Causal EM for counterfactual inference, with an application to palliative care

(me)

Marco Zaffalon

IDSIA

Alessandro Antonucci

IDSIA

Rafael Cabañas

Universitas Almeriensis

David Huber

IDSIA

Dario Azzimonti

IDSIA

Artificialy

Work from:

PGM 2020
WHY-21 @ NeurIPS
PGM 2022

IDSIA common
institute of USI & SUPSI
since 2000

SUPSI

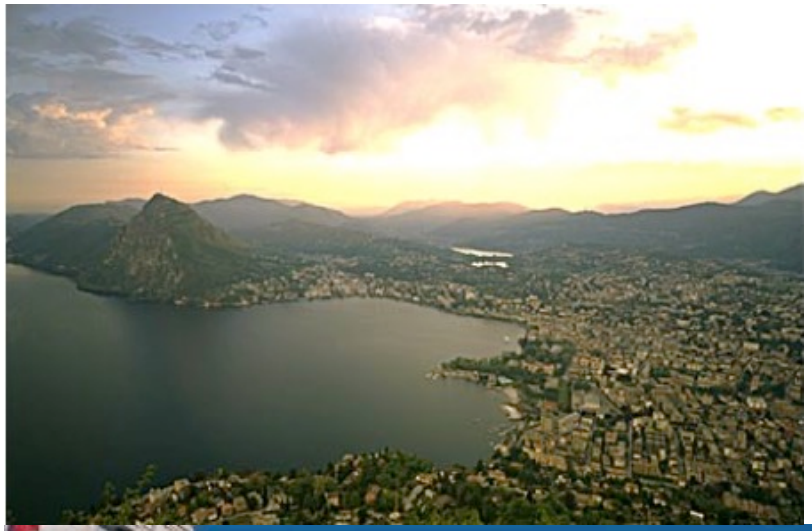
Istituto Dalle Molle di studi sull'intelligenza artificiale

IDSIA is a research institute on **Artificial Intelligence** founded in 1988 in Lugano, Switzerland

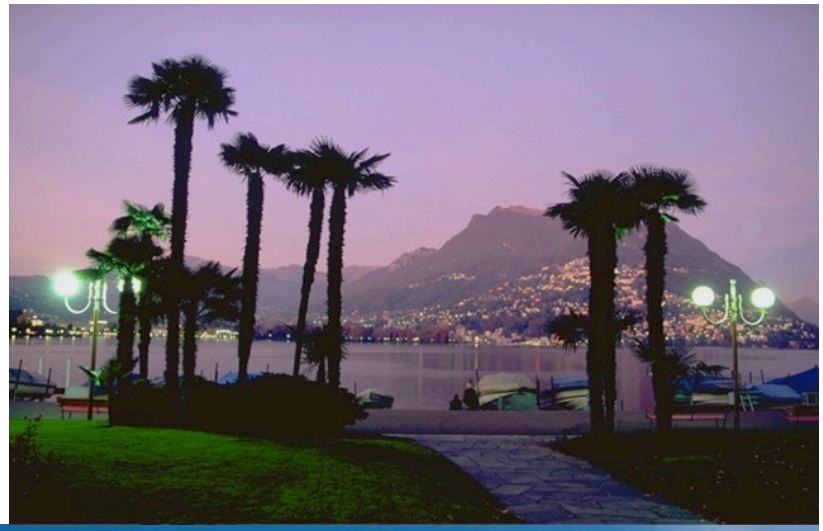
121 people = 16 Professors, 51 Researchers, 45 PhD students, 9 Research assistants

Thanks to Italian philantropist Angelo Dalle Molle (1908-2002)

«*The progress of science in general, and that of the emerging computer science in particular, should not subjugate people, but rather benefit them*»

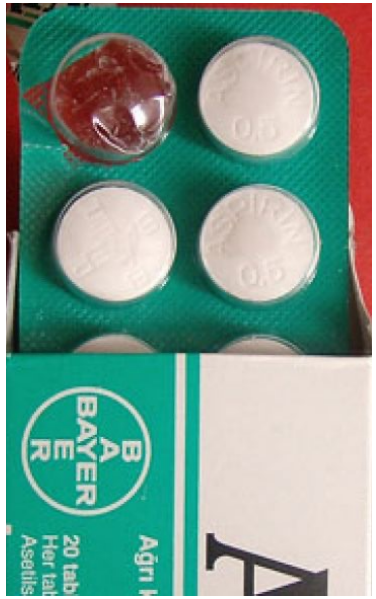


Lugano



Headache

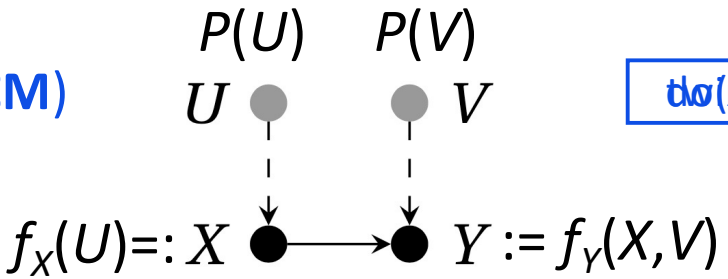
- You take an aspirin (X=1) and your headache vanishes (Y=1)



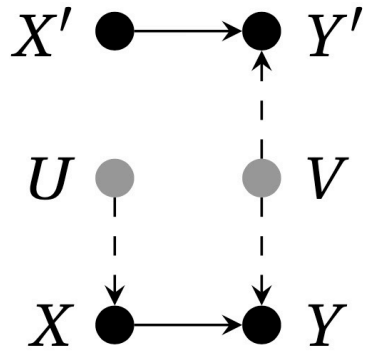
- What is the probability that this has been due to the aspirin?
  - Data says: $P(Y{=}0|X{=}0){=}0.5$ and $P(Y{=}0|X{=}1){=}0.1$

- What if I had **not** taken the aspirin, would have headache stayed?
- Probability of necessity (*PN*): $P(Y_{X=0} = 0|Y = 1, X = 1)$

- We need a fully specified structural causal model (**SCM**) to compute it
  - **What if not?**
  - **Yet *P*(*X*,*Y*) is available**



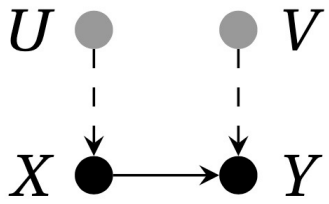
f_X, f_Y are so-called **structural equations**



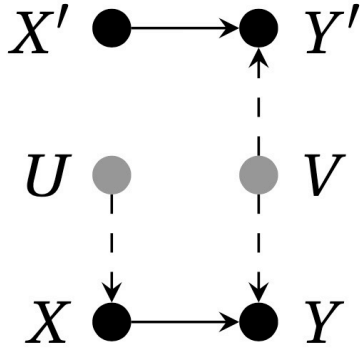
PN = $P(Y'|X'{=}0,X{=}1,Y{=}1)$

Causal inference
(via *credal nets* = sets of Bayesian nets)

- Select number of states for U and V
- No knowledge:
  - Conservative specification (canonical partition) |U| = 2, |V| = 4



- Deterministic functions written via 0-1 valued probabilities for P(X|U) and P(Y|X,V)
- Propagate P(X,Y) back to find out P(U) and P(V)
- We get P(U) = P(X) and P(V) = [t, 0.4 + t, 0.5 - t, 0.1 - t], t∈[0,0.1] (we call this a **credal set** = a set of distributions)
- Create its *twin net* where PN = P(Y'|X'=0,X=1,Y=1)

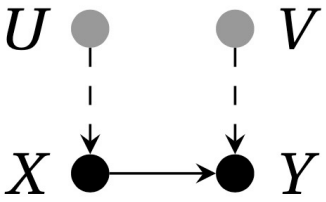


- Run an *exact* credal net algorithm to finally get 4/9 <= PN <= 5/9

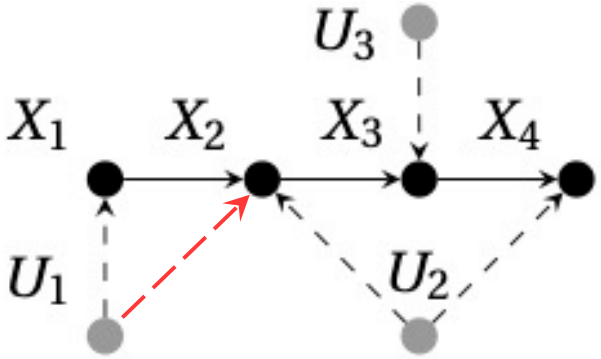
One can *in principle* solve all counterfactuals with this methodology

Problem

- The previous exact approach works well 😊
  with Markovian structural causal models (SCMs)



- So-so 🙂 with quasi-Markovian ones
- And does not 😕 for non-quasi-Markovian SCMs



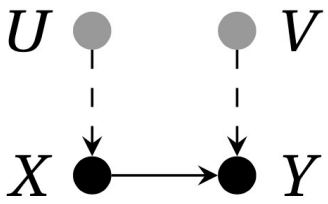
- Thm.: Causal inference (interventions) is NP-hard even in polytree-SCMs
  – Hardly surprising if you're in credal nets

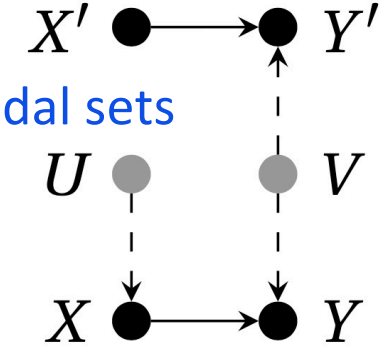
- Let's approximate!
- Idea:

the exogenous variables are missing at random ⇒ use the **EM**!

EM for Causal Computation (EMCC)

- Say we have a data set D of iid (x,y) instances
- Randomly initialise P(U) and P(V)
- Run the EM up to convergence





- Cor.: At convergence, P(U) and P(V) belong to their corresponding credal sets
  - EM samples the space of compatible fully specified SCMs!
  - On each of them, we can compute PN on its twin graph via Bayesian nets

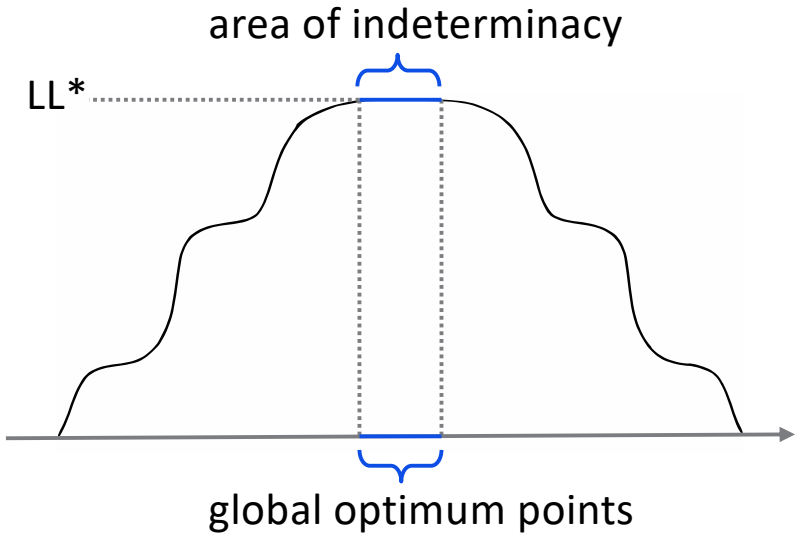
- Repeat k times: random initialisation + EM up to convergence + BN algorithm
  - You get $\{P_i(U),P_i(V)\}$ and PN_i, for i=1, …, k
  - A set of k points inside the interval [4/9,5/9]

- Take min and max and you get an **inner** approximation: $[a,b] \subseteq [4/9,5/9]$
  - k=20 already gives a pretty good approximation

works for any semi-Markovian SCM with categorical endogenous variables

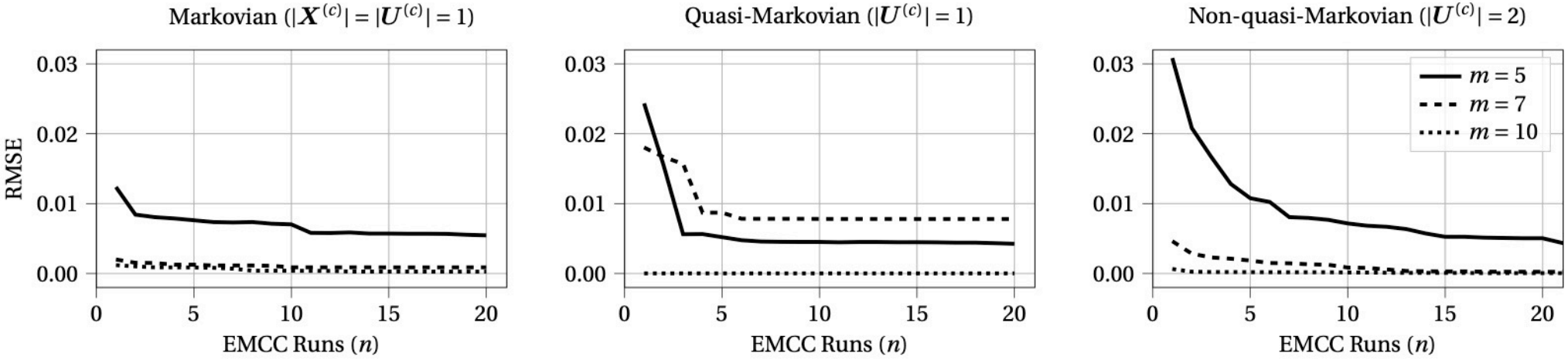
Why does it work?

- Thm.: The (log-)likelihood is unimodal
- Cor.: The global optimum points are IFF $P(U_1)$, …, $P(U_m)$ in their resp. credal sets



How well does it work?

(code available at github.com/idsia/credici)



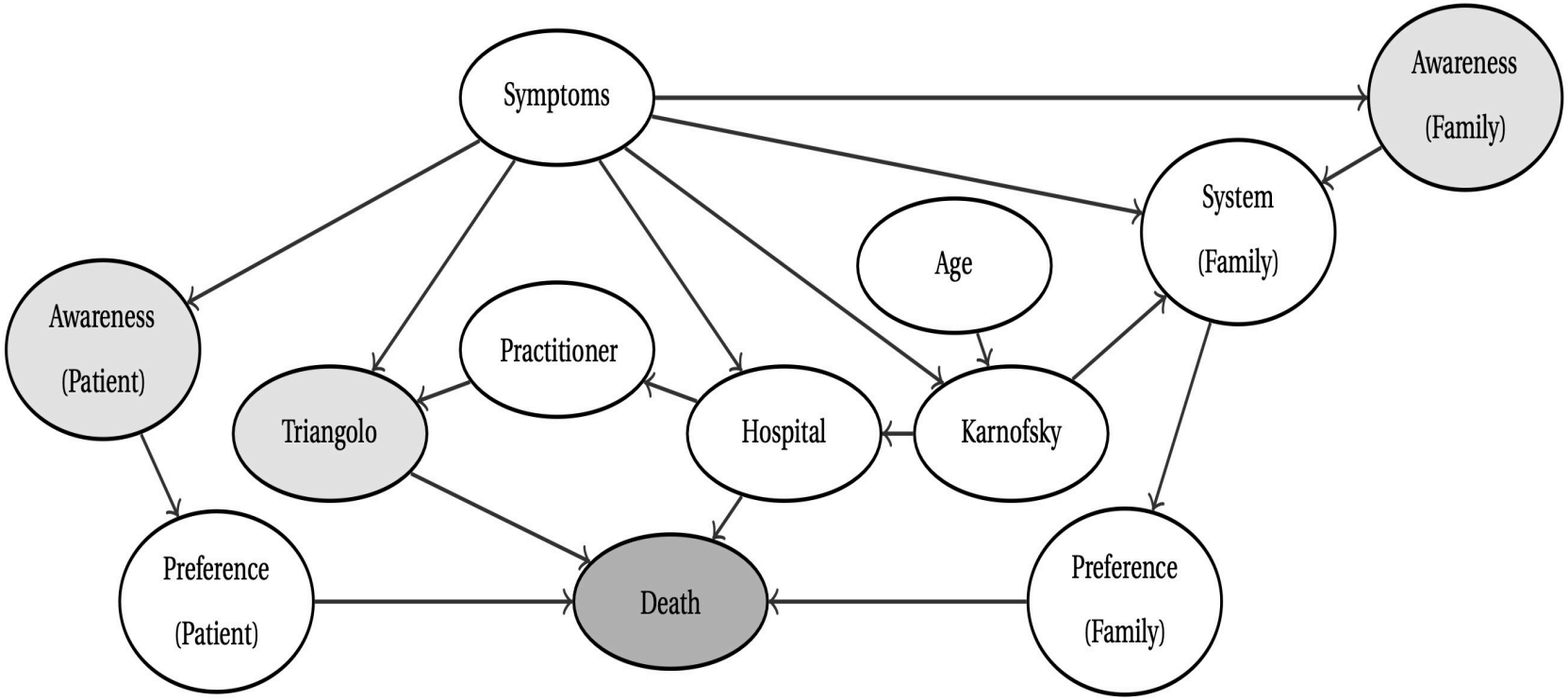
- Thm: $$P\left(a - \varepsilon L \leq a^* \leq b^* \leq b + \varepsilon L \,\middle|\, \rho\right) = \frac{\int_0^{\delta/2}\int_0^{\delta/2} P(x, y; L, \alpha, \beta, k)\,\mathrm{d}x\,\mathrm{d}y}{\int_0^{a+(1-b)}\int_0^{a+(1-b)-y} P(x, y; L, \alpha, \beta, k)\,\mathrm{d}x\,\mathrm{d}y}$$

$$P(x, y; L, \alpha, \beta, k) = \left(\frac{(L+x)^\alpha {}_2F_1(\alpha, 1-\beta, \alpha+1, \frac{L+x}{L+x+y}) - x^\alpha {}_2F_1(\alpha, 1-\beta, \alpha+1, \frac{x}{L+x+y})}{\alpha(L+x+y)^\alpha B(\alpha, \beta)}\right)^k$$

  - ${}_2F_1$ is the ordinary Gaussian hypergeometric function and *B* denotes a beta function
  - Remember [a,b] ⊆ [a*,b*]
  - k=20–30 runs already fairly good approximation

- **Corollary 3** *If $a = b$, i.e., all k runs in ρ are equal, then* $P(a^* = b^* | \rho) = 1 + 9/3^k - 8/2^k$
  - 9 equal runs => identifiable at 99% confidence

An application in palliative care

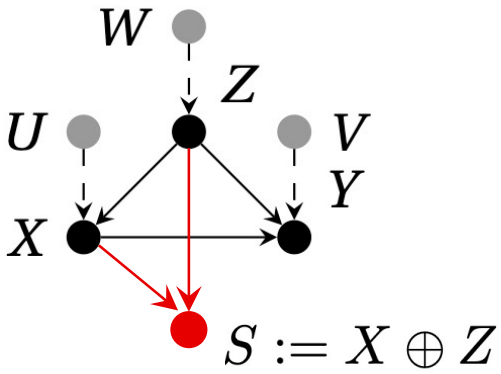


- Data from 116 patients about all the Boolean variables in the network => X_1, …, X_{12}
- No latent confounders => structural causal model is Markovian => U_i -> X_i (i=1, …,12)
  - Use the conservative specification
- Compute *PNS* := $P(Death_{X=yes} = \text{home}, Death_{X=no} = \text{hospital})$ w.r.t. the controllable X's
- EMCC: Triangolo 27−35%; family's awareness 4−11%; patient's awareness 3−11%
  - By the very Triangolo variable we can change the fate for ~30% of patients

EMCC extended to selection bias

- Consider (*X*,*Y*,*Z*) = (Treatment, Outcome, Gender) with SCM and data as shown

Treatment (X)	Recovery (Y)	Gender (Z)	#
0	0	0	2
1	0	0	41
0	1	0	114
1	1	0	313
0	0	1	107
1	0	1	109
0	1	1	13
1	1	1	1



- Treated males (*X*=1,*Z*=1) and untreated females (*X*=0,*Z*=0) systematically not reported
  - Case of **selection bias**
  - Can we still say something about the overall population?

EMCC extended to selection bias

- Consider (*X*,*Y*,*Z*) = (Treatment, Outcome, Gender) with SCM and data as shown

$\mathcal{D}$

Z	X	Y	S	#
0	0	0	0	2
0	0	1	0	114
0	1	0	1	41
0	1	1	1	313
1	0	0	1	107
1	0	1	1	13
1	1	0	0	109
1	1	1	0	1

$\mathcal{D}_1$

W	U	V	Z	X	Y	S	#
*	*	*	0	1	0	1	41
*	*	*	0	1	1	1	313
*	*	*	1	0	0	1	107
*	*	*	1	0	1	1	13

$\mathcal{D}_0$

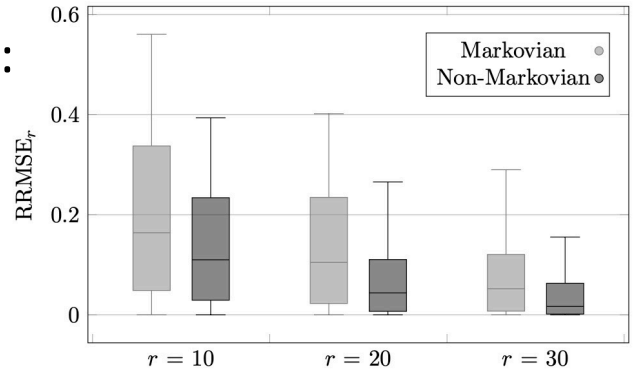
W	U	V	Z	X	Y	S	#
*	*	*	*	*	*	0	226

- EMCC is applied as before with the only difference that the iteration becomes

$$P_{t+1}(U) \leftarrow \frac{d_0 P_t(U|S=0) + \sum_{\boldsymbol{x}\in\mathcal{D}_1} P_t(U|\boldsymbol{x})}{(d_0 + d_1)}$$

- Main results hold as before:
  - Thm. 1: The (log-)likelihood is unimodal
  - Cor. 1: The global optimum points are IFF $P(U_1)$, …, $P(U_m)$ in their resp. credal sets

Experiments:



Conclusions

- The EMCC is based on simple ideas and tools
  - It should not be too difficult to join it to other models
  - Or to extend it to the continuous case
  - It might lead to `simple', while general, ways to join causal inference with machine learning

- It delivers guaranteed inner approximations
  - Outer ones are safer but tend to be more difficult to yield without becoming loose
  - Yet EMCC is anytime and can easily be made parallel
  - And we can yield credible intervals to increase safety with some guarantee

- More work is certainly needed on all these fronts

For now, we have automated counterfactual computation also under selection bias