A Bayesian Nonparametric Approach for Causal Inference with MAR covariates

J. Roy, K. Lum, M. Daniels, B. Zeldow, J. Dworkin, V. Lo Re

UPenn and UT-Austin

HDDA-2017, June 15-19 2017, CIMAT, Guanajuato, Mexico

Approach to problem of causal inference

- Except in very simple situations (e.g., randomized trial with perfect compliance), untestable (from the observed data) assumptions need to be made for drawing causal inferences
- Similar to missing data problems, the problem can be partitioned into two components:
  1. a model for the observed data
  2. a set of (reasonable) assumptions that allow identification and estimation of causal estimands given the observed data.
- Given that the second component is not checkable from the observed data, uncertainty about these assumptions is essential
- These two components can be handled most naturally in the Bayesian paradigm - in context of causal mediation (D et al. 2012; Kim et al. 2016)

Bayesian approach I

- Requires modelling conditional distribution of outcome given confounders OR joint distribution of outcome and confounders
- since the number of potential confounders might be large, ample opportunity for mis-specification

Bayesian approach II

- for the former approach, can directly model the conditional distribution using a dependent Dirichlet process (MacEachern, 1999) and estimate the marginal distribution of confounders using its empirical distribution (Roy, Lum and D (2017)). But
  - computationally expensive (invert $n \times n$ matrices)
  - unclear how to deal with missing potential confounders?

Bayesian approach III

- for the latter (joint model), can use *generative* models
  - refinement of models in Shahbaba and Neal (2009) and Wade et al. (2014)
  - can handle large n and p
  - can handle missing covariates/confounders easily (under an assumption of ignorable missingness)

Bayesian approach IV

- proposed Bayesian approach will
  - offer efficiency gains
  - provide full posterior inference
  - accomodate any causal effect metric
  - accomodate uncertainty about uncheckable assumptions
  - allow missing confounders that are ignorably missing

Notation

- A: treatment
- L: potential (pre-treatment) confounders
- Y: observed outcome
- Y^a: potential outcome if the subject had been assigned to treatment level a.

Causal effects

- average causal effect: $E[Y^1 - Y^0]$
- conditional average causal effect: $E[(Y^1 - Y^0)|V]$
- quantile causal effect: $F_1^{-1}(p) - F_0^{-1}(p)$

Uncheckable causal assumptions

1. Consistency: $Y^a = Y$ among subjects with $A = a$, for all a.
   - implies that $p(Y^a|A = a, L) = p(Y|A = a, L)$.
   - allows us to estimate parameters in $Y^a|L$ model using the observed data
2. Positivity: $p(a|L) > 0$
   - each treatment level has non-zero probability for every confounder level.
3. Ignorability: $\{Y^a : \forall a \in \mathcal{A}\} \perp A|L\}$.
   - implies that $p(Y^a|A = a, L) = p(Y^a|A = a', L)$.
   - 'no unmeasured confounders' assumption.

Three assumptions imply

$$F(y|A = a, L) = F(y^a|A = a, L) = F(y^a|L)$$

Observed data models using BNP I

- model the joint distribution $p(Y, A, L)$
- for simplicity, let $X_i = (A_i^T, L_i^T)^T$ so we model $p(Y, X)$ (or just condition on A)

Observed data models using BNP II

Model the joint distribution of (Y, X) using an enriched Dirichlet process (EDP) mixture (Wade et al, 2011, 2014)

$$\begin{aligned} Y_i|X_i, \theta_i &\sim p(y|x, \theta_i) \\ X_{i,r}|\omega_i &\sim p(x_r|\omega_i), \\ (\theta_i, \omega_i)|P &\sim P \\ P &\sim EDP(\alpha_\theta, \alpha_\omega, P_0). \end{aligned} \tag{1}$$

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

Observed data models using BNP III

- each subject i has their own parameters θ_i and ω_i
  - However, because P is discrete, some clusters of subjects will have the same θ_i and ω_i
- The number of clusters depends on the concentration parameters α_θ and α_ω (low values indicate fewer clusters)
  - typical DP models have a single concentration parameter
- The *enrichment* of the usual DP is to have nested concentration parameters
  - allows for more x-clusters than y-clusters, which is important because the dimension of x will typically be much larger than that of y.
  - keeps cluster membership dependent on both $y|x$ and x through the nesting of the random partition.

Observed data models using BNP IV

We assume a local generalized linear model for $p(y|x, \theta_i)$,

$$p(y|x, \theta_i) = \exp\left\{\frac{Y_i\eta_i - b(\eta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right\}$$

where $g\{b'(\eta_i)\} = X_i\beta_i$ and $g\{\}$ is a link function.
If Y is binary

$$Y_i|X_i, \theta_i \sim \text{Bern}\{\text{logit}^{-1}(X\beta_i)\}$$

where $\theta_i = \beta_i$ and X is the design matrix involving A and L

Observed data models using BNP V

- assumes covariates **X** are locally independent.
  - That is given ω_i, covariates are independent. Two subjects in the same subcluster would have similar values of X.
- local independence assumption
  - makes it easy to include many continuous and discrete confounders, because the joint distribution is just a product of marginal distributions.
  - makes computations considerably faster because covariance matrices for the joint distribution of confounders are not needed.
- while assume that locally the generalized linear model is correctly specified for y and x and that the x's are independent from each other, globally all of the variables are dependent with potentially non-linear relationships.

Observed data models using BNP VI

- The EDP model can equivalently be represented with the square-breaking formulation (Wade et al, 2011) (generalization of the standard stick-breaking representation of DP models (Sethuraman, 1994))

Observed data models using BNP VII

- joint distribution of the observed data for subject i can be written

$$f(y_i, x_i|P) = \sum_{j=1}^{\infty} \gamma_j \sum_{l=1}^{\infty} \gamma_{l|j} K(y_i|x_i, \theta_j) K(x_i|\omega_{l|j}),$$

  where j indexes the y-clusters and the $K()$ are the kernels of the corresponding distributions.

- The weights have priors $\gamma'_j \sim$Beta$(1, \alpha_\theta)$ and $\gamma'_{l|j} \sim$Beta$(1, \alpha_\omega)$, where $\gamma_j = \gamma'_j \prod_{r<j}(1-\gamma'_r)$ and $\gamma_{l|j} = \gamma'_{l|j} \prod_{m<l}(1-\gamma'_{m|j})$.

Observed data models using BNP VIII

- The conditional distribution implied by the joint model is $p(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|x, \theta_j)$, where

$$w_j(x) = \frac{\sum_{l=1}^{\infty} \gamma_{l|j} K(x|\omega_{l|j})}{\sum_{h=1}^{\infty} \gamma_h \sum_{l=1}^{\infty} \gamma_{l|h} K(x|\omega_{l|h})}.$$

- Notice that the weights $w_j(x)$ depend on x. Therefore, even though $K(y|x, \theta_j)$ is a generalized linear model, $p(y|x)$ is a computationally tractable, flexible, non-linear, non-additive model.

Posterior computations

- Gibbs sampler for obtaining draws from the posterior distribution of the parameters (extension of Neal (2000) Algorithm 8 to accomodate nested clustering)
- Data augmentation - sample from conditional distribution of missing covariates at each iteration (valid under ignorable missingness)
- MC integration (over L) for each posterior draw to compute any functional of the distribution of the potential outcomes (can be done in parallel)

Simulation setup I

- We conducted simulation studies to examine the performance of the proposed BNP approach under several scenarios
  1. binary outcome, simple functional form
  2. binary outcome, mixture distribution
  3. continuous outcome, complex functional form
  4. continuous outcome, complex treatment, many potential confounders

Simulation setup II

- compare the BNP approach to
  - inverse probability treatment weighted estimator (IPTW)
  - targeted maximum likelihood estimator (TMLE)
    - used Super Learner (van der Laan et al., 2007); ensemble machine learning method that uses cross-validation to weigh different prediction algorithms
    - used four algorithms (glm, step, gam, randomforest) and implemented TMLE using the R package tmle (Gruber and van der Laan, 2012).
  - parametric Bayesian approach

Scenario 1 I

- binary outcome, simple functional forms.
- IPTW and TMLE uses a correctly specified propensity score
- TMLE uses Super Learner with 3 prediction algorithms for the outcome model.
- The Bayesian parametric (Bayesian par.) approach uses a correctly specified logistic regression model and integrates over confounders using the empirical distribution.
- 'BNP missing data' is the BNP approach, with data augmentation, applied to a data set where approximately 20% of the covariate values were set to missing; the other methods were applied to the full data set with no missing covariate values

Scenario 1 II

	Relative risk, ψ_{rr}				Risk difference, ψ_{rd}			
Method	Bias	Coverage	ESD	CI width	Bias	Coverage	ESD	CI width
				$n = 250$				
IPTW	0.09	0.96	0.43	1.76	0.00	0.96	0.08	0.33
TMLE	0.06	0.92	0.37	1.31	0.00	0.91	0.07	0.25
Bayesian par.	0.05	0.93	0.33	1.29	0.00	0.94	0.06	0.24
BNP	0.03	0.93	0.32	1.20	0.00	0.93	0.06	0.23
BNP missing data	0.05	0.94	0.33	1.31	0.00	0.94	0.07	0.25
				$n = 1000$				
IPTW	0.03	0.97	0.20	0.84	0.00	0.97	0.04	0.17
TMLE	0.01	0.93	0.18	0.65	0.00	0.93	0.04	0.13
Bayesian par.	0.01	0.95	0.15	0.59	0.00	0.94	0.03	0.12
BNP	0.01	0.94	0.15	0.58	0.00	0.94	0.03	0.12
BNP missing data	0.01	0.93	0.16	0.63	0.00	0.93	0.03	0.13

Scenario 2

Table: Results from simulation scenario 2: binary outcome, mixture distribution. IPTW and TMLE use a correctly specified propensity score. The Bayesian parametric (Bayesian par.) approach uses a misspecified logistic regression model.

	Relative risk, ψ_{rr}				Risk difference, ψ_{rd}			
Method	Bias	Coverage	ESD	CI width	Bias	Coverage	ESD	CI width
				$n = 250$				
IPTW	0.02	0.92	0.27	1.27	0.01	0.89	0.13	0.44
TMLE	0.03	0.86	0.32	1.03	0.02	0.83	0.12	0.36
Bayesian par.	0.36	0.65	0.25	0.94	0.12	0.62	0.08	0.29
BNP	0.04	0.93	0.26	0.97	0.01	0.93	0.09	0.34
BNP missing data	0.07	0.95	0.26	1.00	0.02	0.94	0.09	0.35
				$n = 1000$				
IPTW	0.00	0.92	0.19	0.71	0.00	0.92	0.07	0.26
TMLE	0.02	0.91	0.16	0.56	0.01	0.90	0.06	0.20
Bayesian par.	0.33	0.19	0.13	0.46	0.12	0.17	0.04	0.14
BNP	0.04	0.95	0.13	0.54	0.02	0.94	0.05	0.20
BNP missing data	0.02	0.94	0.15	0.58	0.01	0.94	0.05	0.21

Scenario 3

Table: Results from simulation scenario 3: continuous outcome, complex functional forms. IPTW and TMLE both use a correctly specified propensity score.

Method	Bias	Coverage	ESD	CI width
		$n = 1000$		
IPTW	0.00	0.98	0.24	1.18
TMLE	0.00	0.94	0.21	0.79
BNP	0.09	0.91	0.19	0.71
		$n = 3000$		
IPTW	0.00	0.99	0.17	0.68
TMLE	0.00	0.94	0.12	0.47
BNP	0.05	0.93	0.10	0.41

Scenario 4

Table: Results from simulation scenario 4: continuous outcome, complex treatment, 84 covariates (but only 4 confounders). IPTW and TMLE use a misspecified propensity score. BNP is the proposed approach. Results are from 1000 simulated datasets.

Method	Bias	Coverage	ESD	CI width
		$n = 1000$		
IPTW	0.26	0.90	0.23	1.10
TMLE	0.11	0.89	0.20	0.76
BNP	0.05	0.87	0.23	0.72
		$n = 3000$		
IPTW	0.27	0.61	0.13	0.61
TMLE	0.09	0.83	0.13	0.43
BNP	0.06	0.89	0.11	0.41

Simulation conclusions

- In scenario 4,
  - we generated a complex treatment for which the propensity score was misspecified as it would be unlikely for the form of the true propensity model to be implemented in the semiparametric approaches
  - the BNP approach had the smallest bias and ESD.
  - relatively large number of non-confounders and each of the methods displayed some amount of undercoverage.
- For the TMLE method, bootstrapping as opposed to using asymptotic confidence intervals, may improve the coverage.

HIV Data example I

- Antiretroviral therapy (ART) is recommended for all human immunodeficiency virus (HIV) / chronic hepatitis C virus (HCV)-coinfected patients.
- ART regimens often include drugs from the nucleoside reverse transcriptase inhibitor (NRTI) class.
  - concern that some drugs in the NRTI class (didanosine, stavudine, zidovudine, and zalcitabine) might cause depletion of mitochondrial DNA, leading to liver injury.

HIV Data example II

- We apply the proposed BNP approach to compare outcome Y (death within 2 years) among those prescribed mitochondrial toxic NRTI (mtNRTI)-containing ART regimen to those prescribed other NRTI-containing ART regimen.
- data from a study of HIV/HCV patients who newly initiated ART within the Veterans Aging Cohort Study (VACS)
- The study population included co-infected patients who newly initiated an ART-regimen that include NRTIs (either mtNRTIs or other NRTIs) from 2002 to 2009.
- total of $n = 1747$ patients included in the study

HIV Data example III

- $A = 1$: initiating an ART regimen that included an mtNRTI; $A = 0$ initiating an ART regiment that included some other NRTI.
- outcome: all-cause mortality (focused on the event occurring within 2 years of ART initiation)
- There were 76 deaths out of 836 patients in the mtNRTI group, and 89 deaths out of 911 patients in the other NRTI group.
- causal parameter of interest is the relative risk:

$$\psi_{rr} = E(Y^1)/E(Y^0)$$

HIV Data example IV

- Variables that were included in the model as confounders (L) included
  - baseline demographics and clinical variables: age at baseline (years), race/ethnicity, body mass index, diabetes mellitus, alcohol dependence/abuse, injection/non-injection drug abuse, year of ART initiation, and exposure to other antiretrovirals associated with hepatotoxicity (i.e., abacavir, nevirapine, saquinavir, tipranavir).
  - baseline laboratory variables: CD4 count, HIV RNA, alanine aminotransferase (ALT), aspartate aminotransferase (AST), and fibrosis-4 (FIB-4) score.

HIV Data example V

- The percentage of missing data for each variable is as follows: ALT 1.3%, AST 2.5%, CD4 1.8%, FIB-4 3.1%.
- The percentage of patients with at least one missing variable is 4.8%.

Data example: Results I

- use the EDP model with a logistic regression model for the outcome.
- We ran three chains of the Gibbs sampler, each with 20,500 iterations.

Data example: Results II

- The posterior median and 95% credible intervals (CI) for α_θ and α_ω were $0.63(0.21, 1.43)$ and $0.74(0.43, 1.16)$, (larger values leading to more clusters)
- The number of y-clusters, k tended to be about 4, while the number of x-subclusters, k_j tended to range from 1 to 7.
- For example, at the last iteration of the first chain, there were $k = 5$ y-clusters, with the following sample sizes in each subcluster: cluster $s_y = 1$, $(36, 164, 134, 45, 32, 38, 76, 1)$; cluster $s_y = 2$, $(171, 211, 131, 68, 18, 1)$; cluster $s_y = 3$, $(171, 281, 172, 50, 28)$; cluster $s_y = 4$, $(137, 30, 2, 1)$; cluster $s_y = 5$, (2).

Data example: Results III

- The posterior median and 95% CI of the average causal relative risk (RR), ψ_{rr}, were

$$1.16(0.87, 1.54)$$

  - 16% increased risk of death within 2 years comparing mtNRTI-containing ART regimens with other NRTI-containing ART regimens.

Conclusions I

- proposed a Bayesian nonparametric approach for causal inference for large n and p that can handle discrete or continuous outcomes and categorical treatment.
- simulations showed overall good performance of the BNP approach.

Conclusions II

- While the full distribution of outcome, treatment, and confounders is modeled, the proposed BNP approach allows
  - for flexible modeling of these distributions
  - estimation of any functionals of the potential outcome distribution
  - high-dimensional confounding.
  - 'imputation' of missing covariates under ignorable missingness
  - for uncertainty about uncheckable assumptions via informative priors

Ongoing work

- EDP approach allows α_ω to be a function of θ
  - In our analyses we only included a single α_ω parameter
  - explore more complex models
- extension to the time-varying confounding setting
- extension to settings with many covariates that are not actually confounders: explore zero-inflated or shrinkage priors for the coefficients in the BNP model