# Explaining Bayesian networks using MAP-independence: Some new properties

Enrique Valero-Leal
Pedro Larrañaga
Concha Bielza

New Bridges between
Mathemathics and Data Science

8th-11th November 2021

UNIVERSIDAD POLITÉCNICA DE MADRID

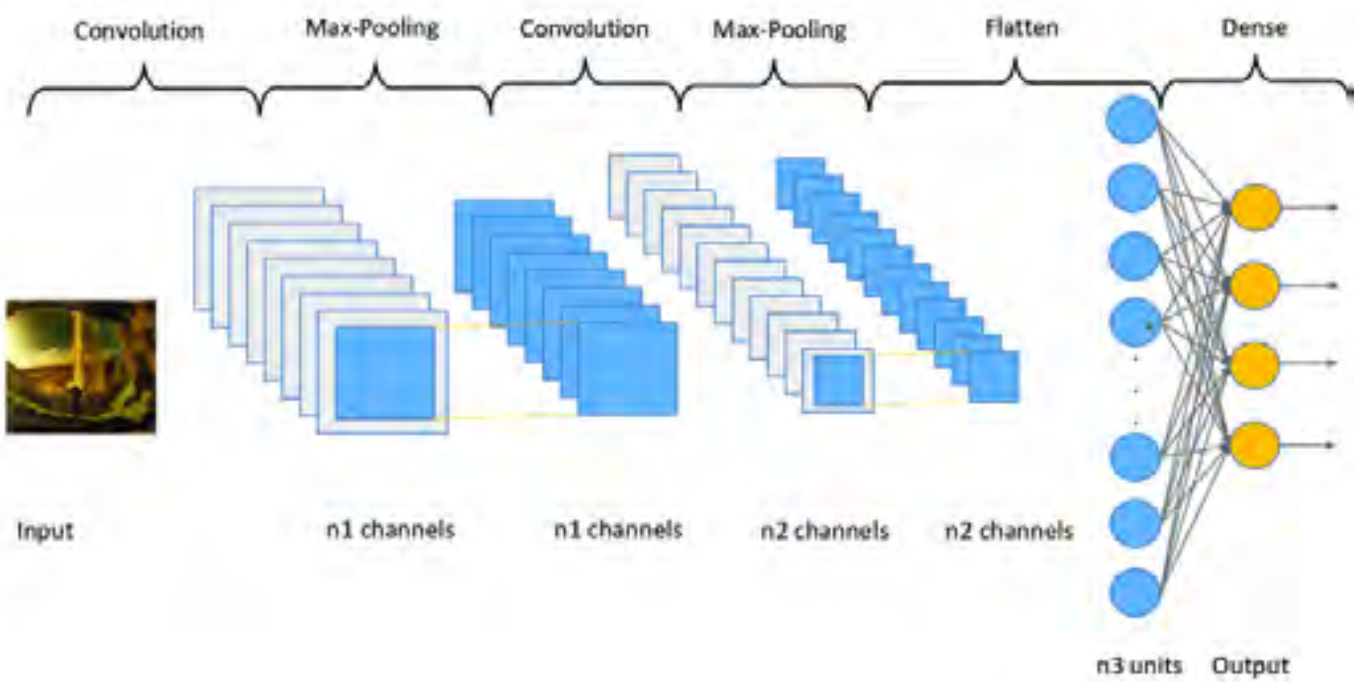
# Index

# Index

# Motivation for Explainable AI

Around 2012, the deep learning revolution took place

- A significant rise in complex and powerful methods
- A further insertion of such methods in our societies




Benítez-Andrades, via ResearchGate

# Motivation for Explainable AI

The "reasoning" of these methods is impossible to understand

- Legal, ethical and performance problems


Explainable AI deals with better interpreting AI systems

- Miller (2019): Interpretability is the degree to which a human can understand the cause of a decision

Regarding simple models are really as simple as we believe?

- Even in interpretable models, explanations might be desirable

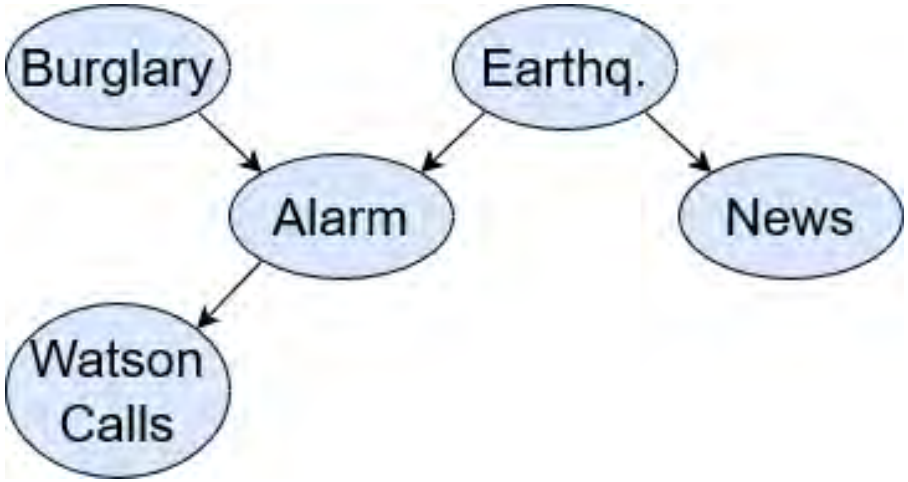
# Motivation for Bayesian networks

A model based on actual mathematical and statistical theory is desirable

- Uncertainty + many variables → intractable joint probability distribution (JPD) and inference

Bayesian networks are probabilistic graphical models that offer:

1. Modularity representing the JPD, thus compacting it
2. Easier and semantically comprehensible inference
3. Interpretable and visual relations

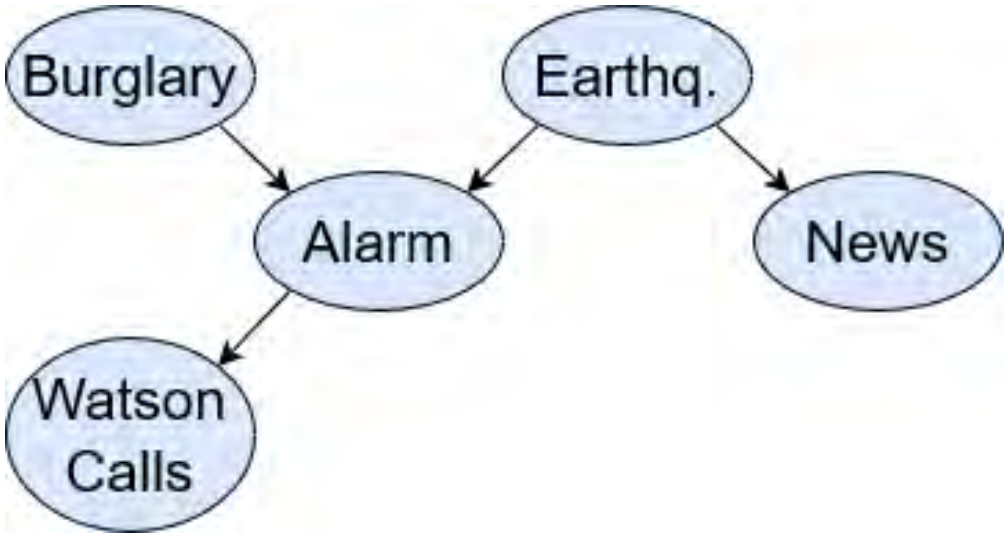


# Index

# Bayesian networks (Pearl, 1988) (Koller and Friedman, 2009)

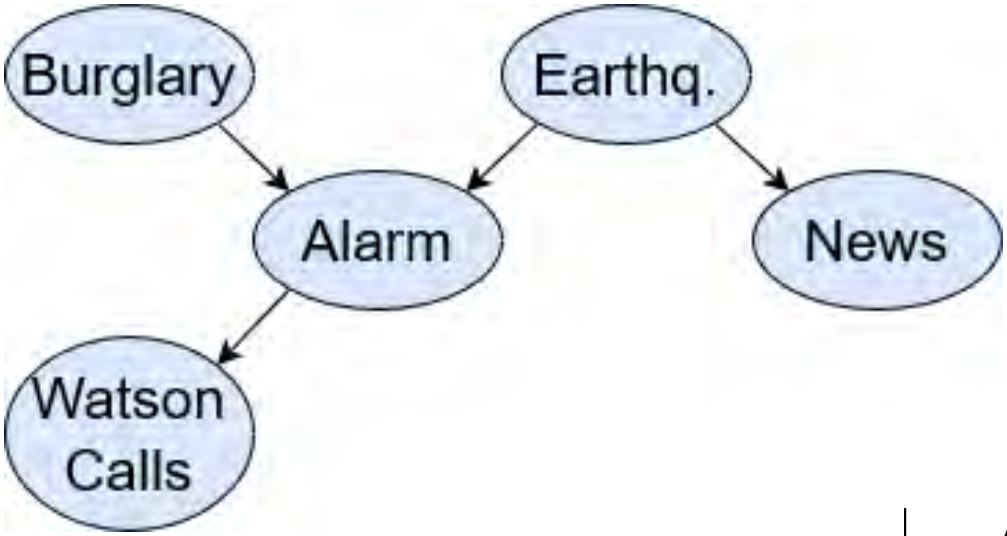
A Bayesian network $\mathcal{B} = (\mathcal{G}, \theta)$ is a probabilistic graphical model that encodes a JPD $P(X_1, ..., X_n)$ over a set of variables $\boldsymbol{X} = \{X_1, ..., X_n\}$.

Qualitatively, a Bayesian network is a directed acyclic graph $\mathcal{G} = (\boldsymbol{X}, A)$ that represents conditional (in)dependences.



Quantitatively, Bayesian networks factorise the JPD $P(X_1, ..., X_n)$ storing only the conditional probability distributions (CPD) $P(X_i|Pa_{X_i})$ of each node.

$$P(B, E, A, N, W) = P(B)P(E)P(A|B, E)P(N|E)P(W|A)$$




$P(B)$
$P(E)$
$P(A|B, E)$
$P(N|E)$
$P(W|A)$

| | News | |
|---|---|---|
| Earthq. | t | f |
| f | 0.1 | 0.9 |
| t | 0.6 | 0.4 |

| | | Alarm | |
|---|---|---|---|
| Burg. | Earthq. | t | f |
| f | f | 0.05 | 0.95 |
| f | t | 0.4 | 0.6 |
| t | f | 0.9 | 0.1 |
| t | t | 0.99 | 0.01 |

| Earthquake | |
|---|---|
| t | f |
| 0.15 | 0.85 |

# Conditional independence and d-separation

$\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are conditionally independent given $\boldsymbol{X}_3$, $I(\boldsymbol{X}_1, \boldsymbol{X}_2|\boldsymbol{X}_3)$, iff $P(\boldsymbol{X}_1, \boldsymbol{X}_2|\boldsymbol{X}_3) = P(\boldsymbol{X}_1|\boldsymbol{X}_3)P(\boldsymbol{X}_2|\boldsymbol{X}_3)$.

Many independences can be identified through **d-separation**
$\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are d-separated by $\boldsymbol{X}_3$ (denoted $\boldsymbol{X}_1 \perp \boldsymbol{X}_2|\boldsymbol{X}_3$) if in all undirected paths between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ there is a node $C$ such that:

1. $C$ and its descendants $\notin \boldsymbol{X}_3$ and $C$ is a converging connection, or
2. $C \in \boldsymbol{X}_3$ and $C$ is not a converging connection.

# Conditional independence and d-separation

$\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are conditionally independent given $\boldsymbol{X}_3$, $I(\boldsymbol{X}_1, \boldsymbol{X}_2|\boldsymbol{X}_3)$, iff $P(\boldsymbol{X}_1, \boldsymbol{X}_2|\boldsymbol{X}_3) = P(\boldsymbol{X}_1|\boldsymbol{X}_3)P(\boldsymbol{X}_2|\boldsymbol{X}_3)$.

Many independences can be identified through **d-separation**
$\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are d-separated by $\boldsymbol{X}_3$ (denoted $\boldsymbol{X}_1 \perp \boldsymbol{X}_2|\boldsymbol{X}_3$) if in all undirected paths between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ there is a node $C$ such that:

1. $C$ and its descendants $\notin \boldsymbol{X}_3$ and $C$ is a converging connection, or
2. $C \in \boldsymbol{X}_3$ and $C$ is not a converging connection.

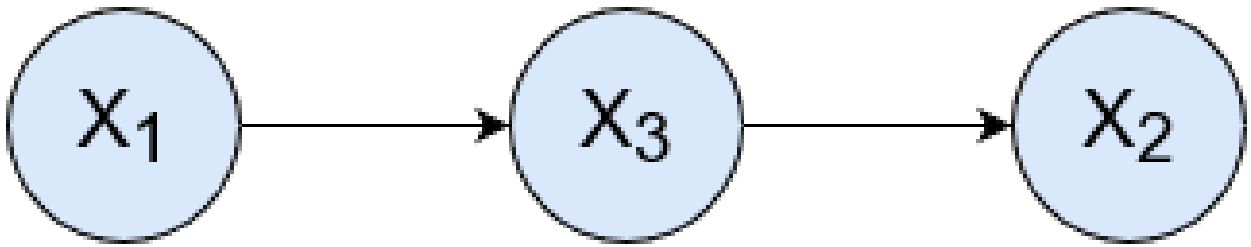
**D-separation theorem (Verma and Pearl, 1990)**

$$\boldsymbol{X}_1 \perp \boldsymbol{X}_2|\boldsymbol{X}_3 \Rightarrow I(\boldsymbol{X}_1, \boldsymbol{X}_2|\boldsymbol{X}_3) \quad \forall \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3 \subseteq \boldsymbol{X} \text{ (disjoint)}$$

## D-separation theorem (Verma and Pearl, 1990)

$$\boldsymbol{X}_1 \perp \boldsymbol{X}_2 | \boldsymbol{X}_3 \Rightarrow I(\boldsymbol{X}_1, \boldsymbol{X}_2 | \boldsymbol{X}_3) \quad \forall \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3 \subseteq \boldsymbol{X} \text{ (disjoint)}$$

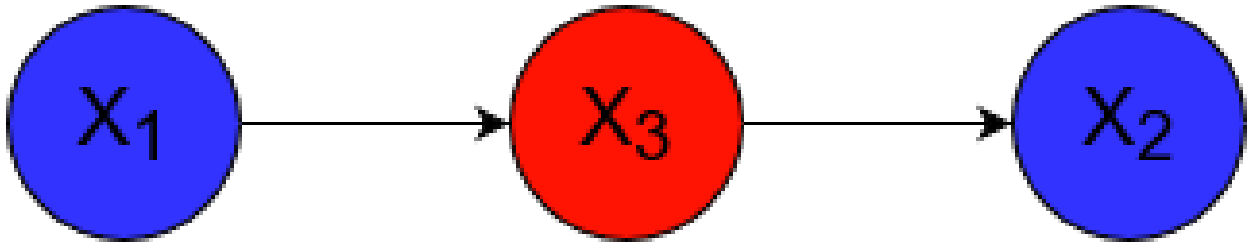
Example:



## D-separation theorem (Verma and Pearl, 1990)

$$\boldsymbol{X}_1 \perp \boldsymbol{X}_2 | \boldsymbol{X}_3 \Rightarrow I(\boldsymbol{X}_1, \boldsymbol{X}_2 | \boldsymbol{X}_3) \quad \forall \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3 \subseteq \boldsymbol{X} \text{ (disjoint)}$$

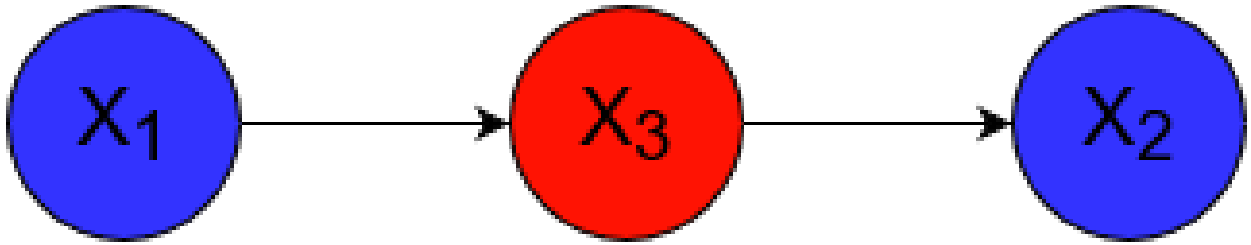
Example:



## D-separation theorem (Verma and Pearl, 1990)

$$\boldsymbol{X}_1 \perp \boldsymbol{X}_2 | \boldsymbol{X}_3 \Rightarrow I(\boldsymbol{X}_1, \boldsymbol{X}_2 | \boldsymbol{X}_3) \quad \forall \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3 \subseteq \boldsymbol{X} \text{ (disjoint)}$$

Example:

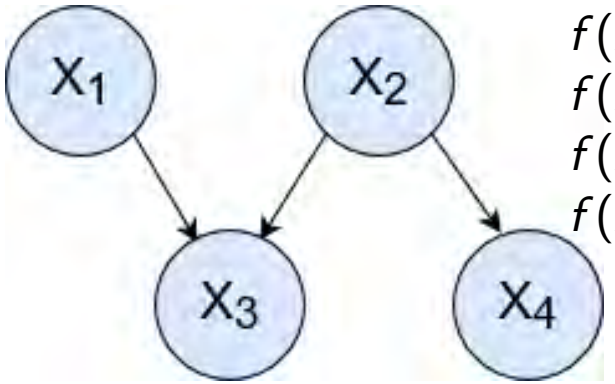



Observation: If $I(\boldsymbol{X}_1, \boldsymbol{X}_2 | \boldsymbol{X}_3)$ and $\boldsymbol{X}_3$ is not observed, a change in $\boldsymbol{X}_1$ is propagated to $\boldsymbol{X}_2$ *via* $\boldsymbol{X}_3$

# Continuous Bayesian networks (Shachter and Kenley, 1989) (Geiger and Heckerman, 1994)

In continuous Bayesian networks, all the variables are continuous.
One of the most widely used are linear Gaussian Bayesian networks, which factorise a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}; \Sigma)$ into a set of linear conditional Gaussian variables.

Conditional density of $Y$: $f(Y|Pa_Y) = \mathcal{N}(\beta_Y + \sum_{X_i}^{Pa_Y} \beta_{YX_i}(x_i - \mu_i)\ ; v_Y)$




$$f(X_1) = \mathcal{N}(\beta_1; v_1)$$
$$f(X_2) = \mathcal{N}(\beta_2; v_2)$$
$$f(X_3|X_1, X_2) = \mathcal{N}(\beta_3 + \beta_{31}(x_1 - \mu_1) + \beta_{32}(x_2 - \mu_2); v_3)$$
$$f(X_4|X_2) = \mathcal{N}(\beta_4 + \beta_{42}(x_2 - \mu_2); v_4)$$

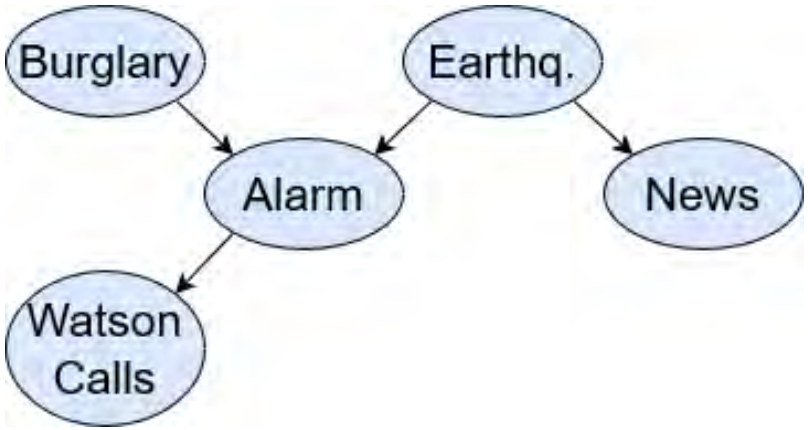
# Inference

Since in a Bayesian network the JPD is factorised:

- $P(X_i) = \sum_{\boldsymbol{x} \setminus X_i} \prod_{j=1}^{n} P(X_j | Pa_{X_j})$

Probability of the alarm, $P(A)$, in the "Watson calls" network

$$P(A) = \sum_{B,E,N,W} P(B)P(E)P(A|B,E)P(N|E)P(W|A)$$



# Inference: Abduction

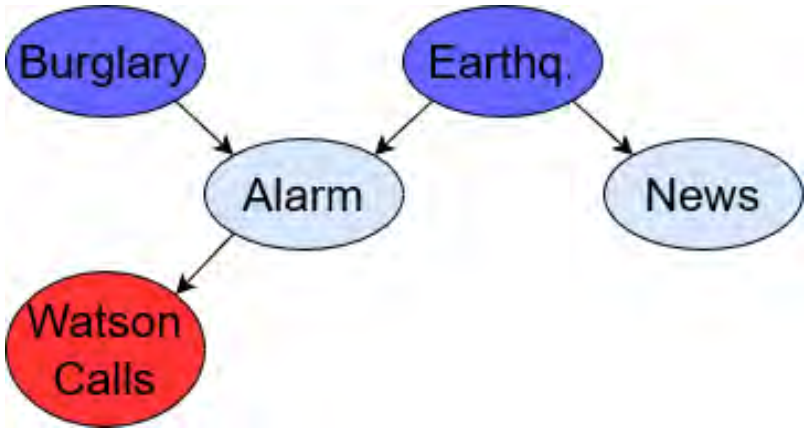
Given a set of evidence nodes $\boldsymbol{E} \subset \boldsymbol{X}$ with a value assignment $\boldsymbol{e}$ and a set of explanation nodes $\boldsymbol{H} \subseteq \boldsymbol{X} \setminus \boldsymbol{E}$, get the most probable explanation $\boldsymbol{h}^*$ for the evidence $\boldsymbol{e}$.

**Maximum a posteriori (MAP)**

$\boldsymbol{h}^* \leftarrow arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{e}),\ \boldsymbol{H} \subseteq \boldsymbol{X} \setminus \boldsymbol{E}$

Example: Watson called, what is the most probable explanation of $B$, $E$?

Answer: $arg\ max_{B,E} P(B, E|w)$

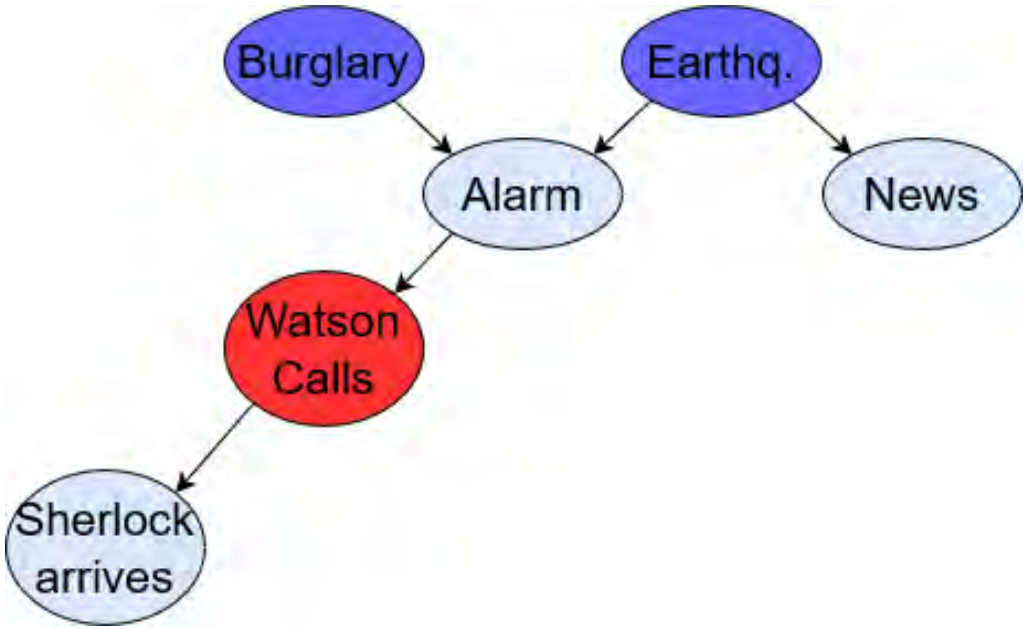


# Index

# MAP-independence: Motivation

Traditionally, the variables that are conditionally dependent from the explanation set $\boldsymbol{H}$ given the evidence $\boldsymbol{e}$ are considered to be relevant for the explanation (Pearl and Paz, 1985).

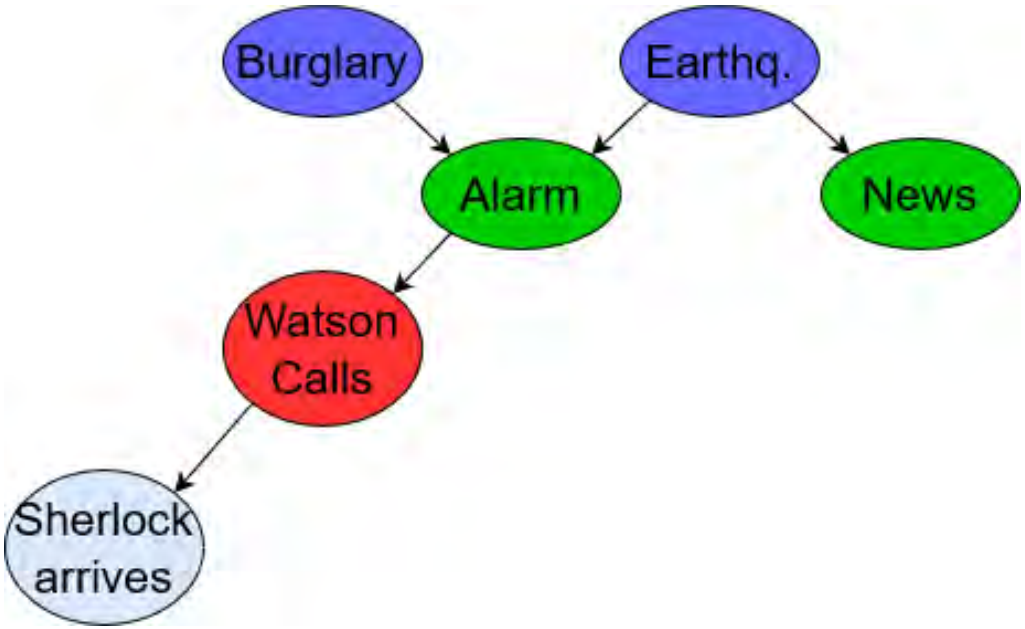
$$I(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{E}) \Leftrightarrow \boldsymbol{R} \textit{ not relevant}$$



# MAP-independence: Motivation

Traditionally, the variables that are conditionally dependent from the explanation set $\boldsymbol{H}$ given the evidence $\boldsymbol{e}$ are considered to be relevant for the explanation (Pearl and Paz, 1985).

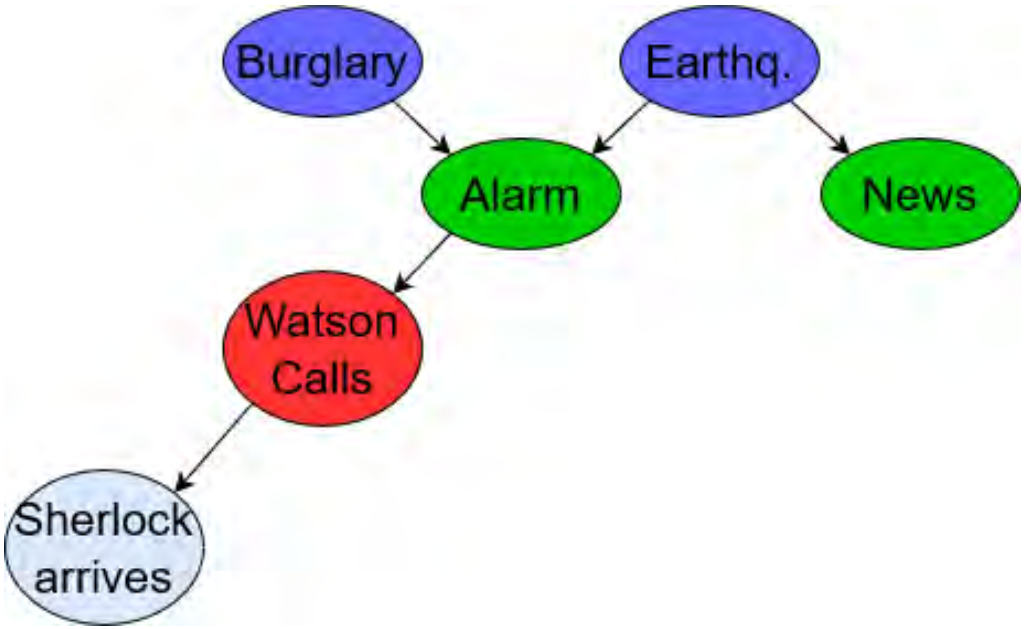
$$I(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{E}) \Leftrightarrow \boldsymbol{R} \textit{ not relevant}$$



# MAP-independence: Motivation

Traditionally, the variables that are conditionally dependent from the explanation set $\boldsymbol{H}$ given the evidence $\boldsymbol{e}$ are considered to be relevant for the explanation (Pearl and Paz, 1985).

$$I(\boldsymbol{H}, \boldsymbol{R}|\boldsymbol{E}) \Leftrightarrow \boldsymbol{R} \textit{ not relevant}$$




Are really all conditionally dependent variables relevant in a specific context?

# MAP-independence

Given a maximum a posteriori (MAP) $\boldsymbol{h}^* = arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{e})$ and a non-empty set of nodes $\boldsymbol{R}$,

**MAP-independence (Kwisthout, 2021)**

Is $\forall \boldsymbol{r} \in \Omega(\boldsymbol{R}),\ arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{r}|\boldsymbol{e}) = \boldsymbol{h}^*$ ?

If the answer to the decision problem is *Yes*, then the explanation $\boldsymbol{h}^*$ is MAP-independent from the set $\boldsymbol{R}$.

# MAP-independence

Given a maximum a posteriori (MAP) $\boldsymbol{h}^* = arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{e})$ and a non-empty set of nodes $\boldsymbol{R}$,

**MAP-independence (Kwisthout, 2021)**

Is $\forall \boldsymbol{r} \in \Omega(\boldsymbol{R}),\ arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{r}|\boldsymbol{e}) = \boldsymbol{h}^*$ ?

If the answer to the decision problem is *Yes*, then the explanation $\boldsymbol{h}^*$ is MAP-independent from the set $\boldsymbol{R}$.

Observations that motivate this work:

- The properties of MAP-independence are yet to be explored.
- What about continuous variables?

# Index

# MAP-independence and subsets

## MAP-independence and subsets

If $\boldsymbol{R_i} \subseteq \boldsymbol{R}$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R}$, then $\boldsymbol{h}^*$ is also MAP-independent from $\boldsymbol{R_i}$

# MAP-independence and subsets

## MAP-independence and subsets

If $\boldsymbol{R_i} \subseteq \boldsymbol{R}$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R}$, then $\boldsymbol{h}^*$ is also MAP-independent from $\boldsymbol{R_i}$

Let $\boldsymbol{R_j} = \boldsymbol{R} \setminus \boldsymbol{R_i}$:

$$\forall(\boldsymbol{r_i}, \boldsymbol{r_j}) \in \Omega(\boldsymbol{R_i}) \times \Omega(\boldsymbol{R_j}),\ \ arg\,max_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e}) = \boldsymbol{h}^*$$

Make the *arg max* explicit.

$$\forall \bar{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ \forall(\boldsymbol{r_i}, \boldsymbol{r_j}) \in \Omega(\boldsymbol{R_i}) \times \Omega(\boldsymbol{R_j}),\ P(\boldsymbol{h}^*, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e}) > P(\bar{\boldsymbol{h}}, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e})$$

# MAP-independence and subsets

## MAP-independence and subsets

If $\boldsymbol{R_i} \subseteq \boldsymbol{R}$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R}$, then $\boldsymbol{h}^*$ is also MAP-independent from $\boldsymbol{R_i}$

Let $\boldsymbol{R_j} = \boldsymbol{R} \setminus \boldsymbol{R_i}$:

$$\forall(\boldsymbol{r_i}, \boldsymbol{r_j}) \in \Omega(\boldsymbol{R_i}) \times \Omega(\boldsymbol{R_j}),\ \ arg\,max_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e}) = \boldsymbol{h}^*$$

Make the *arg max* explicit.

$$\forall\bar{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ \forall(\boldsymbol{r_i}, \boldsymbol{r_j}) \in \Omega(\boldsymbol{R_i}) \times \Omega(\boldsymbol{R_j}),\ P(\boldsymbol{h}^*, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e}) > P(\bar{\boldsymbol{h}}, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e})$$

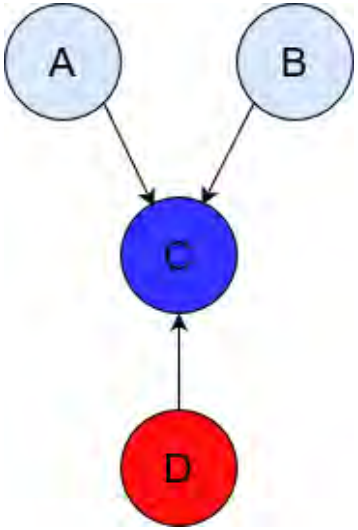
Marginalize $P(\boldsymbol{h}^*, \boldsymbol{r_i}|\boldsymbol{e})$

$$\forall\bar{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ \forall\boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \sum^{\boldsymbol{R_j}} P(\boldsymbol{h}^*, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e}) > \sum^{\boldsymbol{R_j}} P(\bar{\boldsymbol{h}}, \boldsymbol{r_i}, \boldsymbol{r_j}|\boldsymbol{e})$$

$$\forall\bar{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ \forall\boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ P(\boldsymbol{h}^*, \boldsymbol{r_i}|\boldsymbol{e}) > P(\bar{\boldsymbol{h}}, \boldsymbol{r_i}|\boldsymbol{e})$$

Introduce *arg max*

$$\forall\boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ arg\,max_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{r_i}|\boldsymbol{e}) = \boldsymbol{h}^*$$




$P(A)$

| A | |
|---|---|
| t | f |
| $\epsilon$ | $1-\epsilon$ |

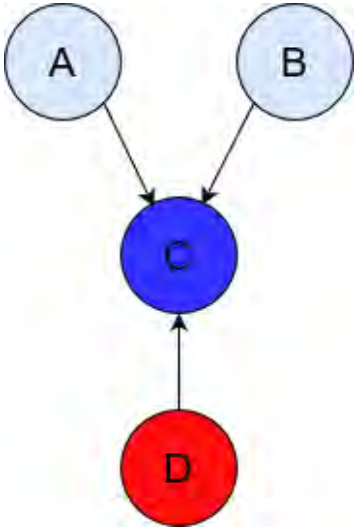
$P(B)$

| B | |
|---|---|
| t | f |
| $\mu$ | $1-\mu$ |

$P(C|A,B,d)$

| | | C | |
|---|---|---|---|
| A | B | t | f |
| f | f | 0.55 | 0.45 |
| f | t | 0.6 | 0.4 |
| t | f | 0.65 | 0.35 |
| t | t | 0.99 | 0.01 |

MAP query: If $d$, most probable value assignment for $C$?
$arg\ max_C P(C|d) = c$




$P(A)$

| A | |
|---|---|
| t | f |
| $\epsilon$ | $1-\epsilon$ |

$P(B)$

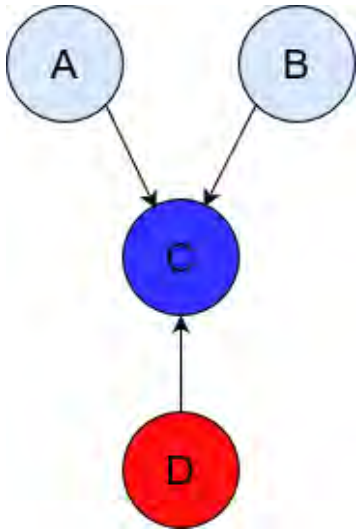
| B | |
|---|---|
| t | f |
| $\mu$ | $1-\mu$ |

$P(C|A,B,d)$

| | | C | |
|---|---|---|---|
| A | B | t | f |
| f | f | 0.55 | 0.45 |
| f | t | 0.6 | 0.4 |
| t | f | 0.65 | 0.35 |
| t | t | 0.99 | 0.01 |

MAP query: If $d$, most probable value assignment for $C$?
$arg\ max_C P(C|d) = c$

Check MAP-independence of the explanation $c$ from $A$ and $B$
$\forall (a,b) \in \Omega(A) \times \Omega(B),\ arg\ max_C P(C|A,B,d) = arg\ max_C P(C|d) = c$




$P(A)$

| A | |
|---|---|
| t | f |
| $\epsilon$ | $1-\epsilon$ |

$P(B)$

| B | |
|---|---|
| t | f |
| $\mu$ | $1-\mu$ |

$P(C|A,B,d)$

| A | B | C | |
|---|---|---|---|
| | | t | f |
| f | f | 0.55 | 0.45 |
| f | t | 0.6 | 0.4 |
| t | f | 0.65 | 0.35 |
| t | t | 0.99 | 0.01 |

MAP query: If $d$, most probable value assignment for $C$?
$arg\ max_C P(C|d) = c$

Check MAP-independence of the explanation $c$ from $A$ and $B$
$\forall (a,b) \in \Omega(A) \times \Omega(B),\ arg\ max_C P(C|A,B,d) = arg\ max_C P(C|d) = c$

Check MAP-independence of $c$ from just $A$
$\forall a \in \Omega(A),\ arg\ max_C\ P(C|a,d)$

$P(C|a,d) = \mu \cdot 0.99 + (1-\mu) \cdot 0.65 > 0.5;\ arg\ max_C P(C|a,d) = c$

$P(C|\bar{a},d) = \mu \cdot 0.6 + (1-\mu) \cdot 0.55 > 0.5;\ arg\ max_C P(C|\bar{a},d) = c$

# MAP-independence and conditional independence

## MAP-independence and conditional independence

If $I(\boldsymbol{H}, \boldsymbol{R_j}|\boldsymbol{R_i})$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_i}$, then $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_j}$.

# MAP-independence and conditional independence

## MAP-independence and conditional independence

If $I(\boldsymbol{H}, \boldsymbol{R_j}|\boldsymbol{R_i})$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_i}$, then $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_j}$.

Since $P(\boldsymbol{h}, \boldsymbol{r_i}|\boldsymbol{e}) \propto P(\boldsymbol{h}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \boldsymbol{h}^* = arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \forall \overline{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ P(\boldsymbol{h}^*|\boldsymbol{r_i}, \boldsymbol{e}) > P(\overline{\boldsymbol{h}}|\boldsymbol{r_i}, \boldsymbol{e})$

# MAP-independence and conditional independence

## MAP-independence and conditional independence

If $I(\boldsymbol{H}, \boldsymbol{R_j}|\boldsymbol{R_i})$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_i}$, then $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_j}$.

Since $P(\boldsymbol{h}, \boldsymbol{r_i}|\boldsymbol{e}) \propto P(\boldsymbol{h}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \boldsymbol{h}^* = arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \forall \overline{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ P(\boldsymbol{h}^*|\boldsymbol{r_i}, \boldsymbol{e}) > P(\overline{\boldsymbol{h}}|\boldsymbol{r_i}, \boldsymbol{e})$

For every probability distribution $P(\boldsymbol{R_i})$, the following holds

$\forall \overline{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ \sum^{\boldsymbol{R_i}} P(\boldsymbol{h}^*|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i}) > \sum^{\boldsymbol{R_i}} P(\overline{\boldsymbol{h}}|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i})$

$arg\ max_{\boldsymbol{H}} \sum^{\boldsymbol{R_i}} P(\boldsymbol{H}|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i}) = \boldsymbol{h}^*$

# MAP-independence and conditional independence

## MAP-independence and conditional independence

If $I(\boldsymbol{H}, \boldsymbol{R_j}|\boldsymbol{R_i})$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_i}$, then $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_j}$.

Since $P(\boldsymbol{h}, \boldsymbol{r_i}|\boldsymbol{e}) \propto P(\boldsymbol{h}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \boldsymbol{h}^* = arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \forall \overline{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ P(\boldsymbol{h}^*|\boldsymbol{r_i}, \boldsymbol{e}) > P(\overline{\boldsymbol{h}}|\boldsymbol{r_i}, \boldsymbol{e})$

For every probability distribution $P(\boldsymbol{R_i})$, the following holds

$\forall \overline{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ \sum^{\boldsymbol{R_i}} P(\boldsymbol{h}^*|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i}) > \sum^{\boldsymbol{R_i}} P(\overline{\boldsymbol{h}}|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i})$

$arg\ max_{\boldsymbol{H}} \sum^{\boldsymbol{R_i}} P(\boldsymbol{H}|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i}) = \boldsymbol{h}^*$

i.e. any probability distribution $P(\boldsymbol{R_i})$ over $\boldsymbol{R_i}$ cannot alter $\boldsymbol{h}^*$

# MAP-independence and conditional independence

## MAP-independence and conditional independence

If $I(\boldsymbol{H}, \boldsymbol{R_j}|\boldsymbol{R_i})$ and $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_i}$, then $\boldsymbol{h}^*$ is MAP-independent from $\boldsymbol{R_j}$.

Since $P(\boldsymbol{h}, \boldsymbol{r_i}|\boldsymbol{e}) \propto P(\boldsymbol{h}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \boldsymbol{h}^* = arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{r_i}, \boldsymbol{e})$

$\forall \boldsymbol{r_i} \in \Omega(\boldsymbol{R_i}),\ \forall \overline{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ P(\boldsymbol{h}^*|\boldsymbol{r_i}, \boldsymbol{e}) > P(\overline{\boldsymbol{h}}|\boldsymbol{r_i}, \boldsymbol{e})$

For every probability distribution $P(\boldsymbol{R_i})$, the following holds

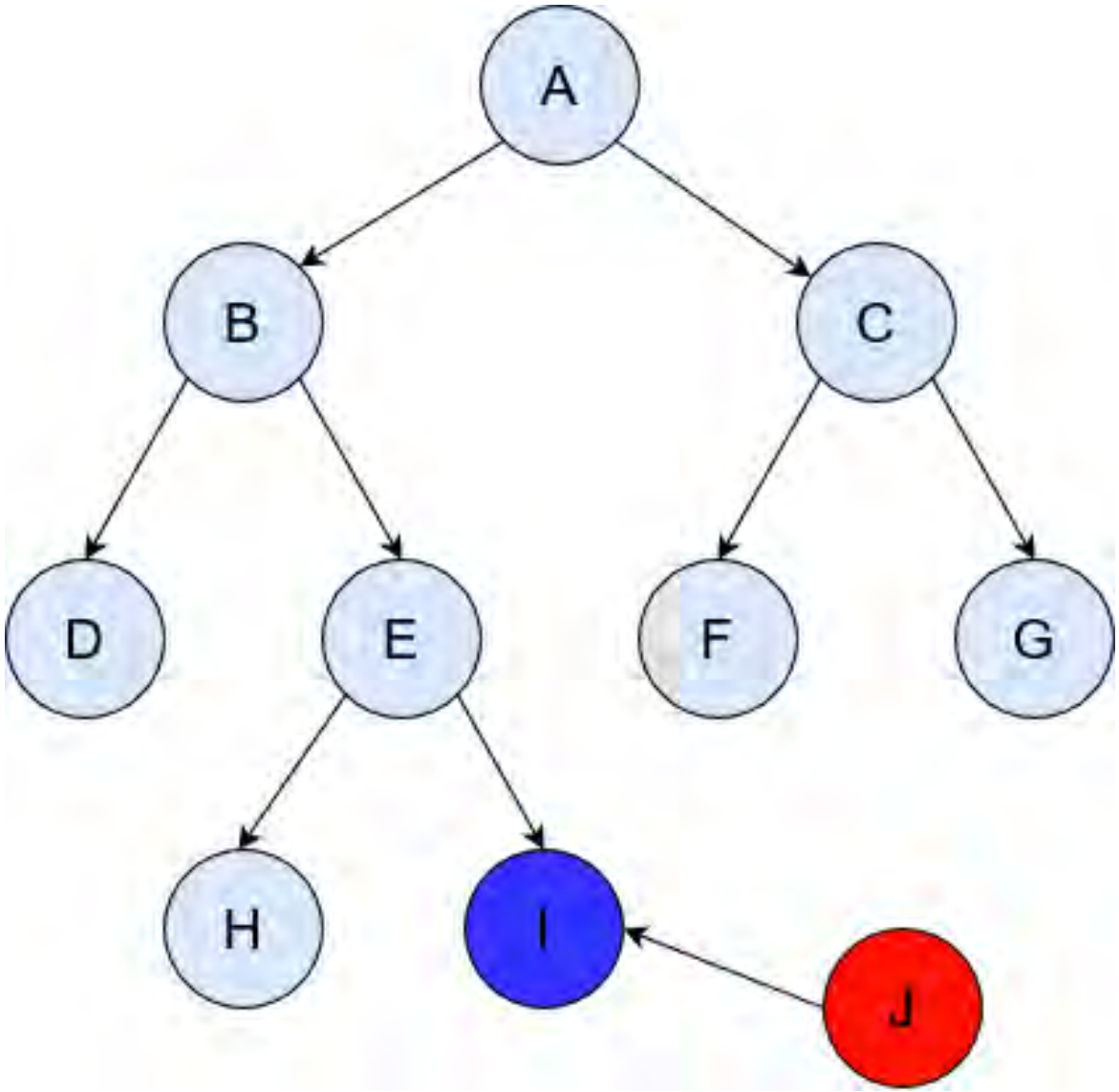
$\forall \overline{\boldsymbol{h}} \in \Omega(\boldsymbol{H}) \setminus \boldsymbol{h}^*,\ \sum^{\boldsymbol{R_i}} P(\boldsymbol{h}^*|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i}) > \sum^{\boldsymbol{R_i}} P(\overline{\boldsymbol{h}}|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i})$

$arg\ max_{\boldsymbol{H}} \sum^{\boldsymbol{R_i}} P(\boldsymbol{H}|\boldsymbol{r_i}, \boldsymbol{e})P(\boldsymbol{r_i}) = \boldsymbol{h}^*$

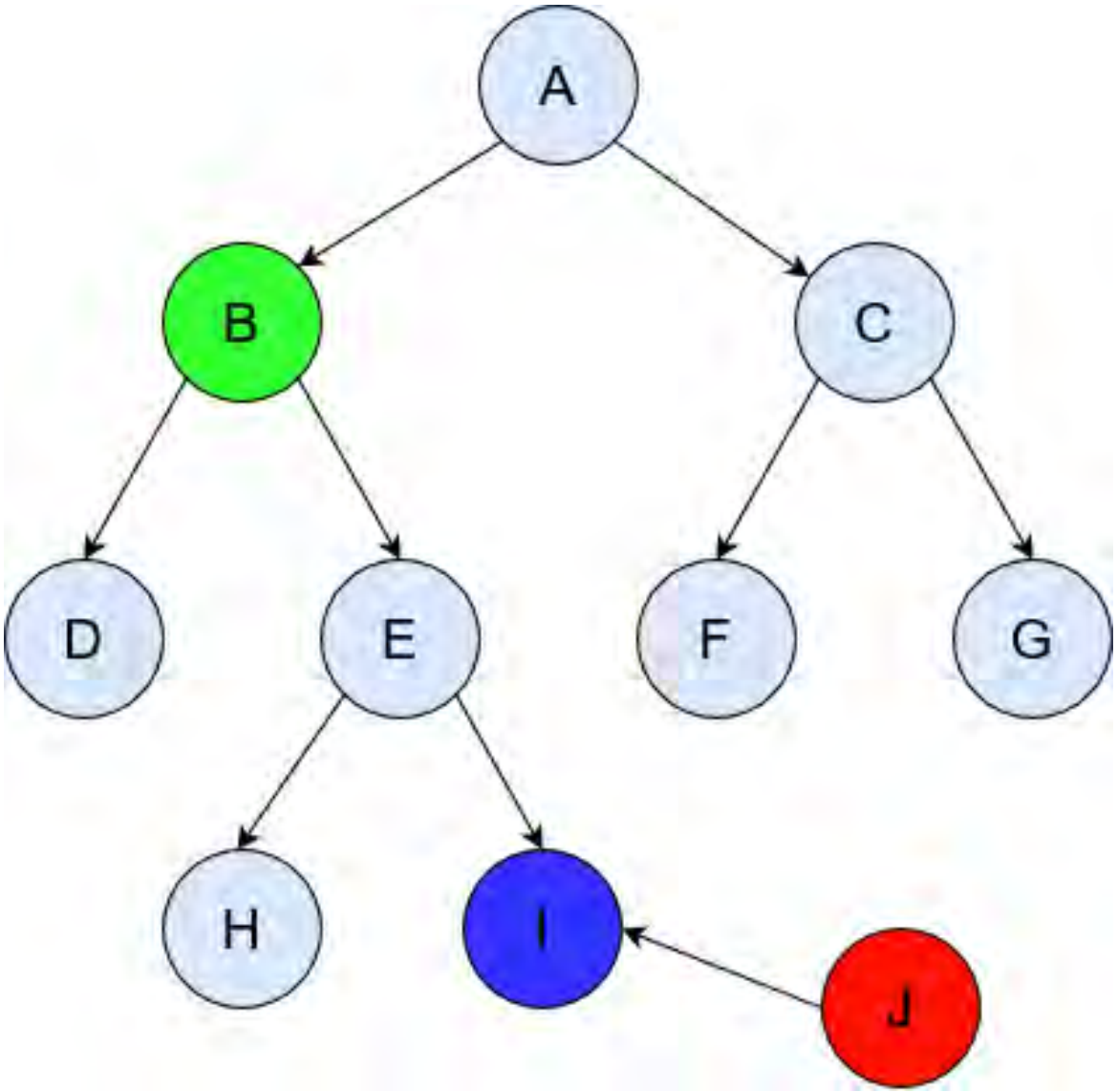
i.e. any probability distribution $P(\boldsymbol{R_i})$ over $\boldsymbol{R_i}$ cannot alter $\boldsymbol{h}^*$

Since $\boldsymbol{R_j}$ only affects $\boldsymbol{H}$ through $\boldsymbol{R_i}$ (c.i. given $\boldsymbol{R_i}$), $\boldsymbol{R_j}$ cannot alter $\boldsymbol{h}^*$.

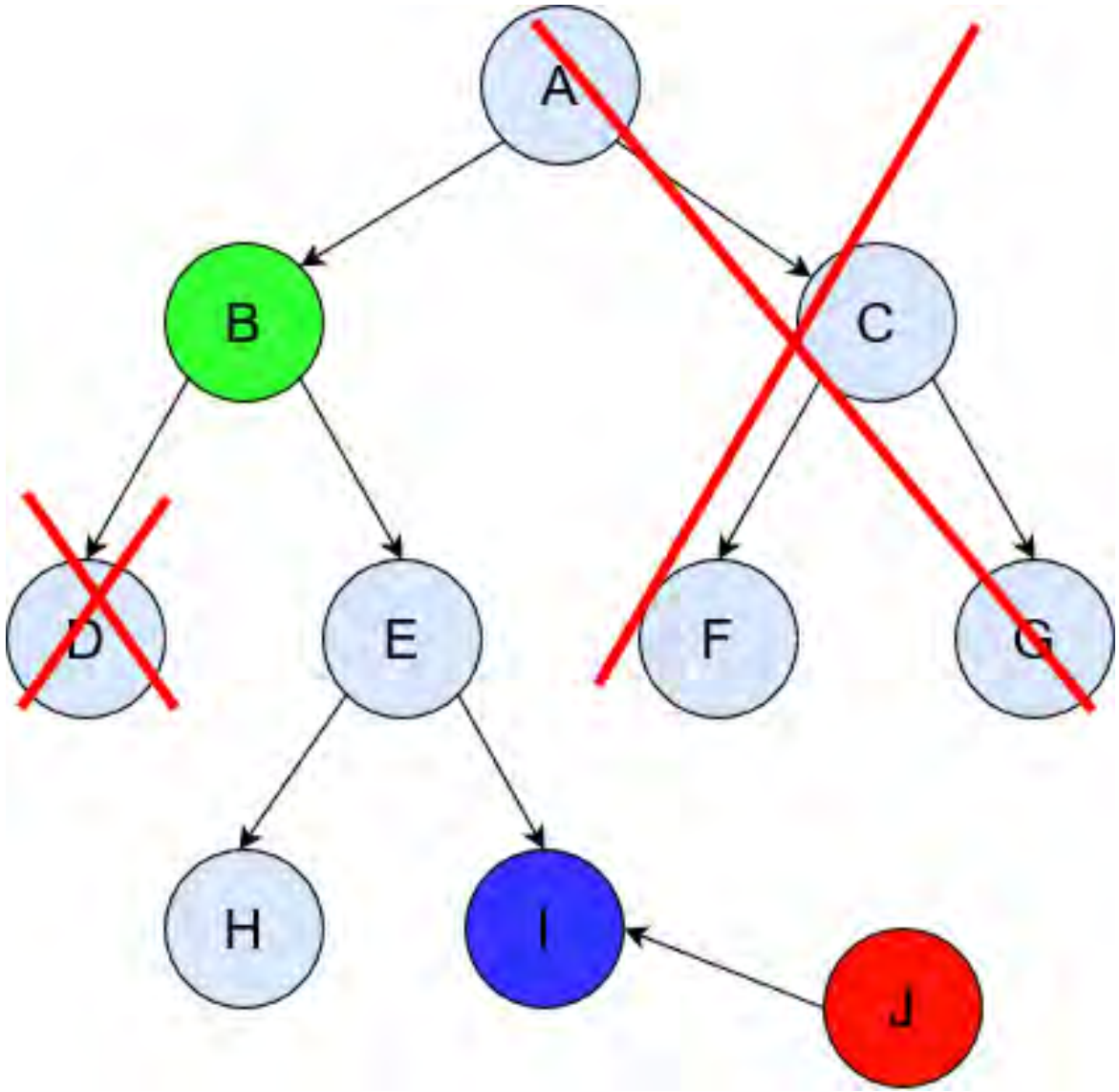
# Evidence J, MAP query over I yields $i^*$



# $i^*$ is MAP-independent from B



$i^*$ is also MAP-independent from the nodes d-separated from I given B



# Index

# MAP-independence in continuous Bayesian networks

## MAP-independence

Is $\forall \boldsymbol{r} \in \Omega(\boldsymbol{R}) arg\ max_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{r}|\boldsymbol{e}) = \boldsymbol{h}^*$ ?

*arg max* of a continuous distribution is a real number or a vector of real numbers (the mode or a vectorial mode).

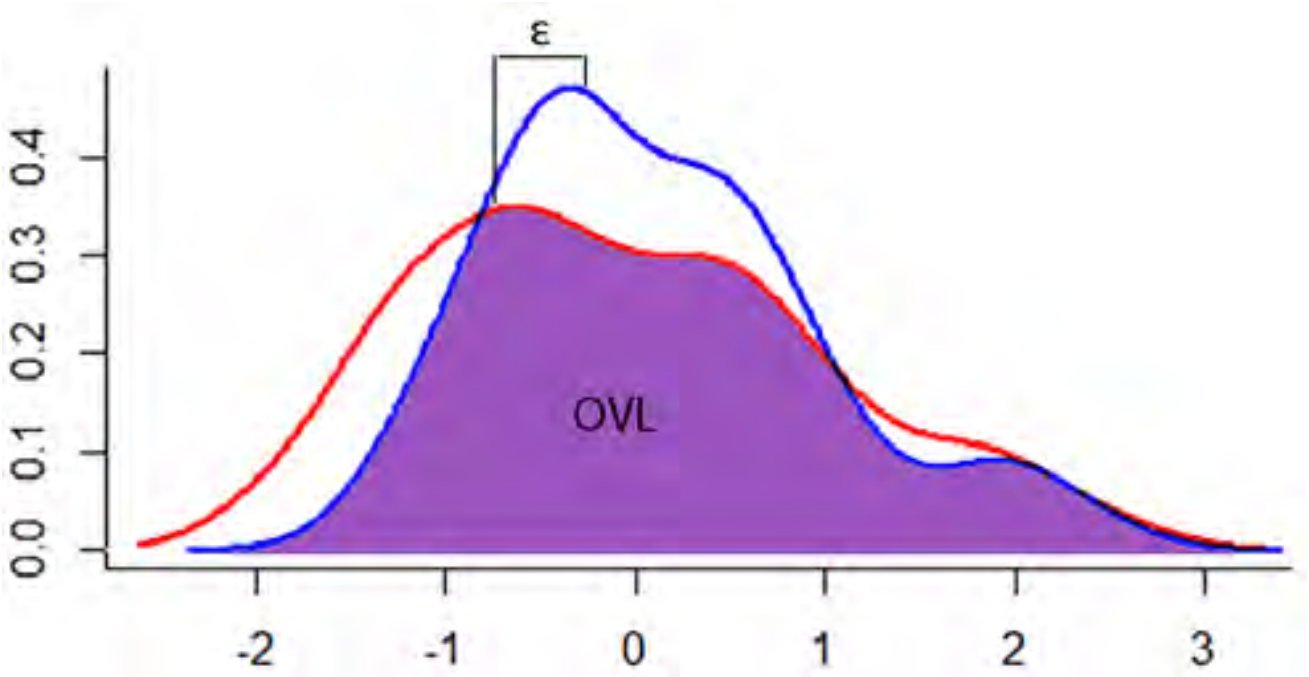
- Use a threshold of tolerance $\epsilon$ to compare the modes
- Another option is to compare the probability distributions directly.
- Both ideas can be captured within an abstract function *dist*()

Change in the formulation

- Is $\forall \boldsymbol{r} \in \Omega(\boldsymbol{R})\ dist(P(\boldsymbol{H}|\boldsymbol{e}), P(\boldsymbol{H}|\boldsymbol{e}, \boldsymbol{r})) < \epsilon$?

Some examples of *dist*() functions (we could define others):

1. sum($|arg\,max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{e}) - arg\,max_{\boldsymbol{H}} P(\boldsymbol{H}|\boldsymbol{e},\boldsymbol{r})|$).
2. The Jensen-Shannon divergence, $JSD(P(\boldsymbol{H}|\boldsymbol{e})||P(\boldsymbol{H}|\boldsymbol{e},\boldsymbol{r}))$.
3. The overlapping coefficient between the two distributions, $OVL(P(\boldsymbol{H}|\boldsymbol{e}), P(\boldsymbol{H}|\boldsymbol{e},\boldsymbol{r}))$.




8th-11th November 2021

# $\epsilon$-MAP-independence

**MAP-independence**

Is $\forall \boldsymbol{r} \in \Omega(\boldsymbol{R})$ $arg\, max_{\boldsymbol{H}} P(\boldsymbol{H}, \boldsymbol{r}|\boldsymbol{e}) = \boldsymbol{h}^*$ ?

Iterating over all values of $\boldsymbol{R}$ ($\forall \boldsymbol{r} \in \Omega(\boldsymbol{R})$) $\rightarrow$ not possible (continuous).

- Sample from $P(\boldsymbol{R})$ and check if on average the condition holds.

Merging the two proposals:

**$\epsilon$-MAP-independence**

Is $dist(P(\boldsymbol{H}|\boldsymbol{e}), P(\boldsymbol{H}|\boldsymbol{e}, \boldsymbol{r})) < \epsilon$, where $\boldsymbol{r}$ is sampled from $P(\boldsymbol{R})$?

# Gaussian linear Bayesian networks

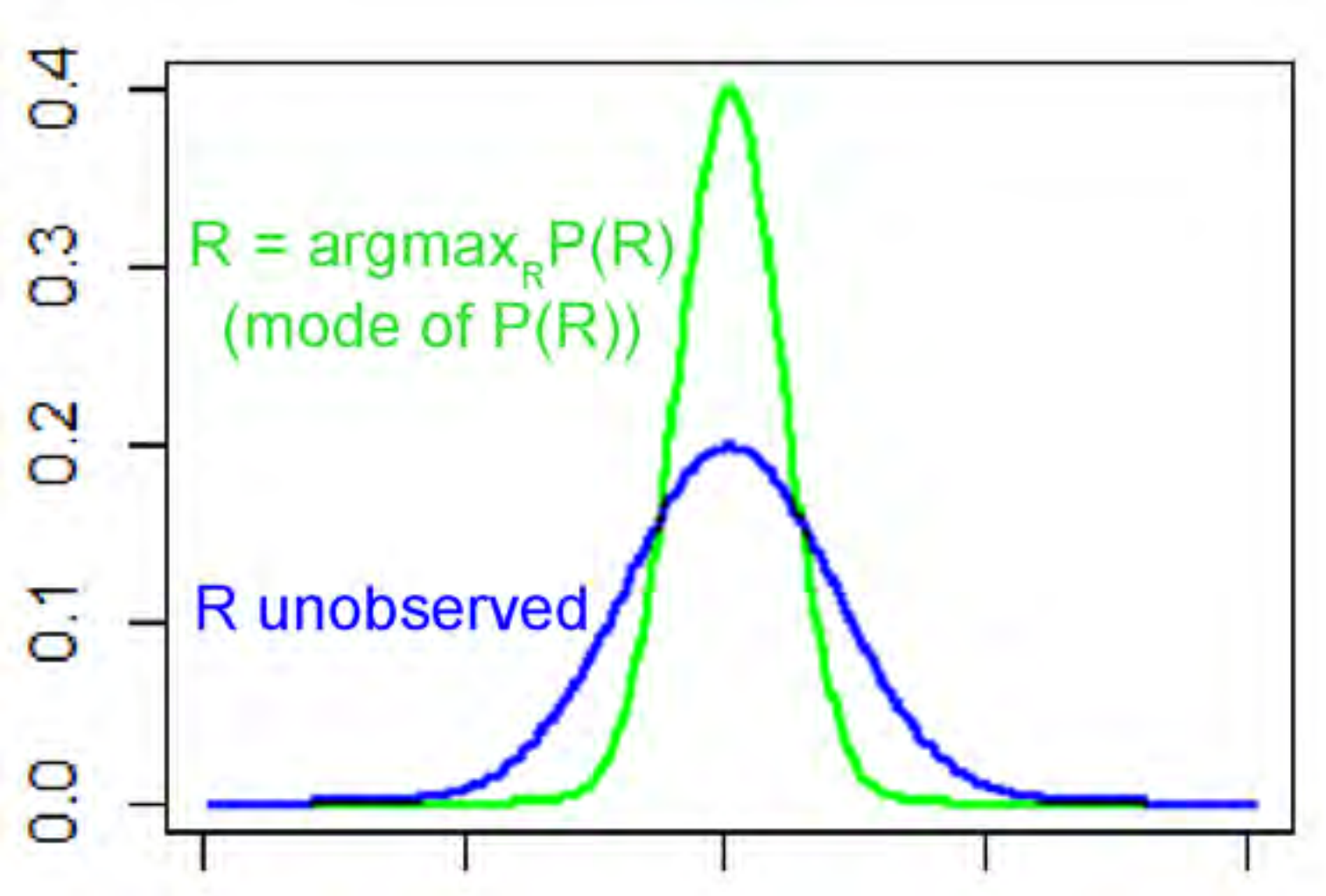
In the particular case of Gaussian Bayesian networks, we can take advantage of the linear nature of the node means.

Conditional density of a node $Y$ with parents $Pa_Y$:

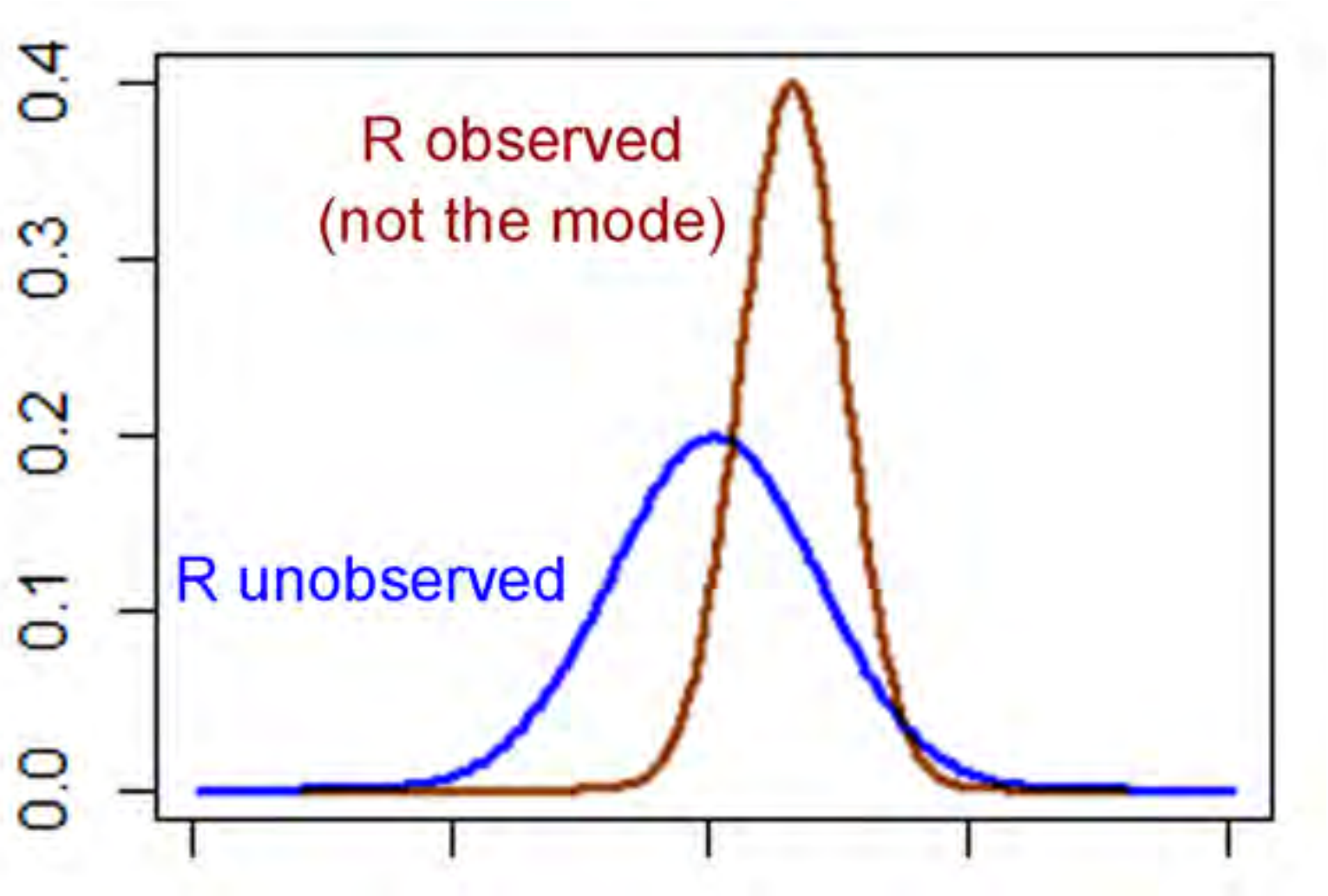
$$f(Y|Pa_Y) = \mathcal{N}(\beta_Y + \sum_{X_i}^{Pa_Y} \beta_{YX_i}(x_i - \mu_i) \; ; v_Y)$$

Some prior observation regarding inference:

If $\boldsymbol{r} = arg\ max_{\boldsymbol{R}} P(\boldsymbol{R})$, then $dist(P(\boldsymbol{H}|\boldsymbol{e}), P(\boldsymbol{H}|\boldsymbol{e}, \boldsymbol{r}))$ is minimum.



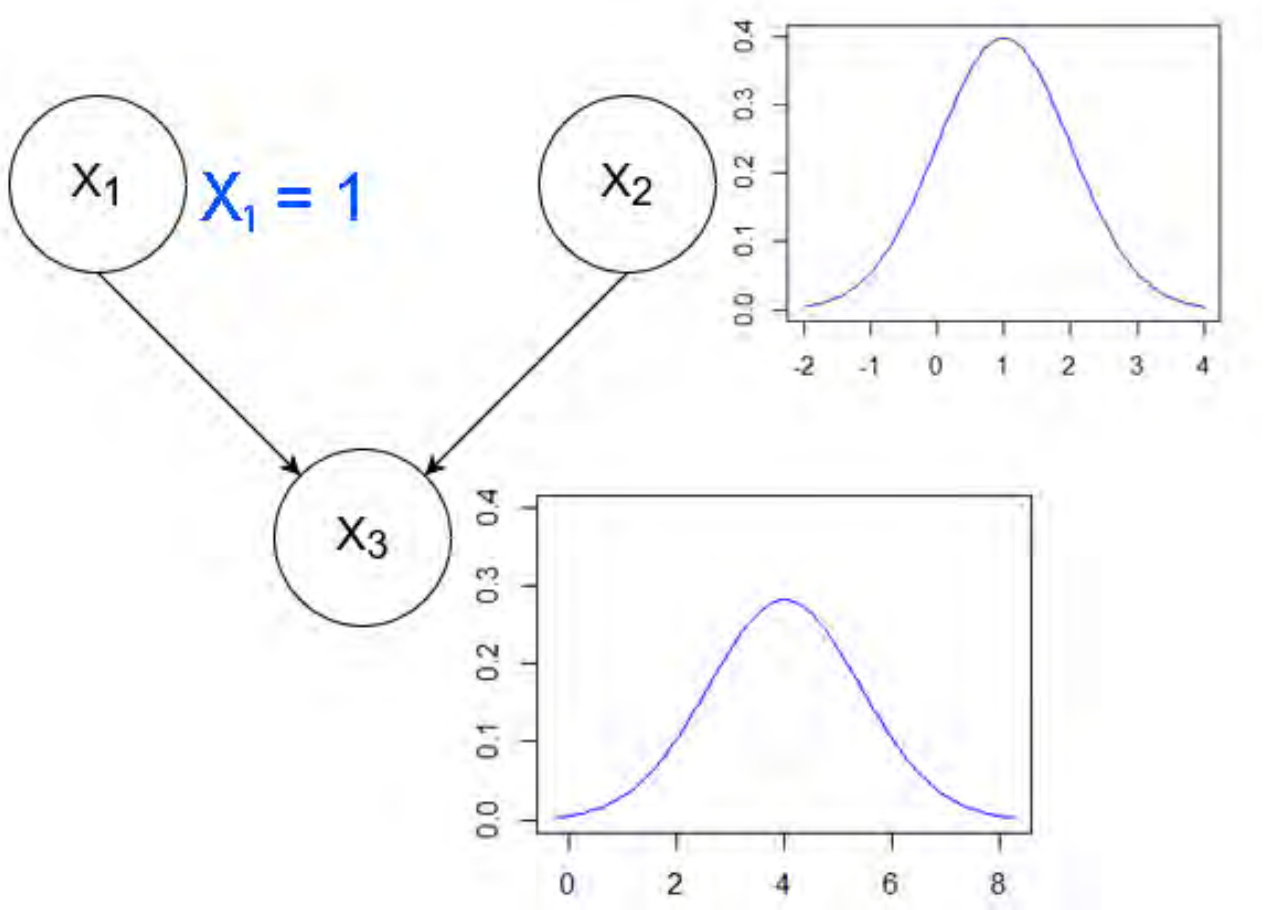
If $\boldsymbol{r} = arg\ max_{\boldsymbol{R}} P(\boldsymbol{R})$, then $dist(P(\boldsymbol{H}|\boldsymbol{e}), P(\boldsymbol{H}|\boldsymbol{e}, \boldsymbol{r}))$ is minimum.




8th-11th November 2021

# Gaussian linear Bayesian networks

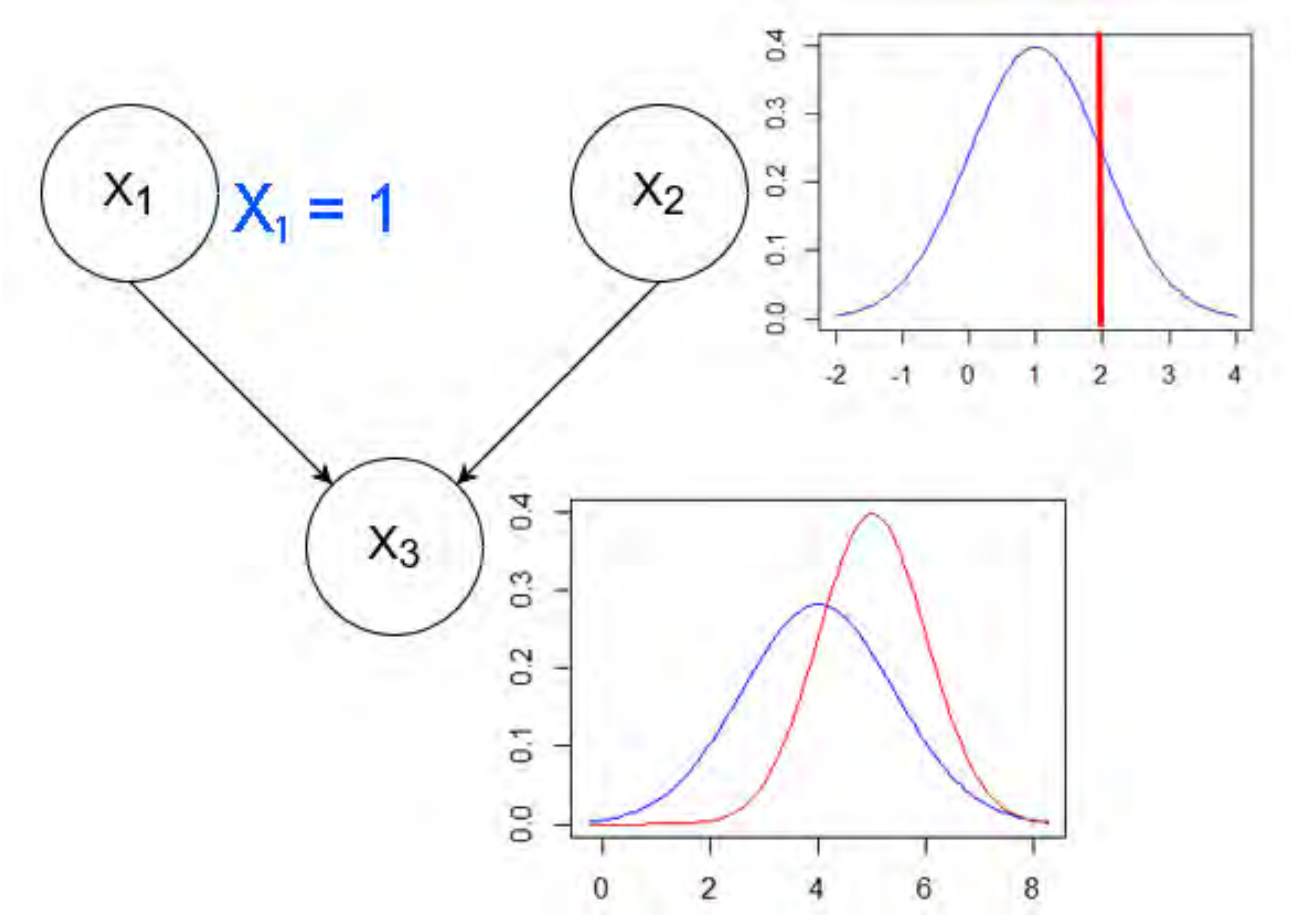
We observe a certain evidence $X_1 = 1$. Observe $P(X_2)$ and $P(X_3|X_1 = 1)$




8th-11th November 2021

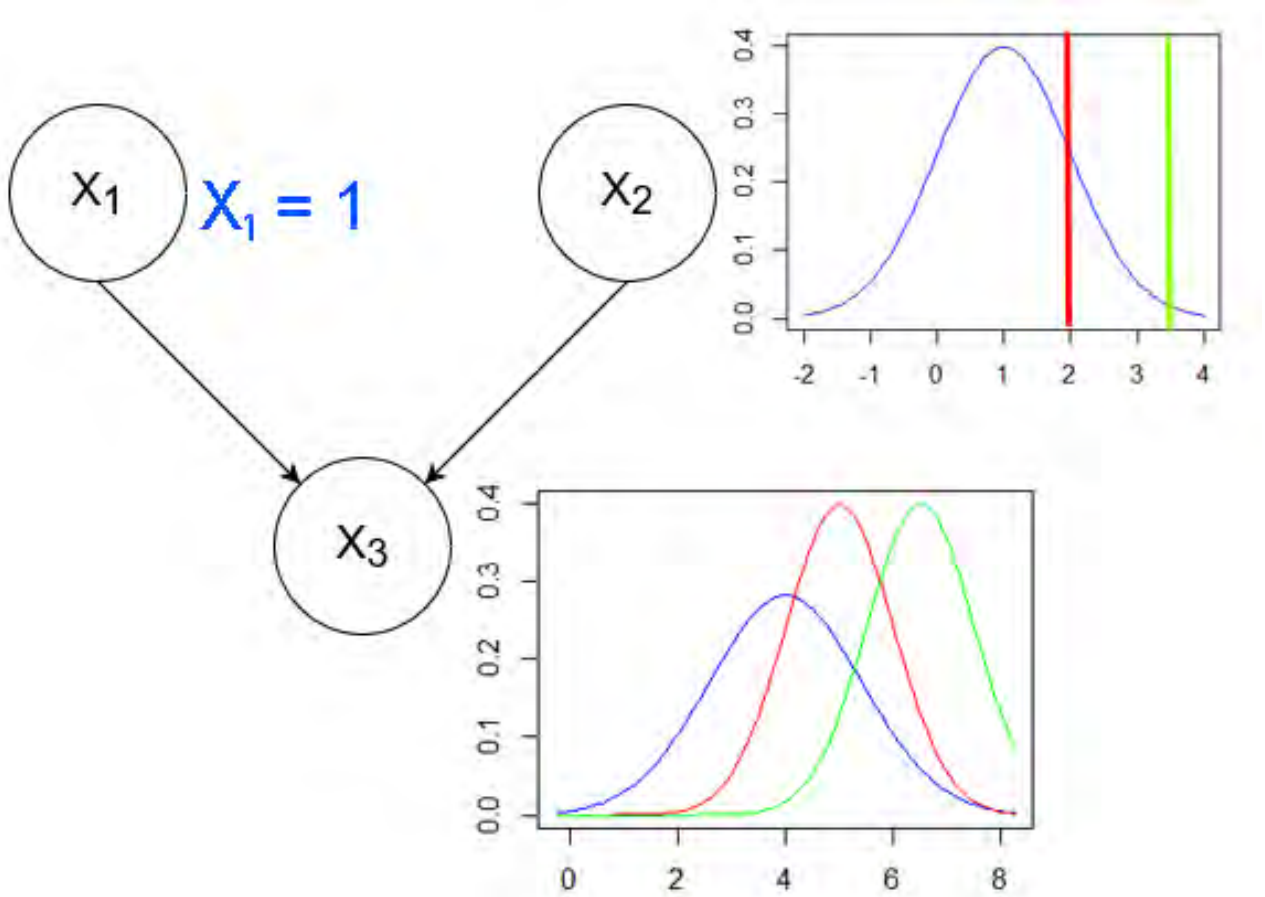
# Gaussian linear Bayesian networks

We sample "2" from $P(X_2)$. In red, $P(X_3|X_1 = 1, X_2 = 2)$



# Gaussian linear Bayesian networks

Since $X_2 = 3.5$ is further from the $X_2$ mode ($arg\ max_{X_2} P(X_2) = 1$) than $X_2 = 2$, then $P(X_3|X_1 = 1, X_2 = 3.5)$ will be more dissimilar (in green)



## $\epsilon$-MAP-independence

Is $dist(P(\boldsymbol{H}|\boldsymbol{e}), P(\boldsymbol{H}|\boldsymbol{e}, \boldsymbol{r})) < \epsilon$, where $\boldsymbol{r}$ is sampled from $P(\boldsymbol{R})$?

A single sample $\boldsymbol{r'}$ from $P(\boldsymbol{R})$ gives us plenty of information

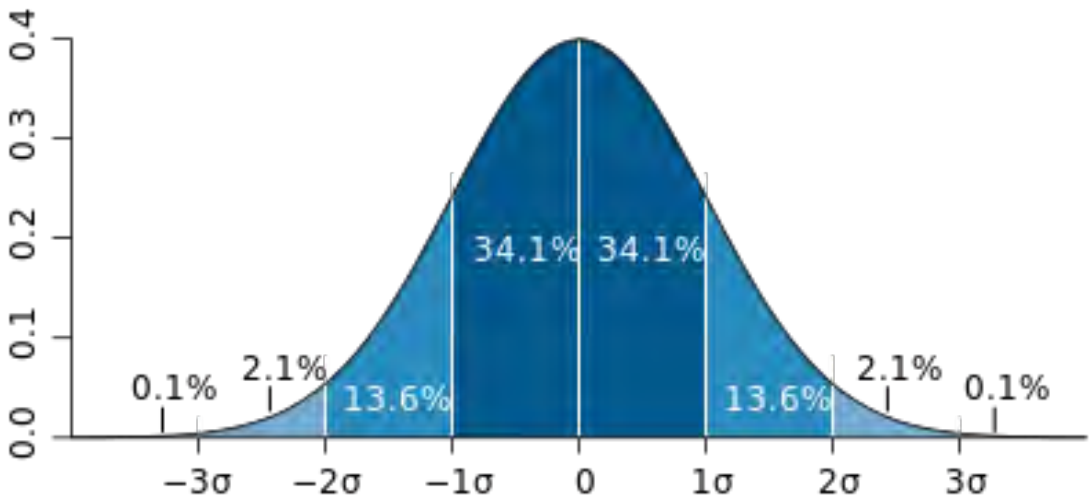
- If $\boldsymbol{r'}$ meets the $\epsilon$-MAP-independence criterion, any sample in range $[arg\ max_{\boldsymbol{R}} P(\boldsymbol{R}), \boldsymbol{r'}]$ will.
- Otherwise, not all values in range $[arg\ max_{\boldsymbol{R}} P(\boldsymbol{R}), \boldsymbol{r'}]$ meet it.

The question remains:

- How many and which $\boldsymbol{r'}$s to select?

We want a value that gives us information about a large part of $P(\boldsymbol{R})$

- Checking only $2\sigma$ and $-2\sigma$ let us check the 95% of its probability.
- We need $2^{|\boldsymbol{R}|}$ samples (combinations of $\pm 2\sigma$ for each $R \in \boldsymbol{R}$).




wikimedia

# Index

# Conclusion

The interest on MAP-independence is two-fold:

1. It lets us study the node relevance and explanation robustness with a stricter criterion.
2. The methodology is extremely easy to understand $\rightarrow$ XAI principle, legislation...

Although our extension to the continuous domain relies heavily on heuristics, simulation methods are still useful and will enable MAP-independence to be applied in many more domains.

# Future ways

Most of the datasets nowadays are either continuous or contain continuous and discrete variables

- Extend the methodology to hybrid Bayesian networks.

We can use the formulated MAP-independence properties to design an efficient algorithm

- Start from the closest nodes to $\boldsymbol{H}$ and explore "outwards" (i.e., starting from the Markov blanket).
- Start from a reduced number of variables.

The definition of MAP-independence is related to counterfactuals (Molnar, 2020)

- Check if a change in $\boldsymbol{R}$ produces a counterfactual in $\boldsymbol{H}$.

# Acknowledgements

Acknowledgements:

- Spanish Ministry of Science and Innovation through the PID 2019-109247GB-I00 project,
- The BBVA Foundation (2019 Call) through the "Score-based nonstationary temporal Bayesian networks. Applications in climate and neuroscience" project,
- The BBVA Foundation's grants (2020 Call) for Scientific Investigation Teams SARS-CoV-2 and COVID-19 through the "Outcome prediction and treatment efficiency inpatients hospitalized with Covid-19 in Madrid: A Bayesian network approach" project.

# References I


Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 235–243. Elsevier.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Kwisthout, J. (2021). Explainable AI Using MAP-independence. In *European Conference on Symbolic and Quantitative Approaches with Uncertainty*, pages 243–254. Springer.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Molnar, C. (2020). *Interpretable Machine Learning*. Lulu. com.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. and Paz, A. (1985). *Graphoids: A graph-based logic for reasoning about relevance relations*. University of California (Los Angeles). Computer Science Department.

Shachter, R. D. and Kenley, C. R. (1989). Gaussian influence diagrams. *Management Science*, 35(5):527–550.

Verma, T. and Pearl, J. (1990). Causal networks: semantics and expressiveness. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, pages 69–78.