin(۲\pi x)/(۲-\sin(۲\pi x)),$$
$$g_۴(x) = ۰/۱\sin(۲\pi x) + ۰/۲\cos(۲\pi x) + ۰/۳\sin(۲\pi x)^۲ + ۰/۴\cos(۲\pi x)^۳ + ۰/۵\sin(۲\pi x)^۳,$$

و متغیرهای توضیحی دارای توزیع حاشیه‌ای نرمال استاندارد و همبستگی $Cov(X_i, X_j) = ۰/۸^{|i-j|}$ هستند. سه مقدار مختلف $d_۳ = ۳[n/\log(n)]$, $d_۲ = ۲[n/\log(n)]$, $d_۱ = [n/\log(n)]$ 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 گزارش شده است. همان‌طور که ملاحظه می‌شود، با افزایش مقدار d عملکرد RDC-SIS بهبود می‌یابد. برای $n = ۱۰۰$، عملکرد RDC-SIS به ازای $d_۱$ و $d_۲$ چندان مطلوب نیست و امکان حذف شدن متغیر مهم $X_۴$ وجود دارد، لذا برای حجم نمونه کوچک مقدار $d_۳$ را پیشنهاد می‌شود. اما برای $n = ۲۰۰, ۴۰۰$ و هر دو نوع توزیع خطا، هر یک از مقادیر $d_۱$، $d_۲$ و $d_۳$ را می‌توان استفاده کرد.

در مثال‌های ۲ تا ۴، 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.

جدول ۲: چندک‌های M، احتمال تجربی P_j و احتمال S در مدل ۱

S	P					M					روش	c	خطا
	۵	۴	۳	۲	۱	٪۹۵	٪۷۵	٪۵۰	٪۲۵	٪۵			
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	RDC-SIS	۱	$N(۰,۱)$
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	DC-SIS		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	SIS		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	SIRS		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	RDC-SIS	۲	
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	DC-SIS		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	SIS		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	SIRS		
۰/۹۹	۰/۹۹	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱۹	۵	۵	۵	۵	RDC-SIS	۱	(۳t(
۰/۶۷	۰/۷۱	۰/۷۷	۰/۸۶	۰/۸۴	۰/۸۲	۵۸۴	۱۰۹	۱۹	۷	۵	DC-SIS		
۰/۱۰	۰/۱۶	۰/۲۰	۰/۲۱	۰/۲۱	۰/۲۰	۹۶۶	۹۱۶	۸۰۶	۴۶۷	۳۵	SIS		
۰/۹۸	۰/۹۸	۰/۹۹	۰/۹۹	۱/۰۰	۱/۰۰	۲۵	۶	۵	۵	۵	SIRS		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	RDC-SIS	۲	
۰/۸۷	۰/۹۰	۰/۹۴	۰/۹۴	۰/۹۴	۰/۹۵	۱۵۶	۶	۵	۵	۵	DC-SIS		
۰/۲۳	۰/۳۰	۰/۴۲	۰/۴۴	۰/۴۳	۰/۴۴	۹۸۱	۸۶۵	۴۹۴	۹۰	۵	SIS		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۵	۵	۵	۵	SIRS		

مثال ۲: دو مدل به صورت

مدل ۱: $Y = c\beta^T X + \sigma\varepsilon$ مدل۲: $Y = X_۱ + ۲X_۲^۲ + ۳X_۳^۳ + ۴X_۴^۴ + \varepsilon$ در نظر بگیرید، که در آن $\beta = (۱, ۰/۸, ۰/۶, ۰/۴, ۰/۲, ۰, \ldots, ۰)^T$ و $\sigma^۲ = ۶/۸۳$. در مدل ۱، به منظور کنترل نسبت سیگنال به نوفه [۱۴] مقادیر مختلفی برای c در نظر گرفته شده است. مقادیر $c = ۱, ۲$ که متناظر با $R^۲ = ٪۵۰, ٪۸۰$ هستند، انتخاب شده‌اند. در هر دو مدل، بردار متغیرهای توضیحی از توزیع نرمال چند متغیره با میانگین صفر و ماتریس کوواریانس $\Sigma = (\sigma_{ij})_{p\times p}$ تولید می‌شود که $\sigma_{ij} = ۰/۵^{|i-j|}$. در مدل ۱، برای خطا دو توزیع نرمال استاندارد و توزیع تی-استودنت 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 و ۳ ارائه شده‌اند.

با توجه به جدول ۲، 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 متغیرهای مهم را به درستی شناسایی می‌کنند، اما برای خطای غیر نرمال نتایج کاملاً متفاوت است. برای توزیع خطای تی-استودنت، 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. همچنین روش SIRS نیز در شناسایی متغیرهای مهم خوب عمل می‌کند و عملکرد دو روش RDC-SIS و SIRS

[14] Signal-to-noise ratio

جدول ۳: چندک‌های M، احتمال تجربی P_j و احتمال S در مدل ۲

S	P				M					روش	خطا
	۴	۳	۲	۱	٪۹۵	٪۷۵	٪۵۰	٪۲۵	٪۵		
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۴	۴	۴	۴	۴	RDC-SIS	$N(۰,۱)$
۰/۹۹	۱/۰۰	۱/۰۰	۱/۰۰	۰/۹۹	۹	۶	۵	۵	۴	DC-SIS	
۰/۹۳	۱/۰۰	۱/۰۰	۱/۰۰	۰/۹۳	۱۰۲	۱۱	۶	۵	۵	NIS	
۰/۸۵	۰/۸۵	۱/۰۰	۱/۰۰	۱/۰۰	۲۹۹	۱۹	۵	۴	۴	SIRS	
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۵	۴	۴	۴	۴	RDC-SIS	(۳t(
۰/۹۴	۰/۹۸	۰/۹۸	۰/۹۶	۰/۹۴	۶۴	۷	۵	۵	۴	DC-SIS	
۰/۵۵	۰/۸۹	۰/۸۵	۰/۸۱	۰/۶۵	۸۶۴	۸۲	۱۱	۶	۵	NIS	
۰/۶۹	۰/۷۸	۱/۰۰	۱/۰۰	۰/۹۹	۴۶۸	۳۴	۷	۴	۴	SIRS	
۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	۴	۴	۴	۴	۴	RDC-SIS	SN
۰/۹۹	۱/۰۰	۱/۰۰	۱/۰۰	۰/۹۹	۸	۶	۵	۵	۴	DC-SIS	
۰/۹۲	۱/۰۰	۱/۰۰	۱/۰۰	۰/۹۲	۱۰۴	۱۱	۶	۵	۵	NIS	
۰/۸۷	۰/۸۷	۱/۰۰	۱/۰۰	۱/۰۰	۳۴۵	۲۳	۵	۴	۴	SIRS	

تقریبا یکسان است. با توجه به مقادیر P_j و S در جدول ۲ 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 که حاوی نتایج مدل ۲ است، نشان می‌دهد که برای هر سه نوع توزیع خطا، عملکرد روش RDC-SIS بسیار خوب است و نسبت به روش‌های دیگر بهتر عمل می‌کند. برای $j = ۱, \ldots, ۴$ 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 این موارد بهتر است از یک روش تکراری مناسب برای حذف متغیرهای بی‌اهمیت و بازگرداندن متغیرهای مهم به مدل استفاده شود.

مثال ۳: در این مثال، عملکرد روش RDC-ISISرا با روش‌های SIS، ISIS، SIRS، DC-SIS و RDC-SIS در مدل خطی (۵) مقایسه می‌شود. علاوه براین، یک مدل خطی دیگر، با نسبت سیگنال به نوفه ضعیف تر، به‌صورت

$$Y = ۲/۵X_۱ + ۲/۵X_۲ + ۲/۵X_۳ - ۷/۵\sqrt{\rho}X_۴ + \varepsilon \qquad (۹)$$

در نظر گرفته شده و مدل (۵) را مدل ۱ و مدل (۹) را مدل ۲ می‌نامیم. در این مثال، $\varepsilon \sim N(۰,۱)$

جدول ۴: احتمال تجربی P_j و احتمال S در مثال ۳

ρ	روش	مدل ۱					مدل ۲				
		P				S	P				S
		۱	۲	۳	۴		۱	۲	۳	۴	
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	۱/۰۰	۱/۰۰	۰/۰۱	۰/۰۱
	DC-SIS	۱/۰۰	۱/۰۰	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

و سه مقدار مختلف $\rho = ۰/۲, ۰/۵, ۰/۸$ برای ضریب همبستگی در نظر گرفته شده و شبیه‌سازی ۵۰۰ بار تکرار شده است. در این مدل، متغیر $X_۴$ بطور توأم مهم اما بطور 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 ملاحظه می‌شود روش RDC-ISISدر هر دو مدل خطی و به ازای مقادیر مختلف ρ در انتخاب متغیر $X_۴$ بسیار تواناست. ما همچنین RDC-ISISرا با ISIS، نسخه تکراری SIS، مقایسه کرده‌ایم. نتایج نشان می‌دهد که در برخی موارد روش RDC-ISISدر انتخاب $X_۴$ حتی از ISIS نیز بهتر عمل می‌کند. به عنوان 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-ISISاز 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-ISISیک روش غربال‌گری آزاد مدل است و می‌تواند برای مدل‌های رگرسیونی مختلف جهت تشخیص روابط خطی و غیر خطی به‌کار رود.

مثال ۴: در این مثال، مدل رگرسیونی جمعی ناپارامتری به صورت

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

است، که در آن توابع پایه‌ای بصورت

$$f_۱(x) = \exp(۲x/۳), \qquad f_۲(x) = \begin{cases} x+۴ & x < -۲ \\ |x| & |x| \le ۲ \\ ۴-x & x > ۲, \end{cases}$$
$$f_۳(x) = \frac{\sin(۳\pi x/۴ + ۳/۲)}{۲ - \sin(۳\pi x/۴ + ۳/۲)}, \quad f_۴(x) = \log(x^۲).$$

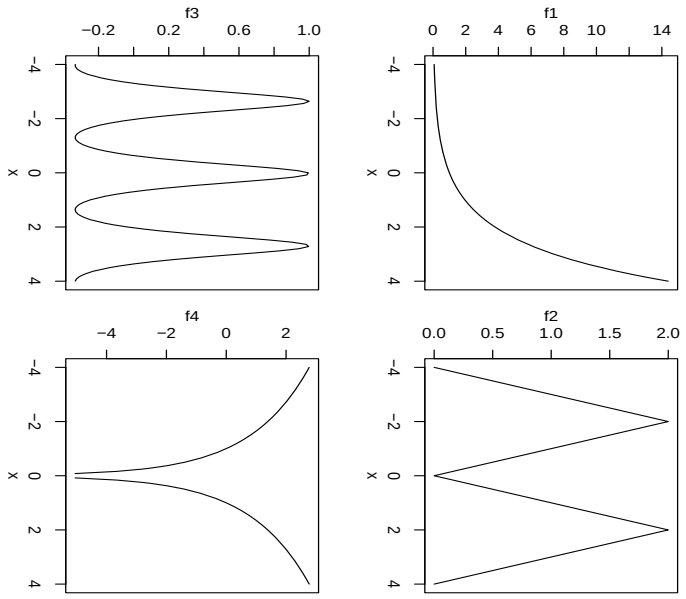
تعریف شده و نمودارهای آنها در شکل ۱ رسم شده‌اند. با توجه به مدل، متغیرهای $(X_۱, X_{۱۰۱}, X_{۲۰۱}, X_{۲۰۲})$ متغیرهای مؤثر بر پاسخ هستند. متغیرهای توضیحی از توزیع نرمال چند متغیره با میانگین صفر و همبستگی $Cov(X_i, X_j) = \rho^{|i-j|}$ تولید شده‌اند. سپس متغیر X_k با متغیر جدید $X_k = ۰/۸X_۱ + \zeta_k$ جایگزین می‌شود، که در آن $\zeta_k \sim N(۰, ۱)$ و $k = ۲, \ldots, ۱۰۰$، یعنی 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 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 یعنی $(X_۲, \ldots, X_{۱۰۰})$ 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 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.

مثال ۵: در این مثال، ابتدا روش RDC-SIS برای کاهش بعد به کار رفته، سپس با استفاده از دو تابع

جدول ۵: احتمال تجربی P_j و احتمال S در مثال ۴

ρ	روش	$\varepsilon \sim N(۰,۱)$					$\varepsilon \sim t(۲)$				
		S	P				S	P			
			۱	۲	۳	۴		۱	۲	۳	۴
۰٫۵	SIS	۰/۰۰	۱/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	۰/۹۹	۰/۰۱	۰/۰۰	۰/۰۱
	ISIS	۰/۰۰	۰/۹۸	۰/۰۷	۰/۰۶	۰/۰۵	۰/۰۰	۰/۹۸	۰/۰۶	۰/۰۵	۰/۰۵
	NIS	۰/۵۰	۱/۰۰	۰/۸۲	۰/۸۳	۰/۷۴	۰/۳۲	۰/۹۸	۰/۷۵	۰/۷۱	۰/۶۱
	SIRS	۰/۰۰	۱/۰۰	۰/۰۱	۰/۰۱	۰/۰۰	۰/۰۰	۱/۰۰	۰/۰۰	۰/۰۱	۰/۰۰
	DC-SIS	۰/۰۲	۱/۰۰	۰/۲۵	۰/۲۹	۰/۲۴	۰/۰۳	۱/۰۰	۰/۳۱	۰/۳۴	۰/۲۷
	RDC-SIS	۰/۱۱	۱/۰۰	۰/۳۴	۰/۴۰	۰/۳۳	۰/۱۱	۱/۰۰	۰/۳۰	۰/۳۶	۰/۳۵
	RDC-ISIS	۰/۹۶	۱/۰۰	۰/۹۸	۰/۹۹	۰/۹۹	۰/۸۲	۱/۰۰	۰/۹۴	۰/۹۹	۰/۸۸
۰٫۸	SIS	۰/۰۰	۱/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	۰/۰۰	۰/۹۹	۰/۰۰	۰/۰۱	۰/۰۰
	ISIS	۰/۰۰	۰/۹۳	۰/۰۷	۰/۰۴	۰/۰۴	۰/۰۰	۰/۹۷	۰/۰۹	۰/۰۵	۰/۰۵
	NIS	۰/۸۱	۱/۰۰	۰/۸۵	۰/۹۹	۰/۹۶	۰/۵۹	۰/۹۹	۰/۷۲	۰/۹۳	۰/۸۸
	SIRS	۰/۰۰	۱/۰۰	۰/۰۰	۰/۰۱	۰/۰۱	۰/۰۰	۱/۰۰	۰/۰۰	۰/۰۱	۰/۰۰
	DC-SIS	۰/۱۳	۱/۰۰	۰/۲۸	۰/۷۸	۰/۶۱	۰/۱۴	۱/۰۰	۰/۳۰	۰/۷۲	۰/۶۲
	RDC-SIS	۰/۳۱	۱/۰۰	۰/۳۵	۰/۷۸	۰/۷۲	۰/۳۲	۱/۰۰	۰/۳۶	۰/۷۹	۰/۷۲
	RDC-ISIS	۰/۹۰	۱/۰۰	۱/۰۰	۰/۹۸	۰/۹۳	۰/۸۳	۰/۹۸	۰/۹۱	۰/۹۸	۰/۹۵



شکل ۱: نمودارهای توابع پایه‌ای برای مدل جمعی ناپارامتری مثال ۴ در فاصله $[-۴,۴]$.

تاوان SCAD به‌طور همزمان یک مدل جمعی خطی-جزیی برازش داده می‌شود. داده‌ها از مدل

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

تولید شده‌اند، که در آن $f_۱(x) = ۳\sin(۲\pi x)/(۲ - \sin(۲\pi x))$، $f_۲(x) = ۶x(۱ - x)$، $f_۳(x) = ۲x$، $f_۴(x) = x$، $f_۵(x) = -x$، 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 همانند مثال ۲ تولید شده‌اند. برای نشان دادن کارایی روش ارائه شده، برای خطا دو توزیع نرمال استاندارد و توزیع تی-استودنت با ۲ درجه آزادی که برای تولید خطای دم سنگین استفاده می‌شود، در نظر گرفته شده و شبیه‌سازی با $n = ۳۰, ۷۰, ۲۰۰, ۴۰۰$ و $p = ۱۰۰۰, ۲۰۰۰$ به تعداد ۲۰۰ بار تکرار شده است. نتایج انتخاب متغیر و تشخیص ساختار مدل در جدول ۶ 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. 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 به عنوان مؤلفه غیرصفر ناپارامتری تشخیص می‌دهد، اما برای خطای غیرنرمال در برخی از تکرارها بعضی از مؤلفه‌های ناپارامتری به عنوان مؤلفه صفر یا خطی تشخیص داده می‌شود. همچنین در برخی از تکرارها، مؤلفه‌های خطی به عنوان مولفه‌های صفر یا غیرخطی انتخاب می‌شوند. با توجه به جدول، هنگامی که حجم نمونه بسیار کوچک است، این روش عملکرد چندان مطلوبی ندارد. همانطور که می بینید برای $n = ۳۰$ عملکرد این روش در انتخاب متغیر و تشخیص ساختار مدل چندان رضایت بخش نیست. 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.

۵ مثال کاربردی

مجموعه داده این مثال توسط بشل و همکاران (۲۰۰۷) مورد تحلیل و بررسی قرار گرفته و در بسته mixOmics نرم‌افزار R قابل دسترسی است. این داده‌ها مربوط به بیماری مسمومیت کبد بوده و شامل سطوح بیان ۳۱۱۶ ژن و ۹ اندازه گیری بالینی برای ۶۴ موش است. بنابراین در این داده‌ها ۳۱۱۶ متغیر توضیحی و ۹ متغیر پاسخ کمی وجود دارد. متغیرهای پاسخ به اختصار عبارتند از: BUN، نیتروژن اوره؛ TP، پروتئین کل؛ ALB، آلبومین؛ ALT، آلانین آمینوترانسفراز؛ SDH، سوربیتول دهیدروژناز؛

جدول ۶: نتایج تشخیص ساختار مدل در مثال ۵

NLT	NL	NNT	NN	N	خطا	n	p
۲/۶۰(۰/۸۷)	۲/۷۶(۱/۲۲)	۲(۰)	۲/۹۱(۱/۱۶)	۵/۶۷	$N(۰,۱)$	۳۰	
۲/۶۱(۰/۹۸)	۳/۶۵(۱/۲۶)	۱/۹۵(۰/۲۳)	۳/۰۹(۱/۳۱)	۶/۷۴	$t(۲)$		
۲/۶۹(۰/۶۳)	۲/۷۴(۰/۸۶)	۲(۰)	۲/۶۸(۱/۰۴)	۵/۴۲	$N(۰,۱)$	۷۰	
۲/۶۲(۰/۸۵)	۳/۶۵(۱/۱۵)	۱/۹۵(۰/۱۹)	۲/۹۶(۱/۱۰)	۶/۶۱	$t(۲)$		۱۰۰۰
۲/۷۴(۰/۳۲)	۲/۹۳(۰/۸۸)	۲(۰)	۲/۳۱(۰/۹۶)	۵/۲۴	$N(۰,۱)$	۲۰۰	
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	$N(۰,۱)$	۴۰۰	
۲/۷۴(۰/۷۵)	۳/۲۸(۰/۶۶)	۱/۹۹(۰/۰۸)	۲/۴۹(۰/۵۴)	۵/۷۷	$t(۲)$		
۲/۶۱(۱/۳۶)	۳/۶۴(۱/۳۹)	۲(۰)	۳/۱۱(۱/۳۳)	۶/۷۵	$N(۰,۱)$	۳۰	
۲/۴۲(۱/۲۴)	۳/۶۸(۱/۵۴)	۱/۸۳(۱/۰۶)	۳/۷۰(۱/۴۶)	۷/۳۸	$t(۲)$		
۲/۸۵(۱/۰۵)	۳/۲۳(۱/۰۱)	۲(۰)	۲/۷۱(۱/۰۹)	۵/۹۴	$N(۰,۱)$	۷۰	
۲/۷۲(۰/۹۸)	۳/۳۸(۱/۴۸)	۱/۹۰(۰/۵۸)	۳/۲۶(۱/۳۱)	۶/۶۴	$t(۲)$		۲۰۰۰
۲/۸۴(۰/۶۴)	۳/۱۲(۰/۹۴)	۲(۰)	۲/۴۶(۰/۸۴)	۵/۵۸	$N(۰,۱)$	۲۰۰	
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	$N(۰,۱)$	۴۰۰	
۲/۸۱(۰/۶۳)	۲/۹۷(۰/۸۵)	۱/۹۹(۰/۰۶)	۲/۵۴(۰/۷۴)	۵/۵۱	$t(۲)$		

جدول ۷: چهار ژن بسیار مهم انتخاب شده توسط روش‌های مختلف

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

AST، آسپارتات آمینوترانسفراز؛ ALP، آلکالین فسفاتاز؛ TBA، کل اسیدهای صفراوی؛ و CHOL، کلسترول. در این بخش ALT به عنوان متغیر پاسخ است و هدف پیدا کردن مؤثرترین ژنها در پیش‌بینی ALT و نوع تاثیر این ژن‌هاست.

جدول ۷ نشان دهنده ژن‌هایی است که توسط روش‌های غربالگری مختلف در رتبه‌های اول تا چهارم قرار می‌گیرند. به عنوان 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 به عنوان مهمترین ژن انتخاب می‌کنند. برای مقایسه کارایی این روش‌ها، می‌توان یک مدل جمعی با ۴ متغیر به‌صورت

$$Y = g_{k۱}(X_{k۱}) + g_{k۲}(X_{k۲}) + g_{k۳}(X_{k۳}) + g_{k۴}(X_{k۴}) + \varepsilon_k, \qquad k = ۱, ۲, ۳, ۴$$

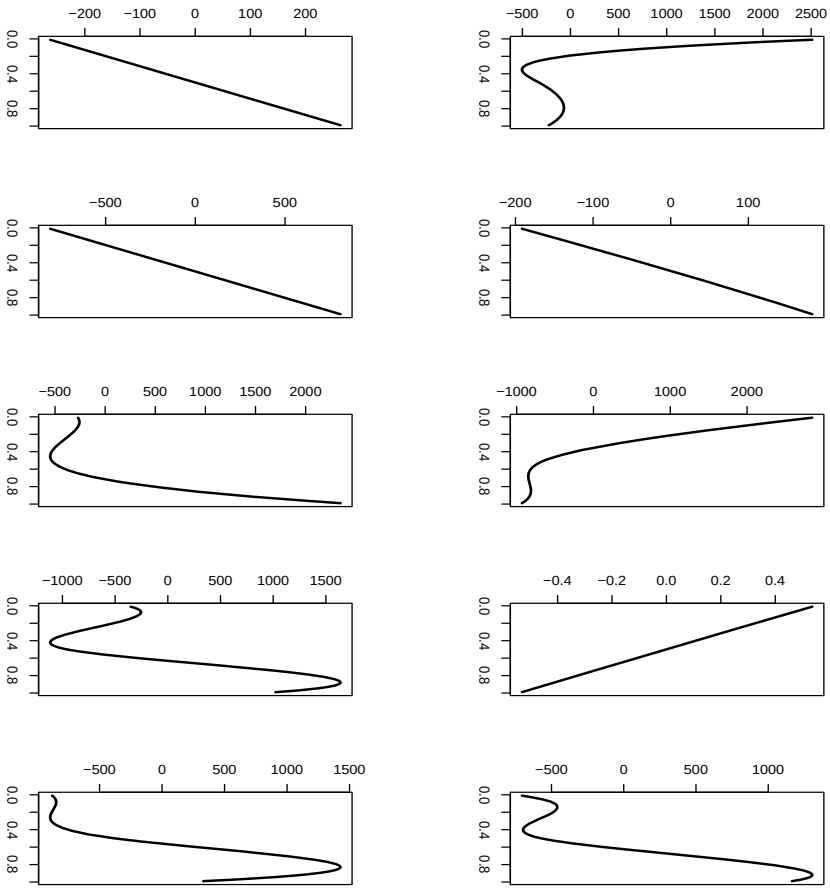
برازش داد، که در آن $X_{k۱}$، $X_{k۲}$، $X_{k۳}$ و $X_{k۴}$ سطوح بیان ژن‌های انتخاب شده توسط هر کدام از روش‌های غربالگری، $g_{kj}(.), j = ۱, ..., ۴$ توابع پیوند نامعلوم هستند. برای برازش مدل فوق از

جدول ۸: 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

تابع gam در بسته mgcv نرم‌افزار R استفاده می‌کنیم. تابع gam برای برازش مدل جمعی تعمیم‌یافته استفاده می‌شود. برای مقایسه عملکرد نیکویی برازش از R^2 تعدیل شده مدل‌های جمعی برازش شده فوق استفاده می‌شود. پس از محاسبه R^2 تعدیل شده داریم: $R^2_{RDC-SIS} = ۰/۹۴$، $R^2_{SIRS} = ۰/۸۵$، $R^2_{NIS} = ۰/۹۹$ $R^2_{DC-SIS} = ۰/۹۹$. با توجه به مقادیر R^2، برای این مجموعه داده، روش‌های NIS و DC-SIS عملکرد بسیار خوبی دارند و توانایی RDC-SIS نیز قابل قبول است، اما SIRS دارای عملکرد نسبتاً ضعیفی است. بنابراین RDC-SIS یک روش کارا برای کاهش بعد داده‌های با بعد بالاست. برای انتخاب مدل نهایی، در ابتدا نسخه تکراری RDC-SIS، یعنی RDC-ISIS با $d = ۲[n/\log(n)] = ۲۸$ و $p_۱ = d/۲$ به کار رفته و بعد داده‌ها را به $d = ۲۸$ کاهش داده شده است. سپس از دو تابع تاوان SCAD برای انتخاب متغیر و تشخیص ساختار مدل استفاده شده است. 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. 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 آنها در جدول ۸ گزارش شده است. نمودار توابع اثرات ژن‌های انتخاب شده را نیز در شکل ۲ می‌توان ملاحظه نمود.

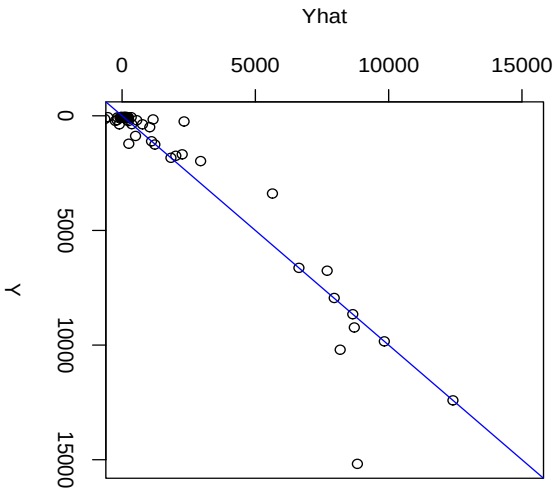
برای ارزیابی روش دو مرحله‌ای ارائه شده، از اعتبارسنجی متقابل استفاده می‌کنیم. ابتدا یک مشاهده را به عنوان داده آزمون و سایر مشاهدات را به عنوان داده‌های آموزشی در نظر بگیرید. پس از برازش مدل با استفاده از داده‌های آموزشی، مقدار متغیر پاسخ به ازای داده آزمون پیش بینی می‌شود. این روند برای تمام مشاهدات تکرار شده و نمودار مقادیر مشاهده شده متغیر پاسخ در مقابل مقادیر پیش بینی شده در شکل ۳ رسم شده است. با توجه به نمودار، روش معرفی شده در پیش بینی مقادیر ALT عملکرد خوبی دارد.



شکل ۲: توابع رگرسیونی برازش‌شده برای ژن‌های مهم

بحث و نتیجه‌گیری

در این مقاله، یک روش دو مرحله‌ای برای انتخاب متغیر و تشخیص ساختار در مدل‌های جمعی خطی-جزیی با بعد بالا ارائه شد. در مرحله اول، یک روش غربالگری مستقل مطمئن برای کاهش بعد مدل استفاده شد که در آن متغیرهای توضیحی براساس میزان همبستگی فاصله‌ای آنها با تابع توزیع حاشیه‌ای متغیر پاسخ رتبه‌بندی می‌شوند. کارایی این روش غربالگری در مطالعه‌ای 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 بالا،



شکل ۳: نمودار مقادیر مشاهده شده متغیر پاسخ در مقابل مقادیر پیش بینی شده

معمولا 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 تاوان 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 مقاله مدل جمعی خطی-جزیی مد نظر بوده است اما روش ارائه شده برای غربالگری و تشخیص ساختار را می‌توان به مدل‌های نیمه‌پارامتری دیگر مانند مدل‌های با ضریب متغیر خطی-جزیی و مدل‌های جمعی تعمیم یافته خطی-جزیی تعمیم داد. همچنین لازم به ذکر است که در این مقاله متغیرهای توضیحی و پاسخ صرفاً از نوع کمی در نظر گرفته شده است. موضوع غربالگری در حالتی که این متغیرها از نوع کیفی و چند سطحی است، می‌تواند موضوعی برای تحقیقات آینده در نظر گرفته شود. البته تحت یک روش غربالگری متفاوت، مسئله غربالگری متغیرهای کیفی توسط هوآنگ و همکاران (۲۰۱۴) مورد بررسی قرار گرفته است.

در خصوص تقلیل بعد بالای فضای متغیر های توضیحی به بعد مرتبه ی d توسط روش غربالگری بکار رفته در این مقاله می‌توان گفت که انتخاب d، همانند انتخاب پارامتر تاوان 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 تاوان در تشخیص ساختار مدل، مسئله دیگری است که بایستی به آن توجه شود. اگر چه ما از تابع تاوان SCAD استفاده کرده‌ایم، اما می‌توان از ترکیب روش غربالگری با سایر توابع تاوان نظیر LASSO، MCP یا LASSO تطبیقی نیز استفاده کرد.

تقدیر و تشکر

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

مراجع

Bushel, P., Wolfinger, R. D., Gibson, G. (2007), Simultaneous Clustering of Gene Expression Data with Clinical Chemistry and Pathological Evaluations Reveals Phenotypic Prototypes, *BMC Systems Biology*, **1**. doi:10.1186/1752-0509-1-15.

Cheng, Q. (2010), A Sparse Learning Machine for High-dimensional Data with Application to Microarray Gene Analysis, *IEEE/ACM transactions on computational biology and bioinformatics* , **7**, 636-646.

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.

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

Fan, J., and Lv, J. (2008), Sure Independence Screening for Ultrahigh Dimensional Feature Space, *Journal of the Royal Statistical Society: Series B*, **70**, 849-911.

Fan, J., Samworth, R. J., and Wu, Y. (2009), Ultrahigh Dimensional Feature Selection: Beyond the Linear Model, *Journal of Machine Learning Research*, **10**, 1829-1853.

Fan, J., Feng, Y., and Song, R. (2011), Nonparametric Independence Screening in Sparse Ultrahigh-Dimensional Additive Models, *Journal of the American Statistical Association*, **106**, 544-557.

Fan, J., and Song, R. (2010), Sure Independence Screening in Generalized Linear Models with NP-dimensionality, *Annals of Statistics*, **6**, 3567-3604.

Ferraty, F., and Hall, P. (2015), An Algorithm for Nonlinear, Nonparametric Model Choice and Prediction, *Journal of Computational and Graphical Statistics*, **24**, 695-714.

Guha, S., and Baladandayuthapani, V. (2016), A Nonparametric Bayesian Technique for High-dimensional Regression. *Electronic Journal of Statistics*, **10**, 3374-3424.

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 & Data Analysis*, **65**, 56-67.

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, 533–550.

Huang, J., Wei, F., and Ma, S. (2012), Semiparametric Regression Pursuit, *Statistica Sinica*, **22**, 1403-1426.

Huang, D., Li, R., and Wang, H. (2014), Feature Screening for Ultrahigh Dimensional Categorical Data with Applications, *Journal of Business and Economic Statistics*, **32**, 237-244.

Kazemi, M., Shahsavani, D., and Arashi, M. (2017), A Sure Independence Screening Procedure for Ultra-high Dimensional Partially Linear Additive Models, arXiv preprint arXiv:1708.08604.

Li, R.Z., Zhong, W., and Zhu, L.P. (2012), Feature Screening via Distance Correlation Learning, *Journal of the American Statistical Association*, **107**, 1129-1139.

Lian, H. (2012), Variable Selection in High-dimensional Partly Linear Additive Models, *Journal of Nonparametric Statistics*, **24**, 825-839.

Lian, H. (2012), Shrinkage Estimation for Identification of Linear Components in Additive Models, *Statistics & Probability Letters*, **82**, 225-231.

Lian, H., Chen, X., & Yang, JY. (2012), Identification of Partially Linear Structure in Additive Models with an Application to Gene Expression Prediction from Sequences. *Biometrics* **68**, 437–445.

Lian, H., Liang, H., and Ruppert, D,. (2015), Separation of Covariates into Nonparametric and Parametric Parts in High-dimensional Partially Linear Additive Models, *Statistica Sinica*, **25**, 591-607.

Liu, X., Wang, L., and Liang, H., (2011), Estimation and Variable Selection for Semiparametric Additive Partial Linear Models, *Statistica Sinica*, **21**, 1225-1248.

Lv, J., Yang, H., and Guo, C. (2016), Variable Selection in Partially Linear Additive Models for Modal Regression, *Communication in Statistics-Simulation and Computation*, DOI: 10.1080/03610918.2016.1171346.

Opsomer, J. D., and Ruppert, D. (1999), A Root-n Consistent Backfitting Estimator for Semiparametric Additive Modeling. *Journal of Computational and Graphical Statistics*, **8**, 715-732.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003). Regression Approach for Microarray Data Analysis, *Journal of Computational Biology*, **10**, 961-980.

Zhao, S. D., and Li,Y. (2012), Principled Sure Independence Screening for Cox Models with Ultra-High-Dimensional Covariates, *Journal of Multivariate Analysis*, **105**, 397–411.

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), Measuring and Testing Dependence by Correlation of Distances, *Annals of Statistics*, **35**, 2769-2794.

Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society: Series B*, **58**, 267-288.

Wang, H. (2009), Forward Regression for Ultra-high Dimensional Variable Screening, *Journal of the American Statistical Association*, **104**, 1512-1524.

Zhang, C. H. (2010), Nearly Unbiased Variable Selection under the Minimax Concave Penalty, *Annals of Statistics*, **83**, 894-942.

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**, 1099-1112.

Zhu, L.P., Li, L., Li, R., and Zhu, L.X., (2011), Model-free Feature Screening for Ultrahigh Dimensional Data, *Journal of the American Statistical Association*, **106**, 1464-1475.