

Controlling Confounding Bias

Modelos Gráficos Causales

Sebastián Bejos, Ivan Feliciano

11 de junio de 2019

Instituto Nacional de Astrofísica Óptica y Electrónica

Criterio de Puerta Trasera

- Supongamos que se tiene un diagrama causal G , junto con datos no experimentales sobre un subconjunto \mathbf{V} de variables observadas en G .
- Supongamos que deseamos estimar $P(Y | do(\mathbf{x}))$ para un conjunto de variables \mathbf{Y} , con $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, a partir de una estimación de muestra de $P(v)$, dadas las suposiciones codificadas en G .
- El **criterio de puerta trasera** es una prueba sencilla que se puede aplicar directamente sobre la gráfica causal para verificar si un conjunto $Z \subseteq V$ de variables, es suficiente para identificar $P(Y | do(\mathbf{x}))$.

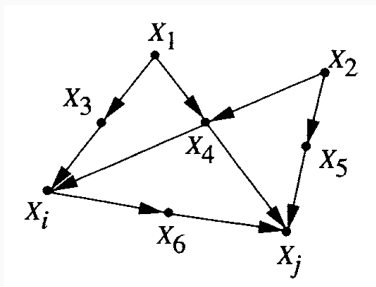
Definición

Un conjunto de variables \mathbf{Z} satisface el **criterio de puerta trasera** relativo a un par ordenado de variables (X_i, X_j) en una DAG G si:

- i *ningún* nodo en Z es descendiente de X_i y
- ii \mathbf{Z} boquea toda trayectoria entre X_i y X_j **que contiene un arco hacia X_i .**

De manera similar, si \mathbf{X} y \mathbf{Y} son dos conjuntos disjuntos de nodos en G , entonces se dice que \mathbf{Z} satisface el criterio de puerta trasera relativo a (\mathbf{X}, \mathbf{Y}) si satisface el criterio para cualquier par (X_i, X_j) tal que $X_i \in \mathbf{X}$ y $X_j \in \mathbf{Y}$.

Criterio de Puerta Trasera



Los conjuntos $\mathbf{Z}_1 = \{X_3, X_4\}$ y $\mathbf{Z}_2 = \{X_4, X_5\}$ cumplen con el criterio de puerta trasera, pero $\mathbf{Z}_3 = \{X_4\}$ no lo hace porque X_4 no bloquea la trayectoria $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$.

Teorema (Ajuste de Puerta Trasera)

Si un conjunto de variables \mathbf{Z} satisface el criterio de puerta trasera relativo a (\mathbf{X}, \mathbf{Y}) , entonces el efecto causal de \mathbf{X} sobre \mathbf{Y} es identificable y está dado por la fórmula

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z})P(\mathbf{z}) \quad (1)$$

Criterio de puerta delantera (Front-Door Criterion) i

- La condición i) de la definición de puerta trasera refleja que “las observaciones concomitantes no deben ser afectadas por el tratamiento”.
- Esta sección demuestra como concomitantes que son afectados por el tratamiento pueden usarse para facilitar la inferencia causal.
- Consideremos la Figura 1. Aunque Z no satisface ninguna de las condiciones de puerta trasera, medidas de Z permiten una estimación consistente de $P(y|do(x))$. Esto se muestra al reducir la expresión para $P(y|do(x))$ a fórmulas que son calculables de la función de distribución observada $P(x, y, z)$.

Criterio de puerta delantera (Front-Door Criterion) ii

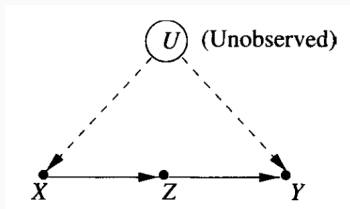


Figura 1

- La distribución conjunta asociada con la Figura 1 se puede descomponer como

$$P(x, y, z, u) = P(u)P(x|u)P(z|x)P(y|z, u).$$

Criterio de puerta delantera (Front-Door Criterion) iii

- De la factorización truncada, la intervención $do(x)$ elimina el factor $P(x|u)$ e induce la distribución post-intervención

$$P(y, z, u|do(x)) = P(y|z, u)P(z|x)P(u).$$

- Sumando sobre z y u tenemos

$$P(y|do(x)) = \sum_z P(z|x) \sum_u P(y|z, u)P(u).$$

- Para eliminar u del lado derecho utilizamos las dos suposiciones de independencia condicional codificadas en el grafo

$$P(u|z, x) = P(u|x), \tag{2}$$

$$P(y|x, z, u) = P(y|z, u). \tag{3}$$

- Esto produce las siguientes igualdades

$$\begin{aligned}\sum_u P(y|z, u)P(u) &= \sum_x \sum_u P(y|z, u)P(u|x)P(x) \\ &= \sum_x \sum_u P(y|x, z, u)P(u|x, z)P(x) \\ &= \sum_x P(y|x, z)P(x)\end{aligned}$$

- y esto produce una forma que sólo involucra cantidades observadas:

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x'). \quad (4)$$

Todos los factores del lado derecho son estimables de datos no experimentales, por lo que $P(y|do(x))$ también es estimable. Entonces, tenemos un estimando no paramétrico sin sesgo para el efecto causal de X sobre Y cuando podamos encontrar una variable de medicación Z que cumpla las condiciones de las ecuaciones (2) y (3).

La ecuación 4 puede interpretarse como una aplicación en dos pasos de la fórmula de la puerta trasera.

1. Encontramos el efecto causal de X sobre Z ; como no hay un camino de puerta trasera no bloqueado de X a Z .

$$P(z|do(x)) = P(z|x).$$

2. Calculamos el efecto causal de Z sobre Y , el cual no sigue siendo igual a $P(y|z)$ porque existe un camino de puerta trasera $Z \leftarrow X \leftarrow U \rightarrow Y$ de Z a Y . Sin embargo, como X bloquea (d-separa) este camino, X puede tomar el rol de un concomitante en el criterio de puerta trasera, lo que permite calcular el efecto causal de Z en Y , $P(y, do(z)) = \sum_{x'} P(y|x', z)P(x')$.
3. Finalmente, combinamos los dos efectos causales a través de

$$P(y|do(x)) = \sum_z P(y|do(z))P(z, do(x)).$$

Definition (Puerta delantera)

Un conjunto de variables Z se dice que satisface el criterio de puerta delantera relativo a un par ordenado de variables (X, Y) si:

1. Z intercepta todos los caminos directos de X a Y ;
2. no existe un camino de puerta trasera desbloqueado de X a Z ; y
3. todos los caminos de puerta trasera de Z a Y son bloqueados por X .

Theorem (Ajuste de puerta delantera)

Si Z satisface el criterio de puerta delantera relativo a (X, Y) y si $P(x, z) > 0$, entonces el efecto causal de X en Y es identificable y es dado por la siguiente fórmula

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z) P(x'). \quad (5)$$

Ejemplo: Fumar y la teoría del genotipo i

- Consideremos el debate de la relación entre **fumar** (X) y el **cáncer de pulmón** (Y). La industria tabacalera ha logrado prevenir la legislación antitabaco argumentando que la correlación observada entre fumar y el cáncer de pulmón podría explicarse por algún tipo de **genotipo cancerígeno** (U) que involucra el deseo innato de nicotina. La **cantidad de alquitrán** (Z) depositada en los pulmones de una persona es una variable que promete cumplir las condiciones enumeradas en la definición previa, ajustándose así a la estructura de la Figura 1.
- Para cumplir con la condición 1., debemos suponer que fumar cigarrillos no tiene ningún efecto sobre la producción de cáncer de pulmón, excepto cuando está mediado por los depósitos de alquitrán.

Ejemplo: Fumar y la teoría del genotipo ii

- Para cumplir con las condiciones 2. y 3., debemos suponer que, incluso si un genotipo está agravando la producción de cáncer de pulmón, no tiene ningún efecto sobre la cantidad de alquitrán en los pulmones, excepto de manera indirecta (a través de fumar).
- Del mismo modo, debemos suponer que ningún otro factor que afecte al depósito de alquitrán influye en el hábito de fumar.
- Finalmente, la condición $P(x, z) > 0$ del teorema anterior requiere que los altos niveles de alquitrán en los pulmones sean el resultado no solo del consumo de cigarrillos, sino también de otros factores (por ejemplo, la exposición a contaminantes ambientales) y que el alquitrán pueda estar ausente en algunos fumadores (debido quizás a un mecanismo extremadamente eficiente para desechar el alquitrán).

Ejemplo: Fumar y la teoría del genotipo iii

- Para demostrar cómo podemos evaluar el grado en que el consumo de cigarrillos aumenta (o disminuye) el riesgo de cáncer de pulmón, asumiremos un estudio hipotético en el que las tres variables X , Y , Z se midieron simultáneamente en una muestra grande y seleccionada al azar, de la población. Se muestra en la siguiente tabla:

		$P(x, z)$	$P(Y = 1 x, z)$
	Group Type	Group Size (% of Population)	% of Cancer Cases in Group
$X = 0, Z = 0$	Nonsmokers, No tar	47.5	10
$X = 1, Z = 0$	Smokers, No tar	2.5	90
$X = 0, Z = 1$	Nonsmokers, Tar	2.5	5
$X = 1, Z = 1$	Smokers, Tar	47.5	85

Ejemplo: Fumar y la teoría del genotipo iv

- La tabla muestra que el 95 % de fumadores y 5 % de no fumadores han desarrollado altos niveles de alquitrán en sus pulmones. Además, 81 % de los sujetos con depósitos de alquitrán han desarrollado cáncer de pulmón, comparado con el 14 % de aquellos sin depósitos de alquitrán. Finalmente, dentro de estos dos grupos (con y sin alquitrán), los fumadores tienen un mayor porcentaje de cáncer que aquellos no fumadores.
- Los resultados pueden probar que fumar es un factor principal para el cáncer de pulmón. Sin embargo, la industria tabacalera podría argumentar que la tabla cuenta una historia diferente: *que fumar realmente disminuye el riesgo de cáncer de pulmón*. Su argumento es el siguiente. Si decides fumar, entonces tus posibilidades de acumular depósitos de alquitrán son del 95 %, en comparación con el 5 % si decide no fumar. Para evaluar el efecto de los depósitos de alquitrán, observamos por separado dos grupos, fumadores y no

Ejemplo: Fumar y la teoría del genotipo v

fumadores. La tabla muestra que los depósitos de alquitrán tienen un efecto protector en ambos grupos: en los fumadores, los depósitos de alquitrán reducen las tasas de cáncer del 90 % al 85 %; en los no fumadores, reducen las tasas de cáncer del 10 % al 5 %. Por lo tanto, independientemente de si tengo un deseo natural por la nicotina, debería buscar el efecto protector de los depósitos de alquitrán en mis pulmones, y fumar ofrece un medio muy eficaz de adquirir esos depósitos.

- Para resolver la disputa entre las dos interpretaciones, ahora aplicamos la fórmula de la puerta delantera sobre los datos de la tabla.
- Queremos calcular la probabilidad de que al seleccionar aleatoriamente a una persona, ésta desarrollará cáncer, bajo cada una de las siguientes dos acciones: fumar (estableciendo $X = 1$) o no fumar (estableciendo $X = 0$).

Ejemplo: Fumar y la teoría del genotipo vi

- Sustituyendo los valores por $P(z|x)$, $P(y|x, z)$ y $P(x)$, tenemos

$$\begin{aligned}P(Y = 1|do(X = 1)) &= .05(.10 \times .50 + .90 \times .50) \\ &\quad + .95(.05 \times .50 + .85 \times .50) \\ &= .05 \times .50 + .95 \times .45 = .4525\end{aligned}$$

$$\begin{aligned}P(Y = 1|do(X = 0)) &= .95(.10 \times .50 + .90 \times .50) \\ &\quad + .05(.05 \times .50 + .85 \times .50) \\ &= .95 \times .50 + .05 \times .45 = .4975\end{aligned}$$

Ejemplo: Fumar y la teoría del genotipo vii

- El propósito de este ejercicio fue demostrar cómo las suposiciones cualitativas razonables sobre el funcionamiento de los mecanismos, junto con datos no experimentales, pueden producir evaluaciones cuantitativas precisas de los efectos causales. En la realidad, esperaríamos estudios observacionales que involucren variables mediadoras para refutar la teoría del genotipo mostrando, por ejemplo, que las consecuencias mediadoras del fumar (como los depósitos de alquitrán) tienden a aumentar, no a disminuir, el riesgo de cáncer en fumadores y no fumadores por igual.
- La estimación de (5) puede utilizarse para cuantificar el efecto causal del hábito de fumar en el cáncer.