Measuring a causal effect on a network

Antoine Chambaz

Modal'X
(Université Paris-Ouest Nanterre)

April 11th, 2014

2014 ENBIS-SFdS Spring Meeting

Graphical causality models: trees, Bayesian networks and big data

Networks and causality

Setting:

- each observation is a network with a context
- observations are i.i.d.
- an underlying causal model is at play

Conceptual examples:

- type A or type B breast cancer
  - each woman has her personal history and characteristics
  - a genomic study of sample tissues yields a gene-interaction network

  $\rightarrow$ what is the "effect" of type on network?
- randomized trial of low-calory diet in obese women
  - each woman has her personal history and characteristics
  - one flips a coin to assign either low-calory diet or pseudo-diet
  - after a few months, a genomic study of blood samples yields a gene-interaction network

  $\rightarrow$ what is the "effect" of diet on network?
- many more... (possibly with continuous "actions")

Formalization

Nested settings:

- statistical model: $O^1, \ldots, O^n \overset{\text{iid}}{\sim} P_0 \in \mathcal{M}$

  $O = (W, A, Y)$: context W, action A, binary network Y **(fixed set of G edges)**
- causal model: $X = (W, A, Y_0, Y_1) \sim \mathbb{P}_0 \in \mathbb{M}$
  - (Y_0, Y_1, A) mutually conditionally independent given W
  - defining $Y \equiv Y_A$ yields $O = (W, A, Y) \sim P_0$

    $\rightarrow P_0$ is a deterministic function of $\mathbb{P}_0$

Parameters:

- causal parameter: say $\Phi : \mathbb{M} \rightarrow [0; 1]$,

$$\Phi(\mathbb{P}) = E_{\mathbb{P}}\{\|Y_1 - Y_0\|^2\}$$

  - notation: $\langle u, v \rangle = \frac{1}{G(G-1)} \sum_{1 \leq k < l \leq G} u_{kl} v_{kl}$, $\|u\|^2 = \langle u, u \rangle$
- *associated* statistical parameter: $\Psi : \mathcal{M} \rightarrow [0; 1]$,

$$\Psi(P) = E_P\{\langle \theta_P(1, W) + \theta_P(0, W) - 2\theta_P(1, W)\theta_P(0, W), \mathbf{1} \rangle\}$$

  - notation: $\theta_P(A, W) = E_P(Y|A, W)$, $g_P(W) = P(A = 1|W)$

identifiability: $\Psi(P_0) = \Phi(\mathbb{P}_0)$!

Analysis of Ψ

Ψ is pathwise-differentiable at every $P \in \mathcal{M}$ wrt $L_0^2(P)$

its efficient influence curve satisfies $\nabla_P \Psi(O) = \langle \overrightarrow{\nabla_P \Psi}(O), \mathbf{1} \rangle$, where

$$\begin{aligned}\overrightarrow{\nabla_P \Psi}(O) \quad = \quad & [\theta_P(1, W) + \theta_P(0, W) - 2\theta_P(1, W) * \theta_P(0, W) - \Psi(P)] \\ & + \frac{A}{g_P(W)}(Y - \theta_P(1, W)) * (1 - 2\theta_P(0, W)) \\ & + \frac{1 - A}{1 - g_P(W)}(Y - \theta_P(0, W)) * (1 - 2\theta_P(1, W))\end{aligned}$$

$\nabla_P \Psi$ conveys crucial information on Ψ, *e.g.*

- Cramér-Rao bound:

  $\mathrm{Var}_P(\nabla_P \Psi(O))$ is the smallest asymptotic variance of a regular estimator of $\Psi(P)$ under P

- robustness:

  if $E_P\{\nabla_{P'} \Psi(O)\} = 0$ and either $g_P = g_{P'}$ or $\theta_P = \theta_{P'}$ then $\Psi(P') = \Psi(P)$

$\nabla_P \Psi$ can drive the elaboration of a **targeted** SP inference procedure

Targeted SP inference procedure

Targeted minimum loss estimation (TMLE):

- coined by van der Laan and Rubin [2006]
  studied, refined/applied in/to many different settings and problems [van der Laan & Rose, 2011]
- SP-based methodology
  parenthood with *estimating equations methodology*, *Huber one-step inference*
  huge body of literature [Robins, 1980-. . . ; Bickel et al., 1993; van der Vaart, 1998]

Roadmap:

1. *initialization:*
   - estimate θ_{P_0}, g_{P_0}, and $P_{0,W}$ (with θ_n^0, g_n^0 and empirical measure)
   - choose any P_n^0 such that $\theta_{P_n^0} = \theta_n^0$, $g_{P_n^0} = g_n^0$ and $P_{n,W}^0 = P_{n,W}$
   - set $k \leftarrow 0$
2. *iterative updates:* while criterion not met, repeat
   - 2.1 define $\frac{dP_n^k(\varepsilon)}{dP_n^k} = 1 + \langle \nabla_{P_n^k} \Psi, \varepsilon \rangle$
   - 2.2 compute MLE of ε: $\varepsilon_n^k = \arg\max_\varepsilon P_n \log P_n^k(\varepsilon)$
   - 2.3 set $P_n^{k+1} = P_n^k(\varepsilon_n^k)$ and $k \leftarrow k + 1$
3. at final step $k = K_n$, define TMLE $\psi_n^\star = \Psi(P_n^{K_n})$

Typical theoretical results

Consistency: under empirical processes typical conditions

- if either $\theta_{P_n^{K_n}}$ or $g_{P_n^{K_n}}$ converges to the truth then $\psi_n^\star$ consistently estimates $\Psi(P_0)$

Central limit theorem: (same as above)

- if either $\theta_{P_n^{K_n}}$ or $g_{P_n^{K_n}}$ converges to the truth
  and *if the product of errors is* $O_P(1/\sqrt{n})$ then ψ_n^* satisifes a CLT
- if, in addition, the estimation of g_{P_0} relies on valid parametric model
  then one can conservatively estimate the asymptotic variance of ψ_n^*

MovieLens

Acknowledgement: GroupLens Research (University of Minnesota), public data set online

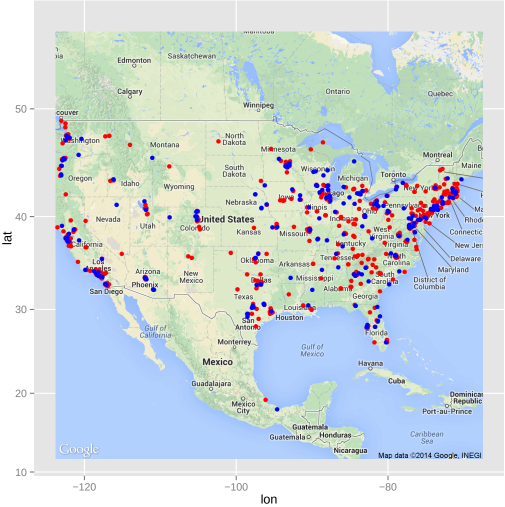
MovieLens: (from the website)

> "**MovieLens** is a movie recommendation website. It uses your ratings to generate personalized recommendations for other movies you will like and dislike, based on . . ."

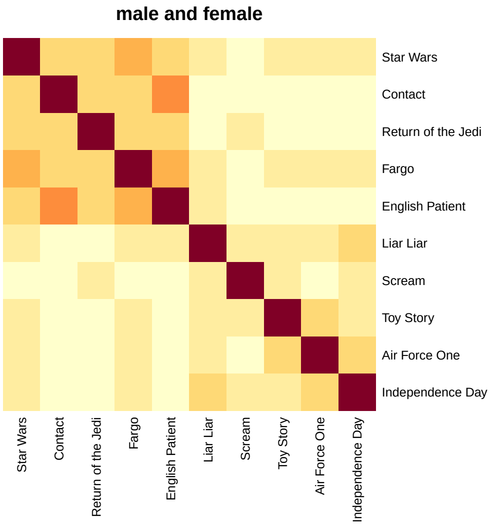
Building a data set:

- W, user information (age, occupation, location)
- A, gender – **why not, this is a toy example**
- Y
  - restriction to the 10 most rated movies
    $Y \in \mathbb{R}^{45}$, $Y_{ij} = 1$ iff movies i and j equally rated by user
    $\rightarrow$ I create a rate 0 ("unrated") so that the network is always well-defined
  - sample size $n = 904$

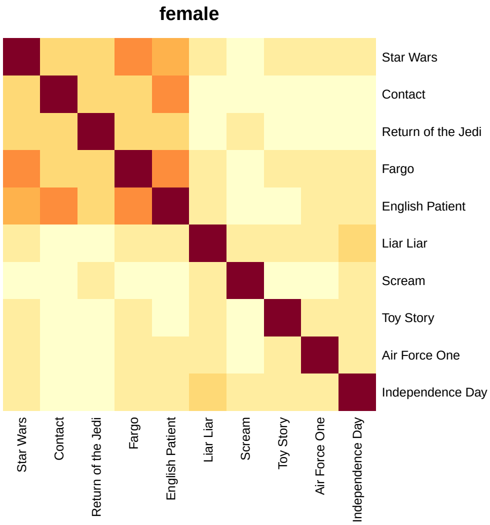
MovieLens: whereabouts and gender of users



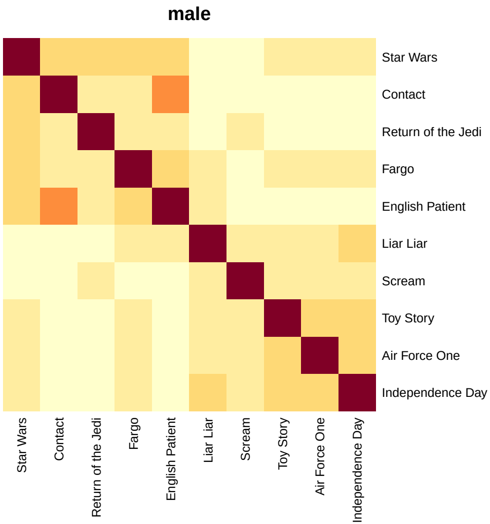
MovieLens: $E_{P_n}\{Y\}$



MovieLens: $E_{P_n}(Y|A=0)$



MovieLens: $E_{P_n}(Y|A=1)$



Method

- initialization based on *stacked logistic regressions*
- stopping criterion: two successive L^2-norms of ε_n^k smaller than a threshold
- brief summary:
  - $\|\varepsilon_n^k\|_2^2$: 0.13591, 0.00205, 0.00231
  - ψ_n^k: 0.193, 0.193, 0.197, 0.196

Discussion

Computational/theoretical challenges...

- cope with
  - larger network, large-dimensional contexts
  - strong dependence structure in conditional law of network given action and context
- elaborate third-order expansions of $\psi_n^\star - \Psi(P_0)$ to derive CLTs that are valid under milder assumptions
  (already done in simpler frameworks, see [van der Laan, 2014])
- extend to time-evolving networks (still on fixed set of edges),
  e.g. by relying on an auto-regressive **working** model

Real-life relevant problems, because

- that would help to come up with a nice simulation scheme
- that always gives insight
- that matters