1. Causal DAGs and intervention
2. Linear structural equation models
3. Estimation of causal effect
4. Structure learning of DAGs
Causal DAGs and intervention

(Reference: Pearl (2000) §3.1 and §3.2; Pearl (1995))

Definition: A causal model among $X_1, \ldots, X_p$ is defined by a DAG $G$ and a distribution $\mathbb{P}(\epsilon) = \mathbb{P}(\epsilon_1, \ldots, \epsilon_p)$.

- Each child-parent relationship in $G$, $(X_j, PA_j)$, represents a functional relationship (structural equation model, SEM):

$$X_j = f_j(\text{PA}_j, \epsilon_j), \quad j = 1, \ldots, p.$$

- The noise variables are jointly independent:

$$\mathbb{P}(\epsilon_1, \ldots, \epsilon_p) = \prod_j \mathbb{P}(\epsilon_j).$$

- $(1)$ and $(2)$ imply that $\mathbb{P}(X_1, \ldots, X_p)$ is Markovian with respect to the DAG $G$:

$$\mathbb{P}(X_1, \ldots, X_p) = \prod_{j=1}^{p} \mathbb{P}(X_j | \text{PA}_j).$$
Causal effect defined via external intervention:

- Consider an atomic intervention that forces $X_i$ to some fixed value $x_i$, which we denote by $do(X_i = x_i)$ or $do(x_i)$ for short.
- Effect of $do(x_i)$: to replace the SEM for $X_i$ by $X_i = x_i$ and substitute $X_i = x_i$ in the other SEMs.
- For two distinct sets of variables $X$ and $Y$, the causal effect of $X$ on $Y$ is determined by the mapping

$$x \mapsto \mathbb{P}[Y \mid do(X = x)] \equiv \mathbb{P}(Y \mid do(x)).$$

Examples of causal effects.

1. linear SEM: Causal effect $\frac{\partial \mathbb{E}(Y \mid do(x))}{\partial x}$.
2. Treatment ($X = 1$) vs control ($X = 0$): Causal effect

$$\mathbb{E}(Y \mid do(X = 1)) - \mathbb{E}(Y \mid do(X = 0)).$$
Model interventions as variables:

- Treat intervention as additional variable in the DAG: $F_j$ for intervention on $X_j$.
- SEM for $X_j$ change to

$$X_j = h_j(\text{PA}_j, F_j, \varepsilon_j) = \begin{cases} f_j(\text{PA}_j, \varepsilon_j), & \text{if } F_j = \text{idle} \\ x, & \text{if } F_j = \text{do}(x). \end{cases} \quad (4)$$

- Augment the parents of $X_j$ to $\text{PA}_j \cup \{F_j\}$:

$$\mathbb{P}(X_j = x_j \mid \text{PA}_j, F_j) = \begin{cases} \mathbb{P}(X_j = x_j \mid \text{PA}_j), & \text{if } F_j = \text{idle} \\ I(x_j = x), & \text{if } F_j = \text{do}(x), \end{cases}$$

assuming all $X_j$ are discrete for convenience.
Computing causal effect (of interventions): To simplify notation, consider discrete $X_j$ and write $\mathbb{P}(X = x) = P(x)$.

- **Truncated factorization** of $P(x_1, \ldots, x_p)$ given $do(X_i = x_i^*)$:

$$P(x_1, \ldots, x_p \mid do(x_i^*)) = I(x_i = x_i^*) \prod_{j \neq i} P(x_j \mid pa_j), \quad (5)$$

where $pa_j = (x_k : k \in PA_j)$.

- Multiple interventions $do(X_S = x^*), S \subset \{1, \ldots, p\}$:

$$P(x_1, \ldots, x_p \mid do(x^*)) = I(x_S = x^*) \prod_{j \notin S} P(x_j \mid pa_j). \quad (6)$$

- Graph structure change when $do(X_i = x_i^*)$: delete edges $X_j \rightarrow X_i$ for all $j \in PA_i$, i.e. change $G$ to $G_{\bar{X}_i}$.  


Causal DAGs and intervention

Difference between $P(y \mid do(x))$ and $P(y \mid x)$.

- Two DAGs $G_1$ and $G_2$ on $X_1, X_2$:

  - $G_1$:
    
    $P(x_1 \mid do(x_2)) = P(x_1)$,
  
  - $G_2$:
    
    $P(x_1 \mid do(x_2)) = P(x_1 \mid x_2)$.

- Find $P(x_1 \mid do(x_2))$ with respect to $G_1$ and $G_2$. 
Causal DAGs and intervention

From (5), putting $x_i = x_i^*$:

$$P(x_{-i} \mid do(x_i^*)) = \prod_{j \neq i} P(x_j \mid pa_j) \cdot \frac{P(x_i^* \mid pa_i)}{P(x_i^* \mid pa_i)}$$

$$= \frac{P(x_1, \ldots, x_p)}{P(x_i^* \mid pa_i)}$$

$$= P(x_j, j \in B \mid x_i^*, pa_i)P(pa_i),$$

(7)

where $B = [p] \setminus \{i, PA_i\}$ and $[p] := \{1, \ldots, p\}$.

- Intervention event (do-operator) not on the right-hand side.
- Compute causal effect (intervention probability) by conditional probabilities (pre-intervention probabilities) that can be estimated from observational data.
Theorem 1 (Adjustment for direct causes)

Let $PA_i$ be the parents of $X_i$ and $Y$ be any set of other variables in a causal DAG $\mathcal{G}$. Then the causal effect of $do(X_i = x_i)$ on $Y$ is given by

$$P(y \mid do(x_i)) = \sum_{pa_i} P(y \mid x_i, pa_i)P(pa_i),$$

(8)

where $P(y \mid x_i, pa_i)$ and $P(pa_i)$ are pre-intervention probabilities.

Proof.

Marginalize out $X_j \notin Y \cup \{X_i\}$ on both sides of (7).
A simple implication of Theorem 1:
If $Y$ is a set of non-descendants of $X_i$, then

$$Y \perp X_i \mid PA_i.$$ 

By Theorem 1

$$P(y \mid do(x_i)) = \sum_{pa_i} P(y \mid x_i, pa_i)P(pa_i) = \sum_{pa_i} P(y \mid pa_i)P(pa_i) = P(y),$$

which is independent of the intervention on $X_i$. Thus, $X_i$ has no causal effect on $Y$. 

Zhou, Q
A causal model \((\mathcal{G}, \mathbb{P}_\varepsilon)\) with linear SEMs:

- A linear model for each child-parent relationship:

\[
X_j = \sum_{i \in PA_j} \beta_{ij} X_i + \varepsilon_j, \quad j = 1, \ldots, p. \tag{9}
\]

- \(\varepsilon_j\)'s are independent and \(\mathbb{E}(\varepsilon_j) = 0\);
- Usually assume \(\varepsilon_j \sim \mathcal{N}(0, \omega_j^2)\). In this case, the DAG is called a Gaussian DAG and the graphical model is called a Gaussian Bayesian network.
Causal effect:

- The causal effect of $X_k$ on $X_j$

$$\gamma_{kj} := \frac{\partial \mathbb{E}(X_j \mid do(X_k = x))}{\partial x}$$

$$= \mathbb{E}(X_j \mid do(X_k = c + 1)) - \mathbb{E}(X_j \mid do(X_k = c)),$$

for any $c \in \mathbb{R}$, due to the linear model assumption.

- Using modified DAG $\mathcal{G}_{\bar{X}_k}$ after intervention,

$$\mathbb{E}(X_j \mid X_k = x; \mathcal{G}_{\bar{X}_k}) = \gamma_{kj} x,$$

where $\mathbb{E}(\bullet; \mathcal{G}_{\bar{X}_k})$ takes expectation with respect to $\mathcal{G}_{\bar{X}_k}$. 
Apply Theorem 1 to find $\gamma_{kj}$:

- Let $Z = PA_k$ and $z$ denote the value of $PA_k$,

$$p(x_j \mid do(X_k = x_k)) = \int_z p(x_j \mid x_k, z)p(z)dz,$$

where the $p$ on the right side is given by the pre-intervention distribution (that of $G$).

- Let $(\beta, \alpha)$ be the regression coefficient of $X_j$ on $(X_k, PA_k)$, that is, $\mathbb{E}(X_j \mid X_k, Z) = \beta X_k + \alpha^T Z$, which can be estimated from observational data.

- Then the causal effect

$$\gamma_{kj} = \frac{\partial}{\partial x_k} \mathbb{E}(X_j \mid do(X_k = x_k))$$

$$= \frac{\partial}{\partial x_k} \int_z \left\{ \beta x_k + \alpha^T z \right\} p(z)dz = \beta.$$
Estimation of causal effect

Reference: Pearl (2000) §3.3.

Problem setup:

- Given a causal DAG $\mathcal{G}$, if $P(y \mid do(x))$ can be uniquely computed from the (pre-intervention) distributions of observed variables in $\mathcal{G}$, then we say the causal effect of $X$ on $Y$ is identifiable.
- Note that we allow unobserved nodes in $\mathcal{G}$.
- Only observational data are collected.
Example: Observed nodes $X \rightarrow Z \rightarrow Y$; hidden node $U$, a common parent of $X$ and $Y$ (sometimes called a confounder).

Can we estimate the causal effect of $X$ on $Y$ or of $Z$ on $Y$ from observational data collected for $(X, Y, Z)$?
Estimation of causal effect

Back-door adjustment:

- Theorem 1 implies: If $X, PA_X, Y$ are observed, then $P(y \mid do(x))$ is identifiable by (8).
- Theorem 1 is a special case of back-door adjustment: $PA_X$ satisfies the back-door criterion relative to $X$ and $Y$.
- **Back-door criterion**: A set of variables $Z$ satisfies the back-door criterion relative to an ordered pair of variables $(X, Y)$ in a DAG $G$ if
  1. no nodes in $Z$ is a descendant of $X$;
  2. $Z$ blocks every path between $X$ and $Y$ that contains an arrow into $X$ (backdoor path).
Theorem 2 (Back-door adjustment)

If $Z$ satisfies the back-door criterion relative to $(X, Y)$. Then the causal effect of $X$ on $Y$ is given by

$$P(y \mid do(x)) = \sum_z P(y \mid x, z)P(z).$$ (11)

Proof.

Add intervention variable $F_X \rightarrow X$ to $\mathcal{G}$:

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

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

Invoke that $(X, Z)$ d-separates $F_X$ and $Y$. 

Zhou, Q Causal DAGs 16/46
Linear SEM: By (11), the causal effect can be identified by regressing $Y$ on $(X, Z)$:

$$\gamma_{X \rightarrow Y} := \frac{\partial}{\partial x} \mathbb{E}(Y \mid do(x)) = \beta_X(Y \sim X + Z).$$

Suppose we have data observed for the three random variables $X, Y, Z$. Then to estimate the causal effect $X$ on $Y$:

1. Discrete data: estimate $P(y \mid x, z)$ and $P(z)$ from data. Then plug into (11).

2. Linear SEM: least-squares regression $Y$ on $(X, Z)$, then

$$\hat{\gamma}_{X \rightarrow Y} = \hat{\beta}_X(Y \sim X + Z).$$
Estimation of causal effect

Example:

By Theorem 2,

\[ P(y \mid do(z)) = \sum_x P(y \mid x, z)P(x), \quad P(z \mid do(x)) = P(z \mid x), \]

without observing \( U \).
Is \( P(y \mid do(x)) \) identifiable? Yes, because:

\[
P(y \mid do(x)) = P(y \mid x; G_{\bar{X}}) \\
= \sum_{z} P(y \mid x, z; G_{\bar{X}})P(z \mid x; G_{\bar{X}}) \\
= \sum_{z} P(y \mid z; G_{\bar{X}})P(z \mid do(x)) \\
= \sum_{z} P(y \mid do(z))P(z \mid x).
\]  \hspace{1cm} (12)

Linear SEMs:

\[
\gamma_{X \rightarrow Y} = \gamma_{Z \rightarrow Y} \times \beta_X (Z \sim X) \\
= \beta_Z (Y \sim Z + X) \times \beta_X (Z \sim X).
\]
Estimation of causal effect

- Eq. (12) is an example of \textit{front-door adjustment} (Theorem 3.3.4, Pearl (2000)):
  1. $Z$ intercepts all directed paths from $X$ to $Y$;
  2. there is no back-door path from $X$ to $Z$; and
  3. all back-door paths from $Z$ to $Y$ are blocked by $X$.

Then $P(y \mid do(x))$ is identifiable

$$P(y \mid do(x)) = \sum_z P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \quad (13)$$

- Rules of do-calculus (Pearl (2000) §3.4): a set of inference rules for transforming intervention and observational probabilities, say to translate causal effect to conditional probabilities.
Estimation of causal effect

Instrumental variable formula (Bowden and Day 1984) (assume linear SEMs)

Observed nodes $Z \rightarrow X \rightarrow Y$, and $U$ is hidden common parent of $X$ and $Y$. Is $\gamma_{X \rightarrow Y} = \alpha_2$ identifiable?
Estimation of causal effect

1. $Z$ has no parents, thus $\alpha_1$ is identifiable by regressing $X$ on $Z$: $\alpha_1 = \beta_Z(X \sim Z)$.

2. Similarly, the causal effect of $Z$ on $Y$, $\alpha_1 \alpha_2$, is also identifiable: $\alpha_1 \alpha_2 = \beta_Z(Y \sim Z)$.

3. Combined we have the instrumental variable formula:

$$\alpha_2 = \frac{\beta_Z(Y \sim Z)}{\beta_Z(X \sim Z)} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}. \quad (14)$$
Estimation of causal effect

Two-stage least-squares:

1. Regress $X$ on $Z$ so $\alpha_1 = \beta_Z(X \sim Z)$ and let $\hat{X} = \alpha_1 Z$.

2. Regress $Y$ on $\hat{X}$ and then $\alpha_2 = \beta_{\hat{X}}(Y \sim \hat{X})$:

$$
\beta_{\hat{X}}(Y \sim \hat{X}) = \frac{\text{Cov}(Y, \alpha_1 Z)}{\text{Var}(\alpha_1 Z)} = \frac{\text{Cov}(Y, Z)}{\alpha_1 \text{Var}(Z)} = \alpha_2.
$$

Note: To estimate $\alpha_2$ from samples of $(X, Y, Z)$, $\beta \rightarrow \text{LSE } \hat{\beta}$. 
Structure learning: Given data $x_i = (x_{i1}, \ldots, x_{ip}) \sim (G, \mathbb{P})$ (causal model), $i = 1, \ldots, n$, how to estimate the DAG $G$?

- Constraint-based methods: Conditional independence tests against $X_i \perp X_j \mid X_S$ for all $i, j, S$.
- Score-based methods: Optimizing a scoring function over graph space.

See, e.g. Aragam and Zhou (2015) Section 1.2 for recent literature.

Data types:

- Observational data (no intervention)
- Experimental data (intervention available)
Structure learning of DAGs

Assumption: \( \mathbb{P}(X_1, \ldots, X_p) \) is faithful wrt \( \mathcal{G} \):

\[
X_A \perp X_B \mid X_S \iff S \text{ separates (d-separates)} A \text{ and } B.
\]

Definition 1

For a graphical model \((\mathcal{G}, \mathbb{P})\), we say the distribution \( \mathbb{P} \) is faithful to the graph \( \mathcal{G} \) if for every triple of disjoint sets \( A, B, S \subset V \),

\[
X_A \perp X_B \mid X_S \iff S \text{ separates (d-separates)} A \text{ and } B.
\]

- Conditional independence (CI) in \( \mathbb{P} \) \( \iff \) d-separation in \( \mathcal{G} \), i.e.

\[
I_{\mathbb{P}}(A, B \mid S) \iff D_{\mathcal{G}}(A, B \mid S).
\]

- Given \( \mathcal{G} \), almost all parameter values in the SEMs will define a faithful \( \mathbb{P} \).

- Structure learning: use CI relations learned from data to infer edges in \( \mathcal{G} \).
Structure learning of DAGs

Suppose we only have observational data. What can be learned?

**Definition 2 (Markov equivalence)**

Two DAGs $\mathcal{G}$ and $\mathcal{G}'$ on the same set of nodes $V$ are Markov equivalent if $D_\mathcal{G}(X, Y|Z) \iff D_{\mathcal{G}'}(X, Y|Z)$ for any $X, Y \in V$ and $Z \subseteq V \setminus \{X, Y\}$.

- Two DAGs are Markov equivalent if and only if they have the same skeletons and the same $v$-structures.
- A $v$-structure is a triplet $\{i, j, k\} \subseteq V$ of the form $i \rightarrow k \leftarrow j$: $i$ and $j$ are nonadjacent; $k$ is called an *uncovered collider*.
- Equivalent DAGs form an equivalence class.
- DAGs in the same equivalence class cannot be distinguished from observational data. Thus we can only learn the equivalence class of $\mathcal{G}$ from observational data.
How to represent an equivalence class? CPDAG (Completed partially DAG).

Two types of edges in a DAG $\mathcal{G}$:
- A directed edge $i \rightarrow j$ is \textit{compelled} in $\mathcal{G}$ if for every DAG $\mathcal{G}'$ equivalent to $\mathcal{G}$, the edge $i \rightarrow j$ exists in $\mathcal{G}'$.
- If an edge is not compelled in $\mathcal{G}$, then it is \textit{reversible}.

**Definition 3 (CPDAG)**

The CPDAG of an equivalence class is the PDAG consisting of a directed edge for every compelled edge in the equivalence class, and an undirected edge for every reversible edge in the equivalence class.

**Examples:**
Theorem 3 (Spirtes et al. (1993))

Suppose \((G, \mathbb{P})\) satisfies the faithfulness assumption. Then there is no edge between a pair of nodes \(X, Y \in V\) if and only if there exists a subset \(Z \subseteq V \setminus \{X, Y\}\) such that \(I_P(X, Y | Z)\).

Constraint-based methods:

1. Find the skeleton of \(G\) by CI tests;
2. Identify \(v\)-structures;
3. Orient other edges.

Output: CPDAG (or PDAG)
Outline of PC algorithm (Spirtes and Glymour 1991):

1. $E \leftarrow$ edge set of the complete undirected graph on $V$.
2. for $(i, j) \in E$ do
   3. Search for a subset $S_{ij}$ of either $N_i(E)$ or $N_j(E)$ such that $X_i \perp X_j | S_{ij}$. If found, $E \leftarrow E \setminus \{(i, j), (j, i)\}$ and store $S_{ij}$.
4. end for
5. Identify $v$-structures based on $E$ and $\{S_{ij}\}$.
6. Orient as many edges in $E$ as possible by Meek’s rules.

Notes:

1. Line 3: $N_i(E) = \{X_k : (i, k) \in E\}$.
2. For loop: implemented in ascending order of $|S_{ij}| = \ell$ for $\ell = 0, \ldots, \ell_{\max}$.
3. Line 1 to 4: Estimate skeleton $sk(\hat{G})$ of $G$. 

Edge orientation steps:

1. Identify $v$-structures (Line 5) given $sk(\hat{G})$:
   - For all nonadjacent pair $(i, j)$ with a common neighbor $k$, orient $i \rightarrow k \leftarrow j$ as $i \rightarrow k \leftarrow j$ if $k \notin S_{ij}$.
   - Because otherwise, $X_i \not\perp X_j \mid S_{ij}$, contradiction.
   - After this step, we obtain a PDAG.

2. Meek’s rules (Line 6): In the resulting PDAG, orient as many undirected edges as possible by repeated application of four rules (Meek 1995).
   - Basic idea: If orienting an undirected edge $i \rightarrow j$ into $i \rightarrow j$ would result in additional $v$-structures or a directed cycle, then orient it into $i \leftarrow j$. 

Structure learning of DAGs

Conditional independence tests \((H_0 : X \perp Y \mid S)\):

- **Gaussian data**: partial correlation \(\text{cor}(X, Y \mid S) = 0\).
  
  1. Sample covariance matrix \(\hat{\Sigma}\) from data columns of \((X, Y, S)\).
  2. \(\hat{\Omega} = (\omega_{ij}) \leftarrow \hat{\Sigma}^{-1}\) and \(\hat{\rho}_{XY \mid S} = -\omega_{12}/\sqrt{\omega_{11}\omega_{22}}\).
  3. Fisher z-transformation,

\[
z(X, Y \mid S) = \frac{1}{2} \log \left( \frac{1 + \hat{\rho}_{XY \mid S}}{1 - \hat{\rho}_{XY \mid S}} \right)
\]

and \(\sqrt{n - |S| - 3} \cdot z(X, Y \mid S) \mid H_0 \sim \mathcal{N}(0, 1)\).

- **Discrete data**: \(G^2\) or \(\chi^2\) test for conditional independence.

\[
G^2(X, Y; S = s) = 2 \sum_{x,y} O_{xys} \log(O_{xys}/E_{xys}),
\]

\[
G^2(X, Y; S) = \sum_{s} G^2(X, Y; S = s) \mid H_0 \sim \chi^2(|X| - 1)(|Y| - 1)|S|,
\]

\(E_{xys}\): expected counts under \(H_0\); \(O_{xys}\): observed counts.
Correctness and consistency:

Let $\hat{G}_n$ be the estimated graph by PC from a sample of size $n$ and $C$ be the CPDAG of $G$. Suppose that $P$ is faithful to $G$.

1. CI oracles (Spirtes et al. 1993; Meek 1995): If all CI tests are perfect (oracle), then $\hat{G}_n = C$.

2. Large-sample limit: When the sample size $n \to \infty$, all CI tests involved will be perfect (no type I or II error) with high probability. Then the PC algorithm estimates the CPDAG of $G$ consistently, i.e.

$$\lim_{n \to \infty} P(\hat{G}_n = C) = 1.$$
Structure learning of DAGs

Score-based methods:

\[ \hat{G} = \arg\max_{G \in \text{Space}} S(G, D). \]  \hspace{2cm} (15)

1. \( D = (x_{ij})_{n \times p} = [X_1 \mid \ldots \mid X_p] \) i.i.d. data from \((G, \mathbb{P})\).

2. \( S(G, D) \) is a scoring function: log-likelihood of \( D \) given a graph \( G \) with a penalty term on model complexity (number of edges or number of free parameters). For example,

\[ S_{\text{BIC}}(G, D) = \log p(D \mid \hat{\theta}, G) - \frac{d}{2} \log n, \]  \hspace{2cm} (16)

\( \hat{\theta} \): MLE of parameters under \( G \), \( d = \) dimension of \( \theta \).

3. Space of graph: DAG space or equivalence class (CPDAGs).
Structure learning of DAGs

BIC score for Gaussian DAGs:

- Liner SEM for data columns $X_j \in \mathbb{R}^n, j \in [p]$:
  \[
  X_j = \sum_{i \in \text{PA}_j} \beta_{ij} X_i + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}_n(0, \omega_j^2 I_n).
  \]

- Decomposable:
  \[
  S_{\text{BIC}}(G, D) = \sum_{j=1}^{p} s(X_j, \text{PA}_j^G) \\
  = \sum_{j} \log p(X_j \mid \hat{\beta}_j, \hat{\omega}_j^2, \text{PA}_j^G) - \frac{1}{2} |\text{PA}_j^G| \log n.
  \]

$(\hat{\beta}_j, \hat{\omega}_j^2)$: MLEs in Gaussian regression $X_j \sim \text{PA}_j^G$. 
Bayesian Dirichlet score for discrete DAGs (Heckerman et al. 1995):

- Multinomial distribution: \( \theta_{ijk} = \mathbb{P}(X_i = k \mid PA_i = j) \).
  Parameter for \([X_i \mid PA_i]\) is a \(q_i \times r_i\) table:

\[
\Theta_i = \left\{ \theta_{ijk} : j \in [q_i], k \in [r_i], \text{such that} \sum_{k=1}^{r_i} \theta_{ijk} = 1 \right\}.
\]

- Assume a conjugate prior over \(\Theta_i\) given \(G\)

\[
\Theta_i \mid PA_i \sim \text{Product-Dirichlet}((\alpha_{ijk})_{q_i \times r_i}) \iff \theta_{ij} = (\theta_{ij1}, \ldots, \theta_{ijr_i}) \mid PA_i \sim_{\text{ind}} \text{Dirichlet}(\alpha_{ij1}, \ldots, \alpha_{ijr_i}).
\]

Choose \(\alpha_{ijk} = \frac{\alpha}{(r_i \cdot q_i)}\).

- Assume a prior over \(G\): \(P(G) \propto \lambda^{d(G)}, \lambda \in (0, 1)\) and \(d(G) = \sum_{i=1}^{p} r_i q_i\) number of parameters.
Given \((G, D)\), how to compute the BD score: \((PA_i \equiv PA_i^G)\)

- Contingency tables: \(N_{ijk} = \#\{PA_i = j \& X_i = k\}\) in \(D\). For each node, a \(q_i \times r_i\) table: \(N_i = \{N_{ijk} : j \in [q_i], k \in [r_i]\}\).

- Marginal likelihood of \(N_{ij}\) (one row) given \(PA_i\):

\[
P(N_{ij} \mid PA_i) = \int P(N_{ij} \mid \theta_{ij})\pi(\theta_{ij} \mid PA_i) d\theta_{ij}
\]

\[
= \frac{\Gamma(\alpha/q_i)}{\Gamma(N_{ij\bullet} + \alpha/q_i)} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha/(q_i r_i))}{\Gamma(\alpha/(q_i r_i))},
\]

where \(N_{ij\bullet} = \sum_k N_{ijk}\) (row sum).

- Marginal likelihood of \(N_i\) (the whole table):

\[
P(N_i \mid PA_i) = \prod_{j=1}^{q_i} P(N_{ij} \mid PA_i).
\]
Structure learning of DAGs

Marginal likelihood of $D$ (all $p$ tables, one for each node):

$$P(D \mid G) = \prod_{i=1}^{p} P(N_i \mid PA_i).$$

Posterior distribution

$$P(G \mid D) \propto P(G)P(D \mid G)$$

$$= \prod_{i=1}^{p} \lambda^{q_ir_i} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha/q_i)}{\Gamma(N_{ij} + \alpha/q_i)} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + \alpha/(q_ir_i))}{\Gamma(\alpha/(q_ir_i))}.$$

BD score is decomposable:

$$S_{BD}(G, D) := \log P(G) + \log P(D \mid G) = \sum_{i=1}^{p} s(N_i, PA_i).$$

(18)
Properties of the scoring functions (17) and (18):

- **Score-equivalent:** For any two Markov equivalent DAGs $G_1$ and $G_2$, we have $S(G_1, D) = S(G_2, D)$.

- **Consistent (Chickering 2002):** A scoring function $S(G, \bullet)$ is *consistent* if the following two properties hold for $D_n \sim_{iid} P$:
  1. If $P \in G \setminus H$, then $\lim_n P\{S(G, D_n) > S(H, D_n)\} = 1$.
  2. If $P \in G \cap H$ and $d(G) < d(H)$, i.e. $G$ has fewer parameters, then $\lim_n P\{S(G, D_n) > S(H, D_n)\} = 1$.

Haughton (1988) established:

1. $S_{BIC}(G, \bullet)$ (16) is consistent for exponential family.
2. $S_{BD}(G, D_n) = S_{BIC}(G, D_n) + O_p(1) = O_p(n) + O_p(1)$.

Thus, both (17) and (18) are consistent scoring functions.
Consistency of score-based learning:

**Theorem 4**

Suppose $\mathbb{P}$ is faithful to $\mathcal{G}$ and $\mathbf{D}_n \sim_{iid} \mathbb{P}$. If $S(G, \bullet)$ is consistent and score-equivalent, then

$$\lim_{n \to \infty} \mathbb{P} \left\{ \text{argmax}_G S(G, \mathbf{D}_n) = C \right\} = 1,$$

where $C : = \{ G : G \simeq \mathcal{G} \}$ is the Markov equivalence class of $\mathcal{G}$. 
Continuous relaxation of the scoring function:

- Consider Gaussian DAGs for simplicity. The BIC score $S_{BIC}(G, D)$ (17) is over a discrete space and hard to optimize.
- $B = (\beta_{ij}) = [\beta_1 | \cdots | \beta_p]$ and $\Omega = \text{diag}(\omega_j^2)$.

Maximum regularized likelihood:

$$ (\hat{B}, \hat{\Omega}) = \arg\max_{B \in B, \Omega} \sum_{j=1}^{p} \log p(X_j | X \beta_j, \omega_j^2) - \lambda_n \rho(\beta_j). \quad (19) $$

1. $B$: weighted adjacency matrices of DAGs, so that $PA_j = \text{supp}(\beta_j)$ and supp$(B)$ defines a DAG $G$.
2. $\rho(\beta_j) = \sum_i \rho(|\beta_{ij}|)$: continuous function, e.g. $\ell_1$ or concave (Fu and Zhou 2013; Aragam and Zhou 2015).
3. Apply continuous function optimization, such as block-wise coordinate descent.
Structure learning of DAGs

Compare regularizers: $\ell_1$, concave, and $\ell_0$.

Black: $\ell_0$ penalty; Teal: $\ell_1$ penalty; Blue: MCP; Red, dashed: Capped-$\ell_1$ penalty.
Score-based learning with experimental data:

- If $X_i$ is under intervention, i.e. $\text{do}(X_i = x^*)$: delete edges $X_k \rightarrow X_i$ for all $k \in \text{PA}_i$.

- Let $\mathcal{O}_i$ be the row indices of the data matrix $\mathbf{D}$ for which node $X_i$ is not under intervention (i.e. observational). Replace $p(X_i \mid \text{PA}_i)$ by $p(X_{\mathcal{O}_i,i} \mid \text{PA}_{\mathcal{O}_i,i})$.

  1. Gaussian data: log-likelihood in (17) and (19) replaced by

     $$\ell(B, \Omega; \mathbf{D}) = \sum_{j=1}^p \log p(X_{\mathcal{O}_j,j} \mid X_{\mathcal{O}_j,i,j}, \omega_j^2).$$  \hspace{1cm} (20)

  2. Multinomial data: Replace $N_{ijk}$ by

     $$N_{ijk}(\mathcal{O}_i) = \#\{\text{rows} \in \mathcal{O}_i : \text{PA}_i = j \& X_i = k\}.$$
Structure learning of DAGs

Identifiability of causal DAGs:

Assumptions:

(A1) The true parameter $\Theta^*$ is faithful to $\mathcal{G}$.

(A2) The parameter for $[X_j | PA_j]$ is identifiable.

(A3) Each node $X_j$ is under intervention for $n_j \gg \sqrt{n}$ data points.

**Theorem 5 (Gu et al. (2019))**

Assume (A1), (A2) and (A3). Denote by $\ell(\Theta; D_n)$ the log-likelihood of the data $D_n$. For any $\Theta \neq \Theta^*$,

$$\lim_{n \to \infty} \mathbb{P}\{\ell(\Theta^*; D_n) > \ell(\Theta; D_n)\} = 1.$$

1 Gaussian data, $\ell(\Theta; D_n) = (20)$.

2 Discrete data, $\ell(\Theta; D_n) = \sum_{i=1}^{P} \sum_{j,k} N_{ijk}(O_i) \log \theta_{ijk}$. 


