Bayesian Nonparametric Model for Zero-Inflated Outcomes

Clustering, Prediction, and Causal Inference

Arman Oganisian

with Jason Roy and Nandita Mitra

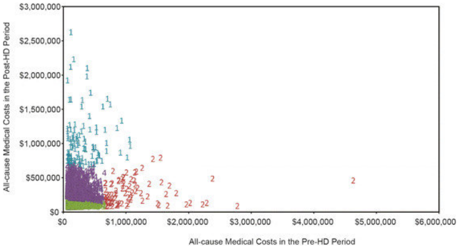
Division of Biostatistics
Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania

July 31, 2019

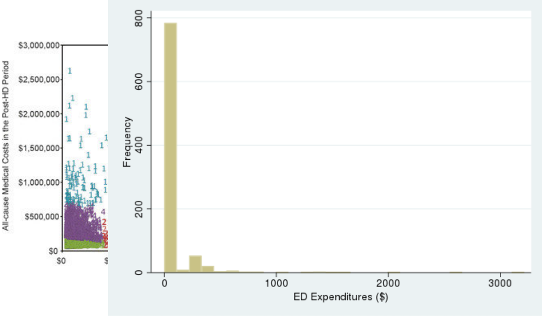
Motivation

- Health policy questions involving costs are complicated:
  - Causality: How different would average costs have been under alternative treatment?
  - Prediction: How much medical costs will subject X likely accumulate?
  - Clustering: Can we identify interesting patient subgroups?
- Cost data are complicated:
  - zero-inflation
  - skewness
  - multimodality
- Complicated questions with complicated data.

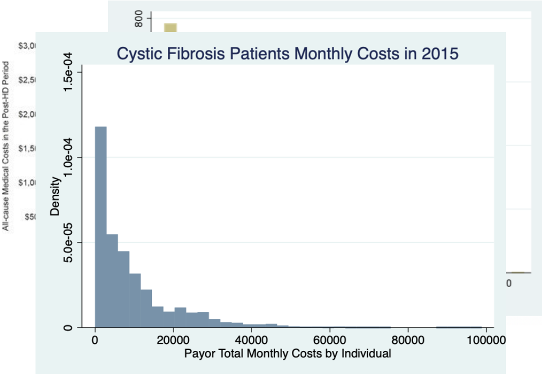
EXAMPLES FROM LITERATURE



Examples from Literature



Examples from Literature



What are our options?

Observed data $D = (D_i)_{i=1:n} = (Y_i, x_i = (A_i, L_i))_{i=1:n}$

- Ignore zeros (efficiency loss, underestimate treatment effect).
- Add a penny. (ad hoc, structural zeros $\rightarrow$ structural pennies)
- Hurdle model (parametric, no clustering)

$$Y_i \mid A_i, L_i \sim \pi\left(x_i'\gamma\right)\delta_0\left(y_i\right) + \left(1 - \pi\left(x_i'\gamma\right)\right) \cdot f\left(y_i \mid x_i'\beta\right)$$

Bayesian Standardization

Under certain identification assumptions, can compute $\Psi = E[Y^1 - Y^0]$.

$$E[Y^a|D] = \int_\beta \int_L E[Y|A = a, L, \beta]p(L)p(\beta|D)dL\, d\beta$$

Zero-Inflated Dirichlet Process Mixture

Generative model for the full joint data $p(D_i|\omega_i)$

$$Y_i \mid A_i, L_i \sim \pi\left(x_i'\gamma_i\right)\delta_0\left(y_i\right) + \left(1 - \pi\left(x_i'\gamma_i\right)\right) \cdot N\left(y_i \mid x_i'\beta_i, \phi_i\right)$$

$$A_i \mid L_i \sim \text{Ber}\left(\text{expit}\left(m_i'\eta_i\right)\right)$$

$$L_i \sim p\left(l_i \mid \theta_i\right)$$

Joint prior for $\omega_i = (\beta_i, \phi_i, \gamma_i, \eta_i, \theta_i)$

$$\omega_i \mid G \sim G$$

$$G \sim DP\left(\alpha G_0\right)$$

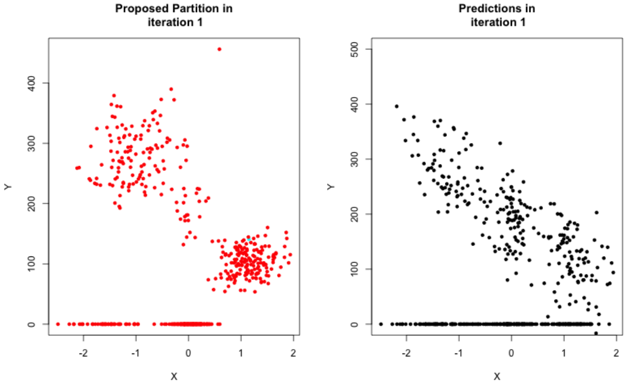
Induced Posterior Clustering

Conditional posterior of i^{th} subject's parameters

$$p\left(\omega_i|\omega_{1:(i-1)}, D\right) \propto \frac{\alpha}{\alpha+i-1} p\left(D_i|\omega_i\right) G_0\left(\omega_i\right) + \frac{1}{\alpha+i-1}\sum_{j<i} p\left(D_i|\omega_j\right) I(\omega_i=\omega_j)$$

- Data adaptive.
- Posterior clustering.
- Flexible predictions by ensembling cluster-specific models.

MCMC Inference: Partitions and Predictions



Some Results

In the paper we have

- expressions for key posterior distributions.
- required Monte Carlo procedures.
- standardization procedure around the model.
- posterior predictive checks assessing positivity.
- hard posterior classification in presence of label switching.
- uncertainty visualization in cluster assignment.

Simulations Results

- Simulated cost data from three distinct Bernoulli-Gamma hurdle model.
- $n = 3,000$ subjects, 5 covariates, single binary treatment.

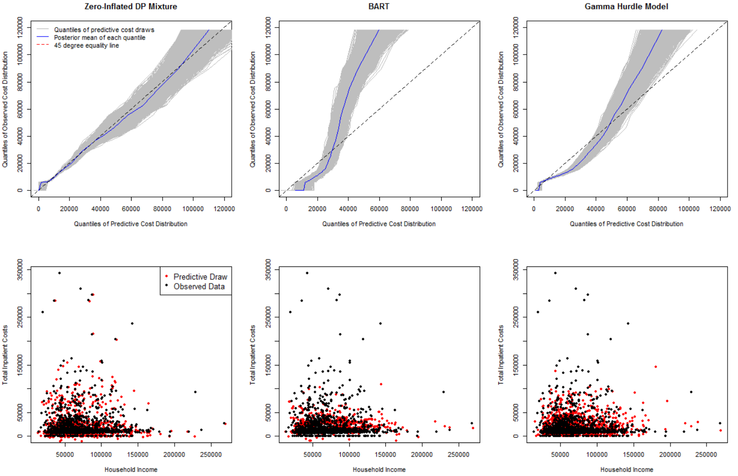
DGP	Model	Bias	Coverage	Rel. Interval Width
Clustered	Zero-Inflated DP	-.081	94.3%	1.10
	BART	-.746	76.2%	1.34
	Doubly Robust	.795	87.1%	1.70
	Gamma Hurdle	-.509	79.8%	1
	Gamma +.01	1.817	4.7%	1.39
Parametric	Zero-Inflated DP	.097	95.1%	1.01
	BART	-.054	96.1%	1.09
	Doubly Robust	-.027	95.9%	1.07
	Gamma Hurdle	-.014	95.1%	1
	Gamma +.01	-.489	100%	2.32

Center for Causal Inference @StableMarkets
DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS
Perelman School of Medicine UNIVERSITY of PENNSYLVANIA

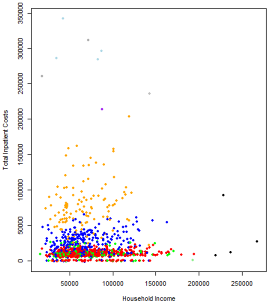
Treatment costs for Endometrial Cancer

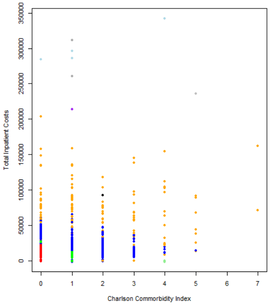
- Data source: SEER-Medicare.
- Endometrial cancer patients (N≈1,000).
- Treatment: post-hysterectomy radiation vs. chemotherapy.
- Outcome: Total inpatient costs over 2 years.
  - Skewed, zero-inflated
  - Chemo arm: 15% zeros; RT arm: 8%
- Measured confounders: tumor grade, cancer stage, CCI.

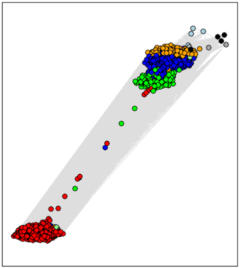
PREDICTIONS



Posterior Clustering



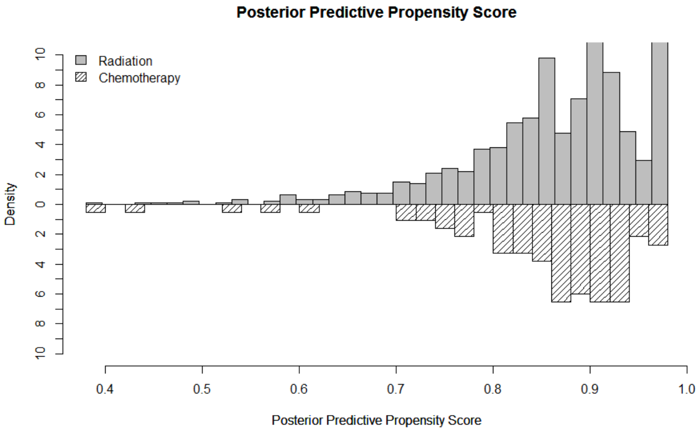




Average Treatment Effects

	Avg. Causal Effect	Median Causal Effect	Zero - Risk Ratio
Zero-Inflated DP	1672.62 (-2566.42, 5722.56)	872.68 (-833.35, 2790.18)	0.498 (0.31, 0.78)
BART	1779.62 (-6085.89, 9797.13)	-	-
Gamma Hurdle	2016.71 (-1499.38, 5593.40)	-	.505 (.34, .76)
Gamma +.01	4889.00 (1004.37, 8795.61)	-	-

Posterior Predictive Propensity Scores



Paper, Tutorial, Software

- arXiv: `https://arxiv.org/abs/1810.09494`
- Interactive DP Tutorial with R Shiny: `https://stablemarkets.shinyapps.io/dpmixapp/`
- ChiRP R package: `https://stablemarkets.github.io/ChiRPsite/index.html`

arXiv | Tutorial | ChiRP