Quantile Regression in the Presence of Monotone Missingness

May 25, 2013

Abstract

1 Introduction

Quantile regression is a simple way to study the relationship between response and covariates when one (or several) quantiles are of interest as compared to mean regression. The dependence between upper or lower quantiles of the response variable and the covariates typically vary differentially relative to that of the mean. This is often of interest in econometrics, educational studies, biomedical studies, and environment studies ((Yu and Moyeed, 2001), (Buchinsky, 1994), (Buchinsky, 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).

Unlike mean regression, quantile regression is more robust to outliers and provides more information about how covariates affect quantiles. For example, as a summary statistic of data, median is more robust than the mean in the presence of outliers. In addition, mean regression only focus on the change of covariates on the mean, while quantile regression can offer 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.

Bayesian approaches offer exact inference. 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. Other than parametric Bayesian approaches, some 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)).

However, above methods focus on complete data without missingness. There are 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. Therefore, their model fully use the whole dataset and have more efficiency. However, they put more focus on the missing covariates rather than missing outcomes, which is of more interested. 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 below: first, we introduce a quantile regression methods to deal with monotone nonignorable dropout in general case in section 2, including sensitivity analysis and computational details. We use simulation studies to demonstrate the performance of our algorithm in section 3. We apply our approach on 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 notations on monotone dropout, then describe our proposed quantile regression model in section 2.1. We provide more details on MAR and MNAR and computation in sections 2.2 and 2.3.

Under monotone dropout, without loss of generality, denote $S_i \in \{1, 2, \ldots, J\}$ to be the follow up time, and $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})^T$ 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 $\boldsymbol{\gamma}_j = (\gamma_{j0}, \gamma_{j2}, \ldots, \gamma_{jp})^T$,

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

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

Let

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

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 deal with missingness. Without loss of clarity, we suppress the subscript i for subject i. Specify the conditional distribution as:

$$\begin{aligned}
&\mathrm{p}_k(y_1) = \mathrm{N}(\Delta_1 + \boldsymbol{x}_1^T \boldsymbol{\beta}_1^{(k)}, \exp(\boldsymbol{x}_1^T \boldsymbol{\alpha}_1^{(k)})), k = 1, \ldots, J, \\
&\mathrm{p}_k(y_j|y_1, \ldots, y_{j-1}) = \begin{cases} \mathrm{N}(\Delta_j + \boldsymbol{x}_{j*}^T \boldsymbol{\beta}_{j*}^{(k)}, \exp(\boldsymbol{x}_j^T \boldsymbol{\alpha}_j^{(k)})), & k < j; \\ \mathrm{N}(\Delta_j + \boldsymbol{x}_{j*}^T \boldsymbol{\beta}_{j*}^{(\geq j)}, \exp(\boldsymbol{x}_j^T \boldsymbol{\alpha}_j^{(\geq j)})), & k \geq j; \end{cases}, \text{ for } 2 \leq j \leq J, \\
&\mathrm{Pr}(S = k) = \pi_k, \\
&\sum_{k=1}^{J} \pi_k = 1,
\end{aligned} \tag{2}$$

where π_k does not depend on covariates, $\boldsymbol{x}_{j*} = (\boldsymbol{x}_j^T, y_1, \ldots, y_{j-1})^T$ is a $(p + j - 1) \times 1$ modified covariates vector, and $\boldsymbol{\alpha}_j^{(k)}$ is a $p \times 1$ vector controlling heterogeneity of response component j under follow up time $S = k$, $\boldsymbol{\beta}_{j*}^{(k)} = (\boldsymbol{\beta}_j^T, \beta_{y_1 j}, \ldots, \beta_{y_{j-1} j})^T$, where $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{pj})^T$ can be regarded as coefficients of interaction of pattern k and modified covariates analog to mean regression with length $(p + j - 1) \times 1$.

We model the heterogeneity parameters $\boldsymbol{\alpha}_j$ inside the exponential because there is no restriction on those heterogeneity parameters, therefore it is computationally more stable under both frequentist and Bayesian framework.

In (2) , Δ_j are functions of $\tau, \boldsymbol{x}_j$ and other parameters and are determined by

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

for $j = 1$ and

$$\begin{aligned}
\tau = \mathrm{Pr}(Y_j \leq \boldsymbol{x}_j^T \gamma_j) &= \sum_{k=1}^{J} \pi_k \, \mathrm{Pr}_k(Y_j \leq \boldsymbol{x}_j^T \gamma_j) \\
&= \sum_{k=1}^{J} \pi_k \int \cdots \int \mathrm{Pr}_k(Y_j \leq \boldsymbol{x}_j^T \gamma_j | y_1, \ldots, y_{j-1}) \, \mathrm{p}_k(y_{j-1}|y_1, \ldots, y_{j-2}) \\
&\quad \cdots \mathrm{p}_k(y_2|y_1) \, \mathrm{p}_k(y_1) dy_{j-1} \cdots dy_1.
\end{aligned} \tag{4}$$

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

The idea is to model the marginal quantile regression coefficients directly, then to involve them in the likelihood through restrictions in the mixture model, and finally to estimate them using likelihood methods. The mixture patterns and heterogeneity between subjects allow the marginal quantile regression coefficients to differ by quantiles . Otherwise, the quantile lines would be parallel to each other for different quantiles.

For identifiability, we need another set of restrictions,

$$\begin{aligned}
&\sum_{k=1}^{J} \beta_{l1}^{(k)} = 0, l = 1, \ldots, p, \\
&(\sum_{k=1}^{j-1} \beta_{lj}^{(k)}) + \beta_{lj}^{(\geq j)} = 0, l = 1, \ldots, p, 2 \leq j \leq J;
\end{aligned}$$

Further details on these restrictions can be found in appendix A.

2.2 Missing Data Mechanism and Sensitivity Analysis

In our mixture model (2), (Molenberghs et al., 1998) shows MAR holds 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 $2 \leq j \leq J, k < j$, Y_j is not observed, thus $\boldsymbol{\beta}_j^{(k)}, k = 1, \ldots, J$ and $\boldsymbol{\alpha}_j^{(k)}$, $\boldsymbol{\beta}_{y,j}^{(k)} = (\beta_{y_1,j}^{(k)}, \ldots, \beta_{y_{j-1},j}^{(k)}), k < j$ can not be identified from the observed data. Denote

$$\begin{aligned}
\boldsymbol{\alpha}_j^{(k)} &= \boldsymbol{\alpha}_j^{(\geq j)} + \boldsymbol{h}_j^{(k)}, \\
\boldsymbol{\beta}_{y,j}^{(k)} &= \boldsymbol{\beta}_{y,j}^{(\geq j)} + \boldsymbol{\eta}_j^{(k)},
\end{aligned}$$

where $\boldsymbol{h}_j^{(k)} = (h_{1j}^{(k)}, \ldots, h_{pj}^{(k)})$ and $\boldsymbol{\eta}_j^{(k)} = (\eta_{y_1,j}^{(k)}, \ldots, \eta_{y_{j-1},j}^{(k)})$ for $k < j$, then $\boldsymbol{\xi}_s = (\boldsymbol{\xi}_\beta, \boldsymbol{\xi}_\alpha)$ could be a group of sensitivity parameters, where $\boldsymbol{\xi}_\beta = (\boldsymbol{\beta}_j^{(k)}, \boldsymbol{\beta}_j^{(\geq j)}, \boldsymbol{\eta}_j^{(k)}), k < j, 2 \leq j \leq J$, and $\boldsymbol{\xi}_\alpha = (\boldsymbol{h}_j^{(k)}), k < j, 2 \leq j \leq J$.

- **Frequentist way:**

  When $\boldsymbol{\xi}_s = \boldsymbol{\xi}_{s0} = \mathbf{0}$, it yields Molenberghs condition (5), therefore MAR condition satisfies. If $\boldsymbol{\xi}_s$ is fixed at $\boldsymbol{\xi}_s \neq \boldsymbol{\xi}_{s0}$, then Molenberghs condition fails, thus the missing mechanism is missing not at random.

- **Bayesian Framework:**

  We put priors on $(\boldsymbol{\xi}_s, \boldsymbol{\xi}_m)$ ($\boldsymbol{\xi}_m$ are identifiable parameters) as :

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

  If we assume MAR with no uncertainty, the prior of $\boldsymbol{\xi}_s$ is $\mathrm{p}(\boldsymbol{\xi}_s = \mathbf{0}) = 1$. Sensitivity analysis can be executed through putting a set of priors on $\boldsymbol{\xi}_s$ to check 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 interface in section 2.3.3.

2.3.1 Calculation of Δ

From equation (3) and (4), Δ_{ij} depends on subject covariates $\boldsymbol{x}_i$, thus Δ_{ij} needs to be calculated for each subject generally. 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 another 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 two approaches:

1. **Assume first order relationship:** We assume

   $$\mathrm{p}(Y_j|S,x,Y_{j-1},\ldots,Y_1) = \mathrm{p}(Y_j|S,x,Y_{j-1}).$$

   After obtaining Δ_{j-1}, for each component in equation (4),

   $$\begin{aligned}\mathrm{p}(Y_j \leq \boldsymbol{x}^T\gamma_j|S=k) = &\int\cdots\int \mathrm{p}(Y_j \leq \boldsymbol{x}^T\gamma_j|S=k,x,Y_{j-1},\ldots,Y_1)\\ &dF(Y_{j-1}|S=k,Y_{j-2},\ldots,Y_1)\cdots dF(Y_2|S=s,Y_1),\\ = &\int \mathrm{p}(Y_j \leq \boldsymbol{x}^T\boldsymbol{\gamma}|S=s,x,Y_{j-1})dF(Y_{j-1}|S=k,Y_{j-2}).\end{aligned}$$

   Thus, only one integral is needed. Furthermore, by lemma 2.1 , we can evaluate the above integral analytically in terms of a normal CDF, without using any numerical approximations.

2. **Recursive Computation:**

   From equation (6), we can find after single integral, the kernel part is still a normal CDF, but with other coefficients. So recursive simplification can be applied. After recursively applying lemma 2.1 $j-1$ times, equation (4) becomes a closed form in terms of normal CDF analytically without calculating integral numerically, thus it can be solved again using standard numerical root-find method for Δ_{ij}.

Option 2 is exact and does not assume a restrictive first order relationship. However, it is computational more intensive than option 1.

2.3.2 Maximum Likelihood Estimation

The observed likelihood for $\boldsymbol{y}_i = (y_1,\ldots,y_k)$ with follow-up time $S=k$ is

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

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 to maximize likelihood is equivalent to minimize the target function $J(\boldsymbol{\xi})$. Under MAR assumption, we fix $\boldsymbol{\xi}_s = \mathbf{0}$, while under MNAR assumption, $\boldsymbol{\xi}_s$ can be chosen to assume there is an intercept shift between the conditional distributions of $Y_j|Y_1,\ldots,Y_{j-1},S$, or there is heterogeneity between those distributions.

During each step in adaptive gradient descent algorithm, Δ_{ij} has to be calculated for each subject and component, as well as partial derivatives for each parameter. Because it is computational expensive, we compile fortran within R for speed.

Details about the maximization algorithm are presented in the Appendix C.

2.3.3 Bayesian Framework

Under a Bayesian framework, we put priors on the parameters $\boldsymbol{\xi}$ and make exact inference.

We use a block Gibbs sampling method to draw samples from the posterior distribution. Denote all the parameters (including sensitivity parameters) to sample as :

$$\boldsymbol{\xi} = \left(\gamma_1, \gamma_2, \ldots, \gamma_J, \boldsymbol{\beta}_{j*}^{(k)}, \boldsymbol{\alpha}_j^{(k)} \text{ for } k = 1, \ldots, J, j = 1, \ldots, J\right).$$

Comma separated parameters are marked to sample as a block in block Gibbs sampling. All updates require Metropolis-Hasting algorithm.

As mentioned in section 2.2, MAR or MNAR assumptions are adopted 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 corresponding subject i, update candidate $\Delta_{ij}^c, j \geq 2$ as well because $\Delta_{ij}, j \geq 2$ depend 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. Obtain Metropolis-Hasting ratio, move the chain or keep the previous one.
- For the rest of the parameters, algorithms for updating the samples are all similar to γ_j.

Computation is still expensive due to need to calculate Δ and likelihood in each iteration. Compiled language is recommended to implement the algorithm.

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.

3.1 MAR

The simulation study included 1000 data sets. Each data set consists 200 bivariate observations $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$ for $i = 1, \ldots, 200$. Y_{i1} was always observed, while some of Y_{i2} were missing. Covariates x were sampled from uniform (0,2). We sampled $\boldsymbol{Y}_i$ from:

$$\begin{aligned} Y_{i1}|R = 1 &\sim N(2 + x, 1 + 0.5x) \\ Y_{i2}|R = 1, y_{i1} &\sim N(1 - x - 1/2y_{i1}, 1) \\ Y_{i1}|R = 0 &\sim N(-2 - x, 1 + 0.5x) \\ Y_{i2}|R = 0, y_{i1} &\sim N(1 - x - 1/2y_{i1}, 1) \\ \mathrm{p}(R = 1) &= 0.5. \end{aligned}$$

When $R = 0$, Y_{i2} is not observed, so $\mathrm{p}(Y_{i2}|R = 0, y_{i1})$ is not identifiable from observed data. Here we assume the distribution of $[Y_{i2}|R = 0, y_{i1}]$ is equal to $[Y_{i2}|R = 1, Y_{i1}]$ (MAR).

By integrating $Y_{i1}|R$ out of $Y_{i2}|R, y_{i1}$, we have

$$Y_{i2}|R = 1 \sim N(-3x/2, 5/4 + x/8),$$
$$Y_{i2}|R = 0 \sim N(2 - x/2, 5/4 + x/8).$$

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

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

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

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

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

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

Mean squared errors are shown in Table 1. Results show that our proposed method has smaller MSE than *rq* function in all cases. Since Y_{i2} 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 only using the observed Y_2 yields larger bias for upper quantile when estimating the marginal quantile estimates.

Table 2: Simulation result: MSE for coefficients estimates of quantiles 0.1, 0.3, 0.5, 0.7, 0.9 under MCAR 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'.

	MAR									
	0.1		0.3		0.5		0.7		0.9	
	MM	RQ	MM	RQ	MM	RQ	MM	RQ	MM	RQ
γ_{01}	0.05	0.09	0.04	0.10	0.03	0.24	0.04	0.10	0.05	0.10
γ_{11}	0.03	0.07	0.02	0.08	0.58	0.74	0.03	0.08	0.03	0.07
γ_{02}	0.04	0.12	0.05	0.07	0.04	0.06	0.05	0.07	0.05	0.11
γ_{12}	0.03	0.09	0.03	0.05	0.03	0.05	0.03	0.05	0.03	0.09

3.2 MCAR and MNAR

We also simulated datasets under MCAR and MNAR . For MCAR, We sampled $\boldsymbol{Y}_i$ from:

$$\begin{aligned}\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \Big| R = 1 \sim N\left(\begin{pmatrix} 1+x \\ 1-x \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right),\\ Y_{i1}|R = 0 \sim \mathrm{N}(-1-x, 1),\\ \mathrm{p}(R = 1) = 0.5.\end{aligned}$$

We conducted simulation study under two different situations: MCAR and MNAR. Under MNAR scenario, we fixed $\boldsymbol{\xi}_s$ at the true value $(1, 0, 0, 0)$, 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 in both scenario, we fit quantile regression for quantiles $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$.

Algorithms were evaluated by mean squared error (MSE) as we described above.

Simulation results show estimates from our algorithm are closer to the true value for all quantiles from 0.1 to 0.9. Table 2 and 3 list the MSE for coefficients estimates of quantile 0.1, 0.3, 0.5, 0.7, 0.9 under MAR and MNAR assumptions. Even for extreme quantiles ($\tau = 0.1$ and $\tau = 0.9$), our algorithm behave as good as for non-extreme quantile ($\tau = 0.3, 0.5, 0.7$) in terms of MSE. Furthermore, 'rq' function did not consider the missing mechanism, so under MNAR assumption, 'quantreg' method led to tremendous MSE and our proposed method were much closer to the true value.

4 Real Data Analysis

Here is the analysis for *tours* data. *Weight2* stands for weight at 6th month after the baseline measure, and *weight3* stands for the one at 18th month after the baseline. There were three treatments and two main races in this study (Treatment M, Treatment O and Treatment T; Race 1(black) and Race 3(white)). Weights at 6th month were always observed and some

Table 3: 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

weights at 18th month were missing (211 observed out of 224 , 94%). All weights are scaled by 1/100.

Figure 1 is the boxplot of weights vs treatments and races.

Quantile regression models were fitted for responses *weight2* and *weight3* together ($\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$) for quantiles (10%, 30%, 50%, 70%, 90%). Covariates are treatments and races, and we assume their effects are additive. Treatment M and Race 1 are baseline references. We fit 1,000 bootstraps to obtain the 95% confidence intervals.

Estimates are presented in table 4. For *weight2*, quantile estimates show there is no significant difference for three treatment group through all the quantiles, because all 95 % confidence intervals include 0. However, when comparing weights from two races, weights at 6th month from race 3 are generally lower than ones from race 1. Estimates of race 3 effect to weights quantiles 10% up to 70% are all significantly away from zero (negative). However, for top weights of two races (90% quantile), the difference is not significant.

For weights at 18th month (weight3), we have similar conclusions. Confidence intervals of treatment effects on weight3 for all quantiles (10% up to 90%) include zero. But after 18 months, weights of patients from race 3 are significantly lower than ones from race 1 only for lower quantiles (10% to 30%). They are not significantly different for quantiles (50% to 90%).

5 Discussion

In this paper, we have developed a marginal quantile regression model for data with monotone dropout missingness. We use a pattern mixture model to explain the missing mechanism. 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 modeling 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 allow different missing data mechanisms . The recursive integration algorithm simplifies computation and can be

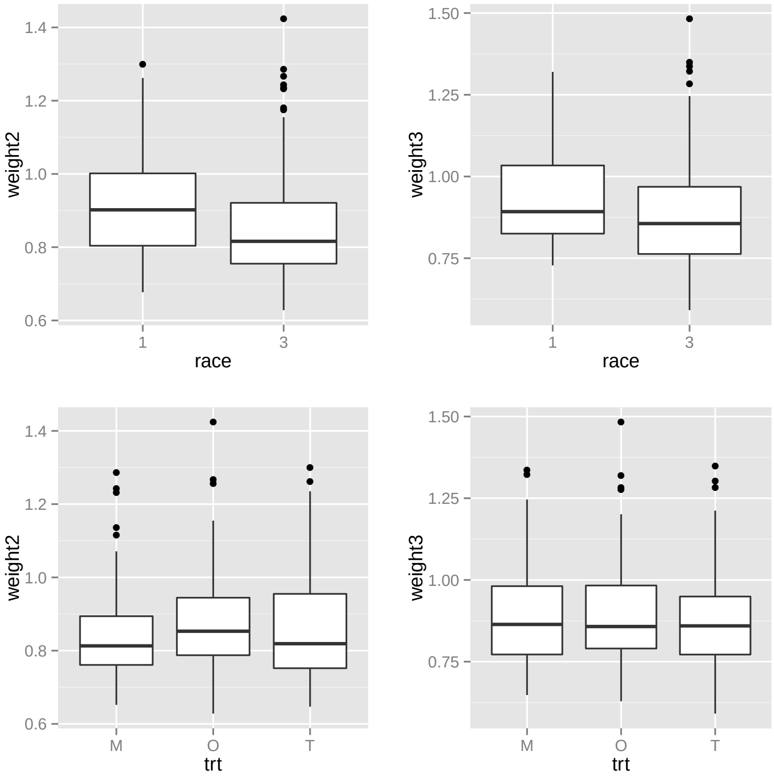

Figure 1: Boxplot of weights at 6th month (weight2) and 18th months (weight3) vs treatments (M, O, T) and race (black:1, white:3)

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 them with a semi-parametric model, for example, the Dirichlet process or Pólya tree. It would also be interesting to assume that mixture probabilities depend on covariates. Our future work will also include proposing a goodness of fit test to check 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.

Table 4: Marginal quantile regression coefficients for Weight2

	Intercept	Trt.O	Trt.T	Race.3
Weight2				
$\tau = 0.1$	0.80 (0.70, 0.86)	0.01 (-0.04, 0.07)	-0.01 (-0.06, 0.06)	-0.13 (-0.19, -0.04)
$\tau = 0.3$	0.83 (0.79, 0.92)	0.04 (-0.02, 0.07)	0.02 (-0.04, 0.05)	-0.07 (-0.16, -0.03)
$\tau = 0.5$	0.85 (0.82, 0.98)	0.05 (-0.03, 0.09)	0.04 (-0.06, 0.07)	-0.03 (-0.14, 0.00)
$\tau = 0.7$	0.95 (0.89, 1.03)	0.03 (-0.02, 0.10)	0.02 (-0.04, 0.08)	-0.04 (-0.12, 0.00)
$\tau = 0.9$	0.98 (0.94, 1.11)	0.07 (-0.02, 0.14)	0.06 (-0.02, 0.14)	-0.01 (-0.10, 0.05)
Weight3				
$\tau = 0.1$	0.78 (0.38, 0.84)	-0.01 (-0.07, 0.06)	-0.04 (-0.10, 0.04)	-0.13 (-0.18, -0.02)
$\tau = 0.3$	0.82 (0.78, 0.93)	0.01 (-0.04, 0.06)	-0.01 (-0.07, 0.05)	-0.06 (-0.16, -0.01)
$\tau = 0.5$	0.88 (0.84, 1.00)	0.02 (-0.06, 0.06)	0.02 (-0.08, 0.06)	-0.03 (-0.13, 0.02)
$\tau = 0.7$	0.99 (0.92, 1.07)	0.00 (-0.05, 0.08)	-0.00 (-0.07, 0.06)	-0.04 (-0.12, 0.01)
$\tau = 0.9$	1.02 (0.98, 1.16)	0.05 (-0.06, 0.11)	0.04 (-0.05, 0.12)	0.00 (-0.10, 0.06)

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.

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 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.

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 consider univariate case with two patterns. 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:

$$\begin{aligned} y|R = 1 &\sim N(\Delta + R_1, \sigma_1) \\ y|R = 0 &\sim N(\Delta + R_0, \sigma_0) \\ \Pr(R = 1) &= \pi \\ E(y) &= \theta \end{aligned}$$

Thus by iterated expectation, we have

$$\begin{aligned} \theta &= \Delta + R_1\pi + R_0(1-\pi) \\ \Delta &= \theta - \pi R_1 - (1-\pi)R_0 \end{aligned}$$

We can see Δ is deterministic by θ, R_1, R_0. If plugged in likelihood, we have

$$\begin{aligned} y|R = 1 &\sim N(\theta + (1-\pi)R_1 - (1-\pi)R_0, \sigma_1) \\ y|R = 0 &\sim N(\theta - \pi R_1 + \pi R_0, \sigma_0) \end{aligned}$$

Denote $\xi_1 = (\theta, R_1, R_0)$, and if $\xi_2 = (\theta, R_1 + 1, R_0 + 1)$, both 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 R_1 and R_0, for example $R_0 = -R_1$, then

$$\begin{aligned} y|R = 1 &\sim N(\theta + 2(1-\pi)R_1, \sigma_1) \\ y|R = 0 &\sim N(\theta - 2\pi R_1, \sigma_0) \end{aligned}$$

thus it 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 pattern mixture model:

$$\begin{aligned} y|R = 1 &\sim N(\Delta + R_1, \sigma_1) \\ y|R = 0 &\sim N(\Delta + R_0, \sigma_0) \\ \Pr(R = 1) &= \pi \\ \mathrm{p}(y \leq \theta) &= \tau \end{aligned}$$

where θ is the quantile estimate of interest and we does not include covariates so far. We will show $\boldsymbol{\xi} = (\theta, R_1, R_0)$ is not identifiable.

Again by iterated expectation, we have

$$\tau = \pi\Phi\left(\frac{\theta - \Delta - R_1}{\sigma_1}\right) + (1-\pi)\Phi\left(\frac{\theta - \Delta - R_0}{\sigma_0}\right)$$

thus Δ is again deterministic by other parameters:

$$\Delta = h(\theta, R_1, R_0, \sigma_1, \sigma_0, \pi, \tau)$$

To show $\boldsymbol{\xi} = (\theta, R_1, R_0, \sigma_1, \sigma_0)$ 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 $\theta^{'}, R_1^{'}, R_0^{'}$ such that

$$\begin{aligned} h(\boldsymbol{\xi}) + R_1 &= h(\boldsymbol{\xi}^{'}) + R_1^{'} \\ h(\boldsymbol{\xi}) + R_0 &= h(\boldsymbol{\xi}^{'}) + R_0^{'} \end{aligned}$$

By substracting previous equations, we have $R_1^{'} - R_0^{'} = R_1 - R_0$, thus denote $R_1^{'} = R_1 + \delta$ and $R_0^{'} = R_0 + \delta$, and let $\theta^{'} = \theta$ such that

$$\Delta^{'} = h(\theta^{'}, R_1, R_0, \sigma_1, \sigma_0, \delta) = h(\boldsymbol{\xi}) - \delta = \Delta - \delta$$

then the new parameter $\boldsymbol{\xi}^{'}$ yields the same distribution with one from $\boldsymbol{\xi}$. Therefore $\boldsymbol{\xi}$ is not identifiable.

Instead, if we put constraint, for example $R_1 = -R_0$, then by calculation, $\mathrm{p}(y|R;\boldsymbol{\xi}) = \mathrm{p}(y|R;\boldsymbol{\xi}^{'})$ yields $\boldsymbol{\xi} = \boldsymbol{\xi}^{'}$.

Now consider the case with covariates. Suppose the model is

$$\begin{aligned} y|R = 1, x &\sim N(\Delta + R_1 + \beta_{x1}x, \sigma_1) \\ y|R = 0, x &\sim N(\Delta - R_1 + \beta_{x0}x, \sigma_0) \\ \Pr(R = 1) &= \pi \\ \mathrm{p}(y \leq \gamma_0 + \gamma_1 x) &= \tau \end{aligned}$$

Still Δ can be determined by

$$\Delta = h(x, \gamma_0, \gamma_1, R_1, \beta_{x1}, \beta_{x0}, \sigma_1, \sigma_0, \pi, \tau)$$

We want to show parameter $\boldsymbol{\xi} = (\gamma_0, \gamma_1, R_1, \beta_{x1}, \beta_{x0}, \sigma_1, \sigma_0, \pi)$ is not identifiable by finding $\boldsymbol{\xi}^{'} \neq \boldsymbol{\xi}$, but $\mathrm{p}(y|R;\boldsymbol{\xi}) = \mathrm{p}(y|R;\boldsymbol{\xi}^{'})$. Still if the last equation holds, first we have $\sigma_1^{'} = \sigma_1, \sigma_0^{'} = \sigma_0$, then to equate the two means, we have

$$\begin{aligned} \Delta + R_1 + \beta_{x1}x &= \Delta^{'} + R_1^{'} + \beta_{x1}^{'}x \\ \Delta - R_1 + \beta_{x0}x &= \Delta^{'} - R_1^{'} + \beta_{x0}^{'}x \end{aligned}$$

By substracting the two equations, we have

$$2R_1 + (\beta_{x1} - \beta_{x0})x = 2R_1^{'} + (\beta_{x1}^{'} - \beta_{x0}^{'})x$$

which holds for all x. Thus $R_1 = R_1^{'}$ and $(\beta_{x1} - \beta_{x0}) = (\beta_{x1}^{'} - \beta_{x0}^{'})$. Then let

$$\begin{aligned} \beta_{x1}^{'} &= \beta_{x1} + \delta \\ \beta_{x0}^{'} &= \beta_{x0} + \delta \end{aligned}$$

and all the other parameters in $\boldsymbol{\xi}^{'}$ keep 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, especially for β_{x1} and β_{x0}. One solution is to restrict $\beta_{x1} = -\beta_{x0}$ to make all the parameters identifiable.

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

$$\begin{aligned} y_2|y_1, x, R = 1 &\sim N(\Delta + R_1 + x\beta_{x1} + \beta_{11}y_1, \sigma_1) \\ y_2|y_1, x, R = 0 &\sim N(\Delta - R_1 - x\beta_{x1} + \beta_{10}y_1, \sigma_0) \end{aligned}$$

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

we are wondering if β_{11} and β_{10} are identifiable, say if there exists $\beta_{11}^{'}$ and $\beta_{10}^{'}$, such that

$$\begin{aligned} \Delta + R_1 + x\beta_x + \beta_{11}y_1 &= \Delta^{'} + R_1^{'} + x\beta_x^{'} + \beta_{11}^{'}y_1 \\ \Delta - R_1 - x\beta_x + \beta_{10}y_1 &= \Delta^{'} - R_1^{'} - x\beta_x^{'} + \beta_{10}^{'}y_1 \end{aligned}$$

still by substracting two equations, we have $R_1 = R_1^{'}$ and $\beta_x = \beta_x^{'}$. Considering 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 in heterogeneous model described in section [ref], it is easy to show there is no trouble with heterogeneity parameters α, analog to the linear model case. For the other parameters, it can be found similar to the above discussion.

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

$$\begin{aligned} I(a,b) &= -\frac{1}{\sqrt{a^2+1}}\int_b^\infty \phi\left(\frac{s}{\sqrt{a^2+1}}\right)ds \\ &= \int_{b/\sqrt{a^2+1}}^\infty \phi(t)dt \\ &= 1 - \Phi(b/\sqrt{a^2+1}). \end{aligned} \tag{8}$$

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

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

- 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 (8)

- For integrating a N(b,a) CDF over another normal distribution (N(μ,σ)):

$$\begin{aligned}\int \Phi\left(\frac{x-b}{a}\right)d\Phi(x;\mu,\sigma) &= \int \Phi\left(\frac{x-b}{a}\right)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx \\ &= \int \Phi\left(\frac{\sigma y+\mu-b}{a}\right)\phi(y)dy \\ &= 1-\Phi\left(\frac{b-\mu}{\sigma}/\sqrt{\frac{a^2}{\sigma^2}+1}\right).\end{aligned} \tag{10}$$

  If $a < 0$,

$$\int \Phi\left(\frac{x-b}{a}\right)d\Phi(x;\mu,\sigma) = \Phi\left(\frac{b-\mu}{\sigma}/\sqrt{\frac{a^2}{\sigma^2}+1}\right). \tag{11}$$

C Maximum Likelihood Estimation Using Adaptive Gradient Descent Algorithm

See section 2.3.2 for notations.

The likelihood can be maximized via the following algorithms:

1. Initialize ξ

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 great than certain number, say 10^{-3}, 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 algorithm step 2. For example, when $j = 1$,

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