Exact Bayesian structure learning from uncertain interventions

D. Eaton, K. Murphy

University of British Columbia

Tipos de modelos en base a las intervenciones.

- Sin intervención.
- Intervención perfecta.
- Intervención imperfecta.
- Intervención incierta.

Sin intervención.

- Se asume que la distribución de probabilidad condicional (CPD) de cada nodo en el grafo esta dada por
  $p(X_i|X_G, \theta, G) = f_i(X_i|X_G, \theta_i)$
  donde G_i son los padres de i, θ_i son los parámetros de i y f_i es la función de densidad de probabilidad (multinomial, gaussiana, etc.).
- Se asume que cada $p(\theta_i)$ es conjugados de f_i.
- La probabilidad marginal:
  $p(X^{1:N}|G) = \int p(X^{1:N}|G, \theta)p(\theta)d\theta$
  donde N es el numero de casos de datos.
- Ejemplo multinomial-Direchlet:

$$
\begin{aligned}
p(x_i^{1:N}|x_{G_i}^{1:N}) &= \int [\prod_{n=1}^{N} p(x_i^n|x_G^n, \theta_i)]p(\theta_i)d\theta_i \\
&= \prod_{j=1}^{r_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{q_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}
\end{aligned}
\tag{1}
$$

Sin intervención.

$$p(x_i^{1:N}|x_{G_i}^{1:N}) = \prod_{j=1}^{r_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} \prod_{k=1}^{q_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{2}$$

- Donde $N_{ijk} = \sum_{n=1}^{N} I(x_i^n = k, x_{G_i}^n = j)$ son los conteos.
- $Nij = \sum_k N_{ijk}$.
- I(e) es la función indicador, donde I(e)=1 si el evento e es cierto y I(e)=0 en otro caso.
- α_{ijk} son las pseudo cuentas de los hiperparametros de Direchlet.
- $\alpha_{ij} = \sum_k \alpha_{ijk}$.
- r_i is el numero de estados discretos para X_i y q_i es el numero de estados para X_{G_i}.
- $\alpha_{ijk} = 1/q_i r_i$
- la probabilidad marginal de cada nodo es:
  $p(X^{1:N}|G) = \prod_{i=1}^{d} p(X_i^{1:N}|X_{G_i}^{1:N})$
  donde d es el numero de nodes.

Intervención perfecta.

- Se establece $X_i^n = x_i^*$, donde x_i^* es el estado objetivo para el nodo i (se asume que es fijo y conocido).
- Se modifica en CPD en este caso
  $p(X_i|X_{G_i}, \theta) = I(X_i = x_i^*)$
- Se ve que X_i es "cut off" de sus padres X_{G_i}

Intervención Imperfecta.

Una forma de modelo de intervención es introducir nodos de intervención, que actúen como “switching parents”.

- Si I_i^n=1, entonces se desarrolla una intervención sobre el nodo i en el caso n.
- Se tiene un conjunto diferente de parámetros en caso de I_i^n=0.
- Esto es:
  $p(X_i|X_{G_i}, I_i = 1, \theta, G) = f_i(X_i|X_{G_i}, \theta_i^1)$ y
  $p(X_i|X_{G_i}, I_i = 0, \theta, G) = f_i(X_i|X_{G_i}, \theta_i^0)$.

Intervención Imperfecta.

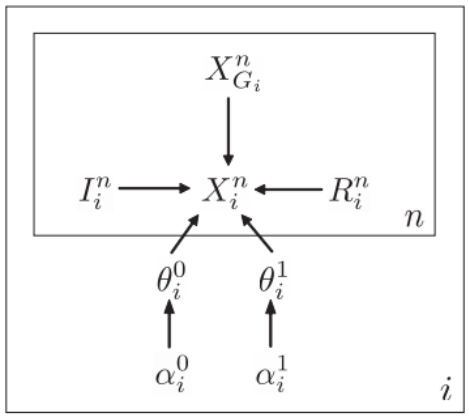


Figure 1: Modelo de cambio de mecanismo

- Si I_i^n=1 representa intervención, entonces X_i usa los parámetros θ_i^1.
- $\alpha_i^{0/1}$ son los hyper-parámetros.
- R_i^n modela el grado de efectividad de la intervención.
- si un nodo i no es intervenible, se coloca I_i^n=0 para todos los n.

Intervención Imperfecta.

Cuando se tiene datos intervenibles, se modifica la función de probabilidad marginal local, particionando la data en los casos en que X_i es parcialmente observada.

$$\begin{aligned} p(x_i^{1:N}|x_{G_i}^{1:N}, I_i^{1:N}) = &\int [\prod_{n:I_i^n=0} p(x_i^n|x_G^n, \theta_i^0)]p(\theta_i^0)d\theta_i^0 \\ &\times \int [\prod_{n:I_i^n=1} p(x_i^n|x_G^n, \theta_i^0)]p(\theta_i^1)d\theta_i^1 \end{aligned} \tag{3}$$

En el caso de intervención perfecta, solo se evalúa el segundo termino.

Intervención Imperfecta.

También se puede añadir al modelo cuando hay casos de intervenciones no confiables.

- Se introduce el indicador R_i^n, donde se R_i^n=1 siempre suceden las intervenciones, R_i^n=0 no suceden.
- En este caso de usa $p(X_i|X_{G_i}, \theta, I_i = 1)$

Intervención imperfecta.

Otro modelo es intervención ”suave”.
Donde una intervención incrementa la probabilidad que un nodo entre en su estado objetivo x_i^*.

- Los parámetros $\theta_i^{1/0}$ tienen hyper-parámetros dependientes: $\theta_i^{1/0} \sim Dir(\alpha_{ij}^{0/1})$.
- $\alpha_{ij}^1 = \alpha_i j^0 + w_i \vec{e}_t$.
- Donde t=x_i^* es el objetivo evaluado.
- $\vec{e}_t$=(0,...0,1,0,...0) con 1 en tº posición.
- w_i es la fuerza de la intervención, con $w_i \to \infty$ se convierte en intervención perfecta.

Intervención incierta. i

Se representa la intervención con objetivos inciertos, así como efectos inciertos.

- Se asume que existe conexiones entre los nodos I_i y los nodos regulares X_i.
- Se asume que I_i puede tener múltiples hijos regulares.
- Se dice que se hace un ”fat hand” debido a que se ”tocan” varios nodos a la vez.
- Si un nodo regular es afectado por varios nodos de intervención, se crea un nuevo vector de parámetros por cada posible combinación de intervenciones.

Intervención incierta.

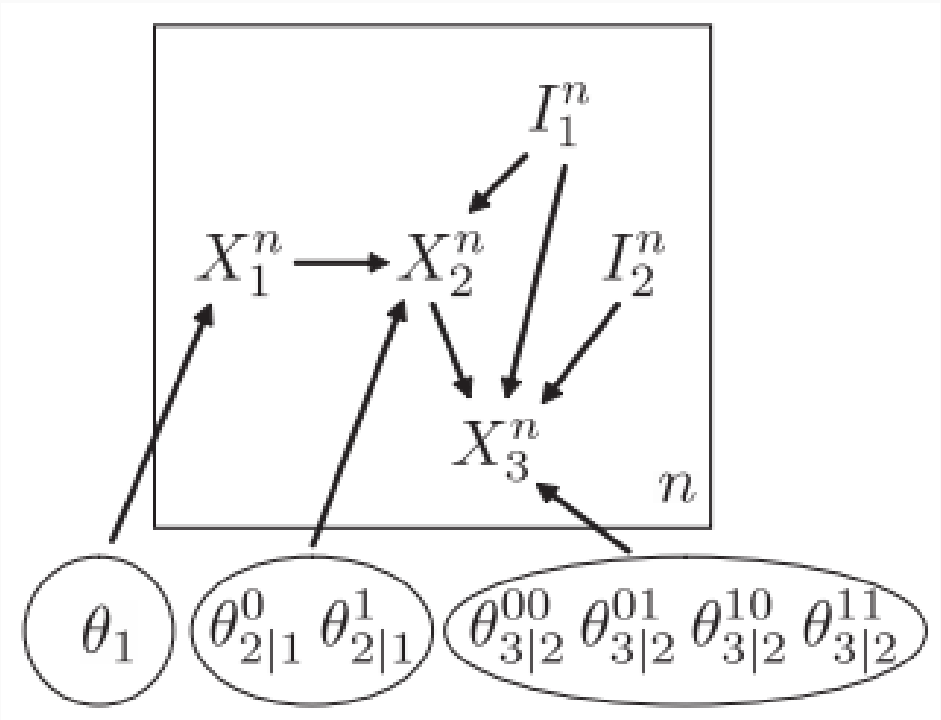


Figure 2: Ejemplo de intervención ”fat hand”

Se asume que los nodos de intervención son exógenos y fijos.
No existen arcos $I_i \to I_j$ ni $X_i \to I_i$

Intervención incierta.

Se modifica la función de probabilidad marginal.

- X_{G_i} son los nodos regulares padres de i.
- I_{G_i} son los nodos de intervención padres de i.
- θ_i^l son los parámetros del nodo i dado los padres de intervención l.

$$p(x_i^{1:N}|x_{G_i}^{1:N}, I_{G_i}^{1:N}) = \prod_l \int [\prod_{n:I_{G_i}^n=l} p(x_i^n|x_{G_i}^n, \theta_i^l)]p(\theta_i^l)d\theta_i^l \tag{4}$$

Algoritmo para el aprendizaje de la estructura

Se utiliza el algoritmo de programación dinámica (DP) para encontrar la estructura de una red bayesiana. Este posee menor complejidad al resto de los algoritmos planteados $O(d2^d)$.

La entrada de este algoritmo es:

- Un prior sobre los nodos ordenados $q_i(U_i)$.
- Un prior de los posibles padres $\rho_i(G_i)$.
- Una función de probabilidad marginal local.

Obtener Priors

- El ordenamiento de nodos $\prec$ esta dado por $(U_1, ..., U_d)$.
- U_i=j:j$\prec$i son el conjunto de nodos que preceden i.
- Se asume un prior uniforme sobre el ordenamiento $q_i(U_i) \propto 1$.
- $\rho_i(G_i) \propto \binom{d-1}{|G_i|}^{-1}$, si $|G_i| \leq K$ sino $\rho_i(G_i) \propto 0$.

$$p(\prec, G) = \frac{1}{Z} \prod_{i=1}^{d} q_i(U_i)\rho_i(G_i) \times I(\prec, G) \tag{5}$$

- G debe ser consistente con $\prec$ y $\prec$ define in orden total sobre G (DAG).
- Z es una constante de normalización.

Obtener Likelihoods

El parámetro final es la función de probabilidad condicional local $p(x_i^{1:N}|x_{G_i}^{1:N}, I_{G_i}^{1:N})$.

Esta dada por formula de la multinomial-Dirichlet.

Layering

Se establece el grafo con los capas de nodos, uno son los nodos regulares y los otros los nodos de intervención.

- $V = X \cup I$

La complejidad computacional del algoritmo DP es ahora $O(d2^{d_x} + d^{k+1}C(N)$, donde C(N) es el costo del probabilidad marginal.

Las capas son cruciales para una eficiente intervención incierta.

Experimentos

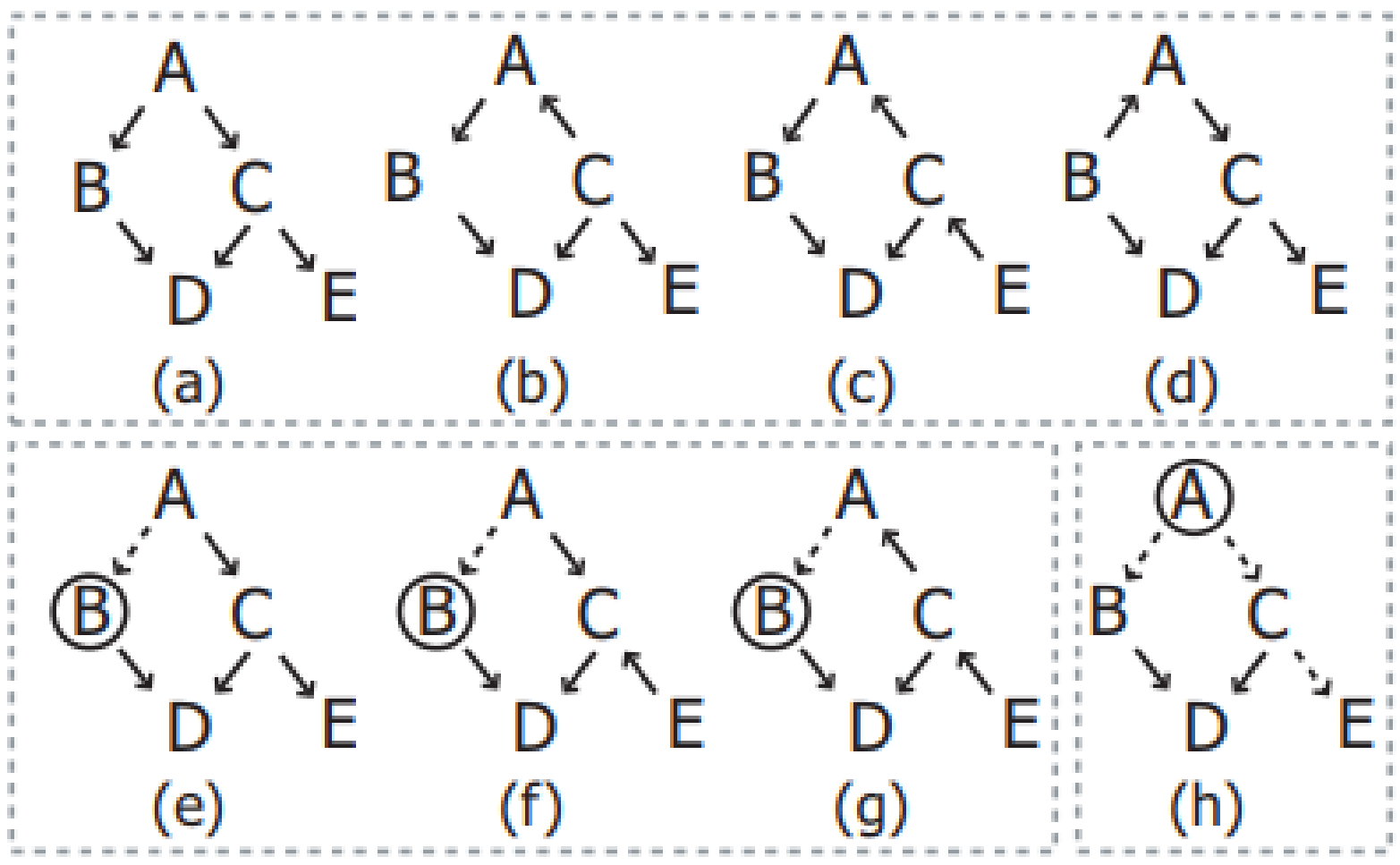


Figure 3: Grafo causal del cáncer.

Experimentos

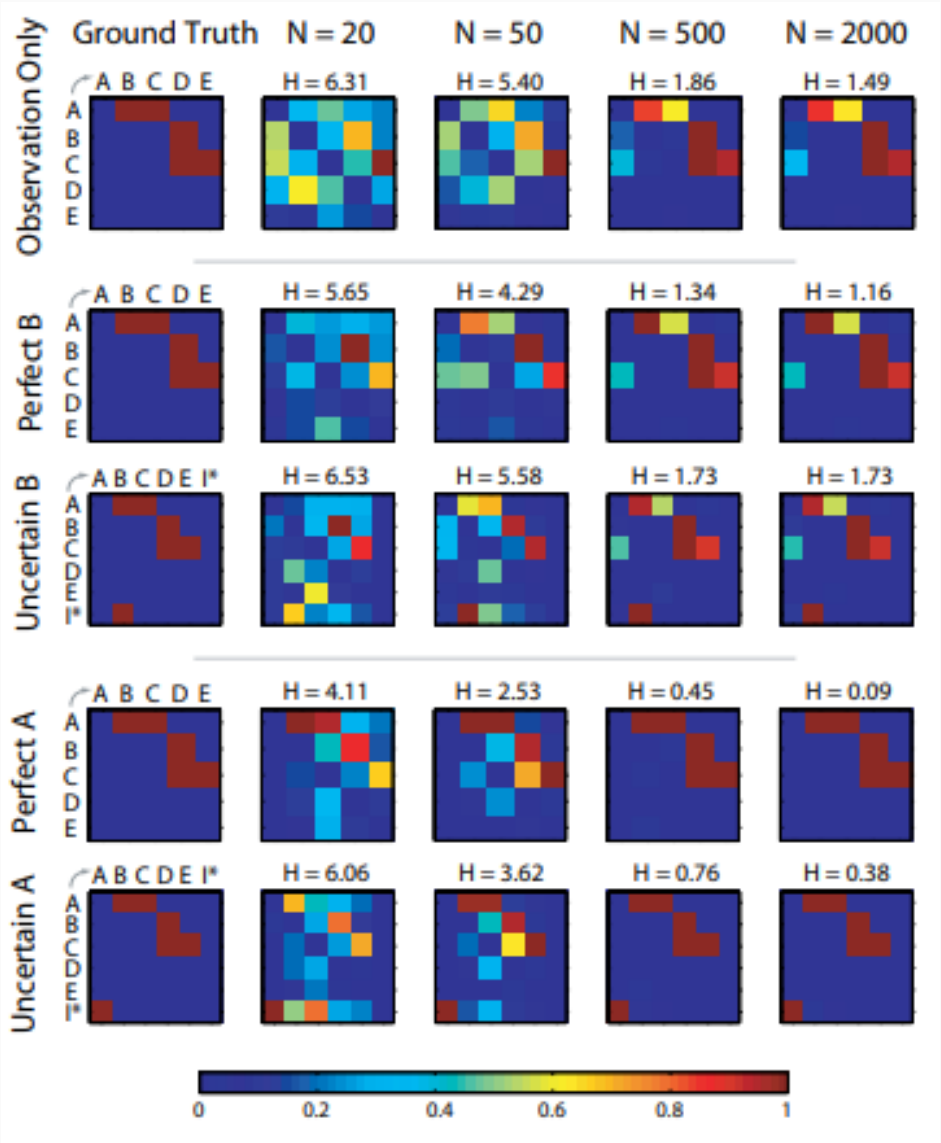
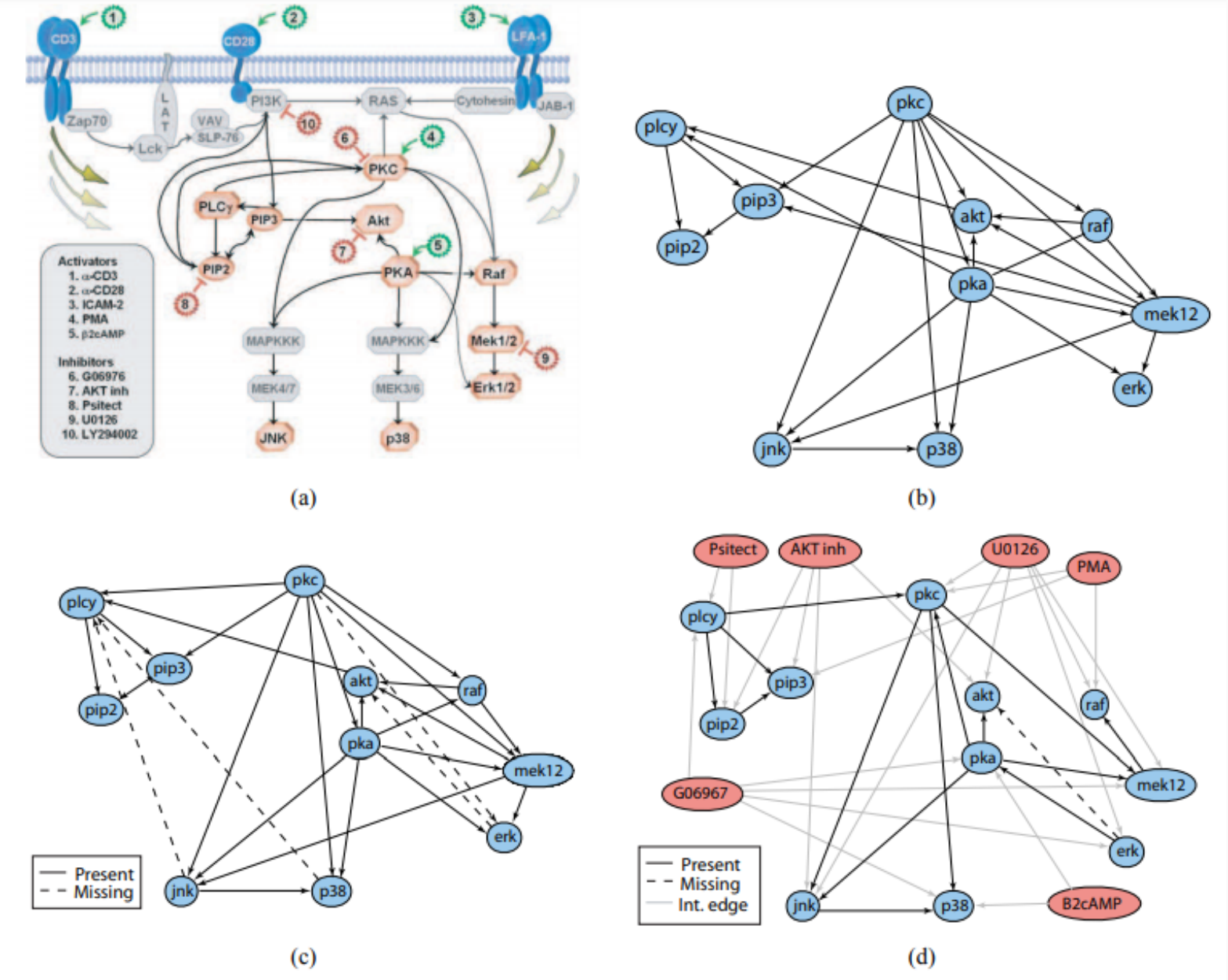


Figure 4: Resultados de hacer diversas intervenciones.

Experimentos



Conclusión.

- Un cuello de botella es la cantidad de variables que permite el algoritmo d=20.
- Layering podría ser usado para aprender redes bayesianas dinámicas.
- Uno problema es la uniformidad de los prior $p(G)$ que el algoritmo DP implícitamente usa.