Causal machine learning

Wei Pan

Division of Biostatistics and Health Data Science, School of Public Health,
University of Minnesota, Minneapolis, MN 55455
Email: panxx014@umn.edu

PubH 8475/Stat 8056

Outline

- Causal inference in the presence of hidden confounding
  - Instrumental variable (IV) regression: 2SLS
  - Mendelian randomization (MR)
  - Network deconvolution (ND)
  - DeepIV
- Causal inference without confounding
  - Counterfactual model
  - Standard approaches
  - New (ML) approaches
  - Causal trees and forests

Chocolate consumption vs Nobel prize winning: Messerli 2012, NEJM

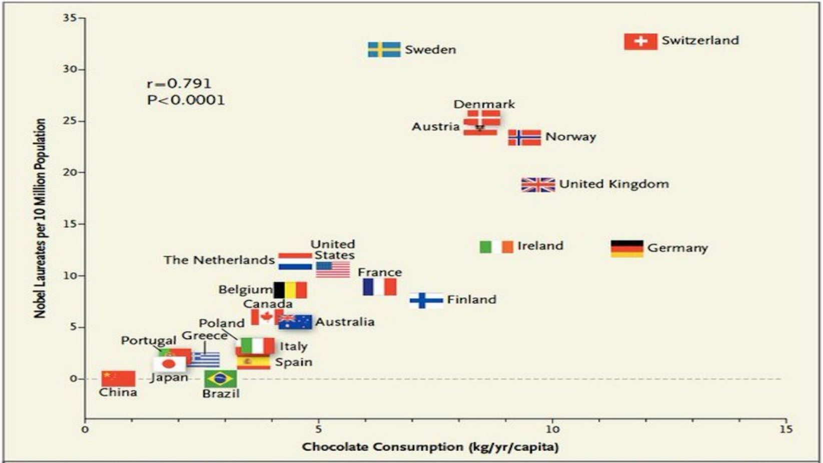


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Chocolate consumption vs Nobel prize winning: flavanoids?

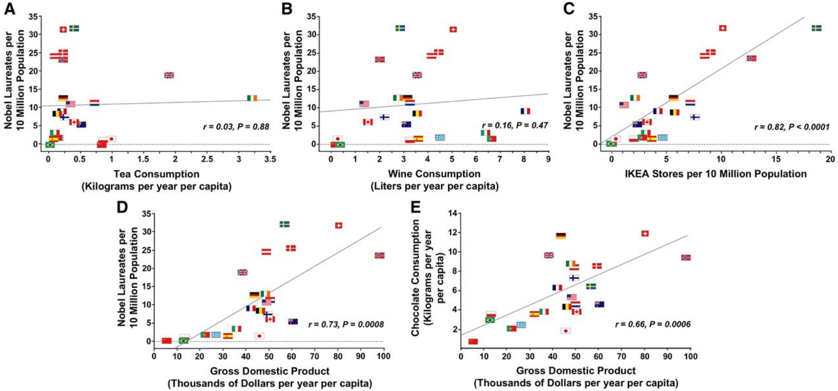


Figure: Fig 1 in Maurage et al 2013, *J Nutrition*.

Introduction: motivation

Adam, D. (2019). The gene-based hack that is revolutionizing epidemiology: Mendelian randomization offers a simple way to distinguish causation from correlation. But are scientists overusing it? *Nature*, 576:196-199.

- Limitations of epi/observational/association studies $\Longrightarrow$
- Failures of multiple $100-million trials!
- Causal inference!
- MR: causal inf with observational data
- Easy to use: wide availability of GWAS summary data
- Magic? No! Strong modeling assumptions...
- Boef (2015, *IJE*): 178 published, $< 1/2$ "adequately discussed these assumptions"
- *"Statistical tools for epidemiology are improving. And although Mendelian randomization does not always offer perfect clarity, it might, at least, point researchers in the right direction."*

Introduction: IV reg and MR

- MR: a special application of instrumental variable (IV) reg.
  2S-2SLS
  using (indep) genetic variants/SNPs as IVs
  GWAS summary data
  Bowden, Burgess, Davey Smith, ...;
- IV reg for causal inference
  Angrist, Imbens won a half of the 2021 Nobel Prize in Economics:
  Angrist, J.D. and G.W. Imbens (1995). "Two-stage least squares estimation of average causal effect in models with variable treatment intensity." *JASA*, 90(430): 431-442.
  Angrist, J.D., G.W. Imbens, and D.B. Rubin (1996). "Identification of causal effects using instrumental variables." *JASA*, 91: 444-472.
  Imbens, G.W. and J.D. Angrist (1994). "Identification and estimation of local average treatment effects." *Econometrica*, 61: 467-476.

IV reg: Basic Idea

- True **causal** model:

$$X = Z\beta_X + U\beta_{XU} + \epsilon_X^*, \qquad Y = X\theta + U\beta_{YU} + \epsilon_Y^*.$$

- θ: parameter of interest; e.g., H_0: $\theta = 0$.
- Key challenge: **hidden** confounder U
  $\Longrightarrow \hat{\theta}$ **biased** in $Y \sim X$.
  Why?
- 2-Stage Least Square (2SLS): under 3 valid IV assumptions,

$$E(Y|Z) = \theta E(X|Z) \Longrightarrow$$

  Stage 1: $\hat{X} = Z\hat{\beta}_X$,
  Stage 2: $Y = \hat{X}\theta + \epsilon_Y$.
  Cnnsistent and AN (but low efficiency)!
- a Key feature: 2-sample 2SLS,
  can infer θ with two independent samples $\{(Z_i, X_i)\}$'s and $\{(Z_i, Y_i)\}$'s!

MR: Basic Idea

- MR: consider one IV,

$$X = Z_i \beta_{Xi} + \epsilon_X, \qquad Y = X\theta + \epsilon_Y^{**} = Z_i \beta_{Yi} + \epsilon_Y$$

  Key: $\beta_{Yi} = \beta_{Xi}\theta$

- MR: under the 3 valid IV assumptions, $\hat{\theta} = \hat{\beta}_{Yi}/\hat{\beta}_{Xi}$ unbiased, consistent, AN, ...
- $\hat{\beta}_{Yi}, \hat{\beta}_{Xi}$ (and $\hat{\sigma}^2_{Yi}$, $\hat{\sigma}^2_{Xi}$) directly available from two indep GWAS summary datasets.
- If multiple indept IVs, combine by meta-analysis: IVW(FE), ...

Assumptions

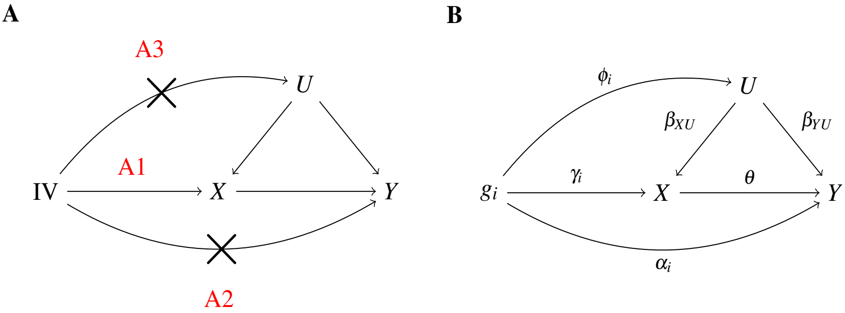


Figure: (A) Three assumptions for valid IVs. (B) Our causal model.

- Violation of A2: uncorrelated pleiotropy; $\beta_{Xi} = \gamma_i, \alpha_i$ uncor.
- Violation of A3: correlated pleiotropy; $\beta_{Xi} = \gamma_i + \phi_i\beta_{XU}, \alpha_i + \phi_i\beta_{YU}$ correlated $\Longrightarrow$ violation of InSIDE required by MR-Egger, IVW(RE), RAPS (treating α_i random).

MR vs RCT

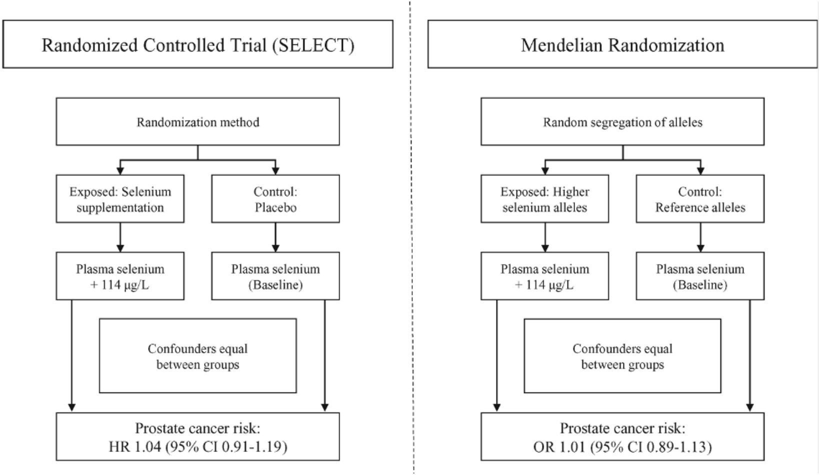

Randomized Controlled Trial (SELECT)
Randomization method
Exposed: Selenium supplementation
Control: Placebo
Plasma selenium + 114 μg/L
Plasma selenium (Baseline)
Confounders equal between groups
Prostate cancer risk: HR 1.04 (95% CI 0.91-1.19)
Mendelian Randomization
Random segregation of alleles
Exposed: Higher selenium alleles
Control: Reference alleles
Plasma selenium + 114 μg/L
Plasma selenium (Baseline)
Confounders equal between groups
Prostate cancer risk: OR 1.01 (95% CI 0.89-1.13)

Figure: Yarmolinsky, James, et al. "Circulating selenium and prostate cancer risk: a Mendelian randomization analysis." JNCI: Journal of the National Cancer Institute 110.9 (2018): 1035-1038.

UVMR-cML (Xue, Shen and Pan (2021, AJHG)

- Key: relax Assumptions (A1), A2 & A3.
- A more general (true causal) model:

$$\begin{aligned}\beta_{Xi} &= \gamma_i + \beta_{XU} \cdot \phi_i, \\ \beta_{Yi} &= \theta \cdot (\gamma_i + \beta_{XU} \cdot \phi_i) + \alpha_i + \beta_{YU} \cdot \phi_i = \theta \cdot \beta_{Xi} + r_i,\end{aligned}$$

- From the GWAS data: $\hat{\beta}_{Xi} \sim N(\beta_{Xi}, \hat{\sigma}^2_{Xi})$ and $\hat{\beta}_{Yi} \sim N(\beta_{Yi}, \hat{\sigma}^2_{Yi})$ for $i = 1, \cdots, m$. All indep
- Log-likelihood:

$$L = -\frac{1}{2} \sum_{i=1}^{m} \left(\frac{(\hat{\beta}_{Xi} - \beta_{Xi})^2}{\hat{\sigma}^2_{Xi}} + \frac{(\hat{\beta}_{Yi} - \theta \cdot \beta_{Xi} - r_i)^2}{\hat{\sigma}^2_{Yi}} \right),$$

- Constrained maximum likelihood (cML):

$$\max L \text{ subject to } \sum_{i=1}^{m} I(r_i \neq 0) = K.$$

- Try $K = 0, 1, 2, ..., m-2$, then use BIC to select $\hat{K}$.
- A sequential algorithm: fast but ...

Theory

- Assumption 1: (Plurality valid condition.) Suppose that B_0 is the index set of the true valid IVs with $K_0 = |B_0|$. For any $B \subseteq \{1, \cdots, m\}$ and $|B| = K_0$, if $B \neq B_0$, then the K_0 ratios $\{\beta_{Yi}/\beta_{Xi}, i \in B\}$ are not all equal.
- Note: valid IVs: $\beta_{Yi}/\beta_{Xi} = \theta$;
  invalid IVs: $\beta_{Yi}/\beta_{Xi} = \theta + r_i/\beta_{Xi} \neq \theta$.
- Assumption 2: (Orders of the variances and sample sizes.) There exist positive constants l_X, l_Y, l_N and u_X, u_Y, u_N such that we have $l_X/N_1 \leq \hat{\sigma}^2_{Xi} \leq u_X/N_1$, $l_Y/N_2 \leq \hat{\sigma}^2_{Yi} \leq u_Y/N_2$, and $l_N \cdot N_2 \leq N_1 \leq u_N \cdot N_2$ for $i = 1, \cdots, m$.
- Note: usually satisfied, e.g. with LSE or MLE.

Theory

- Theorem 1: (Selection consistency.) With Assumptions 1 and 2 satisfied, if $K_0 \in \mathcal{K}$, we have $P(\hat{K} = K_0) \to 1$ and $P(\hat{B}_{\hat{K}} = B_0) \to 1$ as N_1, $N_2 \to \infty$.
- Theorem 2. (Consistency and AN.) With Assumptions 1 and 2 (and some regularity conditions), the cMLE $\hat{\theta}$ is consistent and asymptotically normal.
  Note: similar to the theory in RAPS (Zhao et al 2020, *AoS*). Only valid IVs are used.
- Allowing the presence of weak IVs (i.e. A1 violated). similar to RAPS.
- Theorem 3. (DP/bootstrap is consistent.)

Finite-sample adjustments

- Model averaging (MA) (Buckland et al 1997, *B'cs*):

$$w_K^0 = \exp\left(-\mathrm{BIC}(K)/2\right), \; w_K = w_K^0 / \sum_{K \in \mathcal{K}} w_K^0,$$

$$\hat{\theta}_w = \sum_{K \in \mathcal{K}} w_K \cdot \hat{\theta}(K), \qquad \mathrm{SE}(\hat{\theta}_w) = \ldots$$

- Data perturbation (DP): or, parametric bootstrap, $\hat{\beta}_{Xi}^{(t)} \sim N(\hat{\beta}_{Xi}, \hat{\sigma}_{Xi}^2)$ and $\hat{\beta}_{Yi}^{(t)} \sim N(\hat{\beta}_{Yi}, \hat{\sigma}_{Yi}^2)$ for $i = 1, \ldots, m$ obtain $\hat{\theta}^{(t)}(K)$ for $t = 1, 2, \ldots, T$.

$$\hat{\theta}_{DP}(K) = \frac{\sum_{t=1}^{T} \hat{\theta}^{(t)}(K)}{T}, \qquad \mathrm{SE}\left(\hat{\theta}_{DP}(K)\right) = \mathrm{SD}(\{\hat{\theta}^{(t)}(K)\}).$$

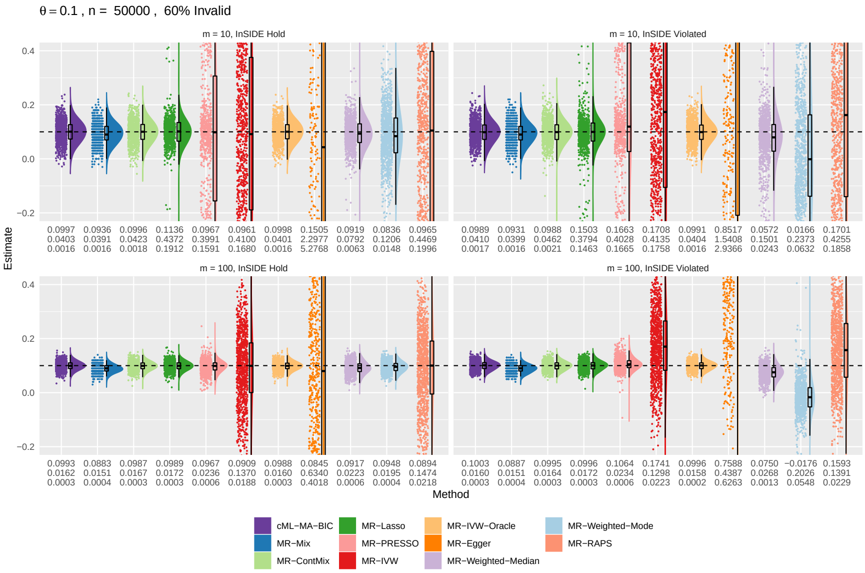
  then apply MA (optional): ......
  Bagging (Breiman 1996 *ML*; Efron 2014 *JASA*).

Simulations

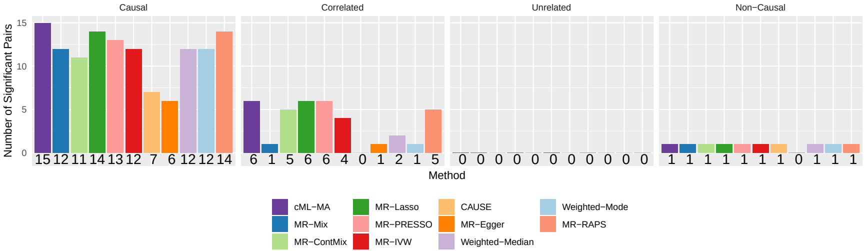
Compared with most state-of-the-art MR methods;

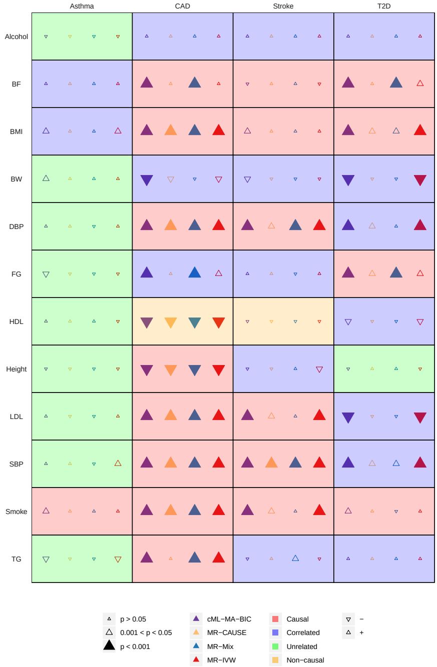
As expected, ...

Simulation results:



Applications:




Asthma
CAD
Stroke
T2D
Alcohol
BF
BMI
BW
DBP
FG
HDL
Height
LDL
SBP
Smoke
TG
p > 0.05
0.001 < p < 0.05
p < 0.001
cML−MA−BIC
MR−CAUSE
MR−Mix
MR−IVW
Causal
Correlated
Unrelated
Non−causal
−
+

Extensions/alternatives

- UVMR-cML-C: allowing overlapping samples (Lin, Xue and Pan 2023, PLOS Genet);
- MVMR-cML: allowing multiple exposures (Lin, Xue and Pan 2023, AJHG)
- Next: apply UVMR-cML-C and ND to infer (general) causal networks.

  allowing cycles; data from different and possibly overlapping samples.
- A limitation: assuming the causal direction is known.

  bi-directional MR
- Steiger's method (Hemani et al 2017, *PLoS Genet*):

  Lemma. If $Z \to X \to Y$ and no hidden confounders, then $\text{corr}(Z, X) > \text{corr}(Z, Y)$.

  With hidden confounders, it may not always hold;

  Only working for one IV.
- Xue & Pan (2020, *PLoS Genet*): extending to multiple IVs.
- Xue & Pan (2022, *PLoS Genet*): Bi-directional CD-cML (and MR-cML).

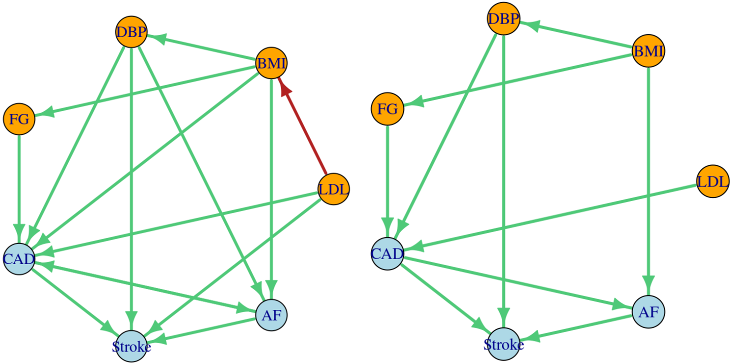
Network deconvolution (ND)

- Feizi et al (2013, Nat Biotechnol.)
- Q: given a total-causal-effect graph G_t, how to estimate the direct-causal-effect graph G_d?
- $$G_t = G_d + G_d^2 + G_d^3 + ... = G_d(I + G_d + G_d^2 + G_d^3 + ...) = G_d(I - G_d)^{-1}$$

  if $\rho(G_d) < 1$.
  Hence, $G_d = G_t(I + G_t)^{-1}$.
- MR-cMLgraph (Lin, Xue and Pan, 2023, *PLOS Genet*): use MR-cML to construct $\hat{G}_t$, then use ND to obtain $\hat{G}_d = \hat{G}_t(I + \hat{G}_t)^{-1}$.
- Theory: $vec(\hat{G}_t)$ and $vec(\hat{G}_d)$ are consistent and AN.
- Can use data perturbation for inference.

Application: BMI might be a ‘minor’ risk factor for CAD, but an indep one for AF

Figure: **Total (left) and direct (right) causal graphs**



ND: continued

- Derivation is for a directed graph; how about for an undirected graph?
- Let Σ be an invertible correlation matrix among a set of variables of interest. If $G_t = \Sigma - I$, then $G_d = I - \Omega$, where $\Omega = \Sigma^{-1}$ is the precision matrix.
  Alipanahi and Frey (2013, Nature Biotechnol).
- Lior Pachter. The network nonsense of manolis kellis, February 2014. `https://liorpachter.wordpress.com/2014/02/11/the-network-nonsense-of-manolis-kellis/`.

DeepIV

- True model:

$$X = Z\alpha + U + \epsilon_X, \qquad Y = g(X) + U + \epsilon_Y$$

- Again fitting $Y \sim X$ leads to biased estimate of $g()$ due to hidden confounding!
- $E(Y|Z) = E[g(X)|Z]$.
- DeepIV (Hartfford et al 2017, ICML): use a FNN $g_\theta(.)$,

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} [Y_i - \int g_\theta(x) dF(x|Z_i)]^2 + P(\theta; \lambda),$$

- Slow: need to use MC sampling, $\int g_\theta(x) dF(x|Z_i) \approx \sum_{j=1}^{M} g_\theta(X_{ij}), \qquad X_{ij} \sim \hat{F}(x|Z_i)$.
- Unstable: ill-posed inverse problem; Fredholm integral equation of the first kind (Newey 2013, Am Econ Rev).

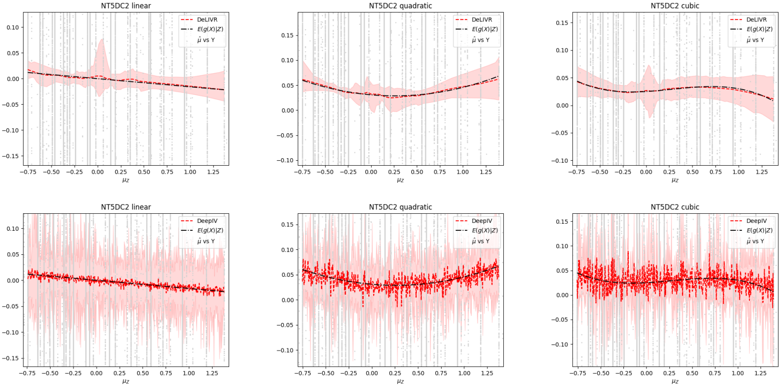
Discussion

- Alternative: DeLIVR (He et al 2023, Biostatistics).
- Several new IV deep learning methods...
- An application: causal feature extraction (Yao et al 2023, Stat in Med).

DeLIVR

- Stage 2 model: $E(Y|Z) = E[g(X)|Z)$
  Problem: estimating $g(X)$.
- **Key**: $E[g(X)|Z) = h(\mu_Z) \neq g(\mu_Z)$, $\mu_Z = E(X|Z)$.
- New: estimating $E[g(X)|Z) = h(\mu_Z)$,
  Assuming $X|Z \sim N(\mu_Z, \sigma^2)$.
- Would this address the original Q?
  Proposition. Suppose $X|Z \sim N(\mu_Z, \sigma^2)$, and $g(X)$ is a univariate function in X (and independent of μ_Z), then
  1. $g(X) = c$, a constant, if and only if $E(g(X)|Z) = c$.
  2. $g(X)$ is linear in X if and only if $E(g(X)|Z)$ is linear in μ_Z.
  3. $g(X)$ is a k-th degree polynomial in X if and only if $E(g(X)|Z)$ is a k-th degree polynomial in μ_Z.
- More generally, if $g(X)$ is locally smooth, by a Taylor expansion, ...
- DeLIVR: estimating an ANN $h_\theta(.)$ for $h(.)$.
- Inference: use independent training and inference subsamples
  ...

Simulation results: DeLIVR more stable than DeepIV



From MRI to AD prediction

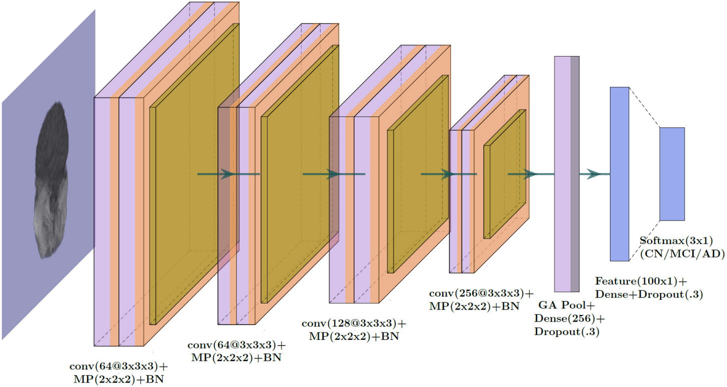


Figure: CNN.

DeepFEIVR

- Deep Feature Extraction via IV Regression (DeepFEIVR).
- X: image; Z: SNPs/IVs; Y: AD status.
- Model:

$$f(X) = ZB + U + \epsilon_X, \qquad Y = f(X)\beta + U + \epsilon_Y$$

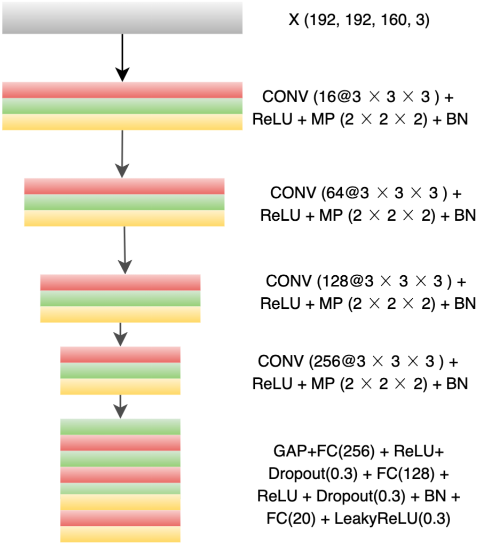
- Key H_0: $\beta = 0$.
- Key challenge: **hidden** confounder U
- 2SLS-like:

$$\hat{f}(X) = Z\hat{B}, \qquad Y = \hat{f}(X)\beta + \epsilon_Y = Z\hat{B}\beta + \epsilon_Y$$

- Contrast to existing nonparametric IV, e.g. deepIV and DeLIVR:

$$X = ZB + U + \epsilon_X, \qquad Y = f(X) + U + \epsilon_Y$$

Network architectures



X (192, 192, 160, 3)

CONV (16@3 × 3 × 3) +
ReLU + MP (2 × 2 × 2) + BN

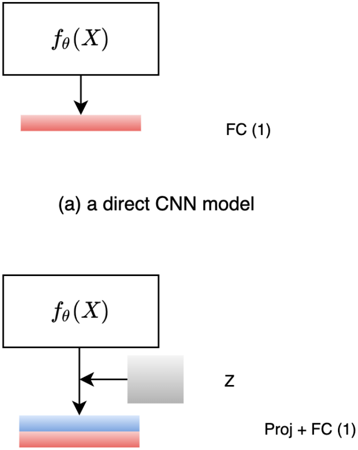
CONV (64@3 × 3 × 3) +
ReLU + MP (2 × 2 × 2) + BN

CONV (128@3 × 3 × 3) +
ReLU + MP (2 × 2 × 2) + BN

CONV (256@3 × 3 × 3) +
ReLU + MP (2 × 2 × 2) + BN

GAP+FC(256) + ReLU+
Dropout(0.3) + FC(128) +
ReLU + Dropout(0.3) + BN +
FC(20) + LeakyReLU(0.3)

f_θ



$f_\theta(X)$

FC (1)

(a) a direct CNN model

$f_\theta(X)$

Z

Proj + FC (1)

(b) DeepFEIVR

Second part: No (hidden) confounding

- Data: $D = \{(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)\}$. $T_i = 0$ or 1.
  Goal: any trt effects?
- Counterfactual model:
  $Y_i(1)$ and $Y_i(0)$ are the responses if individual i is and is not given the treatment, respectively.
  But we can NOT observe both $Y_i(1)$ and $Y_i(0)$!
- Unconfoundedness: $T_i \perp (Y_i(1), Y_i(0)) | X_i$
- individual treatment effect (ITE):

$$\begin{aligned} \tau(x) &:= E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] \\ &= E[Y_i(1)|T_i = 1, X_i = x] - E[Y_i(0)|T_i = 0, X_i = x] \\ &= E[Y_i|T_i = 1, X_i = x] - E[Y_i|T_i = 0, X_i = x], \end{aligned}$$

- average treatment effect (ATE):
  $\tau := E[Y_i(1) - Y_i(0)] = E[\tau(X)]$.
  Note: $E[\bar{Y}(T=1) - \bar{Y}(T=0)] \neq \tau$ in general; why?

Standard approaches

- Old(?) approach: regression!

$$E[Y|T = t, X = x] = t\theta + x'\beta,$$

  which can be fitted using data
  $D = \{(X_i, T_i, Y_i) : i = 1, ..., n\}$.
  Why reasonable? no hidden confounding!
- But it requires ... especially for high-dim data.
- Most popular alternative: Propensity Scores (PS)
  $PS(X_i) := Pr(T_i = 1|X_i)$.
- Rosenbaum and Rubin (1983, Biometrika):
  $T_i \perp (Y_i(1), Y_i(0))|PS(X_i)$.
- Using (X_i, T_i)'s to fit

$$\text{Logit}(Pr(T = 1|X)) = X'\alpha,$$

  $\Longrightarrow e_i := PS(X_i) = \text{Logit}^{-1}(X_i'\hat{\alpha})$.
- Often trim out observations with too small or too large e_i (i.e. outliers).

PS

- PS regression: fit

$$E[Y|T=t, X=x] = t\theta + PS(x)\gamma$$

using data D.

- PS matching:
matching each obs with $T_i = 1$ with one (or more) with $T_i = 0$ by their e_i's, then analysis on matched sets.
- PS stratification:
partitioning the data into subsets/strata based on the distribution of e_i's, then stratified analysis.
- Inverse probability weighting:
each obs is assigned a weight $w_i = 1/e_i$ if $T_i = 1$; $w_i = 1/(1-e_i)$ if $T_i = 0$; then a weighted analysis, e.g.

$$\hat{\tau} = \bar{Y}_w(T=1) - \bar{Y}_w(T=0) = \frac{\sum_{i:T_i=1} w_i Y_i}{\sum_{i:T_i=1} w_i} - \frac{\sum_{i:T_i=0} w_i Y_i}{\sum_{i:T_i=0} w_i}.$$

- But ...

New approaches

- Dorie et al (2019). Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Stat Sci*.
- Simulated data; no hidden confounders,..., as for PS.
  Standard ones: both PS and (regression) mean response modeled by GLMs;
  how about by ML?
- Five competition winners:
  - BART;
  - Superlearner + Targeted MLE: ensemble of glm, gbm, gam, glmnet and splines;
  - calCause: RF or GP by CV;
  - h2o.ai: ensemble of glm, RF, DL (NN), LASSO and ridge reg;
  - GBM + MDIA.

Counterfactual RF

- Lu et al (2018). Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. *JCGS*.
- M1: C-RF: build two RFs, $\hat{f}_1(X)$ and $\hat{f}_0(X)$, using the subsamples of $T_i = 1$ and $T_i = 0$ respectively; then for each $X_i = x \in D$, run

  $$\hat{\tau}(x) = \hat{f}_1(x) - \hat{f}_0(x).$$

  better to use the OOB estimate...
- Or, M2: $\hat{\tau}(x) = RF(x, 1) - RF(x, 0)$, where $RF(X, T)$ is built using all data (X_i, T_i, Y_i)'s.
  Model/assumption: $Y_i = f(T_i, X_i) + \epsilon_i$,
  In contrast to M1: $Y_i = f_t(X_i) + \epsilon_i$ for $T_i = t$.
- Or, M3: $\hat{\tau}(x) = RF(x, 1) - RF(x, 0)$, where $RF(X, T)$ is built using all data $(X_i, T_i, X_i * T_i, Y_i)$'s.
- In analogy, in linear reg:
  M1: $Y_i = X_i'\beta_0 + \epsilon_i$ for $T_i = 0$; $Y_i = X_i'\beta_1 + \epsilon_i$ for $T_i = 1$.
  M2: $Y_i = T_i\theta + X_i'\beta + \epsilon_i$.
  M3: $Y_i = T_i\theta + X_i'\beta + (X_i * T_i)\delta + \epsilon_i$.

Causal trees

- Ref: Athey and Imbens (2016). Recursive partitioning for heterogeneous causal effects. PNAS.
- Goal: partition the data into different subpopulations each with a (alomost) homogeneous treatment effect.
- Key idea: similar to CART, but do "honest" estimation: using two independent data subsets for partitioning and parameter estimation.
  1. Use an independent D^{est}, instead of D^{tr}, to estimate leaf means;
  2. Modify the splitting (and CV) criterion to have an unbiased MSE estimator for the causal treatment effect;
     "fundamental problem of causal inference": the causal effect is not observed.
  3. Account for increasing variance with tree growing.
- Use another independent sample for inference.
- Causal forests (Athey and Wager 2019).

Review: CART for regression

- Y: continuous.
- Key: 1) determin splitting variables and split points (e.g. $x_j < t_j$); $\Longrightarrow R_1, R_2, ...$;
  2) determine c_m in each R_m.
- in 1), use a sequential or greedy searchfor each j and s: find $x_j < s$ s.t.
  $R_1(j,s) = \{x|x_j < s\}$, $R_2(j,s) = \{x|x_j \geq s\}$,
  $\min_{j,s}[\min_{c_1} \sum_{X_i \in R_1(j,s)} (Y_i - c_1)^2 + \min_{c_2} \sum_{X_i \in R_2(j,s)} (Y_i - c_2)^2]$.
- in 2), given R_1 and R_2,
  $\hat{c}_k = \text{Ave}(Y_i|X_i \in R_k\}$ for $k = 1, 2$.
- Repeat the process on R_1 and R_2 respectively, ...