Learning Bayesian network classifiers with completed partially directed acyclic graphs

Bojan Mihaljević, Concha Bielza and Pedro Larrañaga

9th International Conference on Probabilistic Graphical Models
Prague, Czech Republic

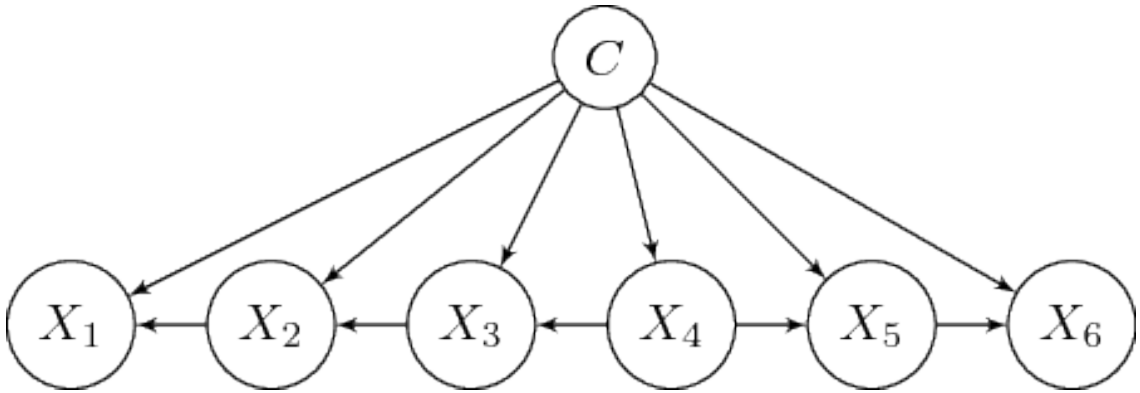
Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid

September 6th, 2018

Table of Contents

Bayesian network classifiers

- Interested in learning $P(c \mid \mathbf{x})$ rather than $P(c, \mathbf{x})$
- Differences with respect to learning $P(c, \mathbf{x})$
- Augmented naive Bayes (ANB) DAG subspace
- Discriminative scores

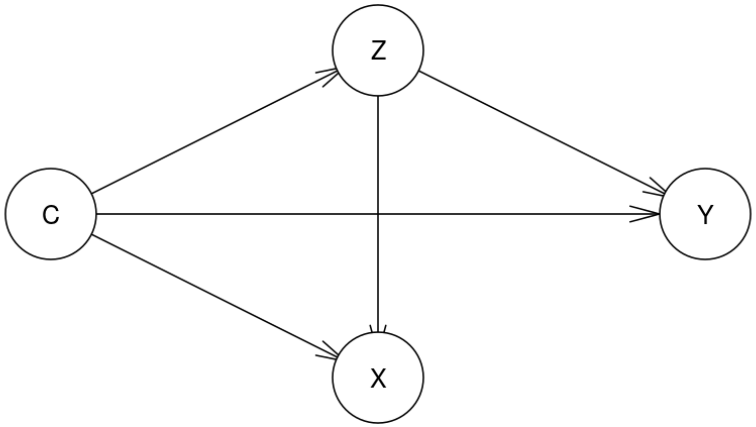


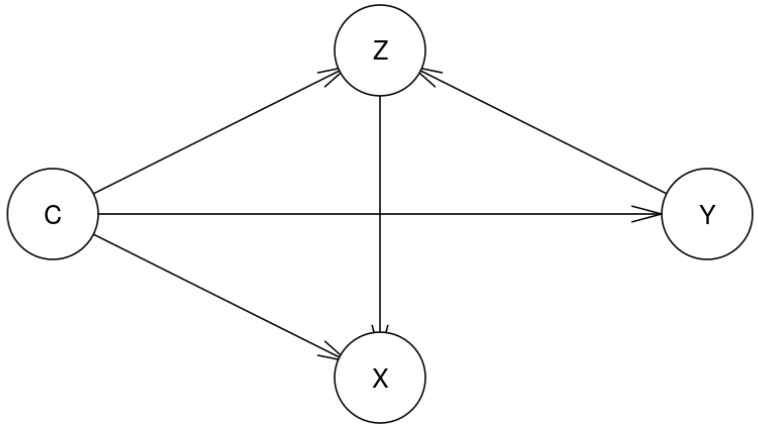
$$p(c, \mathbf{x}) = p(c)p(x_1|c, x_2)p(x_2|c, x_3)p(x_3|c, x_4)p(x_4|c)p(x_5|c, x_4)p(x_6|c, x_5)$$

- We assume data is complete

Search + score in ANB DAG space

- Start from an naive Bayes DAG
- Add arcs among features until stopping criterion





- Candidate DAGs can be encode identical $P(c, \mathbf{x})$

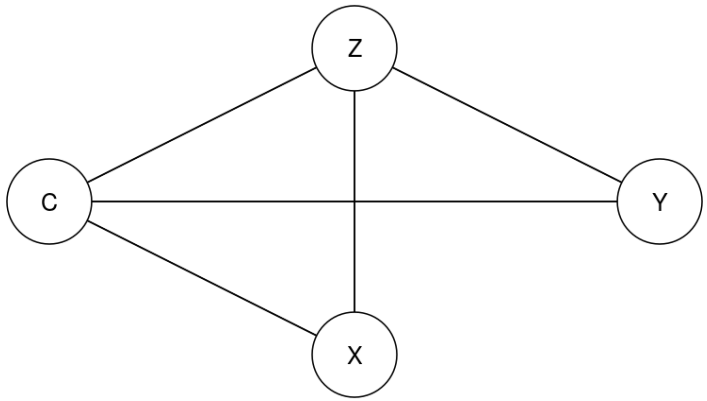
Not accounting for equivalence

- Inefficient: repeated scoring
- Sub-optimal: arbitrary arc direction lead to poorer local optimum

Structure learning: DAG equivalence classes

DAG equivalence class

- Set of DAGs with identical conditional independences
- CPDAG as a canonical representative



GES algorithm

- CPDAGs as search states
- Local updates of decomposable scores
- Greedy, optimal in the large-sample limit
- All possible arcs additions (removals) to each DAG in current class

Bayesian network classifiers with equivalence classes

State-of-the-art

- Proposed by Acid et al. (2005)
- Competitive with NB, TAN, BAN, ...
- No follow-up studies

Can be improved

- Does not use CPDAGs but a non-canonical representation
- Search space not minimal

Our contributions

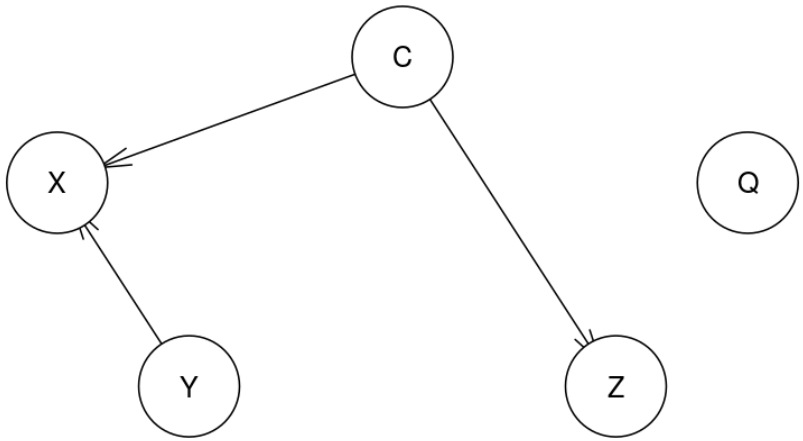
1. Specify the minimal DAG space that suffices with complete data
2. Adapt the GES algorithm to traverse this space
3. Specify how to locally update a discriminative score for our algorithm

Table of Contents

MC-DAGs

Minimal class-focused DAG

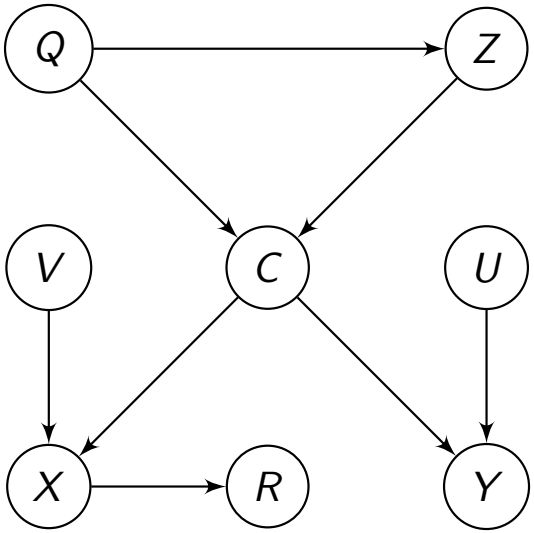
- $X \rightarrow Y$ is in $\mathcal{G}$ if:
  - $X = C$ or
  - $C \rightarrow Y$ in $\mathcal{G}$
- C is a root node



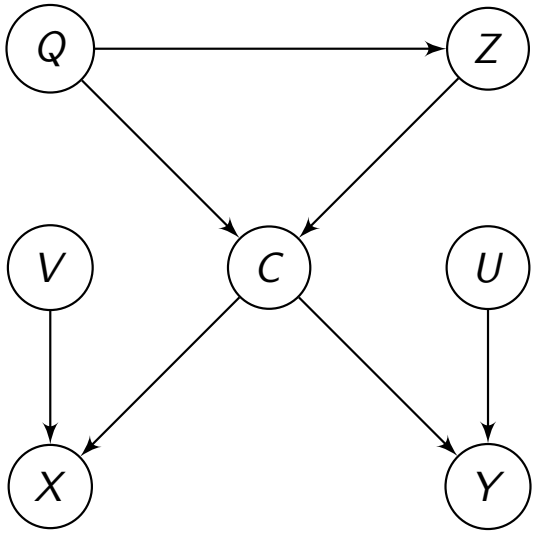
MC-DAG versus ANB

- Roots other than C
- Nodes disconnected from C

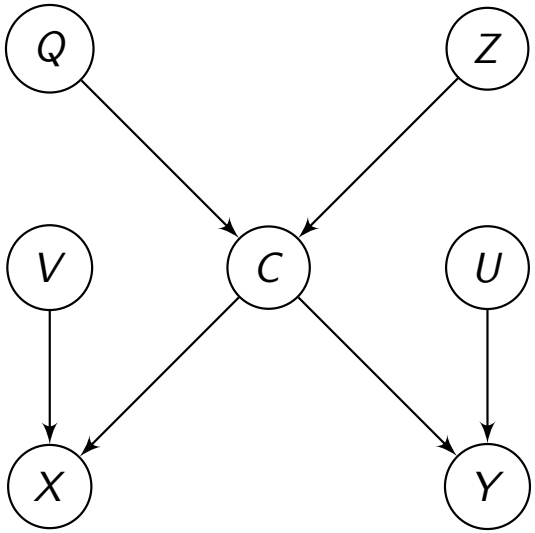
C-DAGs (Acid et al., 2005) cover all class posteriors



$\mathcal{H}$



$\mathcal{G}$: MC(C) subgraph

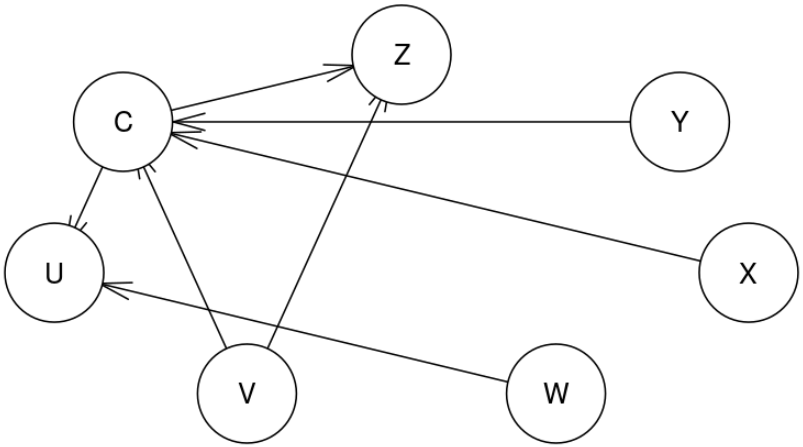


$\mathcal{G}'$: C-DAG subgraph $= \mathcal{G}$ minus arcs among $\mathbf{Pa}_{\mathcal{G}}(C)$

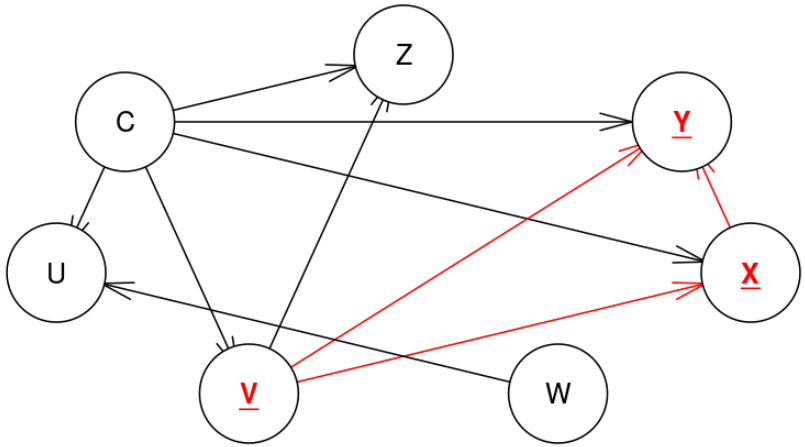
- Any $\mathcal{H}$, $\mathcal{G}$, and $\mathcal{G}'$ are classification-equivalent $P_{\mathcal{H}}(C \mid \mathbf{x}) = P_{\mathcal{G}}(C \mid \mathbf{x}) = P_{\mathcal{G}'}(C \mid \mathbf{x})\ \forall \mathbf{x}$
- C-DAG space encodes $P(c \mid \mathbf{x})$ for any possible DAG

An MC-DAG for any C-DAG

- For any C-DAG $\mathcal{H}$ there is a classification-equivalent MC-DAG $\mathcal{G}$
  - $P_{\mathcal{H}}(C \mid \mathbf{x}) = P_{\mathcal{G}}(C \mid \mathbf{x})\ \forall \mathbf{x}$



$\mathcal{H}$

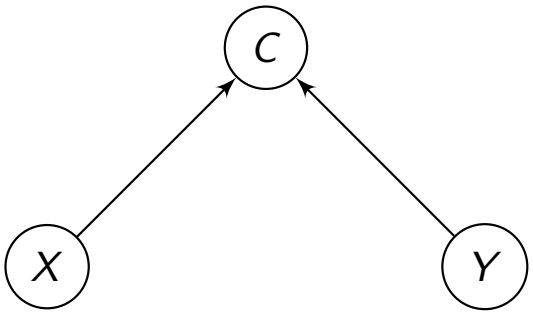


$\mathcal{G}$

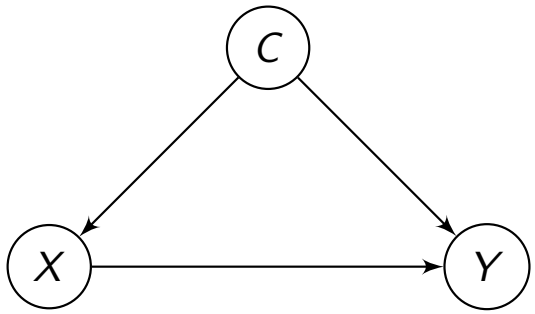
- $\mathcal{H}$ and $\mathcal{G}$ such that:
  - $\mathbf{Pa}_{\mathcal{G}}(C) = \emptyset$
  - $X \in \mathbf{Pa}_{\mathcal{H}}(C)$ in $\mathbf{Ch}_{\mathcal{G}}(C)$
  - An arc in $\mathcal{G}$ for every pair $\{X, Y\}$, $\forall X, Y \in \mathbf{Pa}_{\mathcal{H}}(C)$
- Thus, MC-DAGs encodes $P(c \mid \mathbf{x})$ for any DAG

Only independences conditional on C affect $P(c \mid \mathbf{x})$

$P_{\mathcal{H}}(C \mid \mathbf{x}) = P_{\mathcal{G}}(C \mid \mathbf{x})\ \forall \mathbf{x}$
$P_{\mathcal{H}}(c, \mathbf{x}) = P_{\mathcal{G}}(c, \mathbf{x})\ \forall \mathbf{x}$, only if $\theta_{\mathcal{G}}$ such that $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \emptyset$



$\mathcal{H}$: C-DAG
$X \not\perp\!\!\!\perp_{\mathcal{H}} Y \mid C$
$X \perp\!\!\!\perp_{\mathcal{H}} Y \mid \emptyset$



$\mathcal{G}$: MC-DAG
$X \not\perp\!\!\!\perp_{\mathcal{G}} Y \mid C$
$X \not\perp\!\!\!\perp_{\mathcal{G}} Y \mid \emptyset$

- Dropping $X \perp\!\!\!\perp_{\mathcal{H}} Y \mid \emptyset$ in $\mathcal{G}$ irrelevant ...
- ... because we always condition on C: $P(c \mid \mathbf{x}) \propto P(c)P(\mathbf{x} \mid c)$

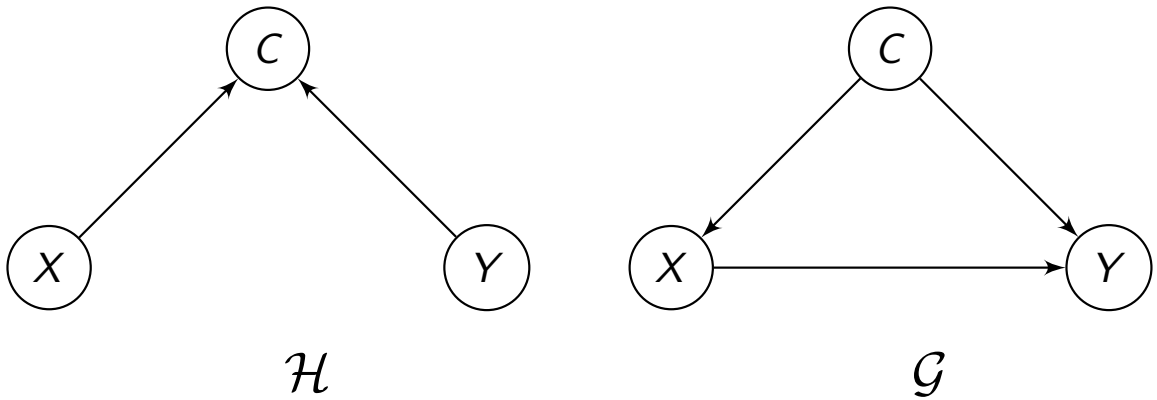
MC-DAG versus C-DAG

MC-DAG search space

- Smaller: a C-DAG is an MC-DAG, vice-versa not necessarily
- Sufficient: a C-DAG has at least one classification-equivalent MC-DAG

Additional parameters in MC-DAG

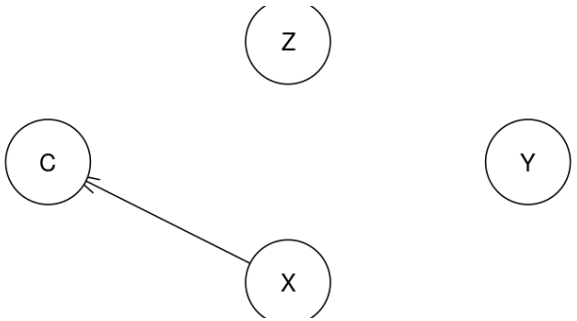
- $\mathcal{G}$ has more parameters than classification-equivalent $\mathcal{H}$
- Additional already in $\mathcal{H}$: marginalizations over $P_{\mathcal{H}}(\cdot \mid C)$
  - $P_{\mathcal{H}}(C \mid \mathbf{x}) = P_{\mathcal{G}}(C \mid \mathbf{x}) \; \forall \mathbf{x}$
- Largest new CPT as big as the CPT of C in $\mathcal{H}$



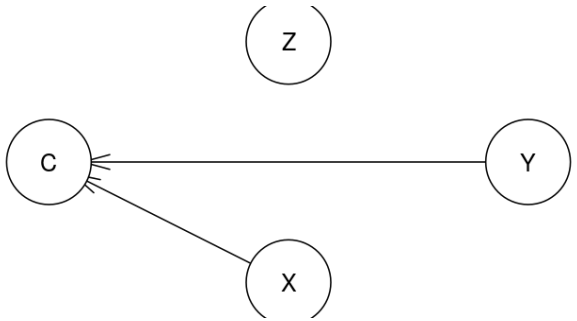
$\mathcal{H}$ $\mathcal{G}$

C-DAG search

- Allows $Y \in \mathbf{Ch}(C)$:
  - Fully dependent on all $\mathbf{Pa}(C) \setminus Y$
  - Single step to add $Y \not\perp\!\!\!\perp C$, $Y \not\perp\!\!\!\perp X \mid C$, ...

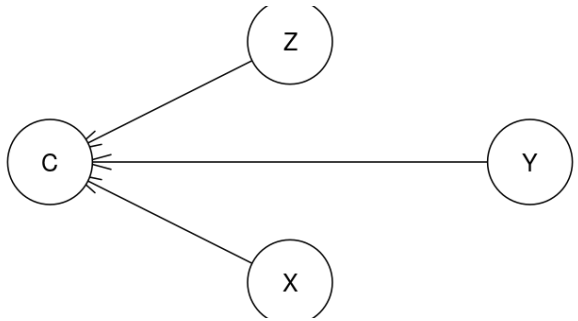


$\mathcal{G}$



$\mathcal{G}'$

$Y \not\perp\!\!\!\perp_{\mathcal{G}'} C$ and
$Y \not\perp\!\!\!\perp_{\mathcal{G}'} X \mid C$

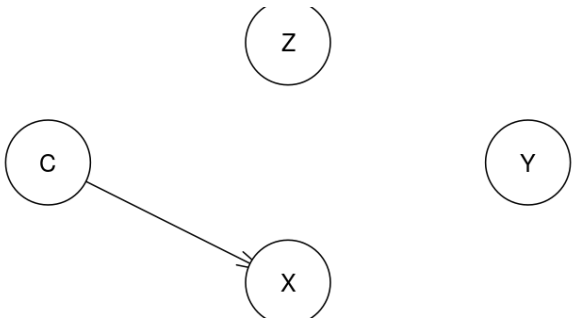


$\mathcal{G}''$

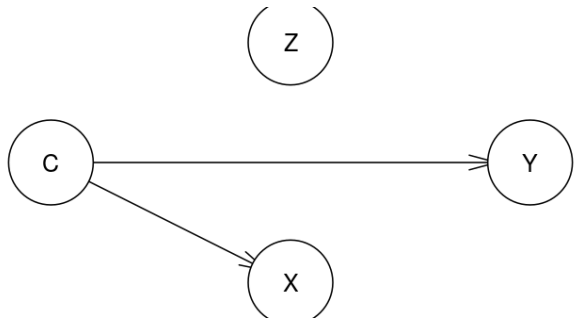
$Z \not\perp\!\!\!\perp_{\mathcal{G}''} C$ and
$Z \not\perp\!\!\!\perp_{\mathcal{G}''} X \mid C$ require
$Z \not\perp\!\!\!\perp_{\mathcal{G}''} Y \mid C$

MC-DAG search

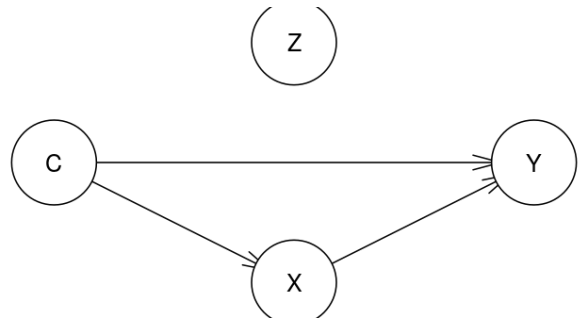
- Enforces $Y \in \mathbf{Ch}(C)$:
  - Flexible dependencies with other variables
  - Separate steps to add $Y \not\perp\!\!\!\perp_{\mathcal{G}'} C$, $Y \not\perp\!\!\!\perp_{\mathcal{G}'} X \mid C$, ...



$\mathcal{G}$



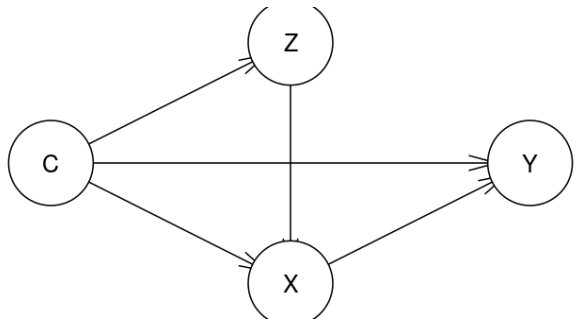
$\mathcal{G}'$



$\mathcal{G}''$

$Y \not\perp\!\!\!\perp_{\mathcal{G}'} C$

$Y \not\perp\!\!\!\perp_{\mathcal{G}'} X \mid C$



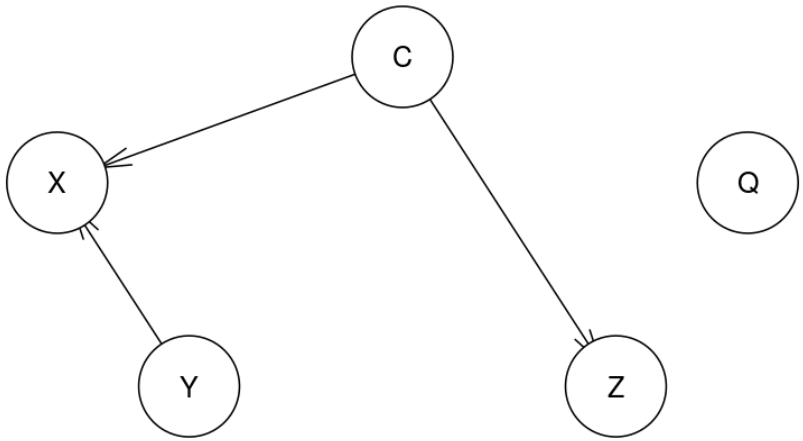
$\mathcal{G}'''$

$Z \not\perp\!\!\!\perp_{\mathcal{G}'''} C$ and $Z \not\perp\!\!\!\perp_{\mathcal{G}'''} X \mid C$ does not require $Z \not\perp\!\!\!\perp_{\mathcal{G}'''} Y \mid C$

ANB space

ANB space versus MC-DAG space

- Smaller: an ANB is an MC-DAG, vice-versa not necessarily
- Not sufficient: no classification-equivalent ANB for some MC-DAGs



Adding $C \rightarrow Q$ and/or $C \rightarrow Y$ changes $P(c \mid \mathbf{x})$

Table of Contents

GES algorithm

Search

1. CPDAG $\mathcal{P} = (\mathbf{V}, \mathbf{E}_{\mathcal{P}} = \emptyset)$
2. Forward phase:
   - Insert$(X, Y, \mathbf{T})$: considers adding all possible arcs to every DAG in $\mathcal{E}(\mathcal{P})$
   - $\mathcal{P}$ is the CPDAG of the best candidate
3. Backward phase:
   - Delete$(X, Y, \mathbf{H})$: considers removing every arc in every DAG in $\mathcal{E}(\mathcal{P})$
   - $\mathcal{P}$ is the CPDAG of the best candidate

GES algorithm for MC-DAGs

Properties

- Visits only MC-DAGs
- Simple conditions to check a PDAG for an MC-DAG extension

Search

1. CPDAG $\mathcal{P} = (\mathbf{V}, \mathbf{E}_{\mathcal{P}} = \emptyset)$
2. Forward phase:
   - Insert$(X, Y, \mathbf{T})$: considers adding all possible arcs to every DAG in $\mathcal{E}(\mathcal{P})$
   - **Discard candidates without an MC-DAG extension**
   - $\mathcal{P}$ is the CPDAG of the best candidate
3. Backward phase:
   - Delete$(X, Y, \mathbf{H})$: considers removing every arc in every DAG in $\mathcal{E}(\mathcal{P})$
   - **Discard candidates without an MC-DAG extension**
   - $\mathcal{P}$ is the CPDAG of the best candidate

GES for MC-DAGs: insert operator condition

- Apply Insert($X, Y, \mathbf{T}$) to CPDAG $\mathcal{P}$ ...

GES Insert($X, Y, \mathbf{T}$) operator

- Adds arc $X \to Y$
- Directs undirected $T - Y$ as $T \to Y$ for $T \in \mathbf{T}$

- ... to obtain PDAG $\mathcal{P}'$ ...
- Does $\mathcal{P}'$ have an MC-DAG as a consistent extension?

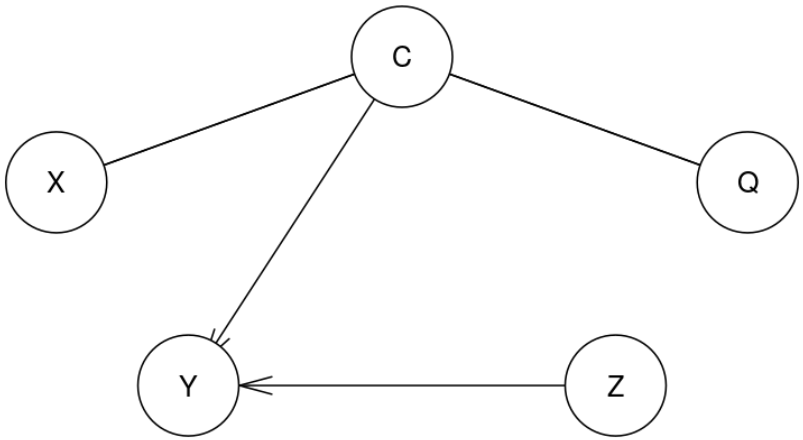
MC-DAG condition

- $Y \in \{\mathbf{Ch}_{\mathcal{P}'}(C), \mathbf{Nbr}_{\mathcal{P}'}(C)\}$

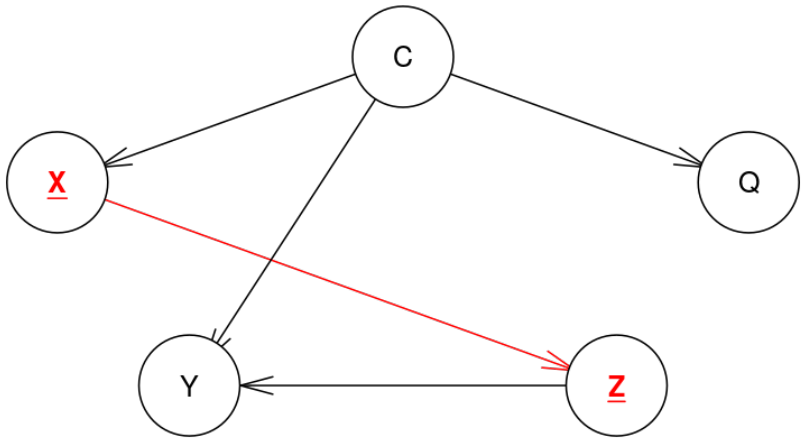
We need $Y \in \mathbf{Ch}_{\mathcal{G}'}(C)$

- If $\mathbf{Ch}_{\mathcal{P}'}(C)$, then done
- If $\mathbf{Nbr}_{\mathcal{P}'}(C)\}$, we can orient $C - Y$ as $C \to Y$

GES for MC-DAGs: insert operator condition



$\mathcal{P}$



$\mathcal{P}'$

Table of Contents

Scoring

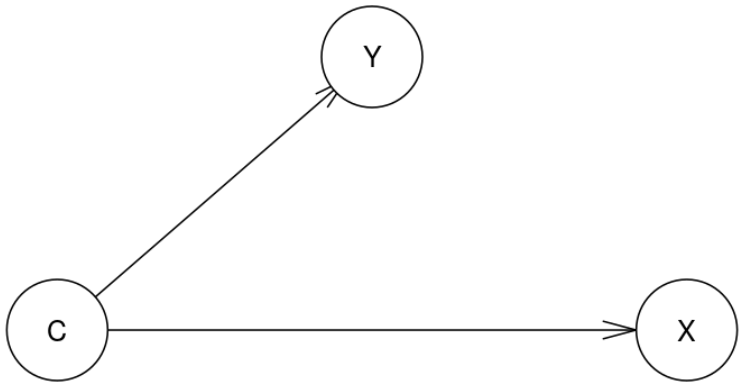
Decomposable and locally consistent scores

- Local computation for Insert($X, Y, \mathbf{T}$) and Delete($X, Y, \mathbf{H}$) specified by (Chickering, 2002)
- GES recovers generating distribution in the large sample limit

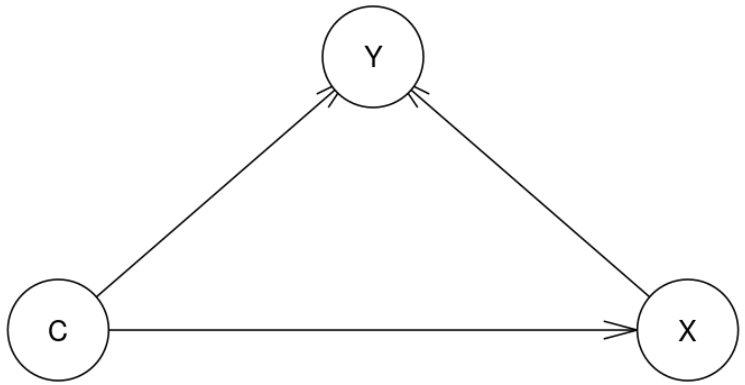
Discriminative scores

- Not decomposable
- We can update locally in time independent of n

Locally update $P(c, \mathbf{x})$ for a DAG



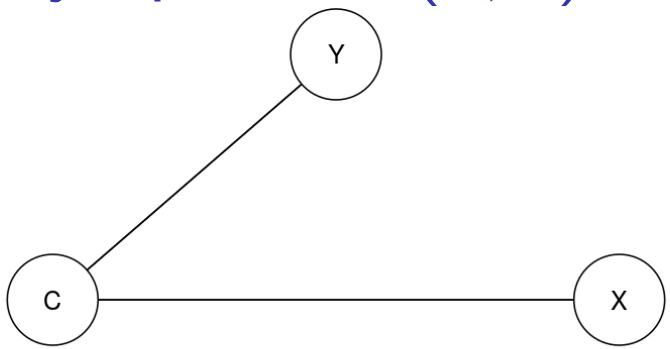
$\mathcal{G}$



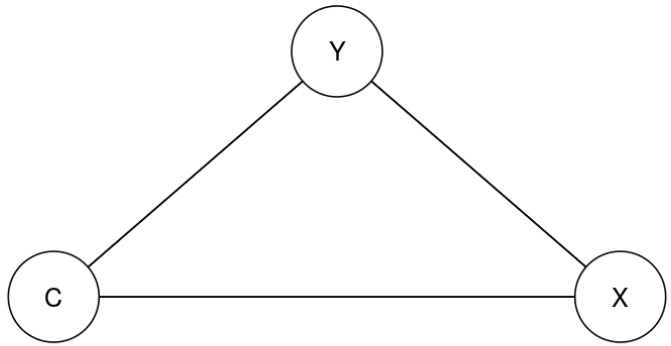
$\mathcal{G}'$

- Store $P_{\mathcal{G}}(c, \mathbf{x})$ for each each $\mathbf{x}$ in $\mathcal{D}$ and each c: $N \times r_c$ matrix
- For $\mathbf{x}$ with $X = x$ and $Y = y$:
  - $P_{\mathcal{G}'}(c_j, x, y) = P_{\mathcal{G}}(c_j, x, y) \frac{P_{\mathcal{G}'}(y|x,c_j)}{P_{\mathcal{G}}(y|c_j)}$
  - Compute for each c_j
- Estimate $P_{\mathcal{G}'}(y \mid x, c_j)$ and $P_{\mathcal{G}}(y \mid c_j)$ from $\mathcal{D}$

Locally update $P(c, \mathbf{x})$ for MC-DAG CPDAG



$\mathcal{P}$



$\mathcal{P}'$

- $\mathcal{G} \in \mathbf{cext}(\mathcal{P})$, $\mathcal{G}' = \mathcal{G}$ plus $X \to Y$ ($\mathcal{G}' \in \mathbf{cext}(\mathcal{P}')$)
- When $X \in \mathbf{Ch}(C)$

$$P_{\mathcal{G}'}(c_j, \mathbf{x}) = P_{\mathcal{G}'}(c_j, \mathbf{x}) \frac{P_{\mathcal{G}'}(y \mid x, \mathbf{t}, \mathbf{pa}^*_{\mathcal{P}}(y), c_j)}{P_{\mathcal{G}'}(y \mid \mathbf{t}, \mathbf{pa}^*_{\mathcal{P}}(y), c_j)}$$

  - $\exists\, \mathcal{G} \in \mathbf{cext}(\mathcal{P})$ s.t. $\mathbf{pa}_{\mathcal{G}}(Y) = \mathbf{pa}^*_{\mathcal{P}}(Y) \cup \mathbf{T} \cup C$
  - Because we can direct all $Y - N$, $N \in \mathbf{Nbr}_{\mathcal{P}}(Y) \setminus \mathbf{T} \setminus C$ as $Y \to N$
- When $X \notin \mathbf{Ch}(C)$: its factor can be ignored as does not depend on C

Table of Contents

Experimental evaluation

Setting

- Compare to k-dependence Bayesian classifiers and Acid et al. (2005)
- Empty graph (MG) and naive Bayes (MG_{NB}) initial state
- At most k parents per feature
- Cross-validated classification accuracy as score

Results

- Won all: car and nursery
- Won all but Acid et al. (2005): mofn-3-7-10
- Lost to TAN and Acid et al. (2005): mushroom
- No differences on other data sets
- No advantage when bounding the number of parents
  - After 10 iterations on voting: MG^1_{NB} 968 states, DB^1_{HC} 961

	MG^1_{nb}	MG^2_{nb}	MG_{nb}	MG^1	MG^2	MG	DB^1_{hc}	DB^2_{hc}	DB^1_{cl}	DB^0	A	N	n	r_c
iris	0.93	0.94	0.91	0.9	0.91	0.94	0.94	0.93	0.93	0.94	**0.95**	150	4	3
car	0.95	0.94	**0.98**	0.86	0.86	0.86	0.94	0.95	0.94	0.86	0.93	1728	6	4
nursery	0.95	0.96	**0.97**	0.9	0.9	0.9	0.95	0.96	0.93	0.9	0.94	12960	8	5
breast	**0.98**	**0.98**	0.97	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	0.96	**0.98**	**0.98**	683	9	2
tictac	0.74	0.89	**0.90**	0.71	0.78	0.75	0.71	**0.90**	0.76	0.69		958	9	2
glass	0.74	**0.75**	0.74	0.74	0.74	0.74	**0.75**	0.72	0.74	0.74	0.73	214	9	6
mofn-3-7-10	0.94	0.94	**1.00**	0.86	0.86	0.86	0.93	0.93	0.93	0.86	**1.00**	1324	10	2
wine	**0.99**	**0.99**	0.98	**0.99**	**0.99**	**0.99**	0.98	**0.99**	0.98	**0.99**		178	13	3
crx	**0.87**	**0.87**	**0.87**	0.85	0.85	0.84	0.86	0.86	0.86	0.86	**0.87**	653	15	2
voting	0.91	0.9	0.91	**0.95**	0.94	**0.95**	0.91	0.9	0.94	0.9		435	16	2
tumor	**0.48**	0.46	0.47	**0.48**	**0.48**	**0.48**	0.47	0.45	0.38	**0.48**		132	17	18
lymphography	0.84	0.84	**0.86**	0.85	0.85	0.84	0.84	0.84	0.84	0.85	0.82	148	18	4
mushroom	0.98	0.98	0.98	0.97	0.97	0.97	0.99	0.98	**1.00**	0.97	**1.00**	5643	22	2
iono	0.92	0.92	0.91	0.92	0.92	0.92	0.92	0.92	**0.93**	0.92		351	34	2
dermatology	0.98	0.98	0.98	0.98	0.98	0.98	**0.99**	0.98	0.97	0.98		358	34	6
soybean	0.91	0.92	0.92	0.91	0.91	0.91	0.91	0.92	**0.93**	0.91	0.9	562	35	19
kr-vs-kp	0.96	0.97	**0.98**	0.88	0.88	0.88	0.96	0.97	0.92	0.88	0.97	3196	36	2
molecular	0.88	0.9	**0.92**	0.91	0.91	0.91	0.9	0.9	0.81	0.91		106	57	2

Summary

Future work

- Further evaluation on synthetic and real-world data sets
  - Analyse behaviour w.r.t. sample size
- Add / complete proofs
- Different search algorithm for bounded in-degree
- Adapting to traverse the space of augmented naive Bayes

Learning Bayesian network classifiers with completed partially directed acyclic graphs

Bojan Mihaljević, Concha Bielza and Pedro Larrañaga

9th International Conference on Probabilistic Graphical Models
Prague, Czech Republic

Departamento de Inteligencia Artificial
Universidad Politécnica de Madrid

September 6th, 2018