تمرین کامپیوتری اوّل ـ استنتاج علّی

بهراد منیری
۹۵۱۰۹۵۶۴
bemoniri@live.com

دانشکده‌ی مهندسی برق ـ دانشگاه صنعتی شریف

۱ سوال اول

۱.۱ بخش الف

- **(مدل خطی با نویز گاوسی)** در این مدل داریم:

$$X \to Y : \begin{cases} X := N_x \\ Y := X + N \end{cases} \qquad N \perp\!\!\!\perp N_x$$

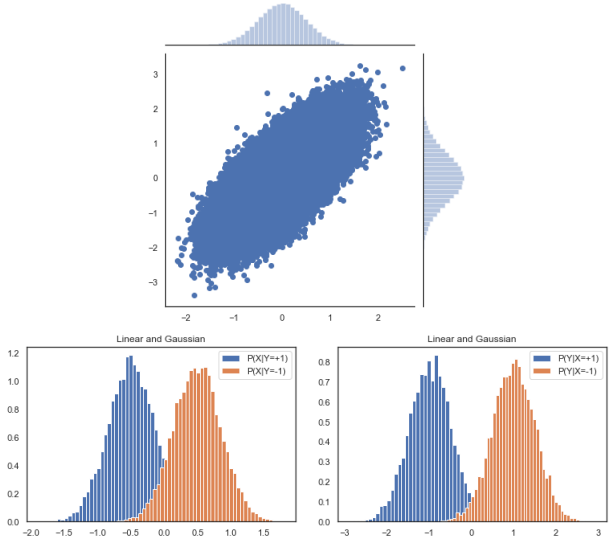
فرض کنید $N_x : \text{Normal}(۰, \frac{۱}{۲})$ و $N = \text{Normal}(۰, \frac{۱}{۲})$. با این فرض، هر ترکیب خطی X و Y نرمال است، در نتیجه آنها مشترکاً نرمال هستند.

$$\forall \alpha, \beta : \quad \alpha X + \beta Y = \text{Normal} \to (X, Y) : \text{Multivariable Normal}(۰, ۰; ۱, \frac{۱}{\sqrt{۲}}; \frac{۱}{\sqrt{۲}})$$

توزیع‌های شرطی یک توزیع مشترکاً نرمال نرمال است.

$$\begin{cases} P(Y|X = x) = \text{Normal}(x, \frac{۱}{۲}) \\ P(X|Y = y) = \text{Normal}(-\frac{y}{۲}, \frac{۱}{۲\sqrt{۲}}) \end{cases}$$

شکل (۱) نمودار توزیع‌های مطرح شده هستند.



شکل ۱: توزیع‌ها با فرض $\sigma_x = \sigma_N = ۰/۵$

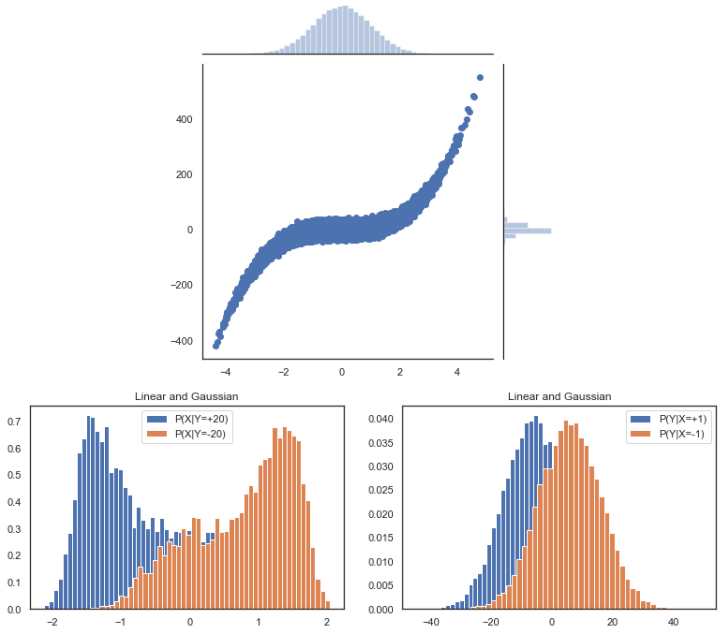
- **(مدل غیرخطی با نویز گاوسی)** در این مدل داریم:

$$X \to Y : \begin{cases} X := N_x \\ Y := X + ۵X^۳ + N \end{cases} \qquad N \perp\!\!\!\perp N_x$$

شکل (۲) نمودارهای خواسته شده‌ی مربوط به این SCM با فرض

$$N_x = \text{Normal}(۰, ۱) \quad N = \text{Normal}(۰, ۱۰)$$

است.

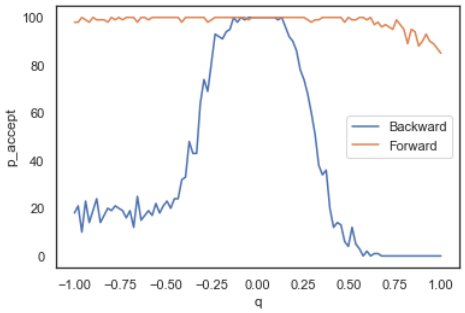


شکل ۲: توزیع‌ها با فرض $\sigma_x = ۱, \quad \sigma_N = ۱۰$

۲.۱ بخش ب

در این بخش می‌خواهیم تاثیر خطی و یا گاوسی بودن مدل را در تشخیص جهت درست علیّت بررسی کنیم.

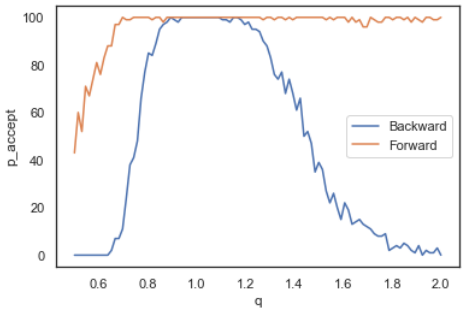
- **بررسی اثر خطی‌بودن** برای بررسی اثرات غیرخطی بودن مدل، b را در بازه‌ی $[-۱, ۱]$ تغییر می‌دهیم. نویز مدل را گاوسی $(q = ۱)$ در نظر می‌گیریم. برای مقدار مختلف b، در دو جهت رگرسیون انجام می‌دهیم و سپس استقلال Residue از علّت را بررسی می‌کنیم. در این تمرین، معیار ما برای استقلال، آزمون HSIC با سطح اطمینان ۲ درصد است. برای هر b، صد این کار را تکرار می‌کنیم و نمودار درصد پذیرش فرض استقلال بر حسب b را رسم می‌کنیم. شکل (۳) نمایش‌دهنده‌ی نتایج ماست.



شکل ۳: بررسی اثر خطی بودن بر تشخیص جهت علیّت

این نتایج با انتظارات ما کاملاً هم‌خوانی دارند زیرا در حالت خطی، و با توجه به گاوسی بودن N و X انتظار داریم در تشخیص جهت علیّت دچار اشتباه شویم (در مدل خطی‌ـ گاوسی، با روش مطرح شده نمی‌توان جهت را تشخیص داد و هر دو جهت ویژگی استقلال را دارا می‌باشند.)

- **بررسی اثر گاوسی بودن** برای بررسی تاثیر گاوسی بودن نویز، مدل را خطی فرض کرده $(b = ۰)$ و مقدار q را در بازه‌ی $[۰٫۵, ۲]$ تغییر می‌دهیم. مشابه بخش قبل، این بار برای مقادیر مختلف q، در دو جهت رگرسیون انجام داده و نمودار درصد پذیرش فرض استقلال بر حسب q را رسم می‌کنیم. شکل (۴) را ببینید.



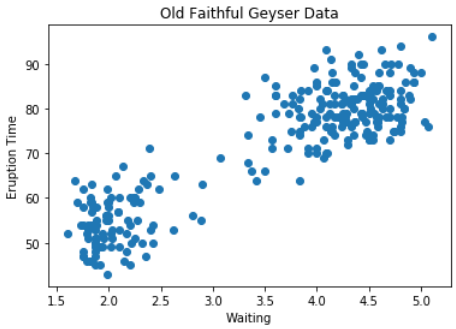
شکل ۴: بررسی اثر گاوسی بودن بر تشخیص جهت علیّت

نتایج با انتظارت هم‌خوانی دارند زیرا طبق آنچه در کلاس بررسی کردیم، در حالت خطی‌ـ گاوسی، می‌توان SCM را، بدون بر هم خوردن شرط استقلال در هر دو جهت نوشت.

۳.۱ بخش ج

۱.۳.۱ دیتاست آبفشان

ابتدا داده‌ها را در یک فضای دوبعدی رسم می‌کنیم تا شهود بهتری نسبت به مسأله پیدا کنیم، شکل (۵).



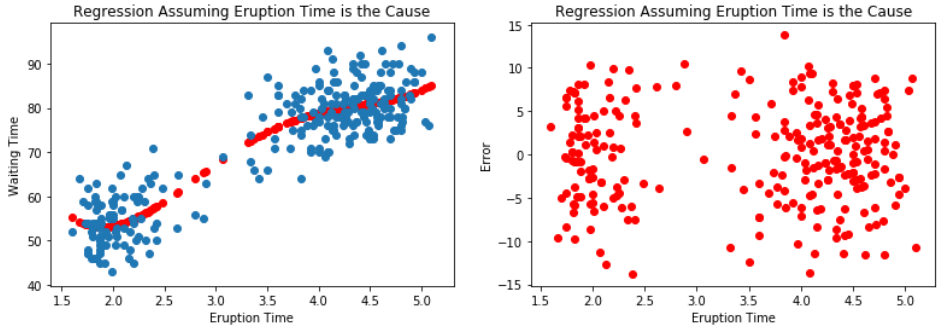
شکل ۵: رسم دیتای مربوط به آبفشان

برای تشخیص جهت درست علّیت، با فرض ANM، مطابق بخش‌های قبل یک‌بار هر یکی از دو متغیّر را علّت فرض کرده و رگرسیون‌های غیرخطی مربوط را انجام می‌دهیم.

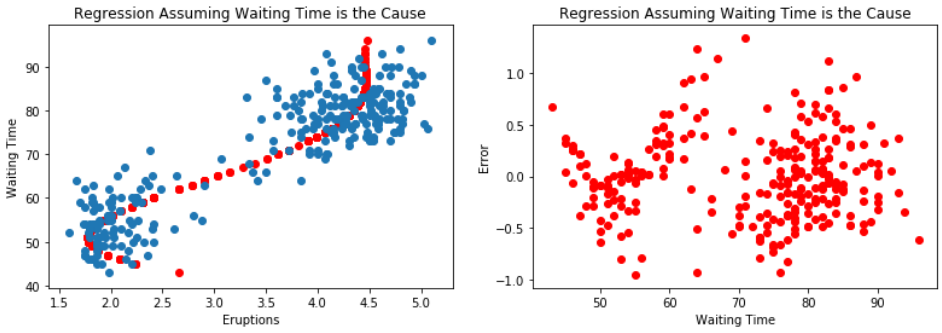
$$\begin{cases} Y = \hat{f}_1(X) + \hat{N}_1 & : X \to Y \text{ با فرض} \\ X = \hat{f}_2(Y) + \hat{N}_2 & : Y \to X \text{ با فرض} \end{cases}$$

انتظار داریم که در جهت درست علّیت، N و متغیّری که عنوان علّت در نظر گرفته‌ایم مستقل شوند. با انجام این فرآیند در هر دو جهت و اعمال آزمون استقلال هیلبرت–اشمیت برای دو کمیت مذکور در هر جهت، جهت درست علّیت را تشخیص می‌دهیم.

شکل (۶) رگرسیون با فرض اینکه زمان فوران کنونی علّت فاصله‌ی زمانی تا فوران بعدی است و شکل (۷) نیز رگرسیون با فرض معکوس است.



شکل ۶: رگرسیون با فرض اینکه زمان فوران کنونی علّت فاصله‌ی زمانی تا فوران بعدی است.

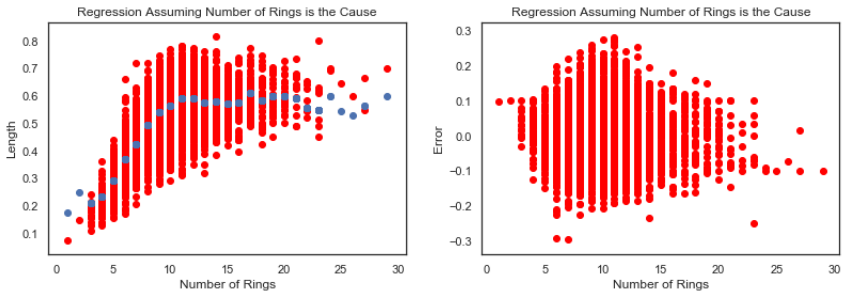


شکل ۷: رگرسیون با فرض اینکه فاصله‌ی زمانی تا فوران بعدی علّت زمان فوران کنونی است.

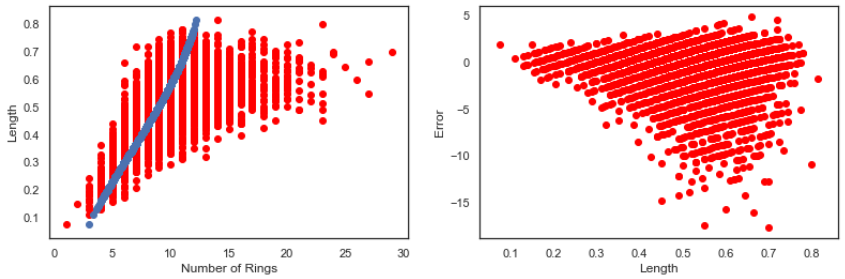
با این کار و انجام آزمون فرضیه‌ی HSIC، متوجه می‌شویم با فرض «**زمان فوران کنونی علّت فاصله‌ی زمانی تا فوران بعدی است**» زیرا همان‌طور که در شکل (۶) دیده می‌شود، بعد از رگرسیون فاصله‌ی زمانی تا فوران بعدی بر حسب طول زمان فوران فعلی، مقدار residue این رگرسوراز فاصله زمانی فوران فعلی مستقل است. این موضوع تا حدی بدیهی است زیرا فوران بعدی، بعد از فوران فعلی رخ داده و نمی‌تواند تاثیر علّی بر فوران فعلی داشته باشد.

۲.۳.۱ دیتاست صدف

در این دیتاست قصد داریم جهت علّی بین طول این نوع صدف و تعداد حلقه‌های آن را پیدا کردیم. می‌دانیم که تعداد حلقه‌ها رابطه‌ای تقریباً غیرتصادفی با سن صدف دارد. از نظر شهودی، سن صدف(یا به عبارتی تعداد حلقه‌ها) علّت طول صدف است. در این بخش این موضوع را به کمک دیتای داده‌شده تایید می‌کنیم. یک بار با فرض سن صدف(یا به عبارتی تعداد حلقه‌ها) علّت طول صدف، الگوریتم را اجرا می‌کنیم. شکل (۸) نتایج این کار هستند. شکل (۹) نیز نتایج اجرای الگوریتم با فرض طول صدف علّت تعداد حلقه‌ها (یا به عبارتی سن صدف) هستند. با توجه این نتایج و اعمال آزمون استقلال و همچنین با توجه به نمودارهای residue بر حسب متغییر علّت فرض شده، فرض ما



شکل ۸: سن صدف(یا به عبارتی تعداد حلقه‌ها) علّت طول صدف

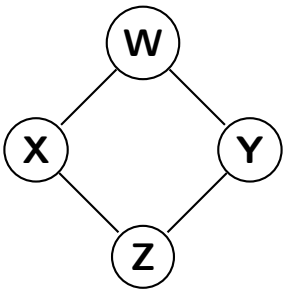


شکل ۹: طول صدف علّت تعداد حلقه‌ها (یا به عبارتی سن صدف)!

مبنی بر اینکه **سن صدف(یا به عبارتی تعداد حلقه‌ها) علّت طول صدف** تایید می‌شود.

۲ سوال دو

می‌دانیم داده‌های این سوال از یک SCM تولید شده‌اند و اسکلت گراف این SCM در شکل روبروآمده است. حال سعی می‌کنیم با چند بار تکرار فرایندی که در سوال قبلی طی شد، در این سوال نیز جهت‌های درست گراف را تشخیص دهیم. این کار را در چند مرحله انجام می‌دهیم. یعنی با فرض‌های مختلف علّیت، رگرسیون انجام داده و چک می‌کنیم که آیا residue این رگرسیون مستقل از علت‌های مفروض هست یا خیر. در جهت درست علّیت این شرط برقرار است.

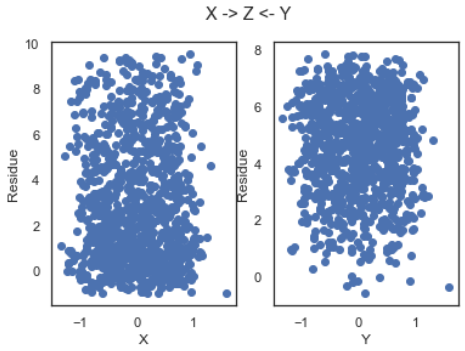


شکل ۱۰: اسکلت گراف مربوط به داده‌ها

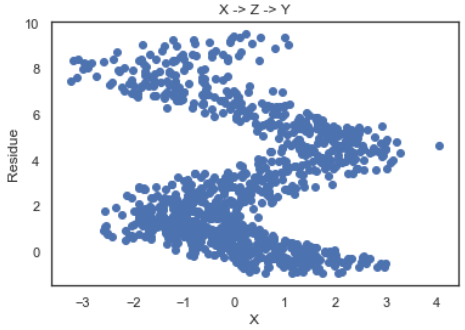
- **تعیین جهت یال‌های متصل به Z** در صورتی که فرض کنیم هر دو یال به این راس وارد می‌شود و رگرسیون $z = \hat{f}(x, y) + \epsilon$ را حساب کنیم، دیده‌می‌شود که داریم:

$$\epsilon \perp\!\!\!\perp X, \quad \epsilon \perp\!\!\!\perp Y$$

که این نشان می‌دهد این، جهت درست علّیت است. با فرض‌های دیگر، نتایجی مخالف انتظارمان از جهت درست علّیت بر خواهیم داشت. در شکل (۱۱) فرض شده است که $X \to Z \leftarrow Y$ و در شکل (۱۲) فرض شده $X \to Z \to Y$. دیده می‌شود در تمام حالت به جز شکل (۱۱) تست‌های استقلال نتایجی سازگار با گراف ندارند و بنابراین حالت (۱۱) را به عنوان حالت صحیح می‌پذیریم.



شکل ۱۱: مقدار ی Residue $Z = \hat{f}(X, Y)$ بر حسب X و Y با فرض $X \to Z \leftarrow Y$

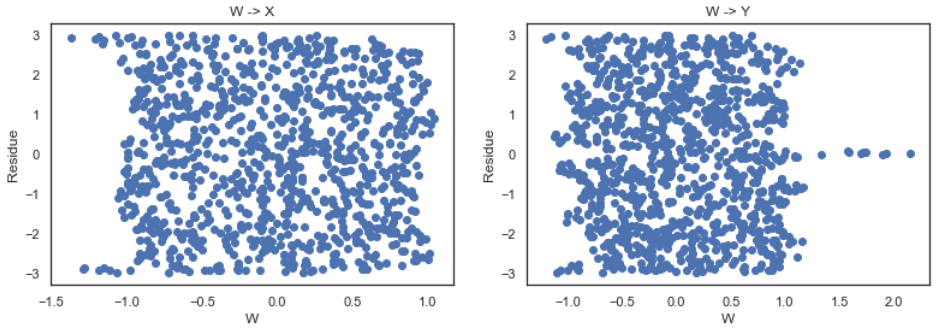


شکل ۱۲: مقدار Residue ی $Z = \hat{f}(X)$ بر حسب X با فرض $X \to Z \to Y$

- **تعیین جهت یال‌های متصل به W** ابتدا فرض می‌کنیم که هر دو یال از W خارج شوند. با این فرض برای X و Y داریم:

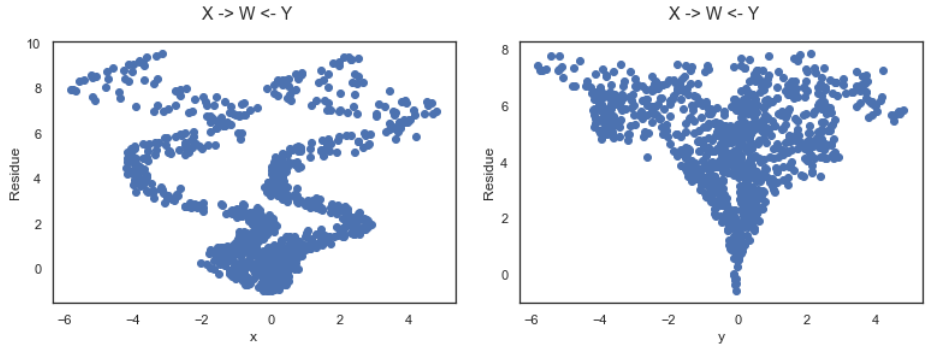
$$\begin{cases} X := f_1(W) + N_x \\ Y := f_2(W) + N_y \end{cases} \quad W \perp\!\!\!\perp N_x, \;\; W \perp\!\!\!\perp N_y$$

حال سعی می‌کنیم این SCM را بر دیتای داده شده برازش کنیم. با انجام دو رگرسیون و بررسی استقلال Residue از علّت، به نتایج زیر می‌رسیم.



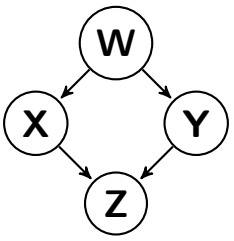
شکل ۱۳: مقدار Residue ی $Y = \hat{f}(W)$ و $X = \hat{g}(W)$ بر حسب W با فرض $Y \leftarrow W \rightarrow X$

این تنها حالتی است که با Confidence Level دو درصد، تمام تست‌های استقلال منجر به نتیجه‌ی مورد نظرمان می‌شوند. شکل (۱۴) یک فرض اشتباه است که در نهایت، Residue ها از کمیّت‌هایی که علت W در نظر گرفته شده‌اند مستقل نشده است.



شکل ۱۴: مقدار Residue $W = \hat{f}(X, Y)$ بر حسب X و Y با فرض $Y \rightarrow W \leftarrow X$

به طور کلی، ایده‌ی ما در این سوال این بود که برای راس‌ها جهت‌های مختلف فرض کرده و سپس با انجام رگرسیون‌های مربوطه و بررسی روابط استقلال، صحت جهت انتخاب شده را بررسی کنیم. در نهایت با توجه به موارد مطرح شده، گراف زیر را به عنوان یافته‌ی خود اعلام می‌کنیم.



شکل ۱۵: گراف نهایی