1 Introduction to Quantile Regression

Typically in regression analyses, the standard linear model is as follows:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{1}$$

where $f(x_i)$ is thought of as the conditional mean of y_i given the predictor variables $\mathbf{x}_i$. The error term is assumed to be homoskedastic with mean zero which facilitates the inference for the parameters. While the loss function taken for the mean regression is the quadratic error

$$\min_{\beta} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2, \tag{2}$$

the least squares method simply will not cut it for quantile regression which calls for another type of a loss function. We will call this the *check function* following the literature of Kneib and is expressed as

$$\rho_p(x) = x(p - I(x < 0)), \tag{3}$$

where p is the conditional quantile of interest that is being modeled. Equivalent representations exist:

$$\begin{aligned} \rho_p(x) &= x(pI(x > 0) - (1-p)I(x < 0)) \\ &= \frac{1}{2}(|x| + (2p-1)x). \end{aligned}$$

Now the loss function transforms to

$$\min_{\beta} \sum_{i=1}^{n} \rho_p(y_i - f(\mathbf{x})). \tag{4}$$

1.1 Asymmetric Laplace distribution

Although the frequentist design of quantile regression does not allow for a likelihood function for the response variables since they directly model the loss function, Bayesian approach does not stand a chance without it as it is an indispensable machinery for the tools of Bayesian inference. Therefore, we introduce a density that we can play with whose maximization is shown to be equivalent to the minimization of the frequentist loss function. Somehow Bayesian methodology is inextricably dependent upon the frequentist formulation. What a shame. Anyway, the

density is called the *asymmetric Laplace distribution* whose density is as follows:

$$f_X(x \mid p) = p(1-p)\exp(-\rho_p(x)), \tag{5}$$

where, as before, $0 < p < 1$, which becomes the quantile of interest, becomes the parameter and ρ is the check function. Setting $p = 1/2$ puts back the asymmetric version to the symmetric Laplace distribution. It can also embrace both location and scale parameters like the Gaussian distribution:

$$f_X(x \mid p, \mu, \sigma) = \frac{p(1-p)}{\sigma}\exp\left(-\rho_p\left(\frac{x-\mu}{\sigma}\right)\right) \text{ if } X \sim \text{ALD}(\mu, \sigma, p). \tag{6}$$

1.2 Alternative Representations for ALD

There arise situations where, for instance, the mixing speed of an MCMC algorithm is too slow and another representation is needed to accelerate inference. For such circumstances, we provide useful alternatives to the vanilla density function of ALD.

Let $U \sim \text{Exp}(\sigma^{-1})$ and $Z \sim \mathcal{N}(0,1)$ be independeant random variables. (Exponential distribution is rate-parameterized.) Then $Y \sim \text{ALD}(\mu, \sigma, p)$ can be represented by

$$Y \overset{d}{\sim} \mu + \nu_p U + \tau_p\sqrt{\sigma U}Z, \tag{7}$$

where $\nu_p = (p(1-p))^{-1}(1-2p)$ and $\tau_p^2 = 2(p(1-p))^{-1}$.

(7) allows a much handier mixture representation of ALD.

$$\begin{aligned} Y \mid U = u &\sim \mathcal{N}\left(\mu + \nu_p u, \tau_p^2 \sigma u\right), \\ U &\sim \text{Exp}\left(\sigma^{-1}\right). \end{aligned}$$

It naturally follows that the conditional distribution of U given Y, is $U \mid Y = y \sim \text{GIG}(2^{-1}, \delta, \gamma)$ where $\delta = \tau_p^{-1}\sigma^{-1/2}|y-\mu|$ and $\gamma = \sigma^{-1/2}\left(2 + \tau_p^{-2}\nu_p^2\right)^{1/2} = 2^{-1}\sigma^{-1/2}\tau_p$. $\text{GIG}(\omega, a, b)$ denotes the *generalized inverse Gaussian distribution* with pdf

$$f(x \mid \omega, a, b) = \frac{(b/a)^\omega}{2K_\omega(ab)}x^{\omega-1}\exp\left(-\frac{1}{2}\left(a^2x^{-1} + b^2x\right)\right), \quad x > 0,\ \omega \in \mathbb{R},\ a, b > 0, \tag{8}$$

where $K_\omega(\cdot)$ is a modified Bessel function of the third kind. The moments of X are given by

$$\mathrm{E}\left(X^k\right) = \left(\frac{a}{b}\right)^k \frac{K_{\omega+k}(ab)}{K_\omega(ab)}, \quad k \in \mathbb{R}. \tag{9}$$

For the time being, keep in mind some basic properties of the Bessel function of the third kind:

- $K_\omega(x) = K_{-\omega}(x)$
- $K_{\omega+1}(x) = \frac{2\omega}{x} K_\omega(x) + K_{\omega-1}(x)$
- $K_{r+1/2}(x) = \sqrt{\frac{\pi}{2x}} \exp(-x) \sum_{k=1}^{r} \frac{(r+k)!\,(2x)^{-k}}{(r-k)!k!}, \quad \forall r \in \mathbb{Z}_+ \cup \{0\}$
- $K_{1/2}(x) = \sqrt{\frac{\pi}{2x}} \exp(-x)$

2 Quantile Regression

As we always do, if we know how to set up the likelihood function and prior distributions, then we can come up with some sort of inference mechanism whether be it MCMC or VB. Although it is possible to keep the ALD density within the likelihood function, it is hard to deal with indicator functions hidden inside the check function. This is where the mixture representation comes in handy. We will cosine-transform the covariate and impose the same priors as in Lenk and Choi(2015).

2.1 Likelihood

Using the mixture representation,

$$L\left(u_1, \ldots, u_n, \boldsymbol{\beta}, \boldsymbol{\theta}_J, \gamma, \tau^2 | \mathbf{y}\right) \propto \prod_{i=1}^{n} u_i^{-1/2} \cdot \exp\left(-\sum_{i=1}^{n} \frac{\left(y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - \boldsymbol{\varphi}_i^\top \boldsymbol{\theta}_J - \nu_p u_i\right)^2}{2\tau_p^2 u_i}\right) \tag{10}$$

$$\prod_{i=1}^{n} \exp(-u_i) \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta^0\right)^\top \boldsymbol{\Sigma}_\beta^{0\,-1}\left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta^0\right)\right) \tag{11}$$

$$\prod_{j=1}^{J} \left(\tau^2 e^{-j\gamma}\right)^{-1/2} \exp\left(-\frac{e^{j\gamma}}{2\tau^2}\theta_j^2\right) \cdot \exp(-\omega_0 \gamma) \tag{12}$$

$$\tau^{2-(A+1)} \exp\left(-B/\tau^2\right). \tag{13}$$

As can be seen, the specifications are as follows separately:

- $y_i \,|\, u_i, \boldsymbol{\beta}, \boldsymbol{\theta}_J \sim \mathcal{N}\left(\boldsymbol{w}_i^\top \boldsymbol{\beta} + \boldsymbol{\varphi}_i^\top \boldsymbol{\theta}_J + \nu_p u_i, \tau_p^2 u_i\right)$
- $u_i \sim \mathrm{Exp}\,(1)$
- $\boldsymbol{\beta} \sim \mathcal{N}\left(\boldsymbol{\mu}_\beta^0, \boldsymbol{\Sigma}_\beta^0\right)$
- $\boldsymbol{\theta}_J \sim \mathcal{N}\left(0, \tau^2 \exp\left(-j\gamma\right)\right)$
- $\gamma \sim \mathrm{Exp}\,(\omega_0)$
- $\tau^2 \sim \mathrm{InvGam}\,(A, B)$

2.2 MCMC scheme

Every parameter excluding γ is conditionally conjugate enough to achieve a Gibbs sampler. The inference for γ requires MH updates.

- $u_i \,|\, \boldsymbol{\beta}, \boldsymbol{\theta}_J, y_i \sim \mathrm{GIG}\left(2^{-1}, \tau_p^{-1}\left|y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - \boldsymbol{\varphi}_i^\top \boldsymbol{\theta}_J\right|, 2^{-1}\tau_p\right)$
- $\boldsymbol{\beta} \,|\, \{u_i\}_{i=1}^n, \boldsymbol{\theta}_J \sim \mathcal{N}\left(\boldsymbol{\mu}_\beta^{\mathrm{pos}}, \boldsymbol{\Sigma}_\beta^{\mathrm{pos}}\right)$ where
  - $\boldsymbol{\Sigma}_\beta^{\mathrm{pos}} = \left(\dfrac{1}{\tau_p^2}\displaystyle\sum_{i=1}^n \dfrac{1}{u_i}\boldsymbol{w}_i\boldsymbol{w}_i^\top + \boldsymbol{\Sigma}_\beta^{0\,-1}\right)^{-1}$
  - $\boldsymbol{\mu}_\beta^{\mathrm{pos}} = \boldsymbol{\Sigma}_\beta^{\mathrm{pos}}\left(\dfrac{1}{\tau_p^2}\displaystyle\sum_{i=1}^n \dfrac{\boldsymbol{w}_i\left(y_i - \boldsymbol{\varphi}_i^\top \boldsymbol{\theta}_J - \nu_p u_i\right)}{u_i} + \boldsymbol{\Sigma}_\beta^{0\,-1}\boldsymbol{\mu}_\beta^0\right)$
- $\boldsymbol{\theta}_J \,|\, \{u_i\}_{i=1}^n, \gamma, \boldsymbol{\beta}, \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_\theta^{\mathrm{pos}}, \boldsymbol{\Sigma}_\theta^{\mathrm{pos}}\right)$ where
  - $\boldsymbol{\Sigma}_\theta^{\mathrm{pos}} = \left(\dfrac{1}{\tau_p^2}\displaystyle\sum_{i=1}^n \dfrac{1}{u_i}\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\top + \boldsymbol{E}\right)^{-1}$, where $\boldsymbol{E} = \dfrac{1}{\tau^2}\mathrm{diag}\left(e^\gamma, e^{2\gamma}, \ldots, e^{J\gamma}\right)$
  - $\boldsymbol{\mu}_\theta^{\mathrm{pos}} = \boldsymbol{\Sigma}_\theta^{\mathrm{pos}}\left(\dfrac{1}{\tau_p^2}\displaystyle\sum_{i=1}^n \dfrac{1}{u_i}\boldsymbol{\varphi}_i\left(y_i - \boldsymbol{w}_i^\top \boldsymbol{\beta} - \nu_p u_i\right)\right)$
- $\tau^2 \,|\, \boldsymbol{\theta}_J, \gamma \sim \mathrm{InvGam}\left(A + \dfrac{J}{2}, \dfrac{1}{2}\boldsymbol{\theta}_J^\top \boldsymbol{E}^* \boldsymbol{\theta}_J + B\right), \quad \boldsymbol{E}^* = \mathrm{diag}\left(e^\gamma, e^{2\gamma}, \ldots, e^{J\gamma}\right)$
- $p\left(\gamma \,|\, \tau^2, \boldsymbol{\theta}_J\right) \propto \exp\left(\left(\dfrac{J\left(J+1\right)}{4} - \omega_0\right)\gamma - \dfrac{1}{2\tau^2}\displaystyle\sum_{j=1}^J e^{j\gamma}\theta_j^2\right)$

The full conditional of γ resembles the Gompertz density with canonical parameters

$$f\left(x \,|\, \eta, b\right) = b\eta e^{\eta} e^{bx} \exp\left(-\eta e^{bx}\right) \quad \text{for } x \geq 0,\ \eta, b > 0 \tag{14}$$

but is impossible to separate out the parameters; hence, we construct an MH chain with a standard exponential proposal. If we choose the standard exponential distribution to be our proposal, then the algorithm is as follows:

1. Generate $\gamma_{\text{cand}} \sim \text{Exp}\left(1\right)$.
2. Take $\gamma^{(t+1)} = \begin{cases} \gamma_{\text{cand}}, & \text{with } p = 1 \wedge \dfrac{f\left(\gamma_{\text{cand}}\right) g\left(\gamma^{(t)}\right)}{f\left(\gamma^{(t)}\right) g\left(\gamma_{\text{cand}}\right)} \left(= \rho\right) \\ \gamma^{(t)} & \text{otherwise} \end{cases}$, where
   - $f\left(\gamma\right) = \exp\left(\left(\dfrac{J\left(J+1\right)}{4} - \omega_0\right)\gamma - \dfrac{1}{2\tau^2}\displaystyle\sum_{j=1}^{J} e^{j\gamma}\theta_j^2\right)$
   - $g\left(\gamma\right) = \exp\left(-\gamma\right)$.
   - $\wedge$ is 'choose the minimum' operator.

Rearranging the elements,

$$\rho = \exp\left(\left(\frac{J\left(J+1\right)}{4} - \omega_0\right)\left(\gamma_{\text{cand}} - \gamma^{(t)}\right) - \frac{1}{2\tau^2}\sum_{j=1}^{J}\left(e^{j\gamma_{\text{cand}}} - e^{j\gamma^{(t)}}\right)\theta_j^2 - \gamma^{(t)} + \gamma_{\text{cand}}\right) \tag{15}$$

2.3 Variational Inference

Variational inference is an alternative to MCMC which aims for speed improvement at the expense of performance. There is a vast literature on the variational methods so we do not elaborate more on this topic. For those who are interested should be referred to Wand & Ormerod (2010). The modelling of variational inference is different from the MCMC version in that the prior distribution of γ is no more exponential but rather double exponential.

- $q\left(u_i\right) = \text{GIG}\left(\dfrac{1}{2}, a_q, b_q\right)$ where
  - $a_q^2 = \dfrac{1}{\tau_p^2}\left(\left(y_i - \boldsymbol{w}_i^{\top}\boldsymbol{\mu}_{\beta}^{q} - \boldsymbol{\varphi}_i^{\top}\boldsymbol{\mu}_{\theta}^{q}\right)^2 + \text{Tr}\left(\boldsymbol{w}_i\boldsymbol{w}_i^{\top}\boldsymbol{\Sigma}_{\beta}^{q}\right) + \text{Tr}\left(\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^{\top}\boldsymbol{\Sigma}_{\theta}^{q}\right)\right)$
  - $b_q^2 = 2 + \dfrac{\nu_p^2}{\tau_p^2}$

- $q(\boldsymbol{\beta}) = \mathcal{N}\left(\boldsymbol{\mu}_\beta^q, \boldsymbol{\Sigma}_\beta^q\right)$
  - $\boldsymbol{\Sigma}_\beta^q = \left(\frac{1}{\tau_p^2}\sum_{i=1}^{n}\frac{1}{u_i}\boldsymbol{w}_i\boldsymbol{w}_i^\top + \boldsymbol{\Sigma}_\beta^{0\,-1}\right)^{-1}$
  - $\boldsymbol{\mu}_\beta^q = \boldsymbol{\Sigma}_\beta^q\left(\frac{1}{\tau_p^2}\sum_{i=1}^{n}\boldsymbol{w}_i\left(\mathrm{E}\left(\frac{1}{u_i}\right)\left(y_i - \boldsymbol{\varphi_i}^\top\boldsymbol{\mu}_\theta^q\right) - \nu_p\right) + \boldsymbol{\Sigma}_\beta^{0\,-1}\boldsymbol{\mu}_\beta^0\right)$
- $q(\boldsymbol{\theta}_J) = \mathcal{N}(\boldsymbol{\mu}_\theta^q, \boldsymbol{\Sigma}_\theta^q)$
  - $\boldsymbol{\Sigma}_\theta^q = \left(\frac{1}{\tau_p^2}\sum_{i=1}^{n}\mathrm{E}\left(\frac{1}{u_i}\right)\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\top + \boldsymbol{E}\right)^{-1}$ where $\boldsymbol{E} = \mathrm{E}\left(\frac{1}{\tau^2}\right)\mathrm{diag}\left(\mathrm{E}\left(e^{\gamma}\right), \mathrm{E}\left(e^{2\gamma}\right), \ldots, \mathrm{E}\left(e^{J\gamma}\right)\right)$
  - $\boldsymbol{\mu}_\theta^q = \boldsymbol{\Sigma}_\theta^q\left(\frac{1}{\tau_p^2}\sum_{i=1}^{n}\boldsymbol{\varphi}_i\left(\mathrm{E}\left(\frac{1}{u_i}\right)\left(y_i - \boldsymbol{w}_i^\top\boldsymbol{\mu}_\beta^q\right) - \nu_p\right)\right)$
- $q(\tau^2) = \mathrm{InvGam}(A_q, B_q)$
  - $A_q = \frac{J}{2} + A$
  - $B_q = \frac{1}{2}\left(\boldsymbol{\mu}_\theta^{q\top}\boldsymbol{E}^*\boldsymbol{\mu}_\theta^q + \mathrm{Tr}\left(\boldsymbol{E}^*\boldsymbol{\Sigma}_\theta^q\right)\right) + B$ where $\boldsymbol{E}^* = \mathrm{diag}\left(\mathrm{E}\left(e^{\gamma}\right), \ldots, \mathrm{E}\left(e^{J\gamma}\right)\right)$.
- $q(\gamma) = \mathcal{N}\left(\mu_\gamma, \sigma_\gamma^2\right)$

2.3.1 Lower Bound

As is with every optimization problem, variational inference also calls for an objective function to optimize. We will derive the lower bound for our quantile regression model in this section.

- $L_1 = \mathrm{E}\left(\ln p\left(y_i \mid u_i, \boldsymbol{\beta}, \boldsymbol{\theta}_J\right)\right)$

$$L_1 = -\frac{1}{2}\left(\ln\left(2\pi\tau_p^2\right) + \mathrm{E}\left(\ln u_i\right)\right) \tag{16}$$

$$+\frac{1}{\tau_p^2}\mathrm{E}\left(\frac{1}{u_i}\right)\left(\left(y_i - \boldsymbol{w}_i^\top\boldsymbol{\mu}_\beta^q - \boldsymbol{\varphi}_i^\top\boldsymbol{\mu}_\theta^q\right)^2 + \mathrm{Tr}\left(\boldsymbol{w}_i\boldsymbol{w}_i^\top\boldsymbol{\Sigma}_\beta^q\right) + \mathrm{Tr}\left(\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\top\boldsymbol{\Sigma}_\theta^q\right)\right) \tag{17}$$

- $L_2 = \mathrm{E}\left(\ln p\left(u_i\right)\right)$

$$L_2 = -\mathrm{E}\left(u_i\right) \tag{18}$$

- $L_3 = \mathrm{E}\left(\ln p\left(\boldsymbol{\beta}\right)\right)$

$$L_3 = -\frac{p}{2}\ln\left(2\pi\right) - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_\beta^0\right| - \frac{1}{2}\left(\left(\boldsymbol{\mu}_\beta^q - \boldsymbol{\mu}_\beta^0\right)^\top\boldsymbol{\Sigma}_\beta^{0\,-1}\left(\boldsymbol{\mu}_\beta^q - \boldsymbol{\mu}_\beta^0\right) + \mathrm{Tr}\left(\boldsymbol{\Sigma}_\beta^{0\,-1}\boldsymbol{\Sigma}_\beta^q\right)\right) \tag{19}$$

- $L_4 = \mathrm{E}\left(\ln p\left(\boldsymbol{\theta}_J \mid \tau^2, \gamma\right)\right)$

$$L_4 = -\frac{J}{2}\left(\ln\left(2\pi\right) + \ln\left(B_q\right) - \psi\left(A_q\right)\right) + \frac{J\left(J+1\right)}{4}\mathrm{E}\left(|\gamma|\right) + \frac{A_q}{B_q}\boldsymbol{\mu}_\theta^{q\prime}\mathbf{E}\boldsymbol{\mu}_\theta^q \tag{20}$$

where $\mathbf{E} = \mathrm{diag}\left(\mathrm{E}\left(e^{|\gamma|}\right), \ldots, \mathrm{E}\left(e^{J|\gamma|}\right)\right)$ and

$$\mathrm{E}\left(|\gamma|\right) = \sigma_\gamma\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu_\gamma^2}{2\sigma_\gamma^2}\right) + \mu_\gamma\left(1 - 2\Phi\left(-\frac{\mu_\gamma}{\sigma_\gamma}\right)\right) \tag{21}$$

$$\mathrm{E}\left(e^{j|\gamma|}\right) = \exp\left(\frac{\sigma_\gamma^2 j^2}{2} + \mu_\gamma j\right)\left(1 - \Phi\left(-\frac{\mu_\gamma}{\sigma_\gamma} - \sigma_\gamma j\right)\right) + \exp\left(\frac{\sigma_\gamma^2 j^2}{2} - \mu_\gamma j\right)\left(1 - \Phi\left(\frac{\mu_\gamma}{\sigma_\gamma} - \sigma_\gamma j\right)\right) \tag{22}$$

where Φ is the CDF of standard Gaussian. ψ is the digamma function.

- $L_5 = \mathrm{E}\left(\ln p\left(\gamma\right)\right)$

$$L_5 = \ln\left(\frac{\omega_0}{2}\right) - \omega_0\mathrm{E}\left(|\gamma|\right) \tag{23}$$

- $L_6 = \mathrm{E}\left(\ln p\left(\tau^2\right)\right)$

$$L_6 = A\ln B - \ln\Gamma\left(A\right) - \left(A+1\right)\left(\ln B_q - \psi\left(A_q\right)\right) - B\frac{A_q}{B_q} \tag{24}$$

The variational lower bound also requires the entropy of each variational distribution.

- The entropy of u_i

$$\mathrm{H}\left(u_i\right) = \frac{1}{2}\ln\left(\frac{b_q^2}{a_q^2}\right) + \ln\left(2K_{0.5}\left(a_q b_q\right)\right) + \frac{\frac{d}{d\nu}K_\nu\left(a_q b_q\right)\Big|_{\nu=0.5} + a_q b_q\left(K_{1.5}\left(a_q b_q\right) + K_{-0.5}\left(a_q b_q\right)\right)}{2K_{0.5}\left(a_q b_q\right)} \tag{25}$$