Introduction to Bayesian Nonparametrics

Arman Oganisian

@StableMarkets

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

Outline

Outline of the talk

- What is nonparametric Bayesian inference?
- What are some examples of nonparametric priors?
- Dirichlet Process priors.
- Application to cost modeling.

What is Nonparametric Bayes?

$$Y_1, Y_2, \ldots, Y_n \mid G \sim G$$

How to do inference on G?

What is Nonparametric Bayes?

$$Y_1, Y_2, \dots, Y_n \mid G \sim G$$

How to do inference on G?

- Parametric Inference: $G = G_\omega$.

$$G_\omega = N(\mu, \phi), \quad \omega = (\mu, \phi)$$

What is Nonparametric Bayes?

$$Y_1, Y_2, \ldots, Y_n \mid G \sim G$$

How to do inference on G?

- Parametric Inference: $G = G_\omega$.

$$G_\omega = N(\mu, \phi), \quad \omega = (\mu, \phi)$$

  - Prior over $\omega \to$ prior over G_ω.
  - Posterior over $\omega \to$ posterior over G_ω

What is Nonparametric Bayes?

$$Y_1, Y_2, \ldots, Y_n \mid G \sim G$$

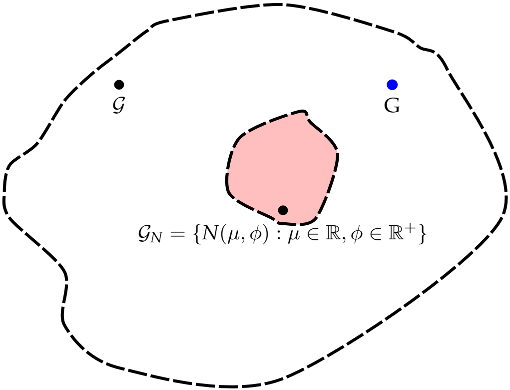
How to do inference on G?

- Parametric Inference: $G = G_\omega$.

$$G_\omega = N(\mu, \phi), \quad \omega = (\mu, \phi)$$

  - Prior over $\omega \rightarrow$ prior over G_ω.
  - Posterior over $\omega \rightarrow$ posterior over G_ω
- Nonparametric Inference: fewer restrictions on form of G.

What is Nonparametric Bayes?



What is Nonparametric Bayes?

Consider the regression problem

$$Y_i \mid X_i \sim N\Big(f(X_i), \phi\Big)$$

$$E[Y \mid X] = f(X)$$

How to do inference on f?

What is Nonparametric Bayes?

Consider the regression problem

$$Y_i \mid X_i \sim N\Big(f(X_i), \phi\Big)$$

$$E[Y \mid X] = f(X)$$

How to do inference on f?

- Parametric Inference: $f = f_\omega$.

$$f_\omega(X) = X'\omega$$

What is Nonparametric Bayes?

Consider the regression problem

$$Y_i \mid X_i \sim N\Big(f(X_i), \phi\Big)$$

$$E[Y \mid X] = f(X)$$

How to do inference on f?

- Parametric Inference: $f = f_\omega$.

$$f_\omega(X) = X'\omega$$

  - Prior over $\omega \to$ prior over f_ω.
  - Posterior over $\omega \to$ posterior over f_ω

What is Nonparametric Bayes?

Consider the regression problem

$$Y_i \mid X_i \sim N\Big(f(X_i), \phi\Big)$$

$$E[Y \mid X] = f(X)$$

How to do inference on f?

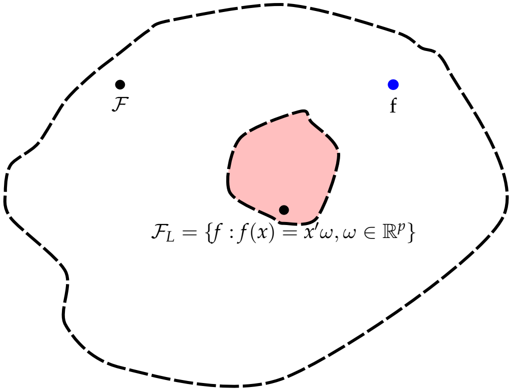
- Parametric Inference: $f = f_\omega$.

$$f_\omega(X) = X'\omega$$

  - Prior over $\omega \rightarrow$ prior over f_ω.
  - Posterior over $\omega \rightarrow$ posterior over f_ω
- Nonparametric Inference: impose less structure on f.

What is Nonparametric Bayes?

Inference for unknown function $f(x)$,



What is Nonparametric Bayes?

Our working definition for this talk

- Nonparametric priors: over infinite dimensional spaces.
- Parametric priors: over finitely dimensional spaces.

Key advantages:

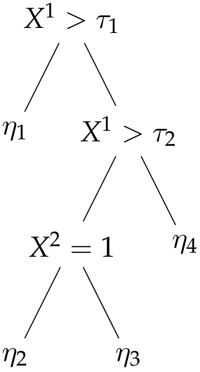
- Flexible: avoids restrictive modeling assumptions.
- Principled: well-defined, probabilistically valid shrinkage.
- Uncertainty quantification: not just point estimation.

Some Nonparametric Priors...

- **BART**: prior over tree functions.
- **Chinese Restaurant Process**: prior over partitions.
- **Dirichlet Process**: prior over probability distributions.
- **Gaussian Process**: prior over functions.
- **Gamma Process**: prior over non-decreasing functions.
- Much, much, more ...

Bayesian Additive Regression Trees

Tree T_j with terminal node parameters $M_j = \{\eta_1, \eta_2, \eta_3, \eta_4\}$



Bayesian Additive Regression Trees

Prior on tree depth, d: Probability that node at depth d is non-terminal

$$\frac{\alpha}{(1+d)^{\beta}}$$

For $\alpha \in (0,1)$ and $\beta \geq 0$

Bayesian Additive Regression Trees

Inference for unknown regression function $f(X)$,

$$f(X) = \sum_{j=1}^{m} f_j\left(X; T_j, M_j\right)$$

- Sum of m trees.
- $f_j(\cdot)$ maps X to some $\eta_k \in M_j$ of tree T_j

Bayesian Additive Regression Trees

Inference for unknown regression function $f(X)$,

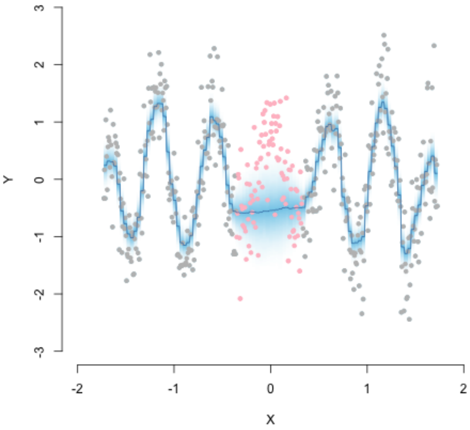
$$f(X) = \sum_{j=1}^{m} f_j\left(X; T_j, M_j\right)$$

- Sum of m trees.
- $f_j(\cdot)$ maps X to some $\eta_k \in M_j$ of tree T_j

Putting it all together, we say

$$f \sim BART$$

Bayesian Additive Regression Trees



Prior Over Partitions

Consider problem of clustering $\{y_{1:n}\}$

$$y_i \mid \mu_{c_i}, c_i \sim N(\mu_{c_i}, \phi)$$

$$\mu_{c_i} \sim G$$

- Want inference on $c_{1:n} = (c_1, c_2, \ldots, c_n)$.
- How to specify a prior over $c_{1:n}$?

Prior Over Partitions

Consider problem of clustering $\{y_{1:n}\}$

$$y_i \mid \mu_{c_i}, c_i \sim N(\mu_{c_i}, \phi)$$

$$\mu_{c_i} \sim G$$

- Want inference on $c_{1:n} = (c_1, c_2, \ldots, c_n)$.
- How to specify a prior over $c_{1:n}$?
- Equivalently: how to specify a partition of set $\{1, 2, \ldots, n\}$.

Chinese Restaurant Process

$$
\begin{aligned}
p(c_{1:n}) &= p(c_1)p(c_2 \mid c_1)p(c_3 \mid c_1, c_2)\ldots p(c_n \mid c_{1:n-1}) \\
&= p(c_1)\prod_{i=2}^{n} p(c_i \mid c_{1:i-1})
\end{aligned}
$$

Chinese Restaurant Process

$$
\begin{aligned}
p(c_{1:n}) &= p(c_1)p(c_2 \mid c_1)p(c_3 \mid c_1, c_2)\dots p(c_n \mid c_{1:n-1}) \\
&= p(c_1)\prod_{i=2}^{n} p(c_i \mid c_{1:i-1})
\end{aligned}
$$

- $c_1 = 1$
- For $i > 1$,

$$
p(c_i = j \mid c_{1:i-1}) = \begin{cases} \frac{n_{i-1,j}}{\alpha+i-1} & j \in c_{1:i-1} \\ \frac{\alpha}{\alpha+i-1} & j \notin c_{1:i-1} \end{cases}
$$

Chinese Restaurant Process

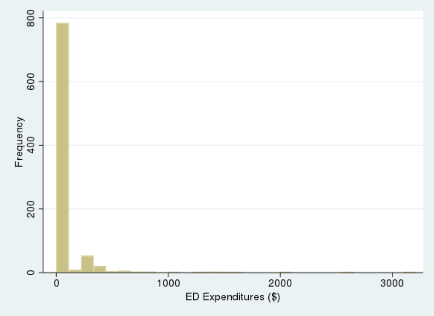
This process generates a random partition.

$$c_{1:n} \sim CRP(\alpha)$$

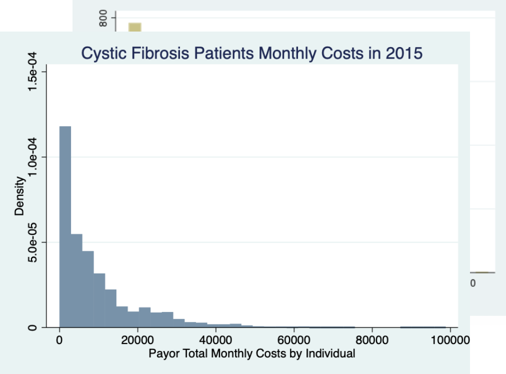
To Recap...

Parametric Bayes		Nonparametric Bayes
	Models	
Low-dimensional	→	High-dimensional
	Priors	
Distributions	→	Processes

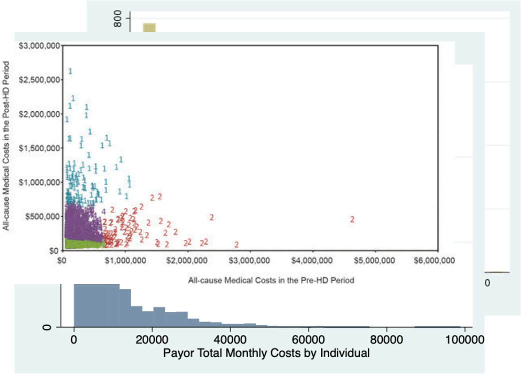
Application to Medical Cost Modeling



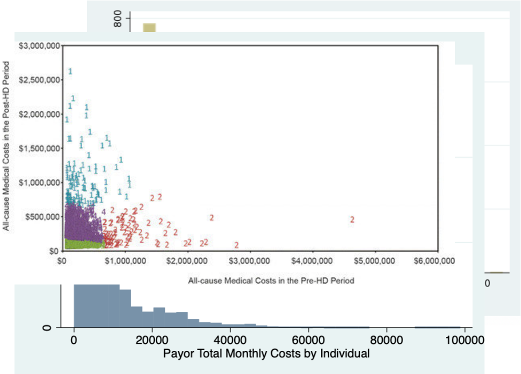
Application to Medical Cost Modeling



Application to Medical Cost Modeling



Application to Medical Cost Modeling



What are our options?

Observed data $D = \left\{ Y_i, X_i = (A_i, L_i) \right\}_{i=1:n}$

What are our options?

Observed data $D = \left\{ Y_i, X_i = (A_i, L_i) \right\}_{i=1:n}$

$$Y_i \mid X_i, \omega \sim \pi\left(X_i'\gamma\right)\delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i'\gamma\right)\right) \cdot p_+\left(Y_i \mid X_i, \theta\right)$$

What are our options?

Observed data $D = \left\{Y_i, X_i = (A_i, L_i)\right\}_{i=1:n}$

$$Y_i \mid X_i, \omega \sim \pi\left(X_i'\gamma\right)\delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i'\gamma\right)\right) \cdot p_+\left(Y_i \mid X_i, \theta\right)$$

Define $\omega = (\gamma, \theta)$

What are our options?

Observed data $D = \left\{ Y_i, X_i = (A_i, L_i) \right\}_{i=1:n}$

$$Y_i \mid X_i, \omega \sim \pi\left(X_i'\gamma\right)\delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i'\gamma\right)\right) \cdot p_+\left(Y_i \mid X_i, \theta\right)$$

Define $\omega = (\gamma, \theta)$

- $\pi(\cdot) = expit(\cdot)$ or $\pi(\cdot) = \Phi(\cdot)$

What are our options?

$$Y_i \mid X_i \sim \pi\left(X_i'\gamma\right)\delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i'\gamma\right)\right) \cdot p\left(Y_i \mid X_i, \theta\right)$$

Define $\omega = (\gamma, \theta)$

- E.g. $p_+\left(Y_i \mid X_i, \theta\right)$ is Normal with $\theta = (\beta, \phi)$

$$E[Y \mid X_i, \beta] = X_i'\beta$$

What are our options?

$$Y_i \mid X_i \sim \pi\left(X_i'\gamma\right)\delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i'\gamma\right)\right) \cdot p\left(Y_i \mid X_i, \theta\right)$$

Define $\omega = (\gamma, \theta)$

- E.g. $p_+\left(Y_i \mid X_i, \theta\right)$ is Normal with $\theta = (\beta, \phi)$

$$E[Y \mid X_i, \beta] = X_i'\beta$$

- E.g. $p_+\left(Y_i \mid X_i, \theta\right)$ is log-Normal with $\theta = (\beta, \phi)$

$$E[Y \mid X_i, \beta] = \exp\left(X_i'\beta + \frac{\phi}{2}\right)$$

A Nonparametric Approach

$$Y_i \mid X_i, \omega \sim \pi\left(X_i'\gamma\right)\delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i'\gamma\right)\right) \cdot p_+\left(Y_i \mid X_i, \theta\right)$$

- Very restrictive structure assumed.
- Are covariate effects really linear? additive?
- Interactions? Multimodality? Skewness?

Go high-dimensional: $\omega \rightarrow \omega_i = (\gamma_i, \theta_i)$

$$Y_i \mid X_i, \omega_i \sim \pi\left(X_i' \gamma_i\right) \delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i' \gamma_i\right)\right) \cdot p_+\left(Y_i \mid X_i, \theta_i\right)$$

A Bayesian Nonparametric Approach...

Go high-dimensional: $\omega \to \omega_i = (\gamma_i, \theta_i)$

$$Y_i \mid X_i, \omega_i \sim \pi\left(X_i'\gamma_i\right)\delta_0\left(Y_i\right) + \left(1 - \pi\left(X_i'\gamma_i\right)\right) \cdot p_+\left(Y_i \mid X_i, \theta_i\right)$$

$$\omega_{1:n} \mid G \sim G$$

A Nonparametric Prior over Distributions

$$\omega_1, \ldots, \omega_n \mid G \sim G$$

Nonparametric prior on G

A Nonparametric Prior over Distributions

$$\omega_1, \ldots, \omega_n \mid G \sim G$$

Nonparametric prior on G

$$G \mid \alpha, G_0 \sim DP(\alpha G_0)$$

A Nonparametric Prior over Distributions

$$\omega_1, \ldots, \omega_n \mid G \sim G$$

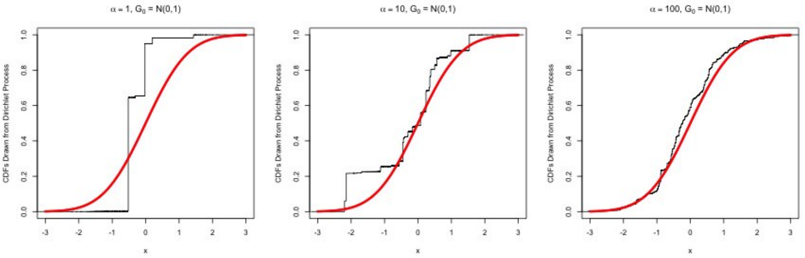
Nonparametric prior on G

$$G \mid \alpha, G_0 \sim DP(\alpha G_0)$$

- Each "realization" or "draw" is a random distribution.
- G_0: mean of these realizations.
- α: controls dispersion/spread around G_0.
- Distributions draw from the DP are discrete.

The Dirichlet Process

$$G \mid \alpha, G_0 \sim DP(\alpha G_0)$$



Pólya Urn Process

Conditional posterior of i^{th} subject

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

Pólya Urn Process

Conditional posterior of i^{th} subject

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

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

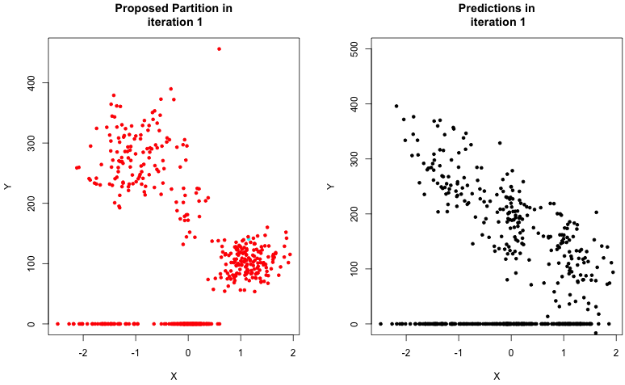
MCMC Inference via auxiliary parameters

Auxiliary variable scheme via $c_{1:n}^{(m)}$ at iteration m.

1. Update $c_{1:n}^{(m)} \mid \omega_{1:n}^{(m-1)}, D$.
2. Update $\omega_{1:n}^{(m)}, \mid c_{1:n}^{(m)}, D$.

Output: Posterior draws $\left\{\omega_{1:n}^{(m)}, c_{1:n}^{(m)}\right\}_{1:M}$

MCMC Inference via auxiliary parameters



Data Description

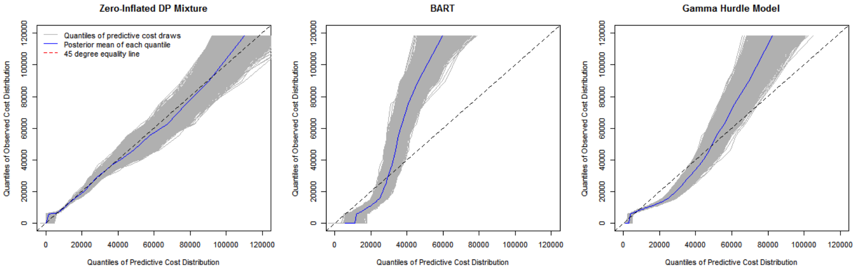
- Data source: SEER-Medicare.
- Endometrial cancer patients (N≈1,000).
- Treatment: post-hysterectomy radiation vs. chemotherapy.
- Outcome: Total inpatient costs over 2 years.
  - Skewed, zero-inflated
  - Chemo arm: 15% zeros; RT arm: 8%
- Covariates: tumor grade, cancer stage, CCI.

Sample Characteristics

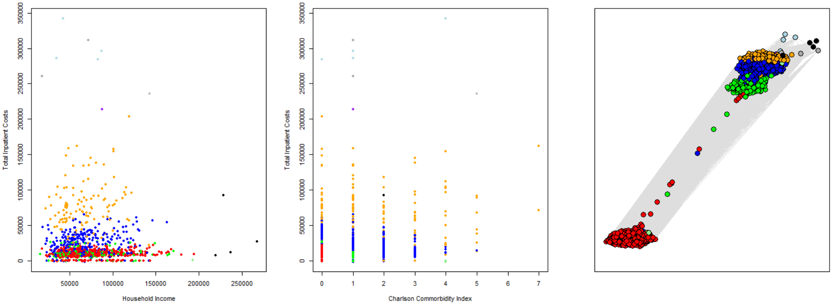
	Chemotherapy (n=92)	Radiation Therapy (n=952)	SMD
Total Inpatient Costs ($)	22131.59 (28608.07)	23370.63 (34453.31)	.039
Zero Costs	14 (15.2%)	75 (7.9%)	
Age (years)	73.68 (6.98)	73.25 (5.98)	.066
Household Income ($)	64368.36 (32422.55)	56785.29 (26166.79)	.257
White	76 (82.6%)	835 (87.8%)	.147
Diabetic	20 (21.7%)	197 (20.7%)	.026
CCI			.350
0	49 (53.3%)	529 (55.6%)	
1	22 (23.9%)	260 (27.3%)	
≥ 2	21 (22.8%)	131 (13.8%)	
Grade = 1	28 (30.4%)	208 (21.8%)	.196
FIGO Stage I-N0 or I-A	63 (68.5%)	357 (37.5%)	.653

Notes: Means and standard deviations are reported for continuous variables. Counts and percentages are reported for categorical variables. All monetary amounts are in 2018 U.S. Dollars.

Predictions



Posterior Clustering



Thank you!