Bayesian Nonparametric Conditional Density Estimation

Sara Wade

University of Cambridge

March 5, 2013
Outline

1 Introduction

2 Density Estimation

3 Conditional Density Estimation
   - Joint approach
   - Conditional Approach: Covariate Dependent Atoms
   - Conditional Approach: Covariate Dependent Weights

4 Discussion
Outline

1. Introduction
2. Density Estimation
3. Conditional Density Estimation
4. Discussion
Introduction

- **Aim**: Conditional density estimation, $\{f(y|x)\}_{x \in X}$. Many methods for flexible regression (e.g., splines, wavelets, Gaussian Processes), but typically assume iid Gaussian errors, $Y_i = m(x_i) + \epsilon_i | \sigma^2 \iid \sim N(0, \sigma^2)$. For iid data, mixture models are a useful tool for flexible density estimation. → Extend mixture models for conditional density estimation.
Aim: Conditional density estimation, \( \{ f(y|x) \}_{x \in X} \).

Many methods for flexible regression (ex. splines, wavelets, Gaussian Processes), but typically assume iid Gaussian errors,

\[
Y_i = m(x_i) + \epsilon_i \quad \epsilon_i | \sigma^2 \text{ iid } \sim N(0, \sigma^2).
\]
Aim: Conditional density estimation, \( \{ f(y|x) \}_{x \in X} \).

Many methods for flexible regression (ex. splines, wavelets, Gaussian Processes), but typically assume iid Gaussian errors,

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

For iid data, mixture models are a useful tool for flexible density estimation.
Aim: Conditional density estimation, \( \{ f(y|x) \} \in X \).

Many methods for flexible regression (e.g., splines, wavelets, Gaussian Processes), but typically assume iid Gaussian errors,

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

For iid data, mixture models are a useful tool for flexible density estimation.

→ Extend mixture models for conditional density estimation.
Outline

1 Introduction

2 Density Estimation

3 Conditional Density Estimation

4 Discussion
Density Estimation

- Assume $Y_i$ are iid from some unknown density $f$. 

Classical Methods (see Scott (1992))

- Histogram: let $B_j = [g_j, g_j+1)$ with $g_{j+1} - g_j = \lambda$ for $j = 0, \ldots, J$, then if $y \in B_j$, 
  $$\hat{f}(y) = \frac{1}{n\lambda} \sum_{i=1}^{n} \mathbf{1}(y_i \in B_j).$$
  Parameters: $g_0, \lambda$.

- Multivariate Extension: let $B_{j_1}, \ldots, j_k = \times_{h=1}^p [g_{j_h}, g_{j_h}+1)$ with $g_{j_h+1} - g_{j_h} = \lambda_h$ for $j_h = 0, \ldots, J$, if $y \in B_{j_1}, \ldots, j_k$, 
  $$\hat{f}(y) = \frac{1}{n\lambda_1 \cdots \lambda_p} \sum_{i=1}^{n} \mathbf{1}(y_i \in B_{j_1}, \ldots, j_k).$$
  Parameters: $g_0, \lambda$. 

Sara Wade (Cambridge)
Density Estimation

- Assume $Y_i$ are iid from some unknown density $f$.
- **Classical Methods (see Scott (1992))**
  - Histogram: let $B_j = [g_j, g_{j+1})$ with $g_{j+1} - g_j = \lambda$ for $j = 0, \ldots, J$, then if $y \in B_j$,
    \[
    \hat{f}(y) = \frac{1}{n\lambda} \sum_{i=1}^{n} 1(y_i \in B_j).
    \]
    Parameters: $g_0, \lambda$. 

Multivariate Extension: let $B_{j1}, \ldots, j_k = \times_{h=1}^{p} [g_{j_h}, g_{j_h+1})$ with $g_{j_h+1} - g_{j_h} = \lambda_h$ for $j = 0, \ldots, J$, if $y \in B_{j1}, \ldots, j_k$,
\[
\hat{f}(y) = \frac{1}{n^{1\ast\ldots\ast 1}} \sum_{i=1}^{n} 1(y_i \in B_{j1}).
\]
Parameters: $g_0, \lambda$. 

Sara Wade (Cambridge)  
BNP Conditional Density Estimation  
March 5, 2013  6 / 45
Density Estimation

- Assume $Y_i$ are iid from some unknown density $f$.
- **Classical Methods** (see Scott (1992))
  - **Histogram**:
    
    Let $B_j = [g_j, g_{j+1})$ with $g_{j+1} - g_j = \lambda$ for $j = 0, \ldots, J$, then if $y \in B_j$,
    \[
    \hat{f}(y) = \frac{1}{n\lambda} \sum_{i=1}^{n} 1(y_i \in B_j).
    \]

    Parameters: $g_0$, $\lambda$.

- **Multivariate Extension**: let $B_{j_1, \ldots, j_k} = \times_{h=1}^{p} [g_{j_h}, g_{j_h+1})$ with $g_{j_h} - g_{j_h+1} = \lambda_h$ for $j = 0, \ldots, J$, if $y \in B_{j_1, \ldots, j_k}$,
    \[
    \hat{f}(y) = \frac{1}{n\lambda_1 \cdots \lambda_p} \sum_{i=1}^{n} 1(y_i \in B_{j_1, \ldots, j_k}).
    \]

    Parameters: $g_0$, $\lambda$. 
Density Estimation

- **Classical Methods (see Scott (1992))**
  - Kernel estimators: let $K(z)$ be a non-negative, integrable function that is symmetric and integrates to 1.

\[
\hat{f}(y) = \frac{1}{n\lambda} \sum_{i=1}^{n} K \left( \frac{y - y_i}{\lambda} \right).
\]

Parameters: $\lambda$. 

Multivariate Extension:

\[
\hat{f}(y) = \frac{1}{n\lambda_1 \cdots \lambda_p} \sum_{i=1}^{n} \prod_{h=1}^{p} K \left( \frac{y_h - y_{ih}}{\lambda_h} \right).
\]

Parameters: $\lambda$. 

Sara Wade (Cambridge)
Density Estimation

- Classical Methods (see Scott (1992))
  - Kernel estimators: let $K(z)$ be a non-negative, integrable function that is symmetric and integrates to 1.

$$\hat{f}(y) = \frac{1}{n\lambda} \sum_{i=1}^{n} K \left( \frac{y - y_i}{\lambda} \right).$$

Parameters: $\lambda$.

Multivariate Extension:

$$\hat{f}(y) = \frac{1}{n\lambda_1 \ldots \lambda_p} \sum_{i=1}^{n} \prod_{h=1}^{p} K \left( \frac{y - y_{i,h}}{\lambda_h} \right).$$

Parameters: $\lambda$. 
Mixture Models

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

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

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

Note: if $P = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$ and $K(y|\theta) = \frac{1}{\lambda} K(y - \theta \lambda) \rightarrow$ kernel estimate.

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

\[ P = \sum_{j=1}^{J} w_j \delta_{\theta_j}, \]

\[ f_P(y) = \sum_{j=1}^{J} w_j K(y|\theta_j). \]

Interpretation of components:

- Heterogeneous population $\rightarrow$ physical interpretation of components.
- Homogeneous population $\rightarrow$ "kernel" method.
Mixture Models

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

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

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

- Note: if $P = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$ and $K(y|\theta) = \frac{1}{\lambda} K \left( \frac{y-\theta}{\lambda} \right) \rightarrow$ kernel estimate.
Mixture Models

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

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

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

- Note: if $P = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$ and $K(y|\theta) = \frac{1}{\lambda} K \left( \frac{y-\theta}{\lambda} \right) \rightarrow$ kernel estimate.

- Bayesian setting, define a prior for $P$, 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 $P$, assume $Y_i$ are iid with density

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

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

- Note: if $P = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$ and $K(y|\theta) = \frac{1}{\lambda} K \left( \frac{y-\theta}{\lambda} \right) \rightarrow$ kernel estimate.

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

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

- Interpretation of components:
  - Heterogeneous population $\rightarrow$ physical interpretation of components.
  - Homogeneous population $\rightarrow$ ”kernel” method.
Mixture Models

- How many components?
  - Fixed $J < \infty$: 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.
BNP Mixture Model: Asymptotic Properties

- **Notation:**
  - $\mathcal{F}$: sets of densities wrt Lebesgue measure.
  - $U_\epsilon(f_0)$: a nbhd of $f_0 \in \mathcal{F}$ of size $\epsilon > 0$
  - $Q_f$ prior of $f_P$ under BNP mixture model, where $P \sim \Pi$ (ex. $\Pi = DP(\alpha, P_0)$).

- **Posterior consistency:** $Q_f$ is consistent at $f_0$ if
  \[
  Q_f(U_\epsilon(f_0)|Y_{1:n}) \rightarrow 1 \quad \text{a.s. } P_{f_0}^\infty,
  \]
  for any $\epsilon > 0$.

- **What conditions on $f_0$, $K(\cdot|\theta)$, and $\Pi$ imply posterior consistency?**
BNP Mixture Model: Asymptotic Properties

- **Neighborhoods:**
  - **Weak:**
    \[
    U_\epsilon(f_0) = \left\{ f \in \mathcal{F} : \left| \int g_i(y) f(y) dy - \int g_i(y) f_0(y) dy \right| < \epsilon, \right\},
    \]
    where \(g_i(\cdot), i = 1, \ldots, I\) are bounded, continuous functions on \(\mathcal{Y}\).
  - **\(L_1\) (Strong):**
    \[
    U_\epsilon(f_0) = \left\{ f \in \mathcal{F} : \int \left| f(y) - f_0(y) \right| dy < \epsilon \right\}.
    \]

- **Convergence rates:** \(Q_f(\cdot|Y_{1:n})\) converges at (at least) rate \(\epsilon_n \downarrow 0\) if for any \(M_n \to \infty\) s.t \(M_n \epsilon_n \to 0\),
  \[
  Q_f(U_{M_n\epsilon_n}(f_0)|Y_{1:n}) \to 1 \quad \text{a.s } P_{f_0}^\infty.
  \]
BNP Mixture Model: Asymptotic Properties

- Ghosal, Ghosh, and Rammamoorthi (1999)- weak and strong consistency of univariate normal location DP mixtures, where a) $f_0$ is mixture of normals where mixing measure is compactly supported, b) $f_0$ is compactly supported.

- Ghosal and van der Vaart (2001)- convergence rates of univariate normal location and location-scale DP mixtures, where $f_0$ is mixture of normals where mixing measure is compactly supported.

- Tokdar (2006)- weak and strong consistency of univariate normal location-scale DP mixtures for general class of continuous $f_0$.

- Ghosal and van der Vaart (2007)- convergence rates of univariate normal location DP mixtures for general class of continuous $f_0$ (slower rates).
BNP Mixture Model: Asymptotic Properties

- Wu and Ghosal (2008) - weak consistency of multivariate location-scale mixtures for general class of continuous $f_0$.
- Wu and Ghosal (2010) - strong consistency of multivariate normal location mixtures for general class of continuous $f_0$.
- Tokdar (2010) - convergence rates for multivariate normal location mixtures for $f_0$ a) infinitely divisible or b) compactly supported.
Outline

1 Introduction

2 Density Estimation

3 Conditional Density Estimation
   - Joint approach
   - Conditional Approach: Covariate Dependent Atoms
   - Conditional Approach: Covariate Dependent Weights

4 Discussion
Extending Mixture Models

Two main lines of research:

- **Joint approach:** Given $P$, assume $(Y_i, X_i)$ 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.

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

  $$f_{P_x}(y|x) = \int K(y|\theta)dP_x(\theta) = \sum_{j=1}^{\infty} w_j(x) K(y|\theta_j(x)).$$
Extending Mixture Models

Two main lines of research:

- **Joint approach:** Given $P$, assume $(Y_i, X_i)$ 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.

- **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 $\{P_x\}$ and $x_i$, the $Y_i$ are independent with density

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

Large literature on finite mixtures for conditional density estimation (ex. mixture of experts, smooth mixture of regressions)
Joint approach

- First studied by 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
  - the inverse Wishart is poorly parametrized.
Joint Approach

- Extension by Shahbaba and Neal (2009):

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

\[P \sim DP(\alpha P_{0\theta} \times P_{0\psi}),\]

where \(x = (1, x')\); \(\theta = (\beta, \sigma^2_{y|x})\); \(\psi = (\mu, \sigma^2_{x})\); 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 \(X\).
Joint Approach

- Extension by Shahbaba and Neal (2009):

\[
    f_P(y, x) = \int \prod_{h=1}^{p} N(x_h|\mu_h, \sigma_{x,h}^2) \, dP(\theta, \psi),
\]

\[
    P \sim DP(\alpha P_0 \theta \times P_0 \psi),
\]

where \( x = (1, x') \); \( \theta = (\beta, \sigma_{y|x}^2) \); \( \psi = (\mu, \sigma_x^2) \); 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 \( X \).

The DP-mixture model is equivalently expressed as

\[
(Y_i, X_i) \mid \theta_i, \psi_i \overset{ind}{\sim} K(y \mid \theta_i, x)K(x \mid \psi_i)
\]

\[
(\theta_i, \psi_i) \mid P \overset{iid}{\sim} P
\]

\[
P \sim \text{DP}(\alpha P_0)
\]
DP mixture model

- The DP-mixture model is equivalently expressed as:

\[
(Y_i, X_i) \mid \theta_i, \psi_i \overset{\text{iid}}{\sim}\ K(y \mid \theta_i, x)K(x \mid \psi_i)
\]

\[
(\theta_i, \psi_i) \mid P \overset{\text{iid}}{\sim}\ P
\]

\[
P \sim\ DP(\alpha P_0)
\]

- It is often useful to marginalize over \(P\).
- The marginal of \(((\theta_1, \psi_1), \ldots, (\theta_n, \psi_n))\) is characterized by the Blackwell and MacQueen urn scheme:

\[
(\theta_1, \psi_1) \sim P_0,
\]

\[
(\theta_{n+1}, \psi_{n+1}) \mid (\theta_1, \psi_1), \ldots, (\theta_n, \psi_n) \sim \frac{\alpha}{\alpha + n} P_0 + \sum_{j=1}^{k} \frac{n_j}{\alpha + n} \delta(\theta^*_j, \psi^*_j),
\]

where \(k\) is the number of species observed in the sample \(((\theta_1, \psi_1), \ldots, (\theta_n, \psi_n))\) and \((\theta^*_j, \psi^*_j)\) is the \(j^{th}\) species of size \(n_j\).
DP mixture model

- The parameters \(((\theta_1, \psi_1), \ldots, (\theta_n, \psi_n))\) can be reparametrized as a random partition ("clustering") \(\rho_n = (s_1, \ldots, s_n)\) and the unique values \((\theta^*_j, \psi^*_j)\).

- From the B+M urn scheme, we have (Antoniak (1974))

\[
p(\rho_n) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^k \prod_{j=1}^{k} \Gamma(n_j),
\]

and posterior is

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

where \(g_y(y_j^*|x_j^*) = \int \prod_{i:s_i=j} N(y_i|x_i|\beta, \sigma^2) dP_{0\theta}(\theta)\).
Under the quadratic loss, the conditional estimate at $y$ given $X_{n+1} = x$ and the data is

$$
\hat{f}(y|x) = \mathbb{E}[f(y|x, \rho_n, \theta_{1:k}^*, \psi_{1:k}^*)|y_{1:n}, x_{1:n}, x] \\
= \frac{\mathbb{E}[f(y, x|\rho_n, \theta_{1:k}^*, \psi_{1:k}^*)|y_{1:n}, x_{1:n}]}{\mathbb{E}[f(x|\rho_n, \psi_{1:k}^*)|x_{1:n}]} = \frac{\hat{f}(y, x)}{\hat{f}(x)} \\
= \mathbb{E}[\tilde{w}_{k+1}(x_{n+1})g_y(y|x) + \sum_{j=1}^{k} \tilde{w}_j(x_{n+1})N(y|x_{n+1} \beta_j^*, \sigma_j^2)|y_{1:n}, x_{1:n}],
$$

where

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

$$
\tilde{w}_{k+1}(x_{n+1}) = c^{-1} \alpha g_x(x_{n+1}),
$$

and $c = (\alpha + n) * \hat{f}(x)$. 


Asymptotic Properties

- First, establish posterior consistency results for the joint density.
- Then, study implications on the conditional density and regression function.

Notice that the kernels studied here are more general than previous multivariate kernels studied in BNP mixture model literature on posterior asymptotics. Results do not directly apply. Hannah et al. (2011) show weak consistency of the joint for \( f_0 \) compactly supported. With some additional mild conditions, they show that the estimated regression function converges pointwise to the true regression function a.s. \( P_\infty f_0 \).

Stronger properties and implications for the conditional density are an open problem.
Asymptotic Properties

- First, establish posterior consistency results for the joint density.
- Then, study implications on the conditional density and regression function.
- Notice that the kernels studied here are more general than previous multivariate kernels studied in BNP mixture model literature on posterior asymptotics → results do not directly apply.
- Hannah et al. (2011) show weak consistency of the joint for \( f_0 \) compactly supported.
- With some additional mild conditions, they show that the estimated regression function converges pointwise to the true regression function a.s. \( P_{f_0}^\infty \).
- Stronger properties and implications for the conditional density are an open problem.
Joint Approach

Advantages:

- Computations are straightforward - can use well known sampling algorithms for BNP mixtures (marginalization, truncation, slice sampling, retrospective sampling).
- Easy to extend to other types of $Y$ and $X$. 

Disadvantages:

- Posterior inference is based on joint when interest is in conditional; this is problematic because posterior on $\rho_n$ will place too much emphasis on fitting the marginal of $x$, especially for large $p$.
- This can result in degraded predictive performance and slow down computations.

First issue address by W. et al. (2013), where DP prior is replaced with EDP prior, which allows for a nested clustering structure.
Joint Approach

Advantages:
- Computations are straightforward - can use well known sampling algorithms for BNP mixtures (marginalization, truncation, slice sampling, retrospective sampling).
- Easy to extend to other types of $Y$ and $X$.

Disadvantages:
- Posterior inference is based on joint when interest is in conditional; this is problematic because posterior on $\rho_n$ will place too much emphasis on fitting the marginal of $x$, especially for large $p$.
  → can result in degraded predictive performance and slow down computations.

First issue address by W. et al. (2013), where DP prior is replaced with EDP prior, which allows for a nested clustering structure.
Joint Approach

- **Advantages:**
  - Computations are straightforward - can use well known sampling algorithms for BNP mixtures (marginalization, truncation, slice sampling, retrospective sampling).
  - Easy to extend to other types of $Y$ and $X$.

- **Disadvantages:**
  - Posterior inference is based on joint when interest is in conditional; this is problematic because posterior on $\rho_n$ will place too much emphasis on fitting the marginal of $x$, especially for large $p$.
    → can result in degraded predictive performance and slow down computations.
  - First issue address by W. et al. (2013), where DP prior is replaced with EDP prior, which allows for a nested clustering structure.
Conditional Approach

- Conditional density is modelled directly as

\[ f_{P_x}(y|x) = \int K(y; x, \theta) dP_x(\theta). \]

- A prior is defined for the random discrete p.m. \( \{P_x\} \) so that they are dependent across \( x \). In general,

\[ P_x = \sum_{j=1}^{\infty} w_j(x) \delta_{\theta_j}(x). \]

This general construction was first proposed by MacEachern (1999) (but some early approaches are closely related).

- Dependent Dirichlet Process (DDP): marginally

\[ w_j(x) = v_j(x) \prod_{j' < j} (1 - v_{j'}(x)) \text{ and } v_j(x) \overset{iid}{\sim} \text{Beta}(1, \alpha(x)). \]
Conditional Approach

Most proposals in literature fall into one of two important subclasses:

1) models with covariate dependent atoms but simple weights,
2) models with covariate dependent weights and simple atoms.

Why? Computations and interpretations are easier and desirable theoretical properties are still available.
Covariate Dependent Atoms

Model:

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

where \( P_x = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}(x) \).
Covariate Dependent Atoms

- Model:

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

where \( P_x = \sum_{j=1}^{\infty} w_j \delta_{\theta_j(x)}. \)

- Single-p DDP: \( w_j = v_j \prod_{j' < j} (1 - v_{j'}) \) and \( v_j \overset{iid}{\sim} \text{Beta}(1, \alpha). \)

- Very popular: regression (MacEachern (2000)), ANOVA (Delorio et al. (2004)), spatial (Gelfand et al. (2005)), time series (Rodriguez and Horst (2008)), discriminant analysis (De la Cruz et al. (2007)), longitudinal analysis (Müller et al. (2005)), and survival analysis (Jara et al. (2010)).

- Inference can be carried out using established algorithms for DP mixtures.
Covariate Dependent Atoms

- Example:

\[
f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j \mathcal{N}(y|m_j(x), \sigma_j^2),
\]

where \(m_j(x)\) may be modelled via
  - basis function expansion,
  - Gaussian processes.
Covariate Dependent Atoms

- Example:

\[ f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j N(y|m_j(x), \sigma_j^2), \]

where \( m_j(x) \) may be modelled via
  - basis function expansion,
  - Gaussian processes.

- \( \{P_x\} \) can be marginalized, and parameters are \( \rho_n \) and \( (\theta_j^*(x)) \) with posterior

\[
p(\rho_n, \theta_{1:k}^*|y_{1:n}, x_{1:n}) \propto \alpha^k \prod_{j=1}^k \Gamma(n_j)p_0(\theta_j^*) \prod_{i:s_i=j} N(y_i|m_j^*(x_i), \sigma_j^{2*}).
\]
Under the quadratic loss, the conditional estimate at $y$ given $x$ and the data is

$$
\hat{f}(y|x) = \mathbb{E}[f(y|x, \rho_n, \theta_{1:k}^{*})|y_{1:n}, x_{1:n}, x]
$$

$$
= \frac{\alpha}{\alpha + n} \mathbb{E} [N(y|m(x), \sigma^2)] + \mathbb{E} \left[ \sum_{j=1}^{k} \frac{n_j}{\alpha + n} N(y|m(x), \sigma^2)|y^*_j, x^*_j \right],
$$
For $i, = 1, \ldots, n$, let

$$Y_i = x_i^2 + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 1)$$

and

$$X_i \sim \text{U}(-5, 5).$$

Consider the model

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j \text{N}(y|m_j(x), \sigma_j^2),$$

where $m_j(x) = x \beta_j$. 

Sara Wade (Cambridge) BNP Conditional Density Estimation March 5, 2013 28 / 45
A Cautionary Tale

Three partitions with the highest estimated posterior probability.
A Cautionary Tale

Estimated regression function (in red) with the true function (in black).
Doesn’t even interpolate the data!
Asymptotic Properties

- Most posterior consistency results for the conditional density rely on theorems for the joint density.
- Assume $Y_i | x_i$ are independently generated from $f_0 x_i$ and $X_i$ are generated from $h(\cdot)$ with joint data generating density $f_0(y, x) = f_0(x(y)h(x)$.
- Notation: $f_x = (f_x)_{x \in X}$ and $\mathcal{F}_x$ set of conditional densities.
- Posterior consistency: $Q_{f_x}$ is consistent at $f_0x$ if

$$Q_{f_x}(U_\epsilon(f_0x)|Y_{1:n}, X_{1:n}) \to 1 \quad \text{a.s } P_{f_0}^\infty,$$

for any $\epsilon > 0$.
- What conditions on $f_0$, $K(\cdot | \theta(x))$, and law of $\{P_x\}$ imply posterior consistency?
Asymptotic Properties

**Neighborhoods for conditional densities:**

- **Weak:**
  \[
  U_\epsilon(f_0 X) = \{ f_X \in \mathcal{F}^X : \int g_i(y, x) f(y|x) h(x) dy dx - \int g_i(y, x) f_0(y|x) h(x) dy dx < \epsilon \},
  \]
  
  \[g_i(\cdot), i = 1, \ldots, I\text{ are bounded, continuous functions on } \mathcal{Y} \times \mathcal{X}.\]

- **$L_1$ (Strong):**
  \[
  U_\epsilon(f_0 X) = \{ f_X \in \mathcal{F}^X : \int \left( \int | f(y|x) - f_0(y|x) | dy \right) h(x) dx < \epsilon \}.
  \]
  
  or as
  \[
  U_\epsilon(f_0 X) = \{ f_X \in \mathcal{F}^X : \sup_{x \in \mathcal{X}} \int | f(y|x) - f_0(y|x) | dy < \epsilon \}.
  \]
Asymptotic Properties

- Pati et al. (2013) show weak consistency for a general class of bounded $f_0$ with certain tail conditions. For $m_j(x)$, they require continuity and approximation properties, and for $w_j$, no constraints.

- Pati et al. (2013) show strong consistency with more stringent conditions for $m_j(x)$ and $w_j$. In particular,

$$m_j(x) = X\beta_j + \eta_j(x),$$

$$\eta_j(x)|\tau \sim \text{GP}(0, c),$$

$$c(x_1, x_2) = c \exp(-\tau ||x_1 - x_2||^2),$$

$$\tau^{p(1+\eta_2)/\eta_2} \sim \text{Gamma}(a, b),$$

where $c, \eta_2, a, b$ are fixed positive constants, and with further conditions on priors of $\beta_j$ and $\sigma_j^2$. The weights must decay rapidly enough, and the usual DP weights do not satisfy this.
Conditional Approach: Covariate Dependent Atoms

- **Advantages:**
  - Posterior inference is based on the conditional.
  - Straightforward computations - can use well known sampling algorithms for BNP mixtures - but cost increases with flexibility in $\theta_j(x)$.

- **Disadvantages:**
  - Choice of $\theta_j(x)$ has strong implications. If too inflexible, may result in (extremely) poor estimates, and if overly flexible, will overfit.
  - Need to carefully monitor the posterior of $\rho_n$ to determine if it depends on $x$ - may be hard in higher dimensions.
Conditional Approach: Covariate Dependent Atoms

Advantages:
- Posterior inference is based on the conditional.
- Straightforward computations - can use well known sampling algorithms for BNP mixtures- but cost increases with flexibility in $\theta_j(x)$.

Disadvantages:
- Choice of $\theta_j(x)$ has strong implications. If too inflexible, may result in (extremely) poor estimates, and if overly flexible, will overfit.
  Need to carefully monitor the posterior of $\rho_n$ to determine if it depends on $x$ - may be hard in higher dimensions.
Covariate Dependent Weights

- Model:

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

where \( P_x = \sum_{j=1}^{\infty} w_j(x) \delta_{\theta_j}. \)
Covariate Dependent Weights

Model:

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

where \( P_x = \sum_{j=1}^{\infty} w_j(x) \delta_{\theta_j} \).

Need to define \( w_j(x) \) s.t. \( \sum_j w_j(x) = 1 \) a.s.

Main technique is through a stick-breaking construction

\[ w_1(x) = v_1(x), \]

\[ w_j(x) = v_j(x) \prod_{j'<j} (1 - v_{j'}(x)) \quad \text{for } j > 1. \]
Proposals

- **Order-based DDP (Griffin and Steele (2006)):**
  - Assume $v_j \sim \text{Beta}(1, \alpha)$ and associate every $(v_j, \theta_j)$ with $\psi_j$ taking values in $X$.
  - For every $x$, reorder $\psi_j$ based on their distance to $x$. This defines a permutation, $\pi_x$, of $(v_j, \theta_j)$.
  - Define $v_j(x) = v_{\pi_x}(j)$.

- **Kernel Stick-Breaking (Dunson and Park (2008)):**
  - Define $v_j(x) = v_j K(x|\psi_j)$, for some positive kernel with parameter $\psi_j$ that integrates 1.
  - Examples of kernels (see Reich and Fuentes (2007), Griffin and Steele (2010), Chung and Dunson (2011)):
    
    $$K(x|\psi_j) = \exp(-\tau_j \|x - \mu_j\|^2).$$

    $$K(x|\psi_j) = \prod_{h=1}^{p} 1(|x_h - \mu_{j,h}| < \tau_j^{-1}).$$
Proposals

- Stick-breaking through sigmoidal transformations:

\[ v_j(x) = l(\psi_j(x)), \]

where \( l : \mathbb{R} \rightarrow [0, 1] \) is a monotone, differentiable link function and \( \psi_j(x) \) is a random, real-valued function on \( \mathcal{X} \).

- \( l(\cdot) \) is commonly chosen to be the probit or logit link function.

- \( \psi_j(x) \) may be defined as a simple linear function, as a linear combination of basis functions, or assigned Gaussian process prior.

Proposals

- Stick-breaking through sigmoidal transformations:
  
  \[ v_j(x) = l(\psi_j(x)), \]

  where \( l : \mathbb{R} \rightarrow [0, 1] \) is a monotone, differentiable link function and \( \psi_j(x) \) is a random, real-valued function on \( \mathcal{X} \).

- \( l(\cdot) \) is commonly chosen to be the probit or logit link function.

- \( \psi_j(x) \) may be defined as a simple linear function, as a linear combination of basis functions, or assigned Gaussian process prior.


- Difficulties with stick-breaking construction: hard to interpret weights, choose hyperparameters, and include discrete and continuous covariates flexibly.
Proposals

- Normalized weights (Antoniano, W., Walker (2013)):

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

for some positive kernel with parameter \( \psi_j \) that integrates 1.

- Intrepretable structure for weights that are implied by the joint model, however we are modelling the conditional directly!

→ selection of hyperparameters is easier.

- Inclusion of discrete and continuous \( x \) is simple.
Prior simulations

(a) KSB

(b) KSB

(c) KSB

(d) NW

Sq. Exp. Kernel SB with a common bandwidth

Sq. Exp. Kernel SB

Normalized Weights
Note that models with covariate dependent weights implicitly define a covariate dependent random partition, but a closed form for the distribution is typically unavailable.

Under the quadratic loss, the conditional estimate at $y$ given $x$ and the data is

$$\hat{f}(y|x) = \mathbb{E}[f(y|x, P_x)|y_{1:n}, x_{1:n}],$$

which can be computed based on posterior inference on $\{P_x\}$. 
Norets and Pelenis (2013) show weak and strong consistency for a general class of $f_0$ under the kernel stick-breaking method with

$$K(-\tau_j \|x - \mu_j\|^2),$$

and location-scale kernel for $y$.

Strong consistency requires additional constraints on priors of $\nu_j, \theta_j, \psi_j$; in particular, a large prior mass on values of $\nu_j$ close to 1 are required.
Pati et al. (2013) show weak and strong consistency for a general class of $f_0$ under the probit stick-breaking method with

$$v_j(x) = \Phi(\psi_j(x)), \quad \psi_j(x) \overset{iid}{\sim} \text{GP}(0, c).$$

They require continuity and approximation properties of $\psi_j(x)$ for weak consistency, and further conditions for strong consistency; in particular

$$\psi_j(x) | \tau_j \overset{iid}{\sim} \text{GP}(0, c),$$

$$c(x_1, x_2) = c \exp(-\tau_j ||x_1 - x_2||^2),$$

where $\tau_j$, are required to decay to zero at a fast enough rate.
Conditional Approach: Covariate Dependent Weights

- **Advantages:**
  - Posterior inference is based on the conditional.
  - Flexible and implicitly incorporate notion of covariate proximity in partition.

- **Disadvantages:**
  - Posterior inference can be quite complicated.
  - Lack of interpretable weights that can accommodate mixed...
Conditional Approach: Covariate Dependent Weights

- **Advantages:**
  - Posterior inference is based on the conditional.
  - Flexible and implicitly incorporate notion of covariate proximity in partition.

- **Disadvantages:**
  - Posterior inference can be quite complicated.
  - Lack of interpretable weights that can accommodate mixed $x$.
    → can overcome with normalized weights.
Outline

1. Introduction
2. Density Estimation
3. Conditional Density Estimation
4. Discussion
In summary, we have discussed the advantages and drawbacks of three types of BNP mixture models for conditional density estimation:

- **Joint models**—flexible and computationally straightforward, but also requires modelling of $x$.
- **Covariate dependent atoms**—directly model conditional and computationally straightforward, but $\theta_j(x)$ need to be carefully specified and $\rho_n$ needs to be monitored for dependency on $x$ in the posterior.
- **Covariate dependent weights**—flexible and directly model conditional, but computationally expensive.

Some recent work on theoretical properties, but this needs to be further explored.