
بسم الله الرحمن الرحيم

High-dimensional & screening

درس مربوطه: مدل های خطی

ارایه دهندگان:
زهرا عسگریان–پگاه طبیبی

توسعه ی تکنولوژی ،منبع نیرومند داده ، رشته کامپیوتر باعث تولید داده شده ،ذخیره سازی و فرآیند آن به صورت نمایی زیاد می شوند داده ها همه جا حضور دارند .در علم و پزشکی و بیزینس (تجارت) و سرمایه و حکومت موجود است .مثلا در آگاهی از آب و هوا و تغییر زیستی و بازاریابی کردن برای شخص

یک کاراکتر عمده داده های مدرن این است که اغلب آن ها همزمان غالب بر هزار تا میلیون خصوصیت روی هر موضوع یا شخص ضبط می شوند. این قبیل داده ها را داده های با بعد بالا می گویند.

داده های بیوتکنولوژی: (مثلا میکرو آرایی دی ان ای) در چنین داده های زیست فناوری تعداد متغیر ها از تعداد افراد درگیر آزمایش بیشتر است.

عکس ها (ویدئو): بانک اطلاعاتی بزرگی از عکس ها و جمع آوری مداوم آن در جهان است. شامل عکس های پزشکی ، عکس های فشرده نجومی ، عکس های نظارت فیلم و غیره

داده های مصرف کننده:به عنوان مثال سیستم توصیه (پیشنهاد) (برای فیلم و کتاب ها و موزیک و غیره) دسته بندی مشتری روی محصولات متفاوت جمع آوری کرده و با بعضی داده های شخصی (سن ، جنسیت،مکان)و حدس این که کدام محصول برای مشتری جالب است را با هم در نظر می گیرد.

داده های تجارت: این داده ها بیشتر تلاش برای پیش بینی دقیق تقاضا در آینده تمرکز دارد . شرکت های بیمه ایی نیز از این دسته اند .

ازدحام منبع داده ها : وبسایت ها در ضبط داده ها دخالت کرده و همراه با گسترش گوشی های موبایل قادر خواهند بود به صورت آنلاین مجموعه داده های بزرگ را ضبط کند. برای مثال آزمایشگاه پرنده شناسی و جامعه شهروندی جاناودبان (پرنده شناس) مشترکا یک برنامه

Crowdsourcing را عرضه کردند.

http://ebird .org

برای تماشاگران پرنده از شمال آمریکا جذاب است که از طریق آنلاین تمام پرنده هایی که دیده و یا در طی آخرین جلسه تماشای پرندگان در زیستگاه خودشان فیلم گرفته و ضبط کنند.هدف این برنامه تماشای فراوانی پدیده ها و سیر تکاملی آن ها در سراسر آمریکا شمالی است در سال 2014 این برنامه شامل 10 هزار نفر همراه بود تاکنون میلیون ها مشاهده را ضبط کرده است

مقدمه

دلیل انتخاب مدل

انواع انتخاب مدل

Tast based (stepwise , Autometrics)

Penalty based

Screening based

داده های با بعد بالا فرصتی پدید آورده اند که روش های نوین انتخاب متغیر آماری و یادگیری ماشین بتوانند هنگامی که

p بسیار بزرگتر از مشاهده است.در این نوع داده ها

P یک تابع نمایی از مشاهده است به عبارت دیگر

$Log(p)=o(n^{\alpha})$

که در آن $0>\alpha$

قالب کلی راه حل به این صورت است که ابتدا توسط یک روش غربالگری مستقل کارا تعداد متغیر ها کاهش یافته و سپس با یکی از روش های پنالتی میکس می شود و انتخاب مدل انجام می شود.

معیار های مورد استفاده :

Signal to noise (SNR)

$$SNR=\frac{var(x'B)}{var\ (\varepsilon)}$$

معیاری جهت نمایش میزان سیگنال مفید در مقابل سیگنال مزاحم یا نویز در سیستم است .مقدار کمتر از 12 نشان دهنده مشکل جدی نویز ،مقدار بالاتر از 20 رضایت بخش و مقدار بالاتر از 30 مناسب است .

Correlation between variable (RHO)

رتبه بندی ضریب همبستگی به مجموعه (0و0.3و0.6و0.9)

Score test

اگر $\beta_j = 0$یعنی j امین متغیر کمکی در نتیجه تاثیری ندارد پس مهم نیست .

یک مجموعه مهم ضرایب بوسیله

M=$\{j, \beta_j \neq 0\}$

ما فرض می کنیم که اندازه این مجموعه کوچک است.

پیشنهاد Score test screening به صورت زیر است:
همه ی متغیر های کمکی مرکزی شده و استاندارد شده با میانگین صفر و واریانس 1 است.
برای هر متغیر کمکی j ساختار معادله برآورد برای β_j فرض میشود که تمام متغیر های کمکیدیگر به نتیجه وابسته نباشد.
حاشیه ای معادله برآورد بویسله $U_j^M(\beta_j)$ معنی دار می شود.

نگه داشتن پارامتر های

$\widehat{M}=\{j: |U_j^M(0)| \geq \gamma\}$

برای آستانه γ.

هر $|U_j^M(0)|$ یک شمارنده آماره

است بنابراین یک آماره معقول برای غربالگری است.

Score test

Distance correlation sure independent screening

با توجه به اهمیت روش‌های غربالگری آزاد-مدل، یک روش دیگر موسوم به غربالگری مستقل مطمئن براساس همبستگی فاصله‌ای (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 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.

غربالگری مستقل مطمئن نیرومند بر اساس همبستگی فاصله‌ای

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

$$dcov^2(U,V) = S_1 + S_2 - 2S_3$$

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.

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: 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.

مقادیر S_3, S_2, S_1 با روش گشتاوری به صورت

$$\hat{S}_{k,1} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \| X_{ik} - X_{jk} \|_q \| F_n(Y_i) - F_n(Y_j) \|_r,$$
$$\hat{S}_{k,2} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \| X_{ik} - X_{jk} \|_q \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \| F_n(Y_i) - F_n(Y_j) \|_r,$$
$$\hat{S}_{k,3} = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \| X_{ik} - X_{lk} \|_q \| F_n(Y_j) - F_n(Y_l) \|_r,$$

برآورد می‌شود، که در آن $F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq y)$ تابع توزیع تجربی Y است. بنابراین، برآورد $dcov^2\big(X_k, F(Y)\big)$ به صورت

$$\widehat{dcov^2}(X_k, F(Y)) = \hat{S}_{k,1} + \hat{S}_{k,2} - 2\hat{S}_{k,3}$$

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 $\hat{\omega}_k = \widehat{dcorr}\left(X_k, F(Y)\right)$. 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

$$\hat{\mathbb{M}}_{\nu_n} = \{1 \leq k \leq p : \hat{\omega}_k \geq \nu_n\}, \qquad (۴)$$

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.

استفاده از روش غربالگری RDC-SIS مستلزم تعیین یک مقدار آستانه 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، 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 به کاربردن 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 به کار 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 $d = ۲\lfloor n/\log(n) \rfloor$ استفاده می‌شود.

• گام ۲: فرض کنید $\mathbf{X}_1$ ماتریس طرح متناظر با متغیرهای مجموعه $\mathcal{M}_1$ و $\mathbf{X}_1^c$ ماتریس طرح متناظر با متغیرهای مجموعه $\mathcal{M}_1^c$ است. ماتریس $\mathbf{X}_{new} = \{I_n - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\}\mathbf{X}_1^c$ 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_2 متغیر توضیحی انتخاب می‌شوند و مجموعه اندیس متغیرهای انتخاب شده را با $\mathcal{M}_2$ نشان می‌دهیم. مجموعه $\mathcal{M}_1$ را با $\mathcal{M}_1 \cup \mathcal{M}_2$ بروز رسانی کنید.

• گام ۳: گام ۲ را $k-1$ بار تکرار کنید تا تعداد متغیرهای توضیحی انتخاب شده از d تجاوز کند، یعنی $d_1 + \ldots + d_k \geq d$. در پایان مجموعه متغیرهای توضیحی انتخاب شده $\mathcal{M}_1 \cup \cdots \cup \mathcal{M}_k$ 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 مقاله ژو و همکاران (۲۰۱۱) برای غربالگری آزاد-مدل استفاده شده است. در اینجا مقدار d_1 در گام ۱ توسط کاربر تعیین می‌شود. در عمل، d_1 به عنوان یک پارامتر تنظیم‌کننده در نظر گرفته می‌شود و مقدار بهینه آن با مینیم کردن میانگین مربع خطای پیش‌بینی تعیین می‌گردد.

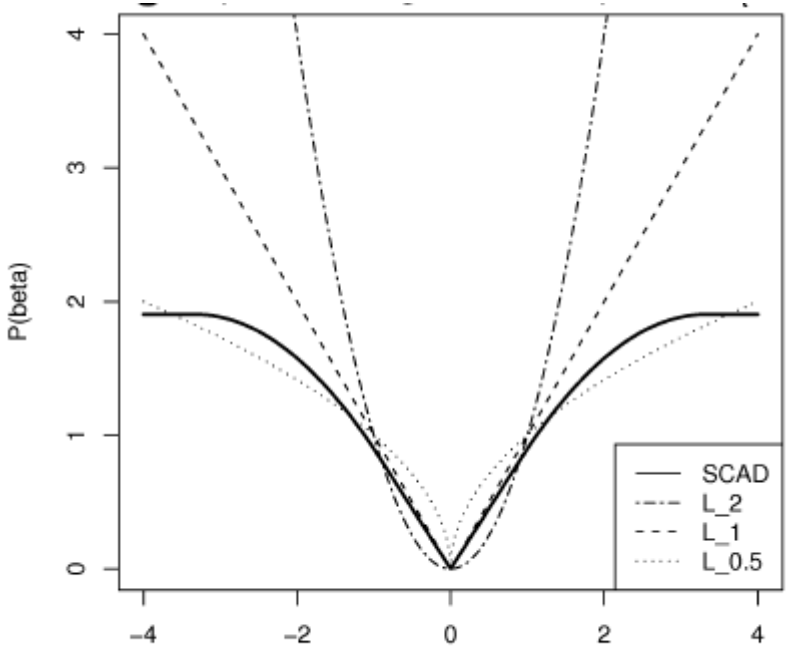
تابع تاوان SCAD

فن و لی [۸] ویژگی‌های یک تابع تاوان خوب را به صورت (i) ناااریبی، (ii) تُنکی و (iii) پیوستگی بیان کردند. هر کدام از این ویژگی‌ها در ارتباط با یکی از روش‌های معرفی شده‌ی قبلی می‌باشند. همانطور که در بخش سوم اشاره شد، روش LASSO با تاوانیدن ضرایب بزرگ سبب بیش‌برآوردی می‌گردد و در نتیجه LASSO از ویژگی ناااریبی بی‌بهره است. رگرسیون ستیغی از ویژگی تنکی بی‌بهره است زیرا توانایی خارج کرن ضرایب بی‌اهمیت از مدل را ندارد. ویژگی سوم نیز به معیارهایی نظیر AIC اشاره دارد که استفاده‌ی از آنها موجب بی‌ثباتی در انتخاب متغیر می‌شود [۳]. دلیل این بی‌ثباتی نیز عدم استفاده از یک تابع تاوان پیوسته است [۸].

با توجه به این سه ویژگی، فن و لی [۸] تابع تاوان SCAD را به صورت زیر معرفی کردند:

$$P_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & \circ \le |\beta_j| < \lambda \\ \frac{(a^2-1)\lambda^2-(|\beta_j|-a\lambda)^2}{2(a-1)} & \lambda \le |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| \ge a\lambda \end{cases} \quad (۷)$$

به طوریکه $\lambda > 0$ و $a > 2$. پارامتر a شکل تابع تاوان SCAD را کنترل می‌کند به طوری که هر چه $a \to \infty$ شکل تابع SCAD به شکل L_1 میل می‌کند و عملکرد تابع تاوان SCAD مشابه عملکرد LASSO خواهد شد. نمودار SCAD به همراه نمودار توابع L_1، $L_{0/5}$ و L_2 در شکل ۱ نشان داده شده است. تابع تاوان SCAD برخلاف LASSO، برای ضرایبی که از یک مقدار به خصوص $(a\lambda)$ بزرگتر باشند تاوان ثابتی را در نظر می‌گیرد (شکل ۱ را ببینید.) . در تابع (۷) دو پارامتر a و λ وجود دارد. فن و لی [۸] با یک مطالعه‌ی بیزی نشان دادند که مقدار ۳/۷ برای a یک انتخاب مناسب است. به همین دلیل تابع SCAD را تنها با اندیس λ نشان‌گذاری می‌کنند.



فن و لی [۸] برای برآورد β از مجموع توان‌های دوم باقیمانده‌ی تاوانیده‌ی زیر:

$$\frac{1}{n}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+\sum_{j=1}^{d}P_\lambda(|\beta_j|), \qquad (۸)$$

یا به طور معادل

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+n\sum_{j=1}^{d}P_\lambda(|\beta_j|), \qquad (۹)$$

استفاده کردند. که در آنها $P_\lambda(|\beta_j|)$ در (۷) تعریف شده است.

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.

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

$\varepsilon \sim t(1)$					$\varepsilon \sim N(0,1)$					d	n	p
S	p_4	p_3	p_2	p_1	S	p_4	p_3	p_2	p_1			
0.41	0.45	0.92	1.00	0.84	0.53	0.60	0.96	1.00	0.91	d_1	100	
0.58	0.62	0.96	1.00	0.94	0.73	0.75	0.98	1.00	0.98	d_2		
0.67	0.70	0.97	1.00	0.96	0.81	0.81	0.99	1.00	1.00	d_3		
0.93	0.93	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	d_1	200	1000
0.97	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_2		
0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_3		
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_1	400	
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_2		
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_3		
0.42	0.55	0.90	1.00	0.77	0.55	0.61	0.96	1.00	0.92	d_1	100	
0.58	0.66	0.96	1.00	0.89	0.71	0.72	0.99	1.00	0.96	d_2		
0.65	0.70	0.98	1.00	0.92	0.78	0.79	0.99	1.00	0.99	d_3		
0.92	0.97	1.00	1.00	0.99	0.97	0.97	1.00	1.00	1.00	d_1	200	2000
0.97	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_2		
0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_3		
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_1	400	
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_2		
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	d_3		

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 گزارش شده است. همان‌طور که ملاحظه می‌شود، با افزایش مقدار d عملکرد RDC-SIS بهبود می‌یابد. برای $n = 100$، 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 را پیشنهاد می‌شود. اما برای $n = 200, 400$ و هر دو نوع توزیع خطا، هر یک از مقادیر d_1، d_2 و d_3 را می‌توان استفاده کرد.

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

خطا	σ	روش	M					P					S
			٪۵	٪۲۵	٪۵۰	٪۷۵	٪۹۵	۱	۲	۳	۴	۵	
$N(۰,۱)$	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
(۳٤(	۱	RDC-SIS	۵	۵	۵	۵	۱۹	۱/۰۰	۱/۰۰	۱/۰۰	۱/۰۰	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

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 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 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.

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

خطا	روش	M					P				S
		٪۵	٪۲۵	٪۵۰	٪۷۵	٪۹۵	۱	۲	۳	۴	
$N(۰,۱)$	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	۱/۰۰	۰/۸۵	۰/۸۵
(۳t(	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	۱/۰۰	۰/۸۷	۰/۸۷

نتیجه:

در مرحله اول، یک روش غربالگری مستقل مطمئن برای کاهش بعد مدل استفاده شد که در آن متغیرهای توضیحی براساس میزان همبستگی فاصله ای آنها با تابع توزیع حاشیه ای متغیر پاسخ رتبه بندی می شوند .کارایی این روش غربالگری در مطالعه ای شبیه سازی و تحلیل یک مجموعه داده واقعی مورد ارزیابی قرار گرفت که نتایج حاکی از عملکرد مطلوب روش ارائه شده است .

.

در اینجا 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، می تواند موضوعی برای تحقیقات آینده در نظر گرفته شود.

در خصوص تقلیل بعد بالای فضای متغیر های توضوحی به بعد مرتبه d توسط روش غربالگری بکار رفته در اینجا میتوان گفت که انتخاب d همانند انتخاب پارامتر تاوان در روش های انقباضی دارای اهمیت بسزایی است

Refrence:
1- Introduction to high-dimensional statistics chap man & hall chapter 1 peges1 to 3
2- scare test variable screening paper sihai dare zhao
3- sezkely G J .Rizzo , M.L..Measuring and testing dependence by carrelation of distance page 2769-2794

4- FAN and LV(2008) sure independence screening for ultrahigh dimensiond feuture space
5- two tales of variable selection for high dimensional data screening and model building page 23-29
6- انتخاب متعیر و تشخیص ساختار در بعد بالا برای مدل های جمعی خطی –جزیی

7- انتخاب متغیر با استفاده از تابع توان

تمام

با تشکر از همراهی شما