From the probabilistic marginal problem to the causal marginal problem (beyond conditional independences)

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany
Department Empirical Inference

29 June 2016

(probabilistic) marginal problem

Let $S := \{X_1, \ldots, X_d\}$ be random variables and $S_j \subset S$

Question:

Given the marginal distributions P_{S_j}, is there a joint distribution $P_{X_1,\ldots,X_d}$ having P_{S_j} as marginal distributions?

Vorobev: Consistent families of measures and their extensions (1962), Kellerer: Maßtheoretische Marginalprobleme (1964)

Simple negative example

- $P_{X,Y}$ such that $cor(X, Y) \approx 1$
- $P_{Y,Z}$ such that $cor(Y, Z) \approx 1$
- $P_{X,Z}$ such that $cor(X, Z) \approx -1$

$P_{X,Y,Z}$ cannot exist because covariance matrix would not be positive

Causal marginal problem

(vague formulation)

Given distributions P_{S_j} with causal model M_j, is there a distribution $P_{X_1,\ldots,X_d}$ with causal model M whose marginals on S_j coincide with M_j?

Vague because

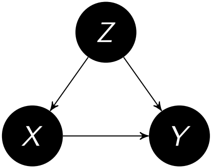
- 'causal model' not defined (DAG, MAG, PAG, structural equation model, semi-Markovian model, potential outcome, chain graph...)

- definition of 'marginalization of a causal model' not obvious

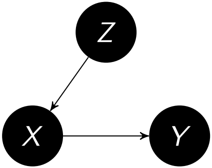
Note: marginal problems with focus on conditional independences have been considered elsewhere, e.g. Tsamardinos, Triantafillou, Lagani: Towards Integrative Causal Analysis of Heterogeneous Data Sets and Studies, JMLR 2012

Examples instead of definitions...

$Z \rightarrow Y$ can be the marginalization of

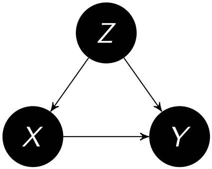


or



DAGs not closed under marginalization

usually, marginalization of



to X, Y is no longer a DAG because X, Y are confounded

then, marginal is a semi-Markovian model, for instance

Defining marginalization of a causal model

(for a special case)

Definition:

Given a causal DAG with nodes $X_1, \dots, X_d$ and joint distribution $P_{X_1,\dots,X_d}$. The DAG $X_i \to X_j$ is said to be a valid marginalization if $p(x_j|do(x_i)) = p(x_j|x_i)$.

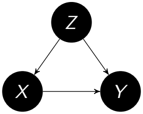
Note: here the trivial case $p(x_j|do(x_i)) = p(x_j) = p(x_j|do(x_i))$ is also allowed

Example for a causal marginal problem

Given:

- $P_{X,Y}$ with causal DAG $X \to Y$
- $P_{Z,X}$ with causal DAG $Z \to X$
- $P_{Z,Y}$ wih causal DAG $Z \to Y$

can they be obtained from marginalizing some $P_{X,Y,Z}$ with DAG

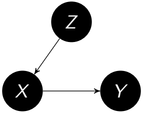


?

1) a natural solution

Let $P_{X,Y}, P_{Z,X}, P_{Z,Y}$ be Gaussians such that their covariance matrices $\Sigma_{XY}, \Sigma_{ZY}, \Sigma_{Z,X}$ yield a positive matrix $\Sigma_{X,Y,Z}$ (solving the probabilistic marginal problem)

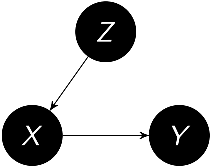
If the partial correlation $\rho_{XY|Z}$ vanishes (can be checked from pairwise covariances alone), the DAG



with Gaussian $P_{X,Y,Z}$ solves the causal marginal problem

Generalization to non-Gaussians?

The DAG



is only possible if

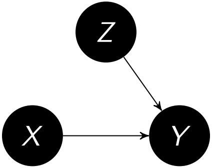
$$I(Z : Y) \leq \min\{I(X : Y), I(Z : X)\}$$

(data processing inequality, follows from $I(Z : Y|X) = 0$)

necessary condition, I'm pretty sure it's not sufficient

2) a boring solution

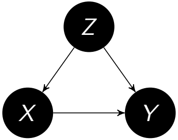
If $P_{ZX} = P_Z P_X$ the DAG



solves our causal marginal problem

3) a paradox solution

'improper confounding'



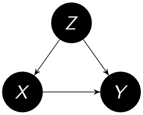
- let Z contain 2 independent bits, one influencing X and one influencing Y.
- relabel the states $Z = 00, 01, 10, 11$ as $Z = 1, 2, 3, 4$ to make the example more serious.
- clearly $p(y|do(x)) = p(y|x)$.

Z is a confounder in the sense that it really influences both X and Y, but it does not really *confound* the influence of X on Y

Inconsistent causal marginals

Let $P_{X,Y}, P_{Z,X}, P_{Z,Y}$ be Gaussians such that their covariance matrices $\Sigma_{XY}, \Sigma_{ZY}, \Sigma_{Z,X}$ yield a positive matrix $\Sigma_{X,Y,Z}$ (solving the probabilistic marginal problem)

If the partial correlation $\rho_{XY|Z}$ does not vanish, the DAG



with Gaussian $P_{X,Y,Z}$ does *not* solve the causal marginal problem because $p(y|do(x)) \neq p(y|x)$

Simple marginal problem

- let X, Y, Z be binary
- define $P_{X,Y}$ by

$$p(x,y) = \begin{cases} \frac{1}{2} - \epsilon & \text{for } x = y \\ \epsilon & \text{for } x \neq y \end{cases}$$

- $p(y,z)$ and $p(x,z)$ defined similarly with the same ϵ
- causal models $Z \to X$, $X \to Y$, and $Z \to Y$

Conjecture: inconsistent marginals, at least for small ϵ

Question:

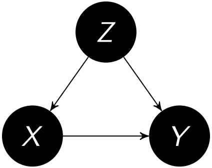
What are necessary and sufficient conditions in the above tripartite case?

Structural causal marginal problem

Given:

- $P_{X,Y}$ with structural equation $Y = f_Y(X, N_Y)$
- $P_{Z,X}$ with structural equation $X = f_X(Z, N_X)$
- $P_{Z,Y}$ with structural equation $Y = f_Y(Z, N_Y)$

Are there structural equations for the DAG



that are consistent with the above (even regarding counterfactual statements)?

Example with binaries

1. $Y = X \oplus N_X$ with unbiased N_X for $X \to Y$ (hence $X \perp\!\!\!\perp Y$)
2. $X = Z \oplus N_Z$ with unbiased N_Z for $Z \to X$ (hence $Z \perp\!\!\!\perp X$)
3. $Y = N_Y$ with unbiased N_Y for $Z \to Y$ (hence $Z \perp\!\!\!\perp Y$)

counterfactual statements inconsistent:

- for fixed values of N_X, N_Y, N_Z, inverting Z has no effect on Y according to (3)
- according to (1) and (2), inverting Z inverts Y

Applications of the causal marginal problems

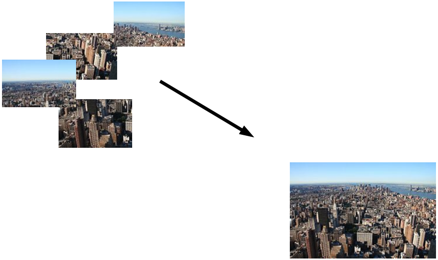
Given observations from different data sets with overlapping variables, do they correspond to the same background conditions?

- if the marginal problem is not solvable they correspond to different conditions
- causal information can render the extension of marginals to a joint unique or *less ambiguous*

causal information helps in putting pieces of the world together

Science as a puzzle:

putting pieces together without causal information:



Science as a puzzle...

with causal information: