بسمه تعالی

استنتاج علّی
گزارش پروژه تحقیقاتی نظری

نیمسال اول ۹۸-۹۷
شماره دانشجویی: ۹۵۱۰۹۵۶۴

Additive Noise Models: Identifiability, Learning Algorithms, Hidden Variables, Cycles and Time Series

بهراد منیری

۱ مقدمه

در این پروژه به بررسی مدل‌های نویز جمعی، Additive Noise Models، خواهیم پرداخت. به طور خاص، قضایای قابل شناسایی بودن جهت درست علّی را بررسی خواهیم کرد، مروری بر الگوریتم‌های یادگیری ساختار علّی خواهیم داشت و همچنین درباره‌ی مدل‌های نویز جمعی در وجود متغیرهای پنهان، حلقه‌های فیدبک و در سری‌های زمانی بحث خواهیم کرد.

مدل‌های نویز جمعی از اهمیت بالایی در استنتاج علّی برخوردارند. این روش که معادل جست و جو در یک فضای توابع محدود شده است شباهیت زیادی با یادگیری مشاین پیدا می‌کند. وجود قضایای قابل شناسایی بودن گراف با داشتن توزیع مشاهداتی که به ما امکان دستیابی به اطلاعاتی بیش از کلاس هم‌ارزی مارکوف گراف مولد را می‌دهد، یکی از جذابیت‌های مدل های نویز جمعی است [۲]. الگوریتم‌های بهینه‌تری نسبت به الگوریتم‌هایی مانند PC برای یادگیری ساختار علّی SCM هایی که در آن‌ها مدل‌های نویز جمعی وجود دارد [۵]. به نظر می‌رسد که در جهان واقعی، خیلی از پدیده‌ها از مدل نویز جمعی پیروی می‌کنند [۵، ۶]. به طرز شگفت‌آوری، مدل‌های مبتنی بر نویز جمعی در یادگیری سری‌های زمانی نیز توانسته‌اند به نتایج تجربی امیدوارکننده‌ای دست پیدا کنند [۴].

۲ تعریف مسئله

یک SCM، C، در نظر بگیرید. می‌گوییم این SCM از مدل نویز جمعی پیروی می‌کند اگر

$$S_j: \quad f_j = f_j(\mathbf{PA}_j) + N_j, \qquad j = 1, 2, ..., p \tag{1}$$

که در آن $\mathbf{PA}_j$ مجموعه‌ی والدین X_j هستند و متغیرهای نویز، N_j دارای چگالی‌احتمالی اکیداً مثبت می‌باشند و مستقل‌اند. همچنین در همه‌ی بخش‌ها فرض می‌شود گراف مولد داده‌ها یک DAG است مگر به طور صریح خلاف این موضوع ذکر شود.

۳ قضایای قابل‌شناسایی بودن

در این بخش به بررسی مدل‌های نویز جمعی در حالاتی که هیچ Common Cause مشاهده‌نشده‌ای در سیستم وجود ندارد می‌پردازیم. در مدل‌های نویز جمعی، هدف محدود کردن مجموعه‌ی توابعی است که در آن به جست‌وجوی توابع تولید کننده SCM می‌گردیم. مراجع ما در این بخش، مقالات [۲]، [۵] و [۶] هستند. این مقالات به بررسی قضایای Identifiability در مدل‌های نویز جمعی پرداخته اند. ما در این بخش به بررسی قضایای مطرح شده در این مقالات پرداخته و در آخر محدودیت‌های آن‌ها را با هم مقایسه می‌کنیم.

قضیه ۱.۳. درمدل‌های نویز جمعی، شرط *Causal Minimality* معادل این است که توابع f_j نسبت به هیچ یک از متغیرهایشان ثابت نباشند.

۱.۳ حالت دو متغیره

Hoyer et al. در مقاله‌ی [۲] قضیه‌ی زیر را در مورد قابل شناسایی بودن ANM در حالت دو متغیره اثبات می‌کند. در ادامه به بررسی دقیق این قضیه پرداخته و بحث خواهیم کرد که در چه شرایطی، در یک مدل جمعی، از روی چگالی احتمال مشترک قادر به شناسایی کامل جهت علّی نیستیم.

قضیه ۲.۳. فرض کنید داشته باشیم

$$\begin{cases} X = N_x \\ Y = f(X) + N_y \end{cases}$$

اگر برای هر x, y با شرط $\nu''(y - f(x))f'(x) \neq 0$ حل معادله‌ی دیفرانسیل زیر نباشد:

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'} , \tag{2}$$

که در آن $\xi := \log(p_X)$ و $\nu := \log(p_{N_Y})$ است، مدلی با نویز جمعی پیوسته در جهت دیگر وجود نداشته و جهت علّی، با داشتن توزیع مشترک متغیرها قابل شناسایی است. در شرط فوق، آرگومان توابع و مشتق توابع ν، ξ و f به ترتیب $y - f(x)$، x و x هستند.

اثبات. فرض کنید که گراف قابل شناسایی نباشد و دو گراف با جهت‌های متضاد بر این داده‌ها قابل برازش باشد. در این صورت خواهیم داشت:

$$p(x, y) = p_n(y - f(x))p_x(x) = p_{\tilde{n}}(x - g(y))p_y(y) . \tag{3}$$

π را برابر تابع log-liklihood در نظر بگیرید:

$$\pi(x, y) := \log p(x, y) = \nu(y - f(x)) + \xi(x) , \tag{4}$$

از طرف دیگر، بنا بر معادله‌ی (۳) داریم:

$$\pi(x, y) = \tilde{\nu}(x - g(y)) + \eta(y) \tag{5}$$

با مشتق گیری از این معادله خواهیم داشت:

$$\frac{\partial^2 \pi}{\partial x \partial y} = -\tilde{\nu}''(x - g(y))g'(y) \quad \text{و} \quad \frac{\partial^2 \pi}{\partial x^2} = \tilde{\nu}''(x - g(y)) .$$

در نتیجه داریم:

$$\frac{\partial}{\partial x} \left(\frac{\partial^2 \pi / \partial x^2}{\partial^2 \pi / (\partial x \partial y)} \right) = -\frac{\partial}{\partial x} g'(y) = 0 . \tag{6}$$

از معادله‌ی (۴) به دست می‌آید:

$$\frac{\partial^2 \pi}{\partial x \partial y} = -\nu''(y - f(x))f'(x) , \tag{7}$$

و همچنین

$$\frac{\partial^2 \pi}{\partial x^2} = \frac{\partial}{\partial x} \left(-\nu'(y - f(x))f'(x) + \xi'(x) \right) = \nu''(f')^2 - \nu' f'' + \xi'' , \tag{8}$$

که در آن آرگومان‌ها برای خوانایی بیشتر حذف شده‌اند. از معادله‌ی (۷) و (۸)

$$\frac{\partial}{\partial x} \left(\frac{\frac{\partial^2 \pi}{\partial x^2}}{\frac{\partial^2 \pi}{\partial x \partial y}} \right) = -2f'' + \frac{\nu' f'''}{\nu'' f'} - \xi''' \frac{1}{\nu'' f'} + \frac{\nu' \nu''' f''}{(\nu'')^2} - \frac{\nu' (f'')^2}{\nu'' (f')^2} - \xi'' \frac{\nu'''}{(\nu'')^2} + \xi'' \frac{f''}{\nu'' (f')^2} .$$

بر اساس معادله‌ی (۶)، عبارت فوق برابر صفر است یعنی:

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'}$$

با بازنویسی این معادله بر حسب توابع اصلی مساله داریم:

$$p_x''' = p_x'' p_n'' f' \left(-\frac{p_n'''}{(p_n'')^۲} + \frac{f''}{p_n''(f')^۲} \right) + p_n'' f' \left(-۲f'' + \frac{p_n' f'''}{p_n'' f'} + \frac{p_n' p_n''' f''}{(p_n'')^۲} - \frac{p_n'(f'')^۲}{p_n''(f')^۲} \right).$$

□

سوال مهمی که در اینجا مطرح می‌شود این است که در چه شرایطی، قادر به تشخیص جهت از روی چگال مشترک نیستیم. قضیه‌ی زیر به ما می‌گوید که در صورتی که به صورت کاملاً تصادفی از مدل‌های نویز جمعی یک مدل را انتخاب کنیم، احتمال قابل شناسایی نبودن جهت علّی صفر است.

قضیه ۳.۳. اگر برای مجموعه‌ی توابع f و توزیع نویزهای خارجی داده‌شده، برای یک y خاص، $v''(y - f(x))f'(x) = ۰$ شمارا جواب داشته باشد، یا جوابی نداشته باشد، توزیع $\mathbf{X}$ در یک فضای سه بعدی زندگی می‌کند.

از آنجا که مجموعه‌ی توزیع‌های پیوسته بی‌نهایت بعدی است، برای «اکثر» مدل‌های نویز جمعی، جهت علّی قابل شناسایی است.

اثبات. y فیکس‌ای در نظر بگیرید به طوری که $\nu''(y - f(x))f'(x) \neq ۰$ در همه‌ی xها به جز تعدادی شمارا برقرار باشد. برای هر f, ν داده‌شده، بنا بر قضیه‌ی ۲.۳ یک معادله‌ی دیفرانسیل برای ξ به دست می‌آوریم:

$$\xi'''(x) = \xi''(x)G(x, y) + H(x, y)\,, \tag{۹}$$

که در آن H و G برابرند با

$$G := -\frac{\nu''' f'}{\nu''} + \frac{f''}{f'}$$

و

$$H := -۲\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu'(f'')^۲}{f'}\,,$$

با حل این معادله‌ی دیفرانسیل برای ξ'' داریم

$$\xi''(x) = \xi''(x_۰)e^{\int_{x_۰}^{x} G(\tilde{x}, y)d\tilde{x}} + \int_{x_۰}^{x} e^{\int_{\hat{x}}^{x} G(\tilde{x}, y)d\tilde{x}} H(\widehat{x}, y)d\widehat{x}\,. \tag{۱۰}$$

مجموعه‌ی توابع ξای که در معادله‌ی دیفرانسیل مذکور صدق می‌کنند، در یک زیرفضای سه بعدی آفین زندگی می‌کنند که با سه عدد $\xi(x_۰), \xi'(x_۰), \xi''(x_۰)$ تابع به طور یکتا تعیین می‌شود. در نتیجه اثبات کردیم که برای یک تابع و مجموعه‌ی توزیع نویزهای خارجی داده‌شده، مجموعه‌ی تویزیع احتمال Xهایی که به اجازه‌ی وجود یک مدل برعکس می‌دهند، در یک زیرفضای سه بعدی از فضای بی‌نهایت بعدی توزیع‌های پیوسته هستند. □

با وجود اینکه احتمال اینکه برای یک مدل نویز جمعی، مدل نویز جمعی دیگری در جهت مخالف وجود داشته باشد بسیار نادر است، این سوال مطرح است که در چه شرایطی برای یک مدل نویز جمعی چنین اتفاقی رخ می‌دهد. Zhang et al. در [۶] پنج دسته مدل نویز جمعی معرفی می‌کند و اثبات می‌کند هر مدل نویز جمعی غیر قابل شناسایی از تابع چگالی احتمال، به ناچار در یکی از این دسته‌ها قرار می‌گیرد.

قضیه ۴.۳. فرض کنید $X_۲ = f_۲(X_۱) + N_۲$ باشد و $N_۲$ نیز نویزی $full\ support$ و مستقل از $X_۱$ باشد، تابع $f_۲$ سه بار مشتق‌پذیر بوده و همچنین معادله‌ی $\frac{d}{dx_۱} f_۲(x_۱) \frac{d^۲}{dx_۱^۲} \log(p_{N_۲}(x_۲)) = ۰$ تنها در تعدادی متناهی نقطه‌ی $(x_۱, x_۲)$ برقرار باشد.
در صورتی که یک مدل در جهت بر عکس وجود داشته باشد، به این معنا که $X_۱ = g_۱(X_۲) + \tilde{N}_۱$ که $X_۲$ و $\tilde{N}_۱$ در آن مستقل باشد، یکی از پنج حالت زیر برقرار است.

تعاریف دقیق عبارات به کار رفته در جدول فوق را در تعریف زیر آورده‌ایم:

تعریف ۱.۳. فرض کنید p چگالی احتمال یک توزیع پیوسته P باشد.

Table 1: All situations in which the PNL causal model is not identifiable.

	p_{e_2}	p_{t_1} $(t_1 = g_2^{-1}(x_1))$	$h = f_1 \circ g_2$	Remark
I	Gaussian	Gaussian	linear	h_1 also linear
II	log-mix-lin-exp	log-mix-lin-exp	linear	h_1 strictly monotonic, and $h_1' \to 0$, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
III	log-mix-lin-exp	one-sided asymptotically exponential (but not log-mix-lin-exp)	h strictly monotonic, and $h' \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$	—
IV	log-mix-lin-exp	generalized mixture of two exponentials	Same as above	—
V	generalized mixture of two exponentials	two-sided asymptotically exponential	Same as above	—

- P یک *log-mix-lin-exp* است اگر وجود داشته باشند $c_۱, c_۲, c_۳, c_۴$ به نحوی که $c_۱ < ۰$ و $c_۲c_۳ > ۰$ به صورتی که:

$$\log p(x) = c_۱ \exp(c_۲ x) + c_۳ x + c_۴.$$

- P *one-sided asymptotically exponential* است اگر وجود داشته باشد $c \neq ۰$ به نحوی که

$$\frac{d}{dx} \log p(x) \to c$$

وقتی $x \to -\infty$ یا $x \to \infty$.

- P یک *two-sided asymptotically exponential* است اگر وجود داشته باشند $c_۱ \neq ۰$ و $c_۲ \neq ۰$ به نحوی که

$$\frac{d}{dx} \log p(x) \to c_۱$$

وقتی $x \to -\infty$ و

$$\frac{d}{dx} \log p(x) \to c_۲$$

وقتی $x \to \infty$.

- P یک *generalized mixture of two exponentials* است اگر $d_۱, d_۲, d_۳, d_۴, d_۵, d_۶$ وجود داشته باشند به نحوی که $d_۴ > ۰$، $d_۳ > ۰$، $d_۱ d_۵ > ۰$ و $d_۲ < -\frac{d_۱}{d_۵}$ و داشته باشیم:

$$\log p(x) = d_۱ x + d_۲ \log(d_۳ + d_۴ \exp(d_۵ x)) + d_۶.$$

۲.۳ حالت چند متغیره

تا اینجا برای حالت دو بعدی، نشان دادیم در حالت generic یک توزیع احتمال، اجازه‌ی وجود مدل نویز جمعی در هر دو طرف را نمی‌دهد. در این بخش به تعمیم این قضیه از دوبعدی به حالت چندبعدی می‌پردازیم. مرجع اصلی ما در این بخش مقاله‌ی [۵] است. این مقاله قضیه‌ای بسیار جالب را مطرح می‌کند که عنوان می‌کند هنگامی یک قضیه‌ی Identifiability دو بعدی داریم در چه صورتی می‌توان آن را به حالت چندبعدی تعمیم داد. برای ورود به این بحث مقاله‌ی [۵] مثالی جالب را مطرح کرده که در این گزارش نیز به همان شیوه‌ی مقاله عمل می‌کنیم.

مثال ۱.۳. SCM زیر را در نظر بگیرید:

$$\begin{cases} X_۱ = N_۱ \\ X_۲ = f_۲(X_۱) + N_۲ \\ X_۳ = f_۳(X_۱) + aX_۲ + N_۳ \end{cases} \tag{۱۱}$$

که در آن $N_۱ \sim \mathrm{t-student}(\nu = ۳)$، $N_۲ \sim \mathrm{Normal}(۰, \sigma_۲^۲)$ و $N_۳ \sim \mathrm{Normal}(۰, \sigma_۳^۲)$. در اینجا $X_۲$ و $X_۳$ غیر گاوسی هستند اما

$$X_۳ | X_۲ = x_۲ = c + aX_۲ | X_۱ = x_۱ + N_۳$$

برای هر $x_۱$ یک معادله‌ی خطی-گاوسی است در حالی که هیچ‌یک از معادلات اصلی SCM گاوسی-خطی نیستند.
می‌توانیم SCM دیگری بسازیم که در توزیع مشاهداتی تفاوتی با SCM اصلی نداشته باشد:

$$\begin{cases} X_۱ = M_۱ \\ X_۲ = g_۲(X_۱) + bX_۳ + M_۲ \\ X_۳ = g_۳(X_۱) + aX_۲ + M_۳ \end{cases} \tag{۱۲}$$

به نظر می‌رسد باید شرطی بر روی توزیع‌های شرطی قرار دهیم!

برای بیان شرط قابل شناسایی بودن از توزیع احتمال مشاهداتی، به یک تعریف نیاز داریم:

تعریف ۲.۳. یک مدل نویز جمعی با n متغیر را در نظر بگیرید. این مدل را یک مدل نویز جمعی محدودشده می‌نامیم اگر برای هر $j \in \mathbf{V}$ و $i \in \mathbf{PA}_j$ و تمام مجموعه‌های $\mathbf{S} \subseteq \mathbf{V}$ به طوری که $\mathbf{PA}_j \setminus \{i\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_j \setminus \{i, j\}$
وجود داشته باشد $x_\mathbf{S}$ که $p_\mathbf{S}(x_\mathbf{S}) > ۰$ به نحوی که

$$\Big(f_j(x_{\mathbf{PA}_j \setminus \{i\}}, \underbrace{\cdot}_{X_i}), P(X_i | X_\mathbf{S} = x_\mathbf{S}), P(N_j) \Big)$$

در شرایط قضیه‌ی (۲.۳) صدق کند.

قضیه ۵.۳. فرض کنید که $\mathbf{X} = \{X_۱, X_۲, \ldots, X_n\}$ توسط یک مدل نویز جمعی محدودشده با گراف $G_۰$ تولید شده‌باشند و فرض کنید $P(\mathbf{X})$ نسبت به $G_۰$ شرط $Causal\ Minimality$ را ارضا کند. در این صورت $G_۰$ از روی توزیع احتمال $P(\mathbf{X})$ قابل شناسایی است.

برای اثبات این قضیه نیاز به یک لم گرافی داریم که Chickering در سال ۱۹۹۵ اثبات کرده است [۱].

لم ۱.۳. فرض کنید G و G' دو DAG روی مجموعه‌ی متغیرهای $\mathbf{X}$ باشند. فرض کنید $P(\mathbf{X}$ چگالی احتمالی همواره مثبت دارد که نسبت به G و G' مارکوف هستند و شرط $Causal\ Minimality$ را ارضا می‌کنند. در این صورت متغیرهای L و Y وجود دارند که برای مجموعه‌های $\mathbf{Q} := \mathbf{PA}_L^G \setminus \{Y\}$، $\mathbf{R} := \mathbf{PA}_Y^{G'} \setminus \{L\}$ و $\mathbf{S} := \mathbf{Q} \cup \mathbf{R}$ داشته باشیم:

- $Y \to L$ در G و $L \to Y$ در G'.
- $\mathbf{S} \subseteq \mathbf{ND}_L^G \setminus \{Y\}$ و $\mathbf{S} \subseteq \mathbf{ND}_Y^{G'} \setminus \{L\}$

اثبات. **قضیه‌ی** (۵.۳)
برهان خلف: فرض کنید دو مدل نویز جمعی محدود شده با این توزیع احتمال وجود داشته باشند. یکی با گراف G و دیگری با گراف متفاوت G'. دو متغیر L و Y را بر اساس لم فوق انتخاب می‌کنیم و مشابه لم فوق، می‌گیریم $\mathbf{Q} := \mathbf{PA}_L^G \setminus \{Y\}$، $\mathbf{R} := \mathbf{PA}_Y^{G'} \setminus \{L\}$ و $\mathbf{S} := \mathbf{Q} \cup \mathbf{R}$ هر تحقق دلخواه $\mathbf{s} = (\mathbf{q}, \mathbf{r})$ را در نظر بگیرید و بنویسید $L^* := L|\mathbf{S} = \mathbf{s}$ و $Y^* := Y|\mathbf{S} = \mathbf{s}$ از لم فوق داریم: $\mathbf{S} \subseteq \mathbf{ND}_L^G \setminus \{Y\}$ و این بدین معناست که $\mathbf{S} \cup \{Y\} \subseteq \mathbf{ND}_L^G$ در نتیجه $N_L \perp\!\!\!\perp \{Y\} \cup \mathbf{S}$ در نتیجه می‌توان به معادله‌ی زیر رسید:

$$L^* = f_L(\mathbf{q}, Y^*) + N_L \quad N_L \perp\!\!\!\perp Y^*$$

برای G' هم با استدلال مشابه می‌توان به معادله‌ی زیر رسید:

$$Y^* = g_Y(\mathbf{r}, L^*) + N_Y \quad N_Y \perp\!\!\!\perp L^*$$

و این در تناقض با فرض مدل نویز جمعی محدود شده است پس فرض خلف باطل بوده و G و G' برابرند. □

نکته‌ی بسیار جالبی که در این قضیه وجود دارد این است که می‌توان آن را برای هر قضیه‌ی Indentifiability دیگر نیز به کار برد، به شرط عدم وجود دور. به طور مثال به کمک آن می‌توان LINGAM و یا Post Non-Linear Additive Noise Model، که در ادامه به آن پرداخته خواهد شد، را از حالت دو بعدی به حالت چند بعدی تعمیم داد.

۴ الگوریتم‌های یادگیری

در این بخش به بررسی الگوریتم‌های یادگیری گراف مربوط به یک SCM از داده‌ی محدود به فرض اینکه در SCM مدل نویز جمعی برقرار باشد، پرداخته و الگوریتم‌های مختلف مطرح شده را با هم مقایسه می‌کنیم. مراجع ما در این بخش مقالات [۵] و [۳] هستند.

۱.۴ الگوریتم RESIT

این الگوریتم توسط Peters et al.، سال ۲۰۱۴، در [۵] معرفی شده است. ایده‌ی اصلی این الگوریتم این است که برای هر X_i اگر X_i یک Sink Node باشد داریم $N_i \perp\!\!\!\perp \mathbf{X} \setminus \{X_i\}$ به طور کلی برای هر $Y \in \mathbf{X}$، $N_Y \perp\!\!\!\perp \mathbf{ND}_Y$.

Algorithm 1 Regression with subsequent independence test (RESIT)

1: **Input:** i.i.d. samples of a p-dimensional distribution on $(X_1, \ldots, X_p)$
2: $S := \{1, \ldots, p\}, \pi := [\ ]$
3: PHASE 1: Determine causal order.
4: **repeat**
5:     **for** $k \in S$ **do**
6:         Regress X_k on $\{X_i\}_{i \in S \setminus \{k\}}$.
7:         Measure dependence between residuals and $\{X_i\}_{i \in S \setminus \{k\}}$.
8:     **end for**
9:     Let k^* be the k with the weakest dependence.
10:     $S := S \setminus \{k^*\}$
11:     $\mathrm{pa}(k^*) := S$
12:     $\pi := [k^*, \pi]$ (π will be the causal order, its last component being a sink)
13: **until** $\#S = 1$
14: PHASE 2: Remove superfluous edges.
15: **for** $k \in \{2, \ldots, p\}$ **do**
16:     **for** $\ell \in \mathrm{pa}(\pi(k))$ **do**
17:         Regress $X_{\pi(k)}$ on $\{X_i\}_{i \in \mathrm{pa}(\pi(k)) \setminus \{\ell\}}$.
18:         **if** residuals are independent of $\{X_i\}_{i \in \{\pi(1), \ldots, \pi(k-1)\}}$ **then**
19:             $\mathrm{pa}(\pi(k)) := \mathrm{pa}(\pi(k)) \setminus \{\ell\}$
20:         **end if**
21:     **end for**
22: **end for**
23: **Output:** $(\mathrm{pa}(1), \ldots, \mathrm{pa}(p))$

الگوریتم RESIT در هر مرحله یک Sink Node را تشخیص داده و حذف می‌کند. برای تشخیص یک Sink نیز از ویژگی $N_i \perp\!\!\!\perp \mathbf{X} \setminus \{X_i\}$ استفاده می‌کند.

الگوریتم RESIT دو فاز دارد. در فاز اول (خط ۳ تا ۱۳)، یک Causal Order پیدا می‌شود. با رگرس کردن هر متغیر روی بقیه‌ی متغیرهای گراف هر مرحله، متغیری که باقی مانده‌ی رگرسیون مربوط به از دیگر متغیرها مستقل‌تر(مثلاً با معیار p-value آزمون HSIC) باشد را به عنوان یک sink در نظر می‌گیریم. با حذف این راس، مجدداً یک DAG دیگر به وجود می‌آید که در آن همین روند را روی آن تکرار می‌کنیم. با این کار می‌توان به یک Causal Order برای متغیرها رسید. در فاز دوم، برای شروع فرض می‌شود که اگر $\pi(i) < \pi(j)$، از i به j یک یال وجود دارد . از این گراف شروع کرده. هر بار یک متغیر، $X_{\pi(k)}$، را در نظر گرفته و آن را بر روی parent هایش به جز یک parent، X_l، رگرس می‌کنیم به نحوی که هر parent یک بار از رگرسیون کنار گذاشته شود. در هر رگرسیون، اگر باقی‌مانده رگرسیون از متغیرهایی که در Causal Order بالاتر از $X_{\pi(k)}$ هستند مستقل شد، ارتباط X_l و $X_{\pi(k)}$ را حذف می‌کنیم.

الگوریتم RESIT در مرحله‌ی اول خود $O(n^۲)$ تست آماری انجام می‌دهد و در مرحله‌ی دوم نیز تعداد تست های آماری $O(n)$ است. چند جمله‌ای بودن این الگوریتم بسیار عجیب است زیرا مسائل معمول در Bayesian Network Learning اکثراً -NP Hard هستند. با این وجود الگوریتم RESIT برای n های بزرگ قابل استفاده نیست زیرا در صورتی که در انجام تست آماری دچار خطا شویم، خطا به شدت در مراحل بعد منتشر شده و باعث می‌شود به طور قابل ملاحظه‌ای از گراف اصلی دور شویم.

۲.۴ روش Brute Force و GDS

یک دسته‌ی دیگر از الگوریتم‌های یادگیری در مدل‌های نویز جمعی، الگوریتم‌های مبتنی بر score هستند.

برای یادگیری ساختار مدل‌های نویز جمعی محدود شده، می‌توان تمام DAG ها را enumerate کرد و بررسی کرد که آیا استقلال‌هایی که از داده‌ها استخراج می‌شوند در گراف نیز وجود دارند یا خیر. یک ایراد این روش این است که این روش لزوما یک گراف Causal Minimal به ما نمی‌دهد. برای حل این مشکل یک penalized independence score تعریف کرده و آن را برای گراف‌ها محاسبه می‌کنیم و این معیار را مبنای انتخاب گراف قرار می‌دهیم.

$$\widehat{G} = \text{argmin}_G \sum_{i=۱}^{n} \text{DM}(res_i^{G,\text{RM}}, res_{-i}^{G,\text{RM}}) + \lambda \#\text{edges}$$

در آن RM روش رگرسیون ما و DM یک معیار استقلال است. res_i مقدار باقی مانده‌ی رگرسیون X_i است وقتی آن را بر روی تمام parent هایش رگرس می‌کنیم و res_{-i} باقی‌مانده‌ی رگرسیون است. واضح است که این الگوریتم برای گراف‌های بزرگ به هیچ وجه قابل استفاده نخواهد بود.

یک راه معقول برای کاهش پیچیدگی محاسباتی الگوریتم فوق، استفاده از روش‌های حریصانه است [۵]. دو گراف را مجاور می‌گوییم اگر تنها با یک تغییر جهت، افزودن یال یا کاستن یال به یکدیگر تبدیل شوند. در الگوریتم حریصانه، در هر مرحله score گراف آن مرحله را حساب کرده و با score گراف‌های مجاور مقایسه می‌شود. هرجا امتیاز یکی از این گراف‌های مجاور از گراف اصلی بالاتر بود، گراف اصلی را برابر گراف مجاور می‌گذاریم و الگوریتم را ادامه می‌دهیم. برای اینکه استپ‌ها کمی بهتر شوند، به صورت تصادفی امتیاز همسایه‌ها را محاسبه نمی‌کنیم بلکه از همسایه‌ای شروع می‌کنیم که باقی‌مانده آن نسبت به باقی مانده‌ی دیگر راس‌ها کمتر مستقل است. تضمینی وجود ندارد که این روش به بهترین گراف برسد [۵].

۳.۴ مقایسه‌ی الگوریتم‌های RESIT، Brute Force و GDS با سایر الگوریتم‌ها

مقاله‌ی [۵] به صورت مفصل الگوریتم‌های RESIT، Brute Force، GDS، PC GES و LiNGAM را مقایسه می‌کند. در ادامه به بررسی این مقایسه و نقد آن می‌پردازیم. **ایراد بزرگ این مقایسه این است که بر روی داده‌هایی انجام شده که به صورت مصنوعی تولید شده‌اند.** معیار مورد استفاده‌ی ما برای مقایسه‌ی الگوریتم‌ها **Structural Hamming Distance (SHD)** است. همان‌طور که از اسم این معیار بر می‌آید، این معیار تعداد یال‌های اشتباه را می‌شمارد. وجود هر یال اضافه یا جهت‌دار به دست‌آوردن یال بی‌جهت نیز یک خطا در نظر گرفته می‌شود. در جدول‌های مقایسه‌ای که در بخش‌های بعد خواهند آمد، ردیف DAG مقایسه‌ی DAG واقعی با گراف به دست آمده و ردیف CDAG مقایسه‌ی CDAG و گراف حاصل است.

۱.۳.۴ حالت SCM خطی

در حالت خطی، ضرایب β_{jk} به صورت تصادفی با توزیع یکنواخت $[۰/۱, ۲] \cup [-۲, -۰/۱]$ انتخاب می‌شوند. نویز های برونی مستقل هستند و با توزیع $K_j \cdot \text{sign}(M_j) \cdot |M_j|^{\alpha_j}$ تولید می‌شوند که در آن $M_j \sim N(۰, ۱)$، $K_j \sim U([۰/۱, ۰/۵])$ و $\alpha_j \sim U([۲, ۴])$. جدول (۱) مقایسه‌ی SHD این الگوریتم‌ها برای گراف‌ها با تعداد راس مختلف است. در این جدول، روش

	GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
	$p = 4, n = 100$							
DAG	0.7 ± 0.9	0.6 ± 0.8	1.2 ± 1.3	1.9 ± 1.2	3.5 ± 1.5	3.6 ± 1.4	3.1 ± 1.7	4.4 ± 1.0
CPDAG	1.1 ± 1.5	0.9 ± 1.4	1.5 ± 1.7	2.4 ± 1.5	2.4 ± 1.7	2.3 ± 1.6	2.0 ± 2.0	4.3 ± 1.4
	$p = 4, n = 500$							
DAG	0.2 ± 0.6	0.1 ± 0.3	0.6 ± 0.8	0.5 ± 0.8	3.1 ± 1.4	3.2 ± 1.4	2.9 ± 1.6	4.1 ± 1.2
CPDAG	0.3 ± 0.9	0.2 ± 0.5	0.9 ± 1.3	0.8 ± 1.2	1.9 ± 1.8	1.6 ± 1.7	1.6 ± 1.9	3.9 ± 1.4
	$p = 15, n = 100$							
DAG	12.2 ± 5.3	–	25.2 ± 8.3	11.1 ± 3.7	13.0 ± 3.6	13.7 ± 3.7	12.7 ± 4.2	57.4 ± 26.4
CPDAG	13.2 ± 5.4	–	27.0 ± 8.5	12.4 ± 3.9	10.7 ± 3.5	10.8 ± 3.8	12.4 ± 4.9	58.5 ± 27.1
	$p = 15, n = 500$							
DAG	6.1 ± 6.4	–	51.2 ± 17.8	3.4 ± 2.8	10.2 ± 3.8	10.8 ± 4.2	8.7 ± 4.6	57.6 ± 24.2
CPDAG	6.8 ± 6.9	–	54.5 ± 18.5	4.5 ± 3.8	8.2 ± 4.6	7.5 ± 4.4	7.1 ± 5.6	58.9 ± 25.0

شکل ۱: SHD الگوریتم‌های مختلف به ازای گراف‌هایی با سایزهای مختلف - مدل خطی

RAND به این نحو است که به صورت کاملاً تصادفی یک گراف انتخاب کرده و به داده‌ها نسبت می‌دهیم.

○ همان‌طور که انتظار می‌رفت، در گراف‌های کوچک، بررسی تمام گراف‌ها بهترین نتیجه را داشته ولی برای گراف‌های بزرگ، امکان اجرای این الگوریتم نبوده است.

○ در گراف‌های بزرگ‌تر روش LiNGAM بهترین نتیجه را دارد. نکته‌ی جالب این است که روش DAG نیز حاصلی شبیه به روش LiNGAM دارد که نشان می‌دهد روش GAS در کمینه‌های موضعی گیر نیفتاده است.

○ مطابق انتظار، الگوریتم REST در حالت خطی بسیار ناموفق عمل کرده است. این عدم موفقیت، به خصوص در گراف‌هایی با تعداد رأس زیاد بسیار شدید است. دلیل این امر این است که این الگوریتم، در این حالت در فاز اول خود یال‌های اضافی زیادی نگه می‌دارد که در فاز دوم قابل حذف نیستند.

۴.۴ حالت SCM غیرخطی

در این حالت، تابع به صورت تصادفی از یک فرآیند گاوسی با پهنای باند ۱ انتخاب شده و نویزها گاوسی هستند و واریانس آن تصادفی انتخاب شده است. در حالت غیرخطی نیز مشابه حالت خطی، در گراف‌های کوچک روش Brute Force بهترین نتیجه را می‌دهد. روش GDS نیز مشابه روش Brute Force عمل می‌کند. با بزرگ‌تر شدن سایز گراف، دیگر روش Brute Force قابل استفاده نیست و روش GDS هم کارایی خود را از دست داده و گرفتار کمینه‌های موضعی می‌شود. به نظر می‌رسد که در گراف‌های بزرگ‌تر الگوریتم RESIT موفق‌تر از سایر روش‌ها عمل می‌کند. بهتر بود که این مقاله برای گراف‌های بزرگ‌تر نیز این الگوریتم‌ها را تست می‌کرد تا بتوانیم با اطمینان بیشتری گزاره‌ی بهتر بودن روش RESIT از سایر روش‌ها در گراف‌های بزرگ‌تر را مطرح کنیم.

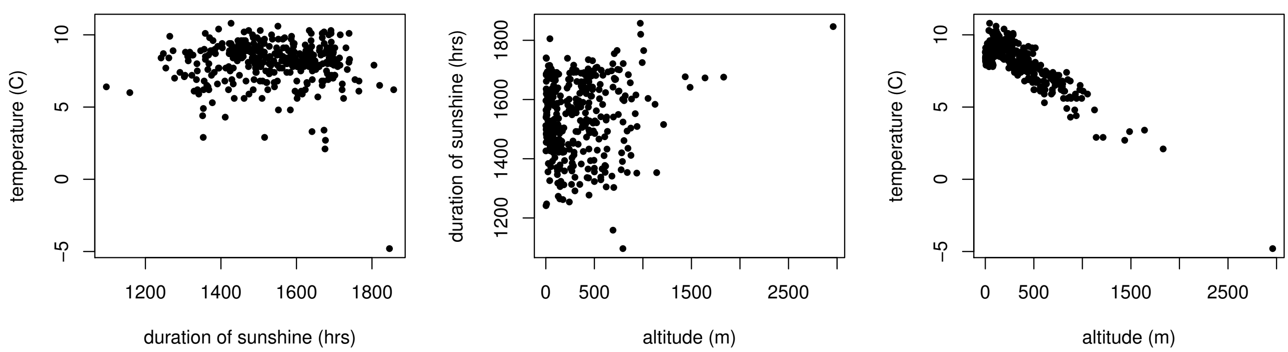
	GDS	BF	RESIT	LiNGAM	PC	CPC	GES	RAND
				$p = 4, n = 100$				
DAG	1.5 ± 1.4	1.0 ± 1.0	1.7 ± 1.3	3.5 ± 1.2	3.5 ± 1.5	3.8 ± 1.4	3.5 ± 1.3	4.0 ± 1.3
CPDAG	1.7 ± 1.7	1.2 ± 1.4	2.0 ± 1.6	3.0 ± 1.4	2.9 ± 1.5	2.7 ± 1.4	3.4 ± 1.7	3.9 ± 1.4
				$p = 4, n = 500$				
DAG	0.5 ± 0.9	0.3 ± 0.5	0.8 ± 0.9	3.7 ± 1.2	3.5 ± 1.5	3.8 ± 1.5	3.3 ± 1.5	4.1 ± 1.2
CPDAG	0.6 ± 1.1	0.6 ± 1.0	1.0 ± 1.3	3.0 ± 1.7	3.1 ± 1.9	2.8 ± 1.8	3.4 ± 1.9	3.8 ± 1.6
				$p = 15, n = 100$				
DAG	14.3 ± 4.9	–	15.4 ± 5.7	15.4 ± 3.6	14.2 ± 3.5	15.5 ± 3.6	24.8 ± 6.3	56.8 ± 24.1
CPDAG	15.1 ± 5.4	–	16.5 ± 5.9	15.3 ± 4.0	13.3 ± 3.6	13.3 ± 4.0	26.4 ± 6.5	58.0 ± 24.7
				$p = 15, n = 500$				
DAG	13.0 ± 8.4	–	10.1 ± 5.7	21.4 ± 6.9	13.9 ± 4.5	15.1 ± 4.8	26.8 ± 8.5	56.1 ± 26.8
CPDAG	14.2 ± 9.2	–	11.3 ± 6.3	21.1 ± 7.3	13.7 ± 4.9	13.4 ± 5.1	28.6 ± 8.8	57.0 ± 27.3

شکل ۲: SHD الگوریتم‌های مختلف به ازای گراف‌هایی با سایزهای مختلف - مدل غیرخطی

۵.۴ بررسی داده‌های واقعی؟!

در ادامه این مقاله تلاشی مذبوحانه(!) برای اثبات کارآمدی الگوریتم RESIT انجام می‌دهد.

داده‌ی مورد استفاده، سه متغیر مشاهده شده دارد: دمای شهر، ارتفاع شهر و مدت زمان تابش خورشید. این داده از ۳۴۹ مرکز هواشناسی آلمان به دست آمده است.



شکل ۳: نمودار پراکنش داده‌های مسئله

Method	Graph
LiNGAM	$T \rightarrow A$
PC	$T \rightarrow A \leftarrow DS$
CPC	$T \rightarrow A \leftarrow DS$
GES	Fully Connected
GDS	$T \leftarrow A \rightarrow DS$
BF	$T \leftarrow A \rightarrow DS$
RESIT	$T \leftarrow A \rightarrow DS$

شکل ۴: گراف پیشنهادی الگوریتم‌های مختلف

دیده می‌شود که الگوریتم‌های LiNGAM، (C)PC، GES گراف‌هایی را پیشنهاد می‌کنند که به وضوح غلطند! زیرا Intervention برو روی دمای شهر و یا زمان تابش خورشید در یک شهر، تاثیری بر روی ارتفاع آن نخواهد داشت. سه روش GES، BF و RESIT جواب‌های مشابهی می‌دهند.

به نظر من این گراف‌ها غلط هستند زیرا علاوه بر دو یالی که RESIT می‌یابد، چنین یالی نیز در گراف موجود باشد: $T \leftarrow DS$. نویسنده‌ی مقاله بیان می‌کند که گرافی که هر سه یال را دارد، در الگوریتم های GES و BF دومین بالاترین score را می‌یابد و ادعا می‌کند که این اشتباه به دلیل وجود confounder هایی از جنس متغیرهای جغرافیایی است.

به نظر من این تحلیل به هیچ عنوان تحلیل جامع و دقیقی برای مقایسه‌ی الگوریتم‌ها نیست. اولاً این مطالعه تنها بر روی یک دیتاست انجام شده است، حتی برای بدترین لگوریتم‌ها نیز می‌توان یک دیتاست یافت که خروجی آن الگوریتم جواب مناسبی باشد. دوماً جواب این الگوریتم خیلی جواب خوبی نیز نیست! سوماً اینکه **تعداد متغیرهای این دیتاست بسیار محدود است و برای آزمودن الگوریتمی که ادعای چندمتغیره بودن دارد و با پشتوانه‌ی داده‌های مصنوعی، در مورد آن این ادعا شده است که در گراف‌هایی با تعداد متغیر بالا از الگوریتم‌های گذشته بهتر عمل می‌کند، کافی نیست.**

مراجع

[1] Chickering, D. M. A transformational characterization of equivalent bayesian network structures. in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (1995), UAI'95, pp. 87–98.

[2] Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. in *Advances in Neural Information Processing Systems 21*. 2009, pp. 689–696.

[3] Nowzohour, C., and Bühlmann, P. Score-based causal learning in additive noise models. *Statistics 50*, 3 (2016), 471–485.

[4] Peters, J., Janzing, D., and Schölkopf, B. Causal inference on time series using restricted structural equation models. in *Advances in Neural Information Processing Systems 26*. 2013, pp. 154–162.

[5] Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res. 15*, 1 (Jan. 2014), 2009–2053.

[6] Zhang, K., and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Arlington, Virginia, United States, 2009), UAI '09, pp. 647–655.