Bayesian Nonparametric Regression through Mixture Models

Sara Wade

Bocconi University
Advisor: Sonia Petrone

October 7, 2013
Outline

1. Introduction
2. Enriched Dirichlet Process
3. EDP Mixtures for Regression
4. Restricted DP Mixtures for Regression
5. Normalized Covariate Dependent Weights
6. Discussion
1 Introduction

2 Enriched Dirichlet Process

3 EDP Mixtures for Regression

4 Restricted DP Mixtures for Regression

5 Normalized Covariate Dependent Weights

6 Discussion
Aim: flexible regression, where both the mean and error distribution may evolve flexibly with $x$. 
• Aim: flexible regression, where both the mean and error distribution may evolve flexibly with $x$.

• Many approaches for flexible regression (ex. splines, wavelets, Gaussian processes), but typically assume iid Gaussian errors,

$$Y_i = m(x_i) + \epsilon_i, \quad \epsilon_i | \sigma^2 \sim iid \ N(0,\sigma^2).$$
• Aim: **flexible regression**, where both the mean and error distribution may evolve flexibly with $x$.

• Many approaches for flexible regression (ex. splines, wavelets, Gaussian processes), but typically assume iid Gaussian errors,

$$Y_i = m(x_i) + \epsilon_i, \quad \epsilon_i|\sigma^2 \sim \text{N}(0, \sigma^2).$$

• For i.i.d. data, mixtures are a useful tool for flexible density estimation.
• **Aim**: flexible regression, where both the mean and error distribution may evolve flexibly with $x$.

• Many approaches for flexible regression (ex. splines, wavelets, Gaussian processes), but typically assume i.i.d Gaussian errors,

$$ Y_i = m(x_i) + \epsilon_i, \quad \epsilon_i | \sigma^2 \overset{iid}{\sim} N(0, \sigma^2). $$

• For i.i.d. data, mixtures are a useful tool for flexible density estimation.

$\Rightarrow$ Extend mixture models for conditional density estimation.
• Given $f$, assume $Y_i$ are iid with density

$$f_p(y) = \int K(y|\theta) dP(\theta),$$

for some parametric density $K(y|\theta)$. 
Mixture Models

- Given $f$, assume $Y_i$ are iid with density

$$f_P(y) = \int K(y|\theta) dP(\theta),$$

for some parametric density $K(y|\theta)$.

- Bayesian setting, define a prior on $P$, where typically $P$ is discrete a.s.,

$$P = \sum_{j=1}^{J} w_j \delta_{\theta_j}, \quad f_P(y) = \sum_{j=1}^{J} w_j K(y|\theta_j).$$
Mixture Models

- Given $f$, assume $Y_i$ are iid with density
  
  $$f_P(y) = \int K(y|\theta) dP(\theta),$$
  
  for some parametric density $K(y|\theta)$.
- Bayesian setting, define a prior on $P$, where typically $P$ is discrete a.s.,
  
  $$P = \sum_{j=1}^{J} w_j \delta_{\theta_j}, \quad f_P(y) = \sum_{j=1}^{J} w_j K(y|\theta_j).$$
- How many components?
  - Fixed $J$: overfitted mixtures (Rousseau and Mengersen (2011)) or post-processing techniques.
  - Random $J$: reversible jump MCMC.
  - $J = \infty$: Dirichlet process (DP) mixtures and developments in BNP.
Two main lines of research:

- **Joint approach**: Augment the observations to include $X$ and assume that given $f_P$, the $(Y_i, X_i)$ are iid with density

$$f_P(y, x) = \int K(y, x|\theta)dP(\theta) = \sum_{j=1}^{\infty} w_j K(y, x|\theta_j).$$

Conditional density estimates are obtained as a by-product through

$$f_P(y|x) = \frac{\sum_{j=1}^{\infty} w_j K(y, x|\theta_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x|\theta_{j'}).}$$

- **Conditional approach**: Allow the mixing measure to depend on $x$ and assign a prior so that $\{P_x\}_{x \in \mathcal{X}}$ are dependent. Assume that given $f_{P_x}$ and $x$, the $Y_i$ are independent with density

$$f_{P_x}(y|x) = \int K(y|\theta)dP_x(\theta) = \sum_{j=1}^{\infty} w_j(x) K(y|\theta_j(x)).$$
Many proposals in literature fall into these two general categories:

1. **Joint Approach:** Müller, Erkanli, and West (1996), Shahbaba and Neal (2009), Hannah, Blei, and Powell (2011), Park and Dunson (2010), Müller and Quintana (2010), Bhattacharya, Page, and Dunson (2012), Quintana (2011)...

2. **Conditional Approach:**
   a. **Covariate-dependent atoms:** MacEachern (1999; 2000; 2001), Delorio et al. (2004), Gelfand, Kottas, and MacEachern (2005), Rodriguez and Horst (2008), De La Cruz, Quintana, and Müller (2007), De Iorio et al. (2009), Jara et al. (2010)...
      **Example:** \( f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j N(y|\mu_j(x), \sigma_j^2) \).
   b. **Covariate-dependent weights:** Griffin and Steele (2006; 2010), Dunson and Park (2008), Ren et al. (2011), Chung and Dunson (2009), Rodriguez and Dunson (2011)...
      **Example:** \( f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) N(y|x \beta_j, \sigma_j^2) \).
• How to choose among the large number of proposals for the application at hand?
How to choose among the large number of proposals for the application at hand?

- Computational reasons

Our aim is to answer this question from a Bayesian perspective. We do so through a detailed study of predictive performance based on finite samples, identifying potential sources of improvement, and developing novel models to improve predictions.
Mixture Models for regression

- How to choose among the large number of proposals for the application at hand?
  - Computational reasons

- Our aim is to answer this question from a Bayesian perspective.
  We do so through
  - a detailed study of predictive performance based on finite samples,
  - identifying potential sources of improvement,
  - developing novel models to improve predictions.
Motivating Application

• Motivating application: study of Alzheimer’s Disease based on neuroimaging data.

• Definite diagnosis isn’t known until autopsy, and current diagnosis is based on clinical information.

• Aim: show that neuroimages can be used to improve diagnostic accuracy and may be better outcomes measures for clinical trials particularly in early stages of the disease.

• Prior knowledge, need for dimension reduction, non-homogenity $\rightarrow$ BNP approach.
1 Introduction

2 Enriched Dirichlet Process

3 EDP Mixtures for Regression

4 Restricted DP Mixtures for Regression

5 Normalized Covariate Dependent Weights

6 Discussion
In many applications, the DP is a prior for a multivariate random probability measure.

DP has many desirable properties (easy elicitation of parameters: $\alpha, P_0$, conjugacy, large support).

However, when considering set of probability measures on $\mathbb{R}^k$, a DP prior can be quite restrictive in the sense that the variability is determined by a single parameter, $\alpha$.

*Enriched conjugate priors* (Consonni and Veronese (2001)) have been proposed to address an analogous lack of flexibility of standard conjugate priors in a parametric setting.

Solution: Construct a enriched DP that is more flexible, yet still conjugate by extending the idea of enriched conjugate priors to a nonparametric setting.
EDP: Definition

- **Sample space**: $\Theta \times \Psi$, where $\Theta$, $\Psi$ are complete separable metric spaces.

- **Enriched Dirichlet Process**: $P_0$ is a probability measure on $\Theta \times \Psi$, $\alpha_\theta > 0$, and $\alpha_\psi(\theta) > 0 \ \forall \theta \in \Theta$. Assume:
  1. $P_\theta \sim \text{DP}(\alpha_\theta P_0 \theta)$.
  2. $\forall \theta \in \Theta$, $P_\psi|\theta(\cdot|\theta) \sim \text{DP}(\alpha_\psi(\theta)P_0 \psi|\theta(\cdot|\theta))$.
  3. $P_\psi|\theta(\cdot|\theta)$, $\theta \in \Theta$ are independent among themselves.
  4. $P_\theta$ is independent of $\{P_\psi|\theta(\cdot|\theta)\}_{\theta \in \Theta}$.

The law of the random joint is obtained from the joint law of the marginal and conditionals through the mapping 
$(P_\theta, P_\psi|\theta) \rightarrow \int (.) P_\psi|\theta(\cdot|\theta)dP_\theta(\theta)$.

- Many more precision parameters, yet conjugacy and many other desirable properties maintained.
Outline

1. Introduction
2. Enriched Dirichlet Process
3. EDP Mixtures for Regression
4. Restricted DP Mixtures for Regression
5. Normalized Covariate Dependent Weights
6. Discussion
Summary

• DP mixtures models are popular tools in BNP - known properties and analytically computable urn scheme.

• However, as $p$ increases, $x$ tends to dominate the posterior of the random partition, negatively affecting the prediction.

• **Solution**: replace the DP with the EDP, to allow local clustering, yet maintain an analytically computable urn scheme.
Joint DP mixture model

- Model (Müller et al. (1996)):
  
  \[ f_P(y, x) = \int N(y, x | \mu, \Sigma) dP(\mu, \Sigma), \quad P \sim \text{DP}(\alpha P_0), \]

  where \( P_0 \) is the conjugate normal-inverse Wishart.

- Problems: requires sampling the full \( p + 1 \times p + 1 \) covariance matrix and the inverse Wishart is poorly parametrized.

- Improvements: 1) eases computations, 2) more flexible prior for the variances, 3) flexible local correlations between \( Y \) and \( X \), and 4) easy incorporation of other types of \( Y \) and \( X \).
Joint DP mixture model

- Model (Müller et al. (1996)):
  \[ f_P(y, x) = \int N(y, x | \mu, \Sigma) dP(\mu, \Sigma), \quad P \sim DP(\alpha P_0), \]
  where \( P_0 \) is the conjugate normal-inverse Wishart.

- Problems: requires sampling the full \( p + 1 \times p + 1 \) covariance matrix and the inverse Wishart is poorly parametrized.

- Model (Shahbaba and Neal (2009)):
  \[ f_P(y, x) = \int N(y | x \beta, \sigma_y^2 \) \prod_{h=1}^{p} N(x | \mu_h, \sigma_x^2) dP(\theta, \psi), \]
  \[ P \sim DP(\alpha P_0 \theta \times P_0 \psi), \]
  where \( \theta = (\beta, \sigma_y^2 \) \), \( \psi = (\mu, \sigma_x^2 \), and \( P_0 \theta, P_0 \psi \) are the conjugate normal-inverse gamma priors.

- Improvements: 1) eases computations, 2) more flexible prior for the variances, 3) flexible local correlations between \( Y \) and \( X \), and 4) easy incorporation of other types of \( Y \) and \( X \).
Examining the Partition

- $P$ is discrete a.s. → a random partition of the subjects.
- Notation: $\rho_n = (s_1, \ldots, s_n)$-(partition), $(\theta_j^*, \psi_j^*)$-(unique parameters), $x_j^*, y_j^*$-(cluster-specific observations).
- Partition:

$$p(\rho_n|x_1:n, y_1:n) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) \prod_{j=1}^{k} \prod_{h=1}^{p} g_{x,h}(x_j^*,h) \ast g_y(y_j^*|x_j^*),$$

where, for ex. $g_y(y_j^*|x_j^*) = \int \prod_{i:s_i=j} K(y_i|x_i, \theta) dP_{0 \theta}(\theta)$. 
Examining the Partition

• $P$ is discrete a.s. $\rightarrow$ a random partition of the subjects.

• Notation: $\rho_n = (s_1, \ldots, s_n)$-(partition), $(\theta^*_j, \psi^*_j)$-(unique parameters), $x_j^*, y_j^*$-(cluster-specific observations).

• Partition:

$$p(\rho_n | x_1:n, y_1:n) \propto \alpha^k \prod_{j=1}^{k} \Gamma(n_j) \prod_{j=1}^{k} \prod_{h=1}^{p} g_{x,h}(x_j^*,h) * g_y(y_j^*|x_j^*),$$

where, for ex. $g_y(y_j^*|x_j^*) = \int \prod_{i:s_i=j} K(y_i|x_i, \theta) dP_0 \theta(\theta)$.

• As $p$ increases,
  $\rightarrow$ $x$ dominates the partition structure.
  $\rightarrow$ $x$ exhibits increasing departures.
  $\rightarrow$ large $k$ and small $n_1, \ldots, n_k$ with high posterior probability.
Examining the Prediction

- How does this affect prediction?

\[
E[Y_{n+1}, y_{1:n}, x_{1:n+1}] = \sum_{\mathcal{P}_n} \int \tilde{w}_{k+1}(x_{n+1})x_{n+1} \beta_0 + \sum_{j=1}^{k} \tilde{w}_j(x_{n+1})x_{n+1} \beta_j^* \text{d}P(\rho_n, \theta^*, \phi^* | y_{1:n}, x_{1:n+1}),
\]

\[
\tilde{w}_j(x_{n+1}) = c^{-1} n_j \prod_{h=1}^{p} N(x_{n+1,h} | \mu^*_j, h, \sigma_{x,j,h}^2) \text{ for } j = 1, \ldots, k,
\]

\[
\tilde{w}_{k+1}(x_{n+1}) = c^{-1} \alpha g_x(x_{n+1}).
\]

- As \( p \) increases,
  \( \rightarrow \) given the partition, predictive estimates are an average over the large number of components.
  \( \rightarrow \) within component, predictive estimates are unreliable and variable due to small sample sizes.
  \( \rightarrow \) rigid measure of similarity in the covariate space.
• Model:

\[ f_p(y, x) = \int N(y|x\beta, \sigma^2_{y|x}) \prod_{h=1}^{p} N(x|\mu_h, \sigma^2_{x,h}) dP(\theta, \psi), \]

\[ P \sim \text{EDP}(\alpha_\theta, \alpha_\psi(\cdot), P_0\theta \times P_0\psi), \]
An EDP mixture model for regression (with S. Petrone)

- Model:

$$f_P(y, x) = \int \mathcal{N}(y|\mu_\beta, \sigma_x^2) \prod_{h=1}^{p} \mathcal{N}(x|\mu_h, \sigma_{x,h}^2) dP(\theta, \psi),$$

$$P \sim \text{EDP}(\alpha_\theta, \alpha_\psi(\cdot), P_0 \theta \times P_0 \psi),$$

- Defines a nested partition structure $\rho_n = (\rho_{n,y}, (\rho_{n_{j+},x}))$, where $k$ represents the number of $y$-clusters of sizes $n_{j+}$, and $k_j$ represents the number of $x$-clusters within the $j^{th}$ $y$-cluster of sizes $n_{j,1}$.

- Additional notation: $(\theta^*_j, \psi^*_j, l)$-(unique parameters), $(y^*_j, x^*_j, l)$-(cluster specific observations).
• Under the assumption that $\alpha_\psi(\theta) = \alpha_\psi$ for all $\theta \in \Theta$,

$$
p(\rho_n | x_{1:n}, y_{1:n}) \propto \alpha_\theta^k \prod_{j=1}^k \alpha_\psi^{k_j} \frac{\Gamma(\alpha_\psi) \Gamma(n_{j+})}{\Gamma(\alpha_\psi + n_{j+})} \prod_{l=1}^{k_j} \Gamma(n_{j,l}) \prod_{j=1}^k g_y(y_j^* | x_j^*) \prod_{l=1}^{k_j} \prod_{h=1}^p g_x(x_j^*,l,h).
$$

→ likelihood of $y$ determines if a coarser $y$-partition is present
→ smaller $k$ and larger $n_{1+}, \ldots, n_{k+}$ with high posterior probability.
Examining the Prediction

- **Prediction:**

\[
E[Y_{n+1} | y_{1:n}, x_{1:n+1}]
\]

\[
= \sum_{\mathcal{P}_n} \int \tilde{w}_{k+1}(x_{n+1})x_{n+1}\beta_0 + \sum_{j=1}^{k} \tilde{w}_j(x_{n+1})x_{n+1}\beta_j^* dP(\rho_n, \theta^*, \phi^* | y_{1:n}, x_{1:n+1}),
\]

\[
\tilde{w}_j(x_{n+1}) = c_1^{-1} n_j + \left( \frac{\alpha_x}{\alpha_x + n_j} \right) g_x(x_{n+1}) + \sum_{l=1}^{k_j} \frac{n_j,l}{\alpha_x + n_j} \prod_{h=1}^{p} N(x_{n+1,h} | \mu_{l|j,h}, \sigma_{x,l|j,h}^2),
\]

\[
\tilde{w}_{k+1}(x_{n+1}) = c_1^{-1} \alpha_y g_x(x_{n+1}).
\]

→ given the partition, predictive estimates are an average over a **smaller** number of components.

→ within component, predictive estimates are based on **larger** sample sizes.

→ **flexible** measure of similarity in the covariate space.
Examining the Prediction

• Prediction:

\[
E[Y_{n+1}|, y_{1:n}, x_{1:n+1}]
\]

\[
= \sum_{\mathcal{P}_n} \int \tilde{w}_{k+1}(x_{n+1})x_{n+1} \beta_0 + \sum_{j=1}^{k} \tilde{w}_j(x_{n+1})x_{n+1} \beta_j^* dP(\rho_n, \theta^*, \phi^* | y_{1:n}, x_{1:n+1}),
\]

\[
\tilde{w}_j(x_{n+1}) = c_1^{-1} n_j^+ \left( \frac{\alpha_x}{\alpha_x + n_j^+} g_x(x_{n+1}) + \sum_{l=1}^{k_j} \frac{n_{j,l}}{\alpha_x + n_j^+} \prod_{h=1}^{p} N(x_{n+1}, h | \mu^*_{l,j,h}, \sigma^2_{x, l,j,h} ) \right),
\]

\[
\tilde{w}_{k+1}(x_{n+1}) = c_1^{-1} \alpha_y g_x(x_{n+1}).
\]

\[\rightarrow\] given the partition, predictive estimates are an average over a smaller number of components.

\[\rightarrow\] within component, predictive estimates are based on larger sample sizes.

\[\rightarrow\] flexible measure of similarity in the covariate space.

• Summary: More reliable and less variable estimates, yet computations remain relatively simple due to analytically computable urn scheme.
Application to AD data

- **Aim:** Diagnosis of AD.
- **Data:** Y-binary indicator of CN (1) or AD (0); X-summaries of structural Magnetic Resonance Images (sMRI) \( (p = 15) \).
- **Model:**

\[
Y_i | x_i, \beta_i \overset{\text{ind}}{\sim} \text{Bern}(\Phi((x_i \beta_i) / \sigma_{y,i})),
\]

\[
X_i | \mu_i, \sigma_{x,i}^2 \overset{\text{ind}}{\sim} \prod_{h=1}^{p} \text{N}(\mu_{i,h}, \sigma_{x,i,h}^2),
\]

\[
\beta_i, \mu_i, \sigma_{x,i}^2 | P \overset{i.i.d.}{\sim} P,
\]

\[
P \sim Q.
\]

- **Q** = DP or EDP with the same base measure and gamma hyperpriors on the precision parameters.
Results for $n = 185$

- Posterior mode of $k$: 16 (DP) vs. 3 (EDP).
• Posterior mode of $k$: 16 (DP) vs. 3 (EDP).

(c) DP  
(d) EDP

• Prediction for $m = 192$ subjects:
  • Number of subjects correctly classified: 159 (DP) vs. 168 (EDP).
  • Tighter credible intervals.
1 Introduction

2 Enriched Dirichlet Process

3 EDP Mixtures for Regression

4 Restricted DP Mixtures for Regression

5 Normalized Covariate Dependent Weights

6 Discussion
Aim: Analyse the role of the partition in prediction for BNP mixtures and highlight an understated problem: huge dimension of the partition space.
• Aim: Analyse the role of the partition in prediction for BNP mixtures and highlight an understated problem: huge dimension of the partition space.
• Assume both $Y$ and $X$ are univariate and continuous.
• Aim: Analyse the role of the partition in prediction for BNP mixtures and highlight an understated problem: huge dimension of the partition space.
• Assume both $Y$ and $X$ are univariate and continuous.
• Number of ways to partition the $n$ subjects:

$$B_n = \sum_{k=1}^{n} S_{n,k},$$

$$S_{n,k} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^n.$$

• Many of these partitions similar $\rightarrow$ large number of partitions will fit the data.
Covariates typically provide information on the partition structure.

Incorporating this covariate information (as for models with $w_j(x)$) will improve posterior inference on partition.

If the aim is estimation of regression function, desirable partitions satisfy an ordering constraint of $(x_i) \rightarrow$. This reduces the total number of partitions to $2^n - 1$.

Note: if $n = 10$, $B_{10} = 115,975$ and $2^{10} - 1 = 512,000$, which is 44% of the total partitions, and if $n = 100$, $2^{100} - 1 / B_{100} < 10^{16}$.5
• Covariates typically provide information on the partition structure.
• Incorporating this covariate information (as for models with $w_j(x)$) will improve posterior inference on partition.
• But posterior is still diluted $\rightarrow$ this information needs to be strictly enforced.
• Covariates typically provide information on the partition structure.

• Incorporating this covariate information (as for models with $w_j(x)$) will improve posterior inference on partition.

• But posterior is still diluted $\rightarrow$ this information needs to be strictly enforced.

• If aim is estimation of regression function, desirable partitions satisfy an ordering constraint of $(x_i) \rightarrow$ Reduces the total number of partitions to $2^{n-1}$.

• Note: if $n=10$, $B_{10} = 115,975$ and $2^{10-1} = 512$, 0.44% of the total partitions, and if $n = 100$, $2^{100-1}/B_{100} < 10^{-85}$. 
• Let $x_1, \ldots, x_n$ denote the ordered values of $x$, and $s_1, \ldots, s_n$ denote the corresponding values of $s_1, \ldots, s_n$.

• Prior for the random partition:

$$p(\rho_n|x) = \frac{\Gamma(\alpha)\Gamma(n+1)}{\Gamma(\alpha + n)} \frac{\alpha^k}{k!} \prod_{j=1}^{k} \frac{1}{n_j} * l_{s_1 \leq \cdots \leq s_n}.$$ 

• Introduces covariate dependency, greatly reduces the total number of partitions, improves prediction, and speeds up computations.
Data are simulated from:

\[ y_i \mid x_i = \begin{cases} 
-x_i/8 + 5 & \text{if } x_i \leq 6 \\
2x_i - 12 & \text{if } x_i > 6,
\end{cases} \]

\[ x_i = (0, 0.25, 0.5, \ldots, 8.75, 9). \]
Simulated Ex.: 3 partitions with highest posterior mass

No. of partitions with positive post. prob.: 1, 2, 3, 4, 5, 6

\( p(\rho | y, x) = 0.3973 \)  \( p(\rho | y, x) = 0.5317 \)  \( p(\rho | y, x) = 0.9031 \)

\( p(\rho | y, x) = 0.0493 \)  \( p(\rho | y, x) = 0.0695 \)  \( p(\rho | y, x) = 0.0381 \)

\( p(\rho | y, x) = 0.0129 \)  \( p(\rho | y, x) = 0.0418 \)  \( p(\rho | y, x) = 0.0129 \)

(e) DPM  (f) jDPM  (g) rDPM

respectively.
Simulated Example: Prediction

\[ \text{Figure: Prediction for } x = (3.3, 5.9, 6.2, 6.3, 7.9, 8.1, 10). \]

The empirical $L_2$ prediction error,

\[ \left( \frac{1}{m} \sum_{j=1}^{m} (\hat{y}_{m,j,\text{est}} - \hat{y}_{m,j,\text{true}})^2 \right)^{1/2}, \]

for the three models, respectively, is 2.19, 1.66, and 1.09.
BNP mixture models with covariate dependent weights:

\[ f_{P_x}(y|x) = \int K(y|x, \theta) dP_x(\theta), \quad P_x = \sum_{j=1}^{\infty} w_j(x) \delta_{\theta_j}. \]

Construction of covariate dependent weights (in literature):

\[ w_j(x) = v_j(x) \prod_{j'<j} (1 - v_{j'}(x)), \]

where \( v_j(\cdot) \) are independent stochastic processes.
Introduction

- BNP mixture models with covariate dependent weights:
  \[ f_{P_X}(y|x) = \int K(y|x, \theta) dP_X(\theta), \quad P_X = \sum_{j=1}^{\infty} w_j(x) \delta_{\theta_j}. \]

- Construction of covariate dependent weights (in literature):
  \[ w_j(x) = v_j(x) \prod_{j' < j} (1 - v_{j'}(x)), \]
  where \( v_j(\cdot) \) are independent stochastic processes.

- Drawbacks:
  - non-interpretatable weight function, typically exhibiting peculiarities.
  - choice for functional shapes and hyper-parameters.
  - extensions for continuous and discrete covariates are not straightforward.
BNP mixture models with covariate dependent weights:

\[
f_{P_x}(y|x) = \int K(y|x, \theta) dP_x(\theta), \quad P_x = \sum_{j=1}^{\infty} w_j(x) \delta_{\theta_j}.
\]

Construction of covariate dependent weights (in literature):

\[
w_j(x) = v_j(x) \prod_{j' < j} (1 - v_{j'}(x)),
\]

where \(v_j(\cdot)\) are independent stochastic processes.

Drawbacks:

- non-interpretable weight function, typically exhibiting peculiarities.
- choice for functional shapes and hyper-parameters.
- extensions for continuous and discrete covariates are not straightforward.

Aim: construct interpretable weights that can handle continuous and discrete covariates, with feasible posterior inference.
Solution: Since $w_j(x)$ represents the weight assigned to regression model $j$ at covariate $x$, an interpretable construction is obtained through Bayes Theorem by combining

- $P(A|j) = \int K(x|\psi_j)\,dx$, the probability that an observation from regression model $j$ has a covariate value in the region $A \subseteq \mathcal{X}$.
- $P(j) = w_j$, the prior probability that an observation comes from model $j$.

\[ w_j(x) = \frac{w_j K(x|\psi_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x|\psi_{j'})}. \]

Inclusion of continuous and discrete covariates is straightforward via appropriate definition of $K(x|\psi)$. 

Normalized weights (with S.G. Walker and I. Antoniano)
• Likelihood contains an intractable normalizing constant,

\[ f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|x, \theta_j), \]

where \( w_j(x) = \frac{w_j K(x|\psi_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x|\psi_{j'}).} \).
• Likelihood contains an intractable normalizing constant,

\[
f_P(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|x, \theta_j),
\]

where \( w_j(x) = \frac{w_j K(x|\psi_j)}{\sum_{j'=1}^{\infty} w_{j'} K(x|\psi_{j'})} \).

• However, using results for the geometric series, for \(|r| < 1\),

\[
\sum_{k=0}^{\infty} r^k = \frac{1}{1-r},
\]

the likelihood can be written as

\[
f_P(y|x) = \sum_{j=1}^{\infty} w_j K(x|\psi_j) K(y|x, \theta_j) \sum_{k=0}^{\infty} (1 - \sum_{j'=1}^{\infty} w_{j'} K(x|\psi_{j'}))^k.
\]
• The likelihood can be re-written as

\[ f_P(y|x) = \sum_{j=1}^{\infty} w_j K(x|\psi_j) K(y|x, \theta_j) \sum_{k=0}^{\infty} \left( \sum_{j'=1}^{\infty} w_{j'} (1 - K(x|\psi_{j'})) \right)^k. \]

\[ = \sum_{j=1}^{\infty} w_j K(x|\psi_j) K(y|x, \theta_j) \sum_{k=0}^{\infty} \prod_{l=1}^{k} \left( \sum_{j_i=1}^{\infty} w_{j_i} (1 - K(x|\psi_{j_i})) \right). \]

• Introducing latent variables \( k, d, D \), latent model is

\[ f_P(y, k, d, D|x) = w_d K(x|\psi_d) K(y|x, \theta_d) \prod_{l=1}^{k} w_{D_l} (1 - K(x|\psi_{D_l})). \]
• BNP mixture models for regression are highly flexible and numerous.

• This thesis aimed to shed light on advantages and drawbacks of the various models.
• BNP mixture models for regression are highly flexible and numerous.
• This thesis aimed to shed light on advantages and drawbacks of the various models.

1. Joint models:
   + easy computations,
   + flexible and easy to extend to other types of $Y$ and $X$,
   – posterior inference is based on the joint, problematic for large $p$.
   → Can overcome with EDP prior (maintains $+$ + ).
Discussion

2. Conditional models with covariate dependent atoms:
   + moderate computations (increases with flexibility in $\theta_j(x)$),
   + posterior inference is based on the conditional,
   - appropriate definition of $\theta_j(x)$ for mixed $x$ can be challenging,
   - choice of $\theta_j(x)$ has strong implications,
     · inflexible: extremely poor prediction,
     · over flexible: prediction may also suffer.

2. Conditional models with covariate dependent atoms:
   + moderate computations (increases with flexibility in $\theta_j(x)$),
   + posterior inference is based on the conditional,
   − appropriate definition of $\theta_j(x)$ for mixed $x$ can be challenging,
   − choice of $\theta_j(x)$ has strong implications,
     · inflexible: extremely poor prediction,
     · over flexible: prediction may also suffer.

3. Conditional models with covariate dependent weights:
   + posterior inference is based on the conditional,
   + flexible and incorporate covariate proximity in partition,
   − difficult computations,
   − lack of interpretable weights that can accommodate mixed $x$.
   $\rightarrow$ Can overcome with normalized weights.
2. Conditional models with covariate dependent atoms:
   + moderate computations (increases with flexibility in $\theta_j(x)$),
   + posterior inference is based on the conditional,
   − appropriate definition of $\theta_j(x)$ for mixed $x$ can be challenging,
   − choice of $\theta_j(x)$ has strong implications,
     · inflexible: extremely poor prediction,
     · over flexible: prediction may also suffer.

3. Conditional models with covariate dependent weights:
   + posterior inference is based on the conditional,
   + flexible and incorporate covariate proximity in partition,
   − difficult computations,
   − lack of interpretable weights that can accommodate mixed $x$.
     $\Rightarrow$ Can overcome with normalized weights.

- Huge dimension of partition space: posterior of $\rho_n$ is very spread out.
  $\Rightarrow$ Can overcome by enforcing notion of covariate proximity.
Future Work

- Examining models for large $p$ problems.
- Formalizing model comparison of various models,
  - Consistency,
  - Convergence rates,
  - Finite sample bounds.
- AD study:
  - Diagnosis based on sMRI and other imaging techniques.
  - Investigate dynamics of AD biomarkers (jointly + incorporating longitudinal data).


