We propose two modeling approaches to Bayesian quantile regression based on Dirichlet process (DP) mixture models and Gaussian processes (GP). The first approach uses a semiparametric additive formulation, with the regression function modeled parametrically and with nonparametric priors for the error distribution. The prior models include dependent DP-s yielding error distributions that can change nonparametrically with the covariates. The second class of models uses Gaussian process for the regression function and again DP priors for the error distribution. Inference is implemented using posterior simulation methods for DP mixtures.

This is joint work with Athanasios Kottas, Department of Applied Math and Statistics, University of California at Santa Cruz, e-mail: thanos@ams.ucsc.edu.
Quantile Regression: Motivation

- Covariates may have effect on the whole shape of the response distribution and not only on its location. **Quantile regression** (QR) quantifies relationship between covariates and a set of quantiles of the response distribution and enables a better understanding of the effects of covariates.

- Additive regression formulation for the $p$-th quantile, $h(\cdot)$, of the response distribution

\[ y_i = h(x_i) + \epsilon_i, \]

where, $h(x_i) = x_i^T \beta$, for example, and $\epsilon_i$ are independent samples from a distribution having zero $p$-th quantile: $\int_{-\infty}^{0} f_p(\epsilon) d\epsilon = p$.

- **Optimization-based estimation:** In analogy to minimizing sum of squared errors in mean regression, in QR we minimize ”check function” $\rho_p(\epsilon) = \epsilon - \epsilon I(\epsilon < 0)$. To obtain the median of the response, the problem reduces to $\min \sum_{i=1}^{n} |y_i - x_i^T \beta|$. Point estimates only, no likelihood, inference based on asymptotics.

- **Parametric modeling:** explicitly specifies error distribution, enables more complete inference. For example, asymmetric Laplace (AL) distribution, (model $\mathcal{M}_0$):

\[ \epsilon_i \sim iid k_p^{AL}(\epsilon; \mu, \sigma) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{\epsilon - \mu}{\sigma}\right)\right\} \text{ with } \int_{-\infty}^{0} k_p^{AL}(x; 0, \sigma) dx = p. \]

- **Limitation:** one parameter determines both quantile and skewness ($p > 0.5$ left skewed, $p = 0.5$ symmetric), e.g. in median regression, error distribution must be symmetric.
Bayesian Nonparametric Modeling of Error Distribution

• We develop **flexible prior models** for the error distribution based on Dirichlet process (DP) mixtures.

• **Model** $\mathcal{M}_1$, a general scale mixture of asymmetric Laplace densities captures more flexible tail behavior:

$$f_P^1(\epsilon; G) = \int k_P^{AL}(\epsilon; \sigma)dG(\sigma), \quad G \sim DP(\alpha G_0).$$

(3)

The mixing preserves quantiles since $\int_{-\infty}^{0} f_P^1(\epsilon; G)d\epsilon = p$.

• $\mathcal{M}_1$ extends $\mathcal{M}_0$ with regard to tail behavior but the skewness of the mixture $f_P^1(\cdot; G)$ still suffers the same limitation as the kernel $k_P^{AL}(\cdot; \sigma)$.

• Motivated by limitations of model $\mathcal{M}_1$ we propose two more mixture models each of which can capture the shape (skewness, tail behavior) of *any* unimodal error density $f_P(\cdot)$ on $R^1$ with $p$-th quantile zero.
A key result is a representation theorem for non-increasing densities on $R^+$: For any non-increasing density $f(\cdot)$ on $R^+$ there exists a distribution function $G$, with support on $R^+$, such that $f(t; G) = \int \theta^{-1} 1_{[0, \theta)}(t)dG(\theta)$.

This result leads to a mixture representation for any unimodal density on the real line with $p$-th quantile (and mode) equal to zero,

$$\int \int k_p(\epsilon; \sigma_1, \sigma_2)dG_1(\sigma_1)dG_2(\sigma_2),$$

with $G_1$ and $G_2$ supported by $R^+$, and

$$k_p(\epsilon; \sigma_1, \sigma_2) = \frac{p}{\sigma_1} 1_{(-\sigma_1, 0)}(\epsilon) + \frac{(1-p)}{\sigma_2} 1_{[0, \sigma_2)}(\epsilon),$$

with $0 < p < 1, \sigma_r > 0, r = 1, 2$.

**Model $M_2$** is obtained assuming independent DP priors for $G_1$ and $G_2$:

$$f_p^2(\epsilon; G_1, G_2) = \int \int k_p(\epsilon; \sigma_1, \sigma_2)dG_1(\sigma_1)dG_2(\sigma_2), \quad G_r \sim \text{DP}(\alpha_r G_{r0}), r = 1, 2$$

for the error density in (1). Model $M_2$ is sufficiently flexible to capture general forms of skewness and tail behavior, as suggested by the data.
Nonparametric scale mixtures of uniform densities

- **Model $M_2$** in hierarchical form:

$$
\begin{align*}
  y_i \mid \beta, \sigma_{1i}, \sigma_{2i} & \sim \text{ind} \quad k_p(y_i - \mathbf{x}_i^T \beta; \sigma_{1i}, \sigma_{2i}), i = 1, ..., n \\
  \sigma_{ri} \mid G_r & \sim \text{iid} \quad G_r, \ r = 1, 2, \ i = 1, ..., n \\
  G_r \mid \alpha_r, d_r & \sim \quad \text{DP}(\alpha_r G_{r0}), \ r = 1, 2.
\end{align*}
$$

(4)

- An alternative BNP family of error densities is based on a single $G$, with prior $\text{DP}(\alpha G_0^*)$, where $G_0^*$ is bivariate on $R^+ \times R^+$. We use a bivariate lognormal with hyperpriors for location, scale and correlation parameters.

- **Model $M_3$** with a bivariate DP:

$$
 f_p^3(\epsilon; G) = \int \int k_p(\epsilon; \sigma_1, \sigma_2)dG(\sigma_1, \sigma_2), \quad G \sim \text{DP}(\alpha G_0^*).
$$

(5)

- It is straightforward to extend models $M_2$ and $M_3$ to handle censored data.

- Inference is based on variants of established MCMC sampling methods typically used in DP mixture models.
Model $\mathcal{M}_2$: six simulated data sets

Figure 1: Simulated data sets for model $\mathcal{M}_2$; prior and posterior predictive densities plotted with dotted and dashed lines respectively; the contour of data histogram and the true density, both with solid line.
Model $M_1$ vs. model $M_2$

Figure 2: Simulated data set, right skewed and with $p = 0.6$; Posterior predictive densities for models $M_1$ and $M_2$ plotted with dashed and dotted lines respectively; the contour of the data histogram, solid line.
Censored data set

Figure 3: The censored data set (Ying, Jung and Wei, 1995), consists of survival times in days for 121 patients with small cell lung cancer; 23 survival times are right censored. Each patient was randomly assigned to one of two treatments, A and B, achieving 62 and 59 patients respectively. The top panel displays posterior predictive densities for treatment A under each model ($M_0$, solid line, $M_2$, dashed line, parametric Weibull model, dotted line). Bottom panels: CPO plot of fully observed data (left), CPO plot of censored data points (right), with 'o' representing points from $M_0$, '2' the model $M_2$, and 'w' the Weibull model. Displayed are 98 fully observed and 23 censored points for each model.
Error distribution that changes nonparametrically with covariates

• **Motivation**: Under the previous setting, distribution of $\epsilon_i$ is the same for all $x_i$, and distribution of $y_i$ changes with $x_i$ only through the $p$-th quantile $x_i'\beta$. We want to **model nonparametrically** error distribution that changes with covariates.

• Need a prior model for $f_{p,x}(\cdot) = \{f_{p,x}(\cdot) : x \in \mathcal{X}\}$, where $\mathcal{X}$ is the covariate space and $\forall x, \int_{-\infty}^{x} f_{p,x}(\epsilon) d\epsilon = p$.

• $f_{p,x}(\cdot)$ changes with $x$ such that $f_{p,x}(\cdot)$ and $f_{p,x'}(\cdot)$ are **similar** for $x$ close to $x'$ and $f_{p,x}(\cdot)$ are dependent. This allows error distribution to have different shapes for different parts of the covariate space.

• Write $k_p(\epsilon; \theta_1, \theta_2) = \frac{p}{\exp(\theta_1)} 1(-\exp(\theta_1),0)(\epsilon) + \frac{(1-p)}{\exp(\theta_2)} 1(0,\exp(\theta_2))(\epsilon)$ implying that model $\mathcal{M}_2$ may have normal base distribution, $G_{r0}(\cdot) = N(\cdot; \mu_r, \tau^2)$, $r = 1, 2$. 
Error distribution that changes nonparametrically with covariates

- To allow \( f_p(\epsilon | G_1, G_2) \) of model \( M_2 \) to change with \( x \) the mixing distributions \( G_1, G_2 \) need to change with \( x \).
- We employ dependent Dirichlet processes (DDP) of MacEachern to model \( G_{1,x}, G_{2,x} \), where \( X \) is the covariate space.
- In Sethuraman’s definition of Dirichlet process, where \( G = \sum_{l=1}^{\infty} w_l \delta_{\Psi_l} \) and \( \Psi_l \) are iid from the base distribution \( G_0 \), replace \( \Psi_l \) with \( \Psi_{l,x} \) to obtain \( G_X = \sum_{l=1}^{\infty} w_l \delta_{\Psi_{l,x}} \)
- \( \Psi_{l,x} \) are now iid realizations from the base Gaussian stochastic process \( G_{0,x} \) and we write \( G_X \sim \text{DDP}(\alpha G_{0,x}) \).
- For a finite, fixed covariate vector, \( x = (x_1, ..., x_M) \), we have \( G_x \sim \text{DP}(\alpha G_0(x)) \) since \( G_x = \sum_{l=1}^{\infty} w_l \delta_{\Psi_l(x)} \) and \( \Psi_l(x) \) are iid from M-variate normal \( G_0(x) \).
• DDP version of model $\mathcal{M}_2$ in hierarchical form:

\begin{align}
Y_i \mid (\beta_0, \beta_1), \theta_{1i}, \theta_{2i} & \overset{\text{ind}}{\sim} f(y_i; (\beta_0, \beta_1), \theta_{1i}, \theta_{2i}) \\
\theta_{ri} \mid G_{r,x} & \overset{iid}{\sim} G_{r,x} \\
G_{r,x} \mid \alpha_r, \phi, \tau^2_r & \overset{\text{ind}}{\sim} DP(\alpha_r G_{r0}(x; \phi))
\end{align}

(6)

where $\theta_{ri} = (\theta_{r1i}, ..., \theta_{rMi})$ and $\phi$ collects all hyperparameters on priors and parameters of $G_{r0}$. Moreover, for $i = 1, ..., n$, and $r = 1, 2$

$$f(y_i; (\beta_0, \beta_1), \theta_{1i}, \theta_{2i}) = \prod_{m=1}^{M} k_p(y_{im} - (\beta_0 + \beta_1 x_m); \theta_{1,im}, \theta_{2,im}).$$

• A critical advantage of the DDP model is its flexibility in capturing different shapes for different covariates. Moreover, it also provides posterior predictive inference at both observed and unobserved covariate values.
Figure 4: Simulated data set for DDP model: one covariate with 5 observed values $x_m = 0, 5, 20, 50, 100$, median regression ($p = 0.5$), and 100 response values for each $y_{im} = \beta_0 + \beta_1 x_m + \epsilon_{im}$. Errors $\epsilon_{im}$, having various degrees of skewness and variability, are generated using the family of split-normals. Predictive densities at unobserved covariate values exhibit learning from data points observed at nearby covariates.
GPDP Model: Gaussian process for regression and DP for error

- A fully nonparametric additive quantile regression model

\[ y_i = \sum_{k=1}^{K} h_k(x_{ik}) + \epsilon_i, \]  

(7)

where \( h_k(\cdot) \) are quantile regression functions, with moderate \( K \), and errors \( \epsilon_i \) are independent from a distribution with \( p \)-th quantile at zero.

- We use DP priors for errors \( \epsilon_i \), and independent GP priors for regression functions \( h_k \), with \( E(h_k(x)) = 0 \) for all \( x \). GP covariance functions are isotropic, \( \text{Cov}(h_k(x), h_k(x')) = \tau_k^2 \exp(-\phi_k |x - x'|^{a_k}) \) with random \( \tau_k^2 \) and \( \phi_k \), and fixed \( a_k \).

- GPDP model in hierarchical form:

\[
\begin{align*}
  y_i & \mid \sum_{k=1}^{K} h_k(\cdot), \sigma_{1i}, \sigma_{2i} \quad \overset{\text{ind}}{\sim} \quad k_p(y_i - \sum_{k=1}^{K} h_k(\cdot), \sigma_{1i}, \sigma_{2i}), \\
  h_k(\cdot) & \mid \tau_k^2, \phi_k \quad \overset{\text{ind}}{\sim} \quad GP(0, C_k(\tau_k^2, \phi_k)) \\
  \sigma_{ri} & \mid G_r \quad \overset{iid}{\sim} \quad G_r \\
  G_r & \mid \alpha_r, d_r \quad \overset{iid}{\sim} \quad \text{DP}(\alpha_r G_{r0}),
\end{align*}
\]

(8)

with \( i = 1, \ldots, n \), \( k = 1, \ldots, K \), and \( r = 1, 2 \).
GPDP Model: regression posterior and predictive error density

Figure 5: Simulated data set (100 points), GPDP model, $y = h(x) + \epsilon$, with regression function $h(x) = 0.3 + 0.4x + 0.5\sin(2.7x) + 1.1/(1 + x^2)$, and symmetric error, $\epsilon \sim 0.07N(0, 1) + 0.5N(0, 0.1)$. 

Bayesian Nonparametric Modeling in Quantile Regression 13/14
GPDP Model: regression posterior and predictive error density

Figure 6: Simulated data set (150 points), GPDP model: \( y = h(x) + \epsilon \), with regression function \( h(x) = 0.3 + 0.4 * x + 0.5 * \sin(2.7 * x) + 1.1/(1 + x * x) \), and zero-median, skewed error obtained using split Gaussians: \( \epsilon \sim 0.5N(0, 0.3)1(\epsilon < 0) + 0.5N(0, 0.8)1(\epsilon > 0) \).