INNOV 2025, Sep. 28- Oct. 2, 2025. Lisbon, Portugal

Marginal Information and its use for Structure Learning

Sung-Ho Kim
Korea Advanced Institute of Science and Technology (KAIST)
email: sungkim@kaist.ac.kr

.

September 28, 2025

Short Biography

Current Position: Professor Emeritus of Statistics, Dept of Mathematical Sciences, Korea Advanced Institute of Science and Technology(KAIST).

Education: B.Sc. in Math (Seoul National Univ, 1980, S. Korea)/ Ph.D. in Statistics (Carnegie Mellon University, U.S.A, 1989)

Research fields: Multivariate Analysis, Graphical modelling, Structure learning, Structure combination, Large scale modelling

Professional experience: Educational Testing Service (research scientist, 1989-1993, USA.), KAIST (Professor, since 1993, S. Korea).

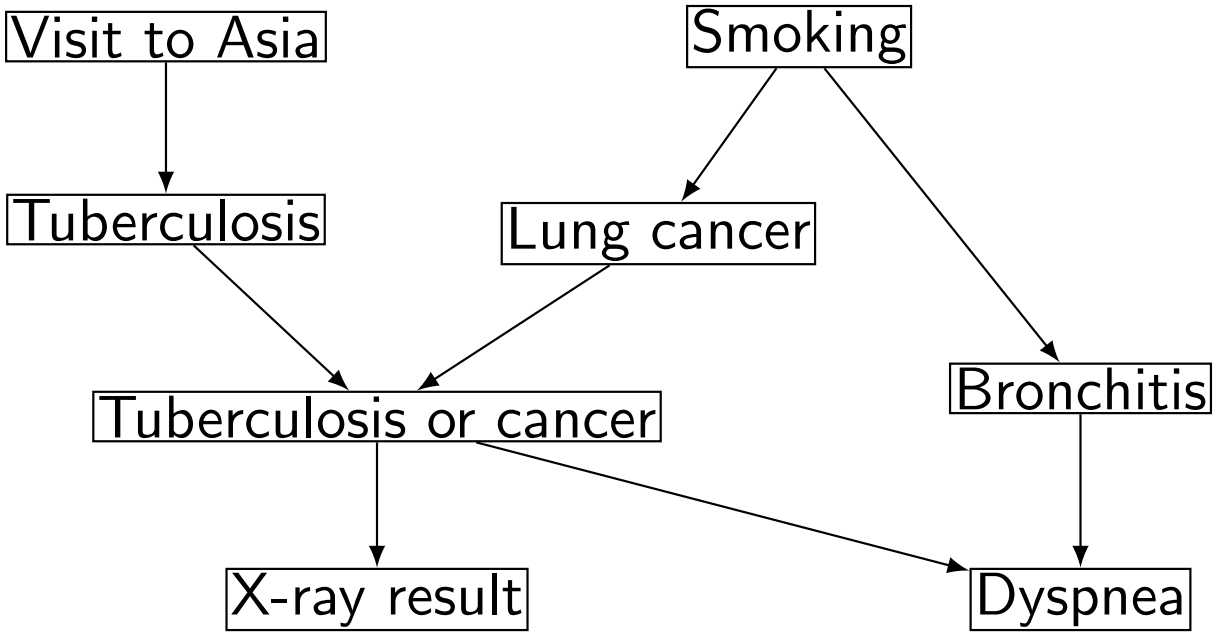
Professional membership: International Statistical Institute, Elected member, Since Jan. 2003.

Professional activity: Over 70 research articles in peer-reviewed journals, over 30 invited talks in conferences and universities, and over 200 evaluations of research articles.

Outline

1. Introduction

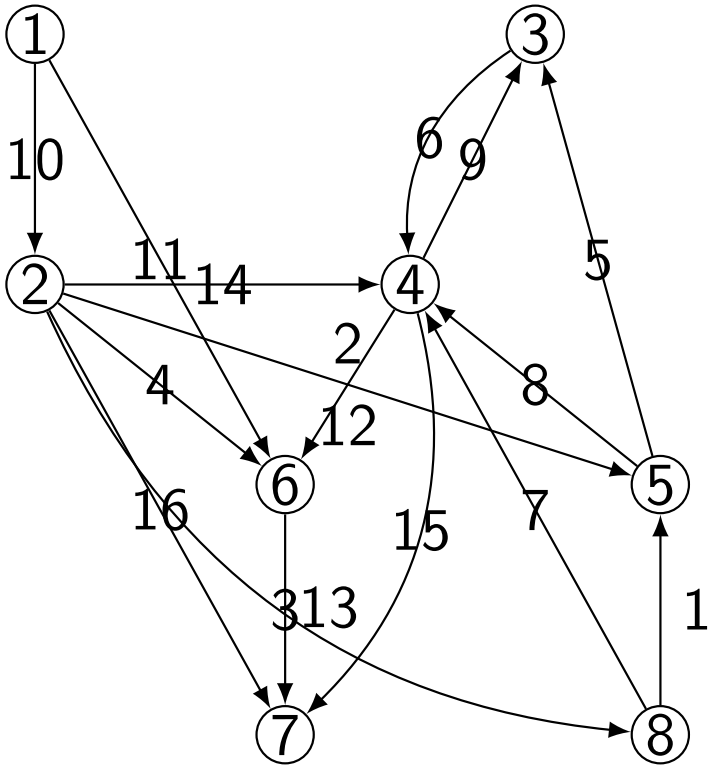
Introduction



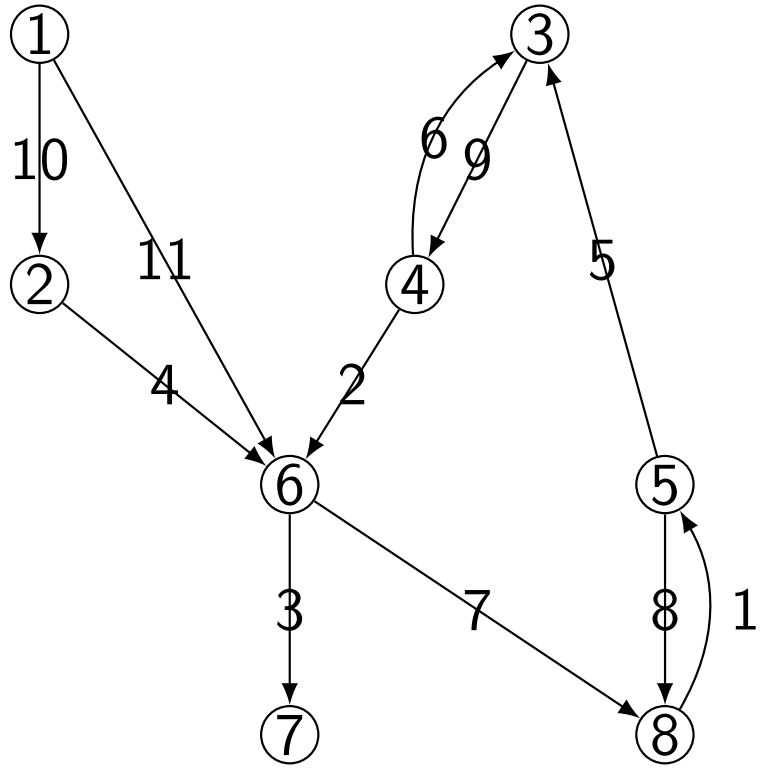
(a) Asia Network

Visit \ *Tuberculosis*	No	Yes
No	0.9	0.1
Yes	0.3	0.7

Table 1: Conditional probability of "Tuberculosis" given "Visit to Asia"

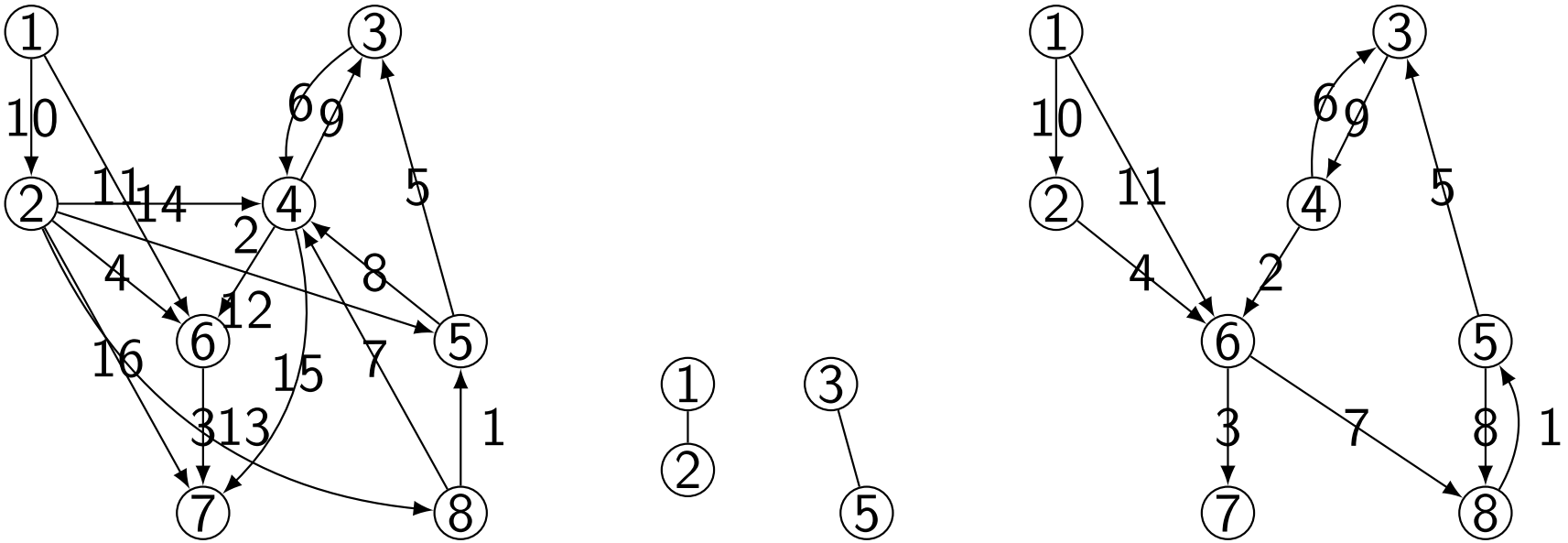


(b) G_{GHC}



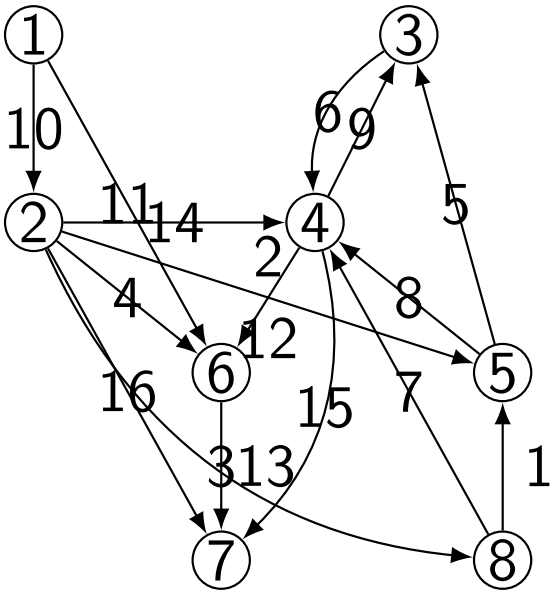
(c) G_{MCB}

Figure 1: The sequence of structure learning with *Asia* data. The final result of the learning is in Fig. 8. The numbers on edges indicate the order of edge appearance.

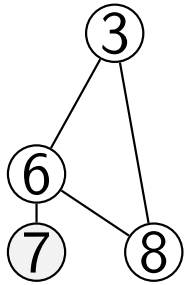


(a) G_{GHC} (b) Marginal structure (c) G_{MCB}

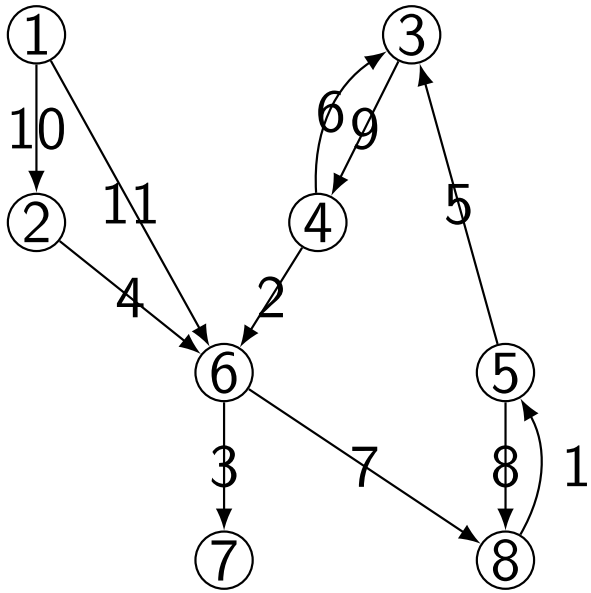
Figure 2: Effect of Marginal Information for Structure Learning



(a) G_{GHC}

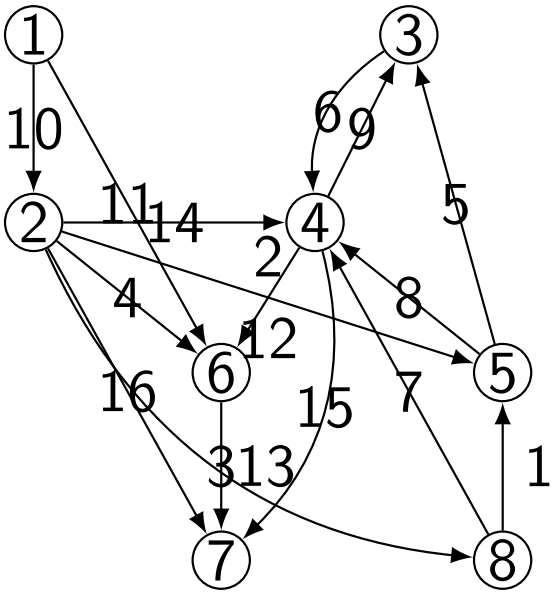


(b) Marginal structure

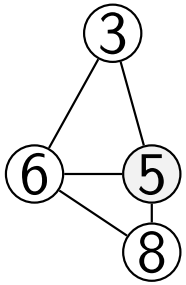


(c) G_{MCB}

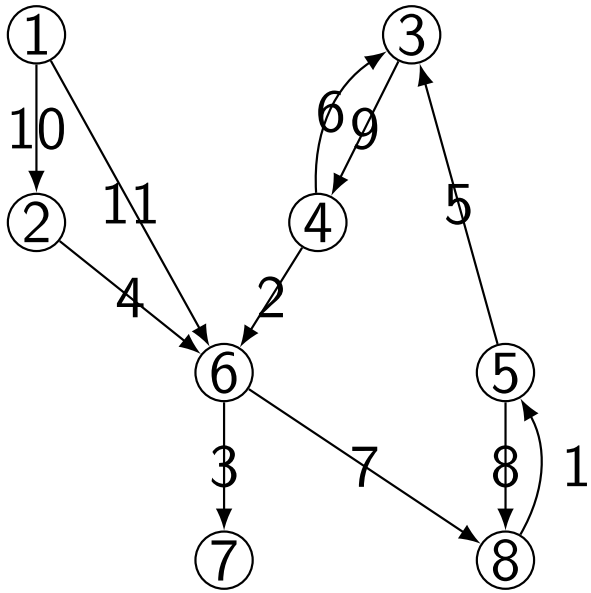
Figure 3: Effect of Marginal Information for Structure Learning



(a) G_{GHC}



(b) Marginal structure



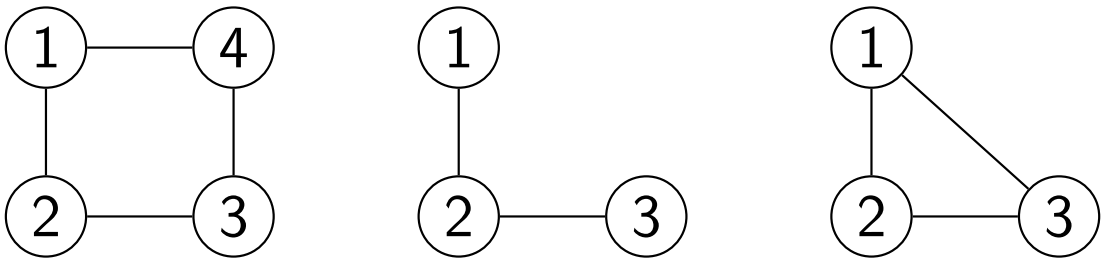
(c) G_{MCB}

Figure 4: Effect of Marginal Information for Structure Learning

2. Graphs and Markov Properties

Graph Basics(1/2)

- Graph $G = (V, E)$: V is a set of nodes, $E \subset V \times V$ is a set of edges.
- Induced subgraph: $G_A = (A, E \cap (A \times A))$ for $A \subset V$.

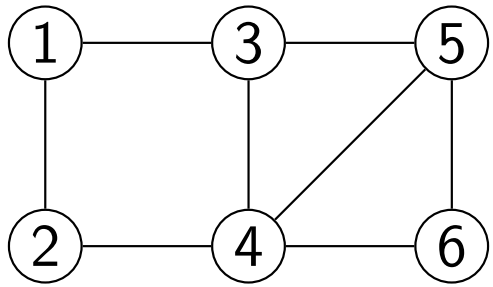


(a) G (b) G_A (c) $G_{\setminus A}$

Figure 5: Two types of subgraph of G on $A = \{1, 2, 3\}$.

Graph basics(2/2)

• Separation: A and B are separated by S if all the paths from A to B pass through S.



Markov Properties(1/5)

- Let $X = \{X_1, \ldots, X_n\}$ be random variables with joint distribution P.
- For $A = \{1, 2\}$, we will write X_A for (X_1, X_2).
- Let $V = \{1, 2, \cdots, n\}$. For two random vectors X_A and X_B with $A, B \subseteq V$ and $A \cap B = \emptyset$, we say that X_A and X_B are stochastically independent if

$$P(x_{A \cup B}) = f_1(x_A) f_2(x_B).$$

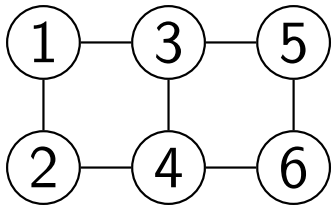
- Now suppose $A \cap B \neq \emptyset$ and let $A \cap B = C$. We say that $X_{A \setminus B}$ and $X_{B \setminus A}$ are conditionally independent given X_C (or $X_{A \setminus B} \perp X_{B \setminus A} | X_C$) if

$$P(x_{A \setminus B}, x_{B \setminus A} | x_C) = f_3(x_{A \setminus B}) f_4(x_{B \setminus A}).$$

Markov Properties(2/5)

- P is globally Markov w.r.t. G if:
  $X_A \perp X_B \mid X_S$ whenever A and B are separated by S in G.
- $M(G)$: A set of distributions globally Markov to G.
- G is a perfect map of P if $P \in M(G)$ and all conditional independencies in P are encoded in G.

Markov Properties(3/5)

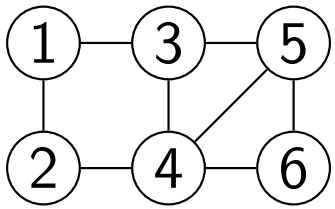


(a) G_1

$$
\begin{aligned}
P_1(x_1, \cdots, x_6) &= g_1(x_1, x_2) g_2(x_1, x_3) g_3(x_2, x_4) g_6(x_3, x_4) g_7(x_3, x_5) \\
&\quad \times g_8(x_4, x_6) g_9(x_5, x_6).
\end{aligned}
$$

$$
\begin{aligned}
P_1(x_1, \cdots, x_6 | x_3, x_6) &= g_1(x_1, x_2) g_2(x_1, x_3) g_3(x_2, x_4) g_6(x_3, x_4) g_7(x_3, x_5) \\
&\quad \times g_8(x_4, x_6) g_9(x_5, x_6) / P(x_3, x_6). \\
&= h_1(x_1, x_2) h_2(x_2, x_4) h_3(x_5)
\end{aligned}
$$

Markov Properties(4/5)

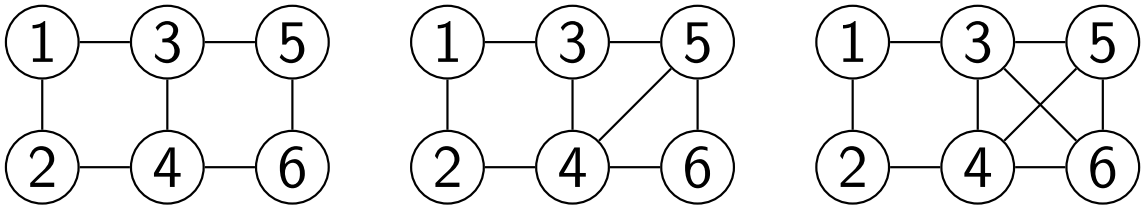


(b) G_2

$$
\begin{aligned}
P_2(x_1, \cdots, x_6) &= g_1(x_1, x_2) g_2(x_1, x_3) g_3(x_2, x_4) g_4(x_3, x_4, x_5) \\
&\quad \times g_5(x_4, x_5, x_6).
\end{aligned}
$$

$$
\begin{aligned}
P_2(x_1, \cdots, x_6 | x_3, x_6) &= g_1(x_1, x_2) g_2(x_1, x_3) g_3(x_2, x_4) g_4(x_3, x_4, x_5) \\
&\quad \times g_5(x_4, x_5, x_6) / P(x_3, x_6). \\
&= h_1(x_1, x_2) g_3(x_2, x_4) h_4(x_4, x_5)
\end{aligned}
$$

Markov Properties(5/5)



(c) G_1 (d) G_2 (e) G_3

Figure 6: Graphs of 6 nodes

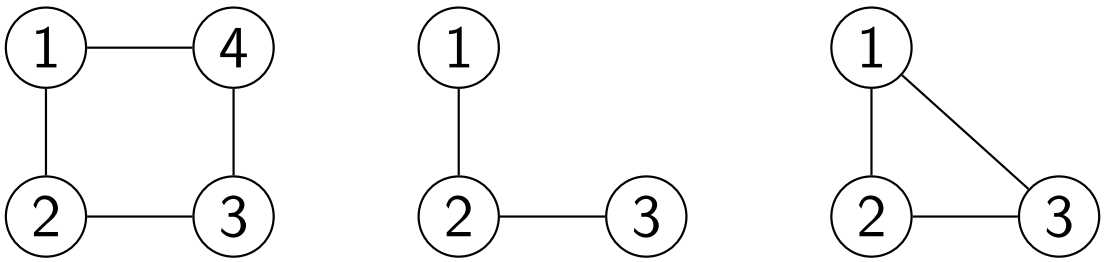
$$
\begin{aligned}
P_1(x_1, \cdots, x_6) &= g_1(x_1, x_2)g_2(x_1, x_3)g_3(x_2, x_4)g_6(x_3, x_4)g_7(x_3, x_5) \\
&\quad \times g_8(x_4, x_6)g_9(x_5, x_6). \\
P_2(x_1, \cdots, x_6) &= g_1(x_1, x_2)g_2(x_1, x_3)g_3(x_2, x_4)g_4(x_3, x_4, x_5)g_5(x_4, x_5, x_6). \\
P_3(x_1, \cdots, x_6) &= g_1(x_1, x_2)g_2(x_1, x_3)g_3(x_2, x_4)g_{10}(x_3, x_4, x_5, x_6).
\end{aligned}
$$

==>

1. G_i is a perfect map of P_i for $i = 1, 2, 3$.
2. P_1, P_2, and P_3 are all in $M(G_3)$.
3. P_1, P_2 are in $M(G_2)$. $P_1 \in M(G_1)$.

Markovian Subgraphs

- Given $G = (V, E)$ and $A \subset V$, define the Markovian subgraph $G_{\_A} = (A, E_{\_A})$:
  $(i, j) \in E_{\_A}$ if $(i, j) \in E$ or there exists a $(V \setminus A)$-path between i and j in G.
- Independence properties of marginal distribution P_A are fully captured by $G_{\_A}$.
- If P is globally Markov w.r.t. G, then P_A is globally Markov w.r.t. $G_{\_A}$. In other words, if $P \in M(G)$, then $P_A \in M(G_{\_A})$.



(a) G (b) G_A (c) $G_{\_A}$

Marginal Model Structure of a DAG

The *ancestral set* of A, $An(A)$, is defined as

$$An(A) = \bigcup_{\alpha \in A} an(\alpha) \cup A.$$

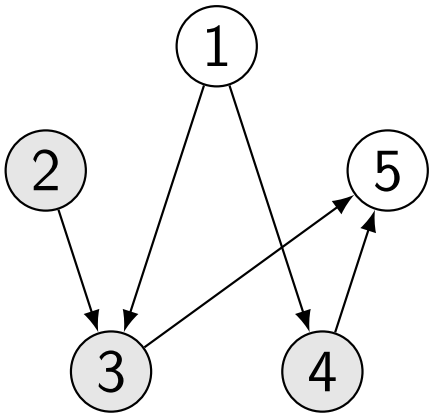
The *moral ancestral graph* of A is defined as the moral graph of the induced subgraph of G on $An(A)$, i.e., $(G_{An(A)})^m$.

Theorem 1

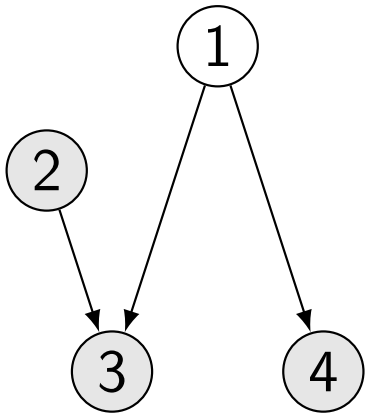
[a] *Let G be a DAG with its set of nodes V and let A_1, A_2 and S be disjoint subsets of V. Then A_1 and A_2 are d-separated by S in G if and only if A_1 and A_2 are separated by S in $(G_{An(A_1 \cup A_2 \cup S)})^m$.*

[a] Lauritzen S (1996) Graphical Models. Oxford, United Kingdom: Clarendon Press

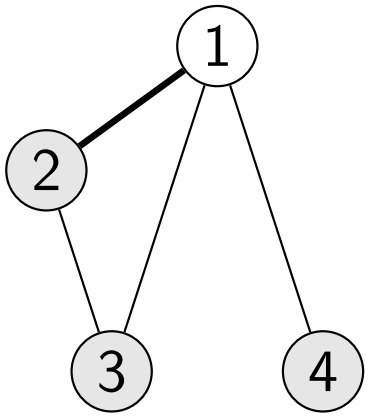
Moral Ancestral Graph



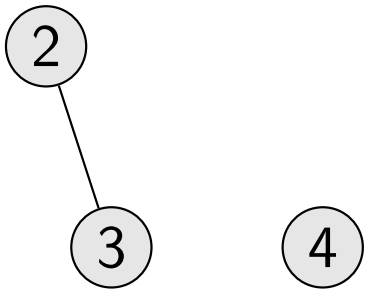
(d) G



(e) $G_{An(A)}$



(f) $(G_{An(A)})^m$



(g) $((G_{An(A)})^m)_A$

Figure 7: A process from G to $((G_{An(A)})^m)_A$ with $A = \{2, 3, 4\}$

The Markovian subgraph connects a DAG and its moral ancestral graph in the following sense.

Proposition 1

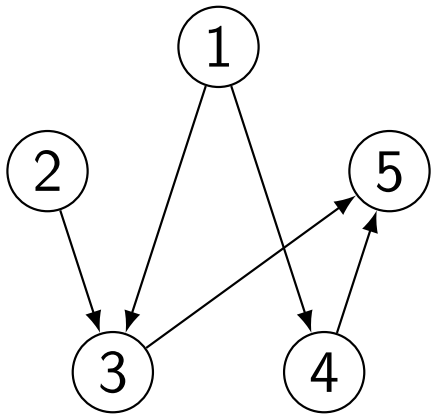
If the distribution P of a random vector X_V is faithful to a DAG G, then P_A is globaly Markov wrt $G_{\_A}$.

3. Factorization

Factorization wrt Bayesian Network

We consider a set of random variables $X = \{X_1, \cdots, X_n\}$ where X_i's are categorical or finitely discrete. A *Bayesian network* is a directed acyclic graph $G = (V, E)$ together with X and a set of conditional probability tables $\Theta = \{\theta_{x_i|x_{pa_G(i)}}\}$ where $\theta_{x_i|x_{pa_G(i)}} = P(X_i = x_i | X_{pa_G(i)} = x_{pa_G(i)})$. We say a probability distribution P on X *factorizes* over G if

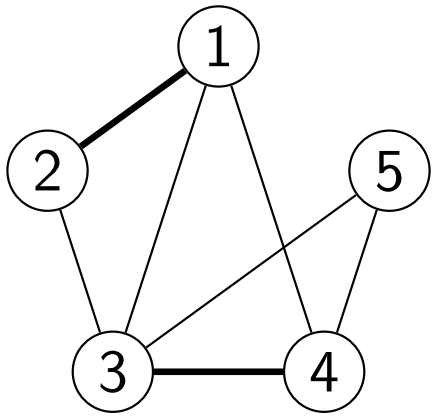
$$P(X = x) = \prod_{i=1}^{n} P(X_i = x_i | X_{pa(i)} = x_{pa(i)}).$$



Factorization wrt UDG

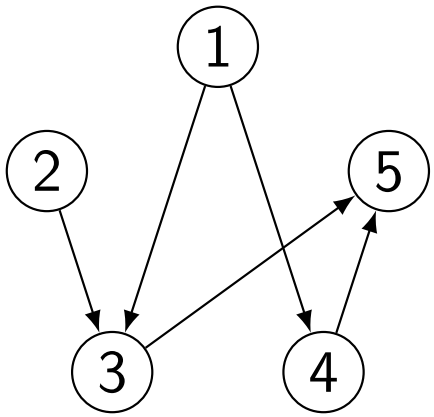
We say a probability distribution P on X *factorizes* over an UDG G if for each maximally complete subsets (i.e., cliques) $c \subset V$, there exists a non-negative function g_c that depends on X_c such that

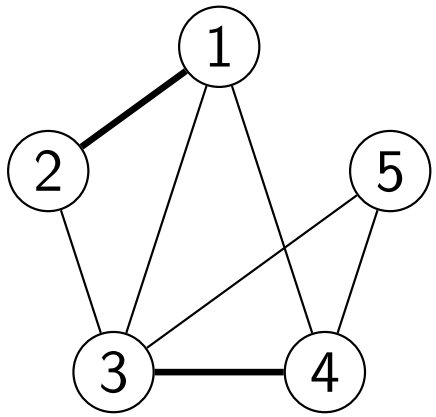
$$P(X = x) = \prod_{c:\ \text{clique in } G} g_c(X_c = x_c).$$



Bayesian network and moral graph

$$P(X = x) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_1)P(x_5|x_3, x_4).$$
$$= f_1(x_1, x_2, x_3)f_2(x_1, x_4)f_3(x_3, x_4, x_5)$$





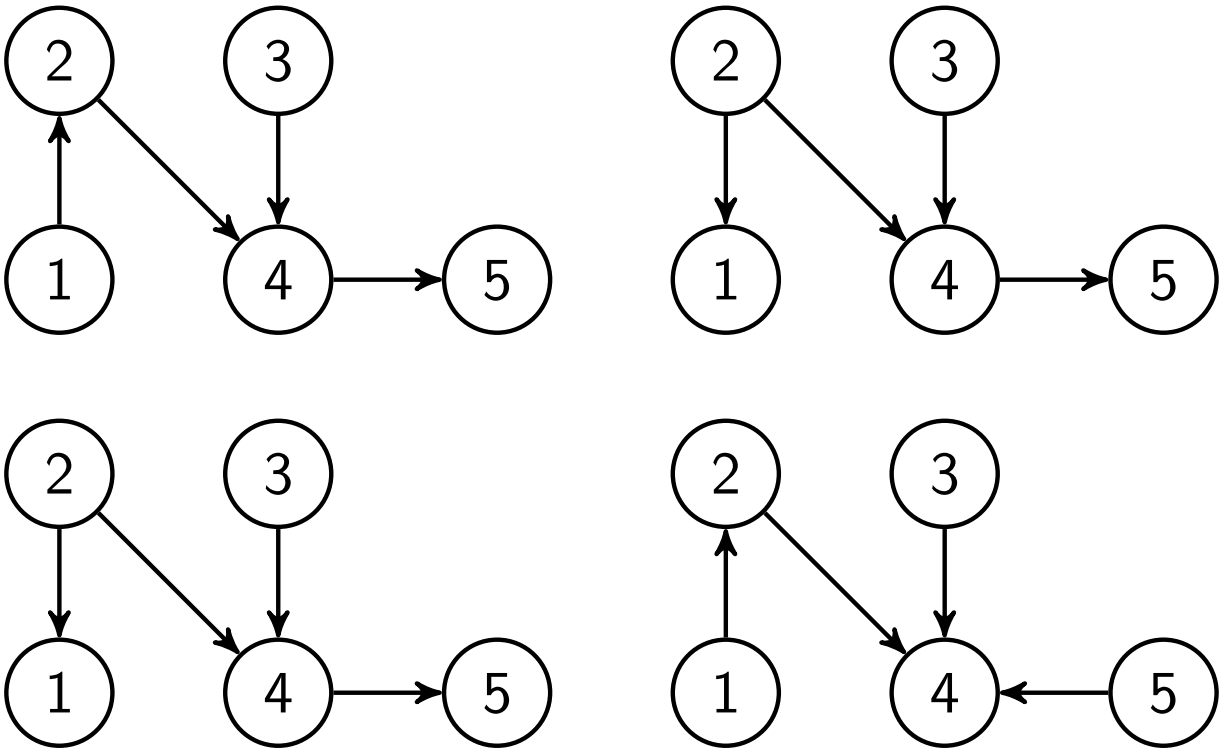
4. Markov/Score equivalence

Markov/Score equivalence

Theorem 2

(Markov equivalence)[a] *Two DAGs are Markov equivalent if and only if they have the same skeletons and the same v-structures.*

[a] Verma T, Pearl J (1991) Equivalence and synthesis of causal models. Uncertain Artif Intell 6:220–227



A property of a score function is defined as

Definition 3

(Score Equivalence)[a] A score function f is said to satisfy *score equivalence* if for given data D and two DAGs G and G' which are Markov equivalent,

$$f(G, D) = f(G', D).$$

[a] Heckerman D, Geiger D, Chickering D (1994) Learning bayesian networks: the combination of knowledge and statistical data. Proc Uncertain Artif Intell 10:293–301

5. Score Function

Score function

The posterior probability of a DAG G given the data D and the marginal structures $\hat{G}_i$'s can be computed by

$$
\begin{aligned}
&P(G|D, \hat{G}_1, \cdots, \hat{G}_k) \\
&\quad \propto P(D, G, \hat{G}_1, \cdots, \hat{G}_k) \\
&\quad = P(G)P(D|G)P(\hat{G}_1, \cdots, \hat{G}_k|G, D) \\
&\quad = P(G)P(D|G)\prod_{i=1}^{k} P(\hat{G}_i|G, D) \\
&\quad = P(G)P(D|G)\prod_{i=1}^{k} P(\hat{G}_i|G_i, D). \qquad (1)
\end{aligned}
$$

Let p_i be the edge-error probability for graph G_i. Then

$$P(\hat{G}_i|G_i, D) = p_i^{\delta_i}(1-p_i)^{m_i-\delta_i} \tag{2}$$

where $m_i = \frac{1}{2}|A_i|(|A_i|-1)$ is the total number of the edges of the complete UDG on A_i and

$$\delta_i = |(E_i \cup \hat{E}_i) - (E_i \cap \hat{E}_i)| \tag{3}$$

is the size of the structural difference between G_i and $\hat{G}_i$.

If we apply the logarithm to the expression in Eq. (1),

$$\log P(G|D, \hat{G}_1, \cdots, \hat{G}_k)$$

$$\propto \log P(G)P(D|G) + \sum_{i=1}^{k} \delta_i \log \frac{p_i}{1-p_i} + \sum_{i=1}^{k} m_i \log(1-p_i).$$

For structure learning, we use the *marginally corrective Bayesian*(MCB) score:

$$\text{MCB-score}(G|D, \hat{G}_1, \cdots, \hat{G}_k)$$

$$= \log P(G)P(D|G) + \sum_{i=1}^{k} \delta_i \log \frac{p_i}{1-p_i} \tag{4}$$

where the term $\log P(G)P(D|G)$ is a traditional Bayesian score function such as the BDeu (Bayesian Dirichlet equivalent uniform) score.

No Data Solution

Suppose that data are not available for p_i's. Then, assuming the Beta distribution $Beta(\alpha_i, \beta_i)$ as a prior distribution on p_i in Eq. (2), the probability $P(\hat{G}_i|G_i, D)$ is obtained by

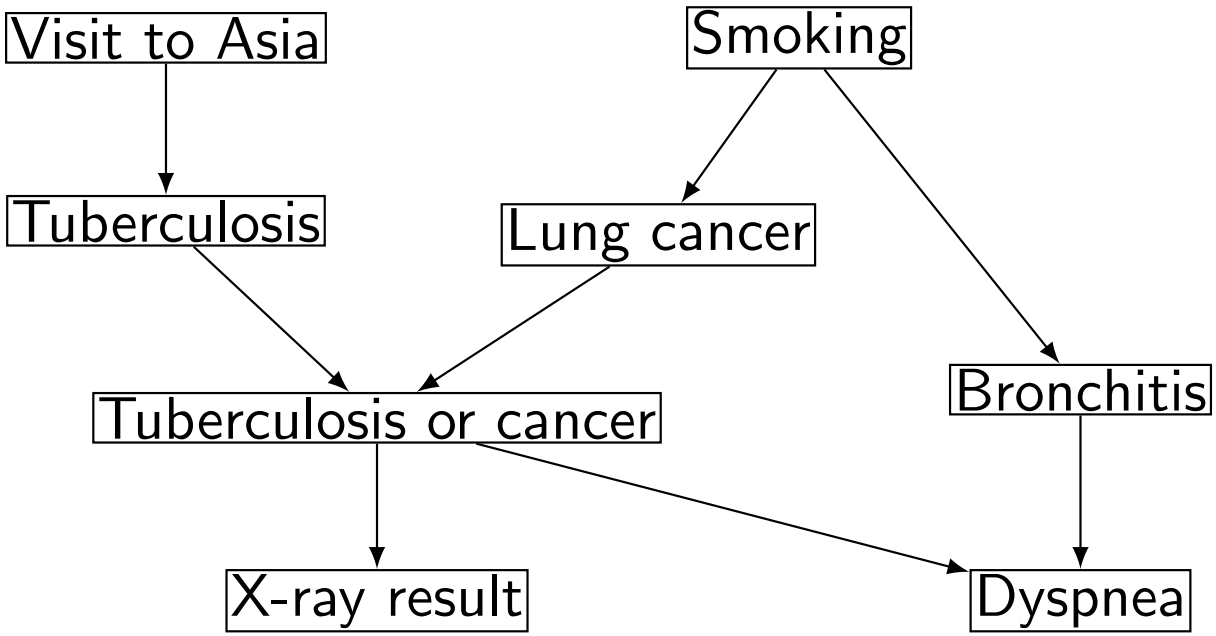
$$\begin{aligned} & P(\hat{G}_i|G_i, D) \\ &= \int_0^1 P(\hat{G}_i|G_i, p_i)\pi(p_i)dp_i \\ &= \int_0^1 p_i^{\delta_i}(1-p_i)^{m_i-\delta_i}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}p_i^{\alpha_i-1}(1-p_i)^{\beta_i-1}dp_i \\ &= \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}\frac{\Gamma(\alpha_i+\delta_i)\Gamma(m_i+\beta_i-\delta_i)}{\Gamma(\alpha_i+\beta_i+m_i)}. \end{aligned} \tag{5}$$

We define the *MCB score without estimation* (MCB* score) as:

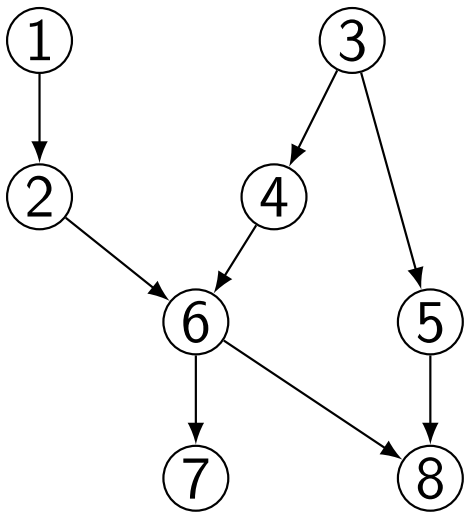
$$\text{MCB*-score}(G|D, \hat{G}_1, \cdots, \hat{G}_k)$$
$$= \log P(G)P(D|G) + \sum_{i=1}^{k}\{\log\Gamma(\alpha_i + \delta_i) + \log\Gamma(m_i + \beta_i - \delta_i)\}. \quad (6)$$

6. Experiment

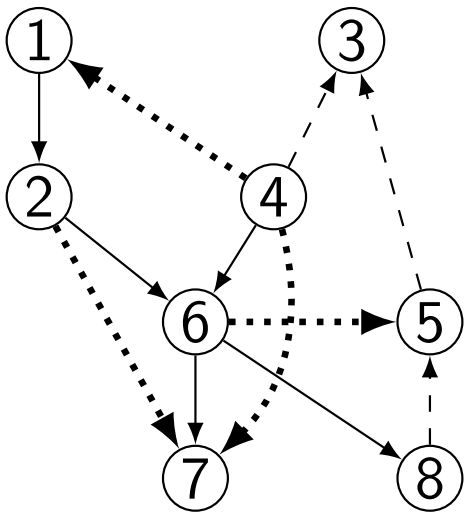
Asia network



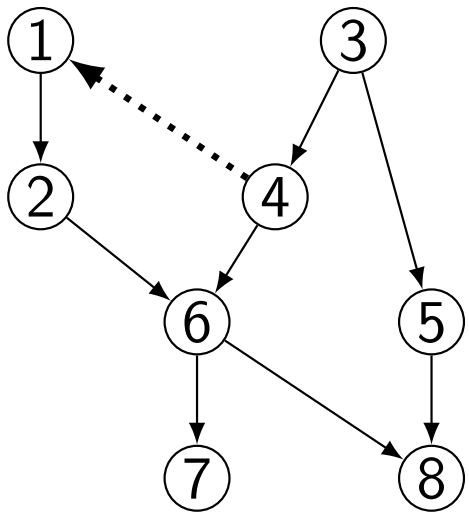
(a) Asia Network



(b) G



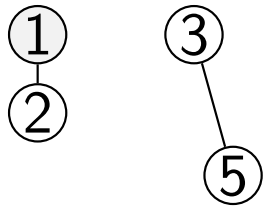
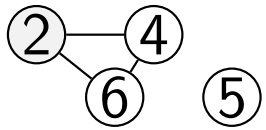
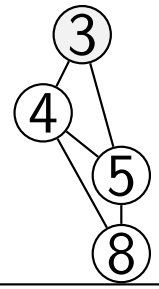
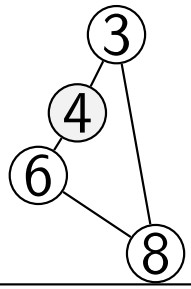
(c) G_{HC}

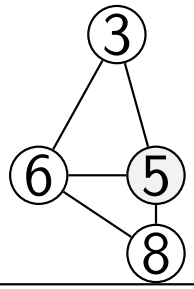
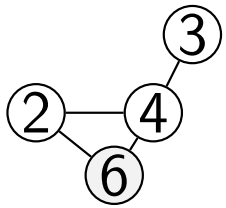
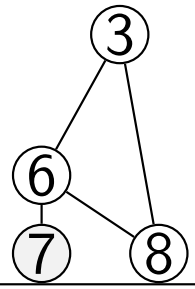
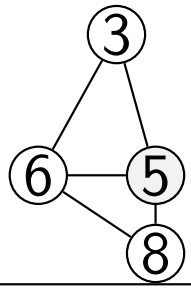


(d) G_{MCB}

Figure 8: The DAGs for *Asia* datasets; G is the true DAG, G_{HC} is the DAG obtained by GHC and G_{MCB} is the DAG obtained by MCB-HC.

Table 2: Marginal structures with the estimates $\hat{p}_i$ of their edge-error probabilities based on *Asia* data

Node	1	2	3	4
$\widehat{G}_i$				
$\hat{p}_i$	0.05	0.19	0.02	0.08

Node	5	6	7	8
$\widehat{G}_i$				
$\hat{p}_i$	0.05	0.01	0.06	0.04

7. Summary

1. Structure learning is made in a sequential manner. Thus the order of edge/arrow selection during the learning process affects significantly the final model structure.
2. Pieces of marginal structure information are helpful in fixing local structure errors during learning by using proper score function.
3. The marginal model structures may be provided as directed or undirected graphs for learning Bayesian networks. This is because the graph separateness is the same between the two types of graph except the nodes involved in v-structures.
4. The idea of using marginal structures for structure learning is a good example of expanding the notion of prior information for statistical learning. The prior information used to be parametric given in probability distributions. But it could also be non-parametric as given in marginal structures.
5. This line of research is in need of better score functions so that we could use various types of marginal structure information.

Selected References

1. Amirkhani H, Rahmati M, Lucas P, Hommersom A (2017) Exploiting experts knowledge for structure learning of bayesian networks. IEEE Trans Pattern Anal Mach Intell 39(11):2154–2170.
2. Chickering D (2002) Optimal structure identification with greedy search. J Mach Learn Res 3:507–554.
3. Dawid, A.P. and Lauritzen, S.L. (1993) Hyper Markov laws in the statistical analysis of decomposable graphical models, *Annals of Statistics* 21: 1272–1317.
4. Gámez J, Mateo J, Puerta J (2011) Learning bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood. Data Min Knowl Discov 22(1-2):106–148.
5. Kim, S. (2006) Properties of Markovian subgraphs of a decomposable graph, *MICAI 2006, Lecture Notes in Artificial Intelligence, LNAI* 4293. Advances in Artificial Intelligence. Alexander Gelbukh and Carlos Alberto Reyes-Garcia (Eds.), pp. 15-26.

Selected References (continued)

6. Kim, G. and Kim, S. (2020) Marginal information for structure learning, *Statistics and Computing* 30(2): 331-349.
7. Kim, S. (2006) Properties of markovian subgraphs of a decomposable graph. In: Alexander Gelbukh and Carlos Alberto Reyes-Garcia (eds) MICAI 2006, Lecture Notes in Artificial Intelligence, LNAI 4293 Advances in Artificial Intelligence pp 15–26.
8. Massa, M.S. and Lauritzen, S.L. (2010) Combining statistical models, *Contemporary Mathematics* 516: 239-259.
9. Tsamardinos I, Brown L, Aliferis C (2006) The max-min hill-climbing bayesian network structure learning algorithm. Mach Learn 65(1):31–78.
10. Tsamardinos I, Triantafillou S, Lagani V (2012) Towards integrative causal analysis of heterogeneous data sets and studies. J Mach Learn Res 13:1097–1157.

Thanks a lot for Your Attention!