

Variational Inference for Bayesian Density Regression.

Eric Chuu

Department of Statistics, Texas AM University

ericchuu@tamu.edu

1. Introduction

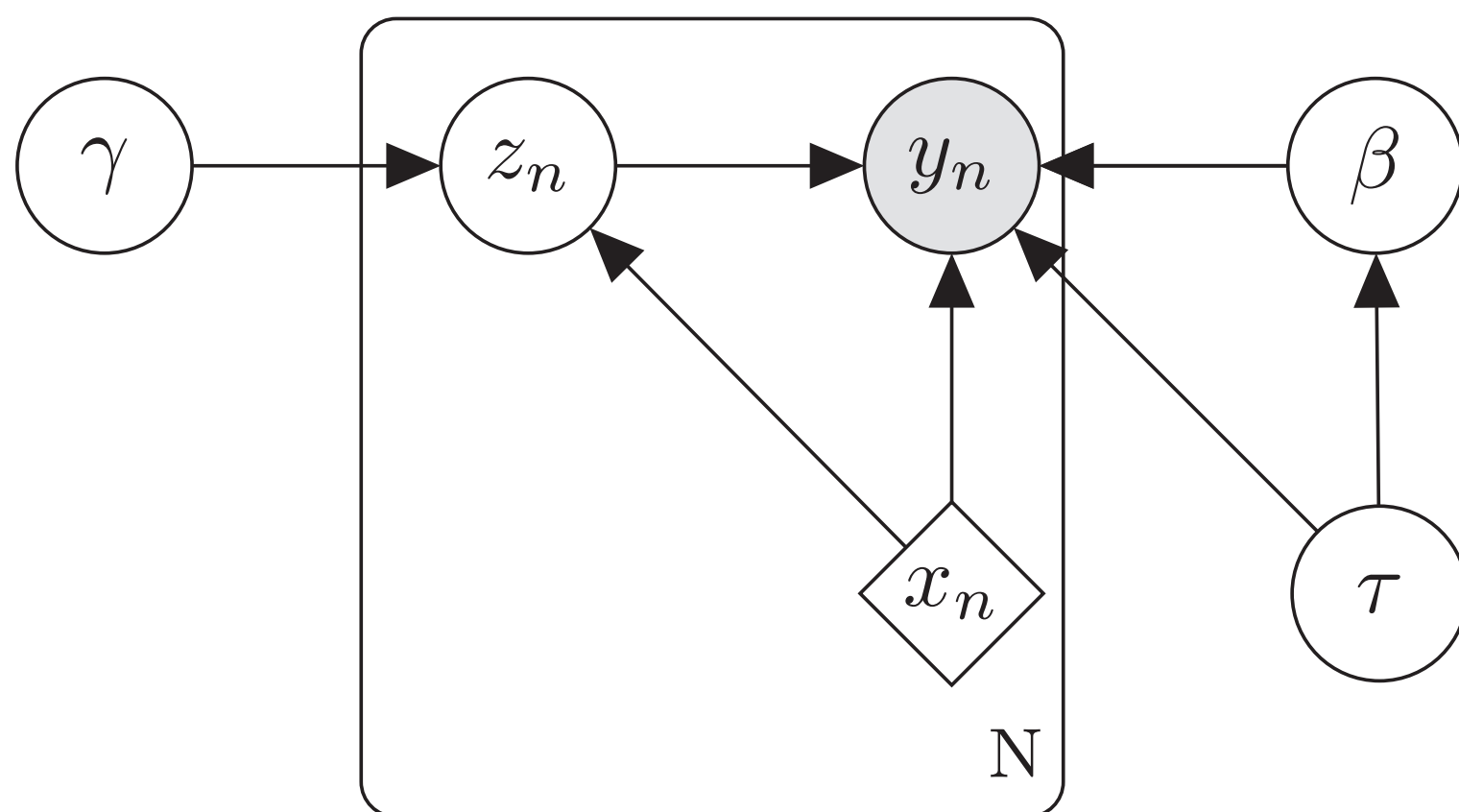
In the Bayesian density regression problem, we observe data $(y_n, x_n), n = 1, \dots, N$, and the goal is to estimate the conditional density of $y | x$. We model the density using a mixture of Gaussians for which covariates enter the weights through a logit link function.

$$f(y | x) = \sum_k^K \pi_k(x) \cdot \mathcal{N}(y | \mu_k(x), \tau_k^{-1})$$

where $\mu_k(x) = x^\top \beta_k$ and $\pi_k(x) \propto \exp(x^\top \gamma_k)$. While this increases the flexibility of the model, it also increases the computational complexity. In order to perform scalable inference on the model parameters, we propose a variational approach that uses a tangential approximation of the softmax function to achieve fast, closed form updates for the coordinate ascent algorithm.

2. Notation

- Data: $\mathbf{y} = \{y_{1:N}\}, \mathbf{X} = \{x_{1:N}\} \subseteq \mathbb{R}^D$
- Coefficients: $\beta = \{\beta_{1:K}\}, \gamma = \{\gamma_{1:K}\}$
- Precision (Gaussian): $\tau = \{\tau_{1:K}\}$
- Cluster Indicator: $\mathbf{Z} = \{z_{1:N}\} \subseteq \mathbb{R}^K$



3. Model Setup

We use conjugate priors to ease computation.

$$p(\mathbf{y} | \mathbf{X}, \beta, \tau, \mathbf{Z}) = \prod_n \prod_k \mathcal{N}(y_n | x_n^\top \beta_k, \tau_k^{-1})^{z_{nk}}$$

$$p(\mathbf{Z} | \mathbf{X}, \gamma) = \prod_n \prod_k \left[\frac{e^{x_n^\top \gamma_k}}{\sum_{j=1}^K e^{x_n^\top \gamma_j}} \right]^{z_{nk}}$$

$$p(\gamma) = \prod_k \mathcal{N}(\gamma_k | 0, \mathbf{I}_D)$$

$$p(\beta, \tau) = \prod_k p(\beta_k | \tau_k) p(\tau_k)$$

$$p(\beta_k | \tau_k) = \mathcal{N}(\beta_k | m_0, (\tau_k \Lambda_0)^{-1})$$

$$p(\tau_k) = \text{Gamma}(\tau_k | a_0, b_0)$$

4. Variational Approximation

We approximate the posterior distribution with:

$$q(\mathbf{Z}, \beta, \tau, \gamma) = q(\mathbf{Z})q(\beta, \tau, \gamma)$$

Then the distribution for each of the variational parameters can be found by taking the expectation of the joint likelihood with respect to the *other* variational parameters.

$$q^*(z_{nk}) = r_{nk}^{z_{nk}}$$

$$q^*(\gamma_k) = \mathcal{N}(\gamma_k | \mu_k, \mathbf{Q}_k^{-1})$$

$$q^*(\beta_k | \tau_k) = \mathcal{N}(\beta_k | m_k, (\tau_k \mathbf{V}_k)^{-1})$$

$$q^*(\tau_k) = \text{Ga}(\tau_k | a_k, b_k)$$

An issue arises in calculating $q(z_{nk})$ because it requires computing:

$$\varepsilon_n = \mathbb{E}_{q(\gamma)} \left[\ln \sum_j \exp\{x_n^\top \gamma_j\} \right]$$

which is not available in closed form. We resort to the following bound (Bouchard, 2007)

$$\varepsilon_n \leq \alpha_n + \sum_j^K \frac{1}{2} (x_n^\top \mu_j - \alpha_n - \xi_{nj}) + \log(1 + e^{\xi_{nj}})$$

Two additional variational parameters:

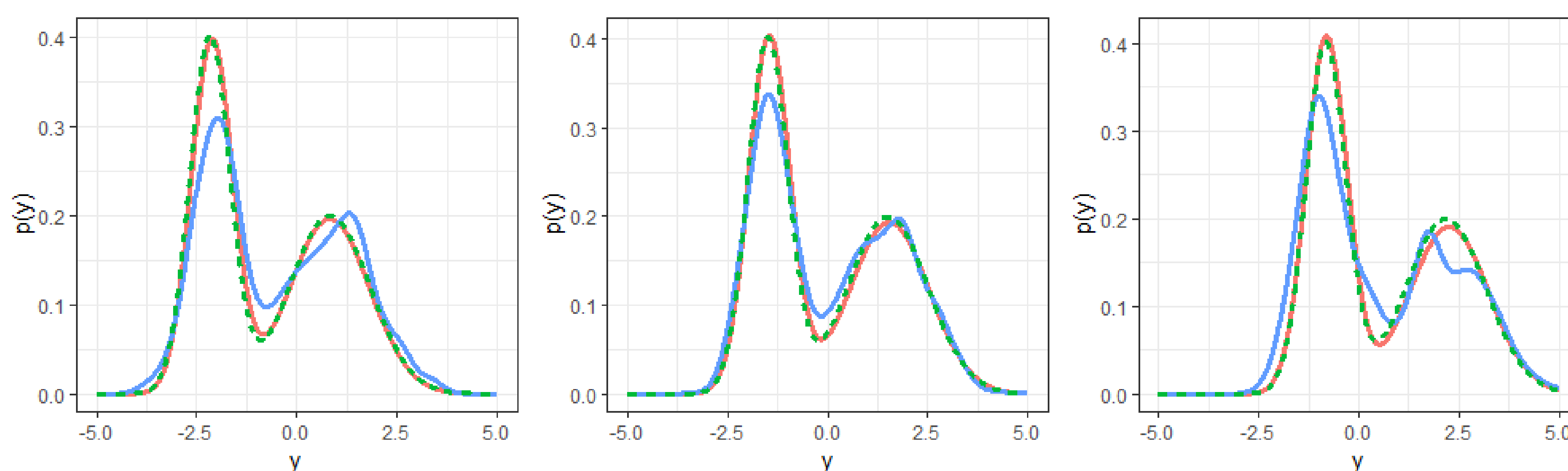
$$\xi_{nj} = \sqrt{(\mu_j^\top x_n - \alpha_n)^2 + x_n^\top \mathbf{Q}_j^{-1} x_n}$$

$$\alpha_n = \frac{\frac{1}{2} \left(\frac{K}{2} - 1 \right) + \sum_{j=1}^K \lambda(\xi_{nj}) \mu_j^\top x_n}{\sum_{j=1}^K \lambda(\xi_{nj})}$$

$$\lambda(\xi) = \frac{1}{4\xi} \tanh\left(\frac{\xi}{2}\right)$$

5. Application to Bimodal Conditional Densities

For the conditional density, $X \sim \mathcal{N}(0, 1), Y | X \sim 0.5\mathcal{N}(X - 1.5, 0.5^2) + 0.5\mathcal{N}(X + 1.5, 1^2)$, we examine samples of size 1000 and look at the conditional density at the three quartiles of the predictor support: $x = -0.6745$ (left), $x = 0$ (center), $x = 0.6745$ (right). For a sample size of 1000, we plot the approximations below. The true density is a dashed green line, the variational approximation is in red, and the kernel density estimate is in blue.



6. Application to Speedflow Data

We consider the following bimodal conditional density, $X \sim \mathcal{N}(0, 1), Y | X \sim 0.5\mathcal{N}(X - 1.5, 0.5^2) + 0.5\mathcal{N}(X + 1.5, 0.5^2)$. In particular, we look at the conditional density at the three quartiles of the predictor support: $x = -0.6745$ (left), $x = 0$ (center), $x = 0.6745$ (right). For a sample size of 1000, we plot the approximations below, where the true conditional density is a dashed green line, the variational approximation is shown in red, and the kernel density estimate is shown in blue.

