Quantile Regression in the Presence of Monotone Missingness

June 20, 2013

Abstract

1 Introduction

Quantile regression is used to study the relationship between a response and covariates when one (or several) quantiles are of interest as opposed to mean regression. The dependence between upper or lower quantiles of the response variable and the covariates often vary differentially relative to that of the mean. How quantiles depend on covariates is of interest in econometrics, educational studies, biomedical studies, and environment studies (Yu and Moyeed, 2001; Buchinsky, 1994, 1998; He et al., 1998; Koenker and Machado, 1999; Wei et al., 2006; Yu et al., 2003). A comprehensive review of applications of quantile regression was presented in Koenker (2005).

Quantile regression is more robust to outliers than mean regression and provides information about how covariates affect quantiles, which offers a more complete description of the conditional distribution of the response. Different effects of covariates can be assumed for different quantiles.

The traditional frequentist approach was proposed by Koenker and Bassett (1978) for a single quantile with estimators derived by minimizing a loss function. The popularity of this approach is due to its computational efficiency, well-developed asymptotic properties, and straightforward extensions to simultaneous quantile regression and random effect models. However, asymptotic inference may not be accurate for small sample sizes and not naturally extend to missing data.

Bayesian approaches offer exact inference in small samples. Motivated by the loss (check) function, Yu and Moyeed (2001) proposed an asymmetric Laplace distribution for the error term, such that maximizing the posterior distribution is equivalent to minimizing the check function. Semiparametric methods have been proposed for median regression. Walker and Mallick (1999) used a diffuse finite Pólya Tree prior for the error term. Kottas and Gelfand (2001) modeled the error by two families of median zero distribution using a mixture Dirichlet process priors, which is very useful for unimodal error distributions. Hanson and Johnson (2002) adopted mixture of Pólya Tree prior in median regression, which is more robust in terms of multimodality and skewness. Other recent approaches include quantile pyramid priors, mixture of Dirichlet process priors of multivariate normal distributions and infinite mixture of Gaussian densities which put quantile constraints on the residuals (Hjort and Petrone, 2007; Hjort and Walker, 2009; Kottas and Krnjajić, 2009; Reich et al., 2010).

The above methods focus on complete data without missingness. There are a few more articles about quantile regression with missingness. Yuan and Yin (2010) introduced a Bayesian quantile regression approach for longitudinal data with nonignorable missing data. They used random effects to explain the within-subject correlation and applied a l_2 penalty in the traditional quantile regression check function to shrink toward the common population values. However, the quantile regression coefficients are conditional on the random effects, which is not of interest if we are looking into the marginal relationship. Wei et al. (2012) proposed a multiple imputation method for quantile regression model when there are some covariates missing at random. They impute the missing covariates by specifying the its conditional density given observed covariates and outcomes, which comes from the estimated conditional quantile regression and specification of conditional density of missing covariates given observed ones. However, they put more focus on the missing covariates rather than missing outcomes. Bottai and Zhen (2013) illustrated a new imputation method by estimated conditional quantiles of missing outcomes given observed data. Their approach does not make distribution assumptions. Their method also has advantages as robustness to outliers and invariance to transformations. Roy and Daniels (2008) proposed a pattern mixture model for data with nonignorable dropout, which borrowed idea from Heagerty (1999). But their methods examine the marginal covariate effects on the mean. We will use these ideas for quantile regression models.

The structure of this article is as follows. First, we introduce a quantile regression method to address monotone nonignorable dropout missingness in section 2, including sensitivity analysis and computational details. We use simulation studies to evaluate the performance of the model in section 3. We apply our approach to data from a recent clinical trial in section 4. Finally, discussion and conclusions are in section 5.

2 Model

In this section, we first introduce some notation , then describe our proposed quantile regression model in section 2.1. We provide details on MAR and MNAR and computation in sections 2.2 and 2.3.

Under monotone dropout, without loss of generality, denote $S_i \in \{1, 2, \dots, J\}$ to be the follow up time, and $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ})^T$ to be the response vector for subject i, where J is the maximum follow up time. We assume Y_{i1} is always observed. We are interested in the τ-th marginal quantile regression coefficients $\boldsymbol{\gamma}_j = (\gamma_{j0}, \gamma_{j2}, \dots, \gamma_{jp})^T$,

$$\Pr(Y_{ij} \leq \boldsymbol{x}_i^T \boldsymbol{\gamma}_j) = \tau, \text{ for } j = 1, \dots, J, \tag{1}$$

where $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates for subject i.

Let

$$\begin{aligned} \mathrm{p}_k(Y) &= \mathrm{p}(Y|S = k), \\ \mathrm{p}_{\geq k}(Y) &= \mathrm{p}(Y|S \geq k) \end{aligned}$$

be the densities of response $\boldsymbol{Y}$ given follow-up time $S = k$ and $S \geq k$. And Pr_k be the corresponding probability given $S = k$.

2.1 Mixture Model Specification

We adopt a pattern mixture model to jointly model the response and missingness. Let the subscript i stand for subject i. Specify the conditional distribution as:

$$\left.\begin{array}{l} \mathrm{p}_k(y_{i1}) = \mathrm{N}(\Delta_{i1} + \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1^{(k)}, \exp(\boldsymbol{x}_{i1}^T\boldsymbol{\alpha}_1^{(k)})), k = 1, \ldots, J, \\ \mathrm{p}_k(y_{ij}|\boldsymbol{y}_{i1:i(j-1)}) = \begin{cases} \mathrm{N}\big(\Delta_{ij} + \boldsymbol{x}_{ij}^T\boldsymbol{h}_j^{(k)} + \boldsymbol{y}_{i1:i(j-1)}^T\boldsymbol{\beta}_{y,j-1}^{(k)}, \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha}_j^{(k)})\big), & k < j; \\ \mathrm{N}\big(\Delta_{ij} + \boldsymbol{y}_{i1:i(j-1)}^T\boldsymbol{\beta}_{y,j-1}^{(\geq j)}, \exp(\boldsymbol{x}_{ij}^T\boldsymbol{\alpha}_j^{(\geq j)})\big), & k \geq j; \end{cases}, \text{ for } 2 \leq j \leq J, \\ S_{ij} = k|\boldsymbol{x}_{ij} \sim \mathrm{Multinomial}(1, \boldsymbol{\pi}), \end{array}\right\} \tag{2}$$

where $\boldsymbol{y}_{i1:i(j-1)} = (y_{i1}, \ldots, y_{i(j-1)})^T, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_J), \boldsymbol{h}_j^{(k)} = (h_{j1}^{(k)}, \ldots, h_{jp}^{(k)})$, $\boldsymbol{x}_j$ is a $p \times 1$ covariate vector, $\boldsymbol{\beta}_{y,j-1}^{(k)} = \big(\beta_{y_1,j-1}^{(k)}, \ldots, \beta_{y_{j-1},j-1}^{(k)}\big)^T$ and $\boldsymbol{\alpha}_j^{(k)}$ is a $p \times 1$ vector controlling heterogeneity of conditional variance of response component j.

In (2), Δ_{ij} are functions of $\tau, \boldsymbol{x}_{ij}, \boldsymbol{\alpha}_j$ and other parameters and are determined by the marginal quantile regressions,

$$\tau = \mathrm{Pr}(Y_{ij} \leq \boldsymbol{x}_{ij}^T\gamma_j) = \sum_{k=1}^{J} \pi_k \mathrm{Pr}_k(Y_{ij} \leq \boldsymbol{x}_{ij}^T\gamma_j), \tag{3}$$

for $j = 1$ and

$$\begin{aligned} \tau = \mathrm{Pr}(Y_{ij} \leq \boldsymbol{x}_{ij}^T\gamma_j) &= \sum_{k=1}^{J} \pi_k \mathrm{Pr}_k(Y_{ij} \leq \boldsymbol{x}_{ij}^T\gamma_j) \qquad (4) \\ &= \sum_{k=1}^{J} \pi_k \int \cdots \int \mathrm{Pr}_k(Y_{ij} \leq \boldsymbol{x}_{ij}^T\gamma_j|y_{i1}, \ldots, y_{i(j-1)})\, \mathrm{p}_k(y_{i(j-1)}|y_{i1}, \ldots, y_{i(j-2)}) \\ &\quad \cdots \mathrm{p}_k(y_{i2}|y_{i1})\, \mathrm{p}_k(y_{i1}) dy_{i(j-1)} \cdots dy_{i1}. \end{aligned}$$

for $j = 2, \ldots, J$. More computational details will be given in section 2.3.

The idea is to model the marginal quantile regressions directly, then to embed them in the likelihood through restrictions in the mixture model. The mixture model and heterogeneity of variance between subjects allows the marginal quantile regression coefficients to differ by quantiles. Otherwise, the quantile lines would be parallel to each other.

For identifiability, we apply the following restrictions,

$$\sum_{k=1}^{J} \beta_{l1}^{(k)} = 0, l = 1, \ldots, p,$$

where $\boldsymbol{\beta}_1^{(k)} = (\beta_{11}^{(k)}, \ldots, \beta_{p1}^{(k)})^T$. Further details on these restrictions can be found in appendix A.

2.2 Missing Data Mechanism and Sensitivity Analysis

In the mixture model in (2), MAR holds (Molenberghs et al., 1998) if and only if, for each $j \geq 2$ and $k < j$:

$$\mathrm{p}_k(y_j|y_1,\ldots,y_{j-1}) = \mathrm{p}_{\geq j}(y_j|y_1,\ldots,y_{j-1}). \tag{5}$$

When $2 \leq j \leq J$ and $k < j$, Y_j is not observed, thus $\boldsymbol{h}_j^{(k)}$ and $\boldsymbol{\alpha}_j^{(k)}$, $\boldsymbol{\beta}_{y,j-1}^{(k)} = \left(\beta_{y_1,j}^{(k)},\ldots,\beta_{y_{j-1},j-1}^{(k)}\right)^T$ can not be identified from the observed data. Denote

$$\boldsymbol{\alpha}_j^{(k)} = \boldsymbol{\alpha}_j^{(\geq j)} + \boldsymbol{\delta}_j^{(k)},$$
$$\boldsymbol{\beta}_{y,j-1}^{(k)} = \boldsymbol{\beta}_{y,j-1}^{(\geq j)} + \boldsymbol{\eta}_{j-1}^{(k)},$$

where $\boldsymbol{\delta}_j^{(k)} = \left(\delta_{1j}^{(k)},\ldots,\delta_{pj}^{(k)}\right)$ and $\boldsymbol{\eta}_{j-1}^{(k)} = \left(\eta_{y_1,j-1}^{(k)},\ldots,\eta_{y_{j-1},j-1}^{(k)}\right)$ for $k < j$. Then $\boldsymbol{\xi}_s = \left(\boldsymbol{h}_j^{(k)},\boldsymbol{\eta}_{j-1}^{(k)},\boldsymbol{\delta}_j^{(k)}\right)$ is a set of sensitivity parameters (Daniels and Hogan, 2008), where $k < j, 2 \leq j \leq J$.

When $\boldsymbol{\xi}_s = \boldsymbol{\xi}_{s0} = \mathbf{0}$, MAR holds. If $\boldsymbol{\xi}_s$ is fixed at $\boldsymbol{\xi}_s \neq \boldsymbol{\xi}_{s0}$, the missingness mechanism is MNAR.

For fully Bayesian inference, We can put independent priors on $(\boldsymbol{\xi}_s,\boldsymbol{\xi}_m)$ as :

$$p(\boldsymbol{\xi}_s,\boldsymbol{\xi}_m) = p(\boldsymbol{\xi}_s)p(\boldsymbol{\xi}_m),$$

where $\boldsymbol{\xi}_m = \left(\gamma_j,\boldsymbol{\beta}_{y,j-1}^{(\geq j)},\boldsymbol{\alpha}_j^{(\geq j)},\boldsymbol{\pi}\right)$.

If we assume MAR with no uncertainty, the prior of $\boldsymbol{\xi}_s$ is $\mathrm{p}(\boldsymbol{\xi}_s = \mathbf{0}) \equiv 1$. Sensitivity analysis can be executed by putting point mass priors on $\boldsymbol{\xi}_s$ to examine the effect of priors on the posterior inference about quantile regression coefficients γ_{ij}^{τ}. For example, if MAR is assumed with uncertainty, priors can be assigned as $\mathrm{E}(\boldsymbol{\xi}_s) = \boldsymbol{\xi}_{s0} = \mathbf{0}$ with $\mathrm{Var}(\boldsymbol{\xi}_s) \neq \mathbf{0}$. If we assume MNAR with no uncertainty, we can put priors satisfying $\mathrm{E}(\boldsymbol{\xi}_s) = \Delta_{\xi}$, where $\Delta_{\xi} \neq \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\xi}_s) = \mathbf{0}$. If MNAR is assumed with uncertainty, then priors could be $\mathrm{E}(\boldsymbol{\xi}_s) = \Delta_{\xi}$, where $\Delta_{\xi} \neq \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\xi}_s) \neq \mathbf{0}$.

2.3 Computation

In section 2.3.1, we provide details on calculating Δ_{ij} in (2) for $j = 1,\ldots,J$. Then we show how to obtain maximum likelihood estimates using an adaptive gradient descent algorithm in section 2.3.2. Finally, we present a MCMC sampling algorithm for Bayesian inference in section 2.3.3.

2.3.1 Calculation of Δ

From equation (3) and (4), Δ_{ij} depends on subject-specific covariates $\boldsymbol{x}_i$, thus Δ_{ij} needs to be calculated for each subject. We now illustrate how to calculate Δ_{ij} given all the other parameters $\boldsymbol{\xi} = (\boldsymbol{\xi}_m,\xi_s)$.

- Δ_{i1} : Expand equation (3):

$$\tau = \sum_{k=1}^{J} \pi_k \Phi\left(\frac{\boldsymbol{x}_{i1}^T\gamma_1 - \Delta_{i1} - \boldsymbol{x}_{i1}^T\boldsymbol{\beta}_1^{(k)}}{\exp\left(\boldsymbol{x}_{i1}^T\boldsymbol{\alpha}_1^{(k)}\right)}\right),$$

where Φ is the standard normal CDF. Because the above equation is continuous and monotone in Δ_{i1}, it can be solved by a standard numerical root-finding method (e.g. bisection method) with minimal difficulty.

- $\Delta_{ij}, 2 \leq j \leq J$:

  First we introduce a lemma:

  Lemma 2.1. *An integral of a normal CDF with mean b and standard deviation a over another normal distribution with mean μ and standard deviation σ can be simplified to a closed form in terms of normal CDF:*

$$\int \Phi\left(\frac{x-b}{a}\right) d\Phi(x;\mu,\sigma) = \begin{cases} 1-\Phi\left(\frac{b-\mu}{\sigma}/\sqrt{\frac{a^2}{\sigma^2}+1}\right) & a>0, \\ \Phi\left(\frac{b-\mu}{\sigma}/\sqrt{\frac{a^2}{\sigma^2}+1}\right) & a<0, \end{cases} \tag{6}$$

  where $\Phi(x;\mu,\sigma)$ stands for a CDF of normal distribution with mean μ and standard deviation σ.

  Proof of 2.1 is in Appendix B.

  To solve equation (4), we propose a recursive approach. For the first multiple integral in equation (4), apply lemma 2.1 once to obtain:

$$\begin{aligned} \Pr(Y_j \leq \boldsymbol{x}^T\gamma_j|S=1) &= \int\dots\int \Pr(Y_j \leq \boldsymbol{x}^T\gamma_j|S=1,\boldsymbol{x},Y_{j-1},\dots,Y_1) \\ &\quad dF(Y_{j-1}|S=1,Y_{j-2},\dots,Y_1)\cdots dF(Y_2|S=1,Y_1)dF(Y_1|S=1), \\ &= \int\dots\int \Phi\left(\frac{\boldsymbol{x}^T\gamma_j - \mu_{j|1,\dots,j-1}(y_{j-1})}{\sigma_{j|1,\dots,j-1}}\right) \\ &\quad dF(Y_{j-1}|S=1,Y_{j-2},\dots,Y_1)\cdots dF(Y_2|S=1,Y_1)dF(Y_1|S=1), \\ &= \int\dots\int \Phi\left(\frac{y_{j-2}-b^*}{a^*}\right) dF(Y_{j-2}|S=1,Y_{j-3},\dots,Y_1)\cdots dF(Y_1|S=1). \end{aligned}$$

  Then, by recursively applying lemma 2.1 $(j-1)$ times, each multiple integral in equation (4) can be simplified to single normal CDF. Thus it can be solved using standard numerical root-find method for Δ_{ij} as for $j=1$.

2.3.2 Maximum Likelihood Estimation

The observed data likelihood for an individual $\boldsymbol{y}_i$ with follow-up time $S=k$ is

$$\begin{aligned} L_i(\boldsymbol{\xi}|\boldsymbol{y}_i,S_i=k) &= \pi_k\,\mathrm{p}_k(y_k|y_1,\dots,y_{k-1})\,\mathrm{p}_k(y_{k-1}|y_1,\dots,y_{k-2})\cdots\mathrm{p}_k(y_1), \\ &= \pi_k\,\mathrm{p}_{\geq k}(y_k|y_1,\dots,y_{k-1})\,\mathrm{p}_{\geq k-1}(y_{k-1}|y_1,\dots,y_{k-2})\cdots\mathrm{p}_k(y_1), \end{aligned} \tag{7}$$

where $\boldsymbol{y}_i = (y_1,\dots,y_k)$.

We use an adaptive gradient descent algorithm to compute the maximum likelihood estimates (Riedmiller and Braun, 1993). Denote $J(\boldsymbol{\xi}) = -\log L = -\log \sum_{i=1}^{n} L_i$. Then maximizing the likelihood is equivalent to minimize the target function $J(\boldsymbol{\xi})$. Under an MAR assumption, we fix $\boldsymbol{\xi}_s = \mathbf{0}$, while under MNAR assumption, $\boldsymbol{\xi}_s$ can be chosen as desired.

During each step of the algorithm, Δ_{ij} has to be calculated for each subject and at each time, as well as partial derivatives for each parameter.

As an example of the speed of the algorithm, for 500 bivariate outcomes and 4 covariates, it takes about 11 seconds for 70 iterations to get convergence using R version 2.15.3 (2013-03-01) (R Core Team, 2013) and platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit). Main parts of the algorithm are coded in Fortran such as calculation of numerical derivatives and log-likelihood to quicken computation.

Further details about the maximization algorithm are presented in Appendix C.

2.3.3 Bayesian Framework

For a Bayesian inference, we specify priors on the parameters $\boldsymbol{\xi}$ and use a block Gibbs sampling method to draw samples from the posterior distribution. Denote all the parameters (including sensitivity parameters) to sample as :

$$\boldsymbol{\xi}_m = \left(\gamma_1, \gamma_2, \ldots, \gamma_J, \boldsymbol{\beta}_{y,j-1}^{(\geq j)}, \boldsymbol{\alpha}_j^{(\geq j)} \text{ for } j = 1, \ldots, J\right), \boldsymbol{\xi}_s = \left(\boldsymbol{h}_j^{(k)}, \boldsymbol{\eta}_{j-1}^{(k)}, \boldsymbol{\delta}_j^{(k)} \text{ for } k = 1, \ldots, j; 2 \leq j \leq J\right).$$

Comma separated parameters are marked to sample as a block. Updates of $\boldsymbol{\xi}_m$ require Metropolis-Hasting algorithm, while $\boldsymbol{\xi}_s$ samples are drawn directly from priors as desired for missingness mechanism assumptions.

As mentioned in section 2.2, MAR or MNAR assumptions are implemented via specific priors. For example, if MAR is assumed with no uncertainty, then $\boldsymbol{\xi}_s = \mathbf{0}$ with probability 1. Details for updating parameters are:

- γ_1: Use Metropolis-Hasting algorithm.

  1. Draw (γ_1^c) candidates from candidate distribution;
  2. Based on the new candidate parameter $\boldsymbol{\xi}^c$, calculate candidate Δ_{i1}^c for each subject i as we described in section 2.3.1. If $S > 1$ for subject i, update candidate $\Delta_{ij}^c, j \geq 2$ as well since $\Delta_{ij}, j \geq 2$ depends on Δ_{i1}. (For $S = 1$, we only need to update Δ_{i1}^c);
  3. Plug in Δ_{i1}^c or $(\Delta_{i1}^c, \Delta_{ij}^c, j \geq 2)$ in likelihood (7) to get candidate likelihood;
  4. Compute Metropolis-Hasting ratio, and accept the candidate value or keep the previous value.

- For the rest of the identifiable parameters, algorithms for updating the samples are all similar to γ_j.

- For sensitivity parameters, because we do not get any information from the data, we sample them from priors, which are specified from the missingness mechanism assumptions.

3 Simulation Study

In this section, we compared the performance of our proposed model in section 2.1 with the *rq* function in *quantreg* R package (Koenker, 2012). The *rq* function minimizes the loss (check) function $\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$ in terms of $\boldsymbol{\beta}$, where the loss function $\rho_\tau(u) = u(\tau - I(u < 0))$ and does not make any distributional assumptions.

We considered two scenarios corresponding to MAR and MNAR assumptions. For each scenario, we simulated 1000 data sets. For each set there are 200 bivariate observations $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$ for $i = 1, \ldots, 200$. Y_{i1} were always observed, while some of Y_{i2} were missing. A single covariate x was sampled from Uniform(0,2). The two models for the full data response $\boldsymbol{Y}_i$ were:

1. $Y_{i1}|R = 1 \sim N(2 + x, 1 + 0.5x)$, $Y_{i2}|R = 1, y_{i1} \sim N(1 - x - 1/2y_{i1}, 1)$, $Y_{i1}|R = 0 \sim N(-2 - x, 1 + 0.5x)$, $Y_{i2}|R = 0, y_{i1} \sim N(1 - x - 1/2y_{i1}, 1)$;

2. $(Y_{i1}, Y_{i2})|R = 1 \sim \mathrm{N}\big((1 + x, 1 - x), (\sigma_1 = 1, \rho = 0, \sigma_2 = 1)\big)$, $Y_{i1}|R = 0 \sim \mathrm{N}(-1 - x, 1)$, $Y_{i2}|R = 1 \sim \mathrm{N}(3 - x, 1)$,

where $\mathrm{N}\big((\mu_1, \mu_2), (\sigma_1, \rho, \sigma_2)\big)$ stands for bivariate normal distribution with two marginal normal distribution $\mathrm{N}(\mu_1, \sigma_1)$, $\mathrm{N}(\mu_2, \sigma_2)$ and correlation ρ.

For all cases, $\Pr(R = 1) = 0.5$. When $R = 0$, Y_{i2} is not observed, so $\mathrm{p}(Y_{i2}|R = 0, y_{i1})$ is not identifiable from observed data. The missingness mechanism is MAR in case 1 and MNAR in case 2.

Under MAR assumption, the sensitivity parameter $\boldsymbol{\xi}_s$ is fixed at $(0, 0, 0, 0, 0)$ as discussed in section 2.2. For *rq* function from *quantreg* R package, because only $Y_{i2}|R = 1$ is observed, the quantile regression for Y_{i2} can only be fit from the information of $Y_{i2}|R = 1$ vs x.

Under MNAR scenario, we fixed $\boldsymbol{\xi}_s$ at the true value $(2, 0, 0, 0, 0)$, assuming there was an intercept shift between distribution of $Y_{i2}|Y_{i1}, R = 1$ and $Y_{i2}|Y_{i1}$, $R = 0$.

For each dataset, we fit quantile regression for quantiles $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$.

Parameter estimates were evaluated by mean squared error (MSE),

$$\mathrm{MSE}(\gamma_{ij}) = \frac{1}{1000} \sum_{k=1}^{1000} \left(\hat{\gamma}_{ij}^{(k)} - \gamma_{ij} \right)^2,$$

where γ_{ij} is the true value for quantile regression coefficient, $\hat{\gamma}_{ij}^{(k)}$ is the maximum likelihood estimates in k-th simulated dataset ($(\gamma_{01}, \gamma_{11})$ for Y_{i1}, $(\gamma_{02}, \gamma_{12})$ for Y_{i2}).

Simulation results show estimates from our algorithm are closer to the true value for all quantiles from 0.1 to 0.9. Table 1 and 2 present the MSE for coefficients estimates of quantile 0.1, 0.3, 0.5, 0.7, 0.9 under MAR and MNAR assumptions. They show that our proposed method has smaller MSE than *rq* function in all cases. When data are missing at random, our method provides larger gains over *rq* method, because *quantreg* does not consider the missingness mechanism. The difference in MSE becomes larger for the upper quantiles because $Y_2|R = 0$ tends to be larger than $Y_2|R = 1$; therefore, the *rq* method using only the observed Y_2 yields larger bias for upper quantiles. For the same reason, under the MNAR assumption, 'quantreg' method led to much larger MSEs than our proposed method.

Table 1: Simulation result: MSE for coefficients estimates of quantiles 0.1, 0.3, 0.5, 0.7, 0.9 under MAR assumptions. $(\gamma_{01}, \gamma_{11})$ are quantile regression coefficients for Y_{i1}, and $(\gamma_{02}, \gamma_{12})$ are ones for Y_{i2}. MM stands for our proposed method, and RQ stands for the 'rq' function in R package 'quantreg'.

	MAR									
	0.1		0.3		0.5		0.7		0.9	
	MM	RQ	MM	RQ	MM	RQ	MM	RQ	MM	RQ
γ_{01}	0.09	0.15	0.12	0.19	0.11	1.08	0.16	0.19	0.10	0.15
γ_{11}	0.09	0.15	0.07	0.19	0.14	1.19	0.08	0.20	0.10	0.15
γ_{02}	0.08	0.27	0.07	0.59	0.06	1.08	0.12	1.75	0.24	2.92
γ_{12}	0.06	0.17	0.05	0.13	0.06	0.33	0.07	0.75	0.09	0.96

Table 2: Simulation result: MSE for coefficients estimates of quantiles 0.1, 0.3, 0.5, 0.7, 0.9 under MNAR scenario. $(\gamma_{01}, \gamma_{11})$ are quantile regression coefficients for Y_{i1}, and $(\gamma_{02}, \gamma_{12})$ are ones for Y_{i2}. MM stands for our proposed method, and RQ stands for the 'rq' function in R package 'quantreg'.

	MNAR									
	0.1		0.3		0.5		0.7		0.9	
	MM	RQ	MM	RQ	MM	RQ	MM	RQ	MM	RQ
γ_{01}	0.04	0.09	0.04	0.10	0.03	0.24	0.04	0.10	0.04	0.10
γ_{11}	0.03	0.07	0.02	0.08	0.64	0.74	0.03	0.08	0.03	0.07
γ_{02}	0.04	0.30	0.05	0.52	0.07	1.06	0.05	1.79	0.05	2.59
γ_{12}	0.03	0.09	0.03	0.05	0.03	0.05	0.03	0.05	0.03	0.09

4 Real Data Analysis

We apply our quantile regression approach to data from TOURS, a weight management clinical trial (Perri et al., 2008). This trial was designed to test whether a lifestyle modification program could effectively help people to manage their weights in long term. After finishing the six-month program, participants were randomly assigned to three treatments groups: face-to-face counseling, telephone counseling and control group. Their weights were recorded at baseline (Y_0), 6 months (Y_1) and 18 months (Y_2) after the trial. Here, we are interested in how the distribution of weights at six months and eighteenth months change with covariates. The regressors of interest include AGE (50-75), RACE (black and white) and weights at baseline (Y_0). Weights at the six months (Y_1) were always observed and 13 out of 224 observations (6%) were missing at 18 months (Y_2). All weights were scaled by 1/100 for computation stability.

We fitted regression models for bivariate responses $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$ for quantiles (5%, 30%, 50%, 70%, 95%). We fit 1,000 bootstraps to obtain the 95% confidence intervals.

Estimates are presented in Table 3. For weights of participants at six months, weights of white people are generally lower than ones of black people, because for all quantiles, the coefficients of race are negative. For the extreme quantiles (5% and 95%), the difference is small and not significant. However, when comparing the weights for quantiles 30%, 50% and 70%, white people tend to weight significantly less than black people. The differences are 8kg, 7kg, and 8kg separately. Meanwhile, weights of participants show obvious heterogeneity by age. Participants tend to have less weight when older. The trend is for all quantiles, but the effect is again small and not significant.

Table 3: Estimated marginal quantile regression coefficients with 95% confidence interval for weight of participants at 6 and 18 months. Weight measurement is scaled by 1/100.

	Intercept	Age(Centered)	White
Weight at 6 months			
5%	0.65 (0.34, 0.73)	-0.03 (-0.10, 0.06)	-0.02 (-0.48, 0.18)
30%	0.86 (0.79, 0.90)	-0.02 (-0.03, 0.00)	-0.08 (-0.12, -0.01)
50%	0.92 (0.87, 0.97)	-0.01 (-0.03, 0.01)	-0.07 (-0.12, -0.02)
70%	1.00 (0.94, 1.05)	-0.01 (-0.04, 0.02)	-0.08 (-0.13, -0.01)
95%	1.12 (1.08, 1.26)	-0.01 (-0.06, 0.05)	-0.03 (-0.14, 0.12)
Weight at 18 months			
5%	0.10 (-0.01, 0.64)	-0.00 (-0.05, 0.01)	-0.02 (-0.19, 0.22)
30%	0.84 (0.77, 0.88)	-0.02 (-0.04, -0.00)	-0.06 (-0.11, 0.01)
50%	0.92 (0.87, 0.98)	-0.02 (-0.04, 0.01)	-0.06 (-0.11, -0.00)
70%	1.02 (0.94, 1.06)	-0.01 (-0.04, 0.02)	-0.06 (-0.12, 0.01)
95%	1.15 (1.09, 1.25)	-0.01 (-0.05, 0.04)	-0.02 (-0.13, 0.05)

For weights at 18 months after baseline, we have similar conclusions. White people still have less weight than black people for all quantiles, but the magnitude is smaller than that at 6th month. The differences are 6kg, 6kg and 6kg instead of 8kg, 7kg and 8kg for quantiles 30%, 50% and 70%. It still shows a decreasing trend of weights over ages. However, none of them shows significance for the weight difference.

5 Discussion

In this paper, we have developed a marginal quantile regression model for data with monotone missingness. We use a pattern mixture model to jointly model the full data response and missingness. Here marginal quantile regression coefficients are of interest instead of coefficients conditional on random effects as in Yuan and Yin (2010). In addition, our approach allows non-parallel quantile lines over different quantiles via the mixture distribution and heterogeneity of variance.

Our method allows the missingness to be MNAR. We illustrated how to put informative priors for Bayesian inference and how to find sensitivity parameters to allow different missing data mechanisms. The recursive integration algorithm simplifies computation and can be easily implemented even in high dimension. Simulation study demonstrates that our approach has smaller MSE than the traditional frequentist method and it allows for MAR and MNAR missingness.

Our model assumes a multivariate normal distribution for each component in the pattern mixture model, which might be too restrictive. It is possible to replace that with a semi-parametric model, for example, the Dirichlet process mixture or Pólya tree. It would also be interesting to allow mixture probabilities depend on covariates. Our future work will also include development of a goodness of fit test to the model fit.

6 Acknowledgments

References

Matteo Bottai and Huiling Zhen. Multiple Imputation Based on Conditional Quantile Estimation. *Epidemiology, Biostatistics and Public Health*, 10(1), 2013. ISSN 2282-0930. doi: 10.2427/8758. URL http://ebph.it/article/view/8758.

Moshe Buchinsky. Changes in the US Wage Structure 1963-1987: Application of Quantile Regression. *Econometrica*, 62(2):pp. 405–458, 1994. ISSN 00129682. URL http://www.jstor.org/stable/2951618.

Moshe Buchinsky. Recent advances in quantile regression models: A practical guideline for empirical research. *The Journal of Human Resources*, 33(1):pp. 88–126, 1998. ISSN 0022166X. URL http://www.jstor.org/stable/146316.

Michael J Daniels and Joseph W Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*, volume 109. Chapman and Hall/CRC, 2008.

T. Hanson and W.O. Johnson. Modeling regression error with a mixture of polya trees. *Journal of the American Statistical Association*, 97(460):1020–1033, 2002.

X. He, P. Ng, and S. Portnoy. Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):537–550, 1998. ISSN 1467-9868. doi: 10.1111/1467-9868.00138. URL http://dx.doi.org/10.1111/1467-9868.00138.

Patrick J Heagerty. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698, 1999.

N.L. Hjort and S. Petrone. Nonparametric quantile inference using dirichlet processes. *Advances in statistical modeling and inference*, pages 463–492, 2007.

N.L. Hjort and S.G. Walker. Quantile pyramids for bayesian nonparametrics. *The Annals of Statistics*, 37(1):105–131, 2009.

R. Koenker. *Quantile regression*, volume 38. Cambridge Univ Pr, 2005.

Roger Koenker. *quantreg: Quantile Regression*, 2012. URL http://CRAN.R-project.org/package=quantreg. R package version 4.91.

Roger Koenker and Jr. Bassett, Gilbert. Regression quantiles. *Econometrica*, 46(1):pp. 33–50, 1978. ISSN 00129682. URL http://www.jstor.org/stable/1913643.

Roger Koenker and Jose A. F. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):pp. 1296–1310, 1999. ISSN 01621459. URL http://www.jstor.org/stable/2669943.

A. Kottas and A.E. Gelfand. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96(456):1458–1468, 2001.

A. Kottas and M. Krnjajić. Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36(2):297–319, 2009.

Geert Molenberghs, Bart Michiels, MG Kenward, and Peter J Diggle. Monotone missing data and pattern-mixture models. *Statistica Neerlandica*, 52(2):153–161, 1998.

Michael G Perri, Marian C Limacher, Patricia E Durning, David M Janicke, Lesley D Lutes, Linda B Bobroff, Martha Sue Dale, Michael J Daniels, Tiffany A Radcliff, and A Daniel Martin. Extended-care programs for weight management in rural communities: the treatment of obesity in underserved rural settings (tours) randomized trial. *Archives of internal medicine*, 168(21):2347, 2008.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

B.J. Reich, H.D. Bondell, and H.J. Wang. Flexible bayesian quantile regression for independent and clustered data. *Biostatistics*, 11(2):337–352, 2010.

Martin Riedmiller and Heinrich Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Neural Networks, 1993., IEEE International Conference on*, pages 586–591. IEEE, 1993.

Jason Roy and Michael J Daniels. A general class of pattern mixture models for nonignorable dropout with many possible dropout times. *Biometrics*, 64(2):538–545, 2008.

Stephen Walker and Bani K. Mallick. A bayesian semiparametric accelerated failure time model. *Biometrics*, 55(2):477–483, 1999. ISSN 1541-0420. doi: 10.1111/j.0006-341X.1999.00477.x. URL http://dx.doi.org/10.1111/j.0006-341X.1999.00477.x.

Ying Wei, Anneli Pere, Roger Koenker, and Xuming He. Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25(8):1369–1382, 2006. ISSN 1097-0258. doi: 10.1002/sim.2271. URL http://dx.doi.org/10.1002/sim.2271.

Ying Wei, Yanyuan Ma, and Raymond J Carroll. Multiple imputation in quantile regression. *Biometrika*, 99(2):423–438, 2012.

Keming Yu and Rana A Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.

Keming Yu, Zudi Lu, and Julian Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350, 2003. ISSN 1467-9884. doi: 10.1111/1467-9884.00363. URL http://dx.doi.org/10.1111/1467-9884.00363.

Ying Yuan and Guosheng Yin. Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, 66(1):105–114, 2010.

A Identifiability

First suppose y is univariate and there are two patterns $R = 1$ and $R = 0$.

Before going forward to quantile regression, first we consider identifiability problem in mean regression.

Consider a pattern mixture model:

$$\begin{aligned} Y|R=1 &\sim N(\Delta+\mu_1,\sigma_1), \\ Y|R=0 &\sim N(\Delta+\mu_0,\sigma_0), \\ \Pr(R=1) &= \pi, \\ E(y) &= \theta. \end{aligned} \tag{8}$$

Thus by iterated expectation, we have

$$\begin{aligned} \theta &= \Delta+\mu_1\pi+\mu_0(1-\pi), \\ \Delta &= \theta-\pi\mu_1-(1-\pi)\mu_0. \end{aligned}$$

We can see Δ is determined by θ, μ_1, μ_0. Plugging in (8), we have

$$\begin{aligned} Y|R=1 &\sim N(\theta+(1-\pi)\mu_1-(1-\pi)\mu_0,\sigma_1), \\ Y|R=0 &\sim N(\theta-\pi\mu_1+\pi\mu_0,\sigma_0). \end{aligned}$$

Denote $\xi_1 = (\theta, \mu_1, \mu_0)$, and if $\xi_2 = (\theta, \mu_1 + c, \mu_0 + c)$, both groups of parameters lead to the same distribution of $\mathrm{p}(y, R) = \mathrm{p}(y|R)\,\mathrm{p}(R)$. Therefore, ξ is not identifiable. If we put constraints on μ_1 and μ_0, for example $\mu_0 = 0$, then

$$\begin{aligned} Y|R=1 &\sim N(\theta+\mu_1,\sigma_1), \\ Y|R=0 &\sim N(\theta,\sigma_0). \end{aligned}$$

Thus $\boldsymbol{\xi} = (\theta, \mu_1)$ is identifiable. If $\xi_2 \neq \xi_1$, then $\mathrm{p}_2(y, R) \neq \mathrm{p}_1(y, R)$.

Secondly, we consider quantile regression for a pattern mixture model:

$$\begin{aligned} Y|R=1 &\sim N(\Delta+\mu_1,\sigma_1), \\ Y|R=0 &\sim N(\Delta+\mu_0,\sigma_0), \\ \Pr(R=1) &= \pi, \\ \mathrm{p}(Y\le\theta) &= \tau, \end{aligned}$$

where θ is the quantile estimate of interest. We again show $\boldsymbol{\xi} = (\theta, \mu_1, \mu_0)$ is not identifiable.

Again by iterated expectations, we have

$$\tau = \pi\Phi\left(\frac{\theta - \Delta - \mu_1}{\sigma_1}\right) + (1-\pi)\Phi\left(\frac{\theta - \Delta - \mu_0}{\sigma_0}\right).$$

Thus Δ is again determined by the other parameters:

$$\Delta = h(\theta, \mu_1, \mu_0, \sigma_1, \sigma_0, \pi, \tau).$$

To show $\boldsymbol{\xi} = (\theta, \mu_1, \mu_0, \sigma_1, \sigma_0)$ is not identifiable, we need to find $\boldsymbol{\xi}^{'} \neq \boldsymbol{\xi}$, such that $\mathrm{p}(y|R) = \mathrm{p}^{'}(y|R)$. If the last equation holds, then we must have $\sigma_1^{'} = \sigma_1, \sigma_0^{'} = \sigma_0$, thus we still need to find $\theta^{'}, \mu_1^{'}, \mu_0^{'}$ such that

$$\begin{aligned} h(\boldsymbol{\xi}) + \mu_1 &= h(\boldsymbol{\xi}^{'}) + \mu_1^{'}, \\ h(\boldsymbol{\xi}) + \mu_0 &= h(\boldsymbol{\xi}^{'}) + \mu_0^{'}. \end{aligned}$$

By substracting previous equations, we have $\mu_1^{'} - \mu_0^{'} = \mu_1 - \mu_0$. Denote $\mu_1^{'} = \mu_1 + \delta$ and $\mu_0^{'} = \mu_0 + \delta$, and let $\theta^{'} = \theta$ such that

$$\Delta^{'} = h(\theta^{'}, \mu_1, \mu_0, \sigma_1, \sigma_0, \delta) = h(\boldsymbol{\xi}) - \delta = \Delta - \delta.$$

Then the new parameter $\boldsymbol{\xi}^{'}$ yields the same distribution as $\boldsymbol{\xi}$. Therefore $\boldsymbol{\xi}$ is not identifiable.

If we use a constraint, for example $\mu_1 = -\mu_0$, then $\mathrm{p}(y|R;\boldsymbol{\xi}) = \mathrm{p}(y|R;\boldsymbol{\xi}^{'})$ yields $\boldsymbol{\xi} = \boldsymbol{\xi}^{'}$.

Now consider the situation with covariates. Suppose the model is

$$\begin{aligned} Y|R = 1, x &\sim N(\Delta + \mu_1 + \beta_{x1}x, \sigma_1), \\ Y|R = 0, x &\sim N(\Delta - \mu_1 + \beta_{x0}x, \sigma_0), \\ \Pr(R = 1) &= \pi, \\ \mathrm{p}(Y \leq \gamma_0 + \gamma_1 x) &= \tau. \end{aligned}$$

Δ can still be determined by

$$\Delta = h(x, \gamma_0, \gamma_1, \mu_1, \beta_{x1}, \beta_{x0}, \sigma_1, \sigma_0, \pi, \tau).$$

We want to show the parameter $\boldsymbol{\xi} = (\gamma_0, \gamma_1, \mu_1, \beta_{x1}, \beta_{x0}, \sigma_1, \sigma_0, \pi)$ is not identifiable by finding $\boldsymbol{\xi}^{'} \neq \boldsymbol{\xi}$, but $\mathrm{p}(y|R;\boldsymbol{\xi}) = \mathrm{p}(y|R;\boldsymbol{\xi}^{'})$. If the last equation holds, we have $\sigma_1^{'} = \sigma_1, \sigma_0^{'} = \sigma_0$, and to equate the two means, we have

$$\begin{aligned} \Delta + \mu_1 + \beta_{x1}x &= \Delta^{'} + \mu_1^{'} + \beta_{x1}^{'}x, \\ \Delta - \mu_1 + \beta_{x0}x &= \Delta^{'} - \mu_1^{'} + \beta_{x0}^{'}x. \end{aligned}$$

By substracting the two equations, we have

$$2\mu_1 + (\beta_{x1} - \beta_{x0})x = 2\mu_1^{'} + (\beta_{x1}^{'} - \beta_{x0}^{'})x,$$

which holds for all x. Thus $\mu_1 = \mu_1^{'}$ and $(\beta_{x1} - \beta_{x0}) = (\beta_{x1}^{'} - \beta_{x0}^{'})$. Then let

$$\beta_{x1}^{'} = \beta_{x1} + \delta,$$
$$\beta_{x0}^{'} = \beta_{x0} + \delta,$$

and keep all the other parameters in $\boldsymbol{\xi}^{'}$ the same. We can still have the same distribution of $y|R;\boldsymbol{\xi}$ but with different $\boldsymbol{\xi}$. Therefore, $\boldsymbol{\xi}$ is not identifiable One solution is to restrict $\beta_{x1} = -\beta_{x0}$ or $\beta_{x1} = 0$.

Now consider the bivariate (y_1, y_2) case, and we focus on the identifiability issue especially $y_2|y_1$. Suppose the model is

$$Y_2|y_1, x, R = 1 \sim N(\Delta + \mu_1 + x\beta_{x1} + \beta_{11}y_1, \sigma_1),$$
$$Y_2|y_1, x, R = 0 \sim N(\Delta - \mu_1 - x\beta_{x1} + \beta_{10}y_1, \sigma_0).$$

Here R stands for two different patterns, and missingness is not considered.

Regarding the identifiability of β_{11} and β_{10}, assume there exists $\beta_{11}^{'}$ and $\beta_{10}^{'}$, such that

$$\Delta + \mu_1 + x\beta_x + \beta_{11}y_1 = \Delta^{'} + \mu_1^{'} + x\beta_x^{'} + \beta_{11}^{'}y_1,$$
$$\Delta - \mu_1 - x\beta_x + \beta_{10}y_1 = \Delta^{'} - \mu_1^{'} - x\beta_x^{'} + \beta_{10}^{'}y_1.$$

By substracting two equations, we have $\mu_1 = \mu_1^{'}$ and $\beta_x = \beta_x^{'}$. Since Δ is determined by integrating out y_1, such that matching the two sides of the above equation for coefficient of y_1, we must have $\beta_{11} = \beta_{11}^{'}$ and $\beta_{10} = \beta_{10}^{'}$, therefore, $\boldsymbol{\xi}$ is identifiable.

For identifiability issue with the heterogeneous model described in section 2.1, it is easy to show there is no trouble with the heterogeneity parameters α, analogous to the linear model case. For the other parameters, it can be found similar to the above development.

B Proof of Lemma 2.1

- Denote

  $$I(a, b) = \int \Phi\left(\frac{x - b}{a}\right) \phi(x) dx,$$

  where Φ is the standard normal cdf and ϕ is the standard normal pdf and $a > 0$.

  $$\begin{aligned}\frac{\partial I(a,b)}{\partial b} &= -\frac{1}{a} \int \phi\left(\frac{x - b}{a}\right) \phi(x) dx \\ &= -\frac{1}{\sqrt{2\pi}\sqrt{a^2 + 1}} \exp\left(-\frac{b^2}{2(a^2 + 1)}\right) \\ &= -\frac{1}{\sqrt{a^2 + 1}} \phi\left(\frac{b}{\sqrt{a^2 + 1}}\right).\end{aligned}$$

Since $I(a, \infty) = 0$,

$$\begin{aligned} I(a,b) &= -\frac{1}{\sqrt{a^2+1}} \int_b^{\infty} \phi \left(\frac{s}{\sqrt{a^2+1}} \right) ds \\ &= \int_{b/\sqrt{a^2+1}}^{\infty} \phi(t) dt \\ &= 1 - \Phi(b/\sqrt{a^2+1}). \end{aligned} \tag{9}$$

For $a < 0$,

$$\begin{aligned} \frac{\partial I(a,b)}{\partial b} &= -\frac{1}{a} \int \phi \left(\frac{x-b}{a} \right) \phi(x) dx \\ &= -\frac{sgn(a)}{\sqrt{2\pi}\sqrt{a^2+1}} \exp \left(-\frac{b^2}{2(a^2+1)} \right) \\ &= -\frac{sgn(a)}{\sqrt{a^2+1}} \phi \left(\frac{b}{\sqrt{a^2+1}} \right). \end{aligned}$$

Since $I(a, -\infty) = 0$:

$$\begin{aligned} I(a,b) &= \int_{-\infty}^{b/\sqrt{a^2+1}} \phi(t) dt \\ &= \Phi(b/\sqrt{a^2+1}). \end{aligned} \tag{10}$$

- For integrating over a normal distribution with mean μ and standard deviation σ:

$$\begin{aligned} \int \Phi(x) d\Phi(x; \mu, \sigma) &= \int \Phi(x) \frac{1}{\sigma} \phi \left(\frac{x-\mu}{\sigma} \right) dx \\ &= \int \Phi(\sigma t + \mu) \phi(t) dt \\ &= 1 - \Phi(-\mu/\sigma/\sqrt{1/\sigma^2+1}). \end{aligned}$$

The last equation holds by (9)

- For integrating a $\mathrm{N}(b, a)$ CDF over another normal distribution ($\mathrm{N}(\mu, \sigma)$):

$$\begin{aligned} \int \Phi \left(\frac{x-b}{a} \right) d\Phi(x; \mu, \sigma) &= \int \Phi \left(\frac{x-b}{a} \right) \frac{1}{\sigma} \phi \left(\frac{x-\mu}{\sigma} \right) dx \\ &= \int \Phi \left(\frac{\sigma y + \mu - b}{a} \right) \phi(y) dy \\ &= 1 - \Phi \left(\frac{b-\mu}{\sigma} / \sqrt{\frac{a^2}{\sigma^2} + 1} \right). \end{aligned} \tag{11}$$

If $a < 0$,

$$\int \Phi \left(\frac{x-b}{a} \right) d\Phi(x; \mu, \sigma) = \Phi \left(\frac{b-\mu}{\sigma} / \sqrt{\frac{a^2}{\sigma^2} + 1} \right). \tag{12}$$

C Maximum Likelihood Estimation Using Adaptive Gradient Descent Algorithm

See section 2.3.2 for notation.

The likelihood can be maximized via the following algorithm:

1. initialize $\boldsymbol{\xi}$
2. calculate $\partial J(\boldsymbol{\xi})/\partial \xi_j$ for all j,
3. update $\boldsymbol{\xi}$ by adaptive gradient descent algorithm described in (Riedmiller and Braun, 1993) for all j
4. evaluate new $J(\boldsymbol{\xi})$
5. if the amount of descent of $J(\boldsymbol{\xi})$ is greater than a certain cutoff, then go back to step 2 and repeat. Otherwise, algorithm converges.

We can use numerical approximation to calculate $\partial J(\boldsymbol{\xi})/\partial \xi_j$ in step 2. For example, for $j = 1$,

$$\frac{\partial J(\boldsymbol{\xi})}{\partial \xi_1} \approx \frac{J(\xi_1 + \epsilon, \xi_2, \ldots) - J(\xi_1 - \epsilon, \xi_2, \ldots)}{2\epsilon}.$$