Causal discovery: additional approaches

Charles K. Assaad, Emilie Devijver, Eric Gaussier

eric.gaussier@imag.fr

Table of content

Learning models from data: a Bayesian approach
- Fundamental concepts
- Equivalence between DAGs
- The GES algorithm

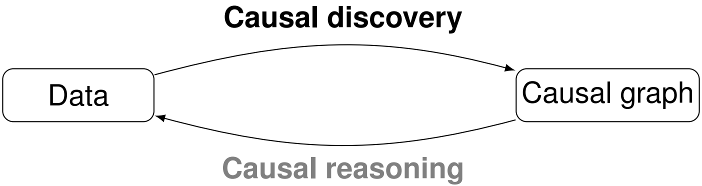
Granger causality

Table of content

Learning models from data: a Bayesian approach

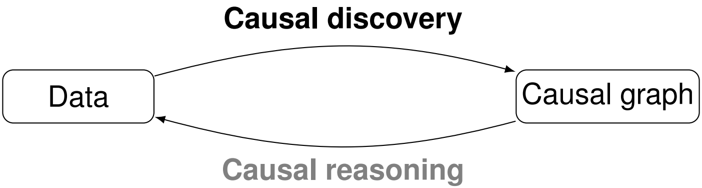
Granger causality

What is our concern?



Infer a causal graph from observed data following a Bayesian approach

What is our concern?



Infer a causal graph from observed data following a Bayesian approach

Table of content

Learning models from data: a Bayesian approach

- Fundamental concepts
- Equivalence between DAGs
- The GES algorithm

Granger causality

Bayesian network models and DAG models

Parametrized Bayesian-network model A pair $(\mathcal{G}, \boldsymbol{\theta})$ where $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a DAG in which nodes correspond to variables and $\boldsymbol{\theta}$ is a set of parameter values that specify all conditional probability distributions ($\boldsymbol{\theta}_i \subset \boldsymbol{\theta}$ subset of parameter values that define the conditional probability if X_i given its parents in $\mathcal{G}$)

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid \mathbf{Pa}_i^{\mathcal{G}} = \mathbf{pa}_i^{\mathcal{G}}, \boldsymbol{\theta}_i) \qquad (1)$$

- The structure $\mathcal{G}$ is a DAG model that represents the independence constraints that must hold in any distribution represented by the network
- The set of independence constraints imposed by $\mathcal{G}$ are represented by the Markov conditions (independence of non-descendants given parents)

Complete network

Bayesian network models and DAG models

Parametrized Bayesian-network model A pair $(\mathcal{G}, \boldsymbol{\theta})$ where $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a DAG in which nodes correspond to variables and $\boldsymbol{\theta}$ is a set of parameter values that specify all conditional probability distributions ($\boldsymbol{\theta}_i \subset \boldsymbol{\theta}$ subset of parameter values that define the conditional probability if X_i given its parents in $\mathcal{G}$)

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \,|\, \mathbf{Pa}_i^{\mathcal{G}} = \mathbf{pa}_i^{\mathcal{G}}, \boldsymbol{\theta}_i) \tag{1}$$

- The structure $\mathcal{G}$ is a DAG model that represents the independence constraints that must hold in any distribution represented by the network
- The set of independence constraints imposed by $\mathcal{G}$ are represented by the Markov conditions (independence of non-descendants given parents)

Complete network

Bayesian network models and DAG models

Parametrized Bayesian-network model A pair $(\mathcal{G}, \boldsymbol{\theta})$ where $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a DAG in which nodes correspond to variables and $\boldsymbol{\theta}$ is a set of parameter values that specify all conditional probability distributions ($\boldsymbol{\theta}_i \subset \boldsymbol{\theta}$ subset of parameter values that define the conditional probability if X_i given its parents in $\mathcal{G}$)

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid \mathbf{Pa}_i^{\mathcal{G}} = \mathbf{pa}_i^{\mathcal{G}}, \boldsymbol{\theta}_i) \qquad (1)$$

- The structure $\mathcal{G}$ is a DAG model that represents the independence constraints that must hold in any distribution represented by the network
- The set of independence constraints imposed by $\mathcal{G}$ are represented by the Markov conditions (independence of non-descendants given parents)

Complete network

Remark: Markov conditions vs faithfulness

absent edge ⇒ conditional independence

Markov conditions   Prob. distribution

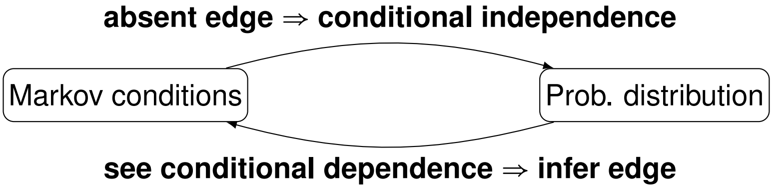
see conditional dependence ⇒ infer edge

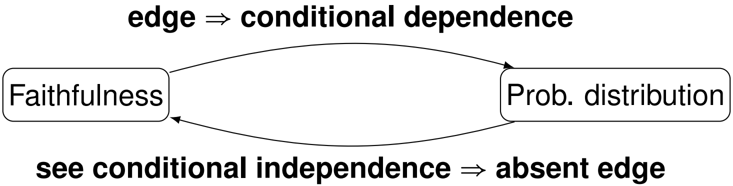
edge ⇒ conditional dependence

Faithfulness   Prob. distribution

see conditional independence ⇒ absent edge

Remark: Markov conditions vs faithfulness





Bayesian-network learning problem (1)

Learning one or more DAG models that fit a set of observed data $\mathbf{D}$ *well according to some scoring criterion* $S(\mathcal{G}, \mathbf{D})$

Hypothesis $\mathcal{G}^h$ for $\mathcal{G}$ The observed data is a set of iid samples from a distribution that contains exactly the independence constraints implied by $\mathcal{G}$

Same hypothesis

Bayesian-network learning problem (1)

Learning one or more DAG models that fit a set of observed data $\mathbf{D}$ *well according to some scoring criterion* $S(\mathcal{G}, \mathbf{D})$

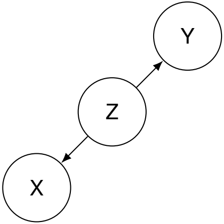
Hypothesis $\mathcal{G}^h$ for $\mathcal{G}$ The observed data is a set of iid samples from a distribution that contains exactly the independence constraints implied by $\mathcal{G}$

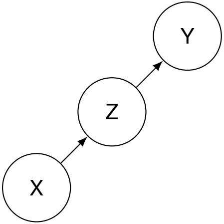
Same hypothesis

Bayesian-network learning problem (1)

Learning one or more DAG models that fit a set of observed data $\mathbf{D}$ *well according to some scoring criterion* $S(\mathcal{G}, \mathbf{D})$

Hypothesis $\mathcal{G}^h$ for $\mathcal{G}$ The observed data is a set of iid samples from a distribution that contains exactly the independence constraints implied by $\mathcal{G}$





Same hypothesis

Bayesian-network learning problem (1)

Learning one or more DAG models that fit a set of observed data $\mathbf{D}$ *well according to some scoring criterion* $S(\mathcal{G}, \mathbf{D})$

Hypothesis $\mathcal{G}^h$ for $\mathcal{G}$ The observed data is a set of iid samples from a distribution that contains exactly the independence constraints implied by $\mathcal{G}$

Same hypothesis

Bayesian-network learning problem (2)

Learning one or more DAG models that fit a set of observed data $\mathbf{D}$ *well according to some scoring criterion* $S(\mathcal{G}, \mathbf{D})$

Hypothesis $\mathcal{G}^h$ for $\mathcal{G}$ The observed data is a set of iid samples from a distribution that contains exactly the independence constraints implied by $\mathcal{G}$ (perfect map)

Perfect map We say that $\mathcal{G}$ is a *perfect map* of P if every independence constraint in P is implied by $\mathcal{G}$ and every independence implied by $\mathcal{G}$ holds in P. In this case, P is *DAG-perfect*

Assumption *Each record in* $\mathbf{D}$ *is an iid sample from a DAG-perfect probability distribution*

Bayesian-network learning problem (2)

Learning one or more DAG models that fit a set of observed data $\mathbf{D}$ *well according to some scoring criterion* $S(\mathcal{G}, \mathbf{D})$

Hypothesis $\mathcal{G}^h$ for $\mathcal{G}$ The observed data is a set of iid samples from a distribution that contains exactly the independence constraints implied by $\mathcal{G}$ (perfect map)

Perfect map We say that $\mathcal{G}$ is a *perfect map* of P if every independence constraint in P is implied by $\mathcal{G}$ and every independence implied by $\mathcal{G}$ holds in P. In this case, P is *DAG-perfect*

Assumption *Each record in* $\mathbf{D}$ *is an iid sample from a DAG-perfect probability distribution*

Bayesian-network learning problem (2)

Learning one or more DAG models that fit a set of observed data **D** *well according to some scoring criterion* $S(\mathcal{G}, \mathbf{D})$

Hypothesis $\mathcal{G}^h$ for $\mathcal{G}$ The observed data is a set of iid samples from a distribution that contains exactly the independence constraints implied by $\mathcal{G}$ (perfect map)

Perfect map We say that $\mathcal{G}$ is a *perfect map* of P if every independence constraint in P is implied by $\mathcal{G}$ and every independence implied by $\mathcal{G}$ holds in P. In this case, P is *DAG-perfect*

Assumption *Each record in* **D** *is an iid sample from a DAG-perfect probability distribution*

Bayesian-network learning problem (4)

Our goal is to infer from observed data the equivalence class of the perfect map using a scoring criterion $S(\mathcal{G}, \mathbf{D})$

Our goal is to infer from observed data the equivalence class of the perfect map using a scoring criterion $S(\mathcal{G}, \mathbf{D})$

Bayesian-network learning problem (4)

Our goal is to infer from observed data the equivalence class of the perfect map using a scoring criterion $S(\mathcal{G}, \mathbf{D})$

Our goal is to infer from observed data the equivalence class of the perfect map using a scoring criterion $S(\mathcal{G}, \mathbf{D})$

Bayesian scoring criterion

Bayesian scoring criterion: $S_B(\mathcal{G}, \mathbf{D}) = \log P(\mathcal{G}^h) + \log P(\mathbf{D} \mid \mathcal{G}^h)$

- $P(\mathcal{G}^h)$: prior probability of $\mathcal{G}^h$
- $P(\mathbf{D} \mid \mathcal{G}^h)$: marginal likelihood obtained by integrating over the unknown parameters the likelihood function (Eq. 1) applied to each record in **D** (illustration on board)

Bayesian information criterion (BIC - Schwarz, 1978) Under some assumptions:

$$S_B(\mathcal{G}, \mathbf{D}) = \underbrace{\log P(\mathbf{D} \mid \hat{\theta}, \mathcal{G}^h) - \frac{d}{2} \log m}_{BIC} + O(1)$$

$\hat{\theta}$: maximum-likelihood values of θ; d: number of free parameters; m: number of records in **D**; $O(1)$: constant

Bayesian scoring criterion

Bayesian scoring criterion: $S_B(\mathcal{G}, \mathbf{D}) = \log P(\mathcal{G}^h) + \log P(\mathbf{D} \mid \mathcal{G}^h)$

- $P(\mathcal{G}^h)$: prior probability of $\mathcal{G}^h$
- $P(\mathbf{D} \mid \mathcal{G}^h)$: marginal likelihood obtained by integrating over the unknown parameters the likelihood function (Eq. 1) applied to each record in **D** (illustration on board)

Bayesian information criterion (BIC - Schwarz, 1978) Under some assumptions:

$$S_B(\mathcal{G}, \mathbf{D}) = \underbrace{\log P(\mathbf{D} \mid \hat{\theta}, \mathcal{G}^h) - \frac{d}{2} \log m}_{BIC} + O(1)$$

$\hat{\theta}$: maximum-likelihood values of θ; d: number of free parameters; m: number of records in **D**; $O(1)$: constant

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, \mathbf{Pa}_i^{\mathcal{G}})$

Is the Bayesian scoring criterion decomposable?

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \to X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \mid \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \mid \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, \mathbf{Pa}_i^{\mathcal{G}})$

Is the Bayesian scoring criterion decomposable?

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \rightarrow X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \,|\, \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \,|\, \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, \mathbf{Pa}_i^{\mathcal{G}})$

Is the Bayesian scoring criterion decomposable?

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \to X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \,|\, \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \,|\, \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Decomposability, local consistency

Decomposability A scoring $S(\mathcal{G}, \mathbf{D})$ is decomposable if $S(\mathcal{G}, \mathbf{D}) = \sum_{i=1}^{n} s(X_i, \mathbf{Pa}_i^{\mathcal{G}})$

Is the Bayesian scoring criterion decomposable?

Local consistency Let $\mathbf{D}$ be m iid samples from distribution P, $\mathcal{G}$ be any DAG and $\mathcal{G}'$ the DAG obtained from $\mathcal{G}$ by adding the edge $X_i \to X_j$. A scoring $S(\mathcal{G}, \mathbf{D})$ is *locally consistent* if the following properties hold:

1. If $X_j \not\perp\!\!\!\perp_P X_i \mid \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) > S(\mathcal{G}, \mathbf{D})$
2. If $X_j \perp\!\!\!\perp_P X_i \mid \mathbf{Pa}_j^{\mathcal{G}}$, then $S(\mathcal{G}', \mathbf{D}) < S(\mathcal{G}, \mathbf{D})$

The Bayesian scoring criterion is locally consistent

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Implications

During the construction of graph inferred from data:

- Bayesian scoring criterion favours addition of edges that eliminate independence constraints not contained in the generative distribution
- Bayesian scoring criterion favours deletion of any unnecessary edge

Table of content

Learning models from data: a Bayesian approach

Fundamental concepts

Equivalence between DAGs

The GES algorithm

Granger causality

Markov equivalence

Theorem (Markov equivalence) Two DAGs are equivalent *iff* they have the same skeleton and the same v-structures

- Markov equivalence defines an equivalence relation (reflexive, symmetric, transitive)
- Equivalence class of $\mathcal{G}$: $\mathcal{E}(\mathcal{G})$

Covered edges An $X \to Y$ is covered in $\mathcal{G}$ if $\mathbf{Pa}(Y) = \mathbf{Pa}(X) \cup X$

Lemma (Chickering, 1995) Let $\mathcal{G}$ be a DAG and let $\mathcal{G}'$ the result of reversing the edge $X \to Y$ in $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{G}'$ are equivalent *iff* $X \to Y$ is covered in $\mathcal{G}$

Markov equivalence

Theorem (Markov equivalence) Two DAGs are equivalent *iff* they have the same skeleton and the same v-structures

- Markov equivalence defines an equivalence relation (reflexive, symmetric, transitive)
- Equivalence class of $\mathcal{G}$: $\mathcal{E}(\mathcal{G})$

Covered edges An $X \rightarrow Y$ is covered in $\mathcal{G}$ if $\mathbf{Pa}(Y) = \mathbf{Pa}(X) \cup X$

Lemma (Chickering, 1995) Let $\mathcal{G}$ be a DAG and let $\mathcal{G}'$ the result of reversing the edge $X \rightarrow Y$ in $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{G}'$ are equivalent *iff* $X \rightarrow Y$ is covered in $\mathcal{G}$

Markov equivalence

Theorem (Markov equivalence) Two DAGs are equivalent *iff* they have the same skeleton and the same v-structures

- Markov equivalence defines an equivalence relation (reflexive, symmetric, transitive)
- Equivalence class of $\mathcal{G}$: $\mathcal{E}(\mathcal{G})$

Covered edges An $X \to Y$ is covered in $\mathcal{G}$ if $\mathbf{Pa}(Y) = \mathbf{Pa}(X) \cup X$

Lemma (Chickering, 1995) Let $\mathcal{G}$ be a DAG and let $\mathcal{G}'$ the result of reversing the edge $X \to Y$ in $\mathcal{G}$. $\mathcal{G}$ and $\mathcal{G}'$ are equivalent *iff* $X \to Y$ is covered in $\mathcal{G}$

Markov equivalence (cont'd)

CPDAG: completed PDAG; PDAG: partially DAG

CPDAG of an equivalence class The CPDAG of an equivalence class consists of a directed edge for every *compelled* edge, and an undirected edge for every *reversible* edge (compelled: exists in all graphs of the equivalence class; reversible: not compelled)

What's the CPDAG of the equivalence class of the following DAG?

Markov equivalence (cont'd)

CPDAG: completed PDAG; PDAG: partially DAG

CPDAG of an equivalence class The CPDAG of an equivalence class consists of a directed edge for every *compelled* edge, and an undirected edge for every *reversible* edge (compelled: exists in all graphs of the equivalence class; reversible: not compelled)

What's the CPDAG of the equivalence class of the following DAG?

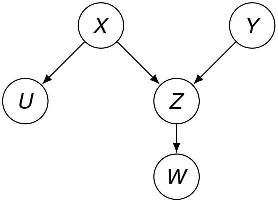


Table of content

Learning models from data: a Bayesian approach
Fundamental concepts
Equivalence between DAGs
The GES algorithm

Granger causality

Fundamental result and neighbour classes

Remark If $\mathcal{G}$ and $\mathcal{H}$ are in the same equivalence class, then $\mathcal{G}^h = \mathcal{H}^h$ and $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{H}, \mathbf{D}) := S_B(\mathcal{E}(\mathcal{G}), \mathbf{D})$

Proposition Let $\mathcal{E}^*$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of records in $\mathbf{D}$. Then in the limit of large m, $S_B(\mathcal{E}^*, \mathbf{D}) > S_B(\mathcal{E}, \mathbf{D})$ for $\mathcal{E} \neq \mathcal{E}^*$

Neighbour classes $\mathcal{E}' \in \mathcal{E}^+(\mathcal{E})$ iff one can transform any DAG $\mathcal{G}$ in $\mathcal{E}$ to any DAG $\mathcal{G}'$ in $\mathcal{E}'$ through a sequence of covered edge reversals followed by a single edge addition followed by a sequence of covered edge reversals (same definition for $\mathcal{E}^-(\mathcal{E})$ with a single edge deletion)

Fundamental result and neighbour classes

Remark If $\mathcal{G}$ and $\mathcal{H}$ are in the same equivalence class, then $\mathcal{G}^h = \mathcal{H}^h$ and $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{H}, \mathbf{D}) := S_B(\mathcal{E}(\mathcal{G}), \mathbf{D})$

Proposition Let $\mathcal{E}^*$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of records in **D.** Then in the limit of large m, $S_B(\mathcal{E}^*, \mathbf{D}) > S_B(\mathcal{E}, \mathbf{D})$ for $\mathcal{E} \neq \mathcal{E}^*$

Neighbour classes $\mathcal{E}' \in \mathcal{E}^+(\mathcal{E})$ iff one can transform any DAG $\mathcal{G}$ in $\mathcal{E}$ to any DAG $\mathcal{G}'$ in $\mathcal{E}'$ through a sequence of covered edge reversals followed by a single edge addition followed by a sequence of covered edge reversals (same definition for $\mathcal{E}^-(\mathcal{E})$ with a single edge deletion)

Fundamental result and neighbour classes

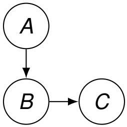
Remark If $\mathcal{G}$ and $\mathcal{H}$ are in the same equivalence class, then $\mathcal{G}^h = \mathcal{H}^h$ and $S_B(\mathcal{G}, \mathbf{D}) = S_B(\mathcal{H}, \mathbf{D}) := S_B(\mathcal{E}(\mathcal{G}), \mathbf{D})$

Proposition Let $\mathcal{E}^*$ denote the equivalence class that is a perfect map of distribution P, and let m be the number of records in **D.** Then in the limit of large m, $S_B(\mathcal{E}^*, \mathbf{D}) > S_B(\mathcal{E}, \mathbf{D})$ for $\mathcal{E} \neq \mathcal{E}^*$

Neighbour classes $\mathcal{E}' \in \boldsymbol{\mathcal{E}}^+(\mathcal{E})$ iff one can transform any DAG $\mathcal{G}$ in $\mathcal{E}$ to any DAG $\mathcal{G}'$ in $\mathcal{E}'$ through a sequence of covered edge reversals followed by a single edge addition followed by a sequence of covered edge reversals (same definition for $\boldsymbol{\mathcal{E}}^-(\mathcal{E})$ with a single edge deletion)

Example

What are the equivalence class $\mathcal{E} = \mathcal{E}(\mathcal{G})$, $\boldsymbol{\mathcal{E}}^{+}(\mathcal{E})$ and $\boldsymbol{\mathcal{E}}^{-}(\mathcal{E})$ of the following graph $\mathcal{G}$?



GES: greedy equivalence search

GES algorithm

1. Initialisation: set $\mathcal{E}$ to the equivalence class corresponding to the DAG with no edge
2. Repeatedly replace $\mathcal{E}$ with the member of $\boldsymbol{\mathcal{E}}^{+}(\mathcal{E})$ that has the highest score, until no such replacement increases the score
3. Repeatedly replace $\mathcal{E}$ with the member of $\boldsymbol{\mathcal{E}}^{-}(\mathcal{E})$ that has the highest score, until no such replacement increases the score
4. Output the current class $\mathcal{E}$

Consistency of GES Let $\mathcal{E}$ denote the equivalence class that results from GES, let P denote the DAG-perfect distribution associated with **D**, and let m denote the number of records in **D**. Then in the limit of large m, $\mathcal{E}$ is a perfect map of P

GES: greedy equivalence search

GES algorithm

1. Initialisation: set $\mathcal{E}$ to the equivalence class corresponding to the DAG with no edge
2. Repeatedly replace $\mathcal{E}$ with the member of $\boldsymbol{\mathcal{E}}^{+}(\mathcal{E})$ that has the highest score, until no such replacement increases the score
3. Repeatedly replace $\mathcal{E}$ with the member of $\boldsymbol{\mathcal{E}}^{-}(\mathcal{E})$ that has the highest score, until no such replacement increases the score
4. Output the current class $\mathcal{E}$

Consistency of GES Let $\mathcal{E}$ denote the equivalence class that results from GES, let P denote the DAG-perfect distribution associated with **D**, and let m denote the number of records in **D**. Then in the limit of large m, $\mathcal{E}$ is a perfect map of P

Remarks

1. Well-founded algorithm with consistency proof first established by Meek (Meek, 1997) based on a conjecture proven by Chickering (Chickering, 2002)
2. Main disadvantage: computational complexity
   - Learning optimal structure with Bayesian scoring criterion is NP-hard (Chickering, 1996)
   - Fast implementations exist when the underlying graph is sparse (Chickering, 2020)
3. Another (faster) approach exists based on the EM (expectation-maximisation) algorithm called MS-EM for *model selection EM* described in (Friedman, 1997)
4. Several other extensions for different data types, *e.g.* for time series (Assaad *et al.*, 2022)

Table of content

Learning models from data: a Bayesian approach

Granger causality

Causality according to Granger

Granger causality A time series X^p Granger-causes X^q if past values of X^p provide unique, statistically significant information about future values of X^q

Standard pariwise version Under the assumption of stationary linear systems, one considers the following autoregression models:

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \xi_t^q \qquad \text{(Mres)}$$

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \sum_{i=1}^{\tau} a_{p,i} X_{t-i}^p + \xi_t^q \qquad \text{(Mfull)}$$

ξ_t^q are uncorrelated rand. var. with 0 mean, a are real coefficients, and τ optimal lag

Causality according to Granger

Granger causality A time series X^p Granger-causes X^q if past values of X^p provide unique, statistically significant information about future values of X^q

Standard pariwise version Under the assumption of stationary linear systems, one considers the following autoregression models:

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \xi_t^q \qquad \text{(Mres)}$$

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \sum_{i=1}^{\tau} a_{p,i} X_{t-i}^p + \xi_t^q \qquad \text{(Mfull)}$$

ξ_t^q are uncorrelated rand. var. with 0 mean, a are real coefficients, and τ optimal lag

Causality according to Granger (cont'd)

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \xi_t^q \qquad \text{(Mres)}$$

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \sum_{i=1}^{\tau} a_{p,i} X_{t-i}^p + \xi_t^q \qquad \text{(Mfull)}$$

If the full model is significantly more accurate than the restricted model, one concludes that X^p Granger-causes X^q

Remarks

- Statistical test such as the F-test can be used to determine whether the full model is significantly better than the restricted one (null hypothesis: X^p does not Granger-cause X^q)
- Optimal lag τ estimated using an information criterion, as AIC (Akaike information criterion) or BIC

Causality according to Granger (cont'd)

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \xi_t^q \quad \text{(Mres)}$$

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \sum_{i=1}^{\tau} a_{p,i} X_{t-i}^p + \xi_t^q \quad \text{(Mfull)}$$

If the full model is significantly more accurate than the restricted model, one concludes that X^p Granger-causes X^q

Remarks

- Statistical test such as the F-test can be used to determine whether the full model is significantly better than the restricted one (null hypothesis: X^p does not Granger-cause X^q)
- Optimal lag τ estimated using an information criterion, as AIC (Akaike information criterion) or BIC

Causality according to Granger (cont'd)

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \xi_t^q \qquad \text{(Mres)}$$

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau} a_{q,i} X_{t-i}^q + \sum_{i=1}^{\tau} a_{p,i} X_{t-i}^p + \xi_t^q \qquad \text{(Mfull)}$$

If the full model is significantly more accurate than the restricted model, one concludes that X^p Granger-causes X^q

Remarks

- Statistical test such as the F-test can be used to determine whether the full model is significantly better than the restricted one (null hypothesis: X^p does not Granger-cause X^q)
- Optimal lag τ estimated using an information criterion, as AIC (Akaike information criterion) or BIC

Associated algorithm

input X a d-dimensional time series, $\tau_{\max} \in \mathbb{N}$ optimal lag
initialisation Form an empty graph $\mathcal{G}$ with d nodes V
Standardize data and check if it is covariance stationary
for $X^q \in V$ **do**
  Fit Mres and compute its residuals
  for $X^p \in V \setminus \{X^q\}$ **do**
    Fit Mfull and compute its residuals
    Compare Mres and Mfull
    if null hypothesis rejected **then** add $X^p \to X^q$ to $\mathcal{G}$
return $\mathcal{G}$

Multivariate extension

$$\mathcal{X}_t^q = a_{q,0} + \sum_{\substack{r=1 \\ r \neq p}}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathcal{X}_{t-i}^p + \xi_t^q \qquad \text{(mvMres)}$$

$$\mathcal{X}_t^q = a_{q,0} + \sum_{r=1}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathcal{X}_{t-i}^r + \xi_t^q \qquad \text{(mvMfull)}$$

If the full model is significantly more accurate than the restricted model (through a statistical test), X^p Granger-causes X^q

Remarks

- Yields better results than previous version
- Computationally costly so that people mostly rely on pairwise version

Multivariate extension

$$\mathcal{X}_t^q = a_{q,0} + \sum_{\substack{r=1 \\ r \neq p}}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathcal{X}_{t-i}^p + \xi_t^q \quad \text{(mvMres)}$$

$$\mathcal{X}_t^q = a_{q,0} + \sum_{r=1}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathcal{X}_{t-i}^r + \xi_t^q \quad \text{(mvMfull)}$$

If the full model is significantly more accurate than the restricted model (through a statistical test), X^p Granger-causes X^q

Remarks

- Yields better results than previous version
- Computationally costly so that people mostly rely on pairwise version

Multivariate extension

$$\mathcal{X}_t^q = a_{q,0} + \sum_{\substack{r=1 \\ r \neq p}}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathcal{X}_{t-i}^p + \xi_t^q \tag{mvMres}$$

$$\mathcal{X}_t^q = a_{q,0} + \sum_{r=1}^{d} \sum_{i=1}^{\tau} a_{r,i} \mathcal{X}_{t-i}^r + \xi_t^q \tag{mvMfull}$$

If the full model is significantly more accurate than the restricted model (through a statistical test), X^p Granger-causes X^q

Remarks

- Yields better results than previous version
- Computationally costly so that people mostly rely on pairwise version

Other extensions

Several other extensions have been proposed (Assaad *et al.*, 2022), including

- Dealing with non-stationary processes (Luo *et al.*, 2015)
- Using deep learning to learn complex, non linear relations (Nauta *et al.*, 2019)

Granger causality is not causality: no explicit way to distinguish causal relations from spurious correlations

Other extensions

Several other extensions have been proposed (Assaad *et al.*, 2022), including

- Dealing with non-stationary processes (Luo *et al.*, 2015)
- Using deep learning to learn complex, non linear relations (Nauta *et al.*, 2019)

Granger causality is not causality: no explicit way to distinguish causal relations from spurious correlations

Conclusion

We have reviewed the major methods for causal discovery

- Constraint-based methods
- Noise-based methods
- Score-based methods
- Granger causality

Other methods exist but are less used: logic-based approaches, topology-based approaches (not true causality), difference-based approaches (Assaad *et al.*, 2022)

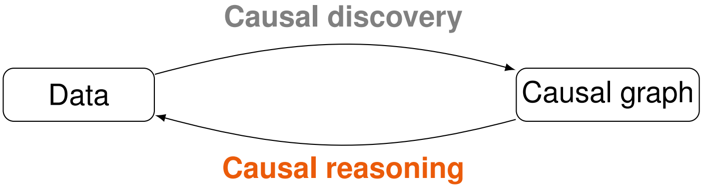
Conclusion

We have reviewed the major methods for causal discovery

- Constraint-based methods
- Noise-based methods
- Score-based methods
- Granger causality

Other methods exist but are less used: logic-based approaches, topology-based approaches (not true causality), difference-based approaches (Assaad *et al.*, 2022)

Next courses



References (1)

1. *Estimating the dimension of a model*, G. E. Schwarz, 1978
2. *A transformational characterization of Bayesian network structures*, D. M. Chickering, 1995
3. *Learning Bayesian networks is NP-complete*, D. M. Chickering, 1996
4. *Learning belief networks in the presence of missing values and hidden variables*, N. Friedman, 1997
5. *Optimal structure identification with greedy search*, D. M. Chickering, 2002
6. *Discovering causal structures from time series data via enhanced Granger causality*, L. Luo, W. Liu, I. Koprinska, F. Chen, 2015

References (2)

7. *Causal discovery with attention-based convolutional neural networks*, M. Nauta, D. Bucur, C. Seifert, 2019

8. *Statistically Efficient Greedy Equivalence Search*, D. M. Chickering, 2020

9. *Survey and evaluation of causal discovery methods for time series*, K. C. Assaad, E. Devijver, E. Gausssier, 2022