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

بهراد منیری
دانشکده‌ی مهندسی برق دانشگاه صنعتی شریف

bemoniri@ee.sharif.edu

فهرست مطالب

۱ 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]، [۴] و [۵] 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.

تعریف ۱.۱. *یک مدل نویز جمعی پیوسته را به صورت زیر تعریف می‌شود:*

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

که در آن $\boldsymbol{PA}_j$ مجموعه‌ی والدین X_j هستند و متغیرهای نویز، N_j دارای چگالی‌احتمالی اکیداً مثبت می‌باشند و مستقل‌اند. همچنین در این بخش فرض می‌شود گراف مولد داده‌ها یک DAG است.

قضیه ۱.۱. *درمدل‌های نویز جمعی، شرط Causal Minimality معادل این است که توابع f_j نسبت به هیچ یک از متغیرهایشان ثابت نباشند.*

۱.۱ 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 ۲.۱. *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:

$$\xi''' = \xi'' \left(-\frac{\nu''' f'}{\nu''} + \frac{f''}{f'} \right) - 2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu' (f'')^2}{f'} , \quad (۲)$$

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:

$$p(x, y) = p_n(y - f(x))p_x(x) = p_{\tilde{n}}(x - g(y))p_y(y) . \quad (۳)$$

π را برابر تابع log-liklihood در نظر بگیرید:

$$\pi(x, y) := \log p(x, y) = \nu(y - f(x)) + \xi(x) , \quad (۴)$$

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

$$\pi(x, y) = \tilde{\nu}(x - g(y)) + \eta(y) \quad (۵)$$

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:

$$\frac{\partial}{\partial x} \left(\frac{\partial^2 \pi / \partial x^2}{\partial^2 \pi / (\partial x \partial y)} \right) = -\frac{\partial}{\partial x} g'(y) = 0 . \quad (۶)$$

از معادله‌ی (۴) به دست می‌آید:

$$\frac{\partial^2 \pi}{\partial x \partial y} = -\nu''(y - f(x))f'(x) , \quad (۷)$$

و همچنین

$$\frac{\partial^2 \pi}{\partial x^2} = \frac{\partial}{\partial x} \left(-\nu'(y - f(x))f'(x) + \xi'(x) \right) = \nu''(f')^2 - \nu' f'' + \xi'' , \quad (۸)$$

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:

$$p_x''' = p_x'' p_n'' f' \left(-\frac{p_n'''}{(p_n'')^2} + \frac{f''}{p_n''(f')^2} \right) + p_n'' f' \left(-2f'' + \frac{p_n' f'''}{p_n'' f'} + \frac{p_n' p_n''' f''}{(p_n'')^2} - \frac{p_n'(f'')^2}{p_n''(f')^2} \right).$$

□

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 ۳.۱. *اگر برای مجموعه‌ی توابع f و توزیع نویزهای خارجی داده‌شده، برای یک y خاص، $v''(y-f(x))f'(x) = 0$ شمارا جواب داشته باشد، یا جوابی نداشته باشد، توزیع $\boldsymbol{X}$ 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 $\nu''(y-f(x))f'(x) \neq 0$ در همه‌ی xها به جز تعدادی شمارا برقرار باشد. برای هر f, ν داده‌شده، بنا بر قضیه‌ی ۲.۱ 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

$$H := -2\nu'' f'' f' + \nu' f''' + \frac{\nu' \nu''' f'' f'}{\nu''} - \frac{\nu'(f'')^2}{f'}\,,$$

با حل این معادله‌ی دیفرانسیل برای ξ'' داریم

$$\xi''(x) = \xi''(x_0) e^{\int_{x_0}^{x} G(\tilde{x},y)d\tilde{x}} + \int_{x_0}^{x} e^{\int_{\hat{x}}^{x} G(\tilde{x},y)d\tilde{x}} H(\hat{x},y) d\hat{x}\,. \tag{۱۰}$$

مجموعه‌ی توابع ξای که در معادله‌ی دیفرانسیل مذکور صدق می‌کنند، در یک زیرفضای سه بعدی آفین زندگی می‌کنند که با سه عدد $\xi(x_0), \xi'(x_0), \xi''(x_0)$ تابع به طور یکتا تعیین می‌شود. در نتیجه اثبات کردیم که برای یک تابع و مجموعه‌ی توزیع نویزهای خارجی داده‌شده، مجموعه‌ی تویزیعع 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 بسیار نادر است، این سوال مطرح است که در چه شرایطی برای یک مدل نویز جمعی چنین اتفاقی رخ می‌دهد. Zhang et al. در [۵] 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 ۴.۱. *فرض کنید $X_2 = f_2(X_1) + N_2$ باشد و N_2 نیز نویزی full support و مستقل از X_1 باشد، تابع f_2 سه بار مشتق‌پذیر بوده و همچنین معادله‌ی $\frac{d}{dx_1} f_2(x_1) \frac{d^2}{dx_1^2} \log(p_{N_2}(x_2)) = 0$ تنها در تعدادی متناهی نقطه‌ی (x_1, x_2) برقرار باشد. در صورتی که یک مدل در جهت بر عکس وجود داشته باشد، به این معنا که $X_1 = g_1(X_2) + \tilde{N}_1$ که X_2 و $\tilde{N}_1$ 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 ۲.۱. *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:*

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

۲.۱ 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 [۴] 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 [۴] مثالی جالب را مطرح کرده که در این گزارش نیز به همان شیوه‌ی مقاله عمل می‌کنیم.

مثال ۱.۱. *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 خطی‌ـ گاوسی است در حالی که هیچ‌یک از معادلات اصلی SCM گاوسی‌ـ خطی نیستند. می‌توانیم SCM دیگری بسازیم که در توزیع مشاهداتی تفاوتی با SCM اصلی نداشته باشد:

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

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 ۳.۱. *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*

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

در شرایط قضیه‌ی (۲.۱) صدق کند.

قضیه ۵.۱. *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 ۱.۱. *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 (۵.۱))**
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 Yرا 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 $\{Y\} \cup \mathbf{S} \subseteq \mathbf{ND}_L^G$ 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.

۲ 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 [۴] و [۳] هستند.

۱.۲ الگوریتم RESIT

این الگوریتم توسط Peters et al.، سال ۲۰۱۴، در [۴] 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 آن 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 $O(n^2)$ 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.

۲.۲ الگوریتم‌های مبتنی بر Independence Score

یک دسته‌ی دیگر از الگوریتم‌های یادگیری در مدل‌های نویز جمعی، الگوریتم‌های مبتنی بر score هستند. یک الگوریتم دیگر برای یادگیری ساختار مدل‌های نویز جمعی محدود شده، این است که تمام DAG ها را enumerate کنیم و تست های استقلال مطرح شده را با آنها چک کنیم ولی مساله 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.

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., 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.

[5] 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.