Bayesian Nonparametric Methods for Causal Inference

Jason Roy
Associate Professor of Biostatistics

Co-Director, Center for Causal Inference

jaroy@upenn.edu

February 23, 2017

Center for
Causal Inference
C→C→I

Perelman
School of Medicine
UNIVERSITY of PENNSYLVANIA

Observational data and the g-formula

Center for
Causal Inference
C → C → I

Perelman
School of Medicine
UNIVERSITY *of* PENNSYLVANIA

Potential outcomes

Exposure: A (continuous or discrete)

Potential outcomes: Y^a

- Outcome if exposure A set to level a

Causal effects

Possible causal (target) parameters (choose one):

- $E(Y^1 - Y^0)$
- $E(Y^1)/E(Y^0)$
- $E(Y^1 - Y^0|V = v)$
- $E(Y^1 - Y^0|A = 1)$
- $F_1^{-1}(p) - F_0^{-1}(p)$

Note: The target parameter might not be a parameter in the model

Linking potential outcomes to observed data

So far we have only talked about hypothetical data (outcomes that would be observed under various exposure levels).

In observational studies, we do not get to set exposure levels.

Instead, we link potential outcomes and observed data by:

- making causal assumptions
- collecting the right kind of data to make the assumptions as plausible as possible (design)

Observed Data

- Outcome: Y
- Exposure: A
- Confounders: L
- Observed data: $\{Y_i, A_i, L_i; i = 1, \cdots, n\}$

Causal assumptions

→ Consistency: if $A = a$ then $Y = Y^a$

→ Positivity: $p(A = a|L) > 0$ if $p(L) > 0$

→ Ignorability: $Y^a \perp\!\!\!\perp A|L$

These 3 assumptions imply

$$E(Y|A = a, L) = E(Y^a|A = a, L) = E(Y^a|L)$$

Causal effects

We can therefore identify various target parameters from $p(Y, A, L)$. For example for average causal effect:

$$\begin{aligned} E(Y^1) - E(Y^0) &= E\{E(Y|A=1,L)\} - E\{E(Y|A=0,L)\} \\ &= \int E(Y|A=1,L)dF(L) - \int E(Y|A=0,L)dF(L) \end{aligned}$$

or, for quantiles,

$$P(Y^a \leq y) = \int_{-\infty}^{y} \int p(Y|A=a,L)dF(L)$$

Target parameter and nuisance parameters

Suppose our target parameter ψ is $\psi = E(Y^1) - E(Y^0)$.

If we know $p(Y, A, L)$ and make causal assumptions, then we can identify ψ.

Suppose $p(Y, A, L)$ is a distribution with parameters θ, $p(Y, A, L|\theta)$.

- θ might be high-dimensional
- We do not care about θ (nuisance parameters)
- We do not care about the form of the joint distribution
- But.... we need correctly specify it (in some sense)

Introduction to Bayesian nonparametrics

Center for
Causal Inference
C→C→I

Perelman
School of Medicine
UNIVERSITY of PENNSYLVANIA

Parametric vs nonparametric

Suppose $x_1, \cdots, x_n \sim P_\theta$. The parameter space of θ is Θ.

Parametric model:

- Θ has finite dimension, e.g., $\Theta \subset \mathbb{R}^d$ where $d \in \mathbb{N}$

Nonparametric model:

- Θ has infinite dimension
- So, a nonparametric model can be thought of as a large parametric model
- Typically, the number of parameters increases with the sample size
- Bayesian nonparametric: priors reflect uncertainty about functional or distributional models (so distributions on uncertain functions or distributions)

Examples

Density estimation:

- **Parametric example**: assume data $N(\mu, \sigma^2)$. The distribution is defined with just 2 parameters. Parameter space is $\mathbb{R} \times \mathbb{R}^+$. Sample size does not affect number of parameters.
- **Nonparametric example**: assume data have distribution $\sum_{k=1}^{\infty} \pi_k N(\mu_k, \sigma_k^2)$. This is an infinite mixture of normal distributions. Prior distributions can induce clustering, reducing the dimension (number of μ's and σ's) for a given sample size n to be unknown, but at most n.

Function estimation. Suppose we have data (x_i, y_i) and we are interested in $E(y_i|x_i;\theta)$.

- **Parametric example**: let $E(y_i|x_i;\theta) = \theta_0 + \theta_1 x_i$. The parameter space is $\mathbb{R}^2$. We simply have 2 parameters to estimate and this does not increase with n. (set of all possible lines)
- **Nonparametric example**: let the parameter space be the set of all functions (possibly with some smoothness constraints)

Overview

We will consider BNP approach for both density estimation and function estimation.

Density estimation:

- We fill focus on mixture models (Dirichlet process mixtures)
- Clustering induced by prior distributions (Dirichlet process prior)
  - Sampling yields a Pólya urn
  - Distribution on the partitions is a Chinese restaruant process
  - The weights have stick breaking representation

Function estimation (later):

- Gaussian process models
- Bayesian adaptive regression trees

Dirichlet distribution

If random variables $x_1, \ldots, x_k$ have a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$, then the probability distribution is

$$\frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1},$$

where $B(\alpha)$ is the beta function

$$B(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)}.$$

- $\sum_{i=1}^{k} x_i = 1; x_i > 0; \alpha_j > 0$
- if $k = 2$ it's a beta distribution
- Using MCMCpack, x$< -$rdirichlet(k,alpha)

Conjugate prior for Multinomial distribution

Suppose $y = (y_1, \ldots, y_k)$ is a vector of counts, i.e., y_1 is the number of observations in category 1. y follows a multinomial distribution with parameters $\pi = (\pi_1, \ldots, \pi_k)$.

A conjugate prior is $p(\pi) \sim Dir(\alpha)$.

The posterior $p(\pi|y)$ is proportional to

$$\left\{ \prod_{i=1}^{k} \pi_i^{y_i} \right\} \left\{ \prod_{i=1}^{k} \pi_i^{\alpha_i - 1} \right\}$$

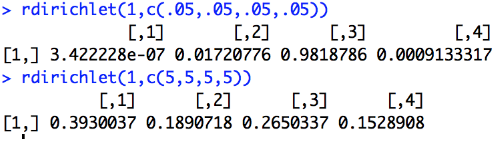
which is $Dir(\alpha + y)$.

If $p(\pi) \sim Dir(\alpha)$ then $\mathrm{E}(\pi_i) = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}$

- thus, a large value for α_i (relative to other α's) suggests π_i is likely to be large

If $\alpha_1 = \cdots = \alpha_k$ then the Dirichlet is said to be symmetric

- the single parameter is called the concentration parameter
- all k components of π have same marginal distribution
- a large value of α concentrates the distribution of π_j at $1/k$
- values of α close to 0 lead to distribution for π_j with about $(k-1)$ out of every k of simulated values being close to 0 and the others being close to 1



```
> rdirichlet(1,c(.05,.05,.05,.05))
             [,1]       [,2]      [,3]         [,4]
[1,] 3.422228e-07 0.01720776 0.9818786 0.0009133317
> rdirichlet(1,c(5,5,5,5))
          [,1]      [,2]      [,3]      [,4]
[1,] 0.3930037 0.1890718 0.2650337 0.1528908
```

Dirichlet process prior

Suppose we want to specify some prior distribution P. For example, $\mu_i \sim P$. A fully parametric choice for P would be $N(\mu_0, \tau)$.

A non-parametric alternative is to assume P follows a Dirichlet process. That is, $P \sim DP(\alpha, P_0)$, where

- $\alpha > 0$ is the concentration parameter
- P_0 is the base distribution

As described on the previously, α characterizes prior precision and clustering.

The base distribution P_0 has to have same support as P.

A Dirichlet process is a stochastic process.

- a draw from a DP can be thought of as a draw from a probability distribution whose domain is a random variable
- even if P_0 is continuous, realization of P are from a mixture of point masses (discrete distribution), with larger values of alpha leading to distributions that more closely approximate a continuous distribution

We can think of P_0 as our best guess for the distribution of P and α as representing confidence in that guess. If α is very large then P is approximately equal to P_0.

Marginalize over P

Let $\phi_1, \cdots, \phi_K$ be the unique values of $\theta_1, \cdots, \theta_n$, where $k \leq n$. Let n_k be the number of times ϕ_k appears in $(\theta_1, \cdots, \theta_n)$. Then,

$$E\{P(A)|\theta_1, \cdots, \theta_n\} = \frac{\alpha P_0(A) + \sum_{k=1}^{K} n_k \delta_{\phi_k}(A)}{\alpha + n}$$

Thus, marginalizing (integrating out) P, results in the following conditional probabilities:

$p(\theta_{n+1} = \phi_j|\phi_1, \cdots, \phi_K) = \frac{n_j}{n+\alpha}$, for any $j \in \{1, \cdots, K\}$

$p(\theta_{n+1} \notin \{\phi_1, \cdots, \phi_K)|\phi_1, \cdots, \phi_K) = \frac{\alpha}{n+\alpha}$, for any $j \in \{1, \cdots, K\}$

Pólya urn

$$\theta_n|\theta_1,\cdots,\theta_{n-1} \sim \frac{1}{n-1+\alpha}\sum_{j\neq i}\delta_{\theta_j} + \frac{\alpha}{n-1+\alpha}P_0$$

This is a Pólya urn scheme! (suppose there are α black balls, n_1 balls of color ϕ_1, n_2 balls of color ϕ_2,..., n_K balls of color ϕ_K, in an urn. If you draw a non-black ball, put that ball back along with another ball of that same color. If you draw a black ball, put the black ball back along with a ball of color randomly drawn from P_0.)

It can be shown that the joint distribution of $(\theta_1,\cdots,\theta_n)$ is invariant to order. In general:

$$\theta_i|\theta_{-i} \sim \frac{1}{n-1+\alpha}\sum_{j\neq i}\delta_{\theta_j} + \frac{\alpha}{n-1+\alpha}P_0$$

Conclusion: DP sampling of θ's is a Pólya urn

Chinese restaurant process

The distribution on partitions induced by a DP prior is a chinese restaurant process

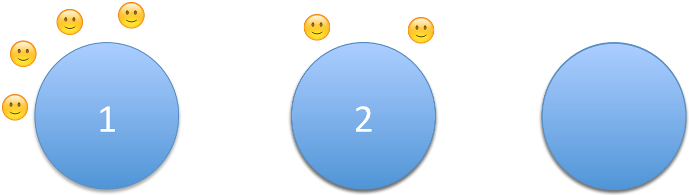
Think of all of the people sitting at table with label ϕ_j as being in cluster j. Suppose there are currently $n-1$ people sitting at K tables. The tables are labeled $\phi_1, \ldots, \phi_K$. There are n_j people at table ϕ_j, etc. There are also infinitely many empty tables.

Consider a new person who needs to decide where to sit. They will sit at occupied table ϕ_j with probability $\frac{n_j}{n-1+\alpha}$

They will sit at a new, unoccupied table with probability $\frac{\alpha}{n-1+\alpha}$

CRP

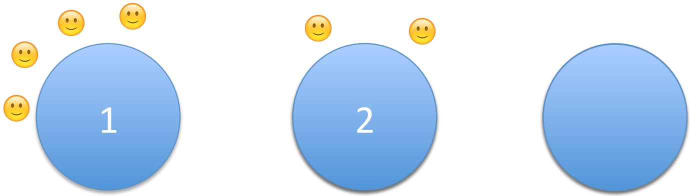
Prior Probabilities (ignores data)



P(table 1)=4/(7-1+α)
P(table 2)=2/(7-1+α)
P(new table)=α/(7-1+α)

CRP

If α=0.1

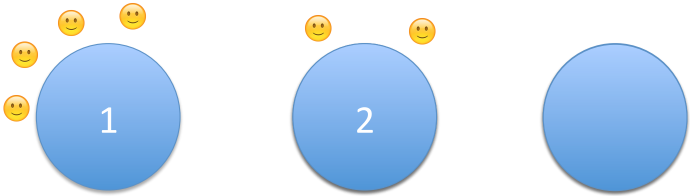


🙂

P(table 1)=0.656
P(table 2)=0.328
P(new table)=0.016

CRP

If α=10



🙂

P(table 1)=0.25
P(table 2)=0.125
P(new table)=0.625

Distribution of partitions

The partition of a DP is described by $\rho_n = (s_1, \cdots, s_n)$ with $s_i = j$ if $\theta_i = \phi_j$, the jth distinct θ-value in order of appearance. As $n \to \infty$, the partition generated by the DP has distribution CRP(α)

$$p(\rho_n|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^K \prod_{j=1}^{K} \Gamma(n_j)$$

Dirichlet process mixture models

We are going to focus on models of the form:

$$\begin{aligned} x_i|\theta_i &\sim p(x_i|\theta_i) \\ \theta_i|P &\sim P \\ P &\sim DP(\alpha P_0). \end{aligned}$$

Today we will focus on a marginal Gibbs sampler, which is a Gibbs sampler obtained after integrating out P. In other words, it's based on the CRP representation (Gibbs sampler based on partition).

Note that we could write this model

$$\begin{aligned} x_i|\theta_i &\sim p(x_i|\theta_i) \\ \theta_i|P &\sim P \\ P &\sim DP(\alpha P_0). \end{aligned}$$

as this model

$$\begin{aligned} x_i|\theta_c^* &\sim p(x_i|\theta_c^*), \text{if } s_i = c \\ \theta_c^*|P &\sim P, \text{for } c \in \rho_n \\ \rho_n &\sim CRP(n, \alpha) \end{aligned}$$

where $\theta^* = (\theta_1^*, \cdots, \theta_K^*)$ are the unique values of θ, and $\rho_n = (s_1, \cdots, s_n)$

Marginal Gibbs sampler

The marginal Gibbs sampler based on CRP will alternate between:

- updating cluster membership given current values of parameters
- updating parameters, given cluster membership

Update cluster membership

We do this one subject at a time. Given the current value of the parameters θ^*, α, the cluser membership of everyone except subject i, and the data, we draw subject i's cluster membership from a multinomial distribution.

Denote by k^{-i} the number of unique clusters if you exclude subject i. Let $n_j^{-i} = (s^{-i} = j)$ for $j = 1, \cdots, k^{-i}$

Recall prior probability of joining existing cluster:

$$P(s_i = c | s^{-i}) = \frac{n_c^{-i}}{\alpha + n - 1}, c = 1, \cdots, K^{-i}$$

And prior probability of starting new cluster:

$$P(s_i \neq s_j \text{for all} j \neq i) = \frac{\alpha}{n - 1 + \alpha}$$

Update cluster membership

For $i = 1, \cdots, n$,

$$P(s_i = c | rest) \propto \begin{cases} n_c^{-i} K(x_i | \theta_c^*), c = 1, \cdots, k^{-i} \\ \alpha \int K(x_i | \theta) dP_0(\theta), c = k^{-i} + 1 \end{cases}$$

where $K(x_i|\theta)$ is the kernel of $p(x_i|\theta)$

- If P_0 is conjugate then $\int K(x_i|\theta)dP_0(\theta)$ should be easy to calculate

Update parameters

Given cluster members $(s_1, \cdots, s_n)$, we can sample θ^* from

$$p(\theta^*_c | rest) \propto P_0(\theta^*_c) \prod_{i:s_i=c} K(x_i|\theta^*_c).$$

This is an ordinary parameter update step, where we do so within each cluster.

Specific example

Suppose we have data $x_1, \cdots, x_n$ from some unknown continuous distribution. We can use a DP approach to model it nonparametrically:

$$\begin{aligned} x_i|\theta_i &\sim N(x_i|\mu_i, \sigma_i^2) \\ \mu_i, \sigma_i^2|P &\sim P \\ P &\sim DP(\alpha P_0). \end{aligned}$$

We will assume (for now) that α is known (often set to 1). For P_0:

$$\begin{aligned} p_0(\mu_i|\sigma_i^2) &= N(\mu_0, \sigma_i^2/c_0), \\ p_0(\sigma_i^2) &\sim Inv - \chi^2(\nu_0, \sigma_0^2). \end{aligned}$$

where μ_0, c_0, ν_0, and σ_0^2 are values that we choose.

Update parameters

For each currently observed cluster, we can then update μ_j^* and $\sigma_j^{2,*}$ from normal and Inv-χ^2 distributions.

$$\sigma_j^{2,*} | rest \sim Inv-\chi^2 \left(\nu_0 + n_j, \frac{\nu_0 \sigma_0^2 + (n_j - 1) s_j^2 + \frac{c_0 n_j}{c_0 + n_j} (\overline{x}_j - \mu_0)^2}{\nu_0 + n_j} \right)$$

$$\mu_j^* | rest \sim N \left(\frac{\frac{c_0}{\sigma_j^{2,*}} \mu_0 + \frac{n_j}{\sigma_j^{2,*}} \overline{x}_j}{\frac{c_0}{\sigma_j^{2,*}} + \frac{n_j}{\sigma_j^{2,*}}}, \frac{1}{\frac{c_0}{\sigma_j^{2,*}} + \frac{n_j}{\sigma_j^{2,*}}} \right)$$

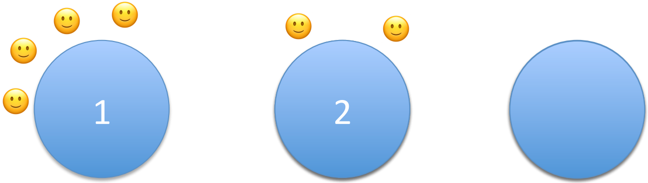
Update clusters

For $i = 1, \cdots, n$,

$$P(s_i = c | rest) \propto \left\{ \begin{array}{c} n_c^{-i} N(x_i | \mu_c^*, \sigma_c^{2,*}), c = 1, \cdots, k^{-i} \\ \alpha \int N(x_i | \mu, \sigma^2) dP_0(\mu, \sigma^2), c = k^{-i} + 1 \end{array} \right.$$

So draw s_i from a multinomial. If a new cluster is opened up, also need to draw $\mu_{k^{-i}+1}^*$ and $\sigma_{k^{-i}+1}^{2,*}$ from the prior

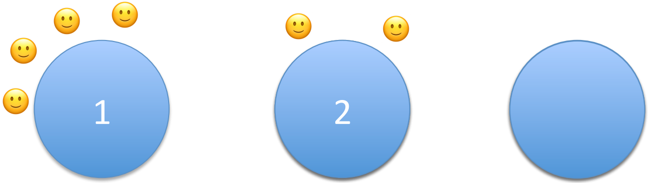
If α=1, prior



🙂

P(table 1)=4/(6+1)=0.57
P(table 2)=2/(6+1)=0.29
P(new table)=1/(6+1)=0.14

If α=1, posterior



μ*=1, $\sigma^{2,*}$=1

μ*=-2, $\sigma^{2,*}$=.25

X=-1

P(table 1) \propto 0.57*N(-1;1,1)=.57*.05

P(table 2) \propto 0.328*N(-1; -2,0.25)=.328*.11

P(new table) \propto 0.016*ave of N(-1; mu, sigma2) over prior

Example 1: BNP approach to marginal structural models

Center for
Causal Inference
C → C → I

Perelman
School of Medicine
University *of* Pennsylvania

Potential outcomes and causal model

Potential outcomes:

- Y^a: outcome if A set to a

MSM:

$$E(Y^a|V = v; \psi) = h_0(v; \psi_0) + h_1(a, v; \psi_1),$$

where

- $h_0()$ and $h_1()$ are known functions
- ψ_0 and ψ_1 are unknown parameters
- ψ_1 are causal parameters of interest

e.g., $E(Y^a|V = v; \psi) = \psi_{00} + \psi_{01}v + \psi_{10}a + \psi_{11}a \times v$

Data

- Outcome: Y
  - We only consider continuous Y
- Treatment: A
- Set of confounders: $L = (V, W)$
  - V are effect modifiers of interest
  - W are other confounders
- $\{Y_i, A_i, V_i, W_i; i = 1, \cdots, n\}$

Data cont'd

Note:

- A could be continuous or discrete
- V is typically of low dimension (1 or 2 variables) and could be continuous or discrete
- W might be high dimensional

DDP for outcome model

Dependent Dirichlet process (DDP; MacEachern (1999)):

$$p(y^a|l) = \sum_{k=1}^{\infty} \gamma_k N(y; \Delta(a, v; \psi, \gamma) + \theta_k(l), \sigma^2)$$

- infinite mixture of normals - stick-breaking representation of DP
- $\Delta(a, v; \psi, \gamma)$, defined later, ensures MSM assumption holds

Gaussian process of θ

$$\theta_k(l) \sim \mathcal{GP}(\mu_k(w), C(l; \eta, \rho))$$

where

$$\mu_k(w) = w\beta_k$$

and ith row and jth column of $C(l; \eta, \rho)$ is

$$\eta \exp\left(-\rho||l_i - l_j||^2\right) + 0.01\delta_{ij},$$

- Large η implies $\theta(l)$ is very different from linear
- Larger η penalized in loglikelihood: $\log|C(l; \eta, \rho)|$
- ρ affects the degree to which the means of subjects who have similar L will have similar $\theta(l)$

Estimators that we compare with BNP

1. Correctly specified regression model (reg)
2. IPTW with correctly specified propensity score (IPTW)
3. IPTW with weights truncated at 2nd and 98th percentiles (IPTWtr)
4. Augmented IPTW (IPTWaug)
5. TMLE with correctly specified propensity score and Super Learner for outcome model (glm, step, gam, randomforest) (TMLE)

Causal parameters and performance metrics

True causal model:
$\mathrm{E}(Y^a|V = v; \psi) = \psi_{00} + \psi_{01}v + \psi_{10}a + \psi_{11}a \times v$

- causal effect parameters: ψ_{10} and ψ_{11}

For each simulation scenario, compare:

- bias
- empirical standard deviation (ESD)
- coverage probability

Simulation scenario: bimodal outcome

$$W_j \sim N(0,1), j = 1, \cdots, 4$$

$$V \sim Bern(0.5),$$

$$A \sim Bern\{\mathrm{logit}^{-1}(m)\}$$

$$m = -0.3 + w_1 - 0.5w_2^2 - 0.8w_3 + 1.2w_4 - 0.2w_1w_4 + 0.5w_2w_3 + v$$

$Y = \Delta(A, V; \psi) + g(W) + 5(B - \overline{B}) + N(0,1)$, where B is Bernoulli(0.5), $g(W) = W_1 + 2W_2 - W_3 - 2W_4$

$\psi = (10, 1, 1, -0.5)$, $n = 200$

Results

Parameter	Method	Bias	Coverage	ESD
ψ_{10}: A	REG	0.02	1.00	0.55
	IPTW	0.01	0.98	0.81
	IPTWtr	0.04	0.98	0.75
	IPTWaug	0.01	0.95	0.60
	TMLE	0.01	0.93	0.58
	BNP	0.00	0.97	0.38
ψ_{11}: $A \times V$	REG	-0.04	1.00	0.79
	IPTW	-0.05	0.95	1.37
.	IPTWtr	-0.07	0.95	1.29
	IPTWaug	-0.02	0.96	0.88
	TMLE	-0.04	0.93	0.87
	BNP	0.00	0.96	0.53

Simulation scenario: near violation of positivity assumption

- Scenario similar to Kang and Schafer (2007)
- $n = 200$
- V ∼Bernoulli(0.5)
- $Z_1, \cdots, Z_4$ iid $N(0, 1)$
- A from Bernoulli $\text{logit}^{-1}(-1Z_1 + 0.5Z_2 - 0.25Z_3 - 0.1Z_4)$
- Y is from a normal with mean $\Delta(A, V; \psi) + Z\beta$ and standard deviation 200, where $\beta = (27.4, 13.7, 13.7, 13.7)$
- $\psi = (210, -50, 50, 20)$

Observed confounders

- $W_1 = \exp(Z_1/2)$
- $W_2 = Z_2/(1 + \exp(Z_1)) + 10$
- $W_3 = (Z_1 Z_3/25 + 0.6)^3$
- $W_4 = (Z_2 + Z_4 + 20)^2$

Results

Parameter	Method		Coverage	ESD
ψ_{10}: A	REG	-7.51	0.92	43.57
	IPTW	-6.17	0.93	54.43
	IPTWtr	-9.59	0.94	45.40
	IPTWaug	-12.37	0.95	139.29
	TMLE	-9.34	0.90	48.52
	BNP	-7.68	0.93	44.35
ψ_{11}: $A \times V$	REG	1.91	0.95	57.35
	IPTW	3.36	0.93	75.64
	IPTWtr	3.01	0.95	62.66
	IPTWaug	2.97	0.96	204.56
	TMLE	4.90	0.92	68.64
	BNP	1.98	0.95	57.43

Study design and variables

- Geisinger Health System EHR
- New initiators of angiotensin-converting enzyme (ACE) inhibitors or angiotensin II receptor blockers (ARBs)
- Women, age 65+, diabetes diagnosis, initiate treatment 2001-2008
- $n = 1,964$
- Confounders: 24 variables including race, age, BMI, BP, history of CKD, MI, CHF, stroke, cancer, etc.
- Outcome: all cause mortatlity

Causal model, results

Our casual model of interest is

$$E(Y^a|\psi) = \psi_0 + \psi_1 a$$

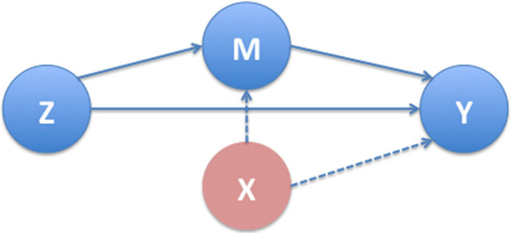
- $A = 1$ if ARB, $= 0$ if ACEI
- Y is log survival time

Results:

- ψ_0: $3.25\ [2.83, 3.72]$
- ψ_1: $0.17\ [-0.19, 0.53]$

Example 2: BNP approach to causal mediation

Mediation



Causal effects

Potential outcomes: $Y_{z,M_{z'}}$ the value of the outcome that would have been observed if an individual had been assigned to intervention z with (possibly hypothetically) mediator M set to its value under z'

- Natural indirect effect $E[Y_{1,M_1} - Y_{1,M_0}]$
- Natural direct effect: $E[Y_{1,M_0} - Y_{0,M_0}]$

Identifiability

Sequential ignorability

$$\begin{aligned}\{Y_{z',m}, M_z\} &\perp Z \mid X = x \\ Y_{z',m} &\perp M_z \mid Z = z, X = x,\end{aligned}$$

$$
\begin{aligned}
E(Y_{1,M_0} \mid \mathrm{X} = \mathrm{x}) &= \int E(Y_{1,m} \mid M_0 = m, Z = 0, \mathrm{X} = \mathrm{x}) dF_{M_0 \mid Z=0, \mathrm{X}=\mathrm{x}}(m) \\
&= \int E(Y_{1,m} \mid Z = 0, \mathrm{X} = \mathrm{x}) dF_{M_0 \mid Z=0, \mathrm{X}=\mathrm{x}}(m) \\
&= \int E(Y_{1,m} \mid Z = 1, \mathrm{X} = \mathrm{x}) dF_{M_0 \mid Z=0, \mathrm{X}=\mathrm{x}}(m) \\
&= \int E(Y_{1,m} \mid M_1 = m, Z = 1, \mathrm{X} = \mathrm{x}) dF_{M_0 \mid Z=0, \mathrm{X}=\mathrm{x}}(m) \\
&= \int E(Y \mid M_1 = m, Z = 1, \mathrm{X} = \mathrm{x}) dF_{M \mid Z=0, \mathrm{X}=\mathrm{x}}(m),
\end{aligned}
$$

BNP observed data model

$$\begin{aligned} (Y^z_{\text{obs},i}, M^z_{\text{obs},i}, \boldsymbol{X}^z_i) &\sim N_q(\boldsymbol{\mu}_{z,i}, \boldsymbol{\Sigma}_{z,i}), \\ (\boldsymbol{\mu}_{z,i}, \boldsymbol{\Sigma}_{z,i}) &\sim G_z, \\ G_z &\sim DP(\alpha_z \mathcal{G}_z), \end{aligned}$$

The base distribution $\mathcal{G}_z$ is taken to be the conjugate normal-inverse-Wishart distribution (NIW).

- R package: BNPmediation (github.com/lit777/BNPMediation)
- Reference: Kim et al (2016) Biometrics

Example 3: BNP approach to causal inference with missing confounders

DP mixture of multivariate normals

Mueller et al. (1996) proposed:

$$\begin{aligned}(y_i, a_i, l_i) &\sim N(\mu_i, \Sigma_i)\\ \mu_i, \Sigma_i &\sim G\\ G &\sim DP(\alpha G_0)\end{aligned}$$

where, typically, G_0 would be normal-inverse-wishart.

- problematic if many covariates
- does not easily handle discrete outcomes and covariates

Joint DP mixture model

Shahbaba and Neal (2009) proposed:

$$\begin{aligned} P &\sim DP(\alpha, P_{0\theta} \times P_{0\omega}) \\ (\theta_i, \omega_i)|P &\sim P \\ X_{i,j}|\omega_i &\sim p(x_j|\omega_i), \\ Y_i|X_i, \theta_i &\sim p(y|x, \theta_i). \end{aligned}$$

- In our causal setting, $X = (A, L)$
- $p(y|x, \theta_i)$ can be a GLM
- covariates locally independent, so computationally friendly
- Potentially puts too much weight on fit of X's at the cost of $y|x$, which is what we care most about

Enriched DP mixture model

Wade et al. (2014) proposed:

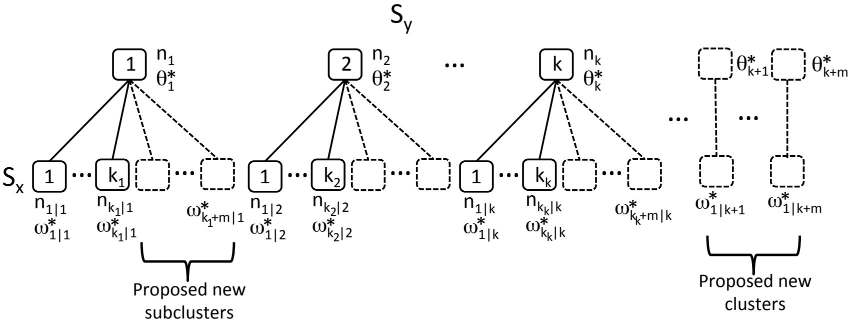
$$\begin{aligned} P &\sim EDP(\alpha_\theta, \alpha_\omega, P_0) \\ (\theta_i, \omega_i)|P &\sim P \\ X_{i,j}|\omega_i &\sim p(x_j|\omega_i), \\ Y_i|X_i, \theta_i &\sim p(y|x, \theta_i). \end{aligned}$$

$P \sim EDP(\alpha_\theta, \alpha_\omega, P_0)$ means $P_\theta \sim DP(\alpha_\theta, P_{0\theta})$ and $P_{\omega|\theta} \sim DP(\alpha_\omega, P_{0\omega|\theta})$ and base measures $P_0 = P_{0\theta} \times P_{0\omega|\theta}$.

Gibbs sampler

1. sample $s_i = (s_{i,y}, s_{i,x})$ given current values of parameters and all other data
   - extension of algorithm 8 of Neal (2000)
2. update ω_j^* and θ_j^* given s
   - This update is generally very easy (either conjugate or standard Bayesian calculations)
3. Update hyperparameters such as α_ω and α_θ
4. Data augmentation step: given cluster membership and current values of parameters, draw from posterior of L_{ij} for any patient i who has missing value for covariate j

S update: general



Simulation 1: binary outcome, simple form

$$
\begin{aligned}
L_1 &\sim \mathrm{Bern}(0.2) \\
L_2 &\sim \mathrm{Bern}\{\mathrm{logit}^{-1}(0.3 + 0.2L_1)\} \\
L_3 &\sim N(L_1 - L_2, 1^2) \\
L_4 &\sim N(1 + 0.5L_1 + 0.2L_2 - 0.3L_3, 2^2) \\
A &\sim \mathrm{Bern}\{\mathrm{logit}^{-1}(-0.4 + L_1 + L_2 + L_3 - 0.4L_4)\} \\
Y &\sim \mathrm{Bern}\{\mathrm{logit}^{-1}(-0.5 + 0.78A - 0.5L_1 - 0.3L_2 + 0.5L_3 - 0.5L_4)\}
\end{aligned}
$$

The true causal parameters are $\psi_{rr} = 1.5$

Results

		Relative risk, ψ_{rr}	
Method	Bias	Coverage	ESD
		$n = 250$	
IPTW	0.09	0.96	0.43
TMLE	0.06	0.92	0.37
Bayesian par.	0.05	0.93	0.33
BNP	0.03	0.93	0.32
BNP missing data	0.05	0.94	0.33

Simulation 2: binary outcome, complex form

$$L \sim N(4, 2^2)$$

$$A|L \sim \mathrm{Bern}\{\mathrm{logit}^{-1}(1.3 - 0.8L)\}$$

$$\begin{aligned} Y|A, L \sim\ & (p)\mathrm{Bern}\{\mathrm{logit}^{-1}(-0.8 - 0.1L + A)\} \\ & + (1-p)\mathrm{Bern}\{\mathrm{logit}^{-1}(-2 + 0.45L)\}, \\ p =\ & \frac{2\exp\{-2(L-4)^2\}}{2\exp\{-2(L-4)^2\} + 2\exp\{-2(L-6)^2\}} \end{aligned}$$

The true causal parameters are $\psi_{rr} = 1.4$

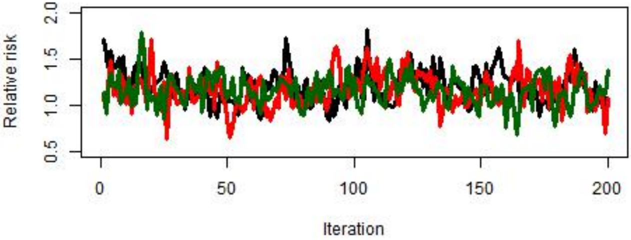
Results

		Relative risk	
Method	Bias	Coverage	ESD
		$n = 1000$	
IPTW	0.00	0.92	0.19
TMLE	0.02	0.91	0.16
Bayesian par.	0.33	0.19	0.13
BNP	0.04	0.95	0.13
BNP missing data	0.02	0.94	0.15

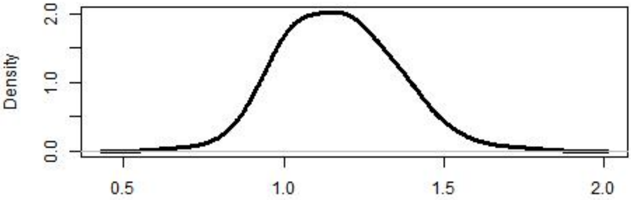
Application: ART for HIV/HCV-coinfected patients

Data from Veterans Aging Cohort Study

- Interested in ART-regiments that include nucleoside reverse transcriptase inhibitor (NRTI)
- Treatment comparison: mitochondrial toxic NRTI (mtNRTI) versus other NRTI
- Population: co-infected patients who newly initiated an ART-regimen that include NRTIs (either mtNRTIs or other NRTIs) from 2002 to 2009
- Outcome: all cause mortality
- Confounders (L): age at baseline (years), race/ethnicity, body mass index, diabetes mellitus, alcohol dependence/abuse, drug abuse, year of ART initiation, exposure to other antiretrovirals associated with hepatotoxicity, CD4 count, HIV RNA, alanine aminotransferase (ALT), aspartate aminotransferase (AST), and fibrosis-4 (FIB-4) score.



Posterior distribution



Discussion

BNP approach: Use BNP models for observed data, then use post-processing steps to obtain causal effects

Not discussed: informative priors on unidentifiable parameters can be used to capture uncertainty about causal assumptions

Software (R packages):

- BNPmediation
- DPpackage
- BayesTree (BART)
- GPfit (Gaussian processes)

Thanks!

Center for
Causal Inference
C → C → I

Perelman
School of Medicine
UNIVERSITY of PENNSYLVANIA