Variational Inference for Bayesian Density Regression.

Eric Chuu
Department of Statistics, Texas AM University
ericchuu@tamu.edu

1. Introduction

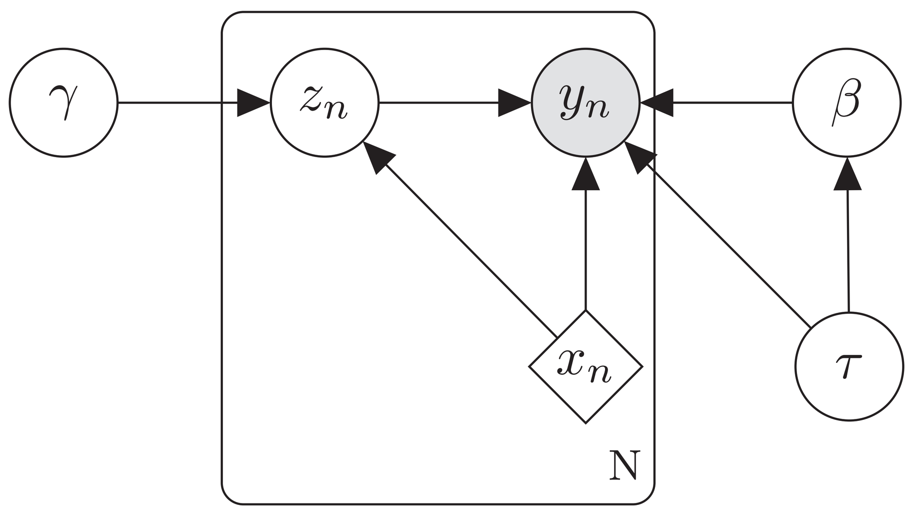
In the Bayesian density regression problem, we observe data $(y_n, x_n), n = 1, \ldots, N$, and the goal is the estimate the conditional density of $y \mid x$. We model the density using a mixture of Gaussians for which covariates enter the weights through a logit link function.

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

where $\mu_k(x) = x^\intercal \beta_k$ and $\pi_k(x) \propto \exp(x^\intercal \gamma_k)$. 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: $\boldsymbol{\beta} = \{\beta_{1:K}\}, \boldsymbol{\gamma} = \{\gamma_{1:K}\}$
- Precision (Gaussian): $\boldsymbol{\tau} = \{\tau_{1:K}\}$
- Cluster Indicator: $\mathbf{Z} = \{z_{1:N}\} \subseteq \mathbb{R}^K$



3. Model Setup

We use conjugate priors to ease computation.

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

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

$$p(\boldsymbol{\gamma}) = \prod_k \mathcal{N}(\gamma_k \mid 0, \mathrm{I}_D)$$

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}) = \prod_k p(\beta_k \mid \tau_k)\, p(\tau_k)$$

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

$$p(\tau_k) = \mathrm{Gamma}(\tau_k \mid a_0, b_0)$$

4. Variational Approximation

We approximate the posterior distribution with:

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

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

$$q^\star(\mathbf{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}}$$

$$q^\star(\gamma_k) = \mathcal{N}\left(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1}\right)$$

$$q^\star(\beta_k \mid \tau_k) = \mathcal{N}\left(\beta_k \mid m_k, (\tau_k \mathrm{V}_k)^{-1}\right)$$

$$q^\star(\tau_k) = \mathrm{Ga}(\tau_k \mid a_k, b_k)$$

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

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

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

$$\varepsilon_n \leq \alpha_n + \frac{1}{2}\left(x_n^\intercal \mu_j - \alpha_n + \xi_{nj}\right)$$

$$+ \sum_j^K \lambda(\xi_{nj}) \left((x_n^\intercal \mu_j - \alpha_k)^2 - \xi_{nj}^2 + x_n^\intercal \mathrm{Q}_k^{-1} x_n\right)$$

$$+ \log(1 + e^{\xi_{nj}})$$

Two additional variational parameters:

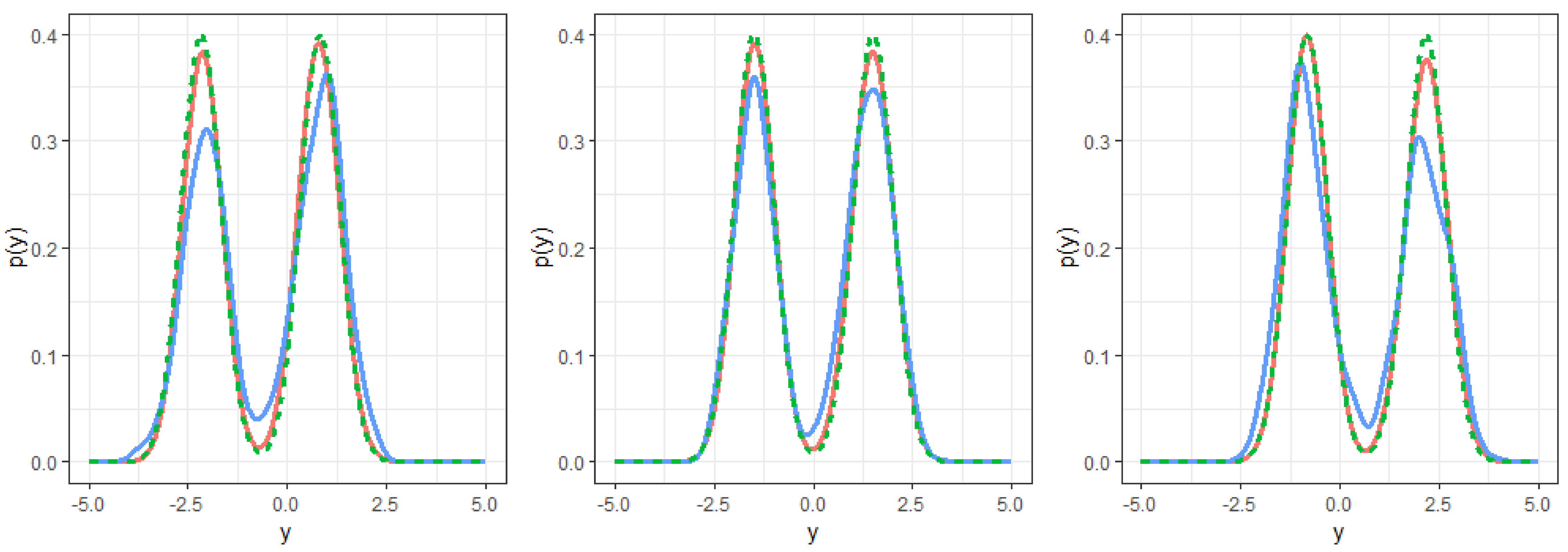
$$\xi_{nk} = \sqrt{\left(\mu_k^\intercal x_n - \alpha_n\right)^2 + x_n^\intercal \mathrm{Q}_k^{-1} x_n}$$

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

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

5. Application to Bimodal Conditional Densities

For the conditional density, $X \sim \mathcal{N}(0,1), Y \mid X \sim 0.5\mathcal{N}\left(X - 1.5, 0.5^2\right) + 0.5\mathcal{N}\left(X + 1.5, 0.5^2\right)$, 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), $x = 0.6475$ (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, $X \sim \mathcal{N}(0,1), Y \mid X \sim 0.5\mathcal{N}\left(X - 1.5, 0.5^2\right) + 0.5\mathcal{N}\left(X + 1.5, 0.5^2\right)$ In particular, we look at the conditional density at the three quartiles of the predictor support: $x = -0.6745$ (left), $x = 0$ (center), $x = 0.6475$ (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.